

# NCBI News

Last Updated: July 12, 2012



National Center for Biotechnology Information (US)  
Bethesda (MD)

National Center for Biotechnology Information (US), Bethesda (MD)

The contents of this newsletter may be reprinted without permission. The mention of trade names, commercial products, or organizations does not imply endorsement by NCBI, NIH, or the U.S. Government.

NLM Citation: NCBI News [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 1991-2012.

## Table of Contents

<b>NCBI News, May 2017</b> .....	1
NCBI to phase out support for non-human organism data in dbSNP and dbVar .....	1
Eleven eukaryotic annotations added to RefSeq in April 2017 .....	1
NCBI to assist with NYGC Genomics Hackathon June 19-21 .....	1
GenBank release 219.0 is available via FTP .....	3
May 10th NCBI Minute: How to Locate and Use Human Genomes and Annotations from NCBI .....	4
<b>NCBI News, April 2017</b> .....	5
April 26th NCBI Minute: Medical Genetics Summaries on the NCBI Bookshelf - a pharmacogenomics resource for clinicians .....	5
Maize ( <i>Zea mays</i> ) genome annotation release 101 is now available! .....	5
dbSNP FTP file format change planned for early 2018 .....	5
dbSNP's human build 150 has doubled the amount of RefSNP records! .....	6
NCBI researchers and collaborators discover novel group of giant viruses .....	6
April 19th NCBI Minute: Magic-BLAST, NCBI's next-gen sequence alignment program .....	8
Six functional prototypes available from the March NCBI hackathon .....	8
Eight new eukaryotic genome annotations added to RefSeq .....	9
New Genome Data Viewer access page .....	9
<b>NCBI News, March 2017</b> .....	11
Sequence Viewer 3.20 is now available .....	11
Conserved Domain Database (CDD) version 3.16 now available online and via FTP .....	11
NCBI to assist with BioFrontiers Hackathon in May .....	11
NCBI will attend the AACR Annual Meeting 2017 .....	12
Genome Workbench 2.11.10 now available .....	12
Tree Viewer version 1.13 implements new search in tree features .....	12
Complete RefSeq genome annotation results represented in UCSC genome browser .....	12
March 29th NCBI Minute: How to Submit Your 16S rRNA Data to NCBI .....	13
NCBI will attend the 2017 Annual Clinical Genetics Meeting .....	13
RefSeq release 81 now public .....	13
GenBank release 218.0 is now available .....	14
Seven new annotations added to RefSeq .....	15
Multiple Sequence Alignment Viewer 1.4 is now available .....	15
Magic-BLAST 1.2.0 now available .....	15

Expression section and bulk datasets added to NCBI Gene .....	15
<b>NCBI News, February 2017</b> .....	17
NCBI Insights   PubMed Citations: A New, Faster Process for Correcting Errors.....	17
March 1st NCBI Minute: Setting up new data alerts with MyNCBI.....	17
Bottlenose dolphin annotation release 101 .....	17
New video on YouTube: Embed the NCBI Sequence Viewer into Your Pages.....	18
NLM Webinar series: "Insider's Guide to Accessing NLM Data: EDirect for PubMed" .....	19
Tree Viewer version 1.12 implements new API to markup trees .....	19
Interim annotation updates for the human GRCh37p.13 and GRCh38.p10 assemblies.....	19
February 22nd webinar: Introducing the Multiple Sequence Alignment Viewer .....	20
SmartBLAST updated to provide more information, database matches.....	20
Sequence Viewer 3.19 is now available .....	21
New NCBI Insights post: New Web Services for Comparing and Grouping Sequence Variants.....	21
NCBI to host genomics hackathon March 20-22 .....	21
<b>NCBI News, January 2017</b> .....	23
Multiple Sequence Alignment Viewer 1.3 is now available.....	23
February 8th NCBI Minute: Finding Gene, Protein and Chemical Names, Aliases and Synonyms.....	23
BLAST+ 2.6.0 offers improved support for accession.version .....	23
New NCBI Insights post: Visualize and Interpret Alignment Data with the Multiple Sequence Alignment Viewer .....	23
January 31st NCBI Minute: New version of E-utilities supports accession.version.....	23
RefSeq release 80 now available; GI identifiers to be removed in next release (March 2017).....	24
New videos on YouTube: Clone DB and clone placements.....	24
GenBank release 217.0 is available via FTP.....	24
Genome Workbench 2.11.7 now available.....	25
<b>NCBI News, December 2016</b> .....	27
New YouTube video: Sequence Viewer: Display Translation Discrepancies.....	27
New NCBI Insights post: Converting Lots of GI Numbers to Accession.version .....	27
CCDS release 21 for mouse is public in Gene.....	27
December 21st NCBI Minute: Bulk Conversion of NCBI Sequence GI Identifiers to accession.version.....	28
Variant interpretations from Illumina double ClinVar data.....	28
The new Human Genome Resources site: a portal for exploration of the human genome.....	28
Sequence Viewer 3.18 is now available .....	29
New on NCBI Insights: Converting GI Numbers to Accession.version.....	29

NCBI Tech Talk and Booth at the American Society for Cell Biology 2016 National Meeting .....	29
<b>NCBI News, November 2016</b> .....	31
Evidence Viewer has been retired .....	31
NCBI, NLM, NHGRI to hold on-campus hackathon January 9-11 .....	31
Genome Workbench 2.11.5 now available .....	32
November 17th webinar: NCBI Resources for Agricultural Research .....	33
RefSeq release 79 now available .....	33
New NCBI Insights post: Identifying and Correlating Chemical Names and Synonyms .....	33
New NCBI Insights blog post: Clearing Up Confusion with Human Gene Symbols and Names .....	33
Permanent redirect to HTTPS will occur on November 10, 2016 .....	34
New video on YouTube: The New MSA Viewer .....	34
<b>NCBI News, October 2016</b> .....	35
November 9th webinar: PubMed for Clinicians .....	35
NLM In Focus blog profiles Dr. Kim Pruitt, NCBI staff scientist .....	35
Genome Workbench 2.11.0 now available .....	35
GI numbers will be removed from sequence record presentations .....	35
New YouTube video: NCBI Staff at ASHG 2016 .....	36
October 26th NCBI Minute: New BLAST Databases Provide Cleaner Results .....	36
Multiple Sequence Alignment Viewer 1.1 is now available .....	36
NLM presents "Insider's Guide" webinar on E-utilities and PubMed on October 19th .....	36
CCDS release 20 for human now public in Gene .....	37
<b>NCBI News, September 2016</b> .....	39
October 11th NCBI Minute: BLAST+ 2.5.0 with Support for HTTPS, accession.version Identifiers and Much More .....	39
BLAST+ 2.5.0 released with support for HTTPS, accession.version and more .....	39
October 5th webinar: NCBI at ASHG 2016 .....	39
Sequence Viewer 3.16 is now available .....	40
GenBank release 215.0 is now available via FTP .....	40
Genomes-announce listserv reactivated .....	40
NCBI to hold Developers' Forum September 28th .....	41
October 4-6: Stream the University of Michigan NCBI workshops .....	41
Introducing Magic-BLAST .....	41
Scheduled: Next Round of HTTPS Tests .....	41
NCBI's Bryant and Bolton receive 2016 Herman Skolnik Award for PubChem database .....	42

September 21st webinar: Update on NCBI's Transition to HTTPS .....	42
New video on YouTube: Tree Viewer - Display Large Trees.....	43
Genomes FTP site update (version 1.3) adds new data formats and more.....	43
RefSeq release 78 is now available.....	45
New on NCBI Insights: The Future of Existing GI Numbers at NCBI.....	45
Leiden Open Variation Database to be retired September 30, 2016 .....	45
New on NCBI Insights: Find, Browse and Follow Biomedical Literature with PubMed Journals.....	46
September 7th webinar: The E-Utilities in an Age without GI Numbers .....	46
October 24-26: Hackathon at Cold Spring Harbor Laboratory .....	46
<b>NCBI News, August 2016</b> .....	49
Genomes FTP site data organization to change on September 20, 2016 .....	49
dbVar July 2016 data release includes new 1000 Genomes Phase III structural variants.....	51
August 31st NCBI Minute: Downloading Genome Data from the NCBI FTP Site.....	51
VAST+ update provides refined alignments .....	51
September 12th class at NLM: EDirect - Command Line Access to NCBI's Biomolecular Databases .....	51
HIV-1 datasets in Gene updated .....	52
<b>NCBI News, July 2016</b> .....	53
HTTPS at NCBI: Guidance for NCBI web API users .....	53
dbSNP build 148 for corn, fruit fly, rice and 8 other organisms available.....	53
August 3rd webinar: NCBI Targeted Loci: RefSeq Ribosomal RNA Sequences for Identification and Phylogenetic Analysis....	53
Tree Viewer 1.10 visualizes large phylogenetic trees up to 100,000 nodes .....	53
NCBI Insights blog post: NCBI is Phasing out Sequence GIs - Here's What You Need to Know .....	54
July 27th NCBI Minute: Important Changes to NCBI Web Protocols.....	54
Sequence Viewer 3.15 is now available .....	54
July 20th NCBI Minute: Important Changes Coming to Sequence Databases.....	54
Conserved Domain Database (CDD) version 3.15 now available online and via FTP.....	55
RefSeq release 77 is now available.....	55
Mouse and zebrafish genome annotations updated .....	56
<b>NCBI News, June 2016</b> .....	59
Genome Workbench 2.10.7 now available.....	59
July 6th NCBI Minute: Quickly Find Coding Sequences Using ORFfinder.....	59
Human genome Annotation Release 108 incorporates new RefSeq sequences, predicts new variants .....	59
GenBank release 214.0 is now available via FTP .....	60

June 29th webinar: Downloading Exon and Coding Region Sequences for Genes .....	61
International HapMap Browser to go offline June 16, 2016.....	61
NCBI to hold hackathon on NIH campus in August.....	61
BLAST+ 2.4.0 now available.....	63
NCBI will transition to HTTPS on September 30, 2016.....	63
Tree Viewer 1.9 visualizes medium-large phylogenetic trees .....	63
June 10th webinar: Finding Systematic Reviews at PubMed Health and PubMed .....	64
MutaBind: Evaluating the effects of sequence variants and disease mutations on protein-protein interactions.....	64
Browse histones, analyze sequences with revamped HistoneDB 2.0.....	64
June 15th webinar: Using NCBI Resources and Variant Interpretation Tools for the Clinical Community.....	64
<b>NCBI News, May 2016</b> .....	67
Epigenomics database to be retired June 1, 2016 .....	67
NCBI launches new Twitter account for NCBI Bookshelf.....	67
New NCBI Insights blog post: Fast Sequence Inspection with ORFfinder and SmartBLAST (PubMed Labs).....	67
RefSeq release 76 is now available.....	68
Genome Browsers section added to Gene.....	68
Sequence Viewer version 3.14 upgrades platform.....	69
New NCBI Variation summary page highlights all organisms in dbSNP or dbVar with full assembly annotations .....	69
NCBI launches web-based iCn3D, a new viewer for 3D macromolecular structures .....	69
NCBI annotates 300th organism with the Eukaryotic Genome Annotation Pipeline.....	70
Genome Workbench 2.10.5 now available .....	70
Preview the new BLAST home page! .....	70
NCBI and RCSB PDB to assist ISCB in Sequence-Structure hackathon at ISMB Orlando 2016 .....	72
May 18th webinar: Using VDB BLAST Clients to Search Whole Genome Shotgun Contigs (WGS) and Transcriptome Shotgun Assembly (TSA) Data at the NCBI.....	73
New NCBI video on YouTube: ProSplign comes to Genome Workbench.....	73
New NCBI video on YouTube: Submitting BioSample Data to NCBI.....	74
GenBank release 213.0 is now available via FTP.....	74
<b>NCBI News, April 2016</b> .....	75
dbSNP build 147 data for human, chicken, soybean and more are available .....	75
Eukaryotic Genome Annotation Pipeline now directly annotates top-level sequences, not scaffolds.....	75
May 4th NCBI Minute: Linking PubMed and ClinicalTrials.gov .....	75
New NCBI video on YouTube: "Sequence Viewer: Display dbVar Supporting Calls" .....	76

Webinars on April 29 present BLAST, human variation & medical genetic records .....	76
NCBI to assist UC Davis in June hackathon .....	77
New NCBI video on YouTube: Navigating the NIH Manuscript Submission Process .....	78
Articles in Nucleic Acids Research Database 2016 Issue discuss NCBI databases, updates and future plans .....	79
Maximizing PubChem: webinar on April 20th will cover new and future features.....	80
<b>NCBI News, March 2016</b> .....	81
Specialized database with unique search interface added to Zika virus resource page .....	81
Register for the April 6th webinar: Using NCBI Databases with Tools that Predict Genomic Variant Effects .....	81
Register for the April 13th webinar, Submitting Data to NCBI and BioSample.....	81
New NCBI video on YouTube provides strategies to search ClinVar efficiently .....	82
RefSeq release 75 is now available.....	82
March 23, 2016: NCBI to offer workshop for advanced SRA and dbGaP users.....	82
Search for WGS Sequences using Stand-alone BLAST!.....	82
First of the New Bookshelf NCBI Insights Blog Posts - New Streptococcus pyogenes book.....	83
NCBI is phasing out sequence GIs - use Accession.Version instead!.....	83
Tree Viewer's Next Update is Available.....	86
<b>NCBI News, February 2016</b> .....	87
NCBI to assist Brandeis University in hosting Boston genomics hackathon in April.....	87
GenBank release 212.0 available via FTP .....	88
March 2nd webinar: NCBI Resources for Cancer Researchers .....	89
Zika virus resource page provides access to nucleotide, protein sequences from latest outbreak .....	89
dbSNP Build 146 for salmon, barrel medic, cottonwood and mouse now available.....	90
Rotavirus resource uses standardized metadata and annotations, suite of tools to make it easier to search, download and analyze sequences .....	90
New video on the NCBI YouTube channel: Eukaryotic Genome Data Curation at NCBI.....	91
NCBI Insights blog post: Professors: "NCBI can help you streamline your teaching and research efforts" .....	91
New video on the NCBI YouTube channel: Viral resources at NCBI .....	91
NCBI to assist Louisiana State University in South and Southeast regional genomics hackathon.....	92
Variation Viewer 1.5 adds facet toggling, updated backend data.....	93
February 17th webinar: "Five ways to submit next-gen sequencing data to NCBI's Sequence Read Archive (SRA)" .....	93
<b>NCBI News, January 2016</b> .....	95
Genome Workbench 2.10 now available.....	95
Sequence Viewer 3.11 now available .....	95



February 3rd webinar: "How to Upload and Analyze dbGaP Data in the Cloud" .....	95
RefSeq Release 74 now available on FTP.....	95
January 28th webinar: "Genomic Data Sharing with dbGaP: Registration and Submission" for IRP investigators.....	95
<b>NCBI News, December 2015</b> .....	97
NCBI staff will attend the International Plant and Animal Genome Conference XXIV in January .....	97
BLAST+ executables 2.3.0 now available .....	97
GenBank release 211.0 is now available via FTP .....	97
January 7th: Explore new graphical viewer track options with The NCBI Minute.....	98
January 5th webinar: Eukaryotic Genome Data Curation at NCBI .....	98
New on NCBI Insights blog: "The NCBI Minute: quick introductions to NCBI resources".....	99
dbSNP Build 146 for non-human organisms is now available .....	99
New NCBI Insights blog posts highlight SRA Toolkit, Run Selector.....	99
December 17th webinar: "Accessing 1000 Genomes Project Data" .....	100
Registration open for December 16th NCBI Minute: "New Faceted Advanced Search in dbGaP Provides Easy Access to Relevant Data" .....	100
<b>NCBI News, November 2015</b> .....	101
dbSNP human build 146 available through Entrez and FTP.....	101
Tree Viewer 1.7.5 now available .....	101
NCBI releases first five lectures of NCBI NOW on YouTube .....	101
December 2nd NCBI Minute webinar: Finding Genes in PubMed.....	101
New video on the NCBI YouTube channel: "Explore Gene pages at NCBI: Variation and Expression" .....	101
PubChem adds a legacy designation for outdated data.....	102
NCBI to hold three-day genomics hackathon in January .....	102
Sequence Viewer 3.10.5 adds support for track sets with non-default options .....	103
RefSeq Release 73 is now available.....	103
Tree Viewer 1.7 now available .....	103
Registration open for November 18 NCBI Minute, "The New ClinVar Submission Wizard" .....	103
Researchers identify potential alternative to CRISPR-Cas genome editing tools .....	104
dbVar publishes October 2015 data release .....	105
Registration open for November 12th webinar, "PubMed for Scientists" .....	106
<b>NCBI News, October 2015</b> .....	107
Variation Viewer 1.4.1 is now available with optimized Variant Filters and Table performance.....	107
New on the NCBI YouTube channel: "LinkOut - Linking to Datasets, Databases and More" .....	107

OSIRIS Version 2.5 is now available.....	107
Outdated Genomes FTP directories will be archived on November 30, 2015 .....	107
GenBank release 210.0 is now available via FTP .....	108
New on the NCBI YouTube channel: Learn how to view track sets and store track collections .....	109
Larger word size in modified algorithm speeds up BLASTP, BLASTX, TBLASTN search .....	109
Variation Viewer 1.4 is now available with faster filter performance, track sets & collections.....	109
Sequence Viewer 3.10 adds support for track sets and track collections, performance optimization and more.....	110
New NCBI Insights blog post: "Troubleshooting GenBank Submissions: Annotating the Coding Region (CDS)" .....	110
NCBI staff to attend and present at ASHG 2015 .....	111
<b>NCBI News, September 2015</b> .....	113
First offering of NCBI NOW (Next generation sequencing Online Workshop) to begin October 13, 2015 .....	113
September 30th NCBI Minute: Preview of NCBI at American Society of Human Genetics 2015.....	114
"Create a Biosketch with SciENCv" webinar recording on YouTube.....	114
October 2nd webinar - LinkOut: Linking to datasets, databases and more.....	115
New NCBI Insights blog post: "Finding Chemical Probes & Modulators - The Hunt for New Chemical Reagents and Medicines" .....	115
NCBI to hold fourth offering of "A Librarian's Guide to NCBI" .....	115
September 16th NCBI Minute: Accessing the Human Genomics Standard Data (Genome in a Bottle) at NCBI .....	116
New NCBI Insights blog post: "Identifying Chemical Targets - Finding Potential Cross-Reactions and Predicting Side Effects" .....	116
RefSeq Release 72 is now available .....	116
HIV-1 interaction datasets in Gene updated.....	117
Genome Workbench 2.9.5 now available.....	117
dbSNP build 145 (pig, chicken, sorghum, gibbon) now available.....	117
<b>NCBI News, August 2015</b> .....	119
NCBI annotates 250th eukaryote with Eukaryotic Genome Annotation Pipeline .....	119
September 2nd NCBI Minute: "Introducing SmartBLAST, a Rapid Protein Identification Tool" .....	119
GenBank release 209.0 is now available via FTP .....	119
Tree Viewer 1.6 now available .....	120
New NCBI video: "NCBI's 1000 Genomes Browser: Introduction" .....	120
August 26th webinar: "Troubleshooting GenBank Submissions: Determining and Annotating Coding Regions (CDS) for Eukaryotic Genes" .....	120
New NCBI Insights blog post: "SciENCv Updated to Support New NIH Biosketch Format" .....	121
Genomes FTP site update (version 1.2) expands taxonomic scope and more.....	121

CCDS release 19 for mouse added to Gene .....	122
<b>NCBI News, July 2015</b> .....	123
New NCBI Insights blog post: Introducing PubMed Labs, an NCBI initiative to include user community in product development from beginning .....	123
August 12th NCBI Minute: Using Variation Reporter to Map and Annotate Your Own Variant Calls .....	123
Sequence Viewer 3.9 adds data upload options to API, improved response time and more .....	123
August 5th NCBI Minute: "Using EDirect's Xtract Utility to Parse NCBI BLAST XML Output" .....	124
July 30th webinar: "Using SciENcv to Create Your NIH Biosketch" .....	124
July 22nd NCBI Minute webinar: Find disease-related variants in ClinVar .....	124
RefSeq Release 71 is now available! .....	124
July 15th webinar: "EDirect: Bringing the E-Utilities to the UNIX Command Line" .....	125
<b>NCBI News, June 2015</b> .....	127
Tree Viewer version 1.5 improves performance .....	127
June 3rd webinar "Troubleshooting GenBank Submissions: Coding Region Annotation" video up on YouTube .....	127
June 10th webinar "Phylogenetic Trees in Genome Workbench" video up on YouTube .....	127
New dbVar webinar available on NCBI YouTube channel .....	127
GenBank release 208.0 is now available via FTP .....	128
NCBI Southern California Regional Workshops to be held June 30 - July 2 .....	128
New YouTube video: "Sequence Viewer: Navigate Objects with Jump Arrows" .....	129
UniVec build 9.0 now available for VecScreen searches and FTP .....	129
BLAST+ stand-alone updated to version 2.2.31 .....	130
Complete MERS coronavirus genomes from China and South Korea are in GenBank .....	130
dbSNP build 144 now available .....	130
NCBI Sequence Viewer version 3.8 available .....	130
June 10th webinar: "Working with Phylogenetic Trees in Genome Workbench" .....	131
Conserved Domain Database (CDD) version 3.14 now available online and via FTP .....	131
The SRA Submission App on BaseSpace lets you submit directly to SRA .....	131
New NCBI YouTube video: "NCBI Minute: Prokaryotic Genome Annotation Update" .....	131
<b>NCBI News, May 2015</b> .....	133
New NCBI YouTube video: "Genome Workbench: Import BAMs and Export Alignments" .....	133
June 3rd webinar: "Troubleshooting GenBank Submissions: Coding Region Annotation" .....	133
NCBI to hold three-day genomics hackathon in August .....	133
New NCBI YouTube Video: NCBI's Tree Viewer .....	134

New NCBI Insights blog post: "NCBI's First Hackathon: Advanced Bioinformatic Analysis of Next-Gen Sequencing Data" ...	134
May 26th webinar: "The NCBI Minute: Prokaryotic Genome Annotation Update" .....	135
New NCBI Insights blog post: "NCBI RefSeq's Antimicrobial Peptide Indexed Field: Facilitating Novel Antibiotic Discovery" .....	135
Export data into Genome Workbench with Tree Viewer version 1.4.....	135
June 9th hands-on workshops at NLM will show users how to search NCBI's molecular databases.....	135
Genome Workbench 2.9.0 now available.....	136
New NCBI Insights blog post - Accessing the Hidden Kingdom: Fungal ITS Reference Sequences".....	136
RefSeq release 70 is now available with re-annotated bacterial genomes for uniformity across genomes and species .....	137
May 13th webinar: "Introducing dbVar, the NCBI Database of Large-Scale Genetic Variation .....	140
<b>NCBI News, April 2015</b> .....	141
May 6th webinar: "The NCBI Minute: Connecting with PubMed Commons".....	141
New NCBI Insights blog post: "NIHMS Users: Do You Know How Often Your Paper is Being Accessed via PMC?" .....	141
April 29th webinar: "The NCBI Minute: Finding Genomes and Annotations by Searching NCBI Assembly".....	142
April 21st webinar: Rebroadcast of "NCBI and the NIH Public Access Policy: PubMed Central Submissions, My NCBI, My Bibliography and SciENcv" .....	142
April 15th webinar: "The NCBI Minute: Finding and Getting the Data You Want from NCBI in Less than Three Minutes - Introducing BioProject" .....	142
NIH issued statement on use of dbGaP in the Cloud .....	142
<b>NCBI News, March 2015</b> .....	145
Updated human and mouse genome annotations now available .....	145
April 8th webinar: "The NCBI Minute: Introducing MOLE-BLAST" .....	145
April 1st webinar: "A Practical Guide to Using NCBI BLAST on the Web" .....	146
dbSNP Build 143 Phase II now available.....	146
New NCBI Insights blog post: "Exploring Entrez Direct: Parsing the XML Output of E-utilities".....	146
NCBI homepage update includes action buttons, category pages.....	146
NCBI Sequence Viewer version 3.6 available.....	147
March 18th webinar: "Using the dbGaP Data Browser to browse aligned reads and genotypes from the Database of Genotypes and Phenotypes".....	147
<b>NCBI News, February 2015</b> .....	151
March 5th webinar: "NCBI and the NIH Public Access Policy: PubMed Central submissions, My NCBI, My Bibliography and SciENcv" .....	151
"A Submitter's Guide to GenBank" webinar parts 1 and 2 on YouTube .....	151
NCBI Insights blog: How to delegate authority to others to edit/create your profile and Collections.....	152
NCBI webinar on February 25: The Next Generation of Access to Sequencing Data: Using NCBI's SRA Toolkit to Access Data from dbGaP and SRA .....	152

NCBI Genomes FTP site update adds analysis sets and other data .....	152
GenBank release 206.0 is now available via FTP .....	153
Mouse, cow and zebrafish added to dbSNP build 142 .....	154
1000 Genomes Browser updated to include Phase 3 May 2013 call set .....	154
<b>NCBI News, January 2015</b> .....	155
NIHMS's new look streamlines the manuscript submission process .....	155
Genome Workbench 2.8.10 available .....	155
Conserved Domain Database (CDD) version 3.13 now available online and via FTP .....	157
NCBI support for SOAP E-Utility ends July 1, 2015 .....	157
GenBank surpasses one trillion total bases of publicly available sequence data .....	157
Nucleic Acids Research Database 2015 Issue illustrates NCBI databases, updates and future plans .....	158
NCBI YouTube channel: A million views and counting! .....	158
NCBI's next webinar is The Statistics of Local Pairwise Sequence Alignment, Parts 1 and 2 .....	158
E-Utilities users: Keep up to date with changes via the Gene database RSS feed .....	159
RefSeq release 69 available on FTP .....	159
NCBI annotates 200th eukaryote .....	160
NCBI staff will attend International Plant and Animal Genome Conference XXIII .....	160
<b>NCBI News, December 2014</b> .....	161
NCBI webinar A Submitter's Guide to GenBank, Part 2 on January 7th .....	161
GenBank release 205.0 is now available via FTP .....	161
Bald eagle and other bird genome sequence and annotation data publicly available at NCBI .....	161
Citation Exporter Feature Now Available in PubMed Central .....	163
New NCBI Insights blog post: Designing exon-specific primers for the human genome .....	164
<b>NCBI News, November 2014</b> .....	165
NCBI to hold two-day genomics hackathon in January .....	165
NCBI BioSample includes curated list of over 400 known misidentified and contaminated cell lines .....	165
NCBI Eukaryotic Genome Annotation Pipeline breaks record; over 100 organisms annotated this year .....	166
NCBI BankIt webinar on December 17th .....	166
NCBI E-Utilities webinar video now on YouTube .....	166
BLAST URL domain change in effect December 1 .....	167
RefSeq release 68 available on FTP .....	167
dbVar releases 1000 Genomes Phase 3 structural variants .....	168
dbVar releases copy number variation (CNV) data from developmental delay study cited in Nature Reviews Genetics .....	168

<b>NCBI News, October 2014</b> .....	169
BLAST+ 2.2.30 released.....	169
New Genome BLAST selector on the BLAST homepage.....	169
Next NCBI webinar on November 5th.....	169
GenBank release 204.0 is now available via FTP.....	170
dbSNP human Build 142 released.....	171
Amazon Web Services (AWS) Marketplace provides the easiest way to start an NCBI BLAST instance.....	171
Variation Reporter version 1.4 released.....	172
Conserved Domain Database (CDD) version 3.12.....	172
Updates to assembly alignments for NCBI Remap service.....	172
New NCBI Insights blog: Sequence updates in human assembly GRCh38: improving gene annotation.....	173
Zebrafish (Danio rerio) GRCz10 now annotated.....	173
dbVar now accepts VCF submissions of structural variation data.....	173
New NCBI Insights blog post: NCBI's medical genetics resources.....	173
NCBI webinar on E-Utilities October 15th.....	174
<b>NCBI News, September 2014</b> .....	175
NCBI Sequence Viewer version 3.4 available.....	175
HIV-1, human interaction database updated.....	175
Virus Variation Resource pages for Ebolavirus, MERS coronavirus give quick and easy access to related sequences and other data.....	177
Simplified FASTA headers included on new NCBI Genomes FTP site.....	178
RefSeq release 67 available on FTP.....	178
Identical Protein Report Display option added to Protein database.....	179
<b>NCBI News, August 2014</b> .....	181
Milestone: NCBI annotates 150th eukaryotic genome.....	181
The new NCBI Genomes FTP site is here!.....	181
New NCBI YouTube video: Downloading FASTA sequences in Sequence Viewer.....	182
Rat annotation release 105 now on Gene, FTP, sequence and BLAST databases.....	182
New NCBI Insights blog: How to comply with NIH Public Access Policy.....	183
"BLAST in the Cloud!" is the newest video on the NCBI Webinars YouTube playlist.....	183
GenBank release 203.0 is now available via FTP.....	183
Genome Workbench 2.8.0 released.....	184
"NCBI's OSIRIS: Quality Assurance for DNA Forensic Profiling" webinar on September 17th.....	184

UniVec build 8.0 now available for VecScreen searches and FTP.....	184
<b>NCBI News, July 2014</b> .....	187
General Research Use collection streamlines access to patient-level data in dbGaP.....	187
NCBI/CDC/FDA/USDA collaboration using whole genome sequencing (WGS) to improve food safety is honored with an HHSinnovates award.....	187
Major revision of the NCBI genomes FTP site this summer.....	188
NCBI webinar "Using the New NCBI Variation Viewer to Explore Human Genetic Variation" on August 13th.....	189
RefSeq release 66 available on FTP site.....	189
NCBI's latest YouTube video presents special features in SciENcv.....	189
"BLAST in the Cloud!" webinar on July 30th showcases NCBI-BLAST Amazon Machine Image.....	190
<b>NCBI News, June 2014</b> .....	191
BLAST machine image (AMI) hosted at Amazon Web Services.....	191
Green monkey annotation release 100 now available.....	191
NCBI's latest YouTube video explores Variation Viewer.....	191
GenBank release 202.0 is now available via FTP.....	191
RefSeq model sequences can now be constructed from genomic and transcript sequences.....	192
Genome Workbench 2.7.19 released.....	192
dbSNP human Build 141 now available.....	192
New features simplify access to annotation information in NCBI's Gene.....	193
<b>NCBI News, May 2014</b> .....	195
BLAST URL domain changes to take effect December 1, 2014.....	195
GTR/ClinVar/MedGen webinar on June 18 will explore NCBI resources.....	195
RefSeq release 65 available on FTP site.....	195
The NIH Genetic Testing Registry now has information on more than 3,500 cancer tests.....	195
SciENcv 2.0 brings major improvements to My NCBI.....	196
NCBI Sequence Viewer version 3.2 available.....	197
<b>NCBI News, April 2014</b> .....	199
Coffee Break tutorial: The promise of PCSK9.....	199
New NCBI Insights Blog: Sequence updates in human genome assembly GRCh38.....	199
HomoloGene release 68 now available.....	199
Milestone: NCBI's Taxonomy database contains over 300,000 species with formal scientific names!.....	199
GenBank release 201.0 is now available via FTP.....	199
Enterococci: From commensals to leading causes of drug resistant infection on Bookshelf.....	200

CCDS release 16 for mouse now public in Gene.....	200
New on the Bookshelf: The Art and Politics of Science, a memoir by Dr. Harold Varmus.....	201
Coffee Break tutorial: Brown fat and obesity.....	201
<b>NCBI News, March 2014</b> .....	203
New NCBI YouTube video: Create custom databases for BLAST.....	203
Come to the NCBI Discovery Workshops on May 6th & 7th!.....	203
NCBI will attend the 2014 ACMG Annual Clinical Genetics Meeting.....	203
NCBI requests feedback on proposed BLAST XML specification update.....	203
RefSeq full release 64 out.....	204
Orthologous genes and gene regions now accessible through Gene.....	204
New dbGaP online system for registering studies and applying for data access introduces time-saving features.....	205
New Sorting and Output Options for E-utilities.....	205
<b>NCBI News, February 2014</b> .....	209
Genome Workbench Update 2.7.15 released.....	209
New CDD Release v.3.11 includes recomputed PSSMs and more.....	209
GenBank has milestone 200th release.....	209
Human genome annotation release 106 now available.....	210
NCBI releases Entrez Direct, the Entrez utilities on the UNIX command line.....	210
Sequence Viewer updated to version 3.1.....	211
<b>NCBI News, January 2014</b> .....	215
Human CCDS release 15 now available on web and FTP.....	215
RefSeq release 63 now available.....	215
Taxonomy database now shows type material, sequences from type specimens and strains now labeled in Entrez.....	215
New NCBI Insights blog: NCBI Remap tool helps you transition to newest human reference genome assembly, GRCh38.....	216
Sequence Viewer PDF rendering available - YouTube video tutorial.....	216
Genome Workbench Update 2.7.12.....	216
VAST+ released: Find similar 3D structures for macromolecular complexes.....	218
NCBI Insights blog: A Librarian's Guide to NCBI - an intensive training course for medical librarians to be offered April 2014.....	219
Mouse genome annotation release 104 available.....	220
BLAST+ 2.2.29 now available.....	220
<b>NCBI News, December 2013</b> .....	221
New human genome assembly (GRCh38) released!.....	221



Annotation reports now generated for recently annotated organisms .....	221
Meet PubMed Commons: The new comments forum in PubMed .....	221
Rat genome annotation release 104 .....	222
New NCBI Handbook chapters: Eukaryotic and prokaryotic genome annotation pipelines .....	222
Sequence Viewer has been updated .....	223
GenBank release 199 now available .....	223
NCBI Video: Submitting manuscripts on NIHMS .....	223
PMCID - PMID - Manuscript ID - DOI Converter Upgraded .....	224
<b>NCBI News, November 2013</b> .....	225
NCBI Insights blog post: Creating custom BLAST databases .....	225
SRA milestone: Over 2 petabases of sequence data .....	225
Planned change in bacterial strain-level information management .....	225
Exploring next-gen sequencing experiments with SRA-BLAST .....	227
NCBI Insights blog post: Saved Searches and E-mail Alerts .....	228
RefSeq release 62 now available .....	228
NCBI's Eukaryotic Annotation Pipeline has now annotated the genomes for 100 different organisms! .....	228
NCBI's 25th Anniversary and The Jim Gray eScience Award .....	229
New SNP data available for several organisms! .....	230
Update on PubMed Commons' comments in the early pilot phase .....	230
<b>NCBI News, October 2013</b> .....	233
Human CCDS release 14 is now available in the Gene database .....	233
New NCBI Insights Blog Post: Joining PubMed Commons - A step-by-step guide .....	233
GenBank Release 198.0 is Available .....	233
PubMed Commons is now live! .....	234
NCBI Staff will be attending the ASHG 2013 Meeting .....	234
Organism BLAST pages now use top-level RefSeq genomic records instead of scaffold records .....	236
<b>NCBI News, September 2013</b> .....	237
Try the new My NCBI Feature: SciENcv .....	237
Comments Requested: NIH genomic data sharing policy .....	237
RefSeq release 61 now available .....	238
A new NCBI Insights post about the use of NCBI Data for scientific discovery .....	238
New PubChem social media sites help keep users up-to-date! .....	238
<b>NCBI News, August 2013</b> .....	239

Human genome annotation release 105 with new splice variants.....	239
dbSNP Build 138, phase III, now available.....	239
Sequence Viewer 2.27: new features, improvements, and help documentation.....	239
>10,000 tests now listed in the NIH Genetic Testing Registry.....	239
GenBank Release 197.0 is Available.....	240
<b>NCBI News, July 2013</b> .....	243
Tenth Anniversary of RefSeq FTP Releases.....	243
RefSeq Release 60 is Available for FTP.....	244
New NCBI Insights Post: "New Pandoravirus Sequences are Accessible in GenBank".....	245
Genome Workbench 7.6 with Publication Quality Graphics Export.....	246
<b>NCBI News, June 2013</b> .....	249
Come to the NCBI Discovery Workshops on July 30 & 31!.....	249
Upload and graphically compare your own data with NCBI Epigenomics tracks.....	249
SRA-BLAST has been updated with new features and performance enhancements.....	250
Welcome to the NCBI News site!.....	250
Dr. David Lipman Receives White House "Open Science" Champions of Change Award on Behalf of NCBI.....	250
GenBank Release 196.0 is Available.....	251
New RefSeq Bacterial Protein Products and Emerging RefSeq Data Model.....	253
<b>NCBI News, May 2013</b> .....	255
Need to Find Information about Genetic Tests? Try GTR!.....	255
RefSeq Release 59 is Available for FTP.....	255
New YouTube Video: Complying with the NIH Public Access Policy with My Bibliography.....	255
<b>NCBI News, April 2013</b> .....	257
New publication: "BLAST: a more efficient report with usability improvements.".....	257
"A Librarian's Guide to NCBI" Course was a Success!.....	257
GenBank Release 195.0 is Available.....	259
New Educational Initiative: A Librarian's Guide to NCBI.....	259
PubChem Releases New and Enhanced Webpage Widgets.....	261
BLAST 2.2.28 now available.....	262
Try it out! The New PubChem Upload Beta Site.....	262
New database options in Microbial Genomes BLAST: Representative Genomes.....	262
<b>NCBI News, March 2013</b> .....	265

New CDD Release v3.10 Includes an Updated PSSM Calculation .....	265
NCBI Presents Genetic Variation and Medical Resources at the ACMG 2013 Meeting .....	265
Genome Workbench 2.7.0 Now Available .....	265
RefSeq Release 58 is Available for FTP .....	267
NCBI now provides interim GFF-formatted updates for human and mouse refseq annotations .....	267
Genome Workbench is the Featured Resource in OpenHelix's "Tip of the Week" .....	267
New Quick Tip on NCBI Insights Blog - How To Format Sequence Data For GenBank Submissions .....	268
<b>NCBI News, February 2013</b> .....	269
New Science Feature on NCBI Insights - Transcriptome of Tasmanian devil and its transmissible cancer .....	269
New Quick Tip on NCBI Insights Blog - how to download bacterial genomes using the Entrez API .....	269
GenBank Release 194.0 is Available .....	269
NCBI Insights' First Quick Tip: How to find functional protein homologs using conserved domains .....	271
New NCBI Insights Blog Explains the IE7 Warning .....	272
PubReader Article View Now In Use By KoreaMed Synapse .....	272
<b>NCBI News, January 2013</b> .....	275
First NCBI Blog Post Highlights New PubReader For PMC Articles .....	275
Now Available: NCBI Insights Blog! .....	275
~300,000 ChemAxon Structures are now in PubChem .....	275
Genetic Testing Registry Records will list Molecular Pathology CPT Codes .....	276
Come to the NCBI Discovery Workshops on February 4 & 5! .....	276
Now in GenBank: Flu Sequences from the Current Influenza Season .....	276
NIH Online Magazine features NCBI Researcher Teresa Przytycka .....	277
RefSeq Release 57 is Available for FTP .....	278
A New Eukaryotic Genome Annotation Status Page Keeps Researchers Informed .....	278
New Rat Genome Available in the MapViewer .....	278
<b>NCBI News, December 2012</b> .....	281
NCBI Introduces PubReader, a New View for Full-Text Articles .....	281
GenBank Release 193.0 is Available .....	281
Now in PubChem: 8+ million Patented Chemicals from the SureChem Database .....	283
SRA Surpasses a PetaBase of Sequence Data .....	283
<b>NCBI News, November 2012</b> .....	285
New version of Genome Workbench is Available .....	285
RefSeq Release 56 is Available for FTP .....	286

New CDD release Available & now Mirrors TIGRFAM v13.....	286
NCBI will be Presenting and Exhibiting at ASHG 2012 .....	286
<b>NCBI News, October 2012</b> .....	289
Human CCDS Release 11 Issued.....	289
NCBI's Genetic Testing Registry at AMP's Annual Genomic Medicine Meeting .....	289
NCBI Genetic Counselors at the National Society for Genetic Counselors' Meeting .....	289
New dbSNP Release for Mouse and Cow.....	290
NLM-Funded Investigator Creates First Complete Computerized Simulation of an Organism .....	290
Try the new PubChem Classification Browser!.....	291
New PubChem Widgets: embed PubChem tables in your own web pages! .....	291
<b>NCBI News, September 2012</b> .....	293
How has the SRA grown!.....	293
RefSeq Release 55 is out!.....	293
Now in PubChem: >6 million chemicals from SCRIPODB with links to USPTO patents .....	293
New CDD release contains 239 new or updated NCBI-curated domains .....	293
PubChem databases and services are now HTTPS compatible .....	294
PubChem's PUG REST 1.0 is now available!.....	294
The BLAST+ User Manual has been revised & updated.....	294
Stand-alone BLAST has been updated.....	294
PubChem reaches milestones on its 8th BDay!.....	295
A new version of Genome Workbench is available.....	295
NCBI is now using Genome Annotation Release numbers.....	295
<b>NCBI News, July 2012</b> .....	297
Registration now open for NCBI Discovery Workshops September 4-5 at NLM .....	297
1000 Genomes Dataset Browser.....	297
PubMed News.....	300
BLAST News .....	301
New HomoloGene Build: Rhesus macaque now included.....	304
Microbial Genomes Update .....	304
GenBank News.....	305
RefSeq News .....	305
GRC Plans New Human Genome Build and Requests Input .....	305
NCBI Now Offers IPv6 Access .....	305

Keeping Up with NCBI.....	305
<b>NCBI News, April 2012</b> .....	307
NCBI Discovery Workshops May 15-16 at NLM: Seats still available.....	307
Assembly: a Companion to the Genome Database.....	307
New Videos on NCBI's YouTube Channel.....	307
The Genetic Testing Registry: Finding Genetic Tests and Related Information.....	310
BLAST News.....	311
Remap and Variation Reporter: Two New Services for Mapping Locations onto Genome Builds.....	314
NCBI Aspera Download Site Available for NCBI Databases and Tools.....	315
1000 Genomes Project Data Now on Amazon Cloud Service.....	315
Microbial Genomes Update.....	315
NCBI Articles in Nucleic Acids Research Database Issue.....	315
GenBank News.....	319
RefSeq News.....	319
Keeping Up with NCBI.....	319
<b>NCBI News, November 2011</b> .....	321
Phase One Rollout of the New Genome Site.....	321
New BLAST videos on NCBI's YouTube channel.....	321
Entrez Utility Changes: New EFetch Version and New alternative ESummary XML.....	323
Highlight Features Link Now on Sequence Records.....	324
New BLAST 16S Prokaryotic Ribosomal RNA Database.....	325
New Phenotype-Genotype Integrator (PheGenI).....	326
Eukaryotic Genome Builds and Updates.....	326
Human Genome Update.....	328
Microbial Genomes Update.....	328
GenBank News.....	328
RefSeq News.....	328
Conserved Domain Database Update.....	329
Sequin Now with Transcriptome Shotgun Assembly and Internal Transcribed Spacer Sequence Submission Wizards.....	329
NCBI C++ Toolkit Major Release.....	330
Announce Lists and RSS Feeds.....	331
<b>NCBI News, August 2011</b> .....	333
NCBI Discovery Workshops September 27-28 at NLM: Seats still available.....	333

Feature Highlight Now Available in Sequence Databases .....	333
New videos on NCBI's YouTube channel .....	335
Updated Genome Workbench (v2.4.0) .....	336
Conserved Domain Database updated (v2.31) .....	336
Microbial Genomes Update .....	336
GenBank News .....	337
RefSeq News .....	337
NCBI will no longer archive new sequencing data from The Cancer Genome Atlas (TCGA) .....	337
The Growth of PubChem .....	337
New Simple Object Access Protocol (SOAP)-based BLAST service .....	337
NCBI at the ICHG/ASHG Meeting in Montreal: Workshop on Medical Genetics .....	337
Announce Lists and RSS Feeds .....	338
<b>NCBI News, June 2011</b> .....	<b>339</b>
Featured Resource: Re-designed PopSet .....	339
New My NCBI Interface .....	340
Transcriptome Shotgun Assembly (TSA) Database Available for BLAST .....	344
New Attributes for Human Variants in dbSNP .....	344
Updated BLAST Genome Search Pages .....	346
NLM Contest: Show off your Apps! Invitation to Submit Applications that Work with NLM Biomedical Data .....	346
New Videos on NCBI's YouTube Channel .....	346
The Sequence Read and Trace Archive Databases to Continue .....	346
BLAST 2.2.25+ Release and New Set-up Instructions .....	347
Microbial Genomes Update .....	347
RefSeq News .....	347
GenBank News .....	347
NCBI Discovery Workshops at Washington University: July 26-27, 2011 .....	347
Announce Lists and RSS Feeds .....	347
<b>NCBI News, March 2011</b> .....	<b>349</b>
PubMed Interface for Mobile Devices Now Available .....	349
NCBI Bookshelf Updated to the New Entrez Design .....	349
New Organism Builds in UniGene .....	350
NCBI YouTube Video Update .....	350
RefSeq News .....	352

GenBank News .....	352
Microbial Genomes Update .....	352
Mouse Genome Annotation Release (build 37.2) and Updated Mouse Consensus Coding Sequence (CCDS) Data .....	352
HomoloGene Release 65 Now Available.....	353
Genome Workbench Version 2.2.2 Release.....	353
NCBI Responds to a Report of Contamination in the Sequence Databases .....	353
NCBI Discontinues the Short Read Archive, Trace Archive, and Peptidome .....	353
<b>NCBI News, January 2011</b> .....	<b>357</b>
NCBI Discovery Workshops: Feb 15-16, 2011 .....	357
Updated Resources for Genomic Libraries and Clones.....	357
New Gene-BioSystems Links Highlight the Gene in the Biological Pathway.....	357
New Organisms in UniGene .....	361
NCBI Databases in Nucleic Acids Research Database Issue .....	361
dbSNP BLAST Pages Updated.....	361
New Mammalian Genomes at NCBI.....	361
Microbial Genomes Update .....	361
New Video on NCBI's YouTube Channel.....	362
GenBank News.....	362
RefSeq News .....	362
Journals Database Now a Part of NLM Catalog .....	362
Announce Lists and RSS Feeds.....	362
<b>NCBI News, October 2010</b> .....	<b>363</b>
New Databases and Tools.....	363
GenBank News.....	365
Updates and Enhancements .....	366
Announce Lists and RSS Feeds.....	370
<b>NCBI News, September 2010</b> .....	<b>371</b>
New Databases and Tools.....	371
GenBank News.....	375
Updates and Enhancements .....	375
<b>NCBI News, August 2010</b> .....	<b>377</b>
New Databases and Tools.....	377
GenBank News.....	377

Updates and Enhancements .....	377
Announce Lists and RSS Feeds .....	377
<b>NCBI News, July 2010</b> .....	<b>379</b>
Featured Resource: Updated Entrez Sequence Database Interfaces .....	379
New Databases and Tools .....	383
GenBank News .....	385
Updates and Enhancements .....	385
Announce Lists and RSS Feeds .....	386
<b>NCBI News, June 2010</b> .....	<b>389</b>
New Databases and Tools .....	389
GenBank News .....	391
Updates and Enhancements .....	391
Announce Lists and RSS Feeds .....	392
<b>NCBI News, May 2010</b> .....	<b>393</b>
New Databases and Tools .....	393
GenBank News .....	393
Updates and Enhancements .....	393
Announce Lists and RSS Feeds .....	394
<b>NCBI News, April 2010</b> .....	<b>395</b>
New Databases and Tools .....	395
GenBank News .....	395
Updates and Enhancements .....	395
Announce Lists and RSS Feeds .....	396
<b>NCBI News, March 2010</b> .....	<b>397</b>
New Databases and Tools .....	397
GenBank News .....	397
Updates and Enhancements .....	397
Announce Lists and RSS Feeds .....	398
<b>NCBI News, February 2010</b> .....	<b>399</b>
New Databases and Tools .....	399
GenBank News .....	399
Updates and Enhancements .....	400



Announce Lists and RSS Feeds.....	400
<b>NCBI News, January 2010</b> .....	401
New Databases and Tools.....	401
GenBank News.....	402
Updates and Enhancements.....	402
Announce Lists and RSS Feeds.....	402
<b>NCBI News, December 2009</b> .....	405
New Databases and Tools.....	405
GenBank News.....	405
Updates and Enhancements.....	405
Announce Lists and RSS Feeds.....	406
<b>NCBI News, November 2009</b> .....	407
Featured Resource: New Discovery-oriented PubMed and NCBI Homepage.....	407
New Databases and Tools.....	414
GenBank News.....	414
Updates and Enhancements.....	414
Announce Lists and RSS Feeds.....	415
<b>NCBI News, October 2009</b> .....	417
New Databases and Tools.....	417
GenBank News.....	417
Updates and Enhancements.....	417
Announce Lists and RSS Feeds.....	418
<b>NCBI News, September 2009</b> .....	419
Featured Resource: The Genome Reference Consortium Human Genome Build 37 now Available.....	419
New Databases and Tools.....	424
GenBank News.....	424
Updates and Enhancements.....	425
Announce Lists and RSS Feeds.....	425
<b>NCBI News, August 2009</b> .....	427
Featured Resource: The NCBI Short Read Archive (SRA) of Next- Generation Sequencing Data.....	427
New Databases and Tools.....	431
GenBank News.....	431

Updates and Enhancements .....	431
Announce Lists and RSS Feeds .....	432
<b>NCBI News, July 2009</b> .....	433
Featured Resource: The BioSystems Database of Biological Pathways .....	433
New Databases and Tools .....	439
GenBank News .....	439
Updates and Enhancements .....	440
Announce Lists and RSS Feeds .....	440
<b>NCBI News, June 2009</b> .....	441
Featured Resource: An Expanded Set of Discovery Components in the Entrez System .....	441
New Databases and Tools .....	445
GenBank News .....	447
Updates and Enhancements .....	447
Announce Lists and RSS Feeds .....	447
<b>NCBI News, May 2009</b> .....	449
Featured Data: 2009 H1N1 Influenza Sequences .....	449
Featured Resource: Protein Multiple Alignment Tool Web Service .....	452
New Databases and Tools .....	454
GenBank News .....	456
Updates and Enhancements .....	456
Announce Lists and RSS Feeds .....	456
References .....	457
<b>NCBI News, April 2009</b> .....	459
Featured Resource: PubChem Now Offers 3-D Small Molecule Structures and a New Conformer Viewer (Pc3D) .....	459
New Databases and Tools .....	461
GenBank News .....	464
Updates and Enhancements .....	464
Announce Lists and RSS Feeds .....	464
<b>NCBI News, March 2009</b> .....	467
Featured Resource: The New Entrez Sequence View has an Emphasis on Discovery .....	467
New Databases and Tools .....	469
GenBank News .....	469

Updates and Enhancements .....	469
Announce Lists and RSS Feeds .....	470
<b>NCBI News, February 2009</b> .....	471
Featured Resource: Improvements to NCBI Services Promote Discovery .....	471
New Databases and Tools .....	472
GenBank News .....	472
Updates and Enhancements .....	473
Announce Lists and RSS Feeds .....	474
<b>NCBI News, January 2009</b> .....	475
Featured Resource: BLAST+, All New BLAST Available on Web Service and for Download .....	475
New Databases and Tools .....	478
GenBank News .....	478
Updates and Enhancements .....	479
Announce Lists and RSS Feeds .....	479
<b>NCBI News, December 2008</b> .....	481
Featured Resource: BLAST 2 Sequences Is Now Part of the Main BLAST Web Service .....	481
New Databases and Tools .....	482
GenBank News .....	483
Updates and Enhancements .....	483
Announce Lists and RSS Feeds .....	484
<b>NCBI News, November 2008</b> .....	485
Featured Resource: Primer-BLAST—NCBI’s Primer Designer and Specificity Checker .....	485
New Databases and Tools .....	488
GenBank News .....	488
Updates and Enhancements .....	489
Announce Lists and RSS Feeds .....	489
<b>NCBI News, August 2008</b> .....	491
Featured Resource: New Graphical Sequence Viewer .....	491
New Databases and Tools .....	492
GenBank News .....	493
Updates and Enhancements .....	493
Announce Lists and RSS Feeds .....	493

<b>Archive Issues, 1994-2008</b> .....	495
January 2008 .....	495
Fall/Winter 2006/07 .....	495
Summer 2006 .....	495
November 2005 .....	495
May 2005 .....	495
Summer/Fall 2004 .....	495
Spring 2004 .....	495
Fall/Winter 2003 .....	496
Summer 2003 .....	496
Spring 2003 .....	496
Fall/Winter 2002 .....	496
Summer 2002 .....	496
Spring 2002 .....	496
Winter 2001 .....	496
Fall 2001 .....	496
Spring 2001 .....	496
Fall / Winter 2000 .....	497
Summer 2000 .....	497
Spring 2000 .....	497
Winter 2000 .....	497
Fall 1999 .....	497
Summer 1999 .....	497
Spring 1999 .....	497
Winter 1999 .....	497
November 1998 .....	497
July 1998 .....	498
February 1998 .....	498
August 1997 .....	498
August 1996 .....	498
March 1996 .....	498
September 1995 .....	498
March 1995 .....	498

August 1994 .....	498
February 1994 .....	498



## NCBI News, May 2017

### NCBI to phase out support for non-human organism data in dbSNP and dbVar

Tuesday, May 09, 2017

Starting September 1, 2017, NCBI will not accept non-human variant data submissions to dbSNP and dbVar. Any non-human data that is already in the databases or that is submitted before September 1, 2017 will continue to be available via the [dbSNP](#) and [dbVar](#) FTP download sites.

NCBI will phase out support for non-human organisms in dbSNP and dbVar following this timeline:

- September 1, 2017 – dbSNP and dbVar stop accepting submissions of non-human variant data.
- November 1, 2017 – dbSNP and dbVar interactive websites and related NCBI services stop presenting non-human variant data. The data will, however, continue to be available for download on the dbSNP and dbVar FTP sites.

We would like to thank all the submitters and users who have supported dbSNP and dbVar throughout the years. If you want to submit non-human variation data now or after September 1, 2017, European Bioinformatics Institute (EBI) – one of our partners in the International Nucleotide Sequence Database Collaboration (INSDC) – is accepting these data in the [European Variation Archive](#).

### Eleven eukaryotic annotations added to RefSeq in April 2017

Monday, May 08, 2017

In April, the [NCBI Eukaryotic Genome Annotation Pipeline](#) released new annotations in RefSeq for the following eleven organisms:

- *Bombus terrestris* (buff-tailed bumblebee)
- *Ceratitidis capitata* (Mediterranean fruit fly)
- *Athalia rosae* (coleseed sawfly)
- *Dendrobium catenatum* (a monocot)
- *Phalaenopsis equestris* (a monocot)
- *Orbicella faveolata* (stony coral)
- *Pogona vitticeps* (central bearded dragon)
- *Oryzias latipes* (Japanese medaka)
- *Sesamum indicum* (sesame)
- *Jatropha curcas* (a eudicot)
- *Amborella trichopoda* (a flowering plant)

See more details on the [Eukaryotic RefSeq Genome Annotation Status page](#).

### NCBI to assist with NYGC Genomics Hackathon June 19-21

Monday, May 08, 2017

From June 19-21, 2017, the NCBI will assist in a bioinformatics hackathon at the [New York Genome Center \(NYGC\)](#). This hackathon will focus on advanced bioinformatics analysis of next generation sequencing (NGS) data, proteomics and metadata. To apply for this hackathon, complete this [application](#) (approximately 10 minutes to complete). Applications are due **Monday, May 22, 2017 by 5 PM ET**.

This event is for researchers, including students and postdocs, who are already engaged in the use of bioinformatics data or in the development of pipelines for bioinformatics analyses from high-throughput experiments. Some projects are available to other non-scientific developers, mathematicians or librarians.

The event is open to anyone selected for the hackathon and able to travel to the NYGC (see address below).

Potential subjects for this iteration are:

- Expanding and publicizing a Shiny app for visualizing protein correlation profiling data,
- building a pipeline for efficient partitioning of barcodes,
- creating a public JBrowse database for all *Staphylococcus aureus* genomes,
- simulating tumor genomes,
- associating somatic mutations with clinical outcomes,
- simplifying access to shared-data repositories from Python, and
- building a pipeline for searching for virus-associated protein domains in NGS datasets.

Please see the [application](#) for specific and evolving team projects.

## Organization

There will be 5-7 teams comprised of 5-6 individuals. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

## Datasets

Datasets will come from public repositories or will be supplied by the project lead. During the hackathon, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the hackathon, we ask that you submit it to a public database within **six months** of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose. Manuscripts describing the design and usage of the software tools constructed by each team may be submitted to an appropriate journal such as the [F1000Research hackathons channel](#).

## Application

To apply, complete this [form](#) (approximately 10 minutes to complete). Applications are due Monday, May 22, 2017 by 5 PM ET.

Participants will be selected based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to apply.

The first round of accepted applicants will be notified on May 24th by 5 pm ET, and have until May 25th at 5 pm ET to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.



## Note:

1. Participants will need to bring their own laptop to this program.
2. A working knowledge of scripting (e.g., Shell, Python, R) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.
3. Applicants must be willing to commit to all three days of the event.
4. No financial support for travel, lodging or meals is available for this event.
5. The hackathon may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

Venue: New York Genome Center, 101 6th Ave, New York, NY 10013

## GenBank release 219.0 is available via FTP

*Thursday, May 04, 2017*

GenBank release 219.0 (4/14/2017) has 200,877,884 traditional records containing 231,824,951,552 base pairs of sequence data. In addition, there are 451,840,147 WGS records containing 2,035,032,639,807 base pairs of sequence data, 165,068,542 TSA records containing 149,038,907,599 base pairs of sequence data, as well as 1,438,349 TLS records containing 636,923,295 base pairs of sequence data.

During the 60 days between the close dates for GenBank releases 218.0 and 219.0, the traditional portion of GenBank grew by 3,105,513,914 base pairs and by 1,536,507 sequence records. During that same period, 173,862 records were updated (an average of 28,506 added and/or updated per day).

Between releases 218.0 and 219.0, the WGS component of GenBank grew by 142,066,331,172 base pairs and by 42,349,750 sequence records. The TSA component of GenBank grew by 15,521,695,495 base pairs and by 13,637,057 sequence records. The TLS component of GenBank did not change.

The total number of sequence data files increased by 42 with this release. The divisions are as follows:

- BCT: 20 new files, now a total of 350
- CON: 3 new files, now a total of 359
- ENV: 2 new files, now a total of 97
- EST: 2 new files, now a total of 483
- INV: 1 new file, now a total of 153
- PAT: 7 new files, now a total of 290
- PHG: 1 new file, now a total of 4
- PLN: 2 new files, now a total of 145
- PRI: 1 new file, now a total of 56
- SYN: 1 new file, now a total of 10
- TSA: 1 new file, now a total of 230
- VRL: 1 new file, now a total of 48

For downloading purposes, please keep in mind that the uncompressed GenBank Release 219.0 flatfile require roughly 818 GB (sequence files only). The ASN.1 data require approximately 685 GB.

More information about GenBank release 219.0 is available in the [release notes](#), as well as in the README files in the [genbank \(ftp.ncbi.nih.gov\)](ftp.ncbi.nih.gov) and [ASN.1 \(ncbi-asn1\)](ncbi-asn1) directories.

## May 10th NCBI Minute: How to Locate and Use Human Genomes and Annotations from NCBI

*Monday, May 01, 2017*

In two weeks, NCBI staff will show you how to quickly find and download human genome annotations from both the web and the command line for incorporation into your workflows. We will also show you how to convert the accessions in these files to those used in other bioinformatics databases, as well as how to visualize these annotations on our [Genome Data Viewer](#).

**Date and time:** Wednesday, May 10, 2017 12:00 PM - 12:30 PM EDT

After [registering](#), you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible from the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## NCBI News, April 2017

### April 26th NCBI Minute: Medical Genetics Summaries on the NCBI Bookshelf - a pharmacogenomics resource for clinicians

*Wednesday, April 19, 2017*

Next Wednesday, April 26, 2017, NCBI staff will introduce the [Medical Genetics Summaries](#), a growing collection of reviews available on the NCBI [Bookshelf](#). Each chapter of this book highlights the impact of genetic variations on response to drugs (pharmacogenomics). By the end of this NCBI Minute, you will be able to use the Medical Genetics Summaries to find information about a particular drug, including known impacts of genetic variation on drug response (efficacy, toxicity, side effects) and identify actionable information, including information about relevant genetic testing and how to interpret the test results in order to optimize therapy based on a patient's genotype.

**Date and time:** Wednesday, April 26, 2017 1:00 PM - 1:30 PM EDT

**Register [here](#).**

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

### Maize (*Zea mays*) genome annotation release 101 is now available!

*Wednesday, April 12, 2017*

A new maize (*Zea mays*) genome annotation has been produced by the [RefSeq eukaryotic genome annotation pipeline](#). In [Annotation Release 101](#) a total of 47,446 genes were annotated, including 37,380 that code for proteins. This data is now [available for download](#) and can be explored in the [Genome Data Viewer](#), with [BLAST](#), and in the [Gene database](#).

This annotation benefited from an improved assembly ([B73 RefGen\\_v4](#)), nearly 2 billion more RNA-Seq reads than the previous annotation, and information from over 750,000 [PacBio Iso-Seq transcripts](#). As a result of the improved assembly and the increase in evidence used for gene prediction, about a third of the genes annotated are either new or substantially changed and 20% of the genes from the previous annotation (release 100) were dropped from the current one. In addition, manual curation of over 4000 genes was added by RefSeq Staff.

A full report on the maize (*Zea mays*) Annotation Release 101 annotation can be found [here](#).

### dbSNP FTP file format change planned for early 2018

*Tuesday, April 11, 2017*

In early 2018, dbSNP will no longer provide relational database table dumps on the FTP site or any general SQL support for future build releases. Instead, dbSNP FTP data will be made available as a cumulative file of RefSNP objects in JSON format. [These files are now available](#), so users can begin migration and testing. Please see the [dbSNP Alert README file](#) for more details.

## dbSNP's human build 150 has doubled the amount of RefSNP records!

*Tuesday, April 11, 2017*

dbSNP's Human Build 150 includes a large number of new submissions from the [Human Longevity, Inc. \(HLI\)](#) and [TopMed](#), increasing the total number of Human RefSNPs in the database from 154 to 324 million. TopMed has also provided new allele frequency data for 163 million RefSNPs.

Human Build 150 Notes:

HLI-submitted data were aligned with the human genome assembly GRCh38. Because dbSNP's pipeline does not support mapping backward to previous assemblies, the rsIDs for these will not appear in the VCF files for GRCh37. We are investigating mechanisms to map these variants from the GRCh38 to GRCh37 assembly and will provide updates.

TopMed-submitted allele frequencies are available for both [GRCh38.p7](#) and [GRCh37.p13](#) in VCF format on the FTP site with the INFO tag 'TOPMED'.

Due to the unexpected increase in the volume of human data and limitations in our systems, dbSNP had to take two temporary actions for this release:

1. The "dbSNP Build 150 (Homo sapiens Annotation Release 108) all data" annotation track for RefSeq genomic sequences will be limited to variants in the gene regions only. This only affect tracks displayed in the NCBI Sequence Viewer and does not impact reporting in dbSNP FTP files or on Reference SNP pages. We are investigating mechanisms to restore complete annotation across the genome and will provide updates.
2. Entrez searching is currently only available for Human Build 150 and a limited number of organisms from previous builds - including mouse, rat, cow, and pig. We will provide updates over the next two weeks as we restore the search capability for other organisms.

For more information, see the [dbSNP-Announce message](#).

## NCBI researchers and collaborators discover novel group of giant viruses

*Thursday, April 06, 2017*

Nearly complete set of translation-related genes lends support to hypothesis that giant viruses evolved from smaller viruses

An international team of researchers, including NCBI's [Eugene Koonin](#) and Natalya Yutin, has discovered a novel group of giant viruses (dubbed "Klosneuviruses") with a more complete set of translation machinery genes than any virus that has been described to date. "This discovery significantly expands our understanding of viral evolution," said Koonin. "These are the most 'cell-like' viruses ever identified. However, the computational analysis of the virus genomes shows that these viruses have not evolved from cells by reductive evolution but rather have evolved from smaller viruses, gradually acquiring genes from their hosts at different stages of their evolution."

The research was published in the journal [Science](#) on April 6, 2017. In addition to biologists from NCBI, the authors include collaborators from the U.S. Department of Energy Joint Genome Institute (DOE JGI), the University of Vienna, and CalTech.

JGI researchers Frederik Schulz and Tanja Woyke unearthed Klosneuvirus while analyzing microcolony sequence data from a wastewater treatment plant sample in Klosterneuburg, Austria. "We expected nitrifier genome sequences in the microcolony sequence data," Woyke said. "Finding a giant virus genome took the project into a completely new and unexpected, yet very exciting, direction." When Schulz noticed that several of the metagenomes were viral in origin, he and Woyke conducted analyses to determine their source. They found that the Klosneuvirus group came from a novel viral lineage affiliated with Mimiviruses, the first giant viruses discovered. A handful of other giant virus groups have been found since the discovery of Mimiviruses in 2003.

Giant viruses are characterized by disproportionately large genomes and virions that house viruses' genetic material. They can encode several genes potentially involved in protein biosynthesis, a unique feature that has led to diverging hypotheses about their origin. Two evolutionary hypotheses have emerged. One posits that giant viruses evolved from an ancient cell, perhaps one from an extinct fourth domain of cellular life. Another — a scenario championed by Koonin — presents the idea that giant viruses descended from smaller viruses. The discovery of Klosneuvirus, Woyke said, supports the latter hypothesis. In this scenario, a smaller virus infected different eukaryote hosts and picked up genes from independent sources over long periods of time through piecemeal acquisition of translational machinery components.

"At first glance, the suite of "cellular" genes in Klosneuvirus seemed to have a common origin, but when we analyzed them in detail, we saw they came from different hosts," Koonin said. "We could infer from the evolutionary trees we built that they have been acquired by the viruses piecemeal, at different stages in their evolution." The Klosneuvirus genes contained aminoacyl-tRNA (transfer ribonucleic acid) enzymes with specificity for 19 out of 20 amino acids, along with more than 20 tRNAs and an array of translation factors and tRNA modifying enzymes—an unprecedented finding among any viruses, including the previously known giant viruses.

Schulz noted that while the metagenomic discovery of Klosneuviruses helped answer important evolutionary questions, the actual biological function of the translation system genes remains elusive—at least until these viruses are grown in the laboratory together with their hosts.

And Koonin believes there are more giant viruses waiting to be discovered in metagenomic data. "I'm quite confident that the current record of the genome size of giant viruses will be broken," he said. "We are going to see the real Goliaths of the giant virus world."

— Many thanks to JGI for their assistance in preparing this news feature.

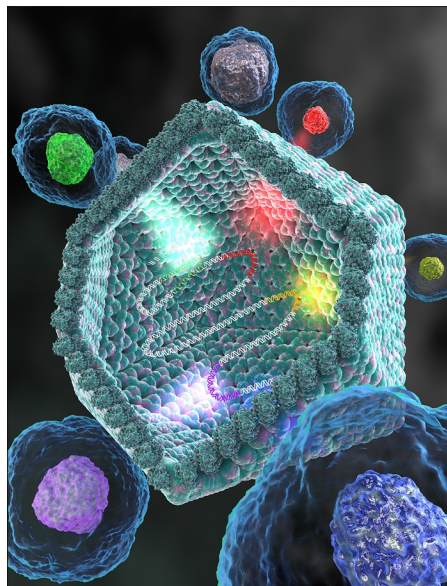


Image credit: Ella Maru studio

## April 19th NCBI Minute: Magic-BLAST, NCBI's next-gen sequence alignment program

Wednesday, April 05, 2017

In two weeks, NCBI staff will introduce you to [Magic-BLAST](#), NCBI's next-gen sequence aligner. You will learn to use magic-BLAST to align next-gen RNA and DNA sequencing runs to genomic and transcript sequences and to understand the options available for magic-BLAST

**Date and time:** Wednesday, April 19, 2017 12:00 PM - 12:30 PM EDT

### Registration

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible from the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## Six functional prototypes available from the March NCBI hackathon

Wednesday, April 05, 2017

At the March 2017 NCBI Genomics Hackathon, participants developed six functional software prototypes, several of which are still under active development. Software is available from the [NCBI-Hackathons GitHub site](#).

1. [Squidstream](#) provides naming consistency by converting sequence feature IDs in entire files (bed, gff3, wig, etc.) to the desired ID format using a single command.
2. [ga4gh-ncbi-api](#) is a method that links NCBI's API and the GA4GH (Global Alliance for Genomics and Health) API, and generates a searchable list of genome datasets from NCBI.
3. [Graph\\_Extraction](#) provides code to implement a simple graph genome browser.
4. [Sidearm](#) searches the SRA database for viruses using the NCBI magicBLAST tool.
5. [Scan2CNV](#) is a commandline tool that generates copy number variation (CNV) calls from raw SNP array data.
6. [Single Cell Reproducible Epigenomics Workflow \(SCREW\)](#) is a single-cell whole-genome bisulfite sequencing (SC-WGBS) pipeline and docker image for performing standard single-cell DNA methylation analyses.



## Eight new eukaryotic genome annotations added to RefSeq

Tuesday, April 04, 2017

In the past month, the NCBI Eukaryotic Genome Annotation Pipeline has released new annotations in RefSeq for the following organisms:

- *Zea mays* (maize)
- *Labrus bergylta* (ballan wrasse)
- *Monopterus albus* (swamp eel)
- *Corvus cornix cornix* (hooded crow)
- *Prunus persica* (peach)
- *Rhincodon typus* (whale shark)
- *Oncorhynchus kisutch* (coho salmon)
- *Pseudomyrmex gracilis* (ant)

See more details on the [Eukaryotic RefSeq Genome Annotation Status page](#).



## New Genome Data Viewer access page

Monday, April 03, 2017

NCBI is pleased to offer a direct entry point to the [NCBI Genome Data Viewer \(GDV\)](#) that supports the exploration, visualization and analysis of eukaryotic RefSeq genome assemblies. The new GDV homepage includes an interactive interface for a quick overview of supported organisms, specific genome searches plus

inter-connectivity to Assembly and RefSeq annotation resources. About 100 genome assemblies are now ready for GDV exploration with more on the way. Stay tuned!

## Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of eukaryotic RefSeq genome assemblies. ⓘ

Select organism

Homo sapiens (human)

### Homo sapiens (human) genome

Search genome

Q

Search examples:

Assembly

▼

Browse genome

BLAST genome

#### Assembly details

<b>Name</b>	GRCh38.p10
<b>RefSeq accession</b>	<a href="#">GCF_000001405.36</a>
<b>GenBank accession</b>	<a href="#">GCA_000001405.25</a>
<b>Download via FTP</b>	<a href="#">RefSeq</a> , <a href="#">GenBank</a>
<b>Submitter</b>	Genome Reference Consortium
<b>Level</b>	Chromosome
<b>Category</b>	Reference genome

#### Annotation details

**Annotation Release** [108](#)

**Release date** 2016-06-06



## NCBI News, March 2017

### Sequence Viewer 3.20 is now available

*Thursday, March 30, 2017*

Sequence Viewer 3.20 has several new features, improvements and bug fixes, including discrete color maps for graph tracks, improved performance in initialization and loading tracks, improved display of overlapping variation features and the addition of a status bar. For a full list of changes, see the Sequence Viewer [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

### Conserved Domain Database (CDD) version 3.16 now available online and via FTP

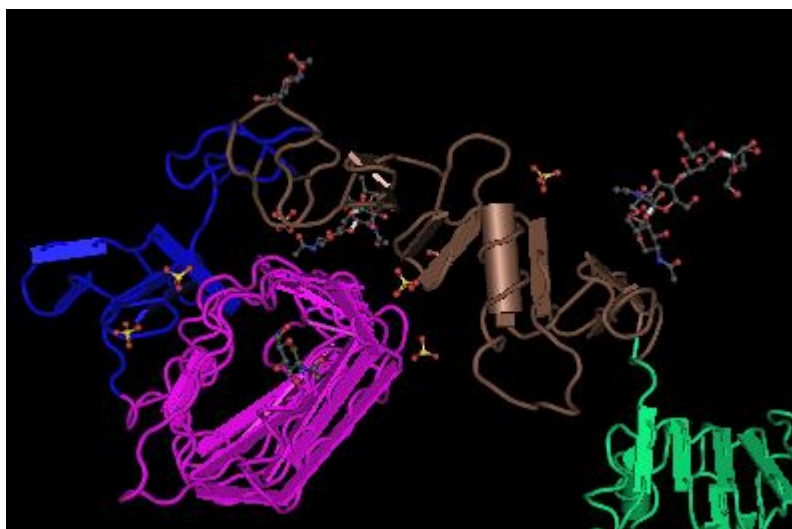
*Thursday, March 30, 2017*

Version 3.16 of the [Conserved Domain Database](#) contains 1,659 new or updated NCBI-curated domains (56,066 total), including models specifically built to annotate structural motifs (accession prefix "sd"), and now mirrors Pfam version 30.

Updates include:

- Fine-grained classification of the 7-membrane GPCR transmembrane subunits.
- Database size parameters for CD-Search have been adjusted, resulting in slightly higher E-values.
- Fewer models are now assigned a multi-domain-model status, affecting the domain annotation of a large number of proteins.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#).



Type 1 Insulin-like Growth Factor Receptor (IIGR), colored by domain.

### NCBI to assist with BioFrontiers Hackathon in May

*Monday, March 27, 2017*

From May 22nd to 24th, NCBI will be assisting with the BioFrontiers Hackathon in Boulder, Colorado. Please see the [BioFrontiers Hackathon website](#) for more information, including what to expect, who should apply, and the application itself. Applications are due by **April 7, 2017**.

## NCBI will attend the AACR Annual Meeting 2017

*Tuesday, March 21, 2017*

From April 2-5, 2017, NCBI will attend the American Association for Cancer Research (AACR) Annual Meeting in Washington, DC. Join us at Exhibit Booth #3230 at the following times:

- Sunday, April 2, 1 – 5pm
- Monday, April 3, 9am – 5pm
- Tuesday, April 4, 9am – 5pm
- Wednesday, April 5, 9am – 12 noon

At the booth, you'll be able to have your questions answered and see demos of NCBI resources pertaining to medical genetics, sequences and their variations, and biomedical literature.

## Genome Workbench 2.11.10 now available

*Monday, March 20, 2017*

The latest version of [Genome Workbench](#) includes a number of new features, fixes and improvements like a critical improvement in HTTPS protocol communication with NCBI and a new coloring scheme in Multiple Alignment View.

For a full list of changes, please see the [Genome Workbench release notes](#).

## Tree Viewer version 1.13 implements new search in tree features

*Monday, March 20, 2017*

Tree Viewer [version 1.13](#) has several improvements, updates and bug fixes, including [search in trees](#), improved automatic subtree collapse, and more. The Tree Viewer [release notes](#) list all updates.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## Complete RefSeq genome annotation results represented in UCSC genome browser

*Friday, March 17, 2017*

We are very pleased to announce the [availability of the complete RefSeq human genome annotation product](#) for the GRCh38 assembly in the University of California, Santa Cruz (UCSC) Genome browser. NCBI and UCSC staff have worked closely to define an improved data exchange process and NCBI is now providing RefSeq genome annotation and alignment data in order to have a more complete reflection of the RefSeq product in the UCSC genome browser. This resolves issues of incomplete data and conflicting placement details between UCSC displays and NCBI displays.

This initial release is for the human reference genome (GRCh38) and does not include NCBI RefSeq annotation for GRCh38 patches added since the initial GRCh38 release. We anticipate working with UCSC to expand on the number of organisms in the future.

NCBI-provided RefSeq data is included in the "NCBI RefSeq" composite track. For the following tracks, the alignments and coordinates are provided by RefSeq:

- RefSeq All – curated and predicted transcript annotations
- RefSeq Curated – curated annotations (transcripts with NM\_ and NR\_ accessions)
- RefSeq Predicted – predicted annotations (transcripts with XM\_ and XR\_ accessions)
- RefSeq Other – annotations not included in RefSeq All such as pseudogenes or other loci
- RefSeq Alignments – alignments of transcripts to the genome provided by RefSeq

By default, only the "RefSeq Curated" subtrack is activated within the "NCBI RefSeq" track, but you may wish to activate the other subtracks to view the complete dataset.

## March 29th NCBI Minute: How to Submit Your 16S rRNA Data to NCBI

*Friday, March 17, 2017*

In two weeks, NCBI staff will guide you through the submission of prokaryotic 16S rRNA sequences to [GenBank](#) using one of the new Submission Wizards.

**Date and time:** Wednesday, March 29, 2017 12:00 PM - 12:30 PM EDT

### Registration

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also learn about future webinars on this page.

## NCBI will attend the 2017 Annual Clinical Genetics Meeting

*Wednesday, March 15, 2017*

Join NCBI staff at the [2017 Annual Clinical Genetics Meeting \(ACMG\)](#) in Phoenix, AZ on March 21-25, 2017. At Exhibit Booth #531, you'll be able to get navigation tips, hands-on help, and handout materials.

In addition, Adriana Malheiro will present NCBI's suite of human genome resources that support the Precision Medicine Initiative in a Platform Presentation titled "In the Clinic with Medical Genetics Summaries (MGS)".

Finally, Melissa Landrum and Adriana Malheiro will present posters titled "ClinVar: For medical practitioners and researchers alike" and "MedGen: Harmonizing phenotypic information into an online, computer-readable resource of medical genetics", respectively.

Please see the [NCBI Conferences & Presentations](#) page, as well as the official [NCBI](#) and [NCBI Clinical](#) Twitter accounts once ACMG starts for further information about presentations and posters, as well as times and locations.

## RefSeq release 81 now public

*Tuesday, March 14, 2017*

RefSeq release 81 is now accessible online, via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript, and protein data available as of March 6, 2017 and contains 121,954,847 records, including 81,027,309 proteins, 18,381,587 RNAs, and sequences from 68,165 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

## GI sequence identifiers removed from flatfile and FASTA formats

Please refer to these NCBI News announcements for more details:

- GI numbers will be removed from sequence record presentations in this release
- NCBI is phasing out sequence GIs - use Accession.Version instead!

## Comprehensive reannotation of prokaryotic genomes

The first phase of the comprehensive reannotation of prokaryotic genomes has been completed, covering *Escherichia*, *Shigella*, *Salmonella*, *Klebsiella*, and *Listeria*.

Reannotation is expected to be completed before the May 2017 RefSeq release.

Information on the improvements to the Prokaryotic Genome Annotation Pipeline 4.1 can be found [here](#).

FTP files for RefSeq prokaryote genomes on the [genomes FTP site](#) will be refreshed upon completion of the reannotation project.

## GenBank release 218.0 is now available

*Tuesday, March 14, 2017*

GenBank release 218.0 (2/13/2017) has 199,341,377 traditional records containing 228,719,437,638 base pairs of sequence data. In addition, there are 409,490,397 WGS records containing 1,892,966,308,635 base pairs of sequence data, 151,431,485 TSA records containing 133,517,212,104 base pairs of sequence data, as well as 1,438,349 TLS records containing 636,923,295 base pairs of sequence data.

During the 60 days between the close dates for GenBank releases 217.0 and 218.0, the traditional portion of GenBank grew by 3,746,377,205 base pairs and by 775,902 sequence records. During the same period, 68,617 records were updated at an average of 14,075 traditional records added and/or updated per day.

Between releases 217.0 and 218.0, the WGS component of GenBank grew by 75,776,742,790 base pairs and by 14,189,221 sequence records. The TSA component of GenBank grew by 8,188,387,596 base pairs and by 9,337,148 sequence records. The TLS component of GenBank grew by 52,225,376 base pairs and by 169,659 sequence records.

The total number of sequence data files increased by 39 with this release. The divisions are as follows:

- BCT: 26 new files, now a total of 330
- CON: 1 less file, now a total of 356
- ENV: 1 new file, now a total of 95
- INV: 1 new file, now a total of 152
- PAT: 3 new files, now a total of 283
- PLN: 6 new files, now a total of 143
- VRL: 2 new files, now a total of 47
- VRT: 1 new file, now a total of 64

For downloading purposes, please keep in mind that the uncompressed GenBank Release 218.0 flatfiles require approximately 818 GB (sequence files only); the ASN.1 data require approximately 677 GB.

More information about GenBank release 218.0 is available in the [release notes](#), as well as in the README files in the genbank (<ftp.ncbi.nih.gov>) and ASN.1 (<ncbi-asn1>) directories.

## Seven new annotations added to RefSeq

*Thursday, March 09, 2017*

In the past month, the NCBI Eukaryotic Genome Annotation Pipeline has released new annotations in [RefSeq](#) for the following organisms:

- *Asparagus officinalis* (garden asparagus)
- *Microcebus murinus* (gray mouse lemur)
- *Aegilops tauschii* (a monocot)
- *Cajanus cajan* (pigeon pea)
- *Castor canadensis* (American beaver)
- *Ananas comosus* (pineapple)
- *Paralichthys olivaceus* (Japanese flounder)

See more details on the [Eukaryotic RefSeq Genome Annotation Status](#) page.

## Multiple Sequence Alignment Viewer 1.4 is now available

*Thursday, March 09, 2017*

The new version of the [Multiple Sequence Alignment Viewer](#) (MSA Viewer) has implemented several bug fixes affecting several features, including zoom on alignments and text import. A full list of bug fixes is available in the MSA Viewer [release notes](#).

## Magic-BLAST 1.2.0 now available

*Monday, March 06, 2017*

The newest version of Magic-BLAST handles multiple SRA accessions, offers improved splice site detection and multi-threading performance, and fixes issues with macOS installation. For more information, see the [release notes](#). The new executables are available on the [NCBI FTP site](#).

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Read more [here](#).

## Expression section and bulk datasets added to NCBI Gene

*Monday, March 06, 2017*

The [Gene](#) resource has a new feature that reports normalized RNA expression levels computed from RNA-Seq data for human, mouse, and rat genes. An expression chart is available on the Gene full report pages, with an additional table view and download option on the new expression report page available through the “See details” link or format menu.

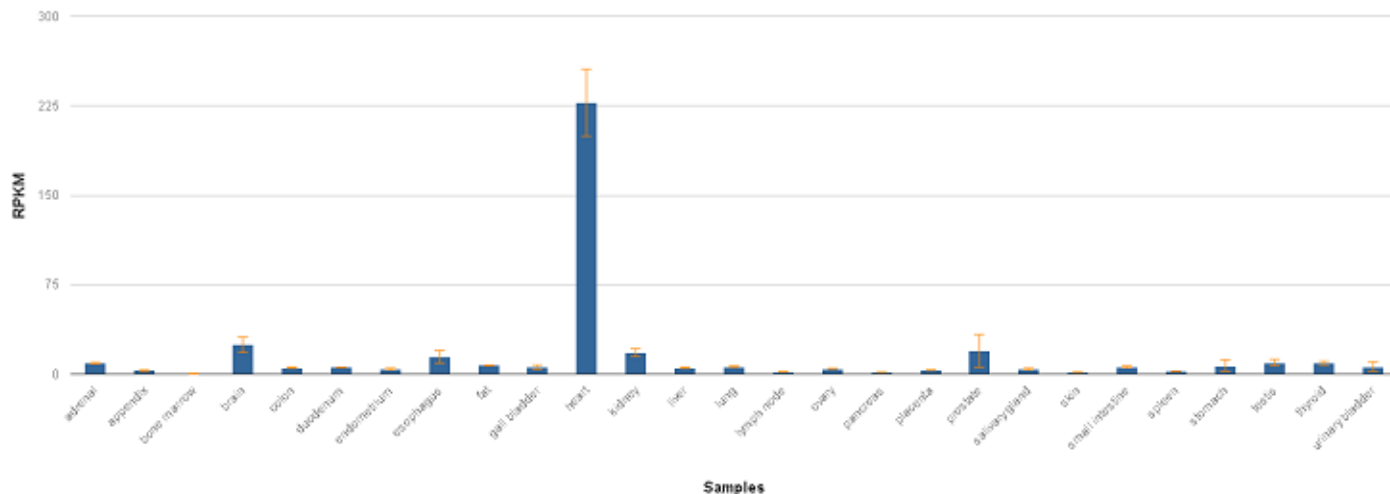
**SLC25A4 solute carrier family 25 member 4 [ *Homo sapiens* (human) ]**

Gene ID: 291, updated on 4-Mar-2017

- Summary ?
- Genomic context ?
- Genomic regions, transcripts, and products ?
- Expression ?

HPA RNA-seq normal tissues [See details](#)

- Project title: HPA RNA-seq normal tissues
- Description: RNA-seq was performed of tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes
- BioProject: [PRJEB4337](#)
- Publication: [PMID 24308836](#)
- Analysis date: Wed Jun 15 11:32:44 2016



**Figure 1. The new Expression section on Gene pages.**

Expression data can provide key insights into where and when a gene may be functioning, for example by exposing the correlation between expression of [human SLC25A4](#) and its established role in heart function.

Bulk datasets will also be available on the Gene FTP site. The RNA-Seq expression coverage graphs for each sample used to compute expression levels are available in the embedded graphical viewer and Genome Data Viewer under the expression category. We welcome questions about this new dataset at [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov) or through the "Contact Help Desk" link available on the Gene full report page.

## NCBI News, February 2017

### NCBI Insights | PubMed Citations: A New, Faster Process for Correcting Errors

*Tuesday, February 28, 2017*

The [latest blog post](#) on NCBI Insights introduces users to the PubMed Data Management System (PMDM), which allows publishers to correct PubMed citation data directly. Authors should contact journal publishers to correct PubMed citation mistakes.

NCBI Insights is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

### March 1st NCBI Minute: Setting up new data alerts with MyNCBI

*Thursday, February 23, 2017*

Next Wednesday, March 1, 2017, NCBI will present a short webinar that will show you how to use MyNCBI alerts to be notified when new citations of interest appear in traditional sequence databases, as well as SRA and GEO.

**Date and time:** Wednesday, March 1, 2017 12:00 PM – 12:15 PM EST

#### Register

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

### Bottlenose dolphin annotation release 101

*Wednesday, February 22, 2017*

Annotation Release 101 for the bottlenose dolphin (*Tursiops truncatus*) is out in RefSeq! This annotation was based on the NIST Tur\_tru v1 assembly, which has a four-fold increase in contiguity from the assembly used in the previous annotation. Over four billion RNA-Seq reads from skin and blood tissue were used for gene prediction. As a result of these improvements, the percent of partially-represented protein-coding genes went down from 24% to 4%. Over 2500 genes that were fragmented in the previous assembly were merged into complete genes. A total of 24,026 genes were annotated, and 17,096 of them were protein-coding. A full report on the annotation can be found [here](#).



Figure 1. *Tursiops truncatus*, the bottlenose dolphin.

This improved genomic resource for the dolphin will allow NIST to develop standardized research methods, produce reference data and tools, and perform bioanalytical measurements on dolphins and other marine organisms. Dolphins are important sentinel organisms for the health status of the marine environment and their study expands knowledge on cognition, communication, acoustics, conservation, and hydrodynamics.

Annotation Release 101 is available for [download](#) and [formatted for BLAST searches](#).

## **New video on YouTube: Embed the NCBI Sequence Viewer into Your Pages**

*Tuesday, February 21, 2017*

The [newest video](#) on the NCBI YouTube channel introduces the [Sequence Viewer](#) embedding API. A few quick examples illustrate how easy it is to embed Sequence Viewer into your own pages.

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.



## NLM Webinar series: "Insider's Guide to Accessing NLM Data: EDirect for PubMed"

*Friday, February 17, 2017*

Beginning February 21, 2017, the National Library of Medicine (NLM) will present the three-part webinar series "Insider's Guide to Accessing NLM Data: EDirect for PubMed."

This series of workshops will introduce new users to the basics of using EDirect to access exactly the PubMed data you need, in the format you need. Over the course of three 90-minute sessions, students will learn how to use EDirect commands in a Unix environment to access PubMed, design custom output formats, create basic data pipelines to get data quickly and efficiently, and develop simple strategies for solving real-world PubMed data-gathering challenges. No prior Unix knowledge is required; novice users are welcome!

This series of classes involves hands-on demonstrations and exercises, and we encourage students to follow along. Before registering for these classes, we strongly recommend that you:

- Watch the first Insider's Guide class "[Welcome to E-utilities for PubMed](#)", or be familiar with the basic concepts of APIs and E-utilities.
- Be familiar with structured XML data (basic syntax, elements, attributes, etc.)
- Have access to a Unix command-line environment on your computer (for more information, see our [Installing EDirect](#) page).
- Install the EDirect software (for more information, see our [Installing EDirect](#) page).

Due to the nature of this class, registration will be limited to 50 students per offering.

Registration is currently open for the February/March 2017 series:

- Part 1: Getting PubMed Data: Tuesday, February 21, 1-2:30 PM ET
- Part 2: Extracting Data from XML: Tuesday, February 28, 1-2:30 PM ET
- Part 3: Building Practical Solutions: Tuesday, March 7, 1-2:30 PM ET

Students are expected to attend Part 1, Part 2, and Part 3 in a single series.

To register, and for more information, visit <https://goo.gl/DVOh6M>.

## Tree Viewer version 1.12 implements new API to markup trees

*Tuesday, February 14, 2017*

Tree Viewer version 1.12 has several improvements, updates and bug fixes, including a new API, PDF rendering, and Tree Viewer macro language. The Tree Viewer [release notes](#) list all updates.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## Interim annotation updates for the human GRCh37.p13 and GRCh38.p10 assemblies

*Tuesday, February 14, 2017*

Updates to the annotation of the human [GRCh37.p13](#) and [GRCh38.p10](#) assemblies are now available for download by anonymous FTP. These annotation updates contain features projected from the current known RefSeq transcripts and curated genomic sequences (with accession prefixes NM\_ or NR\_, and NG\_ respectively) placed on either the GRCh37.p13 or GRCh38.p10 assembly.

The GRCh37.p13 annotation is being provided to help support members of the clinical community who are still dependent on the old GRCh37 (hg19) assembly. However, users should be cautious about using these annotation results, especially in regions that were extensively revised in GRCh38. See the corresponding README file for more details including details on genes that are no longer annotated in the update.

The two annotations started with the same set of RefSeq transcripts, and differences in which RefSeqs are annotated reflect improvements in the GRCh38 assembly, as well as some genes and functionally distinct alleles that were relocated between the chromosomes of the primary assembly and alt loci scaffolds.

Annotation is available in GFF3 format, as well as alignments of current RefSeq transcripts to the genome in both GFF3 and BAM formats. The annotations are also available in NCBI's genome browsers such as Variation Viewer and 1000 Genomes Browser, including in either the "Genes" recommended track set or from the track selection dialog (search for "interim").

Please send questions, comments, and suggestions concerning these updates to [refseq-admin@ncbi.nlm.nih.gov](mailto:refseq-admin@ncbi.nlm.nih.gov) or use the [Feedback link](#) from Entrez Gene reports.

## February 22nd webinar: Introducing the Multiple Sequence Alignment Viewer

*Monday, February 13, 2017*

Next Wednesday, February 22, 2017, NCBI will present a webinar on the [Multiple Sequence Viewer \(MSAV\)](#). In this webinar, you will learn how to display alignment data from many sources, including NCBI BLAST results, as well as precomputed multiple alignments of your own data. You will also see how to embed the viewer in your own web pages or share a link to a particular alignment display.

**Date and time:** Wednesday, February 22, 2017 12:00 PM – 12:30 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/7663489773270563843>

The MSAV is a versatile web application that helps you visualize and interpret multiple sequence alignments for both nucleotide and protein sequences. You can use the viewer to explore sequence conservation, investigate variation or troubleshoot assembly or sequencing errors.

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## SmartBLAST updated to provide more information, database matches

*Monday, February 13, 2017*

The SmartBLAST service has recently been updated to emphasize matches to the [landmark database](#), which comprises the proteomes from 26 well-curated genomic assemblies. The display also now presents more information about conserved domains and details about the query.

SmartBLAST quickly finds the closest relatives to a protein query and evaluates the phylogenetic relationship among the query and matched sequences. You can start a SmartBLAST search from the [SmartBLAST page](#) or the BLAST home page. Read more about SmartBLAST on [NCBI Insights](#).

## Sequence Viewer 3.19 is now available

*Monday, February 13, 2017*

Sequence Viewer 3.19 has several new features, improvements and bug fixes, including a new aggregate track type, improved display of projected features and cleaned alignments, and a new manual for using embedded API. For a full list of changes, see the Sequence Viewer [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## New NCBI Insights post: New Web Services for Comparing and Grouping Sequence Variants

*Thursday, February 09, 2017*

The latest [post](#) on the NCBI Insights blog introduces new web services for comparing and grouping variants. Geneticists, dataflow engineers, and anyone who needs to compare genetic variants can use these [services](#).

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## NCBI to host genomics hackathon March 20-22

*Thursday, February 02, 2017*

From March 20th to 22nd, the NCBI will host a genomics hackathon on the NIH campus. To apply for this hackathon, complete this [application](#) (approximately 10 minutes to complete). Applications are due **February 22nd by 1 PM ET**.

This hackathon will primarily focus on advanced bioinformatics analysis of next generation sequencing data and metadata. This event is for students, postdocs and investigators or other researchers already engaged in the use of genomics data or pipelines for genomic analyses from next generation sequencing data. However, there are some projects available to other non-scientific developers, mathematicians or librarians. The event is open to anyone selected for the hackathon who is able to travel to NIH.

### Organization

There will be 5-7 teams of 5-6 individuals. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure.

The potential subjects for this iteration are:

- GA4GH - NCBI API Integration,
- Global Screening Arrays,
- Graph Genome Information Extraction,
- Single Cell Methylation Data,
- Generation of an automated gff3 parser,
- integration of Immpart metadata with specific genomic datasets,
- And several others.

Please see the [application](#) for specific and evolving team projects.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

## Datasets

Datasets will come from public repositories, primarily those housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis.

Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose.

A manuscript outlining the design and usage of the software tools constructed by each team may be submitted to an appropriate journal such as the [F1000Research hackathons channel](#).

## Application

To apply, complete this [application](#) (approximately 10 minutes to complete). Applications are due **February 22nd by 1 pm ET**.

Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

The first round of accepted applicants will be notified on February 24th by 5 pm ET, and have until February 27th at 1 pm ET to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event.

Please include a monitored email address, in case there are follow-up questions.

## Notes

Participants will need to bring their own laptop to this program.

A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.

Applicants must be willing to commit to all three days of the event.

No financial support for travel, lodging or meals is available for this event.

Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## NCBI News, January 2017

### Multiple Sequence Alignment Viewer 1.3 is now available

*Tuesday, January 31, 2017*

The new version of the [Multiple Sequence Alignment Viewer](#) (MSA Viewer) has implemented a new coloration method and improved tooltips. A full list of new features, improvements, and bug fixes is available in the [MSA Viewer release notes](#).

### February 8th NCBI Minute: Finding Gene, Protein and Chemical Names, Aliases and Synonyms

*Tuesday, January 31, 2017*

Next Wednesday, February 8th, NCBI staff will discuss the systems in the NCBI Gene and PubChem resources that identify and correlate various names used for genes, proteins and chemicals.

**Date and time:** Wednesday, February 8, 2017 12:00 PM - 12:30 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/6498213056303481858>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

### BLAST+ 2.6.0 offers improved support for accession.version

*Thursday, January 26, 2017*

The newest version of the [BLAST+ executables](#) provides improved support for use of `accession.version` as the primary identifier, as well as improved speed of `blastdbcmd` when dumping information from a database.

A number of other bug fixes and improvements are also included in this release. For more information about BLAST+ 2.6.0, please see the [release notes](#).

### New NCBI Insights post: Visualize and Interpret Alignment Data with the Multiple Sequence Alignment Viewer

*Wednesday, January 25, 2017*

The [latest post](#) on the NCBI Insights blog introduces the Multiple Sequence Alignment Viewer (MSAV), a resource for visualizing and interpreting alignments for nucleotide and amino acid sequences. The viewer is easily embedded in web pages, readily customizable, and displays alignment data from many sources. Read on at [NCBI Insights](#).

NCBI Insights is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

### January 31st NCBI Minute: New version of E-utilities supports accession.version

*Monday, January 23, 2017*

Next Tuesday, January 31, 2017, NCBI will present a short webinar that describes and demonstrates new functionality recently introduced to the E-utilities that supports sequence data retrieval.

**Date and time:** Tuesday, January 31, 2017 12:00 PM - 12:30 PM EST

**Registration URL:** <https://attendeegotowebinar.com/register/7530877675754064131>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## **RefSeq release 80 now available; GI identifiers to be removed in next release (March 2017)**

*Friday, January 13, 2017*

RefSeq release 80 is now accessible online, via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript, and protein data available as of January 9, 2017 and contains 118,059,547 records, including 78,028,152 proteins, 17,862,608 RNAs, and sequences from 66,224 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

As announced in March 2016, NCBI has implemented the removal of GI numbers from some presentations of nucleotide and protein sequence records. GI sequence identifiers will be removed from flatfile and FASTA formats in the RefSeq FTP release in March 2017.

RefSeq plans to start a comprehensive reannotation of all prokaryotic genomes in a few weeks, which will be included in its entirety in the May 2017 release.

For more information on RefSeq release 80, please see the [release notes](#).

## **New videos on YouTube: Clone DB and clone placements**

*Friday, January 13, 2017*

Two new videos on the [NCBI YouTube channel](#) demonstrate how to use [Clone DB](#) and clone placements to assess and improve genome assemblies.

The first video, *Using Clone Placements to Interpret Genome Assemblies*, shows you how to interpret and improve genome assemblies with clone placement data.

The second video, *Clone DB FTP Files: Content and Uses for Genome Assemblies*, teaches you how to use Clone DB FTP files to not only improve genome assemblies but also detect potential structural variation and place clones based on clone end sequences.

Subscribe to the NCBI YouTube channel to receive alerts about new videos ranging from quick tips to full webinar presentations.

## **GenBank release 217.0 is available via FTP**

*Wednesday, January 11, 2017*

GenBank release 217.0 (12/15/2016) has 198,565,475 traditional records containing 224,973,060,433 base pairs of sequence data. In addition, there are 395,301,176 WGS records containing 1,817,189,565,845 base pairs of sequence data, 142,094,337 TSA records containing 125,328,824,508 base pairs of sequence data, as well as 1,268,690 TLS records containing 584,697,919 base pairs of sequence data.

During the 65 days between the close dates for GenBank releases 216.0 and 217.0, the traditional portion of GenBank grew by 4,421,745,183 base pairs and by 1,174,784 sequence records. During the same period, 726,256 records were updated at an average of 29,247 traditional records added and/or updated per day.

Between releases 216.0 and 217.0, the WGS component of GenBank grew by 140,951,076,595 base pairs and by 32,087,861 sequence records. The TSA component of GenBank grew by 12,119,598,746 base pairs and by 17,894,740 sequence records.

The total number of sequence data files increased by 57 with this release. The divisions are as follows:

- BCT: 23 new files, now a total of 304
- CON: 5 new files, now a total of 357
- ENV: 1 new file, now a total of 94
- GSS: 2 new files, now a total of 303
- HTG: 2 new files, now a total of 154
- MAM: 2 new files, now a total of 39
- PAT: 12 new files, now a total of 280
- PLN: 2 new files, now a total of 137
- VRL: 1 new file, now a total of 45

For downloading purposes, please keep in mind that the uncompressed GenBank Release 216.0 flatfiles require approximately 809 GB (sequence files only); the ASN.1 data require approximately 671 GB.

More information about GenBank release 217.0 is available in the [release notes](#), as well as in the README files in the genbank (<ftp.ncbi.nih.gov>) and ASN.1 (<ncbi-asn1>) directories.

## Genome Workbench 2.11.7 now available

*Wednesday, January 04, 2017*

The latest version of [Genome Workbench](#) includes a number of new features, fixes and improvements like a critical improvement in HTTPS protocol communication with NCBI, improved rendering for translation discrepancies, and improved handling of tracks.

For a full list of changes, please see the Genome Workbench [release notes](#).





## NCBI News, December 2016

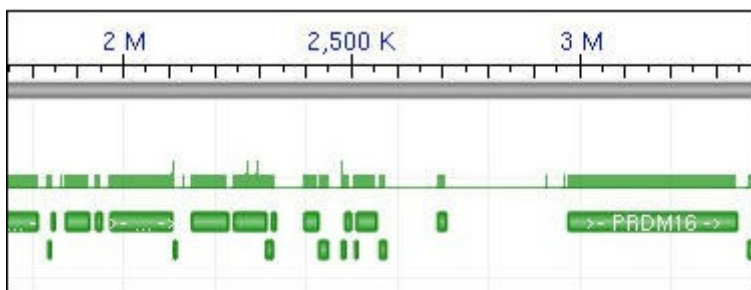
### New YouTube video: Sequence Viewer: Display Translation Discrepancies

Friday, December 23, 2016

The newest video on the [NCBI YouTube channel](#) is a brief introduction to a new set of sequence viewer renderings that better display discrepancies between genomic sequence and annotated features. These discrepancies can occur because RefSeq gene models based on the current genomic sequence can differ from RefSeq transcripts and corresponding proteins that come from our continuous curation efforts. For links to examples used in the video, see the [Sequence Viewer tutorial](#).

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.



### New NCBI Insights post: Converting Lots of GI Numbers to Accession.version

Friday, December 23, 2016

The latest [post](#) on the NCBI Insights blog provides a bulk conversion resource for those who need to convert more than a few thousand GI numbers to accession.version identifiers.

As you may already know, accession.version identifiers, rather than GI numbers, will be the primary identifiers for sequence records at NCBI.

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

### CCDS release 21 for mouse is public in Gene

Thursday, December 22, 2016

The Consensus Coding Sequence (CCDS) update that compares NCBI's *Mus musculus* annotation release 106 to Ensembl's release 86 is now available in [Gene](#). This update adds 938 new CCDS IDs, and adds 137 genes into the mouse CCDS set. CCDS release 21 includes a total of 25,757 CCDS IDs that correspond to 20,354 GeneIDs.

Also, note that the [CCDS survey](#) is still open. NCBI and the CCDS collaboration invite you to take a survey that will help us assess how the human and mouse Consensus CDS data is being accessed and used by the scientific

community. We welcome your feedback and suggestions on this data collection. Data gathered from the survey will help us plan the future direction of the CCDS project.

## December 21st NCBI Minute: Bulk Conversion of NCBI Sequence GI Identifiers to accession.version

*Thursday, December 15, 2016*

Next Wednesday, NCBI will demonstrate how to use a downloadable database and Python script to convert GI identifiers to accession.version. The file and service that will be used are suitable for one-time conversion of very large sets of data.

**Date and time:** Wednesday, December 21, 2016 12:00 PM - 12:15 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/6267508028097746946>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## Variant interpretations from Illumina double ClinVar data

*Monday, December 12, 2016*

On Dec 7, 2016, ClinVar released [138,334 variant interpretations](#) submitted by Illumina Clinical Services Laboratory (ICSL) in San Diego, CA. This dataset represents a 57% increase in the number of submitted interpretations and makes ICSL the largest source of data in ClinVar. The contribution from ICSL also provides 78,590 novel variants to the database, an increase of 45% over the previous total of 173,782 variants.

The data were generated from clinical whole genome sequencing performed in the ICSL; variants were interpreted when the associated gene was in a predefined list of genes associated with Mendelian disorders or when the gene-disease relationship had been manually curated. Information about the criteria that ICSL uses to interpret variants is available on the [NCBI website](#).

ICSL has shared data through other NCBI resources as well, including the [Genetic Testing Registry](#) and the [GeT-RM browser](#).

[ClinVar](#) is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. Interpretations of variants are submitted to ClinVar by clinical testing laboratories, research laboratories, locus-specific databases, genetics clinics, expert panels, and professional societies that establish practice guidelines. The database currently holds 374,018 submitted interpretations representing 263,220 variants. ClinVar provides a public forum for variant interpretations and evidence, so that interpretations may be shared and subjected to peer review.

## The new Human Genome Resources site: a portal for exploration of the human genome

*Monday, December 12, 2016*

The new [Human Genome Resources site](#) offers access to visualization and analysis tools available for the human genome, as well as other relevant tools like [BLAST](#), the [NCBI remapping service](#) and databases that provide human molecular data. The resources are sorted into categories like Find, View, Download and Learn, making it easier to find what you need.

Some specific goals that you can accomplish through the site's guidance are:

- Finding information on individual genes that [NCBI RefSeq staff annotate on the human genome assemblies](#) and are archived in the [Gene database](#).
- Visualizing and analyzing the genome by accessing individual chromosomes in the [Genome Data Viewer](#) and other available viewers.
- Comparing your sequences with the sequences of the [human genome assemblies \(BLAST\)](#).
- Navigating to the clinical and variation data through the complete listing of NCBI's clinical and variation resources.
- Accessing details about the human genome assemblies and annotation.
- Accessing various large datasets for download on the NCBI FTP site.
- Remapping annotation data between different assemblies ([NCBI Genome Remapping Service](#)).

In addition, the portal includes an extensive listing of learning resources that may help you have a better understanding of the wealth of information associated with the human genome.

## Sequence Viewer 3.18 is now available

*Thursday, December 08, 2016*

[Sequence Viewer 3.18](#) has several new features, improvements and bug fixes, including improved handling of [translation discrepancies](#), a new option for "Left-to-right translations" for the six frame translation track, and improved code generation. For a full list of changes, see the [Sequence Viewer release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## New on NCBI Insights: Converting GI Numbers to Accession.version

*Tuesday, December 06, 2016*

The latest blog post on [NCBI Insights](#) shows users how to convert GI numbers to accession.version with EFetch.

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## NCBI Tech Talk and Booth at the American Society for Cell Biology 2016 National Meeting

*Friday, December 02, 2016*

NCBI staff will be participating in the [ASCB 2016 National Meeting](#) from Sunday, December 4 to Tuesday, December 6. We will be at booth #939 from 9AM -4PM PDT Sunday to Tuesday, and will present a Tech Talk on Sunday, December 4 at 5:30PM PDT in Theater 2.

The Tech Talk, "Five Useful Teaching Examples Using NCBI BLAST", will present demonstrations that highlight features of BLAST. These readily adaptable examples are useful for teaching biology principles and techniques including evolution, gene expression analysis and more.

The image shows the NCBI Human Genome Resources webpage. At the top, there is a navigation bar with the NIH logo, 'U.S. National Library of Medicine', the NCBI logo, 'National Center for Biotechnology Information', and a 'Log in' link. Below this is a main header 'Human Genome Resources' with sub-links for 'Find', 'View', 'Download', and 'Learn'. The central part of the page features a 'Search for Human Genes' section with a search input field and a 'Search' button. Below the search bar is a karyotype visualization of human chromosomes, labeled 1 through 22, X, Y, and MT. A text prompt below the karyotype says 'Select a chromosome to access the genome data viewer'. A large downward-pointing arrow is centered below the karyotype. The 'Find' section is divided into three columns: 'Clinical & Variation Resources' featuring 'ClinVar' (Information about genomic variation and its relationship to human health), 'Assemblies & Annotations' featuring 'Genome Reference Consortium (GRC)' (Information on assembly updates and issues from the international collaboration maintaining the human genome), and 'Related Resources' featuring 'RefSeq' (Reference sequences for genomes, transcripts, proteins and more).

Figure 1. The new NCBI Human Genome Resources webpage.

## NCBI News, November 2016

### Evidence Viewer has been retired

*Tuesday, November 22, 2016*

[Evidence Viewer](#), which was designed to show biological evidence supporting curated gene models, has been retired. Current evidence supporting annotated gene structures is included in the "Genomic regions, transcripts and products" section of the [Gene database](#) for organisms annotated using NCBI's [Eukaryotic Genome Annotation Pipeline](#).

The RNA-Seq exon coverage tracks in the graphical [Sequence Viewer](#) show the aggregate exon and intron coverage and individual intron features based on RNA-Seq data in the [SRA database](#). For an example, take a look at the information displayed on the [Human EGFR gene page](#).

### NCBI, NLM, NHGRI to hold on-campus hackathon January 9-11

*Thursday, November 17, 2016*

From January 9th to 11th, the NCBI, with involvement from several NIH institutes, will host a biomedical data science hackathon at the National Library of Medicine. To apply for this hackathon, complete this [application](#) (approximately 10 minutes to complete). Applications are due **December 7th by 4 PM ET**.

This hackathon will primarily focus on:

- Medical informatics,
- Advanced bioinformatics analysis of next generation sequencing data,
- And metadata.

This event is for students, postdocs and investigators or other researchers already engaged in the use of medical informatics data or pipelines for genomic analyses from next generation sequencing data. However, there are some projects available to other non-scientific developers, mathematicians or librarians. The event is open to anyone selected for the hackathon who is able to travel to NIH.

### Organization

There will be 5-7 teams of 5-6 individuals. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure.

The potential subjects for this iteration are:

- HL-7 compliance of myfamilyhealthportrait and GTR,
- Integrating cbioportal with TCGA and dbGaP metadata,
- Variant and gene screening from PubMed,
- Code discovery in PubMed Central,
- Integration of PubChem with other chemical datasets,
- Auto-updating dictionaries for natural language processing,
- And several others.

Please see the [application](#) for specific and evolving team projects. Again, some projects are available to other non-scientific developers, mathematicians or librarians.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

## Datasets

Datasets will come from the public repositories, primarily those housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis.

Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose.

A manuscript outlining the design and usage of the software tools constructed by each team may be submitted to an appropriate journal such as the [F1000Research hackathons channel](#).

## Application

To apply, complete this [application](#) (approximately 10 minutes to complete). Applications are due **December 7th, 2016 by 4 PM ET**.

Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

The first round of accepted applicants will be notified on December 9th by 5 pm ET, and have until December 12th at 4 PM to confirm their participation.

If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event.

Please include a monitored email address, in case there are follow-up questions.

## Notes

Participants will need to bring their own laptop to this program.

A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.

Applicants must be willing to commit to all three days of the event.

No financial support for travel, lodging or meals is available for this event.

Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## Genome Workbench 2.11.5 now available

*Wednesday, November 16, 2016*

The latest version of [Genome Workbench](#) includes a number of new features, fixes and improvements like the use of encrypted HTTPS protocol, multiple feature table loading, and improved exporting.

For a full list of changes, please see the [Genome Workbench release notes](#).

## November 17th webinar: NCBI Resources for Agricultural Research

*Tuesday, November 08, 2016*

On November 17th, NCBI will present a workshop for researchers interested in agriculturally important organisms.

**Date and time:** Thursday, November 17, 2016 1:00 – 2:00 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/42857142156965378>

In the first part of the webinar, participants will learn to effectively use the NCBI website and BLAST to find relevant data including sequence, variation, gene and expression information. The second part of the webinar will focus on accessing large-scale genomics datasets.

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## RefSeq release 79 now available

*Monday, November 07, 2016*

RefSeq release 79 is now accessible online, via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript, and protein data available as of October 31, 2016 and contains 111,024,999 records, including 73,099,060 proteins, 16,967,019 RNAs, and sequences from 64,277 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

As announced in March 2016, NCBI has implemented the removal of GI numbers from some presentations of nucleotide and protein sequence records. The affected presentations are the (default) flat file (GenBank and GenPept) views, along with FASTA views. This change affects only the web views of these two file formats. Note that this change is not reflected in this ftp release. However, GI sequence identifiers will be removed from flatfile and FASTA formats in the March 2017 RefSeq FTP release.

For more details on GI numbers, please see these announcements:

- [The Future of Existing GI Numbers at NCBI](#)
- GI numbers will be removed from sequence record presentations

## New NCBI Insights post: Identifying and Correlating Chemical Names and Synonyms

*Monday, November 07, 2016*

[Identifying and Correlating Chemical Names and Synonyms](#) is the latest post on the NCBI Insights blog. This blog post briefly describes the PubChem system of identifying and correlating varying names for one chemical.

## New NCBI Insights blog post: Clearing Up Confusion with Human Gene Symbols and Names

*Monday, November 07, 2016*

The newest [blog post on NCBI Insights](#) shows you how to use NCBI resources to find and reference official gene names or symbols, as well as synonyms or aliases that refer to the same gene. The blog post also includes a video showing how to correctly import gene symbol data into Excel, avoiding the problems caused by the autocorrect and autoformat functions in spreadsheet applications.

## **Permanent redirect to HTTPS will occur on November 10, 2016**

*Wednesday, November 02, 2016*

Starting on November 10th, NCBI will begin a permanent redirect to HTTPS. More specifically, all HTTP traffic for GET and HEAD requests will be redirected. All other requests will be rejected.

The [HTTPS at NCBI page](#) provides further guidance for NCBI web API users.

## **New video on YouTube: The New MSA Viewer**

*Tuesday, November 01, 2016*

The newest [video](#) on the [NCBI YouTube channel](#) introduces the new [multiple sequence alignment \(MSA\) viewer](#) for amino acid and nucleotide sequences. This short video demonstrates MSA Viewer's basic functions.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.



## NCBI News, October 2016

### November 9th webinar: PubMed for Clinicians

*Friday, October 28, 2016*

On November 9th, NLM staff will show health care professionals how to search [PubMed](#) for the most relevant and recent literature, explore specific clinical research areas, set up email alerts, and more.

**Date and time:** Wednesday, November 9, 2016 1:00 PM - 2:00 PM EST

**Registration URL:** <http://bit.ly/2feyobc>

After registering for the webinar, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also learn about future webinars on this page.

### NLM In Focus blog profiles Dr. Kim Pruitt, NCBI staff scientist

*Tuesday, October 25, 2016*

The [inaugural article](#) in NLM In Focus's new series on NLM scientists features Kim Pruitt, PhD. Dr. Pruitt is a staff scientist at NCBI; she heads the Reference Sequence Database, better known as [RefSeq](#).

In the article, Dr. Pruitt shares her career trajectory as well as pearls of wisdom for young scientists.

### Genome Workbench 2.11.0 now available

*Friday, October 21, 2016*

The latest version of [Genome Workbench](#) includes a number of new features, fixes and improvements like an improved [ProSplign](#) tool, improvements to Graphical Sequence View and new documentation for [using Tree View](#) and other processes.

For a full list of changes, please see the [Genome Workbench release notes](#).

### GI numbers will be removed from sequence record presentations

*Monday, October 17, 2016*

As announced in March 2016, NCBI is now in the process of removing GI numbers from the presentations of nucleotide and protein sequence records. The affected presentations are the (default) flat file (GenBank and GenPept) views, along with FASTA views. NCBI will be releasing these changes on or soon after October 17, 2016.

For web presentations, please see the previous announcement for examples of the new formats. If you would like to obtain formats that include GI numbers, there will be an option in the "Send" menu that allows you to download the former presentations. Once you select "File" as the destination, a "Show GI" checkbox will appear and will be checked by default. When this box is checked, the downloaded data will contain GI numbers as in the past.

These changes also affect presentations obtained using the E-utility EFetch. By default, the new flat file and FASTA presentations will no longer contain GI numbers, just like those obtained on the web. To obtain

presentations that include GI numbers, simply add the parameter & showgi to your EFetch URL and set its value to 1: <https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=nuccore&id=AF123456.2&rettype=gb&showgi=1>.

## New YouTube video: NCBI Staff at ASHG 2016

*Monday, October 17, 2016*

The newest video on the [NCBI YouTube channel](#) is a recording of the October 5th NCBI Minute, which provides a quick overview of NCBI activities at this year's upcoming meeting related to [ClinVar](#), [dbGaP](#), [GRCh38](#) and other topics. In this video, you'll find out why NCBI goes to scientific meetings and how ASHG attendees benefit from having NCBI at ASHG.

NCBI staff members will be at Exhibit booth 521, where attendees can get answers and provide input for the future development of NCBI human genome resources.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## October 26th NCBI Minute: New BLAST Databases Provide Cleaner Results

*Tuesday, October 11, 2016*

On October 26th, NCBI staff will introduce two new BLAST databases: the RefSeq Representative Genomes database and the Model Organisms or Landmark protein database.

**Date and time:** Wednesday, October 26, 2016 12:00 PM - 12:30 PM EDT

**Registration URL:** <http://bit.ly/2di8d0i>

The RefSeq Representative Genomes database contains the NCBI-selected Reference and Representative Genome nucleotide assemblies. The Model Organisms or Landmark protein database contains proteomes from 27 well-characterized model organisms from diverse taxonomic groups.

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also learn about future webinars on this page.

## Multiple Sequence Alignment Viewer 1.1 is now available

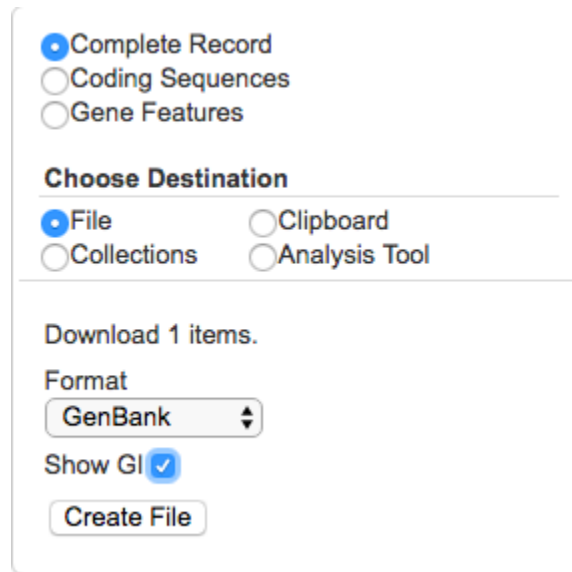
*Thursday, October 06, 2016*

The NCBI [Multiple Sequence Alignment Viewer \(MSA Viewer\)](#) has recently been updated. This new version has an improved rendering mechanism for displaying projected features, improved navigation over gaps in alignment, and improved security and compatibility with HTTPS protocols. A full list of new features, improvements, and bug fixes is available in the [MSA Viewer release notes](#).

The NCBI MSA Viewer is a graphical display for the multiple alignments of nucleotide and protein sequences.

## NLM presents "Insider's Guide" webinar on E-utilities and PubMed on October 19th

*Thursday, October 06, 2016*



Complete Record  
 Coding Sequences  
 Gene Features

**Choose Destination**

File  Clipboard  
 Collections  Analysis Tool

Download 1 items.

Format  
GenBank

Show GI

Create File

**Figure 1.** Revised Send dialog in sequence databases. After choosing "File" as the destination, a new checkbox labeled "Show GI" will appear. When checked, the downloaded data will contain GI numbers as in the past. When unchecked, flat file (GenBank, GenPept) and FASTA formats will not display GI numbers.

On Wednesday, October 19, 2016, the [U.S. National Library of Medicine \(NLM\)](#) will present "Insider's Guide to Accessing NLM Data: Welcome to E-utilities for PubMed".

**Date and time:** Wednesday, October 19, 2016, 2-3 PM EDT

**Registration URL:** <http://bit.ly/2dHKTdy>

Join us for the first in the Insider's Guide series of webinars about more powerful and flexible ways of accessing NLM data, starting with an introduction to the Application Programming Interfaces (APIs) for [PubMed](#) and other NCBI databases.

This series is geared toward librarians and other information specialists who have experience using PubMed via the traditional web interface, but now want to dig deeper. For more information about the first webinar, visit the [registration page](#).

For more information about the Insider's Guide courses, visit the [class page](#).

Questions? Contact [NLM](#).

## CCDS release 20 for human now public in Gene

*Tuesday, October 04, 2016*

[Consensus Coding Sequence \(CCDS\)](#) release 20 compares NCBI's *Homo sapiens* annotation release 108 to Ensembl's release 85, and is now available in [Gene](#). This update adds 1,158 new CCDS IDs and 98 genes into the human CCDS set. CCDS release 20 includes a total of 32,524 CCDS IDs that correspond to 18,892 GeneIDs.



## NCBI News, September 2016

### October 11th NCBI Minute: BLAST+ 2.5.0 with Support for HTTPS, accession.version Identifiers and Much More

*Friday, September 30, 2016*

On October 11th, the NCBI Minute will be a discussion of changes made to the BLAST standalone distribution due to the switch to HTTPS and the transition of the sequence databases to accession.version as the primary identifier.

**Date and time:** Tuesday, October 11, 2016 12:00 PM – 12:30 PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/6522718969008808193>

The new BLAST+ 2.5.0 release supports both, and provides support for composition-based statistics with RPSTBLASTN, and has a new taxonomic organism report. You will learn how these changes improve your BLAST searches and the analysis of results.

After registering for the webinar, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

### BLAST+ 2.5.0 released with support for HTTPS, accession.version and more

*Friday, September 30, 2016*

The new version of the [BLAST+ executables](#) offers support for HTTPS, accession.version as the primary sequence identifier, support for composition-based statistics with RPSTBLASTN, and a new taxonomic organism report. See more details about these updates on [BLAST News](#). A full list of new features, improvements and bug fixes is available in the [release notes](#).

### October 5th webinar: NCBI at ASHG 2016

*Friday, September 30, 2016*

Next Wednesday, October 5th, NCBI staff will give a brief overview of our activities at this year's [ASHG meeting](#) related to ClinVar, dbGaP, GRCh38 and other topics, and how these will benefit ASHG attendees.

**Date and time:** Wednesday, October 5, 2016 1:00 PM - 2:00 PM EDT

**Registration URL:** <http://bit.ly/2dKrCZj>

The ASHG annual meeting will happen October 18-22 in Vancouver, British Columbia, Canada.

After registering for the webinar, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## Sequence Viewer 3.16 is now available

*Thursday, September 29, 2016*

Sequence Viewer 3.16 brings several new features, improvements and bug fixes to the graphical viewer, including improved compliance with [HTTPS protocol requirements](#), a new track type, and improved labeling to avoid label duplication. For a full list of changes, see the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## GenBank release 215.0 is now available via FTP

*Wednesday, September 28, 2016*

GenBank release 215.0 (08/19/2016) has 196,120,831 traditional records containing 217,971,437,647 base pairs of sequence data. In addition, there are 359,796,497 WGS records containing 1,637,224,970,324 base pairs of sequence data, as well as 113,179,607 TSA records containing 103,399,742,586 base pairs of sequence data.

During the 66 days between the close dates for GenBank releases 214.0 and 215.0, the traditional portion of GenBank grew by 4,770,529,828 base pairs and by 1,657,259 sequence records. During the same period, 75,882 records were updated at an average of 26,260 traditional records added and/or updated per day.

Between releases 214.0 and 215.0, the WGS component of GenBank grew by 81,049,025,676 base pairs and by 9,518,416 sequence records. The TSA component of GenBank grew by 8,985,783,667 base pairs and by 8,502,546 sequence records.

The total number of sequence data files increased by 57 with this release. The divisions are as follows:

- BCT: 18 new files, now a total of 267
- CON: 7 new files, now a total of 351
- ENV: 1 new file, now a total of 93
- EST: 1 new file, now a total of 481
- GSS: 1 new file, now a total of 301
- HTG: 2 new files, now a total of 153
- INV: 3 new files, now a total of 144
- PAT: 11 new files, now a total of 263
- PLN: 8 new files, now a total of 134
- PRI: 2 new files, now a total of 55
- SYN: 1 new file, now a total of 9
- TSA: 1 new file, now a total of 229
- VRL: 1 new file, now a total of 48

For downloading purposes, please keep in mind that the uncompressed GenBank release 215.0 flatfiles require approximately 790 GB (sequence files only); the ASN.1 data require approximately 650 GB.

More information about GenBank release 215.0 is available in the [release notes](#).

## Genomes-announce listserv reactivated

*Wednesday, September 28, 2016*

The [Genomes-announce email list](#) has been reactivated. To sign up for the Genomes-announce Listserv, go to <https://www.ncbi.nlm.nih.gov/mailman/listinfo/genomes-announce>.

We will use the Genomes-announce Listserv primarily to announce changes to the [NCBI Genomes FTP site](#) and other genome data download channels. Announcements will be made about new data types, changes to data formats or data organization.

Read more about the Genomes-announce Listserv in [this announcement](#).

## NCBI to hold Developers' Forum September 28th

*Friday, September 23, 2016*

On Wednesday, September 28th, at 3 PM EDT, NCBI will hold a developers' forum for those who use large amounts of NCBI data. The forum will help us provide you with better access to NCBI data.

To join the forum, complete this short [survey](#). An invitation will be extended to 21 people.

## October 4-6: Stream the University of Michigan NCBI workshops

*Thursday, September 22, 2016*

On October 4th, 5th, and 6th, the University of Michigan's Taubman Health Sciences Library will host a series of NCBI workshops that can also be streamed remotely. The workshops are: *Navigating NCBI's Molecular Data Using the Integrated Entrez System and BLAST*, *A Practical Guide to NCBI BLAST* and *EDirect: Command Line Access to NCBI's Biomolecular Databases*. Please see the Taubman Health Sciences Library [Remote Site Registration Page](#) for details.

## Introducing Magic-BLAST

*Thursday, September 22, 2016*

Magic-BLAST is a new tool for mapping large sets of next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Magic-BLAST executables for LINUX, MacOSX, and Windows as well as the source files are available on the [FTP site](#).

Each alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-Seq, locating the candidate introns and adding up the score of all exons. Sequencing reads can be provided as NCBI SRA accessions, FASTA or SRA files.

Magic-BLAST implements ideas developed in the NCBI Magic pipeline using the NCBI BLAST libraries. Magic-BLAST is under active development, and we expect the next few releases to occur on a monthly basis. Read more about Magic BLAST on the [FTP site](#).

## Scheduled: Next Round of HTTPS Tests

*Wednesday, September 21, 2016*

We have scheduled another round of HTTPS tests, following up from the initial tests performed on September 15. More information can be found on an [NCBI Insights Blog post](#).

The schedule for these tests is as follows (all times are EDT):

### Thursday, Sept 22

8:00 AM – 12:00 PM : redirect web pages from HTTP to HTTPS, same as the previous September 15 test

8:00 AM – 9:00 AM : redirect CGI's and API calls to HTTPS where possible, reject where not possible

**Monday, Sept 26**

8:00 AM – 10:00 AM : redirect web pages from HTTP to HTTPS with HSTS activated using a 1-hour expiration

10:00 AM – 12:00 PM : redirect web pages from HTTP to HTTPS without HSTS

**Tuesday, Sept 27**

8:00 AM : Start continually redirecting web pages from HTTP to HTTPS

## **NCBI's Bryant and Bolton receive 2016 Herman Skolnik Award for PubChem database**

*Monday, September 19, 2016*

On August 23, Drs. Stephen Bryant and Evan Bolton received the American Chemical Society (ACS) 2016 Herman Skolnik Award for their work in developing, maintaining, and expanding the National Center for Biotechnology Information's PubChem database of chemical substances and their biological activities. The award was presented at the ACS 252nd National Meeting & Exposition in Philadelphia.

The Herman Skolnik award is named after its first recipient, the founder of the Journal of Chemical Information and Computer Sciences, and "recognizes outstanding contributions to and achievements in the theory and practice of chemical information science and related disciplines," according to ACS.

In its announcement of the award, ACS said: "Under Bryant and Bolton's leadership, the PubChem team has created a world-class resource for chemical and biological information. PubChem is the first major public database to connect cheminformatics to bioinformatics and thereby provide a unique information resource for pharmaceutical research."

Introduced in 2004, [PubChem](#) currently includes more than 220 million chemical substance records for 90 million unique compounds. The database also contains biological screening results from more than 1.2 million bioassays for over 3.5 million tested substances. The information in PubChem is the result of collaborations with more than 250 academic and commercial organizations that have contributed their data. PubChem is integrated with many other NCBI other databases, with links to related information, such as compounds with similar structures, protein sequences, and relevant journal articles. This extensive network of links provides users with vast opportunities for exploration and for making discoveries. Each day, tens of thousands of researchers from university labs and pharmaceutical and biotech companies access PubChem.

## **September 21st webinar: Update on NCBI's Transition to HTTPS**

*Monday, September 19, 2016*

Next Wednesday, September 21st, NCBI staff will discuss plans regarding the move to HTTPS-only services. This past week, we conducted the first of a series of HTTPS tests; in this webinar, we will talk about this and future tests that will help all of us prepare for this change. We will also briefly discuss circumstances surrounding proxy services and software dependent upon NCBI software, such as the SRA and C++ toolkits.

**Date and time:** Wednesday, September 21, 2016 12:00 PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/2680765856528138244>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.





Figure 1. Drs. Bryant and Bolton receive the American Chemical Society 2016 Herman Skolnik Award.

## **New video on YouTube: Tree Viewer - Display Large Trees**

*Thursday, September 15, 2016*

The newest video on the NCBI YouTube channel, *Tree Viewer: Display Large Trees*, demonstrates a new functionality in *Tree Viewer* - the ability to display much larger trees.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## **Genomes FTP site update (version 1.3) adds new data formats and more**

*Wednesday, September 14, 2016*

NCBI has released a comprehensive update of all current genome assemblies in the [Genomes FTP site](#), affecting data reported in the `/genomes/genbank/`, `/genomes/refseq/` and `/genomes/all/` FTP directories. This update adds nucleotide FASTA sequences of CDS and RNA features computed from the genome sequence, expands the scope

of the /genbank/ data to include metagenomes, and more. The FTP content of nearly all "latest" GenBank and RefSeq assemblies was updated to reflect these changes between 5/11/2016 and 6/24/2016.

Genomes FTP version 1.3 includes the following changes:

- New files for genome assemblies with annotation:
  - Files named as \*\_cds\_from\_genomic.fna.gz provide nucleotide FASTA sequences corresponding to all CDS features annotated on the assembly, based on the genome sequence
  - Files named as \*\_rna\_from\_genomic.fna.gz provide nucleotide FASTA sequences corresponding to all RNA features annotated on the assembly, based on the genome sequence
- Metagenomes
  - Metagenomes have been added to the Assembly database and a new genome group directory (metagenomes) is now available: <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/metagenomes/>
- Annotation hashes:
  - Reporting hash values and last changed dates for different aspects of the annotation data are useful to monitor for when annotation has changed in a way that is significant for a particular use case and warrants downloading the updated records. These data can be used to monitor for annotation changes compared to previously downloaded files based on comparison of the hash values, or for annotation changes since a particular point in time
  - A file named annotation\_hashes.txt in each genome assembly directory provides hash values and last changed dates for different aspects of the annotation and descriptor data for that assembly
  - A file named annotation\_hashes.txt in each organism group and species directory provides hash values and last changed dates for all assemblies from the organism group or species
- RepeatMasker files:
  - Data in the \*\_rm.out files is now generated using a newer version of RepeatMasker and repeat libraries (change from RepeatMasker version 3.3.0 to 4.0.6 and from RM database version 20120418 to version 20150807)
- GFF3 format:
  - genomic.gff.gz files for RefSeq eukaryotic genomes annotated with NCBI's Eukaryotic Genome Annotation Pipeline now incorporate 1-2 bp gaps or "micro-introns" to compensate for frameshifting indels in mRNA and CDS features where the indel is thought to represent a genome sequencing error and the gene is likely to produce a functional product
  - Some features are now represented with better or more appropriate Sequence Ontology (SO) terms in the GFF3 files, including the following:

#INSDC term	old SO term	new SO term
misc_feature	region	sequence_feature
variation	sequence_variant	sequence_alteration
conflict	region	sequence_conflict
mobile_element	region	mobile_genetic_element
rep_origin	region	origin_of_replication
telomere	region	telomere
centromere	region	centromere
regulatory/enhancer	region	enhancer
regulatory/promoter	region	promoter
GC_signal	GC_rich_region	GC_rich_promoter_region

*Table continued from previous page.*

N_region	region	N_region
S_region	region	S_region
V_region	region	V_region
unsure	region	sequence_uncertainty
virion	region	viral_sequence

- Feature table report:
  - Small improvements in the feature\_table.txt.gz file, including fixing the occurrence of "?" strand
- GBFF format:
  - Small improvements in formatting of the GenBank flatfiles
- Assembly summary files:
  - A column reporting "Excluded from RefSeq" reasons has been added
- Assembly reports:
  - More metadata fields were added to the headers of the assembly, statistics and regions reports
  - Headers no longer include the the latest/suppressed/replaced status of the assembly
- Gzip compression:
  - A subtle change in gzip compression that has no effect on file contents but does subtly alter file sizes and md5checksums

Additional information about the genomes FTP site can be found in the [genomes FTP README file](#) and in the [genomes FTP FAQ](#). Subscribe to the [genomes-announce mail list](#) to be informed of changes to the NCBI genomes FTP site.

## RefSeq release 78 is now available

*Monday, September 12, 2016*

RefSeq release 78 is accessible online, via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript, and protein data available as of September 6, 2016 and contains 107,045,797 records, including 70,427,238 proteins, 16,172,490 RNAs, and sequences from 62,739 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

More information about release 78 can be found in the [release notes](#). For more information about the RefSeq project, please see the [RefSeq homepage](#).

## New on NCBI Insights: The Future of Existing GI Numbers at NCBI

*Monday, September 12, 2016*

The latest [blog post on NCBI Insights](#) discusses what will happen to existing GI numbers in records now that NCBI is phasing out Sequence GIs.

NCBI Insights is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## Leiden Open Variation Database to be retired September 30, 2016

*Thursday, September 08, 2016*

NCBI is retiring the [Leiden Open Variation Database \(LOVD\)](#) on September 30, 2016. LOVD has been used to capture information about novel human variants. We encourage past submitters of human genetic variations to LOVD to transfer their information to the [ClinVar database](#).

If you would like to add new human variation data, please review our [instructions on submitting to ClinVar](#). You may find our [submission wizard](#) eases the process.

While the LOVD site will be retired on September 30, 2016, an [FTP archive](#) will continue to store LOVD data for download after this date.

## **New on NCBI Insights: Find, Browse and Follow Biomedical Literature with PubMed Journals**

*Wednesday, September 07, 2016*

The latest blog post on [NCBI Insights](#) presents the new [PubMed Journals](#), the latest experiment from [PubMed Labs](#). PubMed Journals allows you to easily find and browse journals of interest, browse new articles, and more. Learn more about PubMed Journals, try it out, and leave us feedback on the [blog post](#).

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## **September 7th webinar: The E-Utilities in an Age without GI Numbers**

*Thursday, September 01, 2016*

Next Wednesday, September 7th, NCBI will present a webinar that briefly describes NCBI's future plans for the E-utilities API in a time where GI numbers are no longer used as the primary identifiers for sequence records. You will learn how to convert GI numbers to accession.version identifiers and how to quickly determine the most recent version of an accession. You'll also learn about a new E-utility parameter (to be released this fall) that allows these tools to work only with accession.version identifiers.

**Date and time:** Wednesday, September 7, 2016 12:00 PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/4529251610671340033>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## **October 24-26: Hackathon at Cold Spring Harbor Laboratory**

*Thursday, September 01, 2016*

From October 24th to 26th, Cold Spring Harbor Laboratory (CSHL), with assistance from NCBI, will host a biomedical data science hackathon immediately before the [Biological Data Science Conference](#) at CSHL. The hackathon will primarily focus on writing functional software for advanced bioinformatics analysis of next generation sequencing data and metadata, but also may include analysis of other types of data, such as images or other molecular measurements.



Figure 1. The PubMed Journals homepage.

This event is for students, postdocs and investigators or other already engaged in the creation of pipelines for genomic analyses from next generation sequencing data, imaging data or metadata.\* The event is open to anyone selected for the hackathon and willing to travel to CSHL.\*\*

\* Some projects are available to other non-scientific developers, mathematicians or librarians.

\*\* Attendees of the CSHL Biological Data Science Conference will be given preference, if space is limited. Also, there will be a nominal fee for attendees, partly to cover refreshments during the events.

## Organization

Working groups of 5-6 individuals will be formed into four to five teams. These teams will build novel pipelines, tools, and visualizations to analyze large datasets within a cloud infrastructure. The potential subjects for this hackathon include:

- structural variant identification & analysis,
- genome assembly,
- single cell transcriptome & epigenome analysis,
- ultrafast genomic mapping,
- data encryption,
- deep learning technologies,
- and image analysis.

Please see the [application](#) for specific team projects.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

## Datasets

Most of the datasets will come from the public repositories, primarily those housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose. A manuscript outlining the design and usage of the software tools constructed by each team may be submitted to an appropriate journal such as the [F1000Research hackathons channel](#). Continued development of the technology after the hackathon is also encouraged.

## Application

To apply, complete this [form](#) (approximately 10 minutes to complete). Applications are due **September 16, 2016 by 4 PM ET**. Prior participants and applicants are especially encouraged to reapply.

The first round of accepted applicants will be notified on September 19th by 5 PM ET, and have until September 22nd at 9 AM ET to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.

## Notes

Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python, R) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.

Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals is available for this event is currently available, although attendees will be notified if that changes.

Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## NCBI News, August 2016

### Genomes FTP site data organization to change on September 20, 2016

Tuesday, August 30, 2016

NCBI is moving the contents of the "all" and "ASSEMBLY\_REPORTS/All" directories on the [Genomes FTP site](#). Currently, listing the contents of these two directories is impractical because they contain many thousands of directories or files.

#### Reorganization of <ftp://ftp.ncbi.nlm.nih.gov/genomes/all>

The genome assembly directories currently directly under "all" will be moved into a new 4-level structure under [genomes/all](#).

Two new directories under "all" will be named for the accession prefix (GCA or GCF). These directories will contain another three levels of directories named for digits 1-3, 4-6 & 7-9 of the assembly accession, creating paths like *genomes/all/GCA/xxx/xxx/xxx/* and *genomes/all/GCF/xxx/xxx/xxx/*. For example:

- The data currently in *genomes/all/GCA\_000001405.23\_GRCh38.p8* will be moved to *genomes/all/GCA/000/001/405/GCA\_000001405.23\_GRCh38.p8*.
- The data currently in *genomes/all/GCF\_001696305.1\_UCN72.1* will be moved to *genomes/all/GCF/001/696/305/GCF\_001696305.1\_UCN72.1*.

#### Schedule of changes

On September 20, 2016:

- New directories [genomes/all/GCA](#), [genomes/all/GCF](#) and the three levels of directories named for groups of digits in the assembly accession will be added.
- Individual genome assembly data directories directly under [genomes/all](#) will be moved into the new directory structure under [genomes/all/GCA](#) & [GCF](#).
- Assembly data directories directly under [genomes/all](#) will be replaced by symbolic links to the corresponding directory in the new structure.
- The old and new data organizations will be maintained in parallel for 6 weeks.

On December 1, 2016:

- The old paths to individual genome assembly data directories directly under [genomes/all](#) will be removed.
- All access to genome assembly data under [genomes/all/](#) will need to use the [genomes/all/GCA/xxx/xxx/xxx/](#) & [genomes/all/GCF/xxx/xxx/xxx/](#) paths.

#### Impact

Users who access genome assembly data by any of the following methods will not be affected by this change:

- Following a link to "Download the GenBank assembly" or "Download the RefSeq assembly" from an Assembly details page
- Navigating the [genomes/genbank](#) or [genomes/refseq](#) paths of the genomes FTP site
- Using the `ftp_path` provided in the `assembly_summary.txt` files provided on the genomes FTP site

Users who mirror all data under [genomes/all](#) will get two copies of the data for each genome assembly during the transition period, unless they modify their scripts to only take data from [genomes/all/GCA](#) & [GCF](#).

Other changes:

- Scripts that retrieve data using hard-coded paths to individual genome assembly directories directly under `genomes/all` will fail after the transition period
- Links from non-NCBI web pages to individual genome assembly directories directly under `genomes/all` will fail after the transition period
- Published paths to individual genome assembly directories directly under `genomes/all` will fail after the transition period

## Removal of [ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY\\_REPORTS/All](ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/All)

First, the assembly reports currently under `genomes/ASSEMBLY_REPORTS/All` will be moved into the assembly data directories in the new directory hierarchy under `genomes/all/GCA` & `genomes/all/GCF` described above, replacing the symbolic links to the assembly report files that currently exist in this location. The assembly report files in the assembly data directories will retain the name previously provided by the symbolic link.

- `{assembly_accession.version}.assembly.txt` will appear as `{assembly_accession.version}_{assembly_name}_assembly_report.txt`
- `{assembly_accession.version}.stats.txt` will appear as `{assembly_accession.version}_{assembly_name}_assembly_stats.txt`
- `{assembly_accession.version}.regions.txt` will appear as `{assembly_accession.version}_{assembly_name}_assembly_regions.txt`

Then, the `genomes/ASSEMBLY_REPORTS/All` directory will be removed.

## Schedule

On September 20, 2016:

- The assembly reports currently under `genomes/ASSEMBLY_REPORTS/All` will be moved into the assembly data directories, replacing the symbolic links currently in the data directories.
- The assembly reports under `genomes/ASSEMBLY_REPORTS/All` will be replaced by symbolic links to the corresponding report in the assembly data directory.
- The old and new data organizations for assembly reports will be maintained in parallel for 6 weeks.

On December 1, 2016:

- The old paths to assembly reports under `genomes/ASSEMBLY_REPORTS/All` will be removed.
- The `genomes/ASSEMBLY_REPORTS/All` directory will be removed.
- All access to assembly reports will need to use the `genomes/all/GCA/`, `genomes/all/GCF`, `genomes/genbank` or `genomes/refseq` paths to the individual assembly data directories.

## Impact

Users who access assembly reports by any of the following methods will not be affected by this change:

- Following a link to "Download the full sequence report" from an Assembly details page
- From an assembly data directory under the `genomes/genbank` or `genomes/refseq` path on the genomes FTP site

Attempts to access assembly reports using the `genomes/ASSEMBLY_REPORTS/All` path will fail after the transition period.

Additional information about the genomes FTP site can be found in the [genomes FTP README file](#) and in the [genomes FTP FAQ](#).



Subscribe to the [genomes-announce mail list](#) to be informed of changes to the NCBI genomes FTP site.

## dbVar July 2016 data release includes new 1000 Genomes Phase III structural variants

*Monday, August 29, 2016*

The dbVar July 2016 data release includes 1,455,032 new Variant regions, 13,961,956 Variant calls and 6 new studies. See a list of the studies, including descriptions and links to the data in the [release notes](#).

Follow the dbVar [RSS feed](#) for monthly releases.

## August 31st NCBI Minute: Downloading Genome Data from the NCBI FTP Site

*Thursday, August 25, 2016*

In the next NCBI Minute, we will teach you how to use the Web and the command line to quickly access and download genomic sequence and annotation files for a species, metagenome or taxonomic group of interest.

**Date and time:** Wednesday, August 31, 2016 12:00 PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/8835228315982188801>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## VAST+ update provides refined alignments

*Tuesday, August 23, 2016*

The new version of VAST+ provides a [refined structure-based alignment](#) of similar macromolecular complexes and displays the 3D superpositions in the recently launched iCn3D.

See the [MMDB news page](#) for more detail about how VAST+ now works.

## September 12th class at NLM: EDirect - Command Line Access to NCBI's Biomolecular Databases

*Monday, August 22, 2016*

On September 12, 2016, NCBI staff will discuss EDirect in a class at the [National Library of Medicine](#). During the optional first hour of this workshop (9-10 AM), you will get a basic introduction to the Unix/Linux command line interface. The main workshop (10 AM - Noon) will cover how to use EDirect to set up simple pipelines to retrieve and process data from [PubMed](#), [Gene](#) and the [Nucleotide](#) and [Protein](#) sequence databases. We will provide access to EDirect installed in a Linux environment on a cloud service.

**Date and time:** Monday, September 12, 2016 9:00 AM EDT

**Registration link:** <https://www.surveymonkey.com/r/5NCWLK6>

NOTE: This is an in-person class at the National Library of Medicine on the NIH campus in Bethesda, MD, USA. **The course is limited to 22 participants. Participants must bring their own laptop.**

The EDirect suite of programs allows easy command line access for searching and retrieving literature (PubMed) and accessing NCBI's biomolecular (Gene, Nucleotide, sequence databases, etc.) records. Its advantages include direct command line access to NCBI's databases without writing Perl or Python scripts, construction of custom pipelines for processing data, built-in batch access, and the ability to generate highly flexible custom output reports.

## HIV-1 datasets in Gene updated

*Thursday, August 11, 2016*

NCBI has updated the HIV-1 interaction datasets available in Gene with data provided by the [Southern Research Institute](#).

The [protein interactions dataset](#) now has:

- 7,762 interactions;
- 15,665 interaction descriptions;
- 3,729 proteins encoded by 3,649 human genes;
- and 6,690 publications.

The [replications interactions dataset](#) now has:

- 1,325 interactions;
- 1,439 interaction descriptions;
- 1,325 proteins encoded by 1,325 human genes;
- and 125 publications.

Data are also available at the [RefSeq HIV-1 website](#) and the [GeneRIF FTP site](#).

## NCBI News, July 2016

### HTTPS at NCBI: Guidance for NCBI web API users

*Wednesday, July 27, 2016*

As originally announced on June 10, NCBI will be moving all web services to the HTTPS protocol on September 30, 2016. Particularly for API users, this move may disrupt any processes that access NCBI APIs using the HTTP protocol. Please see [this document](#) on the [Develop action page](#) for a complete discussion of this move and what you need to do, along with some new test servers to help you confirm whether your code will function after the change to HTTPS.

### dbSNP build 148 for corn, fruit fly, rice and 8 other organisms available

*Tuesday, July 26, 2016*

dbSNP build 148 is accessible on the web and via [FTP](#). This release includes data for cat, corn, cow, fruit fly, grape, horse, rice, sorghum, tomato, turkey and zebra finch. Build 148 provides over 446 million submitted variants and 225 million reference variants for 14 organisms. To see complete build statistics, visit the [SNP summary page](#).

dbSNP, the NCBI Short Genetic Variations database, catalogs short variations in nucleotide sequences from a wide range of organisms.

### August 3rd webinar: NCBI Targeted Loci: RefSeq Ribosomal RNA Sequences for Identification and Phylogenetic Analysis

*Thursday, July 21, 2016*

On August 3rd, NCBI staff will present a webinar on our [targeted loci project](#). You'll learn about the scope of the project and see practical examples of using these data, [BLAST](#) and [MOLE-BLAST](#) to identify organisms and explore their diversity in sequences from environmental and organism-associated communities.

**Date and time:** Wednesday, August 3, 2016 12:00 PM EDT

**Registration:** <https://attendee.gotowebinar.com/register/3690313242319181569>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

The targeted loci project is an NCBI-curated set of marker rRNA sequences for prokaryotes (16S, 5S, 23S) and fungi (18S, 28S and ITS). There are now over 27,000 markers representing diverse sets of bacteria, archaea and fungi.

### Tree Viewer 1.10 visualizes large phylogenetic trees up to 100,000 nodes

*Tuesday, July 19, 2016*

[Tree Viewer version 1.10](#) has several improvements, updates and bug fixes, including the ability to visualize large phylogenetic trees up to 100,000 nodes. In addition, Tree Viewer 1.10 has added functionality for sorting, as well as improved API and zoom functions. The Tree Viewer [release notes](#) list all updates.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## **NCBI Insights blog post: NCBI is Phasing out Sequence GIs - Here's What You Need to Know**

*Friday, July 15, 2016*

The latest [blog post on NCBI Insights](#) answers two questions you may have regarding the switch to accession.version:

1. What pieces of your code will break in September?
2. Are GI numbers gone for good?

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## **July 27th NCBI Minute: Important Changes to NCBI Web Protocols**

*Wednesday, July 13, 2016*

In two weeks, we'll discuss NCBI's upcoming switch to the secure HTTPS protocol. Through this NCBI Minute, you'll learn how this change will affect your access to NCBI pages and services and what you should do to have a smooth transition.

**Date and time:** Wednesday, July 27, 2016 12:00 PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/2000297899730334722>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## **Sequence Viewer 3.15 is now available**

*Wednesday, July 13, 2016*

[Sequence Viewer 3.15](#) brings several new features, improvements and bug fixes to the graphical viewer, including improved PDF graphics for markers, an extended embedding API and improved tooltips. For a full list of changes, see the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## **July 20th NCBI Minute: Important Changes Coming to Sequence Databases**

*Tuesday, July 12, 2016*

In next Wednesday's NCBI Minute, NCBI staff will describe upcoming changes to sequence records. Starting this September, GI identifiers will no longer appear on GenBank and FASTA record views; we will discuss the details and consequences of this change, along with the future of existing GI identifiers.

**Date and time:** Wednesday, July 20, 2016 12:00 PM EDT

**Registration URL:** <https://attendeegotowebinar.com/register/6741098469079346177>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also learn about future webinars on this page.

## Conserved Domain Database (CDD) version 3.15 now available online and via FTP

*Tuesday, July 12, 2016*

Version 3.15 of the [Conserved Domain Database](#) contains 290 new or updated NCBI-curated domains (52,411 total), including models specifically built to annotate structural motifs (accession prefix "sd"), and now mirrors Pfam version 28.

Updates include:

- Fine-grained classification of the beta lactamase-like metallohydrolases
- Conserved domain hits in CD Search are ranked by E-value, without giving preference to NCBI-curated models.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#).



## RefSeq release 77 is now available

*Thursday, July 07, 2016*

RefSeq release 77 is accessible [online](#), via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript and protein data available as of June 29, 2016 and includes:

- 100,678,438 records,
- 65,964,245 proteins,

- 15,563,994 RNAs,
- and sequences from 60,892 organisms.

The release is provided in several directories as a complete dataset and also as divided by logical groupings.

Additional information about release 77 can be found in the [release notes](#). You can also directly receive statistics and a summary of announcements for each RefSeq release by subscribing to the [refseq-announce listserv](#).

## Mouse and zebrafish genome annotations updated

*Wednesday, July 06, 2016*

The mouse ([GRCm38.p4](#)) and zebrafish ([GRCz10](#)) genomes were recently re-annotated by the [Eukaryotic Genome Annotation Pipeline](#). For both, the annotation was performed on the RefSeq assemblies' top-level sequences (chromosomes and unlocalized and unplaced scaffolds). See our previous announcement for details about this change.

### Mouse (*mus musculus*)

The new [annotation](#) includes 2,378 "known" RefSeq transcripts that are new or were modified since the previous release. The proportion of protein-coding genes represented by at least one "known" RefSeq is now 91%.

See [Mus musculus Annotation Release 106](#) in [Gene](#), [BLAST](#), or via [download](#).

### Zebrafish (*Danio rerio*)

The new [annotation](#) includes 2,364 annotated "known" RefSeq transcripts that are new or were modified since the previous release. The proportion of protein-coding genes represented by at least one "known" RefSeq is now 55%.

See [Danio rerio Annotation Release 105](#) in [Gene](#), [BLAST](#), or via [download](#).

**Note:** Genes spanning adjacent scaffolds may now be represented as a single feature. See, for example, mouse genes [Dock2](#), [Rims1](#), or [Immp2l](#) and zebrafish [nduaf11](#).

You can find all annotated organisms on the [Eukaryotic Genome Annotation Pipeline](#) page.

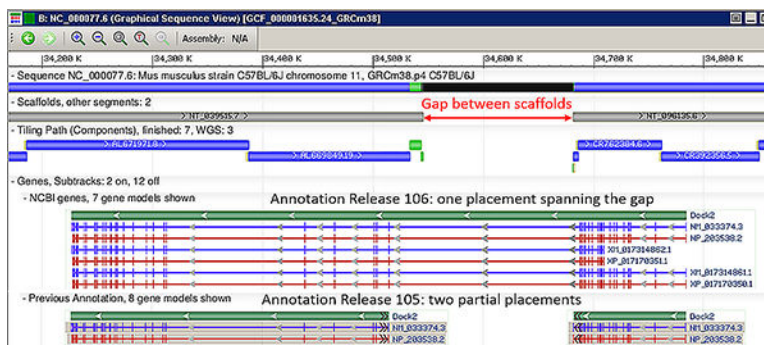


Figure 1. Annotation of mouse Dock2 in Annotation Release 106 and 105





## NCBI News, June 2016

### Genome Workbench 2.10.7 now available

Thursday, June 30, 2016

Genome Workbench 2.10.7 brings a number of new features and fixes like added support for local custom BLAST databases and improvements to Tree View.

For the full list of changes, see the [release notes](#).

### July 6th NCBI Minute: Quickly Find Coding Sequences Using ORFfinder

Monday, June 27, 2016

In one week, we'll show you how to use the redesigned Open Reading Frame Finder (ORFfinder) to quickly identify and analyze complete coding regions on prokaryotic genomic and eukaryotic mRNA sequences.

**Date and time:** Wednesday, July 6, 2016 12:00 PM EDT

**Registration:** <https://attendee.gotowebinar.com/register/8131176934772138753>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also learn about future webinars on this page.

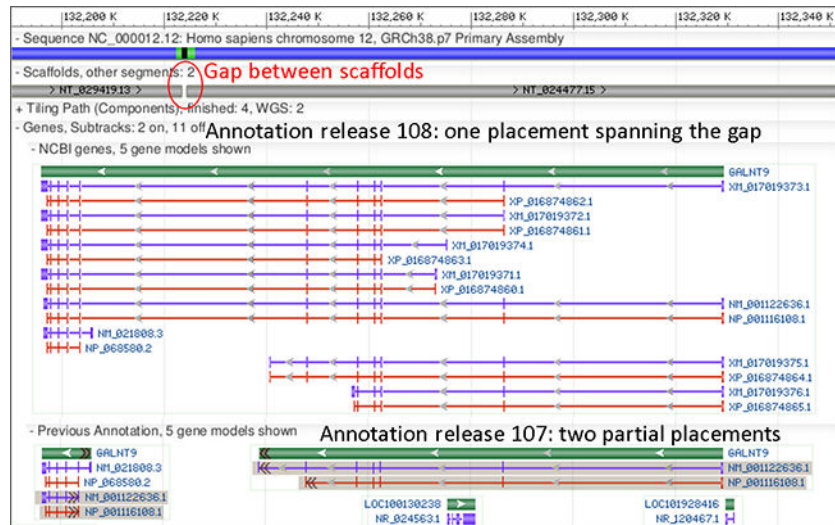
### Human genome Annotation Release 108 incorporates new RefSeq sequences, predicts new variants

Friday, June 24, 2016

The [Eukaryotic Genome Annotation Pipeline](#) recently produced an updated annotation of the human genome. This new annotation, available in RefSeq, incorporates new "known" RefSeq sequences and predicts new alternative variants. For the first time, this annotation was performed on the assemblies' top-level sequences (chromosomes and unlocalized and unplaced scaffolds). See our previous announcement for details about this change.

Noteworthy improvements in Annotation Release 108 include:

- Assemblies annotated: latest GRC assembly, [GRCh38.p7](#) (GCF\_000001405.33, reference) and [CHM1\\_1.1](#) (GCF\_000306695.2)
- RNA-Seq datasets used: The Human Protein Atlas ([PRJEB4337](#)) and new projects, including the multi-tissue [PRJNA280600](#) and fetal development project [PRJNA270632](#)
- 11% increase in coding transcripts for GRCh38.p7 (4,424 known RefSeq and 6,473 model RefSeq)
- 9% increase in non-coding transcripts (1,828 known RefSeq and 2,096 model RefSeq)
- Improved gene annotation across inter-scaffold gaps (for example, [GALNT9](#), [GRK1](#) or [CAPN8](#))



**Figure 1.** Annotation of GALNT9 in Annotation Releases 108 and 107

See Annotation Release 108 in [Gene](#), [BLAST](#), or download it via [FTP](#).

You can find all annotated organisms on the [Eukaryotic Genome Annotation Pipeline](#) page.

## GenBank release 214.0 is now available via FTP

*Thursday, June 23, 2016*

GenBank [release 214.0](#) (06/14/2016) has 194,463,572 traditional records containing 113,200,907,819 base pairs of sequence data. In addition, there are 350,278,081 WGS records containing 1,556,175,944,648 base pairs of sequence data, as well as 104,677,061 TSA records containing 94,413,958,919 base pairs of sequence data.

During the 61 days between the close dates for GenBank releases 213.0 and 214.0, the traditional portion of GenBank grew by 1,776,995,772 base pairs and by 724,061 sequence records. During the same period, 313,786 records were updated at an average of 17,013 traditional records added and/or updated per day.

Between releases 213.0 and 214.0, the WGS component of GenBank grew by 103,968,239,699 base pairs and by 11,353,544 sequence records. The TSA component of GenBank grew by 6,602,795,243 base pairs and by 6,529,495 sequence records.

The total number of sequence data files increased by 32 with this release. The divisions are as follows:

- BCT: 11 new files, now a total of 249
- CON: 10 new files, now a total of 344
- ENV: 1 new file, now a total of 92
- INV: 5 new files, now a total of 141
- PAT: 1 new file, now a total of 252
- PLN: 1 new file, now a total of 126
- VRL: 2 new files, now a total of 42
- VRT: 1 new file, now a total of 61

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 778 GB (sequence files only); the ASN.1 data require approximately 641 GB.

More information about GenBank release 214.0 is available in the [release notes](#).

## June 29th webinar: Downloading Exon and Coding Region Sequences for Genes

*Monday, June 20, 2016*

Next Wednesday, June 29th, NCBI staff will show you how to use the Gene Table report and Graphical Viewer to retrieve exon sequences for genes.

**Date and time:** Wednesday, June 29, 2016 12:30 PM EDT

**Registration:** <https://attendee.gotowebinar.com/register/859302797289474817>

You will also see how to retrieve all exon sequences at once and for multiple genes using the EDirect command line interface to the Entrez search and retrieval system.

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## International HapMap Browser to go offline June 16, 2016

*Thursday, June 16, 2016*

Due to computer security flaws within the HapMap site, it has been decommissioned as of today, June 16, 2016. We regret any inconvenience this sudden removal may cause, but we are acting quickly to ensure security.

The archived HapMap data will continue to be available via [FTP](#). Users can also access the latest data for HapMap samples from the [NCBI 1000 Genomes Browser](#).

**See also:** [NCBI Variation - NCBI retiring HapMap Resource](#).

## NCBI to hold hackathon on NIH campus in August

*Tuesday, June 14, 2016*

The NCBI and several NIH institutes will host a biomedical data science hackathon at the National Library of Medicine from August 15th to 17th. To apply for this hackathon, complete this [form](#) (approximately 10 minutes to complete). **Applications are due July 11th, by 4PM ET.**

This hackathon will primarily focus on advance bioinformatics analysis of next-generation sequencing data and metadata.

This event is for students, postdocs and investigators or other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata; it is open to anyone selected for the hackathon who can travel to NIH.

### Organization

Participants will be grouped into five to seven teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure.

The potential subjects for this iteration are:

- Calling CNVs
- Parsing large scale bacterial samples

- Statistical analysis of variants mapped to biological networks
- Integration of TCGA and dbGaP metadata
- HL-7 compliance of myfamilyhealthportrait
- Incorporating event detection
- Infectious disease phylogenies

Please see the application for specific team projects.

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

## Datasets

Datasets will come from the public repositories, primarily those housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis.

Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose.

A manuscript outlining the design and usage of the software tools constructed by each team may be submitted to an appropriate journal, such as the F1000Research Hackathons channel.

## Application

To apply, complete this [form](#) (approximately 10 minutes to complete). **Applications are due July 11th, 2016 by 4 pm ET.**

Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

The first round of accepted applicants will be notified on July 15th by 5 pm ET, and have until July 18th at 9 am to confirm their participation.

If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event.

Please include a monitored email address, in case there are follow-up questions.

## Notes

Participants will need to bring their own laptop to this program.

A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful.

Applicants must be willing to commit to all three days of the event.

No financial support for travel, lodging or meals is available for this event.

Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## **BLAST+ 2.4.0 now available**

*Monday, June 13, 2016*

Version 2.4.0 of the BLAST+ executables offers improved scoring for selenocysteine residues in the query and database sequences, as well as improved performance for the BLASTP/BLASTX/BLASTN programs.

Download the newest version on the [NCBI FTP site](#). A full list of updates is available in the [release notes](#).

## **NCBI will transition to HTTPS on September 30, 2016**

*Friday, June 10, 2016*

Starting on September 30th, when you visit NCBI pages, you'll see a green lock and **https://** in the address bar instead of **http://**. This lets you know that you are really on an NCBI page – that our server identity is confirmed – and that your communication with our server is encrypted and private.

Here's what to expect if you're a general user or a scripter:

### **For general users**

You will see the changes mentioned above – **https://** and a green lock in the address bar – but you don't have to update or change anything.

You don't need to clear your cache or update any links to NCBI pages that you've put on your own webpages or shared with people. We will redirect all our pages to **https://**.

### **For scripters**

To keep calls from failing, use **https:**, not **http:**.

Scripts that use HTTP POST to send data will not work once we transition from HTTP to HTTPS on September 30th.

If you'd like to know more about this change to HTTPS, please read [The HTTPS-Only Standard](#) from the Federal Chief Information Officers website.

**Note:** This story originally gave the transition date as September 1, 2016. It has been updated to give the correct date, September 30, 2016. We regret the error.

## **Tree Viewer 1.9 visualizes medium-large phylogenetic trees**

*Friday, June 10, 2016*

The latest version of [Tree Viewer](#) can now visualize medium-large phylogenetic trees up to 15,000 nodes. Tree Viewer 1.9 also includes mini URLs in Link to View, and several other improvements and bug fixes. The [Tree Viewer release notes](#) list all updates.

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## June 10th webinar: Finding Systematic Reviews at PubMed Health and PubMed

*Tuesday, June 07, 2016*

This Friday, NCBI will present a brief instructional webinar that will show you how to find systematic reviews using PubMed and PubMed Health.

**Date and time:** Friday, June 10, 2016 12:00-12:30 PM EDT

**Register here:** <https://nih.webex.com/nih/k2/j.php?MTID=tfa9ac8a2a377f6cece016f8dd1c9f6f5>

After registering, you will receive a confirmation email with information about attending the webinar.

## MutaBind: Evaluating the effects of sequence variants and disease mutations on protein-protein interactions

*Friday, June 03, 2016*

MutaBind is a new computational method and server created through NCBI research efforts that maps mutations on a protein structural complex, calculates changes in binding affinity, identifies deleterious mutations and produces a downloadable mutant structural model.

MutaBind guides you through this process, step by step, starting with selecting a protein complex by inputting PDB code or uploading PDB files. You can also retrieve results with a job ID number, view help documents, and review the MutaBind method and [references](#).

## Browse histones, analyze sequences with revamped HistoneDB 2.0

*Wednesday, June 01, 2016*

Created through research efforts at NCBI, [HistoneDB 2.0](#) is a totally overhauled histone database that can be used to explore the diversity of histone proteins and their sequence variants in many organisms.

This resource was established to better understand how sequence variation may affect functional and structural features of nucleosomes. HistoneDB 2.0 includes distinctive sequence alignments of many histone variants, including H2AZ, macroH2A, subH2B, spermH2B, cenH3, and H3.3, as well as canonical histones.

## June 15th webinar: Using NCBI Resources and Variant Interpretation Tools for the Clinical Community

*Wednesday, June 01, 2016*

In two weeks, NCBI will present a webinar that will show you how to use three clinical variant interpretation tools geared to clinicians through an overview of NCBI variation and medical genetics databases. A demonstration using a clinical case to demonstrate a phenotype-driven whole-genome sequence analysis using tools from Golden Helix, Omicia and SimulConsult will follow the overview.

**Date and time:** Wednesday, June 15, 2016 1:00 PM EST

**Registration URL:** <https://attendeegotowebinar.com/register/6974809559951440644>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.





## NCBI News, May 2016

### Epigenomics database to be retired June 1, 2016

*Friday, May 27, 2016*

The Epigenomics database, a public repository that was developed to archive genome-wide maps of DNA and histone modifications, will be retired on June 1, 2016.

All epigenomics data are available in our [GEO resource](#). If you are specifically interested in the NIH Roadmap Epigenomics Project, we will maintain a page for this project's data.

### NCBI launches new Twitter account for NCBI Bookshelf

*Monday, May 23, 2016*

NCBI has a new Twitter feed - [@ncbibooks](#) - to announce new books and documents available on the NCBI Bookshelf. An online resource providing free access to the full text of books and documents in life sciences and health care, the [Bookshelf](#) currently provides access to over 4,500 titles.

The Bookshelf is continuously expanding with new materials as well as receiving updates to existing books & documents. Between May 16, 2016 and May 20, for example, 19 new titles were added. Among the new titles are several Agency for Healthcare Research and Quality reports (for example, a comparative effectiveness report on imaging for pretreatment staging of small cell lung cancer), health technology assessments and systematic reviews from Canadian Agency for Drugs and Technologies in Health, and National Institute for Health Research (UK), and World Health Organization guidelines on daily iron supplementation.

Keep on top of the newest releases by following us on Twitter at [@ncbibooks](#)!

For general NCBI news, follow us on [Twitter](#), [Facebook](#) and [LinkedIn](#).

### New NCBI Insights blog post: Fast Sequence Inspection with ORFfinder and SmartBLAST (PubMed Labs)

*Monday, May 16, 2016*

The latest [blog post on NCBI Insights](#) describes the latest PubMed Labs experiment, ORFfinder. ORFfinder is a graphical analysis tool for reading open reading frames (ORFs). See [NCBI Insights](#) to read about the new features and leave feedback. We look forward to hearing your thoughts on ORFfinder.

PubMed Labs is an initiative for creating innovative and relevant products by involving you, our user community, from the beginning. It is centered upon our user community, experimentation, learning, and conversation.



## RefSeq release 76 is now available

*Monday, May 16, 2016*

RefSeq release 76 is accessible online, via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript and protein data available as of May 9, 2016 and includes 97,792,976 records, 63,971,766 proteins, 14,965,826 RNAs, and sequences from 59,995 organisms. The release is provided in several directories as a complete dataset and also as divided by logical groupings.

More information about release 76 can be found in the [release notes](#). For more information about the RefSeq project, please see the [RefSeq homepage](#).

## Genome Browsers section added to Gene

*Monday, May 16, 2016*

Gene pages now have a new link section in the sidebar called "Genome Browsers". This section provides an easy way to access all of your favorite browsers, including:

- The **NEW** [Genome Data Viewer](#), available for over 300 species
- [Map Viewer](#)
- [Variation Viewer \(human\)](#)
- [1000 Genomes browser \(human\)](#)
- [Ensembl](#)
- [UCSC](#)

Go to the [BRCA1](#) page or search for your favorite gene to try out the links available.

[Gene](#) integrates information from a wide range of species. A gene record may include nomenclature, RefSeqs, maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

The screenshot displays the NCBI Gene record for BRCA1 (breast cancer 1) in Homo sapiens. The right-hand sidebar contains a 'Table of contents' with various sections. The 'Genome Browser' section is highlighted with a red rectangular box. This section includes links to 'Genome Data Viewer', 'Map Viewer', 'Variation Viewer (GRCh37 p13)', 'Variation Viewer (GRCh38)', '1000 Genomes Browser (GRCh37 p13)', 'Ensembl', and 'iCn3D'. Below the main content area, there is a 'Genomic context' section with a table of annotation releases and a visual representation of the gene's location on Chromosome 17. The 'Genomic regions, transcripts, and products' section is also visible at the bottom.

Figure 1. The Genome Browser's section is on the right side of Gene record pages (outlined in red).

## Sequence Viewer version 3.14 upgrades platform

Friday, May 13, 2016

Sequence Viewer 3.14, now available, upgrades the platform to use ExtJS 5. Embedders may need to review their pages and make adjustments to accommodate new APIs.

For a full list of features, improvements and bug fixes, see the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

## New NCBI Variation summary page highlights all organisms in dbSNP or dbVar with full assembly annotations

Friday, May 13, 2016

The NCBI Variation [summary page](#) lists all available organisms in dbSNP and/or dbVar with full assembly annotations. On this new page, you can quickly find out the types and status of genetic variation data for each organism, as well as links to that data. The list is updated regularly with each dbSNP and dbVar new release.

## NCBI launches web-based iCn3D, a new viewer for 3D macromolecular structures

Friday, May 13, 2016

NCBI has created [iCn3D 1.0](#), a new WebGL-based viewer for interactive viewing of 3D macromolecular structures and chemicals on the web. Users no longer need to install a separate application to view structures. With iCn3D 1.0, users can:

- Interactively view 3D structures and corresponding sequence data,
- Interactively view superpositions of similar structures,
- Customize the display of a structure and generate a URL that allows you to share the link,
- And incorporate iCn3D into your own pages.

iCn3D can be accessed from the [molecular graphic](#) that appears on the structure summary for any record in the [Molecular Modeling Database \(MMDB\)](#).

The source code for iCn3D is available from [GitHub](#) for developers who would like to customize the program and/or contribute code, and for users who would like to run the program on their local computer.

For those who still would like to use Cn3D, the executable program version of the 3D structure viewer, it is still available; you can still access structures in the "Download Structure Data" portlet by selecting "Download: ASN.1(Cn3D)".

## NCBI annotates 300th organism with the Eukaryotic Genome Annotation Pipeline

*Thursday, May 12, 2016*

The [NCBI Eukaryotic Genome Annotation Pipeline](#) celebrates the annotation of its 300th organism this month! The lucky 300th is *Sinocyclocheilus anshuiensis*, a cavefish of interest for its adaptation to subterranean habitats. Vertebrates represent about two thirds of the [list of 300](#), but invertebrates and higher plants are also represented. Recently, NCBI has annotated the [rice \(\*Oryza sativa\* subspecies \*Japonica\*\)](#) and the [tobacco \(\*Nicotiana tabacum\*\)](#) genomes.

Data produced by the Eukaryotic Genome Annotation Pipeline is available in the [Reference Sequences \(RefSeq\)](#) collection, [BLAST](#) non-redundant and organism-specific databases, Gene database, and is downloadable from the [NCBI FTP site](#). See the [full list of annotated organisms](#), and request the annotation of your favorite!

## Genome Workbench 2.10.5 now available

*Wednesday, May 11, 2016*

[Genome Workbench 2.10.5](#) brings a number of new features, improvements and fixes including [ProSplign](#) integration, updates to Tree View, and new export functionalities. For the full list of changes, see the [release notes](#).

## Preview the new BLAST home page!

*Tuesday, May 10, 2016*

NCBI has released a [new home page](#) for BLAST. This new page is available through a link on the current home page. The current home page will be replaced by the new page on June 9, 2016.

The new design provides improved navigation, a cleaner look, and easier access to new BLAST services, such as running [BLAST on the cloud](#).

Please take a moment to preview the new design. If you have comments or suggestions, please write to [blast-help@ncbi.nlm.nih.gov](mailto:blast-help@ncbi.nlm.nih.gov).

**BLAST**<sup>®</sup>
Home   Recent Results   Saved Strategies   Help

## Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**


**Searching Whole Genome Shotgun sequences**

It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer.

Wed, 20 Jan 2016 10:00:00 EST

[More BLAST news...](#)


### Web BLAST



**Nucleotide BLAST**  
nucleotide → nucleotide

**blastx**  
translated nucleotide → protein

**tblastn**  
protein → translated nucleotide




**Protein BLAST**  
protein → protein

#### BLAST Genomes


Search

Human
Mouse
Rat
Microbes


### Standalone and API BLAST



**Download BLAST**  
Get BLAST databases and executables



**Use BLAST API**  
Call BLAST from your application



**Use BLAST in the cloud**  
Start an instance at a cloud provider

### Specialized searches

<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>SmartBLAST</b></div> <p>Find proteins highly similar to your query</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>Primer-BLAST</b></div> <p>Design primers specific to your PCR template</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>Global Align</b></div> <p>Compare two sequences across their entire span</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>CD-search</b></div> <p>Find conserved domains in your sequence</p>
<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>GEO</b></div> <p>Find matches to gene expression profiles</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>IgBLAST</b></div> <p>Search immunoglobulins and T cell receptor sequences</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>Vecscreen</b></div> <p>Search sequences for vector contamination</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>CDART</b></div> <p>Find sequences with similar conserved domain architecture</p>
<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>Targeted Loci</b></div> <p>Search markers for phylogenetic analysis</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>Multiple Alignment</b></div> <p>Align sequences using domain and protein constraints</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>BioAssay</b></div> <p>Search protein or nucleotide targets in PubChem BioAssay</p>	<div style="background-color: #00a651; color: white; padding: 5px; margin-bottom: 5px;"><b>MOLE-BLAST</b></div> <p>Establish taxonomy for uncultured or environmental sequences</p>

Figure 1. The new BLAST homepage.

## NCBI and RCSB PDB to assist ISCB in Sequence-Structure hackathon at ISMB Orlando 2016

*Tuesday, May 10, 2016*

From the evening of July 8th to the morning of July 11th, NCBI and the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) will assist the International Society for Computational Biology (ISCB) in hosting a sequence-structure hackathon focused on integrating protein structure information and viewers with genomic sequence and variants, pharmacogenomic binding, and general laboratory practice. To apply for this hackathon, complete [this form](#) (approximately 10 minutes to complete). Applications are due **June 1st, 2016 by 5 pm ET**.

This event is for students, postdocs and investigators or other researchers already engaged in the use of JavaScript-based viewers for protein structure visualization and/or genome/gene/protein sequence browsers. It is open to anyone selected for the hackathon, and registered for the ISMB 2016 meeting.

**Update: Due to the many potential conflicts for ISMB participants, we have moved all hackathon sessions to the evening, after programming.**

Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

- Presenting data to and from JavaScript-viewable protein structures,
- Simplified structure diagrams,
- A digital notebook for structural biologists,
- And pharmacogenomic association in conserved binding sites\*.

\* Some of the projects will build on APIs feeding data to or extracting data from viewers, so applicants need not be working on JavaScript viewers per se. For example, there is an opportunity to interface existing applications and libraries using any programming language with the [Macromolecule Transmission Format \(MMTF\)](#) for ultrafast access, parsing, and processing of PDB structures.

Please see the [application](#) for specific team projects.

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

### Datasets

Datasets will come from the public repositories housed at the NCBI, PDB and elsewhere.

### Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose. A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

### Application

To apply, complete [this form](#) (approximately 10 minutes to complete). Applications are due **June 1st, 2016 by 5 pm ET**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply. The first round of

accepted applicants will be notified on June 3rd by 5 pm ET, and have until June 6th at noon to confirm their participation.

If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. **We understand this hackathon is going on during an information-dense scientific meeting, so we ask that attendees attend at least part of three out of five scientific sessions in the draft daily schedule.** Please include a monitored email address, in case there are follow-up questions.

Note: Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Knowledge of JavaScript will be particularly helpful in this event. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals is available for this event. Also note that the event may extend into the evening hours on Friday and/or Saturday. Please make any necessary arrangements to accommodate this possibility. Awards for achievements like “Most Prolific Code Development” will be given at the ISCB awards ceremony on Tuesday.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## May 18th webinar: Using VDB BLAST Clients to Search Whole Genome Shotgun Contigs (WGS) and Transcriptome Shotgun Assembly (TSA) Data at the NCBI

*Tuesday, May 10, 2016*

On May 18th, NCBI will present a webinar that will show attendees how to use the standalone VDB blast programs (`blastn_vdb` and `tblastn_vdb`), which are part of the SRA Toolkit, as clients to search whole genome shotgun contigs (WGS) and Transcriptome Shotgun Assembly (TSA) data. WGS, which are partially assembled genome sequences, and TSA, which are transcripts assembled from next-gen RNA-Seq data, are two of the fastest growing categories of sequence data available for BLAST searching.

Through this webinar, you will see how to use the WGS/TSA browser or the EDirect utilities to find the WGS and TSA projects that interest you. You'll also learn how to use an NCBI Perl script to retrieve the database prefixes for taxonomic subsets of WGS databases.

**Time and date:** May 18th, 2016 12PM Eastern

**Registration URL:** <http://bit.ly/24Ipwu0>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## New NCBI video on YouTube: ProSplign comes to Genome Workbench

*Friday, May 06, 2016*

The newest video on the NCBI YouTube channel, *Genome Workbench: Use ProSplign for Protein to Genomic Alignments*, shows you how to use [ProSplign](#) within [Genome Workbench](#).

ProSplign is a global alignment tool that produces accurate spliced alignments and locates alignments of distantly related proteins with low similarity.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## New NCBI video on YouTube: Submitting BioSample Data to NCBI

*Tuesday, May 03, 2016*

The newest video on the NCBI YouTube channel, [Submitting BioSample Data to NCBI](#), gives you tips to make the BioSample portion of the data submission process easier.

A BioSample is a description of the biological source materials used in experimental assays.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## GenBank release 213.0 is now available via FTP

*Monday, May 02, 2016*

GenBank [release 213.0](#) (04/14/2016) has 193,739,511 "traditional" (non-WGS, non-CON) records containing 211,423,912,047 base pairs of sequence data. In addition, there are 338,922,537 WGS records containing 1,452,207,704,949 base pairs of sequence data, as well as 98,147,566 TSA records containing 87,811,163,676 base pairs of sequence data.

During the 61 days between the close dates for GenBank releases 212.0 and 213.0, the traditional portion of GenBank grew by 4,405,715,980 base pairs and by 3,489,276 sequence records. During that same period, 539,559 records were updated. An average of 66,046 traditional records were added and/or updated per day.

Between releases 212.0 and 213.0, the WGS component of GenBank grew by 52,342,209,341 base pairs and by 5,909,777 sequence records; during the same period, the TSA component of GenBank grew by 5,878,608,582 base pairs and by 6,015,248 sequence records.

The total number of sequence data files increased by 29 with this release. The divisions are as follows:

- BCT: 14 new files, now a total of 238
- CON: 4 new files, now a total of 334
- ENV: 2 new files, now a total of 91
- GSS: 1 new file, now a total of 300
- PAT: 5 new files, now a total of 251
- PRI: 1 less file, now a total of 53
- TSA: 34 new files, now a total of 228
- VRT: 1 less file, now a total of 60

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 771 GB (sequence files only); the ASN.1 data require approximately 633 GB.

More information about GenBank release 213.0, including current and upcoming changes like the change from GI sequence identifiers to accession.version, is available in the [release notes](#).



## NCBI News, April 2016

### dbSNP build 147 data for human, chicken, soybean and more are available

*Friday, April 29, 2016*

dbSNP Build 147 is accessible on the web and via [FTP](#). This release includes data for human, chicken, tilapia, mallard, sheep, date palm and soybean. Build 147 provides over 745 million submitted variants and 250 million reference variants for 7 organisms. To see complete build statistics, visit the [SNP summary page](#).

dbSNP, the NCBI Short Genetic Variations database, catalogs short variations in nucleotide sequences from a wide range of organisms.

### Eukaryotic Genome Annotation Pipeline now directly annotates top-level sequences, not scaffolds

*Wednesday, April 27, 2016*

The [Eukaryotic Genome Annotation Pipeline](#) has been modified to directly annotate RefSeq assemblies' top-level sequences (chromosomes, and unplaced and unlocalized scaffolds) instead of scaffolds. This change, included in [software release 7.0](#), improves the annotation of features spanning gaps between adjacent scaffolds, and applies to all upcoming annotation releases, including human (scheduled for May 2016).

The consequence is that for genomes assembled to the level of chromosomes, the annotation is no longer reported on placed scaffolds, and is only available on chromosomes. Specific changes include:

#### In Nucleotide:

- GenBank, Graphics and ASN views of RefSeq placed scaffolds no longer show any annotation.
- ASN view of RefSeq chromosomes now include the annotation.

#### On the FTP site:

- GFF files are now only provided for top-level sequences.
- Files in the CHR\_\* directories for nuclear chromosomes no longer include annotation on placed scaffolds.
- Masked spans ([masking\\_coordinates.gz](#)) are now in top-level coordinates.
- Comparison of current to previous annotation are now in top-level coordinates.

Here are examples from [a recent annotation of the platypus genome](#) that illustrate the change:

- [Chromosome](#)
- [Placed scaffold record](#)
- [FTP directory](#)

To see all organisms annotated by the Eukaryotic Genome Annotation Pipeline, click [here](#).

### May 4th NCBI Minute: Linking PubMed and ClinicalTrials.gov

*Tuesday, April 26, 2016*

Next Wednesday, May 4th, NCBI will present a short tutorial that will teach you two ways to filter PubMed searches for publications linked to clinical trials in [clinicaltrials.gov](#); you'll also learn how to use the ClinicalTrials database to get more information on trials of interest.

**Date and time:** May 4, 2016 12:00pm EDT

**Registration link:** <https://attendee.gotowebinar.com/register/8673331823519860737>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical trials.

## **New NCBI video on YouTube: "Sequence Viewer: Display dbVar Supporting Calls"**

*Monday, April 25, 2016*

The newest video on the NCBI YouTube channel, [Sequence Viewer: Display dbVar Supporting Calls](#), demonstrates a new feature for dbVar tracks in [Sequence Viewer](#). You can now toggle the track display to show or hide supporting variant calls, or children, for the parent structural variant.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## **Webinars on April 29 present BLAST, human variation & medical genetic records**

*Friday, April 22, 2016*

On April 29th, NCBI will host two webinars, *A Practical Guide to NCBI BLAST* and *NCBI Human Variation and Medical Genetic Resources*. Each webinar will provide an overview of the respective resources and show you how to use them.

### ***A Practical Guide to NCBI BLAST***

This webinar highlights important features and demonstrates the practical aspects of using the NCBI BLAST service, the most popular sequence similarity service in the world.

You will learn about useful but under-used features of the service, including: access from the Entrez sequence databases; the new genome BLAST service quick finder; the integration and expansion of Align-2-Sequences; organism limits and other filters; re-organized databases; formatting options and downloading options; and TreeView displays.

You will also learn how to use other important sequence analysis services associated with BLAST including Primer BLAST, an oligonucleotide primer designer and specificity checker; the multiple protein sequence alignment tool, COBALT; and SmartBLAST, a new tool for rapid protein identification. These aspects of BLAST provide easier access and results that are more comprehensive and easier to interpret.

**Date and time:** Apr 29, 2016 1:30-2:30pm

**Registration link:** <https://nih.webex.com/nih/onstage/g.php?d=629972037&t=a>

### ***NCBI Human Variation and Medical Genetic Resources***

Through this webinar, you will learn to use and access resources associated with human sequence variations and phenotypes associated with specific human genes and phenotypes. The webinar will emphasize the Gene,

MedGen and ClinVar resources to search by gene, phenotype and variant respectively. You will learn how to map variation from dbSNP and dbVar onto genes, transcripts, proteins, and genomic regions and how to find genetic tests in GTR. You will also gain experience using additional tools and viewers including PheGenI, a browser for genotype associations, the Variation Viewer and the 1000 Genomes Browser. These provide useful ways to search for, map and browse variants as well as upload and download data in genomic context.

**Date and time:** Apr 29, 2016 2:45-4:00pm

**Registration link:** <https://nih.webex.com/nih/onstage/g.php?d=626275627&t=a>

After registering for each webinar, you will receive a confirmation email with information about attending. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#) as well as a schedule of future webinars.

## NCBI to assist UC Davis in June hackathon

*Tuesday, April 19, 2016*

From June 13th to 15th, NCBI will assist the University of California Davis in hosting a biomedical data science hackathon at [the School of Veterinary Medicine](#) in Davis, CA, focusing on advanced bioinformatics analysis of next generation sequencing data and metadata. This event is for students, postdocs, investigators and other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata.\*

\*Some projects are available to other non-scientific developers, mathematicians or librarians.

Researchers and/or data scientists from the west coast of the United States are especially encouraged to apply, but the event is open to anyone selected for the hackathon, and able to travel to Davis. Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

1. Medical informatics
2. Cancer immunogenicity
3. Workflow languages
4. Sequencing contamination
5. Metagenomics
6. Metadata
7. Closing bacterial genomes

Please see the [application](#) for specific team projects.

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

### Datasets

Datasets will come from the public repositories housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository designed for that purpose](#). A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

## Application

To apply, complete [this form](#) (approximately 10 minutes to complete). Applications are due **May 5, 2016 by 5PM Eastern**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

Accepted applicants will be notified on May 9, 2016 by 2PM Eastern, and have until May 12, 2016 at 9AM Eastern to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.

**Note:** Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals is available for this event. Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions. Finally, if you are interested in having NCBI facilitate a regional hackathon hosted at your institution, please fill out [this form](#).

## New NCBI video on YouTube: Navigating the NIH Manuscript Submission Process

*Monday, April 18, 2016*

The newest video on the NCBI YouTube channel, *Navigating the NIH Manuscript Submission Process*, gives you detailed help with submitting, reviewing and approving your manuscript in the [NIH Manuscript Submission \(NIHMS\)](#) system. The NIHMS system supports manuscript depositing into PubMed Central (PMC) as required by the public access policies of NIH and other participating funding agencies.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

U.S. Department of Health & Human Services

About Help

**NIH** Manuscript Submission System

**Sign In**

1 NIH Researchers  

HHMI Researchers 

2 Publishers and Others 

Sign-In Help  
Forgot your sign-in route? [Request E-mail Reminder](#)

The NIH Manuscript Submission (NIHMS) system supports the deposit of manuscripts into PubMed Central (PMC), as required by the public access policies of NIH and other participating funders.

[Learn More](#)

### 3 Manuscript Submission Process



## Articles in Nucleic Acids Research Database 2016 Issue discuss NCBI databases, updates and future plans

Tuesday, April 12, 2016

The 23rd annual edition of the [Nucleic Acids Research Database Issue](#) features several papers from NCBI staff that describe the current state of our databases, recent updates and future plans to improve their use.

The NCBI database articles in NAR are also available from [PubMed](#). To read an article, click on the PMID listed below:

- "The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection" by Daniel J. Rigden, Xose M. Fernandez-Suarez and Michael Y. Galperin (PMID: [26740669](#))
- "Database resources of the National Center for Biotechnology Information" by NCBI Resource Coordinators (PMID: [26615191](#))
- "The International Nucleotide Sequence Database Collaboration" by Guy Cochrane, Ilene Karsch-Mizrachi, Toshihisa Takagi and INSDC (PMID: [26657633](#))
- "[GenBank](#)" by Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers (PMID: [26590407](#))

- "[Assembly](#): a resource for assembled genomes at NCBI" by Paul A. Kitts, Deanna M. Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem et al. (PMID: [26578580](#))
- "Reference sequence ([RefSeq](#)) database at NCBI: current status, taxonomic expansion, and functional annotation" by Nuala A. O'Leary, Matthew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad et al. (PMID: [26553804](#))
- "ClinVar: public archive of interpretations of clinically relevant variants" by Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth Brown, Chen Chao et al. (PMID: [26582918](#))
- "PubChem Substance and Compound databases" by Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu et al. (PMID: [26400175](#))

## Maximizing PubChem: webinar on April 20th will cover new and future features

*Wednesday, April 06, 2016*

In two weeks, NCBI staff will discuss features recently added to [PubChem](#), as well as upcoming changes to the resource.

**Date and Time:** April 20, 2016 1:00 PM ET

**Registration link:** <https://attendee.gotowebinar.com/register/2150693495841803266>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

PubChem, which provides information on the biological activities of small molecules, has been under active development. The resource is organized as three linked databases within the NCBI's [Entrez](#) information retrieval system: PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem also provides a fast chemical structure similarity search tool.

## NCBI News, March 2016

### Specialized database with unique search interface added to Zika virus resource page

Thursday, March 31, 2016

The NCBI Zika virus resource page has been updated with a [specialized database](#). This database uses pipelines to annotate genes, proteins and mature peptides, and standardize sample metadata. With this database, you can:

- Find sequences easily using standardized annotations and normalized metadata terms
- Construct alignments and phylogenetic trees using a suite of online tools
- Download sequences and metadata in a variety of formats and create customized titles/deflines for FASTA file downloads.

The [NCBI Zika virus resource](#), part of the [Virus Variation](#) family of NCBI resources, provides users with a unique, metadata-driven search interface that leverages advanced data management pipelines.

### Register for the April 6th webinar: Using NCBI Databases with Tools that Predict Genomic Variant Effects

Thursday, March 24, 2016

In two weeks, NCBI will give a demonstration of some open-source tools that use NCBI databases to predict effects of variants. We will begin with an overview of where to find and download data, particularly VCF and FASTA files, from NCBI, then show you how to use this data in 10 external tools that predict variant functional consequences, including ANNOVAR, PANTHER, SNAP-2, and Cbio MutationMapper.

**Date and time:** April 6, 2016 1:00 PM EDT

**Registration link:** <https://attendee.gotowebinar.com/register/1347860891564622851>

After registering, you will receive a confirmation email with information about attending the webinar. After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

### Register for the April 13th webinar, Submitting Data to NCBI and BioSample

Wednesday, March 23, 2016

In three weeks, NCBI staff will guide you through the process of submitting sequence data to NCBI BioSample. This webinar will show you how to describe samples and sources, and share tips on making submission to BioSample easier.

**Date and time:** April 13, 2016 1:00 PM EDT

**Registration link:** <https://attendee.gotowebinar.com/register/956885551555521537>

After registering, you will receive a confirmation email with information about attending the webinar.

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also learn about future webinars on this page.

## New NCBI video on YouTube provides strategies to search ClinVar efficiently

*Wednesday, March 23, 2016*

In the newest video on the NCBI YouTube channel, [Search ClinVar with Ease](#), we share search strategies that will help you search [ClinVar](#), our public archive of reports of relationships between human variations and phenotypes, more efficiently. Learn how to search by gene symbol, variant name and disease, and learn how to browse through variants in a genomic region with [Variation Viewer](#).

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full presentations.

## RefSeq release 75 is now available

*Tuesday, March 15, 2016*

RefSeq release 75 is accessible via [FTP](#) and through NCBI's programming utilities. This full release incorporates genomic, transcript and protein data available as of March 7, 2016 and includes 92,936,289 records, 61,034,675 proteins, 14,035,988 RNAs, and sequences from 58,776 organisms.

The release is provided in several directories as a complete dataset and also as divided by logical groupings. More information about release 75 can be found in the [release notes](#). For more information about the RefSeq project, please see the [RefSeq homepage](#).

## March 23, 2016: NCBI to offer workshop for advanced SRA and dbGaP users

*Monday, March 14, 2016*

On March 23 at 12 PM EST, NCBI staff will present a workshop for advanced users of SRA and dbGaP who are interested in using public datasets, and:

- Use and move large genomic datasets,
- Use cloud computing for analyzing genomic datasets,
- Express an interest in doing parallel work on genomic datasets,
- Or are well-versed in RNA-Seq, variant calling, or metagenomics.

The [registration link](#) lists the specific topics the workshop will cover. For a more general explanation of NCBI's genomic resources, please visit [NCBI Learn](#), where we have webinars and factsheets pertaining to dbGaP, SRA, and more.

## Search for WGS Sequences using Stand-alone BLAST!

*Monday, March 07, 2016*

It is now much easier to search WGS (Whole Genome Shotgun) with stand-alone BLAST on your own computer. New tools from the NCBI allow you to BLAST just the WGS projects you are interested in. You can also search a taxonomic subset of WGS (e.g., all human or all bacterial sequences). These new tools for WGS make the existing search mechanism obsolete.

As of August 5, 2016, the current single WGS BLAST database will be retired from the NCBI FTP site and BLAST server. We suggest moving to the new tools as soon as possible.

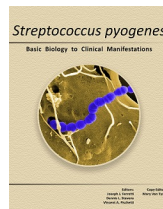


Read more at [ftp://ftp.ncbi.nlm.nih.gov/blast/WGS\\_TOOLS/README\\_BLASTWGS.txt](ftp://ftp.ncbi.nlm.nih.gov/blast/WGS_TOOLS/README_BLASTWGS.txt).

## First of the New Bookshelf NCBI Insights Blog Posts - New *Streptococcus pyogenes* book

*Wednesday, March 02, 2016*

The first of a new series of NCBI Insights blog posts highlighting books and documents is available on NCBI's Bookshelf showcasing a new book: "Streptococcus pyogenes: Basic Biology to Clinical Manifestations".



Published by the University of Oklahoma Health Sciences Center, this new open-access book provides a comprehensive review of research on the bacterium *Streptococcus pyogenes* (Group A *Streptococcus*) which is responsible for diseases such as scarlet fever, pharyngitis, impetigo, cellulitis, necrotizing fasciitis and toxic shock syndrome, as well as the sequelae of rheumatic fever and acute poststreptococcal glomerulonephritis.

"Streptococcus pyogenes: Basic Biology to Clinical Manifestations" is freely available on NCBI's Bookshelf, at <http://www.ncbi.nlm.nih.gov/books/NBK333424/>.

## NCBI is phasing out sequence GIs - use Accession.Version instead!

*Wednesday, March 02, 2016*

As of September 2016, the integer sequence identifiers known as "GIs" will no longer be included in the GenBank, GenPept, and FASTA formats supported by NCBI for sequence records. The FASTA header will be further simplified to report only the sequence accession.version and record title for accessions managed by the International Sequence Database Collaboration (INSDC) and NCBI's Reference Sequence (RefSeq) project. As NCBI makes this transition, we encourage any users who have workflows that depend on GI's to begin planning to use accession.version identifiers instead. After September 2016, any processes solely dependent on GIs will no longer function as expected.

GI numbers have been in use since GenBank release 81.0 (February 1994) as an additional identifier to the accession number to stably refer to a specific version of a sequence record. Version tracking was added to accession numbers in 1997 as an integer suffix that increments with each update to the sequence data within a record. For example, "AC020606.7" indicates that the sequence content of the record has been updated six times since the first release. Thus, sequence versioning information has been provided in a redundant fashion through both the GI and the accession.version. In the past decade, NCBI has continued to receive submissions of new or updated sequences at a rapidly increasing rate. In response to this, we have had to develop new data storage solutions that use accession.version information, rather than GI information, to track updates. Current examples of sequences that lack a GI include unannotated contigs in WGS and TSA projects. This results in a situation where we are conveying version information inconsistently.

Given both the continued increase in the volume of data submissions and the growing inconsistency in record presentation, it is time for us to take the next step and remove the older, redundant GI identifiers and retain a single identifier for sequence versions, the more human-readable accession.version. This change will simplify the

process of tracking sequences without any loss of functionality. This change will also simplify scientific communications by promoting use of accession.version as the preferred sequence identifier. Therefore, over the coming months we will no longer assign GI's to an increasing number of new sequences. Sequence records with existing GI's will retain them in some presentation formats, such as ASN.1 and the 5-column feature table format, but the GI value will no longer be displayed in other presentation formats including GenBank flat file and FASTA formats. NCBI services that accept GI's as input will continue to be supported, and NCBI will be adding support for accession.version identifiers to all services that currently do not support them.

This transition to stop assigning and reporting GIs was first described in the Release Notes for GenBank 199.0 in December 2013 and also described in a [recent GenBank update](#). Please see Section 1.4.1 of the current GenBank release notes for background information: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

The FASTA display for all sequence records exchanged by the INSDC and for all NCBI RefSeq records will also be changed to report only the accession.version and the record title. This will improve compatibility with other file types provided by NCBI, including GFF3, Gene, and dbSNP download files. This FASTA format change has already been made on data available from the redesigned genomes FTP site based on user requests to have a single consistent sequence identifier for both GFF3 and FASTA formats. See the prior announcement of this change: <http://www.ncbi.nlm.nih.gov/news/08-26-2014-new-genomes-FTP-live/> .. At this time, we plan to continue to provide database source information in the FASTA display of sequences from non-INSDC and non-RefSeq sources including: SwissProt, PDB structures, PIR, and patent sequences.

After September 2016, these changes will start to appear on NCBI web views of flat file and FASTA format sequence data, NCBI programming utilities results, and GenBank and RefSeq comprehensive FTP releases.

**Example 1: An INSDC nucleotide record** - In the sample record below, nucleotide sequence AF123456 was assigned a GI of 6633795, and the protein translated from its coding region feature was assigned a GI of 6633796:

```

LOCUS      AF123456                1510 bp   mRNA    linear   VRT 12-APR-2012
DEFINITION Gallus gallus doublesex and mab-3 related transcription factor 1
            (DMRT1) mRNA, partial cds.
ACCESSION  AF123456
VERSION   AF123456.2  GI:6633795
....

CDS                <1..936
                  /gene="DMRT1"
                  /note="cDMRT1"
                  /codon_start=1
                  /product="doublesex and mab-3 related transcription factor
                  1"
                  /protein_id="AA19666.1"
                  /db_xref="GI:6633796"
                  /translation="PAAGKKLPRLPKCARCRNHGYSSPLKGHKRFMWRDCQCKKCSL
                  IAERQRVMAVQVALRRQQAQEEELGISHPVPLPSAPEPVVKKSSSSSSCLLDSSSPA
                  HSTSTVAAAAASAPPEGRMLIQDIPSIIPSRGHLESTSDLVVDSTYYSSFYQPSLYPYY
                  NNLYNYSQYQMAVATESSSSETGGTFVGSAMKNSLRSLPATYMSSQSGKQWQMKGMEN
                  RHAMSSQYRMCSYYPPTSYLQGVGSPTCVTQILASEDTPSYSESKARVFSPPSSQDS
                  GLGCLSSSESTKGDLECEPHQEPGAFVAVSPVLEGE"

```

After September 2016, the accession.version will be the sole indicator of the sequence version. The GI value on the VERSION line and the GI/db\_xref qualifier for the coding region feature will no longer be visible.

```

LOCUS      AF123456                1510 bp    mRNA    linear    VRT 12-APR-2012
DEFINITION Gallus gallus doublesex and mab-3 related transcription factor 1
            (DMRT1) mRNA, partial cds.
ACCESSION  AF123456
VERSION    AAF123456.2

```

....

```

CDS                <1..936
                  /gene="DMRT1"
                  /note="cDMRT1"
                  /codon_start=1
                  /product="doublesex and mab-3 related transcription factor
                  1"
                  /protein_id="AAF19666.1"
                  /translation="PAAGKKLPRLPKCARCRNHGYSSPLKGHKRFMWRDCQCKKCSL
                  IAERQRVMAVQVALRRQQAQEEELGISHPVPLPSAPEFVVKKSSSSSSCLLQDSSSPA
                  HSTSTVAAAAASAPPEGRMLIQDIPSIIPSRGHLESTSDLVVDSTYYSSFYQPSLYPYY
                  NNLYNYSQYQMAVATESSSSETGGTFVGSAMKNSLRSLPATYMSQSGKQWQMKGMEN
                  RHAMSSQYRMC SYYPPTS YLGQGVGSPTCVTQILASEDTPSYSESKARVFSPPSSQDS
                  GLGCLSSSESTKGDLECEPHQEPGAFVAVSPVLEGE"

```

**Example 2: A GenPept protein record** - The current record display includes the GI in the VERSION lines. Note that the coding region feature for GenPept format has never included the display of GI values.

```

LOCUS      AAF19666                311 aa    linear    VRT 12-APR-2012
DEFINITION doublesex and mab-3 related transcription factor 1, partial [Gallus gallus].
ACCESSION  AAF19666
VERSION    AAF19666.1  GI:6633796
DBSOURCE   accession AF123456.2

```

....

```

CDS                1..311
                  /gene="DMRT1"
                  /coded_by="AF123456.2:<1..936"

```

After September 2016, the VERSION line will not include the GI value:

```

LOCUS      AAF19666                311 aa    linear    VRT 12-APR-2012
DEFINITION doublesex and mab-3 related transcription factor 1, partial [Gallus gallus].
ACCESSION  AAF19666
VERSION    AAF19666.1
DBSOURCE   accession AF123456.2

```

....

```

CDS                1..311
                  /gene="DMRT1"
                  /coded_by="AF123456.2:<1..936"

```

**Example 3: Changes to FASTA format: GI and database source values will be removed from FASTA header -**

The current FASTA display, in most resources, currently includes GI and database source information (e.g., 'gb' for GenBank) delimited with a '|'. Downstream analysis tools often require first processing the FASTA header line to simplify the sequence identifier portion to the accession.version or GI. The complex FASTA sequence identifier is highlighted in yellow:

```

>gi|6633795|gb|AF123456.2| Gallus gallus doublesex and mab-3 related transcription factor
1 (DMRT1) mRNA, partial cds
CCGGCGGCGGGCAAGAAGCTGCCGCGTCTGCCCAAGTGTGCCCGCTGCCGCAACCACGGCTACTCCTCGC
CGCTGAAGGGGCAACAAGCGGTTCTGCATGTGGCGGGACTGCCAGTGCAAGAAGTGCCAGCCTGATCGCCGA

>gi|6633796|gb|AAF19666.1| doublesex and mab-3 related transcription factor 1, partial
[Gallus gallus]
PAAGKKLPRLPKCARCRNHGYSSPLKGHKRFMWRDCQCKKCSLIAERQRVMAVQVALRRQQAQEEELGI
SHPVPLPSAPEFVVKKSSSSSSCLLQDSSSPAHSSTVAAAAASAPPEGRMLIQDIPSIIPSRGHLESTSD

```

After September 2016, a simple sequence ID will be provided in the FASTA header for nucleotide and protein records

```
>AF123456.2 Gallus gallus| doublesex and mab-3 related transcription factor 1 (DMRT1)
mRNA, partial cds
CCGGCGGGCGGGCAAGAAGCTGCCGCGTCTGCCCAAGTGTGCCCGCTGCCGCAACCACGGCTACTCCTCGC
CGCTGAAGGGGCACAAGCGGTTCTGCATGTGGCGGGACTGCCAGTGCAAGAAGTGCAGCCTGATCGCCGA

>AAF19666.1 doublesex and mab-3 related transcription factor 1, partial [Gallus gallus]
PAAGKKLPRLPKCARCRNHGYSSPLKGHKRFCMWRDCQCKKCSLIAERQVMAVQVALRRQQAQEEELGI
SHFVPLPSAFEPVVKSSSSSSCLLQDSSSPAHSTSTVAAAAASAPPEGRMLIQDIPSI PSRGHLESTSD
...
```

## Tree Viewer's Next Update is Available

*Wednesday, March 02, 2016*

An updated version (v.1.8.0) of the [NCBI Tree Viewer](#), a tool for viewing your own phylogenetic tree data, has been released which has several new features and improvements, as well as some bug fixes.

These include:

- “Link to View” function to create minimized links to Tree Viewer
- Feedback function to inform NCBI developers about issues and improvements
- Mechanism to customize labels
- Better Zoom navigation with adaptive levels and a new button to quickly zoom to the minimal level where node labels become visible
- API zooming functions for embedded views
- New API for creation of custom labels
- Aspect ratio selection to improve the display of radial trees

In addition, several bugs have been fixed.

To see the full list of changes, see the [Tree Viewer release notes](#).

## NCBI News, February 2016

### NCBI to assist Brandeis University in hosting Boston genomics hackathon in April

Monday, February 29, 2016

From April 25 to 27, NCBI will assist [Brandeis University](#) in hosting a genomics hackathon focusing on advanced bioinformatics analysis of next-generation sequencing data and metadata. This event is for students, postdocs, investigators and other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata.\* Researchers and/or data scientists from the Boston area are especially encouraged to apply, but the event is open to anyone selected for the hackathon and able to travel to Brandeis.

Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

1. Network Analysis of Variants
2. Structural Variation\*
3. RNA-Seq
4. Streaming Data and Metadata\*
5. Neuroscience/Immunity
6. Command-line user-interface design\*

Please see the application for specific team projects.

*\*Some projects are available to other non-scientific developers, mathematicians or librarians.*

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems.

### Datasets

Datasets will come from the public repositories housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

### Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose. A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

### Application

To apply, complete this [form](#) (approximately 10 minutes to complete). Applications are due **March 22nd by 5 PM ET**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior participants and applicants are especially encouraged to reapply.

Accepted applicants will be notified on March 24th by 2 PM ET; applicants have until **March 27th at 9 AM ET** to confirm their participation.

If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.

**Note:** Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. *No financial support for travel, lodging or meals is available for this event.* Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

If you are interested in having NCBI facilitate a regional hackathon hosted at your institution, please fill out this [form](#).

## GenBank release 212.0 available via FTP

*Friday, February 26, 2016*

GenBank release 212.0 (2/13/2016) is now available online, on the [FTP site](#) and through NCBI's [programming utilities](#). Release 212.0 has 190,250,235 non-WGS, non-CON records containing 207,018,196,067 base pairs of sequence data. In addition, there are 333,012,760 WGS records containing 1,399,865,495,608 base pairs of sequence data, as well as 92,132,318 TSA records containing 81,932,555,094 base pairs of sequence data.

During the 61 days between the close dates for GenBank releases 211.0 and 212.0, the traditional (i.e., non-WGS, non-CON) portion of GenBank grew by 3,079,084,996 base pairs and by 1,017,310 sequence records. During that same period, 395,404 records were updated. An average of 23,159 'traditional' records were added and/or updated per day.

Between releases 211.0 and 212.0, the WGS component of GenBank grew by 101,999,877,243 base pairs and by 15,890,603 sequence records; the TSA component of GenBank grew by 4,349,215,918 base pairs and by 4,643,779 sequence records. The total number of sequence data files increased by 29 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 224
- CON: 3 new files, now a total of 330
- ENV: 1 new files, now a total of 89
- EST: 2 new files, now a total of 480
- INV: 1 new files, now a total of 136
- PAT: 4 new files, now a total of 246
- PLN: 3 new file, now a total of 125
- PRI: 4 new file, now a total of 54
- ROD: 1 less file, now a total of 31
- VRL: 1 new file, now a total of 40
- VRT: 1 new file, now a total of 61

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 756 GB (sequence files only); the ASN.1 data require approximately 619 GB. More information about GenBank release 212.0 is available in the [release notes](#) and in the README files in the [genbank](#) and [ASN.1](#) directories.

## March 2nd webinar: NCBI Resources for Cancer Researchers

Wednesday, February 24, 2016

Next Wednesday, NCBI staff will discuss the facets of NCBI resources relevant to cancer in a live webinar. The databases and tools included in this overview are: BLAST, GenBank, DNA-Seq, RNA-Seq, Epigenomics and metagenomics datasets, as well as tools and APIs at NCBI that can be used to extract relevant subsets of data for cancer research.

**Date and time:** March 2, 2016 1:00-2:00 PM EST

**Registration link:** <https://attendee.gotowebinar.com/register/3717666889216708353>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on this page.

## Zika virus resource page provides access to nucleotide, protein sequences from latest outbreak

Tuesday, February 23, 2016

The new [Zika virus resource page](#) makes it easy to find and analyze relevant sequence data. The page includes links to the following Zika virus data at NCBI: nucleotide and protein sequences, the reference genome with updated mature peptide annotation, and publications.



NCBI Resources How To Sign in to NCBI

Virus Variation Search NCBI Search

**Zika Virus Resource**  
Retrieve, view, and download Zika virus nucleotide and protein sequences from a value added database using a specialized search interface.

Zika virus sequences	Other NCBI Zika virus resources	External Zika virus resources
<a href="#">Zika virus nucleotide sequences</a>	<a href="#">Zika virus reference genome</a>	<a href="#">Zika virus health information resources</a>
<a href="#">Zika virus protein sequences</a>	<a href="#">Publications</a>	<a href="#">HealthMap</a>
<a href="#">How to cite us</a>	<a href="#">Genome browser</a>	<a href="#">CDC</a>
<a href="#">Contact us</a>	<a href="#">Taxonomy</a>	<a href="#">WHO</a>
		<a href="#">ViriZone</a>

In addition, a Zika database will be added to the [NCBI Virus Variation resource](#). This database will use specialized pipelines to annotate genes, proteins and mature peptides, and standardize sample metadata. With this specialized database, you'll be able to:

- Quickly find the sequences you need, through an intuitive search interface for all viral sequences using standardized protein/gene names and metadata,
- Select the latest sequences based on date criteria or sorting of results,
- Download sequences in many formats or find links to sequences in NCBI databases, and
- Analyze sequences using multiple sequence alignments and phylogenetic trees.

Stay tuned to NCBI News or our social media channels - particularly [Facebook](#), [Twitter](#) and [LinkedIn](#) - for updates on the specialized Zika virus database.

## dbSNP Build 146 for salmon, barrel medic, cottonwood and mouse now available

Tuesday, February 23, 2016

dbSNP Build 146 data for salmon, barrel medic, cottonwood and mouse are available now on the [web](#) and [FTP](#).

New salmon (*Salmo salar*) information:

- [FTP](#)
- [Entrez](#)
- Assembly: ICSASG\_v2 (GCF\_000233375.1)
- New SS: 1,342,320
- New RS: 1,029,869

New barrel medic (*Medicago truncatula*) information:

- [FTP](#)
- [Entrez](#)
- Assembly: MedtrA17\_3.5 (GCF\_000219495.1)
- New SS: 0
- New RS: 0

New cottonwood (*Populus trichocarpa*) information:

- [FTP](#)
- [Entrez](#)
- Assembly: Poptr1\_1 (GCF\_000002775.1)
- New SS: 17,902,170
- New RS: 9,505,665

New mouse (*Mus musculus*) information:

- [FTP](#)
- [Entrez](#)
- Assembly: GRCm38.p3 (GCF\_000001635.23)
- New SS: 107,682
- New RS: 14,352

dbSNP, NCBI's Short Genetic Variations database, catalogs short variations in nucleotide sequences from a wide range of organisms.

## Rotavirus resource uses standardized metadata and annotations, suite of tools to make it easier to search, download and analyze sequences

Friday, February 19, 2016

The new [NCBI Rotavirus resource](#), part of the [Virus Variation](#) family of NCBI resources, provides users with a unique, metadata-driven search interface that leverages advanced data management pipelines.

Sequence annotations and descriptive metadata are mapped to a standardized vocabulary within this resource, making it much easier to find and analyze sequences of interest. Searches can also be restricted to sequences



from complete genome sets or to sequence sets containing specific combinations of segments and segment genotypes.

Finally, a suite of tools allows users to build alignments and phylogenetic trees from selected sequences, and users can also download sequences with customized titles/defines based on standardized metadata.

## **New video on the NCBI YouTube channel: Eukaryotic Genome Data Curation at NCBI**

*Friday, February 19, 2016*

A [recording](#) of the January 5th webinar is now on the NCBI YouTube channel. In addition, the webinar question and answer session has been summarized in a document available on [FTP](#).

In this webinar, three RefSeq biocurators discuss aspects of eukaryotic organism data curation. Topics covered include sequence analysis, functional annotation, data validation and community collaboration. To make it easier to locate specific sections or topics within the video, we added a table of contents at the 9 second mark.

The video is also included in the [NCBI Webinars playlist](#), which you can save to your own Playlists collection on YouTube for quick access - just click on the plus sign marked Save.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about our videos, which range from quick tips to full presentations.

For information about upcoming webinars, stay tuned to [NCBI News](#) and the [Courses and Webinars page](#).

## **NCBI Insights blog post: Professors: "NCBI can help you streamline your teaching and research efforts"**

*Wednesday, February 17, 2016*

The [latest blog post on NCBI Insights](#) points out the many tasks an NCBI account can help with, from storing and automating searches to creating bibliographic collections and more.

While this post is written to highlight how professors can gain from an NCBI account, these tips apply to everyone who [signs up for an NCBI account](#). We encourage users to bookmark this blog post and refer to it whenever needed.

[NCBI Insights](#) is the official NCBI blog, where we share quick tips and what's new at NCBI.

## **New video on the NCBI YouTube channel: Viral resources at NCBI**

*Thursday, February 11, 2016*

In the newest video on the NCBI YouTube channel, [Viral resources at NCBI](#), Dr. Rodney Brister, head of the viral resources group at NCBI, gives a detailed tour of the many tools and databases publicly available to anyone studying viruses.

Subscribe to [the NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full presentations.

## NCBI to assist Louisiana State University in South and Southeast regional genomics hackathon

Monday, February 08, 2016

From March 21st to 23rd, NCBI will assist Louisiana State University (LSU) in hosting a regional genomics hackathon in Shreveport, LA. This event is for students, postdocs, investigators and other researchers already engaged in the use of pipelines for genomic analyses from next-generation sequencing data or metadata.\* Researchers and/or data scientists from the South and Southeast United States are especially encouraged to apply, but the event is open to anyone selected for the hackathon and able to travel to Shreveport.

\* Some projects are available to other non-scientific developers, mathematicians or librarians.

Working groups of 5-6 individuals will be formed into five or six teams. These teams will build pipelines and tools to analyze large datasets within a cloud infrastructure. The potential subjects for this iteration are:

1. Network Analysis of Variants
2. Structural Variation
3. RNA-Seq
4. Streaming Data and Metadata
5. Neuroscience/Immunity
6. Command-line user-interface design

Please see the application for specific team projects.

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems. This course will take place at the [Louisiana State Health Sciences Center - Shreveport](#) in Shreveport, Louisiana and is hosted by the two LSU institutions in town: LSU Health Sciences Center and LSU Shreveport.

### Datasets

Datasets will come from the public repositories housed at the NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

### Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository designed for that purpose](#). A manuscript outlining the design and usage of the software tools constructed by each team will be submitted to an appropriate journal.

### Application

To apply, complete [this form](#), which takes approximately 10 minutes to complete. Applications are due **February 2/19/16 by 5 pm ET**. Participants will be selected from a pool of applicants based on the experience and motivation they provide on the form. Prior students and applicants are encouraged to reapply. Accepted applicants will be notified on February 22 by 2 pm ET, and have until February 26 at 9 am to confirm their participation. If you confirm, please make sure it is highly likely you can attend, as confirming and not attending bars other data scientists from attending this event. Please include a monitored email address, in case there are follow-up questions.

Note: Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals is available for this event. Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions. Finally, if you are interested in having NCBI facilitate a regional hackathon hosted at your institution, please fill out [this form](#).

## Variation Viewer 1.5 adds facet toggling, updated backend data

*Thursday, February 04, 2016*

Variation Viewer 1.5 provides several new features, improvements and bug fixes, including facet toggling in Variant Table, updated backend data to dbSNP human build 146, dbVar (December 2015) and ClinVar (January 2016), and more. A full list of changes to Variation Viewer is available in the [release notes](#).

[Variation Viewer](#) is a tool for navigating variant data in [dbSNP](#), [dbVar](#) and [ClinVar](#) in a genomic context.

## February 17th webinar: "Five ways to submit next-gen sequencing data to NCBI's Sequence Read Archive (SRA)"

*Wednesday, February 03, 2016*

In two weeks, NCBI will present a webinar on SRA submissions, discussing five different ways to submit next-generation sequencing data to SRA, including the [new SRA submission portal \(beta\)](#) which allows you to submit data via FTP and the Aspera command line, Illumina's BaseSpace, MOTHUR, and the iPlant Collaborative.

**Date and time:** February 17, 2016 1:00-2:00 PM EST

**Registration link:** <https://attendee.gotowebinar.com/register/6510651823186558978>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on [this page](#).



## NCBI News, January 2016

### Genome Workbench 2.10 now available

*Friday, January 29, 2016*

Genome Workbench 2.10 includes a reworked BLAST tool and new functionalities in Tree View. For the full list of features, improvements and fixes, see the [release notes](#).

### Sequence Viewer 3.11 now available

*Wednesday, January 27, 2016*

Sequence Viewer 3.11, now available, contains a number of new features, improvements and bug fixes, including the ability to overlay multiple graphs in one track (for more information, please see the [demo pages](#) and [API documentation](#)), improved track descriptions for better integration with track management, updated SNP tracks, and tooltips for dbVar. A full list of features, improvements and bug fixes is included in the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

### February 3rd webinar: "How to Upload and Analyze dbGaP Data in the Cloud"

*Thursday, January 21, 2016*

In two weeks, NCBI will show you how to upload and analyze dbGaP data through the [SRA Toolkit](#) and Amazon Web Services.

**Date and time:** Feb 3, 2016 1:00-2:00 PM EST

**Registration link:** <https://attendee.gotowebinar.com/register/8258777454794963713>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on this page.

### RefSeq Release 74 now available on FTP

*Wednesday, January 20, 2016*

RefSeq Release 74 is now accessible online, on the [FTP site](#), and through NCBI's programming utilities. This full release incorporates genomic, transcript and protein data available as of January 11, 2016 and includes 89,458,499 records, 56,496,614 proteins, 13,719,136 RNAs, and sequences from 57,993 organisms.

More information can be found in the [release notes](#). For more information about the RefSeq project, please see the [RefSeq homepage](#).

### January 28th webinar: "Genomic Data Sharing with dbGaP: Registration and Submission" for IRP investigators

*Wednesday, January 13, 2016*

In two weeks, NCBI will present a webinar for [Intramural Research Program \(IRP\)](#) investigators engaged in [Genome-Wide Association Studies \(GWAS\)](#) and other genomic research efforts at NIH. Topics covered will include working with your Genomics Program Administrator to register your study in dbGaP and preparing your project metadata files, phenotypes and molecular data for submission to dbGaP.

**Date and time:** Jan 28, 2016 12:30-1:30 PM EST

**Registration link:** <https://attendee.gotowebinar.com/register/6247720109478660865>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also find information about future webinars on this page.

## NCBI News, December 2015

### NCBI staff will attend the International Plant and Animal Genome Conference XXIV in January

*Wednesday, December 30, 2015*

From January 9-13, 2016, NCBI staff will present posters, give lectures and presentations, and lead workshops at the [International Plant and Animal Genome Conference](#). In addition, NCBI will have a booth (#618) where we will answer any questions you may have; we also welcome your suggestions and comments.

#### Selected talks and posters:

- **Plenary Lecture:** *The Genome Era at NCBI – Are We There Yet?* by Dr. James M. Ostell, Chief of the Information Engineering Branch (IEB) of NCBI (Tuesday, January 12, 2016 8:45 AM)
- **NCBI Genome Resources Workshop** (Monday, January 11, 2016 12:50 PM – 3:00 PM)
- **P0294** *dbSNP and dbVar: NCBI Databases of Simple and Structural Variations for All Organisms* (Monday, January 11, 2016 10:00 AM – 11:30 AM)
- **P0295** *Curation of Genome Assemblies by NCBI Tools and Resources* (Monday, January 11, 2016 3:00 PM – 4:30 PM)

To see the complete schedule of NCBI's activities at PAG XXIV, see the [Conferences and Presentations page](#) or visit **Booth #618** while at PAG.

### BLAST+ executables 2.3.0 now available

*Tuesday, December 29, 2015*

A new version (2.3.0) of the stand-alone BLAST executables (Linux, Windows and MacOSX available on [FTP](#)) is now available. The BLAST VM at cloud providers will be updated soon as well.

This new version includes a beta release of SAM format, as well as support for single-file mode for BLAST XML2 and JSON formats; single-file mode means that results are delivered as a single-file regardless of the number of queries. A number of other improvements and bug fixes, including a new versioning policy, are included. Please refer to the [release notes](#) for the full list of improvements, bug fixes, and more.

### GenBank release 211.0 is now available via FTP

*Thursday, December 24, 2015*

GenBank release 211.0 (12/18/2015) has 189,232,925 non-WGS, non-CON records containing 203,939,111,071 base pairs of sequence data. In addition, there are 317,122,157 WGS records containing 1,297,865,618,365 base pairs of sequence data, as well as 87,488,539 TSA records containing 77,583,339,176 base pairs of sequence data.

During the 60 days between the close dates for GenBank releases 210.0 and 211.0, the traditional (i.e., non-WGS, non-CON) portion of GenBank grew by 1,702,029,512 base pairs and by 860,908 sequence records. During that same period, 1,626,191 records were updated and an average of 41,451 traditional records were added and/or updated each day.

Between releases 210.0 and 211.0, the WGS component of GenBank grew by 75,230,350,867 base pairs and by 7,923,214 sequence records; the TSA component of GenBank grew by 6,666,166,232 base pairs and by 5,698,508 sequence records.

The total number of sequence data files increased by 29 with this release. The divisions are as follows:

- BCT: 8 new files, now a total of 216
- CON: 3 new files, now a total of 327
- ENV: 2 new files, now a total of 88
- INV: 3 new files, now a total of 135
- PAT: 7 new files, now a total of 242
- PLN: 3 new files, now a total of 122
- PRI: 1 new file, now a total of 50
- ROD: 1 new file, now a total of 32
- TSA: 1 less file, now a total of 194
- VRL: 1 new file, now a total of 39
- VRT: 1 new file, now a total of 60

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 749 GB (sequence files only); the ASN.1 data require approximately 613 GB. More information about GenBank release 211.0 is available in the [release notes](#).

## January 7th: Explore new graphical viewer track options with The NCBI Minute

*Tuesday, December 22, 2015*

On January 7th, NCBI will present a new NCBI Minute webinar, “New track options for getting the most out of NCBI Graphical Viewers”. In this webinar, you’ll learn how to use these new features to get the most out of Sequence Viewer, Variation Viewer and other NCBI graphical browsers. In addition, we’ll show you how to search and quickly find relevant tracks and upload your own custom data.

**Date and time:** Thursday, January 7, 2016 12:00 – 12:15 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/9028024375893233666>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on this page.

The [NCBI Minute](#) is a series of short webinars that give a brief introduction to a specific topic or NCBI tool.

## January 5th webinar: Eukaryotic Genome Data Curation at NCBI

*Monday, December 21, 2015*

On January 5th, three RefSeq (Reference Sequence) curators will focus on aspects of data curation for eukaryotic organisms. We’ll discuss several aspects of manual curation, including sequence analysis, functional annotation, nomenclature review, data validation and community collaboration. We will also highlight how these curation efforts improve the programmatic approaches used by the genome annotation pipelines, which allow NCBI to handle the ever-increasing amount of data generated by researchers.

**Date and time:** January 5, 2016 1:00 – 2:00PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/5281805741351150082>



After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on this page.

## **New on NCBI Insights blog: "The NCBI Minute: quick introductions to NCBI resources"**

*Thursday, December 17, 2015*

Today's post on [NCBI Insights](#) discusses the [NCBI Minute](#), a series of short webinars that introduce a new NCBI tool or resource or provide quick tips for using a popular NCBI resource.

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## **dbSNP Build 146 for non-human organisms is now available**

*Wednesday, December 16, 2015*

dbSNP Build 146 is accessible via Entrez and [FTP](#). This release includes data for corn, cow, dog, rat, chickpea and rapeseed. Build 146 provides over 380 million submitted variants and 166 million reference variants for 6 species. To see complete build statistics, visit the [SNP summary page](#).

dbSNP, the NCBI Short Genetic Variations database, catalogs short variations in nucleotide sequences from a wide range of organisms.

## **New NCBI Insights blog posts highlight SRA Toolkit, Run Selector**

*Friday, December 11, 2015*

Today, two new blog posts on [NCBI Insights](#) present SRA Toolkit and Run Selector, which allow you to integrate downloaded data sets into pipelines and fine-tune web-based search results, respectively.

### **"SRA Toolkit: the SRA database at your fingertips"**

"SRA Toolkit: the SRA database at your fingertips" briefly explains where to download the SRA Toolkit (on [its own webpage](#), or on [GitHub](#)) and describes the various capabilities of the toolkit's command-line executables, which include:

- Streaming data from the NCBI/SRA servers
- Working with restricted-access data from [dbGaP](#) (*after* applying for and receiving access to this data)

Click over to this post on [NCBI Insights](#) to learn more ways to use the SRA Toolkit.

### **"Fine-tune your web-based search results with SRA Run Selector"**

This [blog post](#) introduces Run Selector, a feature within web-based SRA search that lets you use fields to quickly filter search results to include only data relevant to you. Run Selector also makes it easy to download data or accession lists; this and more is explained in the [blog post](#).

The [Sequence Read Archive \(SRA\)](#) is NCBI's largest growing repository of molecular data. It archives raw sequencing data and alignment information from high-throughput sequencing platforms.

NCBI Insights is the official NCBI blog, where we share science feature stories, quick tips and what's new at NCBI.

## December 17th webinar: "Accessing 1000 Genomes Project Data"

*Thursday, December 03, 2015*

On Thursday December 17th, 2015, NCBI staff will demonstrate how to access 1000 Genomes data through SRA, dbVar, SNP and BioProject, as well as through tracks on annotated human sequences in the graphical sequence viewer and Variation Viewer. Attendees will also learn how to display, search and download individual and genotype level data through the dedicated [1000 Genomes Browser](#) that allows searching by chromosomal position, gene names and other genome markers.

**Date and time:** Dec 17, 2015 1:00 – 2:00 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/5168155820927556866>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also find information about future webinars on this page.

## Registration open for December 16th NCBI Minute: "New Faceted Advanced Search in dbGaP Provides Easy Access to Relevant Data"

*Thursday, December 03, 2015*

On December 16th, the NCBI Minute short webinar will introduce [dbGaP's](#) new faceted advanced search interface and show attendees how to use the new interface to easily find data by study, variables, documents and genotypes.

**Date and time:** Dec 16, 2015, 12:00-12:15 PM EST

**Registration URL:** <https://attendee.gotowebinar.com/register/7869339230869750529>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also find information about future webinars on this page.

The [NCBI Minute](#) is a series of short webinars that give a brief introduction to a specific topic or NCBI tool.

## NCBI News, November 2015

### dbSNP human build 146 available through Entrez and FTP

*Wednesday, November 25, 2015*

dbSNP human Build 146, based on the GRCh38.p2 and GRCh37.p13 assemblies, is now available on the [integrated NCBI Entrez system](#) and through [FTP](#). Build 146 provides 150 million Reference SNP (RS) clusters, including 985,775 new RS clusters and allele frequency data from 1000 Genomes, GO-ESP, and ExAC projects. To see complete build statistics, visit the [dbSNP summary page](#).

### Tree Viewer 1.7.5 now available

*Tuesday, November 24, 2015*

NCBI Tree Viewer version 1.7.5 has several new features, improvements and bug fixes, including improved subtree de-selection function, fixed BLAST Tree View bugs, and a fixed mouse wheel zoom bug. To see the full list of updates, see the [Tree Viewer release notes](#).

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

### NCBI releases first five lectures of NCBI NOW on YouTube

*Tuesday, November 24, 2015*

Today, the first five lectures from the NCBI NOW workshop are available in a [playlist](#) on the NCBI YouTube channel. Last month, NCBI presented this online workshop (more information [here](#)) to 650 participants new to next generation sequencing (NGS) analysis.

Subscribe to the [NCBI YouTube channel](#) to receive notifications about our new videos, which range from quick tips to full webinar presentations.

### December 2nd NCBI Minute webinar: Finding Genes in PubMed

*Monday, November 23, 2015*

Next Wednesday's NCBI Minute will show you how to quickly find literature about a gene of interest using [PubMed](#). NLM staff will highlight the links between gene data and literature and help you leverage the vocabulary used to describe gene information in PubMed to build a better search.

**Date and time:** Dec 2, 2015, 12:00-12:15 PM EST

**Registration URL:** <https://attendeegotowebinar.com/register/6661858858940556801>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses](#) page; you can also find information about future webinars on [this page](#).

The [NCBI Minute](#) is a series of short webinars that give a brief introduction to a specific topic or NCBI tool.

### New video on the NCBI YouTube channel: "Explore Gene pages at NCBI: Variation and Expression"

*Wednesday, November 18, 2015*

The newest video on the NCBI YouTube channel, *Explore Gene Pages at NCBI: Variation and Expression*, provides a walkthrough of how to explore variation and expression data using the Sequence Viewer embedded on [Gene](#) pages. Use these additional track options to quickly enhance your understanding of your genes of interest.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## PubChem adds a legacy designation for outdated data

*Tuesday, November 17, 2015*

PubChem has introduced a “legacy” designation to help users quickly identify records that may have outdated information or links. The designation applies to projects or contributors that appear to be inactive, as well as to their individual records. The [latest post on the PubChem blog](#) explains more about the designation, its impact, and its future in PubChem.

## NCBI to hold three-day genomics hackathon in January

*Friday, November 13, 2015*

From January 4th to 6th, NCBI will host a genomics hackathon focusing on advanced bioinformatics analysis of next generation sequencing data. This event is for students, postdocs and investigators already engaged in the use of pipelines for genomic analyses from next generation sequencing data.\* Working groups of 5-6 individuals will be formed for twelve teams, in the following sections: Network Analysis of Variants, Structural Variation, RNA-Seq, Streaming Data and Metadata, and Neuroscience/Immunity. The working groups will build pipelines to analyze large datasets within a cloud infrastructure. Please see the [application](#) for specific team projects.

\* *Specific projects are available to other developers or mathematicians.*

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems. This course will take place at the [National Library of Medicine](#) on the NIH main campus in Bethesda, Maryland.

### Datasets

Datasets will come from the public repositories housed at NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the hackathon.

### Products

All pipelines and other scripts, software and programs generated in this course will be added to a [public GitHub repository](#) designed for that purpose. A manuscript outlining the design of the hackathon and describing participant processes, products and scientific outcomes will be submitted to an appropriate journal.

### Application

To apply, complete this [form](#) (approximately 10 minutes to complete). Applications are due by **5pm ET on December 1**. Participants will be selected from a pool of applicants; prior students and prior applicants will be

given priority in the event of a tie. Please note: applicants are judged based on the motivation and experience outlined in the form itself.

Accepted applicants will be notified on **December 4th by 2 pm ET**, and have until **5pm on December 7** to confirm their participation. Please include a monitored email address, in case there are follow-up questions.

**Note:** Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals can be provided for this event. Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## Sequence Viewer 3.10.5 adds support for track sets with non-default options

*Friday, November 06, 2015*

Sequence Viewer 3.10.5 is now available with support for track sets with non-default display options. There are also a number of bug fixes, which are listed in the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## RefSeq Release 73 is now available

*Friday, November 06, 2015*

RefSeq Release 73 is now accessible online, on the [FTP site](#), and through NCBI's programming utilities. This full release incorporates genomic, transcript and protein data available as of November 2, 2015 and includes 83,881,439 records, 54,766,170 proteins, 12,998,293 RNAs, and sequences from 55,966 organisms. More information can be found in the [release notes](#).

For more information about the RefSeq project, please take a look at the [RefSeq homepage](#).

## Tree Viewer 1.7 now available

*Thursday, November 05, 2015*

NCBI Tree Viewer version 1.7 includes several new features, improvements and bug fixes, including a new rendering mechanism for displaying very large trees as image tiles. To see the full list of updates, see the Tree Viewer [release notes](#).

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## Registration open for November 18 NCBI Minute, "The New ClinVar Submission Wizard"

*Wednesday, November 04, 2015*

Next Wednesday, November 18th, the NCBI Minute will be an introduction and demonstration of the [new ClinVar Submission Wizard](#), a guided interface for direct data entry made for research laboratories that want to occasionally submit a small number of records.

**Date & time:** Nov 18, 2015, 12:00-12:15 pm EST

**Registration URL:** <https://attendeegotoweinar.com/register/7270975107138338562>

Submission to [ClinVar](#) is usually done through the Variation Submission Portal, which is useful for groups who frequently submit large number of variants, but may not be convenient for infrequent submitters of small numbers of variants. The new Submission Wizard is designed to support all types of submissions to ClinVar, including structural variants, pharmacogenomics variants, somatic variants, as well as interpretations based on functional rather than clinical significance.

After the live presentation, this webinar will be uploaded to the [NCBI YouTube channel](#). Any related materials will be accessible on the [Webinars and Courses page](#); you can also find information about future webinars on this page.

ClinVar is the NCBI archive of submitted interpretations of variants relative to diseases and other phenotypes.

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool.

## Researchers identify potential alternative to CRISPR-Cas genome editing tools

*Wednesday, November 04, 2015*

An international team of CRISPR-Cas researchers has identified three new naturally-occurring systems that show potential for genome editing. The discovery and characterization of these systems is expected to further expand the genome editing toolbox, opening new avenues for biomedical research. The research, published October 22nd in the journal [Molecular Cell](#), was supported in part by the National Institutes of Health.

"This work shows a path to discovery of novel CRISPR-Cas systems with diverse properties, which are demonstrated here in direct experiments," said Eugene Koonin, Ph.D., senior investigator at the [National Center for Biotechnology Information \(NCBI\)](#), National Library of Medicine (NLM), part of the NIH. "The most remarkable aspect of the story is how evolution has achieved a broad repertoire of biological activities, a feat we can take advantage of for new genome manipulation tools."

Enzymes from the CRISPR system are revolutionizing the field of genomics, allowing researchers to target specific regions of the genome and edit DNA at precise locations. "CRISPR" stands for Clustered Regularly Interspaced Short Palindromic Repeats, which are key components of a system used by bacteria to defend against invading viruses. Cas9 - one of the enzymes produced by the CRISPR system - binds to the DNA in a highly sequence-specific manner and cuts it, allowing precise manipulation of a region of DNA. Enzymes such as Cas9 provide researchers with a gene editing tool that is faster, less expensive and more precise than previously developed methods.

The three newly-characterized systems share some features with Cas9 and Cpf1, a recently characterized CRISPR enzyme, but have unique properties that could potentially be exploited for novel genome editing applications. This study highlights the diversity of CRISPR systems, which can be leveraged to develop more efficient, effective, and precise ways to edit DNA.

The researchers took a novel bioinformatics approach to discover the new proteins, provisionally termed C2c1, C2c2, and C2c3, developing a series of computational approaches to search NIH genomic databases and identify new CRISPR-Cas systems. In addition to Koonin, the research team included Feng Zhang of the Broad Institute

of MIT and Harvard and the McGovern Institute for Brain Research at MIT, Konstantin Severinov of Rutgers University – New Brunswick and the Skolkovo Institute of Science and Technology, Omar Abudayyeh, a graduate student at the Harvard- MIT Division of Health Sciences and Technology, and NCBI's Kira Makarova, Sergey Shmakov (also at Skolkovo Institute of Science and Technology), and Yuri Wolf.

"There are multiple ways to modify the search algorithm, so more exciting and distinct CRISPR-Cas mechanisms should be expected soon," said Severinov. "These new mechanisms will undoubtedly attract the attention of basic and applied scientists alike."

Initial experimental work exploring the function of these proteins reveals that they are substantially different from the well-characterized Cas9 protein, which has been widely used for genome editing.

With the analysis of C2c1, C2c2, and C2c3, the team was able to infer the intricate evolutionary pathway of these adaptive defense systems.

"The collaborative nature of this work highlights the power of bringing together top scientists with diverse strengths to innovate at the interface of computation, molecular biology and evolutionary biology," said Zhang.

The Koonin and Zhang groups also recently collaborated on [a project that resulted in the characterization of Cpf1](#), a novel CRISPR nuclease that is expected to become an important genome editing tool.

Feng Zhang, of the Broad Institute and MIT, is supported by the National Institute of Mental Health (5DP-MH100706 and 1R01-MH110049) and by the National Institute of Diabetes and Digestive and Kidney Diseases (5R01DK097760-03).

Konstantin Severinov, of Rutgers University and the Skolkovo Institute of Science and Technology, is supported by National Institute of General Medical Sciences (GM10407).

**About the National Center for Biotechnology Information (NCBI):** NCBI creates public databases in molecular biology, conducts research in computational biology, develops software tools for analyzing molecular and genomic data, and disseminates biomedical information, all for the better understanding of processes affecting human health and disease. NCBI is a division of the National Library of Medicine. For more information, visit [www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov).

**About the National Library of Medicine (NLM):** The world's largest biomedical library, NLM maintains and makes available a vast print collection and produces electronic information resources on a wide range of topics that are searched billions of times each year by millions of people around the globe. It also supports and conducts research, development, and training in biomedical informatics and health information technology. Additional information is available at [www.nlm.nih.gov](http://www.nlm.nih.gov).

**About the National Institutes of Health (NIH):** NIH, the nation's medical research agency, includes 27 Institutes and Centers and is a component of the U.S. Department of Health and Human Services. NIH is the primary federal agency conducting and supporting basic, clinical, and translational medical research, and is investigating the causes, treatments, and cures for both common and rare diseases. For more information about NIH and its programs, visit [www.nih.gov](http://www.nih.gov).

## dbVar publishes October 2015 data release

*Tuesday, November 03, 2015*

The dbVar October 2015 data release has recently been published. This month's release has 109,463 new Variant regions, 165,519 new Variant calls, and 5 new studies, including the following:

- **Decker et al. 2015 (nstd115)**: The authors created the largest existing catalog of canine genome-wide variation and used it to identify somatic variation in the thousands-years-old parasitic cancer clone, canine transmissible venereal tumor (CTVT).
- **User submitted curated variants (nstd51)**: A significant update to this collection of clinically relevant structural variants, curated by NCBI staff from PubMed, OMIM, and GeneReviews.
- **LSDB submitted variants (nstd103)**: A new collection of clinically relevant structural variants submitted by public LSDBs to ClinVar and brokered to dbVar.

Follow the [dbVar RSS feed](#) for monthly releases.

## Registration open for November 12th webinar, "PubMed for Scientists"

*Monday, November 02, 2015*

On November 12th, NCBI will present "PubMed for Scientists", a webinar that will show you how to search biomedical literature more efficiently with PubMed. NCBI staff will teach you how to search by author, explore a subject, use filters to narrow your search, find full text articles, and set up an email alert for new research on your topic. Finally, we will answer your questions about searching PubMed.

**Date and time:** Thursday, November 12, 2015 12:30 PM - 1:30 PM

**Registration URL:** <https://attendee.gotowebinar.com/register/5594790520765285889>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be accessible on the [Webinars and Courses](#) page by clicking the Archived Webinars & Courses tab. You can also check the [Webinars & Courses](#) page to find information about future webinars.



## NCBI News, October 2015

### Variation Viewer 1.4.1 is now available with optimized Variant Filters and Table performance

Thursday, October 29, 2015

Variation Viewer 1.4.1 provides several new features, improvements and bug fixes, including further optimized performance for Variant Filters and Variant Table. A full list of changes to Variation Viewer is available in the [release notes](#).

Variation Viewer is a tool for navigating variant data in [dbSNP](#), [dbVar](#) and [ClinVar](#) in a genomic context.

### New on the NCBI YouTube channel: "LinkOut - Linking to Datasets, Databases and More"

Thursday, October 29, 2015

The recording for the October 2nd webinar "LinkOut - Linking to Datasets, Databases and More" is available on [YouTube](#). The webinar presents an overview of LinkOut and highlights participating resources, with special emphasis on resources beyond full text articles, including databases, datasets and research tools.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

### OSIRIS Version 2.5 is now available

Wednesday, October 28, 2015

OSIRIS, NCBI's open source short tandem repeat (STR) analysis and quality assurance software package, has just been updated to [version 2.5](#). This version brings several improvements, including multiple peak labels and improved allele and artifact calling. A full list of updates is included in the [release notes](#).

The OSIRIS software for Windows/Mac, the User's Guide, and release notes are all freely available for download on the [OSIRIS homepage](#). In addition, an [OSIRIS webinar](#) is available on the NCBI YouTube channel.

OSIRIS was initiated in response to recommendations of a multidisciplinary advisory group empaneled by the U.S. Department of Justice, and was developed in collaboration with state, local and federal forensic laboratories and NIST.

### Outdated Genomes FTP directories will be archived on November 30, 2015

Tuesday, October 27, 2015

At the end of November 2015, many outdated Genomes FTP directories will be archived and no longer updated. If you get Genomes data from the NCBI FTP site, please prepare by checking the detailed list of changes below and updating your bookmarks, links and scripts where necessary before **November 30, 2015**.

#### FTP directories and files moving on November 30, 2015:

- All directories and files from <ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/> will be archived to [ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old\\_genbank/](ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_genbank/);

- The following directories from <ftp://ftp.ncbi.nlm.nih.gov/genomes/> will be archived to [ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old\\_refseq/](ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/);
  - *Aedes\_aegypti*
  - *Anopheles\_gambiae*
  - *Arabidopsis\_lyrata*
  - *Arabidopsis\_thaliana*
  - ASSEMBLY\_BACTERIA
  - Bacteria
  - Bacteria\_DRAFT
  - *Branchiostoma\_floridae*
  - *Caenorhabditis\_elegans*
  - Chloroplasts
  - CLUSTERS
  - *Drosophila\_melanogaster*
  - *Drosophila\_pseudoobscura*
  - Fungi
  - *Medicago\_truncatula*
  - MITOCHONDRIA
  - *Physcomitrella\_patens*
  - PLANTS
  - Plasmids
  - *Populus\_trichocarpa*
  - Protozoa
  - *Sorghum\_bicolor*
- The file *old\_genomeID2nucGI* from <ftp://ftp.ncbi.nlm.nih.gov/genomes/> will be archived to <ftp://ftp.ncbi.nlm.nih.gov/genomes/archive/>;
- The IDS directory from <ftp://ftp.ncbi.nlm.nih.gov/genomes/> will be moved to [ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME\\_REPORTS](ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS).

See the NCBI Genomes FTP FAQ for help on using the newer Genomes FTP directories, <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank> and <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq> that provide replacement content for the directories being archived.

## GenBank release 210.0 is now available via FTP

*Wednesday, October 21, 2015*

GenBank release 210.0 (10/15/15) has 188,372,017 non-WGS, non-CON records containing 202,237,081,559 base pairs of sequence data. In addition, there are 309,198,943 WGS records containing 1,222,635,267,498 base pairs of sequence data, as well as 81,790,031 TSA records containing 70,917,172,944 base pairs of sequence data.

During the 62 days between the close dates for GenBank releases 209.0 and 210.0, the traditional (i.e., non-WGS, non-CON) portion of GenBank grew by 2,413,437,272 base pairs and by 1,306,171 sequence records. During that same period, 280,348 records were updated and an average of 25,573 traditional records were added and/or updated each day.

Between releases 209.0 and 210.0, the WGS component of GenBank grew by 59,359,666,497 base pairs and by 6,243,400 sequence records; the TSA component of GenBank grew by 1,556,418,531 base pairs and by 6,036,982 sequence records.

The total number of sequence data files increased by 38 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 206
- ENV: 1 new file, now a total of 86
- INV: 3 new files, now a total of 132
- MAM: 9 new files, now a total of 37
- PAT: 6 new files, now a total of 235
- PLN: 5 new files, now a total of 119
- VRL: 1 new file, now a total of 38
- VRT: 3 new files, now a total of 59

For downloading purposes, please keep in mind that the uncompressed GenBank flat files require approximately 742 GB (sequence files only); the ASN.1 data require approximately 605 GB.

More information about GenBank release 210.0 is available in the [release notes](#).

## **New on the NCBI YouTube channel: Learn how to view track sets and store track collections**

*Friday, October 16, 2015*

Two new videos on the NCBI YouTube channel will show you how to view track sets in all of the NCBI genome browsers and Sequence Viewer displays and how to store and share custom sets of tracks in track collections.

*NCBI Recommended Tracks* presents track sets, which allow you to instantly tailor your display to a specific need, while *My NCBI Track Collections: Introduction* shows you how to store and share tracks in custom sets called track collections. If you'd like to learn more about track sets and collections, you can read the [FAQ](#) on the Sequence Viewer page.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## **Larger word size in modified algorithm speeds up BLASTP, BLASTX, TBLASTN search**

*Thursday, October 15, 2015*

The [NCBI BLAST webpage](#) now offers faster BLASTP, BLASTX and TBLASTN searches due to a modified algorithm that can use a larger word size. This improvement can make search 2-4 times faster without changing the results most of the time. Please see this [article](#) for more details on the modified algorithm.

**Note:** You may also recover the search's previous behavior by changing the word size from 6 to 3. To do so, expand "Algorithm parameters" at the bottom of the BLAST page and use the Word size menu (see Figure 1).

## **Variation Viewer 1.4 is now available with faster filter performance, track sets & collections**

*Thursday, October 08, 2015*

Variation Viewer 1.4 provides several new features, improvements and bug fixes, including significantly faster performance for Sequence Viewer and Variant Table filters, improved documentation readability, added [track sets and track collections](#), and more. A full list of changes to Variation Viewer is available in the [release notes](#).

[Variation Viewer](#) is a tool for navigating variant data in [dbSNP](#), [dbVar](#) and [ClinVar](#) in a genomic context.

The image shows the NCBI BLAST search interface. At the top, there are tabs for 'blastn', 'blasto', 'blastx', 'tblastn', and 'tblastx'. The main section is titled 'Enter Query Sequence' and includes a text input field for 'Enter accession number(s), gi(s), or FASTA sequence(s)', a 'Clear' button, and a 'Query subrange' section with 'From' and 'To' fields. Below this is an 'Or, upload file' section with a 'Choose File' button and 'No file chosen' text. There are also dropdowns for 'Genetic code' (set to 'Standard (1)') and a 'Job Title' field. A checkbox for 'Align two or more sequences' is present. The 'Choose Search Set' section includes a 'Database' dropdown (set to 'Non-redundant protein sequences (nr)'), an 'Organism' field with an 'Exclude' button, and checkboxes for 'Exclude' options like 'Models (XM/XP)' and 'Uncultured/environmental sample sequences'. There is also an 'Entrez Query' field. A 'BLAST' button is visible. The 'Algorithm parameters' section is expanded, showing 'General Parameters' with 'Max target sequences' (100), 'Expect threshold' (10), 'Word size' (6, highlighted with a red box), and 'Max matches in a query range' (0).

**Figure 1.** You can use the word size menu (outlined in red) to recover BLAST search's previous behavior.

## Sequence Viewer 3.10 adds support for track sets and track collections, performance optimization and more

Monday, October 05, 2015

Sequence Viewer has been updated to version 3.10, bringing new features, improvements and bug fixes. These changes include:

- Support for [track sets and track collections](#)
- Performance optimization allowing faster switching between molecules in genomic browsers and faster Sequence Viewer staging
- A new display option for alignment track

A full list of changes to Sequence Viewer is available in the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## New NCBI Insights blog post: "Troubleshooting GenBank Submissions: Annotating the Coding Region (CDS)"

Friday, October 02, 2015

The [latest blog post](#) on NCBI Insights gives GenBank data submitters a workflow for using BLAST to troubleshoot problems with CDS feature annotation. This information is also available in two webinars on the NCBI YouTube channel: [Coding Region Annotation](#) and [Eukaryotic CDS Annotation](#).

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips, and what's new at NCBI.

Subscribe to the [NCBI YouTube channel](#) to receive alerts about new videos ranging from quick tips to full webinar presentations.

## **NCBI staff to attend and present at ASHG 2015**

*Friday, October 02, 2015*

NCBI will participate in the [ASHG annual meeting](#) in Baltimore, MD (Oct. 6-10). Staff will participate in the Genome Reference Consortium workshop and present twelve different posters on updated tools and resources for clinical genetics, genomics, and human genome assembly and annotation.

NCBI staff members will also be at the NCBI Exhibit Booth (**#2405**), where attendees can get answers and provide input for the future development of NCBI human genome resources.



## NCBI News, September 2015

### First offering of NCBI NOW (Next generation sequencing Online Workshop) to begin October 13, 2015

Wednesday, September 30, 2015

From October 13th to October 23rd, NCBI will present the first iteration of NCBI NOW, a free online experience aimed at those new to next generation sequencing (NGS) analysis. Enrollment in this course is limited to the first 1,000 participants who sign up through the [ORAU Portal](#). Since enrollment is on a first-come, first-served basis, please only sign up for this educational opportunity if you will be able to participate fully.

**NCBI** NIH

**NOW!**

---

An online workshop that introduces next generation sequencing with a hands-on component in the cloud!

7  
online  
lectures

self-guided

hands-on

DNaseq, BLAST, RNaseq, GEO, SRA

Learners will watch 6-7 videos (average video duration: 45-60 minutes) online during the first 7 days of the course. These videos will cover the basics of NGS data, preprocessing, quality control and alignment strategies

for both DNA-Seq and RNA-Seq, as well as a brief discussion of downstream analysis. Additionally, we will demonstrate how to leverage BLAST tools for NGS analysis.

Next, participants will apply a selection of RNA-Seq alignment algorithms over three days (1-2 hours per day), mapping RNA-Seq data to [GRCh38](#) chromosome 20. Finally, participants will compare the results of these mappers for specific genes. Throughout the course, participants will be able to post questions at [Biostars](#); experts from NCBI and elsewhere will be available online to answer questions.

Learners will emerge from the course equipped to map their own RNA-Seq or DNA-Seq data to the human genome, understand the options for downstream analysis, and use their understanding of the basic steps of data processing to interact more effectively with bioinformatician collaborators.

**A note about registration:** When registering for an account, you will be prompted for a partition. Please ensure that NCBI is selected, as this will provide access to the proper materials for this workshop. However, if you have already created an account without selecting this option, there is no need to create a second account. Your account has been reassigned to the NCBI partition.

## September 30th NCBI Minute: Preview of NCBI at American Society of Human Genetics 2015

*Thursday, September 17, 2015*

On September 30th, NCBI staff will provide a quick overview of NCBI activities at this year's [American Society of Human Genetics](#) (ASHG) meeting, including previews of NCBI posters and presentations on tools and resources for clinical genetics, genomics and the assembly and annotation of the human genome.

NCBI will participate in the ASHG annual meeting in Baltimore, MD, October 6-10, 2015. Staff members will be at exhibit booth #2405, where attendees can get answers and provide input for the future development of NCBI human genome resources.

**Date and time:** Wednesday, September 30, 2015 12:00PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/8493856336405913090>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses page](#), where you can also find information about future webinars.

## "Create a Biosketch with SciENcv" webinar recording on YouTube

*Wednesday, September 16, 2015*

The recording of the July 30th webinar (Announcement) on SciENcv and Biosketch is available on [YouTube](#). The webinar shows you how to use SciENcv to maintain your scientific record and generate the new NIH Biosketch; this [NCBI Insights blog post](#) complements the webinar.

To receive alerts about new videos ranging from quick tips to full webinar presentations, subscribe to the [NCBI YouTube channel](#).



## October 2nd webinar - LinkOut: Linking to datasets, databases and more

*Wednesday, September 16, 2015*

On October 2nd, NCBI staff will present a webinar on [LinkOut](#), an NCBI service that allows you to link directly from NCBI databases to a wide range of relevant information beyond the NCBI systems. This webinar will provide an overview of the service and highlight resources that participate in LinkOut, with a special emphasis on resources beyond full text articles, including databases, datasets and research tools.

If you use NCBI databases, produce databases, datasets or resources, or are a librarian supporting research and data science, this webinar is for you.

**Date and time:** Friday, October 2, 2015 12:00PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/5533821500870613249>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses](#) page, where you can also find information about future webinars.

## New NCBI Insights blog post: "Finding Chemical Probes & Modulators - The Hunt for New Chemical Reagents and Medicines"

*Friday, September 11, 2015*

The [latest blog post](#) on [NCBI Insights](#) will show you how to use PubChem to find chemicals that bind to a particular gene or protein target and download a table with that information. In [last week's post](#), we showed you how to do the reverse - find and download a table of gene and protein targets for a particular chemical.

[NCBI Insights](#) is the official NCBI blog, where we share science features, quick tips and updates on what's new at NCBI.

## NCBI to hold fourth offering of "A Librarian's Guide to NCBI"

*Thursday, September 10, 2015*

The NCBI, in partnership with the [National Library of Medicine Training Center](#) (NTC), will once again offer the *Librarian's Guide to NCBI* course on the NIH campus, March 7 - 11, 2016 ([Announcement](#)). If you are a medical or science librarian in the United States who offers bioinformatics education and support services, or are planning to offer such services in the future, please join us for this intensive exploration of modern molecular biology, genetic data and other biomedical data as represented at the NCBI.

The course explains how and why these data are generated, their importance in modern biomedical research, and how to access them through the NCBI website. The [online application](#) is now open. The application deadline is **September 14, 2015**. This is a combined application for the prerequisite online *Fundamentals of Bioinformatics and Searching* and the five-day in-person course. If you have already completed the *Fundamentals* class, please let us know on the application. Anyone who has previously taken *Fundamentals* is eligible to apply for the in-person course in March at NIH.

**Prerequisite:** *Fundamentals of Bioinformatics and Searching*

Because of the fast-paced and intense nature of the course, all applicants for *A Librarian's Guide* must successfully complete the online *Fundamentals of Bioinformatics and Searching* class offered through the NTC. The *Fundamentals* class is an introduction to molecular biology and bioinformatics taught by Diane Rein, Ph.D, MLS that provides essential background for the in-person course.

For those who have not yet taken the online course, a special section of Fundamentals will be offered in the fall of 2015 from **October 26 - December 11**, with classes not held the week of Thanksgiving.

## September 16th NCBI Minute: Accessing the Human Genomics Standard Data (Genome in a Bottle) at NCBI

*Wednesday, September 09, 2015*

In the next NCBI Minute, we will show you how to access the Genome in a Bottle dataset from the NCBI site and present potential use cases.

**Date and time:** Wednesday, September 16, 2015 1PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/1790003099972245250>

The [Genome in a Bottle Consortium](#) aims to provide well-characterized reference materials for human clinical sequencing and spur development and optimization of genomics technology and bioinformatics.

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses page](#), where you can also find information about future webinars.

## New NCBI Insights blog post: "Identifying Chemical Targets - Finding Potential Cross-Reactions and Predicting Side Effects"

*Friday, September 04, 2015*

The latest NCBI Insights [blog post](#) will show you how to use PubChem to download a table of gene and protein targets for a particular chemical, which can help to find potential cross reactions and side effects. Next week, we will show you how to do the opposite - download a table of chemicals that bind to a particular gene or protein target.

[NCBI Insights](#) is the official NCBI blog, where we share science feature stories, quick tips and updates on what's new at NCBI, including [PubMed Labs](#).

## RefSeq Release 72 is now available

*Thursday, September 03, 2015*

RefSeq Release 72 is now available online, on the [FTP site](#), and through NCBI's programming utilities. The full release incorporates genomic, transcript, and protein data available, as of August 27, 2015 and includes 79,189,847 records, 51,933,925 proteins, 12,321,036 RNAs and sequences from 54,937 organisms. More information can be found in the [release notes](#).

Release 72 includes over 43,000 Archaeal and Bacterial RefSeq genomes that were annotated since July 29, 2015 using one version of the prokaryotic genome annotation pipeline software (version 3.0). This re-annotation increases overall consistency across the dataset because all genomes were annotated in a short window of time using the same software and protein alignment data.

This release also incorporates modified policies related to protein annotation that address concerns with very short partial proteins (fragments), partial proteins where the neighboring sequence is good (a/c/g/t bases and no gaps), and proteins that incorporated many 'X' amino acid residues due to low quality genome sequence for the CDS region. These changes in protein annotation resulted in the suppression of 6.4 million non-redundant (WP\_) protein records., the majority of which were partial proteins (>4 million).

At the same time, we created 3.8 million new protein records as we added approximately 12,000 new prokaryotic genomes to the RefSeq collection. An average of 316 new protein records were added for each of the 12,000 new genomes.; the remaining protein complement per genome is based on pre-existing non-redundant protein records. For more information about the RefSeq project, pipelines and data, please take a look at the [RefSeq homepage](#).

## HIV-1 interaction datasets in Gene updated

*Tuesday, September 01, 2015*

NCBI has added data provided by the Southern Research Institute (SRI) to the HIV-1 interaction datasets available in [Gene](#).

The [protein interactions dataset](#) now has:

- 7,567 interactions;
- 15,074 interaction descriptions;
- 3,623 proteins encoded by 3,582 human genes;
- and 6,610 publications.

The [replication interactions dataset](#) now has:

- 1,298 interactions;
- 1,369 interaction descriptions;
- 1,298 proteins encoded by 1,298 human genes;
- and 94 publications.

Data are also available at the [RefSeq HIV-1 website](#) and the [GeneRIF FTP site](#).

## Genome Workbench 2.9.5 now available

*Tuesday, September 01, 2015*

[Genome Workbench](#) 2.9.5 is available, as of August 27th. New features include added support for Mac OS 10.10 (Yosemite) and an added ruler for protein coding regions in graphical sequence view.

For the full list of fixes, improvements and features, see the [Genome Workbench release notes](#).

## dbSNP build 145 (pig, chicken, sorghum, gibbon) now available

*Tuesday, September 01, 2015*

dbSNP build 145 is now available through the integrated NCBI Entrez system and [FTP](#). This release includes data for pig, chicken, sorghum and gibbon. Build 145 provides more than 195 million submitted variants and 84 million reference variants for 4 species. To see complete build statistics, visit the [SNP summary page](#).



## NCBI News, August 2015

### NCBI annotates 250th eukaryote with Eukaryotic Genome Annotation Pipeline

*Friday, August 28, 2015*

This month, the [NCBI Eukaryotic Genome Annotation Pipeline](#) has annotated its 250th organism! Mammals dominate the list of annotated organisms with a total of 95, but NCBI has increased coverage of invertebrates - we've annotated 22 new insects since the beginning of 2015. See the [full list of annotated organisms](#) and request that your favorite organism(s) be annotated!

We make an effort to re-annotate genomes every two years so the latest annotation incorporates recently submitted RNA-Seq and Transcript Shotgun Assemblies as evidence and benefits from the [latest software developments](#). Data produced by the Eukaryotic Genome Annotation Pipeline is available in the [Reference Sequences \(RefSeq\)](#) collection, [BLAST](#) non-redundant and organism-specific databases, and the [Gene](#) database; it is also downloadable from the [NCBI FTP site](#).

### September 2nd NCBI Minute: "Introducing SmartBLAST, a Rapid Protein Identification Tool"

*Thursday, August 20, 2015*

On September 2nd, NCBI staff will introduce [SmartBLAST](#), a faster alternative to ordinary protein-protein BLAST searches for protein query sequence identification.

SmartBLAST reports the top three results from a separate database of high quality protein sequences, as well as the top two hits from nr. SmartBLAST also produces a full multiple alignment of the query sequence and results with mapped conserved domains. SmartBLAST is part of the new [PubMed Labs initiative from NCBI](#).

**Date and Time:** September 2nd, 2015 12:15 PM EDT

**Registration URL:** <https://attendeegotowebinar.com/register/7468878402343662081>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses page](#), where you can also find information about future webinars.

### GenBank release 209.0 is now available via FTP

*Wednesday, August 19, 2015*

[GenBank release 209.0](#) (8/14/2015) has 187,066,846 non-WGS, non-CON records containing 199,823,644,287 base pairs of sequence data. In addition, there are 302,955,543 WGS records containing 1,163,275,601,001 base pairs of sequence data, as well as 87,827,013 TSA records containing 69,360,654,413 base pairs of sequence data.

During the 57 days between the close dates for GenBank releases 208.0 and 209.0, the traditional (i.e., non-WGS/non-CON) portion of GenBank grew by 5,902,601,341 base pairs and by 2,047,494 sequence records. During that same period, 288,641 records were updated. An average of 40,985 traditional records were added and/or updated per day.

Between releases 208.0 and 209.0, the WGS component of GenBank grew by 124,338,390,780 base pairs and by 44,253,405 sequence records; the TSA component of GenBank grew by 8,663,181,843 base pairs and by 10,852,412 sequence records.

The total number of sequence data files increased by 59 with this release. The divisions are as follows:

- BCT: 9 new files, now a total of 196
- CON: 4 new files, now a total of 323
- ENV: 4 new files, now a total of 85
- GSS: 2 new files, now a total of 299
- INV: 1 new file, now a total of 129
- MAM: 19 new files, now a total of 28
- PAT: 6 new files, now a total of 229
- PHG: 1 new file, now a total of 3
- PLN: 2 new files, now a total of 114
- VRL: 1 new file, now a total of 37
- VRT: 10 new files, now a total of 56

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 735GB (sequence files only); the ASN.1 data are approximately 600GB.

More information about GenBank release 209.0 is available in the [release notes](#).

## Tree Viewer 1.6 now available

*Wednesday, August 19, 2015*

NCBI Tree Viewer version 1.6 includes several new features, improvements and bug fixes, including the added ability to download data in Newick and Nexus formats. To see the full list of updates, see the Tree Viewer [release notes](#).

NCBI Tree Viewer is a tool for viewing your own phylogenetic tree data.

## New NCBI video: "NCBI's 1000 Genomes Browser: Introduction"

*Wednesday, August 12, 2015*

The newest video on the [NCBI YouTube channel](#) is an introduction to the [1000 Genomes Browser](#), which allows you to view variation and genotype data, and support sequence reads from the 1000 Genomes Project.

Subsequent videos will cover other functions, such as uploading data. For video updates, subscribe to our [YouTube channel](#).

## August 26th webinar: "Troubleshooting GenBank Submissions: Determining and Annotating Coding Regions (CDS) for Eukaryotic Genes"

*Wednesday, August 12, 2015*

In two weeks, NCBI staff will show you how to use BLAST to determine the locations of the coding sequences in your genomic submissions. You will also learn how to describe these coding regions with the GenBank submission tools BankIt and Sequin, annotate multiple transcript splice variants, and address problems with splice sites, internal stop codons in protein translations, and sequencing gaps that affect coding region annotation.

**Date and Time:** August 26, 2015 1PM EDT

**Registration URL:** <https://attendee.gotowebinar.com/register/3143702023795693569>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses page](#), where you can also find information about future webinars.

## New NCBI Insights blog post: "SciENcv Updated to Support New NIH Biosketch Format"

Monday, August 10, 2015

The latest blog post on [NCBI Insights](#) will show you how to use [SciENcv](#) to convert your existing NIH Biosketch from the old format to the new format required for grant applications submitted with due dates after May 24, 2015.

## Genomes FTP site update (version 1.2) expands taxonomic scope and more

Wednesday, August 05, 2015

NCBI has released a comprehensive update of all current genome assemblies in the [Genomes FTP site](#), affecting data reported in the [/genomes/all/](#), [/genomes/genbank/](#), and [/genomes/refseq/](#) FTP directories. This update expands the taxonomic scope of the [/refseq/](#) data and adds a new report file, a data conversion script, and more. The FTP content of all "latest" GenBank and RefSeq assemblies was updated to reflect these changes.

### Genomes FTP version 1.2 includes the following changes:

- Genome group directories:
  - Assembly summary files have been added to genbank and refseq genome group directories (e.g., [ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/fungi/assembly\\_summary.txt](ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/fungi/assembly_summary.txt))
  - Viral RefSeq genomes have been added to the Assembly database and a new genome group directory, [/viral/](#), is now available: <ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/viral/>
- New feature table report:
  - Files named as `*_feature_table.txt.gz` are tab-delimited files reporting annotated features, coordinates, and attributes (including names). Feature types reported include gene, CDS, RNA (all types), operon, immunoglobulin C/V/N/S regions and V/D/J segments.
- WGS master files:
  - This record type has been added to the FTP release and is provided in GenBank flatfile format using the file name convention `*_wgsmaster.gbff.gz`
- Conversion script:
  - `add_utrs_to_gff.py`: Python script to add explicit UTR exon features, as inferred from the gene, mRNA, exon and CDS features, to GFF3 formatted data. Script location: [ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/add\\_utrs\\_to\\_gff/](ftp://ftp.ncbi.nlm.nih.gov/genomes/TOOLS/add_utrs_to_gff/)
- GBFF format:
  - Genomic records in the CON division now include both a CONTIG line and the sequence. For example, the GenBank flatfile format for GG698602.1, a GenBank scaffold in the *Dialister invisus* DSM 15470 Assembly GCA\_000160055.1, shows both the CONTIG and ORIGIN/sequence data:

```
CONTIG      join(ACIM02000001.1:1..1894898 , gap(unk100) , ACIM02000002.1:1..962)
ORIGIN
1 caaggcttgg agcgacataa aactaatagg tcgaggtctt aacttaggaa caccgagaca (ETC.)
```

- GFF3:
  - Additional information about NCBI's GFF3 files is now available at [ftp://ftp.ncbi.nlm.nih.gov/genomes/README\\_GFF3.txt](ftp://ftp.ncbi.nlm.nih.gov/genomes/README_GFF3.txt)
  - GFF files now include information on the gene biotype
- Assembly summary files:
  - Assembly levels have been simplified to four types: contig, scaffold, chromosome, and complete genome
- Assembly reports:
  - The length of each sequence has been added to the report
  - UCSC style names (e.g., chr1) have been added to the report for those sequences that have been matched to assemblies on the UCSC genomes FTP site (e.g., /genomes/all/GCF\_000001405.30\_GRCh38.p4/GCF\_000001405.30\_GRCh38.p4\_assembly\_report.txt)

Please note that RefSeq prokaryotic genome annotation is currently being refreshed. Once that process is complete, we will update the data in the Genomes FTP site.

## CCDS release 19 for mouse added to Gene

*Tuesday, August 04, 2015*

The [Consensus Coding Sequence \(CCDS\)](#) update that compares NCBI's *Mus musculus* annotation release 105 to Ensembl's release 81 is now reflected in [Gene](#). This update adds 1,003 new CCDS IDs and adds 148 Genes into the mouse CCDS set. CCDS release 19 includes a total of 24,834 CCDS IDs that correspond to 20,215 GeneIDs.

For information about CCDS, please visit the [CCDS homepage](#).



## NCBI News, July 2015

### **New NCBI Insights blog post: Introducing PubMed Labs, an NCBI initiative to include user community in product development from beginning**

*Wednesday, July 29, 2015*

Today on [NCBI Insights](#), we announced PubMed Labs, an initiative for creating innovative and relevant products by involving you, our user community, from the start.

PubMed Labs is centered upon our user community, experimentation, learning and conversation. In "[Introducing PubMed Labs](#)", we describe what you can expect from PubMed Labs, how to find our first new experimental features, SmartBLAST and PubMed also-viewed, and how you can provide us with feedback, which we'll use to improve our services for our users.

To read about SmartBLAST and PubMed also-viewed and try them out, visit [NCBI Insights](#). We look forward to hearing your thoughts on PubMed Labs.

### **August 12th NCBI Minute: Using Variation Reporter to Map and Annotate Your Own Variant Calls**

*Tuesday, July 28, 2015*

In two weeks, NCBI staff will show you how to use [Variation Reporter](#) to submit your own variant calls for analysis and quickly view and interpret mapping results. Variation Reporter is an interface to NCBI's variation resources that quickly provides genomic context, phenotypic assertions and allele frequency for known variants in your data. It also maps and predicts consequences for genes and gene products for variants not in the NCBI databases.

Date and Time: August 12, 2015 12 PM EDT

Registration URL: <https://attendee.gotowebinar.com/register/6563285440761995778>

After the live presentation, the webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any materials will also be archived on the [Webinars and Courses](#) page, where you can also find information about future webinars.

### **Sequence Viewer 3.9 adds data upload options to API, improved response time and more**

*Tuesday, July 28, 2015*

Recent updates to [Sequence Viewer](#) bring the following features and improvements to version 3.9:

- Data upload option added to track and data upload API
- Improved response time
- Improved HTTPS compliance

A full list of features, improvements and bug fixes is available in the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## **August 5th NCBI Minute: "Using EDirect's Xtract Utility to Parse NCBI BLAST XML Output"**

*Wednesday, July 22, 2015*

The next NCBI Minute will introduce the EDirect Xtract XML parser, a useful tool for processing NCBI BLAST XML output. Future NCBI Minute webinars will show additional ways to use the EDirect suite to enhance and customize standalone BLAST.

**Date:** Wednesday, August 5, 2015

**Time:** 12:00PM EDT

### **Registration**

After the live presentation, this webinar will be uploaded to the [NCBI YouTube channel](#). The webinar and any relevant materials will also be archived on the [Webinars and Courses](#) page, where you can also find information about upcoming webinars. For an introduction to EDirect, you can watch our webinar from June 15th on [YouTube](#).

## **July 30th webinar: "Using SciENcv to Create Your NIH Biosketch"**

*Wednesday, July 15, 2015*

In two weeks, NCBI staff will present a webinar on SciENcv, our platform for maintaining your record of research accomplishment in the form of a CV. In this webinar, we'll show you how to use SciENcv to maintain your scientific record and generate the new BioSketch. Register here: <https://bit.ly/1f7wWC8>.

A recording will be posted on the [NCBI YouTube account](#) after the live presentation; subscribe to our YouTube channel to be notified. To see upcoming webinars and materials from past presentations, visit the [Webinars and Courses](#) page.

## **July 22nd NCBI Minute webinar: Find disease-related variants in ClinVar**

*Wednesday, July 15, 2015*

Next Wednesday, July 22, NCBI staff will show you how to quickly find variants related to human disease in the NCBI ClinVar resource, as well as how to download batches of variants and related information in .xml and .vcf formats. To sign up, go here: <https://bit.ly/1gxkBYy>

ClinVar is NCBI's repository for human variation and its relationship to health and disease. ClinVar is an essential resource for basic researchers, clinicians and genetic counselors.

To see upcoming webinars and materials from past presentations, visit the [Webinars and Courses](#) page.

## **RefSeq Release 71 is now available!**

*Monday, July 13, 2015*

RefSeq Release 71 is now available online, on the [FTP site](#), and through NCBI's programming utilities, with 77,730,891 records describing 52,494,032 proteins, 11,803,354 RNAs, and sequences from 55,267 organisms. More information can be found in the [Release Notes](#).

Please note that we plan to comprehensively re-annotate bacterial and archaeal genomes for RefSeq Release 72 (September 2015). This re-annotation is being carried out to reflect improvements in a) management of partial, very short, and fragmented genes and proteins and b) protein name management. It will also result in increased consistency of some textual information applied to RefSeq records. To learn more about the re-annotation project and what NCBI is doing to help users in transitioning to using this new data, please see the [RefSeq Re-annotation Project page](#).

More information about the RefSeq project, pipelines and data, please take a look at the [RefSeq homepage](#).

## **July 15th webinar: "EDirect: Bringing the E-Utilities to the UNIX Command Line"**

*Wednesday, July 01, 2015*

In two weeks, NCBI staff will introduce EDirect, a simple, easy-to-use command-line interface for E-Utilities, the NCBI Entrez API. In this webinar, you will learn how to use EDirect to search, retrieve and process literature and molecular data from NCBI. You will also learn how to set up pipelines for common tasks.

To sign up for this webinar, go here: <https://attendee.gotowebinar.com/register/6848186495038667265>. Like all of our webinars, this presentation will be posted on the [NCBI YouTube account](#) after the live presentation; subscribe to our YouTube channel to be notified of all of our new videos.

To see upcoming webinars, as well as related materials and recordings from past webinars, visit the [NCBI Webinars page](#).



## NCBI News, June 2015

### Tree Viewer version 1.5 improves performance

*Tuesday, June 30, 2015*

NCBI Tree Viewer version 1.5 includes several new features, improvements and bug fixes, including improved tree loading notifications, fixed resizing in full view and more. For a full list of updates, see the [Tree Viewer release notes](#).

NCBI Tree Viewer is a tool for viewing your phylogenetic tree data.

### June 3rd webinar "Troubleshooting GenBank Submissions: Coding Region Annotation" video up on YouTube

*Friday, June 26, 2015*

The recording of the June 3rd webinar on troubleshooting internal stop codon errors has been uploaded to [YouTube](#). In this webinar, you will learn how to troubleshoot internal stop codons encountered during coding region (CDS) annotation. You will also learn how to analyze your sequences and uncover problems with BLAST prior to submitting them to GenBank.

For those who cannot access YouTube, the video is also available via [FTP](#). A .txt file of the video's subtitles is also included within that directory.

Subscribe to the [NCBI YouTube channel](#) to be notified of our new videos, which range from quick tips to full webinar presentations.

### June 10th webinar "Phylogenetic Trees in Genome Workbench" video up on YouTube

*Friday, June 26, 2015*

The recording of the June 10th webinar on Genome Workbench's phylogenetic trees has been uploaded to [YouTube](#). In this webinar, you will learn how to use Genome Workbench to work with phylogenetic trees. You will also see the numerous options for importing trees, for visualising, searching and modifying the trees, and for saving or exporting them.

For those who cannot access YouTube, the video is also available via [FTP](#). A .txt file of the video script is also included within that directory.

Subscribe to the [NCBI YouTube channel](#) to be notified of our new videos, which range from quick tips to full webinar presentations.

### New dbVar webinar available on NCBI YouTube channel

*Thursday, June 25, 2015*

A new webinar highlighting [dbVar](#) - NCBI's database of genomic structural variation - is available and consists of two parts: [Part 1](#) is a slide presentation that explores the dbVar website and demonstrates how to navigate its features and tools, including viewing variants in a genome browser; [Part 2](#) is a live demo, with use cases for finding useful information at dbVar as well as direct links to related information at other NCBI resources - for example, clinical assertions at ClinVar.

Subscribe to the [NCBI YouTube channel](#) to be notified of new videos, which range from quick tips to full webinar presentations.

## GenBank release 208.0 is now available via FTP

*Tuesday, June 23, 2015*

GenBank Release 208.0 (6/18/2015) has 185,019,352 non-WGS, non-CON records containing 193,921,042,946 base pairs for sequence data. In addition, there are 258,702,138 WGS records containing 1,038,937,210,221 base pairs of sequence data, as well as 76,974,601 TSA records containing 60,697,472,570 base pairs of sequence data.

During the 66 days between the close dates for GenBank releases 207.0 and 208.0, the non-WGS/non-CON portion of GenBank grew by 69,834,303,408 base pairs and by 14,922,939 sequence records. During that same period, 792,075 records were updated and an average of 54,889 non-WGS/non-CON records were added and/or updated per day.

Between releases 207.0 and 208.0, the WGS component of GenBank grew by 24,303,492,065 base pairs and by 5,163,496 sequence records; the TSA component of GenBank grew by 4,901,140,135 base pairs and by 4,985,013 sequence records.

The total number of sequence data files increased by 59 with this release. The divisions are as follows:

- BCT: 9 new files, now a total of 187
- CON: 2 new files, now a total of 319
- GSS: 4 new files, now a total of 297
- HTG: 9 new files, now a total of 151
- INV: 2 new files, now a total of 128
- PAT: 4 new files, now a total of 223
- PLN: 5 new files, now a total of 112
- PRI: 1 new file, now a total of 49
- TSA: 20 new files, now a total of 195
- VRL: 2 new files, now a total of 36
- VRT: 1 new file, now a total of 46

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 722 GB (sequence files only). The ASN.1 data require approximately 592 GB.


More information about GenBank release 208.0, including details about important changes included in this release, is available in the [release notes](#).

## NCBI Southern California Regional Workshops to be held June 30 - July 2

*Thursday, June 18, 2015*

NCBI Service Desk staff will present a workshop series (June 30 - July 2, 2015) at the Beckman Research Institute of City of Hope in Duarte, California. These [five individual modules](#) include lectures and hands-on demonstrations intended for physicians and researchers who want to learn more about NCBI resources and how to use them to advance research projects.

The workshop series is free, but it is limited to 140 people. Register online at <http://events.signup4.net/UpdateNCBIResource2015>. More information about the event is also available on the registration page and on the flyer below (click [here](#) to download).



Beckman Research Institute of City of Hope  
 Bioinformatics Core Facility &  
 Lee Graff Medical and Scientific Library

in collaboration with  
**The National Center for Biotechnology Information (NCBI)**

invites you to a hands-on workshop series presented by  
 Peter Cooper, Ph.D. and Wayne Matten, Ph.D.  
 NCBI Service Desk

**“An Update on National Center for  
 Biotechnology Information (NCBI) Resources”**

June 30 – July 2, 2015  
 9:00AM - 4:00PM

For additional information, including a detailed schedule  
 & registration instructions, please visit our website:  
<http://events.signup4.net/UpdateNCBIResource2015>

**Registration Deadline: Friday, June 26, 2015**  
(Limited to the first 140 people)

<p><b>Location:</b> City of              Hope Attyros              Auditorium              Arnold and Mabel Beckman Building              1500 E. Duarte Road, Duarte, CA 91010</p>	<p><b>Contact:</b>              Ryan Chiechi              Research Finance &amp; Shared Services              626-256-4673 x89001</p>
---	---

PLEASE NOTE: The workshop is free but registration is required. The sessions will be recorded and made available to those who registered. Each module will have lecture followed by the hands-on computer practice. We encourage people to bring their own laptops. Please arrive 15 mins before each half-day session to allow extra time for computer setup. A limited number (25) of laptops are available for CCFI attendees on a first come, first served basis, determined by the date we receive the online registration.

Light breakfast and lunch will be served.

## New YouTube video: "Sequence Viewer: Navigate Objects with Jump Arrows"

*Thursday, June 18, 2015*

The newest video on the NCBI YouTube channel, "Sequence Viewer: Navigate Objects with Jump Arrows", introduces jump arrows, a feature recently added to [Sequence Viewer](#). A full list of new features, improvements and fixes can be found in the [release notes](#).

Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

Subscribe to the [NCBI YouTube channel](#) to be notified of our new videos, which range from quick tips to full webinar presentations.

## UniVec build 9.0 now available for VecScreen searches and FTP

*Tuesday, June 16, 2015*

UniVec, NCBI's non-redundant database of vector sequences, has been updated to build 9.0, which enables searches run using NCBI's [VecScreen](#) tool to detect more of the foreign sequences introduced during the cloning or sequencing process. UniVec build 9.0 is also available via [FTP](#).

This build added 252 complete vector sequences and 124 adapter and primer sequences, including many oligonucleotides used in next-generation sequencing protocols, bringing the total number of sequences represented in the UniVec database to 2,658.

UniVec is a non-redundant database of sequences commonly attached to cDNA or genomic DNA during the cloning process. UniVec primarily consists of the unique segment from a large number of vectors but also includes many linker, adapter and primer sequences. Redundant sub-sequences have been eliminated from the database to make searches more efficient and to simplify interpretation of the results. For more details, see the [UniVec page](#).

## BLAST+ stand-alone updated to version 2.2.31

*Tuesday, June 16, 2015*

A new version (2.2.31) of the stand-alone BLAST executables (Linux, Windows and MacOSX on [FTP](#)) is now available. New features include support for BLAST-XML2 specification (information [here](#)) and JSON BLAST output format, as well as several bug fixes and improvements. The BLAST AMI at AWS will also be updated to 2.2.31 (see this BLAST Help page for more [information](#)). For a full list of improvements, see the [release notes](#).

### Related NCBI News stories:

- [June 26, 2014](#): BLAST machine image hosted at Amazon Web Services (AWS)
- [October 16, 2014](#): Amazon Web Services (AWS) Marketplace provides the easiest way to start an NCBI BLAST instance

## Complete MERS coronavirus genomes from China and South Korea are in GenBank

*Wednesday, June 10, 2015*

Two complete MERS coronavirus genomes are in GenBank: one from China ([KT006149](#)) and one from South Korea ([KT029139](#)). In addition, the [MERS coronavirus resource page](#) gives users an easy way to find all sequences related to this pathogen.

The MERS coronavirus resource page has three components designed to support users' discovery activities: the database, the reference genome graphical display, and links to other virus data resources, both external (e.g., CDC, WHO, and HealthMap) and within NCBI.

The database allows you to search for nucleotide and protein sequences by a variety of criteria, including host, sequence patterns, region or country of isolation, and collection or release dates. Using the database, you can:

- Quickly find the sequences you need, through an intuitive search interface, using standardized protein/gene names and metadata
- Select the latest sequences based on date criteria or sorting of results
- Download sequences in many formats and find links to sequences in NCBI databases.

## dbSNP build 144 now available

*Tuesday, June 09, 2015*

dbSNP build 144 data is now available through the integrated NCBI Entrez system and [FTP](#). This release includes data for human, chicken, soybean and horse. Build 144 provides more than 601 million submitted and 191 million reference variants for 4 species. To see complete build statistics, visit the [SNP summary page](#).

## NCBI Sequence Viewer version 3.8 available

*Wednesday, June 03, 2015*

NCBI [Sequence Viewer](#) has recently been updated and now has a new track navigation GUI that allows you to quickly find features like gene, exon, SNP and alignment; it also has a new API option for negative graph values and improved variation tooltips. A full list of new features, improvements and fixes is included in the [release notes](#).



Sequence Viewer is a graphical view of sequences and color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## June 10th webinar: "Working with Phylogenetic Trees in Genome Workbench"

*Wednesday, June 03, 2015*

On June 10th, NCBI staff will show you how to use NCBI's powerful Genome Workbench application to work with phylogenetic trees. You will learn about the many options for importing trees, for visualizing, searching and modifying the trees and for saving or exporting them. We will also answer questions and welcome feedback from our participants on future directions for Genome Workbench.

Click [here](#) to sign up for this webinar. Like all our webinars, this will be posted on the NCBI YouTube account after the live presentation; you can subscribe to [our YouTube channel](#) to be notified of all our new videos.

To see upcoming webinars, as well as related materials and recordings from past webinars, please see the [NCBI Webinars page](#).

## Conserved Domain Database (CDD) version 3.14 now available online and via FTP

*Tuesday, June 02, 2015*

Conserved Domain Database (CDD) version 3.14 is now available with 560 new or updated NCBI-curated domains and 50,648 total domain models from CDD's database providers: Pfam, SMART, COG, TIGRFAMs, Protein Clusters, and the NCBI in-house curation project.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#). You can also learn about the Conserved Domain Database, how it works and is maintained, and its future in the most recent [Nucleic Acids Research database issue](#).

## The SRA Submission App on BaseSpace lets you submit directly to SRA

*Tuesday, June 02, 2015*

If you use Illumina for next-gen sequencing and want or need to share your genomic data by putting it into a public repository, you can now submit directly to [SRA](#) through [BaseSpace](#). You can also submit directly to SRA if you use [Mothur](#) for 16S assembly. Note: it is possible to port data from SRA into BaseSpace; click [here](#) for instructions.

## New NCBI YouTube video: "NCBI Minute: Prokaryotic Genome Annotation Update"

*Monday, June 01, 2015*

The [newest video](#) on the NCBI YouTube channel describes the updates recently made to our prokaryotic genome annotation process. In addition to describing the improvements to RefSeq bacterial and archaeal genome annotation and management, we also provide tips on adapting your workflow and show you how to find more information and help.

Subscribe to the [NCBI YouTube channel](#) to be notified of our new videos, which range from quick tips to full webinar presentations.

## NCBI News, May 2015

### New NCBI YouTube video: "Genome Workbench: Import BAMs and Export Alignments"

*Friday, May 29, 2015*

This [video](#) on the NCBI YouTube channel shows you how to import BAM files, create a BAM file index, and export selected alignments using NCBI's Genome Workbench.

Subscribe to the [NCBI YouTube channel](#) to be notified of our new videos, which range from quick tips to full webinar presentations.

### June 3rd webinar: "Troubleshooting GenBank Submissions: Coding Region Annotation"

*Thursday, May 28, 2015*

Next Wednesday, June 3rd, NCBI staff will show you how to troubleshoot internal stop codon errors encountered during coding region (CDS) annotation. The source of this problem can be in (1) improper frame/strand, or genetic code designation or (2) poor sequence quality. You will learn how to analyze your sequences and uncover problems with BLAST prior to submitting them to GenBank.

To sign up for this webinar, click [here](#). Like all of our webinars, this will be posted on the NCBI YouTube account after the live presentation; you can subscribe to our [YouTube channel](#) to be notified of all our new videos.

To see upcoming webinars, as well as related materials and [recordings](#) from past webinars, please see the [NCBI Webinars page](#).

### NCBI to hold three-day genomics hackathon in August

*Wednesday, May 27, 2015*

From August 3-5, NCBI will host its second genomics hackathon focusing on advancing bioinformatics analysis of next generation sequencing data. This event is for students, postdocs and investigators already engaged in the use of pipelines for genomic analyses from next generation sequencing data.\* Working groups of 5-6 individuals will be formed for twelve teams, in three sections. These groups will build pipelines to analyze large datasets within a cloud infrastructure. The sections for this iteration are: "RNA-Seq Normalization for Every Biologist", "Translational Genomics", and "Democratization of Genomics". Please see the application for specific team projects.

\* Specific projects are available to other developers or mathematicians.

### Organization

After a brief organizational session, teams will spend three days analyzing a challenging set of scientific problems related to a group of datasets. Participants will analyze and combine datasets in order to work on these problems. This course will take place on or near the NIH main campus in Bethesda, Maryland.

## Datasets

Datasets will come from the public repositories housed at NCBI. During the course, participants will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

## Products

All pipelines and other scripts, software and programs generated in this course will be added to a public GitHub repository designed for that purpose. A manuscript outlining the design of the hackathon and describing participant processes, products and scientific outcomes will be submitted to an appropriate journal. A pre-print of the manuscript from the January NCBI/ADDS hackathon is available from [bioRxiv](#).

## Application

To apply, complete this [form](#) (approximately 10-15 minutes to complete). Applications are due **June 6th by 3pm Eastern time**. Participants will be selected from a pool of applicants; prior students and applicants will be given priority in the event of a tie. Please note: applicants are judged based on the motivation and experience outlined in the form itself. Accepted applicants will be notified on June 18th, by 2pm Eastern time, and have until June 22 at 5pm Eastern time to confirm their participation. Please include a monitored email address, in case there are follow-up questions.

**Note:** Participants will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Applicants must be willing to commit to all three days of the event. No financial support for travel, lodging or meals can be provided for this event. Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

## New NCBI YouTube Video: NCBI's Tree Viewer

*Wednesday, May 27, 2015*

This short video, "[NCBI's Tree Viewer](#)" on the NCBI YouTube channel is an introduction to [Tree Viewer](#), a tool for viewing your own phylogenetic tree data. Tree Viewer is customizable and can be embedded in a wide variety of web pages.

Subscribe to the [NCBI YouTube](#) channel to be notified of our new videos, which range from quick tips to full webinar presentations.

## New NCBI Insights blog post: "NCBI's First Hackathon: Advanced Bioinformatic Analysis of Next-Gen Sequencing Data"

*Friday, May 22, 2015*

In the latest [blog post on NCBI Insights](#), we discuss the genomics hackathon NCBI hosted earlier this year, in conjunction with the [NIH Office of Data Science](#). The goal was to have experienced genomics professionals create efficient pipelines for people who are new to this field.

Visit [NCBI Insights](#), the official NCBI blog, for posts on what's new at NCBI, quick tips for using our tools and databases, and science feature stories.

## May 26th webinar: "The NCBI Minute: Prokaryotic Genome Annotation Update"

*Thursday, May 21, 2015*

The next NCBI Minute on Tuesday, May 26th will cover recent improvements to the way we annotate and manage RefSeq bacterial and archaeal genomes at NCBI. We'll introduce you to the new annotation paradigm, provide tips on adapting your workflow, and point out how to find help and more information.

To sign up for this brief webinar, navigate to: <https://attendee.gotowebinar.com/register/1585449954415535618>.

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool. To see upcoming webinars, as well as summaries, recordings on [YouTube](#), and related materials from past webinars, please see the [NCBI Webinars page](#).

## New NCBI Insights blog post: "NCBI RefSeq's Antimicrobial Peptide Indexed Field: Facilitating Novel Antibiotic Discovery"

*Thursday, May 21, 2015*

The latest [blog post on NCBI Insights](#) introduces the RefSeq "*Protein has antimicrobial activity [prop]*" indexed field, which retrieves curated sequence annotations showing naturally occurring antimicrobial peptides (AMPs), making it easier for researchers to identify alternatives to traditional antibiotics.

Visit [NCBI Insights](#), the official NCBI blog, for posts on what's new at NCBI, quick tips for using our tools and databases, and science feature stories.

## Export data into Genome Workbench with Tree Viewer version 1.4

*Tuesday, May 19, 2015*

NCBI Tree Viewer 1.4 implements several new features, improvements and bug fixes, including an updated Download function, which now allows you to export data into [Genome Workbench](#); you can also upload custom user-defined data in ASN.1 and Newick formats. To see the full list of updates, see the Tree Viewer [release notes](#).

NCBI Tree Viewer is a tool for viewing your phylogenetic tree data.

## June 9th hands-on workshops at NLM will show users how to search NCBI's molecular databases

*Monday, May 18, 2015*

On June 9th, 2015, NCBI will present two workshops on searching NCBI molecular databases: "Accessing Genomes, Assemblies and Annotation Products" and "Human Variation and Medical Genetics Resources". Workshops will be held at the Lister Hill Center (Building 38A) Auditorium on the [NIH campus](#).

**NOTE:** Participants must provide their own WiFi-ready laptop with a standard Web browser installed.

## Accessing Genomes, Assemblies and Annotation Products

9am-12pm

You will learn how NCBI processes genome-level data and produces annotation through the prokaryotic and eukaryotic genome annotation pipelines. You will find, browse, and download genome-level data for your organism of interest and for environmental and organismal metagenomes using the [Genome](#), [BioProject](#) and [Assembly](#) resources. In addition to assembled and annotated data, you will retrieve and download draft whole genome shotgun and next-generation sequencing data from the [Nucleotide](#) and [Sequence Read Archive \(SRA\)](#) databases. You will access results of precomputed analyses of genomes, as well as perform your own analyses of assembled and unassembled genomic data using NCBI's genome [BLAST](#) and [SRA-BLAST](#) services.

## Accessing NCBI Human Variation and Medical Genetics Resources

1pm-4pm

You will learn to use and access resources associated with human sequence variations and phenotypes associated with specific human genes and phenotypes. The workshop will emphasize the [Gene](#), [MedGen](#) and [ClinVar](#) resources to search by gene, phenotype and variant respectively. You will learn how to map variation from [dbSNP](#) and [dbVar](#) onto genes, transcripts, proteins and genomic regions and how to find genetic tests in [GTR](#). You will also gain experience using additional tools and viewers including [PheGenI](#), a browser for genotype associations, as well as the new [Variation Viewer](#) and the [1000 Genomes Browser](#). All of these provide useful ways to search for, map, and browse variants.

Register for one or both workshops at <https://www.surveymonkey.com/s/W2ZPW6D>.

If you have any questions, contact [courses@ncbi.nlm.nih.gov](mailto:courses@ncbi.nlm.nih.gov). To see more educational offerings from NCBI, please visit the [Learn](#) page on our website.

## Genome Workbench 2.9.0 now available

*Thursday, May 14, 2015*

As of May 5th, [Genome Workbench 2.9.0](#) is available. New features include custom selections and search support for Tree Viewer, as well as improvements to Graphical Sequence View. For the full list of fixes, improvements and features, see the [Genome Workbench release notes](#).

## New NCBI Insights blog post - Accessing the Hidden Kingdom: Fungal ITS Reference Sequences"

*Monday, May 11, 2015*

NCBI staff, in collaboration with outside mycology experts, are curating a set of fungal sequences from internal transcribed spacer (ITS) regions of nuclear rRNA genes. These ITS sequences are especially useful for identifying and classifying fungal species by morphology, a difficult process when using traditional methods.

Read more about this fungal [RefSeq Targeted Loci BioProject](#) on the NCBI blog, [NCBI Insights](#). To receive notice of new blog posts, you can sign up to the RSS feeds by clicking on the RSS links in the column on the right; you can also click the Follow tab that appears on the bottom of the screen when you visit NCBI Insights.

## RefSeq release 70 is now available with re-annotated bacterial genomes for uniformity across genomes and species

Thursday, May 07, 2015

The full RefSeq release 70 is now available online, on the [FTP site](#), and through NCBI's programming utilities, with 74,720,563 records describing 50,351,119 proteins, 11,310,700 RNAs, and sequences from 54,118 different organisms.

This release reflects a large update of complete bacterial RefSeq genomes, proteins and genes. In order to make genome annotation comparable across genomes and species, NCBI has re-annotated all RefSeq prokaryotic genomes using NCBI's genome annotation pipeline. Previously, it was possible that the same gene, in the same species, with an identical sequence for the gene's genomic region might be annotated with a different protein, simply because it was annotated using different methods. Now, the same gene in the same species with the same sequence will be annotated with exactly the same protein in RefSeq. If you'd like to learn more about the re-annotation project and what NCBI is doing to help you transition to using this new data, please see the [RefSeq Re-annotation Project page](#).

In addition, each annotated CDS used to be tracked with a distinct RefSeq protein accession number. However, due to identical protein sequences being found on multiple re-annotated RefSeq genomes and extensive bacterial genome sequencing, the RefSeq prokaryotic protein dataset rapidly became very redundant. Rather than flood the protein database with thousands of completely identical proteins, NCBI has adopted the use of non-redundant WP proteins for RefSeq prokaryotic genomes annotated with NCBI pipelines, which we first announced in June 2013.

Now, if the identical protein sequence appears on more than one RefSeq genome, NCBI simply reuses the existing WP accession number instead of creating a new accession for each new occurrence and genome. As a result, over 7 million proteins were removed, significantly reducing protein redundancy for the prokaryotic dataset. Removed accessions are reported in [release70.removed-records.gz](#) and a supplemental data mapping file is available in the release-catalog directory ([release70.bacterial-reannotation-report.txt.gz](#)).

Here are some measures for four species that illustrate the significant reduction in protein record redundancy resulting from the use of non-redundant RefSeq proteins (WP\_accessions).

### Counts of annotated proteins:

Species	Genomes	Total Proteins	Total Unique WPs	Total Singleton WPs
Staphylococcus aureus	4,194	11,764,898	222,588	138,284
Escherichia coli	2,685	13,637,370	1,033,617	649,100
Mycobacterium tuberculosis	1,790	7,245,836	139,800	101,255
Salmonella enterica	918	4,099,013	294,106	194,982

### Percent reduction in protein accessions:

Species	Genomes	Percent Reduction (WPs)	Percent Singleton WPs
Staphylococcus aureus	4,194	98%	62%
Escherichia coli	2,685	94%	63%
Mycobacterium tuberculosis	1,790	98%	72%
Salmonella enterica	918	93%	66%

## Singletons per Genome:

Species	Average Protein Count	Singleton WPs per Genome	Percent Singleton per Genome
Staphylococcus aureus	2,814	33	1.17%
Escherichia coli	5,088	241	4.74%
Mycobacterium tuberculosis	4,046	56	1.38%
Salmonella enterica	4,485	212	4.72%

### Definitions:

- "Total Proteins" counts the number of times non-redundant proteins accessions are annotated on the set of genomes for the species.
- "Total Unique WPs" counts the distinct number of non-redundant proteins used across all genomes. This is the truly non-redundant set of proteins for the species.
- "Total Singleton WPs" counts the number of non-redundant proteins used only once in the set of genomes for the species.
- "Percent Reduction" measures the compression in protein identifier space gained by using non-redundant protein accessions (WP\_ prefix).
- "Percent Singleton WPs" measures the percent of all non-redundant proteins for that species that are used only once in that species.

If you'd like to learn more about non-redundant proteins and see an example of this new RefSeq protein record, please see the [RefSeq non-redundant proteins page](#).

This is a first step toward managing data in a world where genomes are sequenced for assays, rather than to discover novel proteins. We appreciate that this is a new and major change for RefSeq prokaryotic genomes, but it is a necessary change to make as the number of disease-outbreak and other isolate sequencing continues to rapidly increase.

## Protein records

In all bacterial genomes, except reference genomes and a small number which have yet to be re-annotated, protein accessions NP/YP have been replaced with non-redundant protein accession numbers (WP\_).

- Over 7 million bacterial YP\_ and NP\_ RefSeq proteins were suppressed as complete bacterial genomes were re-annotated to conform to the new data model.
- Nearly 1 million non-redundant protein records were updated in March and April 2015 to improve protein names. These updates affected CDS "/product=" annotation details for all (>31,000) of the RefSeq bacterial genomes and included typographical corrections, name format standardization, and improved functional information.
- We have initiated a long-term project to validate and improve protein names for non-redundant protein records. In March and April, we validated names for approximately 2 million records using multiple support lines from Swiss-Prot, HMM analysis, domain architecture analysis, and NCBI staff curation.

## Nucleotide records

- Over 6,400 new or re-annotated RefSeq bacterial genomes were released.
- All new complete or draft RefSeq prokaryote genomes now use the accession format rule NZ\_<original\_INDSC\_accession>. Complete genomes that were already accessioned using the 'NC\_' prefix will continue to use that accession number. Thus, the accession prefix is no longer an indicator of a



complete bacterial genome. Information about genome completeness is provided in the record DEFINITION line, the Assembly resource, and FTP reports provided by Assembly and Genome resources.

## Impact to NCBI Gene

Together with this re-annotation effort, the scope of bacterial genomes included in Gene has been changed to include only genomes designated as a "reference genome," or "representative genome" where there is a cluster of related assemblies to indicate that the chosen representative assembly will be stable. Individual gene features on each assembly are identified with a locus\_tag that can be used as a unique identifier for the gene in publications, even if the assembly is out of scope for Gene.

## Using this data:

- a) Strain-specific protein datasets for individual RefSeq genomes can be obtained online, by FTP, and through NCBI's programming utilities. For more detailed instructions, please see the [Prokaryotic RefSeq FAQ](#).
- b) A graphical display of an annotated gene or protein can be accessed from the Nucleotide resource. Starting from a RefSeq genome record of interest, such as [NC\\_002695.1](#), follow the link to 'Graphics', and search for the locus\_tag or protein name of interest.
- c) Conversely, if starting from an individual non-redundant protein record, information about the annotated genomic location and genome taxonomy is available by following the link to the Identical Protein report. When a non-redundant protein record has been annotated on multiple RefSeq genomes, this report page lists the set of genomes that contain that identical protein, the genomic coordinates of the annotated CDS, and the specific organism information of the annotated genomic record. Thus, this report page can be used to identify the taxonomic range in which that identical protein has been found. The protein report can be downloaded in tabular format by using the 'Send to' link, and can be accessed using NCBI's programming utilities.

## Future plans

NCBI's future plans include:

- Organism classification and quality assurance: Work continues to identify misclassified genomes and contaminated genomes. Depending on the specific details of identified issues, additional RefSeq bacterial genomes may be suppressed or updated.
- Re-annotation of complete genomes: A small number of bacterial genomes have not yet been re-annotated at this time and will be in the near future. We also plan to re-annotate the archaeal RefSeq genomes in 2015.
- Protein names: We are working on providing improved names for the non-redundant (WP\_ accessioned) bacterial protein dataset. We are leveraging multiple sources of information, including curated UniProtKB/Swiss-Prot records, HMMs, Domain and domain architecture, publications and manual curation.
- Partial proteins: We are re-examining the prokaryotic genome annotation pipeline logic with regards to providing a non-redundant protein record for partial coding sequences.

## Documentation

NCBI has created documentation to explain these changes in detail:

- [RefSeq Re-annotation Project](#): An explanation of what the re-annotation project is, why and how it was done, and how we will facilitate your transition to the new annotation data.
- [RefSeq non-redundant proteins](#): A description of this new protein record type with examples.

- [Prokaryotic RefSeq Genomes](#): The prokaryotic RefSeq genomes policy, as well as definitions for reference genomes and representative genomes.
- [Prokaryotic annotation pipeline](#): An explanation of the prokaryotic genome annotation process at NCBI.
- [Prokaryotic RefSeq FAQ](#)
- [Supplemental data mapping file](#): An FTP file in the release-catalog directory ([release70.bacterial-reannotation-report.txt.gz](#)) has been prepared for re-annotated complete genomes that have recently transitioned to using the new non-redundant proteins. This file reports the old protein accession and GI, the annotated CDS coordinates, the old locus\_tag and NCBI GeneID values and maps that to the current non-redundant protein accession and GI, the new locus\_tag and NCBI GeneID (if available), the current CDS annotation coordinates, and indicates if the original protein identically matches or is similar to the replacement non-redundant protein.
- [Supplemental report of suppressed assemblies](#): An FTP file in the release-catalog directory ([release70.addedQA-SuppressedAssemblies.txt](#)) reports details for a subset of bacterial genomes that were suppressed in March 2015 following an expansion of QA metrics and curatorial review. This report illustrates some of the reasons for suppression.

If you have more questions or specific questions that are not addressed in the documentation, you can write to the Help Desk at [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov) or use the [feedback form](#) on the RefSeq page.

## May 13th webinar: "Introducing dbVar, the NCBI Database of Large-Scale Genetic Variation"

*Thursday, May 07, 2015*

Next Wednesday, May 13th, NCBI staff will introduce [dbVar](#), NCBI's database of genomic structural variation. In addition to describing the database's scope and features, we will also show you how to find, display and interpret dbVar records of interest.

To sign up for this webinar, click [here](#). Like all of our webinars, this will be posted on the NCBI YouTube account after the live presentation; you can subscribe to [our YouTube channel](#) to be notified of all our new videos.

To see a list of upcoming webinars, as well as [YouTube recordings](#) and related materials from past webinars, please see the [NCBI Webinars page](#).

## NCBI News, April 2015

### May 6th webinar: "The NCBI Minute: Connecting with PubMed Commons"

Monday, April 27, 2015

Next Wednesday, May 6th, NCBI staff will introduce [PubMed Commons](#), a forum for sharing information and perspectives about biomedical publications in PubMed. We will also provide a brief overview of how to participate and highlight the ways users are contributing to scientific discourse.

To sign up for this short webinar, click [here](#). Like all of our webinars, this will be posted on the [NCBI YouTube account](#) after the live presentation.

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool. To see upcoming webinars, as well as summaries, [YouTube recordings](#) and related materials from past webinars, please see the [NCBI Webinars page](#).

### New NCBI Insights blog post: "NIHMS Users: Do You Know How Often Your Paper is Being Accessed via PMC?"

Thursday, April 23, 2015

The latest blog post on [NCBI Insights](#) shows NIHMS users how to view PMC access statistics for any paper with which they are associated in the NIHMS system.

[About](#) ▾ [Help](#) ▾ [Manu](#)



#### Manuscript List for User Name

Needs Your Attention **1** In Process in NIHMS **0** Stalled **0** Completed **5** Available In PMC **14**

14 manuscripts available in PubMed Central

NIHMSID	Title	Status
123456	Title of Author Manuscript In PMC: Example 1 <a href="#">Acad Pediatr</a> March 1, 2014, Grants: XXX XXXXXXXX	Released to PMC (PMCID: PMCXXXXXX) PMC access statistics
112233	Title of Author Manuscript In PMC: Example 2 <a href="#">Appetite</a> February 1, 2014, Grants: XXX XXXXXXXX	Released to PMC (PMCID: PMCXXXXXX) PMC access statistics
654321	Title of Author Manuscript In PMC: Example 3 <a href="#">Mind Brain Educ</a> December 1, 2013, Grants: XXX XXXXXXXX	Released to PMC (PMCID: PMCXXXXXX) PMC access statistics

Visit [NCBI Insights](#), the official NCBI blog, for stories about what's new at NCBI, quick tips for using our tools, and science features.

## **April 29th webinar: "The NCBI Minute: Finding Genomes and Annotations by Searching NCBI Assembly"**

*Wednesday, April 22, 2015*

Next Wednesday, April 29th, NCBI staff will show you how to quickly find a particular genome sequence by using FTP to access our [Assembly](#) database, which houses 25,000 annotated genomes.

To sign up for this short webinar, click [here](#). Like all of our webinars, this will be posted on the [NCBI YouTube account](#) after the live presentation.

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool. To see upcoming webinars, as well as summaries, [YouTube recordings](#) and related materials from past webinars, please see the [NCBI Webinars page](#).

## **April 21st webinar: Rebroadcast of "NCBI and the NIH Public Access Policy: PubMed Central Submissions, My NCBI, My Bibliography and SciENcv"**

*Tuesday, April 14, 2015*

On Tuesday, April 21, NCBI will have an encore presentation of the March 5th webinar for NIH grant holders on My NCBI, My Bibliography and SciENcv. This webinar will include the same material as the March 5th webinar; the recording of the original presentation is available on [YouTube](#).

To register for the April 21st webinar, go [here](#).

To see upcoming webinars, as well as summaries, [YouTube recordings](#) and related materials from past webinars, please see the [NCBI Webinars page](#).

## **April 15th webinar: "The NCBI Minute: Finding and Getting the Data You Want from NCBI in Less than Three Minutes - Introducing BioProject"**

*Friday, April 10, 2015*

Next Wednesday, April 15th, NCBI staff will show you how to use the [BioProject database](#) to quickly find data. When looking for research data using fairly broad search terms like "tuberculosis" or "mouse", BioProject is a great place to start. In this brief presentation, you will learn how to search for studies and quickly identify related PubMed-listed publications and experimental data, such as RNA-seq datasets in the GEO database.

To sign up for this webinar, go [here](#).

The NCBI Minute is a series of short webinars that give a brief introduction to a specific topic or NCBI tool. To see upcoming webinars, as well as summaries, recordings (via [YouTube](#)) and related materials from past webinars, please see the [NCBI Webinars page](#).

## **NIH issued statement on use of dbGaP in the Cloud**

*Thursday, April 02, 2015*

On Monday, the National Institutes of Health announced that it is now allowing investigators to request permission to transfer controlled-access genomic and associated phenotypic data obtained from NIH-designated data repositories, like dbGaP, under the auspices of the [NIH Genomic Data Sharing \(GDS\) policy](#) to public or private cloud systems for data storage and analysis.

Please keep in mind that the responsibility for the security of the dbGaP data is assumed by each investigator and their associated institution who has been approved to access the data, not the cloud provider. To assist in this process, NIH has provided as much information as possible for PIs, institutional signing officials and the IT staff who will be supporting these projects.

The post "[The Cloud, dbGaP and the NIH](#)" on the [NIH Data Science blog](#) discusses the NIH position statement, the Genomic Data Sharing policy, and [best practices](#), as well as NIH's IT security requirements and policies.



## NCBI News, March 2015

### Updated human and mouse genome annotations now available

Tuesday, March 31, 2015

Updated annotations for the [human](#) and [mouse](#) RefSeq genomes produced by the [Eukaryotic Genome Annotation Pipeline](#) are now available. New known RefSeq transcripts (NM\_ and NR\_ accessions) and non-transcribed pseudogenes (NG\_ accessions) were used for these annotations. The number of model RefSeq predictions (XM\_ and XR\_ accessions) also increased through the use of additional RNA-Seq datasets, especially for human where model RefSeq annotated on GRCh38.p2 contain 41% more exonic bases (31 MBp) than the known RefSeq.

#### **Homo sapiens** annotation release 107: [see in Gene](#), [BLAST](#) or [download](#).

- **Assemblies annotated:** [GRCh38.p2](#) (GCF\_000001405.28, reference) and [CHM1\\_1.1](#) (GCF\_000306695.2); note that we removed the [HuRef assembly](#), GCF\_000002125.1, from the RefSeq collection.
- **RNA-Seq datasets used:** The Human Protein Atlas ([PRJEB4337](#)) and BodyMap2 ([PRJEB2445](#))
- **Annotation changes for GRCh38.p2:**
  - 50% more genes with alternative splice variants (an average of 3.52 transcripts per gene)
  - 100% more non-coding genes, 146% more non-coding transcripts
  - 8% more annotated known RefSeq

#### **Mus musculus** annotation release 105: [See in Gene](#), [BLAST](#) or [download](#).

- **Assemblies annotated:** [GRCm38.p3](#) (GCF\_000001635.23, reference) and [Mm\\_Celera](#) (GCF\_000002165.2)
- **RNA-Seq datasets used:** mouse ENCODE transcriptome ([PRJNA66167](#)) and a whole-embryo project ([PRJNA203332](#))
- **Annotation changes for GRCm38.p3:**
  - 3.4% more genes with alternative splice variants (an average of 2.87 transcripts per gene)
  - 16% more non-coding genes, 27% more non-coding transcripts
  - 5.8% more annotated known RefSeq

You can find the annotation runs currently in progress on the [Eukaryotic Genome Annotation Pipeline status page](#).

### April 8th webinar: "The NCBI Minute: Introducing MOLE-BLAST"

Wednesday, March 25, 2015

On April 8th, NCBI will present a five-minute webinar introducing [MOLE-BLAST](#), a tool for clustering targeted sequences, like those from 16s rRNA, with database sequences and providing taxonomic context. MOLE-BLAST can quickly establish taxonomy for sequences from uncultured or environmental sequences. To register, click [here](#).

The NCBI Minute is a series of short webinars that give a brief introduction to an NCBI tool or service, as well as quick tips on using our resources. To see upcoming webinars, as well as summaries, recordings (via [YouTube](#)) and related materials from past webinars, please see the [NCBI Webinars page](#).

## April 1st webinar: "A Practical Guide to Using NCBI BLAST on the Web"

*Tuesday, March 24, 2015*

Next Wednesday, April 1st, NCBI will present a webinar on the NCBI BLAST service. The webinar will highlight important features and demonstrate the practical aspects of using NCBI BLAST, the most popular sequence similarity service in the world. To register, click [here](#).

Some of the useful features that will be discussed include:

- Access from the Entrez sequence databases
- The new genome BLAST service quick finder
- The integration and expansion of Align-2-Sequences
- Organism limits and other filters
- Reorganized databases
- Formatting and downloading options
- TreeView displays

We will also show you how to use other important sequence analysis services associated with BLAST including Primer-BLAST, iGBLAST, and MOLE-BLAST, a new tool for clustering and providing taxonomic content for targeted loci sequences (16S, ITS, 28S). These aspects of BLAST provide easier access and results that are more comprehensive and easier to interpret.

To see upcoming webinars, as well as summaries, recordings (via [YouTube](#)) and related materials from past webinars, please see the [NCBI Webinars page](#).

Update: We will also have a short webinar on MOLE-BLAST on April 8th. Click [here](#) to learn more about it and sign up.

## dbSNP Build 143 Phase II now available

*Tuesday, March 17, 2015*

dbSNP build 143 phase II includes data for cow, *Ciona intestinalis* and prairie vole. Build 143 provides more than 537 million submitted and 299 million reference variants for 9 species. You can access build 143 SNP data through the integrated NCBI Entrez system and through [FTP](#). To see complete build statistics, visit the [SNP summary page](#).

## New NCBI Insights blog post: "Exploring Entrez Direct: Parsing the XML Output of E-utilities"

*Friday, March 13, 2015*

The latest [blog post](#) on [NCBI Insights](#) shows you how to use Entrez Direct's ability to parse and reformat complex XML data returns from EFetch, using PubMed records as an example.

## NCBI homepage update includes action buttons, category pages

*Thursday, March 12, 2015*

The [NCBI homepage](#) now has six new buttons on it: Submit, Download, Learn, Develop, Analyze, and Research. Each of these buttons leads to an action page devoted to a particular set of services.



These action pages will allow you to easily access the pages and resources you need to complete tasks. For instance, you can:

- find information about the Entrez API,
- find an upcoming NCBI webinar,
- find an NCBI tool that designs PCR primers,

and much more.

We've also included a blue Feedback button on the left side of the [Download](#), [Learn](#), [Develop](#) and [Analyze](#) pages so that you can tell us what you think. We look forward to hearing your comments.

On the new action pages, you'll also see 6 categories in the header: [Literature](#), [Health](#), [Genomes](#), [Genes](#), [Proteins](#), and [Chemicals](#). These category pages highlight useful databases, tools and resources for each of the topics all in one place. If you follow us on [LinkedIn](#), these categories will be familiar to you - we've used them as Showcase Pages to group our news stories and announcements by topic.

Stay tuned to [NCBI News](#) and to our blog, [NCBI Insights](#), for more information about the new homepage.

## NCBI Sequence Viewer version 3.6 available

*Wednesday, March 11, 2015*

NCBI Sequence Viewer has recently been updated and now has improved rendering of SNP insertions/deletions and narrow features, as well as better graph track names. A full list of new features, improvements and fixes is included in the [release notes](#).

Sequence Viewer is a graphical view of sequence sand color-coded annotations on regions of sequences stored in the [Nucleotide](#) and [Protein](#) databases.

## March 18th webinar: "Using the dbGaP Data Browser to browse aligned reads and genotypes from the Database of Genotypes and Phenotypes"

*Tuesday, March 03, 2015*

In two weeks, NCBI will present a webinar on the dbGaP Data Browser. This webinar will show you how to use the Data Browser to access aligned reads and genotypes, using the last exon of the APOE gene from an Alzheimer's disease study as an example. To register, click [here](#).

The dbGaP Data Browser provides access to aligned reads and genotypes from a variety of sequencing studies from the [Database of Genotypes and Phenotypes \(dbGaP\)](#). The browser shows sample-level alignments - in the context of the genome sequence - with variants from dbSNP and known clinical variants such as those from [ClinVar](#), as well as differences from the reference genome sequence. The browser allows you to filter the subjects by a variety of indexable values and, depending on your level of access, view-only or downloadable access to reads and genotypes.

To see upcoming webinars, as well as summaries, recordings via [YouTube](#), and related materials from past webinars, please see the [NCBI Webinars page](#).

The image shows the NCBI homepage with a red rectangular box highlighting a central section of action buttons. The buttons are arranged in two rows of three. The top row contains 'Submit', 'Download', and 'Learn'. The bottom row contains 'Develop', 'Analyze', and 'Research'. Each button includes a descriptive sentence and a representative icon. The 'Submit' button features an upward-pointing arrow, 'Download' has a downward-pointing arrow, 'Learn' shows a stack of books, 'Develop' displays a grid of squares, 'Analyze' shows a network diagram, and 'Research' depicts a microscope.

**Submit**  
Deposit data or manuscripts into NCBI databases

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class or watch a tutorial

**Develop**  
Use NCBI APIs and code libraries to build applications

**Analyze**  
Identify an NCBI tool for your data analysis task

**Research**  
Explore NCBI research and collaborative projects

Other visible elements on the page include a top navigation bar with 'NCBI Resources' and 'How To' menus, a search bar, a left sidebar with a 'Resource List (A-Z)' menu, and a right sidebar with 'Popular Resources' and 'NCBI Announcements' sections. The footer contains a 'You are here' breadcrumb, a 'Write to the Help Desk' link, and several categorized lists of links.

Figure 1. The NCBI homepage. The new action buttons are outlined in red.

**U.S. National Library of Medicine** | **NCBI National Center for Biotechnology Information**

NCBI HOME | **LITERATURE** | HEALTH | GENOMES | GENES | PROTEINS | CHEMICALS | POPULAR RESOURCES

All Databases | Search NCBI | Search

## Learn

NCBI provides the user community with a variety of educational resources including courses, workshops, webinars, training materials and documentation.

**UPCOMING EVENTS**

- The Statistics of Local Pairwise Sequence Alignment, Part I  
JANUARY 22, 2015  
Online
- The Statistics of Local Pairwise Sequence Alignment, Part II  
JANUARY 29, 2015  
Online
- A Librarian's Guide to NCBI  
MARCH 9-13, 2015  
Bethesda, MD
- American Society for Microbiology, 2015  
MAY 31-JUNE 2, 2015  
New Orleans, LA

**Webinars & Courses**

In-person courses, live webinars and webinar recordings

**Exhibits & Presentations**

Booth exhibits and workshops at scientific conferences

**Tutorials**

Tutorials: Training materials in HTML, PDF and video formats

**Documentation**

Online manuals, handbooks, fact sheets and FAQs

**News, Blog & Social Media**

Keep up with the latest NCBI news and follow NCBI on social media sites, including FaceBook, Twitter, Google+, LinkedIn and the NCBI Insights blog.

<b>NCBI</b> About NCBI Submit Download Learn Develop Analyze	<b>Literature</b> PubMed PMC Books NLM Catalog	<b>Genomes</b> Genome Nucleotide SRA Assembly dbSNP dbVar	<b>Genes</b> Gene Nucleotide GenBank RefSeq TPA GEO	<b>Proteins</b> Protein RefSeq TPA HomoloGene CDD Protein Clusters	<b>Chemicals</b> PubChem BioAssay Substance Compound BioSystems
--	--	---	---	--	--

**Figure 2.** The [Learn](#) page. The six category pages are linked at the top, in the header. On the left side of the page, an arrow points to the feedback tab, which you can use to comment.



## NCBI News, February 2015

### March 5th webinar: "NCBI and the NIH Public Access Policy: PubMed Central submissions, My NCBI, My Bibliography and SciENcv"

*Wednesday, February 25, 2015*

Next Thursday, March 5th, NCBI will host a webinar outlining how to use My NCBI to report public access policy compliance for NIH grant holders. Topics will include the NIH Public Access Policy, NIHMS and PubMed Central submissions, creating My NCBI accounts, use of My Bibliography to report compliance to eRA Commons and using SciENcv to create BioSketches.

To register for this webinar, go [here](#).

Please see the [NCBI Webinars](#) page for a list of upcoming webinars as well as recordings (on [YouTube](#)) and related materials from past webinars.

### "A Submitter's Guide to GenBank" webinar parts 1 and 2 on YouTube

*Friday, February 20, 2015*

If you missed the recent "A Submitter's Guide to GenBank" webinar series, [Parts 1](#) and [2](#) have been uploaded to our [official YouTube account](#). In addition, we have prepared a PDF of the question and answer sessions conducted after each presentation, available via [FTP](#).

## A Submitter's Guide to GenBank, Part 1

Using BankIt for Small-Scale Nucleotide Sequence Submissions



These webinars outline the process of using [BankIt](#), a web-based submission tool at NCBI, to submit sequence data to the [GenBank database](#). The first part is a demonstration on using BankIt forms to complete a submission

of a single or a few nucleotide sequences, while the second part shows you how to use BankIt file inputs to complete a submission of nucleotide sequences that require multiple features for each sequence.

The [NCBI YouTube channel](#) provides presentations and tutorials about our biomolecular and biomedical literature databases and tools. See the [Webinars playlist](#) to watch presentations you may have missed or to rewatch your favorite video. A list of past and upcoming webinars, as well as related materials, is available on the [NCBI Webinars page](#).

## **NCBI Insights blog: How to delegate authority to others to edit/create your profile and Collections**

*Thursday, February 19, 2015*

The [latest blog post](#) on the NCBI Insights blog shows you how to send a delegate invitation that will allow a colleague to view and edit your My Bibliography collection as well as view, edit and create profiles in your SciENcv.

NCBI Insights offers posts that cover new developments and events at NCBI, as well as tips on using our resources and tools. We encourage you to join the conversation at NCBI Insights.

## **NCBI webinar on February 25: The Next Generation of Access to Sequencing Data: Using NCBI's SRA Toolkit to Access Data from dbGaP and SRA**

*Wednesday, February 18, 2015*

Next Wednesday, February 25th, NCBI staff will present a webinar on the SRA Toolkit, a system for accessing the approximately 3.4 Petabytes of next-generation genomic and expressed sequence data housed in the NCBI Sequence Read Archive (SRA).

As data sets grow larger, mining information and performing comparisons directly from structured databases becomes increasingly necessary. The SRA Toolkit is not only capable of dumping data out as fastq or sam files, but also provides direct analysis and comparison from specific genomics regions across hundreds or thousands of samples.

In the webinar, we will show examples of configuration and use of the Toolkit for both public SRA and controlled access data associated with studies in the Database of Genotypes and Phenotypes (dbGaP).

To register for this webinar, please go [here](#).

## **NCBI Genomes FTP site update adds analysis sets and other data**

*Wednesday, February 18, 2015*

Several improvements have recently been implemented in three sections of the NCBI Genomes FTP site: [GenBank](#) and [RefSeq](#) (both browsable), and the "all" genomes FTP directory (not browsable; however, it can be used for scripted downloads based on assembly directory name).

A range of new content is available for download on the FTP site:

- Analysis sets for human GRCh38 and mouse GRCm38.p3 are in the GenBank assembly directories. These sets contain FASTA and GFF files with modified sequence identifiers and index files, which make these

data convenient for analysis with next generation sequencing tools. Please refer to the provided documentation ([human](#) | [mouse](#)) for a complete description.

- A text file, `assembly_summary.txt`, has been added to each species directory. This file is a species-specific subset of the comprehensive assembly summary files provided in the "[Assembly Reports](#)" folder. The file content includes information on release dates, submitter and assembly names, assembly accession version, assembly status, RefSeq category, full FTP path (see below) and associated meta-data, including BioProject and BioSample identifiers. Example: *Saccharomyces cerevisiae*
- The full FTP path has been added as the last column in `assembly_summary_refseq.txt` and `assembly_summary_genbank.txt` files provided in the "[Assembly Reports](#)" folder.
- A small number of assemblies that have a large number of contigs were omitted from the first release of the new FTP site. These assemblies are now available and include *Triticum aestivum* (see [Assembly GCA\\_000334095.1](#)) and *Locusta migratoria* (see [Assembly GCA\\_000516895.1](#)).

In addition, files for all "latest" assemblies were regenerated to make the following changes:

- Removal of erroneously reported CDD features in RNA flat files
- Inclusion of missing strand information for some features on the forward strand and added plus signs ("+") in column 7 in updated GFF3 files
- Correct representation of multi-interval non-trans-spliced tRNA features on GFF3 files. Each multi-interval non-trans-spliced tRNA feature is now represented by a single feature (line) of type tRNA and multiple nested features of type exon (one for each interval).

NCBI staff continues to work on fully replacing the original `/genbank/genomes/` and `/genomes/` FTP content. As previously announced, we plan to remove content from the older FTP directories by the end of March. We will not remove content from the historical areas until it is available in the new areas. Note that some content may be available in a different file name or format or sub-directory in the newer FTP directories.

Please refer to the FTP README.txt files and the [NCBI Genomes FTP FAQs](#) to learn more.

## GenBank release 206.0 is now available via FTP

*Tuesday, February 17, 2015*

[Release 206.0](#) (2/13/2015) has 181,336,445 non-WGS, non-CON records containing 187,893,826,750 base pairs of sequence data. In addition, there are 205,465,046 WGS records containing 873,281,414,087 base pairs of sequence data, as well as 66,706,014 TSA records containing 49,765,340,047 base pairs of sequence data.

During the 63 days between the close dates for GenBank releases 205.0 and 206.0, the non-WGS, non-CON portion of GenBank grew by 2,955,763,136 base pairs and by 2,040,676 sequence records. During that same period, 164,936 records were updated; an average of 35,010 non-WGS, non-CON records were added and/or updated per day. Between releases 205.0 and 206.0, the WGS component of GenBank grew by 24,303,492,065 base pairs and by 5,163,496 sequence records. The TSA component of GenBank also grew; 3,708,919,144 base pairs and 4,070,397 sequence records were added.

The total number of sequence data files increased by 46 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 169
- CON: 11 new files, now a total of 303
- ENV: 2 new files, now a total of 80
- GSS: 4 new files, now a total of 293
- INV: 1 new file, now a total of 133
- PAT: 3 new files, now a total of 217
- PLN: 1 new file, now a total of 96

- TSA: 13 new files, now a total of 172
- VRL: 1 new file, now a total of 45

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 700 GB (sequence files only), and the ASN.1 data require approximately 572 GB.

More information about GenBank release 206.0, including important changes in this release and upcoming changes, is available in the [release notes](#).

## Mouse, cow and zebrafish added to dbSNP build 142

*Thursday, February 12, 2015*

Three organisms are now available in dbSNP build 142: mouse, cow and zebrafish. This data is indexed in [Entrez](#) and is available by FTP.

New mouse (*mus musculus*) information on [FTP](#) and [Entrez](#):

- Assembly: GRCm38.p2 (GCF\_000001635.22)
- New RS: 9323191
- Total RS: 80429085

New cow (*bos taurus*) information on [FTP](#) and [Entrez](#):

- Assembly: Bos\_taurus\_UMD\_3.1 (GCF\_000003055.4)
- New RS: 11509794
- Total RS: 85027819

New zebrafish (*danio rerio*) information on [FTP](#) and [Entrez](#):

- Assembly: Zv9 (GCF\_000002035.4)
- New RS: 16326757
- Total RS: 17765748

## 1000 Genomes Browser updated to include Phase 3 May 2013 call set

*Tuesday, February 10, 2015*

[1000 Genomes Browser](#) version 3.4 is now available. This update includes variant and genotype calls from the Phase 3 May 2013 call set. For a full list of browser features, see the [Release Notes](#). A detailed [browser user guide](#) is also available.

The browser will continue to provide access to data from the Phase 1 March 2012 call set.



## NCBI News, January 2015

### NIHMS's new look streamlines the manuscript submission process

Thursday, January 29, 2015

Today, the [NIH Manuscript Submission \(NIHMS\) system](#) gets a new interface design, as well as updates that streamline the login and manuscript submission processes and provide relevant help information directly on each screen.

#### Homepage

The NIHMS sign-in routes will now be available from the homepage. Select a route based on your funding agency (1) or sign in through NCBI if you are starting a deposit on an author's behalf(2).

The homepage also includes a graphic overview of the NIHMS process (3). You can hover over each step for more information or click "Learn More" to read the complete overview in the FAQ.

*Note:* The steps of the NIHMS conversion process will remain the same. An author or PI (i.e., Reviewer) will still need to complete the Initial Approval and Final Approval steps. Updated [help documentation](#) and [FAQs](#) will help you navigate the process.

#### Managing Manuscripts

Once you are signed into NIHMS, you will be directed to your Manuscript List. From this page, you can manage and track your existing submissions (1), submit a new manuscript (2), and search for a record (3). You can also click on any headings in the information box (4) to expand a topic and read the help text.

#### Deposit a Manuscript

The initial deposit still requires you to enter a manuscript and journal title, deposit complete manuscript files, and specify funding information and the embargo.

Key updates include:

- assigning an NIHMSID to a record only *after files have been uploaded*, i.e., at the Check Files step (1);
- a streamlined deposit process with clearly defined and explained actions in each step (2);
- requiring the Submitter to *open the PDF Receipt* to review the uploaded files and confirm that the submission is complete before advancing to the next step (3);
- relevant help information available on each page, as in the previous example (4); and
- requiring the Reviewer to *add funding* before approving the initial deposit (not pictured).

**Questions?** Contact [nihms-helpdesk@ncbi.nlm.nih.gov](mailto:nihms-helpdesk@ncbi.nlm.nih.gov).

### Genome Workbench 2.8.10 available

Monday, January 26, 2015

[Genome Workbench 2.8.10](#) is available, as of January 16th. New features include added support for Ubuntu 14.04 and automatic project save. For the full list of fixes, improvements and features, see the [Genome Workbench release notes](#).

U.S. Department of Health & Human Services

About ▾ Help ▾

# NIH Manuscript Submission System


## Sign In

**1** NIH Researchers login commons

HHMI Researchers hhmi

**2** Publishers and Others NCBI


[Sign-In Help](#)  
[Forgot your sign-in route? Request E-mail Reminder](#)



The NIH Manuscript Submission (NIHMS) system supports the deposit of manuscripts into PubMed Central (PMC), as required by the public access policies of NIH and other participating funders.

[Learn More](#)

## **3** Manuscript Submission Process



The process consists of six steps: 1. Deposit Files (Submitter), 2. Initial Approval (Reviewer), 3. NIHMS Conversion (Gears), 4. Final Approval (Reviewer), 5. PMCID Assigned (1473589), and 6. Available in PMC (User with tablet). A 'READY FOR APPROVAL' envelope icon is shown between steps 3 and 4.

[Learn More](#)

Figure 1. The new NIH Manuscript Submission system homepage.

## Manuscript list for Kathryn Funk

**1** Needs Your Attention **2** In Process in NIHMS **64** Stalled **3** Completed **18** Available in PMC **3**

**2** [Submit New Manuscript](#)

**3** Find a manuscript

NIHMSID:

**4** [How do I submit a manuscript?](#)  
[What does the Status of a manuscript record mean?](#)  
[How can I search for manuscripts already in the NIHMS system?](#)  
[Why don't I see my manuscript?](#)

NIHMSID	Title	Status
15656	test deryd 3335	<span style="color: red;">Stalled</span> Awaiting reviewer's initial approval
14015	The discourse of physical power and biological knowledge in the 1930s: a reappraisal of the Rockefeller Foundation's 'policy' in molecular biology.	Awaiting submitter's initial approval or designation of reviewer

Figure 2. The Manuscript List.

Test Submission for NIHMS Announcement  
Journal: Journal Title NIHMSID 16049 [Provide citation](#) **1**

1. Title Information 2. Add Funding 3. Upload Files 4. Check Files 5. Set Reviewer & Embargo **2**

This PDF Receipt shows the content of all text, figure, and table files, in addition to placeholders for any supplemental files that you uploaded in the previous step.

You must review the PDF receipt file to advance to the next step. **3**  
PDF Receipt [2014-12-09 13:56:51, 84.6 KB]

Please return to Upload Files if any files are missing.

[Save & Exit](#) [Cancel Submission](#) [← Upload Files](#) [Set Reviewer & Embargo ►](#)

**4** What should I look for in the PDF Receipt?  
Please open and review the PDF Receipt to confirm that you have provided all the materials that make up your manuscript and that are referenced in the text, including any placeholders for supplemental files (if applicable).

What if the figures appear corrupt or damaged in the PDF Receipt?  
What if the PDF Receipt has not generated properly?

Figure 3. A sample submission.

## Conserved Domain Database (CDD) version 3.13 now available online and via FTP

Monday, January 26, 2015

Conserved Domain Database (CDD) version 3.13 is now available with 286 new or updated NCBI-curated domains and 50,415 total domain models from CDD's database providers: Pfam, SMART, COG, TIGRFAMs, Protein Clusters, and the NCBI in-house curation project.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#). You can also learn about the Conserved Domain Database, how it works and is maintained, as well as future plans for the database in [this paper](#) from the most recent Nucleic Acids Research database issue.

## NCBI support for SOAP E-Utility ends July 1, 2015

Thursday, January 22, 2015

On July 1, 2015, the NCBI E-Utility SOAP web service, along with the SOAP web service for BLAST, will no longer be supported by NCBI. Any requests to these services after that date will not function.

Please consider using the [URL interface](#) to the E-Utilities to retrieve NCBI data or the [BLAST REST API](#) to submit BLAST searches.

## GenBank surpasses one trillion total bases of publicly available sequence data

Thursday, January 22, 2015

Last October, GenBank (Release 204) exceeded an astounding 1 Terabase of assembled sequence data.

GenBank is a comprehensive database that contains publicly available nucleotide sequences for over 300,000 formally described species. To learn how NCBI builds GenBank and ensures its uniformity and comprehensiveness, see this [recently published paper](#).

## **Nucleic Acids Research Database 2015 Issue illustrates NCBI databases, updates and future plans**

*Wednesday, January 21, 2015*

The 22nd annual edition of the [Nucleic Acids Research Database Issue](#) features nine papers from NCBI staff that present recent updates to our databases, including [GenBank](#), [Gene](#), and [RefSeq](#).

These papers describe the state of NCBI databases as well as future plans to improve their use, from new reference resources created to improve the usability of [viral sequence data](#) to in-house curation efforts in the [Conserved Domain Database](#), and much more.

The NCBI database articles in NAR are also available from [PubMed](#). To read an article, click on the PMID listed below:

- "Database Resources of the National Center for Biotechnology Information" by NCBI Resource Coordinators. (PMID: [25398906](#))
- "GenBank" by Dennis A. Benson et al. (PMID: [25414350](#))
- "Gene: a gene-centered information resource at NCBI" by Garth R. Brown et al. (PMID: [25355515](#))
- "CDD: NCBI's conserved domain database" by Aron Marchler-Bauer et al. (PMID: [25414356](#))
- "Expanded microbial genome coverage and improved protein family annotation in the COG database" by Michael Y. Galperin, Kira S. Makarova, Yuri I. Wolf and Eugene V. Koonin. (PMID: [25428365](#))
- "HIV-1, human interaction database: current status and new features" by Danso Ako-Adjei et al. (PMID: [25378338](#))
- "NCBI Viral Genomes Resource" by J. Rodney Brister, Danso Ako-Adjei, Yiming Bao and Olga Blinkova. (PMID: [25428358](#))
- "Update on RefSeq microbial genomes resources" by Tatiana Tatusova et al. (PMID: [25510495](#))
- "Type material in the NCBI Taxonomy Database" by Scott Federhen. (PMID: [25398905](#))

## **NCBI YouTube channel: A million views and counting!**

*Friday, January 16, 2015*

As of December 31, 2014, we have passed the 1 million mark for lifetime views on our [YouTube channel](#)! The NCBI YouTube channel provides presentations and tutorials about our biomolecular and biomedical literature databases and tools.

Subscribe to our [YouTube channel](#) and stay up to date on all the [tutorials](#) and [webinars](#) we offer.

## **NCBI's next webinar is The Statistics of Local Pairwise Sequence Alignment, Parts 1 and 2**

*Tuesday, January 13, 2015*

On Thursday, January 22nd, Stephen Altschul of NCBI will present the first part of a discussion of the statistical theory for local sequence alignments like those produced by the [BLAST](#) database search programs. It will cover the statistical parameters for local alignment scoring systems, and the formulas for calculating bit scores and asymptotic E-values and p-values from raw alignments scores.



Figure 1. Total number of bases from April to December 2014.

This presentation will continue the following Thursday, January 29th. Part 2 will be a discussion of the considerations that go into the construction and selection of amino acid and nucleic acid scoring systems for pairwise local sequence alignment. It will briefly cover the PAM and BLOSUM series of amino acid substitution matrices, and also the concepts of relative entropy and efficiency for substitution matrices.

To sign up, click here: [Part 1](#) and [Part 2](#).

## E-Utilities users: Keep up to date with changes via the Gene database RSS feed

*Monday, January 12, 2015*

If you use E-Utilities/ESummary with the Gene database and have not subscribed to the Gene News RSS feed, you probably missed an important announcement about a few impending changes:

**A new element in the DocSum** has now been added for Gene records. The new Organism element consolidates taxonomic information from the previously used items: Scientific name, common name and TaxID.

**The new XML format for ESummary** is simpler and more compact, using each field name as the XML tag. For example, <TaxID>9606</TaxID>.

**To keep up to date with changes** affecting the Gene database, please consider signing up for the Gene Announce RSS feed, which can be found on [this page](#) listing all of NCBI's RSS feeds and email listservs.

## RefSeq release 69 available on FTP

*Wednesday, January 07, 2015*

The full RefSeq release 69 is now available on the [FTP site](#) with 74 million records describing 52,276,468 proteins, 9,973,568 RNAs, and sequences from 51,661 organisms.

More details about the RefSeq release 69 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the accessions included in the [release](#) and the [files installed](#) are available.

## NCBI annotates 200th eukaryote

*Tuesday, January 06, 2015*

The NCBI Eukaryotic Genome Annotation Pipeline has passed a new landmark: the completion of the annotation of 200 different organisms including 76 mammals, 51 birds, 26 other vertebrates, 21 invertebrates and 26 plants. Over half of these were annotated with the help of RNA-Seq evidence available in the [Sequence Read Archive](#). The lucky 200th organism is a fish, the [large yellow croaker](#) (*Larimichthys crocea*). See the full list of annotated organisms [here](#), and request the annotation of your favorite!

Data produced by the Eukaryotic Genome Annotation Pipeline is available in the Reference Sequences (RefSeq) collection, BLAST non-redundant and organism-specific databases, Gene database, and on the NCBI FTP site.

## NCBI staff will attend International Plant and Animal Genome Conference XXIII

*Monday, January 05, 2015*

Next week, NCBI staff will present posters and lead a workshop at the [International Plant and Animal Genome Conference](#). In addition, NCBI will have a booth (Booth 618). Staff will be at the booth to answer any questions you may have.

To see a full schedule of NCBI's activities at PAG XXIII, including our annual Genome Resources workshop, click [here](#) or visit us at Booth 618.

## NCBI News, December 2014

### NCBI webinar A Submitter's Guide to GenBank, Part 2 on January 7th

*Wednesday, December 31, 2014*

On January 7th, NCBI will present the continuation of the December 17th webinar on using BankIt for GenBank submissions. Part 2 will cover how to use BankIt file inputs to complete a submission of a single or a few nucleotide sequences that require multiple features for each sequence. We will also describe how to create and use Feature Table files to add information about sequence data.

This webinar will stay at a basic level for sequence submissions, but future webinars that illustrate more complex sequence submissions will be considered depending on the feedback received from this presentation.

To register, click [here](#). To see materials and videos from previous webinars, as well as descriptions of upcoming webinars, see the [NCBI Webinars](#) page.

### GenBank release 205.0 is now available via FTP

*Tuesday, December 16, 2014*

Release 205.0 (12/12/2014) has 179,295,769 non-WGS, non-CON records containing 184,938,063,614 base pairs of sequence data. In addition, there are 200,301,550 WGS records containing 848,977,922,022 base pairs of sequence data.

During the 55 days between the close dates for GenBank releases 204.0 and 205.0, the non-WGS/non-CON portion of GenBank grew by 3,374,386,696 base pairs and by 973,516 sequence records. During that same period, 614,225 records were updated; an average of 28,868 non-WGS/non-CON records were added and/or updated per day. Between releases 204.0 and 205.0, the WGS component of GenBank grew by 43,428,754,314 base pairs and by 4,251,576 sequence records.

The total number of sequence data files increased by 35 with this release. The divisions are as follows:

- BCT: 7 new files, now a total of 159
- CON: 6 new files, now a total of 292
- ENV: 2 new files, now a total of 78
- GSS: 2 new files, now a total of 289
- INV: 8 new files, now a total of 132
- PLN: 6 new files, now a total of 95
- TSA: 3 new files, now a total of 159
- VRL: 1 new file, now a total of 33

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 688 GB (sequence files only). The ASN.1 data require approximately 562 GB.

More information about GenBank release 205.0 is available in the [release notes](#).

### Bald eagle and other bird genome sequence and annotation data publicly available at NCBI

*Thursday, December 11, 2014*

A series of press releases yesterday, including one by Science Publishing, announced the first findings of the [Avian Phylogenomics Consortium](#), who analyzed genome sequence and annotation data for 48 bird genomes representing all of the bird taxonomic orders. All of the sequenced genomes, along with any annotation provided by the submitter, are available in NCBI resources including [Assembly](#), [Nucleotide](#), [Protein](#), the [Sequence Read Archive](#) (SRA), and [BLAST](#), or from species-specific GenBank genomes [FTP directories](#). RNA-Seq data for some of the bird species can be found in SRA.

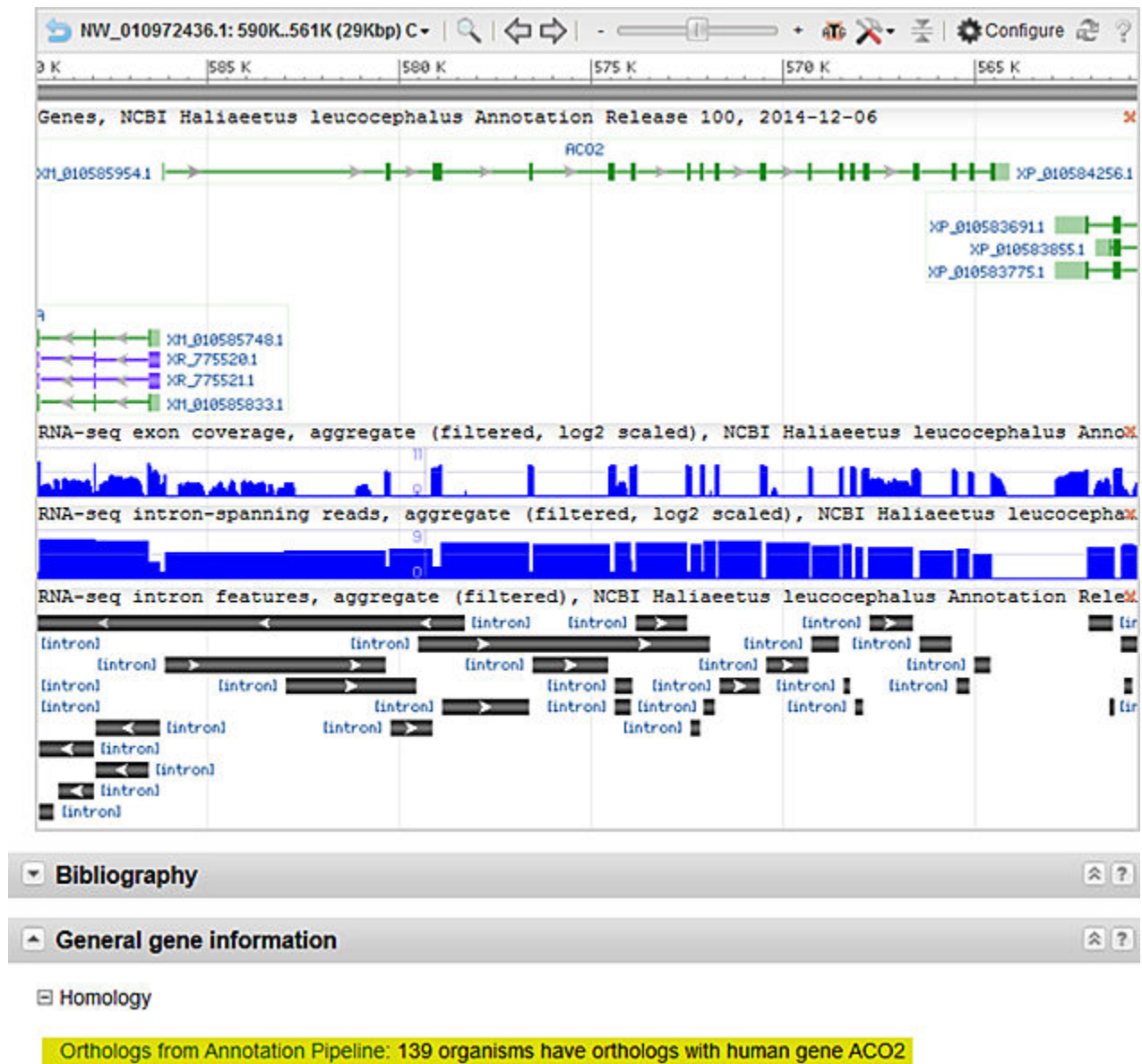
With the exception of three very fragmented assemblies, NCBI annotated the genome assemblies submitted by the Avian Phylogenomics Consortium using NCBI's [Eukaryotic Genome Annotation Pipeline](#), and these annotations are now part of the RefSeq project. The RefSeq project also generated annotations for an additional 6 bird assemblies, for a total of 51 RefSeq genomes. A summary of all the bird genomes that have RefSeq annotation is [here](#).

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links
<i>Acanthisitta chloris</i> (rifleman)	ASM69581v1 (GCF_000695815.1)	100	2014-09-03	2014-09-05	F B AR
<i>Anas platyrhynchos</i> (mallard)	BGI_duck_1.0 (GCF_000355885.1)	100	2013-06-20	2013-06-26	F MV B
<i>Apaloderma vittatum</i> (bar-tailed trogon)	ASM70340v1 (GCF_000703405.1)	100	2014-10-22	2014-10-24	F B AR
<i>Aptenodytes forsteri</i> (emperor penguin)	ASM69914v1 (GCF_000699145.1)	100	2014-09-18	2014-09-22	F B AR
<i>Balearica regulorum gibbericeps</i> (East African grey crowned-crane)	ASM70989v1 (GCF_000709895.1)	100	2014-11-17	2014-11-18	F B AR

**Figure 1.** A selection of the bird genomes with RefSeq annotation. At the top right is a legend describing resource links for each bird genome. Detailed annotation reports, accessible through the "AR" link in the far right column, are available for those genomes annotated in 2014. RefSeq annotation is on organism-specific BLAST pages (the "B" link) and on FTP (the "F" link). Click on the picture to go to the summary table.

RNA-Seq data was used to generate annotations for 12 of the 51 bird assemblies. The number of protein-coding genes per genome ranges from >13,300 to >21,100 (chicken) with an average of 14,932 protein-coding genes. Orthology to human proteins was also calculated, using simple metrics of local synteny and sequence similarity, and on average, roughly 11,000 orthologous proteins were identified per avian genome. These results are shown in the Homology section of NCBI Gene records (see Figure 2 below).





**Figure 2.** A portion of the NCBI Gene report for the bald eagle *ACO2* gene. The graphical display includes information about the gene structure, the RefSeq transcript and protein models, and RNA-Seq coverage graphs produced by the annotation pipeline. The Homology section is highlighted, showing 139 organisms, including the bald eagle, with orthology to the human *ACO2* gene.

### Related stories:

- Revised Genomes FTP site: More information about GenBank and RefSeq sequence and annotation data on the FTP site.

## Citation Exporter Feature Now Available in PubMed Central

Tuesday, December 09, 2014

PubMed Central (PMC) has added a citation exporter, which makes it easy to retrieve styled citations that you can copy and paste into your manuscripts or download in a format compatible with your bibliographic reference manager software.

When viewing an Entrez search results page, each result summary includes a "Citation" link. When clicked, this will open a pop-up window that you can use to easily copy/paste citations formatted in one of three popular styles: AMA (American Medical Association), MLA (Modern Library Association), or APA (American Psychological Association). In addition, the box has links at the bottom that can be used to download the citation information in one of three machine-readable formats, which most bibliographic reference management software programs can import.

The same citation box can also be invoked from an individual article, either in classic view (with the "Citation" link among the list of formats) or the PubReader view, by clicking on the citation information just below the article title in the banner.

These human-readable styled citations and machine-readable formats will be available through a public API, and we will be providing more details about that in another announcement on the [pmc-utils-announce mailing list](#).

## **New NCBI Insights blog post: Designing exon-specific primers for the human genome**

*Tuesday, December 02, 2014*

The [latest blog post](#) on NCBI Insights shows you how to use NCBI Reference Sequences and Primer-BLAST, NCBI's primer designer and specificity checker, to design a pair of primers that will amplify a single exon, using the human breast cancer 1 gene (BRCA1) as an example.

## NCBI News, November 2014

### NCBI to hold two-day genomics hackathon in January

*Wednesday, November 26, 2014*

From January 5th to 7th, NCBI will host a genomics hackathon focusing on advanced bioinformatics analysis of next generation sequencing data. This event is for students, postdocs and investigators already engaged in the use of pipelines for genomic analyses from next generation sequencing data. Working groups of 5-6 individuals will be formed for DNA-Seq/multiomics, RNA-seq, metagenomics and Epigenomics. These groups will build pipelines to analyze large datasets within a cloud infrastructure.

#### Organization

After a basic organizational session, teams will spend 2.5 days analyzing a challenging set of scientific problems related to a group of datasets. Students will analyze and combine datasets in order to work on these problems. This course will take place on the NIH main campus in Bethesda, Maryland.

#### Datasets

Datasets will come from the public repositories housed at NCBI. During the course, students will have an opportunity to include other datasets and tools for analysis. Please note, if you use your own data during the course, we ask that you submit it to a public database within six months of the end of the event.

#### Products

All pipelines and other scripts, software and programs generated in this course will be added to a public GitHub repository designed for that purpose. A manuscript outlining the design of the hackathon and describing participant processes, products and scientific outcomes will be submitted to an appropriate journal.

#### Application

To apply, complete this [form](#) (approximately 10-15 minutes to complete). Applications are due December 1st by 5pm EST. Participants will be selected from a pool of applicants; prior students will be given priority in the event of a tie. Accepted applicants will be notified on December 10th by 9am EST, and have until December 12th at noon to confirm their participation. Please include a monitored email address, in case there are follow-up questions.

**Note:** Students will need to bring their own laptop to this program. A working knowledge of scripting (e.g., Shell, Python) is necessary to be successful in this event. Employment of higher level scripting or programming languages may also be useful. Also note that the course may extend into the evening hours on Monday and/or Tuesday. Please make any necessary arrangements to accommodate this possibility.

Please contact [ben.busby@nih.gov](mailto:ben.busby@nih.gov) with any questions.

### NCBI BioSample includes curated list of over 400 known misidentified and contaminated cell lines

*Monday, November 24, 2014*

The NCBI BioSample database now includes a [curated list](#) of over 400 known misidentified and contaminated cell lines. Scientists should check this list before they start working with a new cell line to see if that cell line is known to be misidentified.

Continuous cell lines are used widely in research as model systems for normal cellular processes and disease states. However, as noted by many (e.g. PubMed [23235867](#), [20143388](#), [19003294](#), [18072586](#), and [17522957](#)), cell line cross-contamination or misidentification represents a serious and widespread problem, and researchers should take great care to check that their cell line is what they think it is. Cell lines can be easily mislabeled or become overgrown by cells derived from a different individual, tissue or species.

This problem is so common it is thought that thousands of misleading and potentially erroneous papers have been published using cell lines that are incorrectly identified (PubMed [20448633](#)). The first step in combating this problem is to make sure your cell line is not on the list of known misidentified and cross-contaminated cell lines. Detailed information about how to test your cell lines is provided by the [International Cell Line Authentication Committee](#).

## NCBI Eukaryotic Genome Annotation Pipeline breaks record; over 100 organisms annotated this year

*Thursday, November 20, 2014*

The NCBI Eukaryotic Genome Annotation Pipeline has broken a record and completed the annotation of [over 100 organisms since the beginning of 2014!](#)

As of today, 81 of this year's 104 annotation releases in RefSeq were first annotations, while 23 were updates. RNA-Seq data was used for gene prediction for 73 of the 104 organisms.

Related links:

- [Request a genome annotation](#)
- [Browse all eukaryotes annotated by NCBI](#)

## NCBI BankIt webinar on December 17th

*Thursday, November 20, 2014*

On December 17th, NCBI will have a webinar entitled "A Submitter's Guide to GenBank: Using BankIt for Small-Scale Nucleotide Sequence Submissions". This presentation will outline the process of using BankIt, a web-based submission tool at NCBI, to submit sequence data to the GenBank database.

Presenters will demonstrate how to use BankIt forms to complete a submission of a single or a few nucleotide sequences, and how to format and upload text input files needed for submissions of multiple sequences or for sequences with multiple genes.

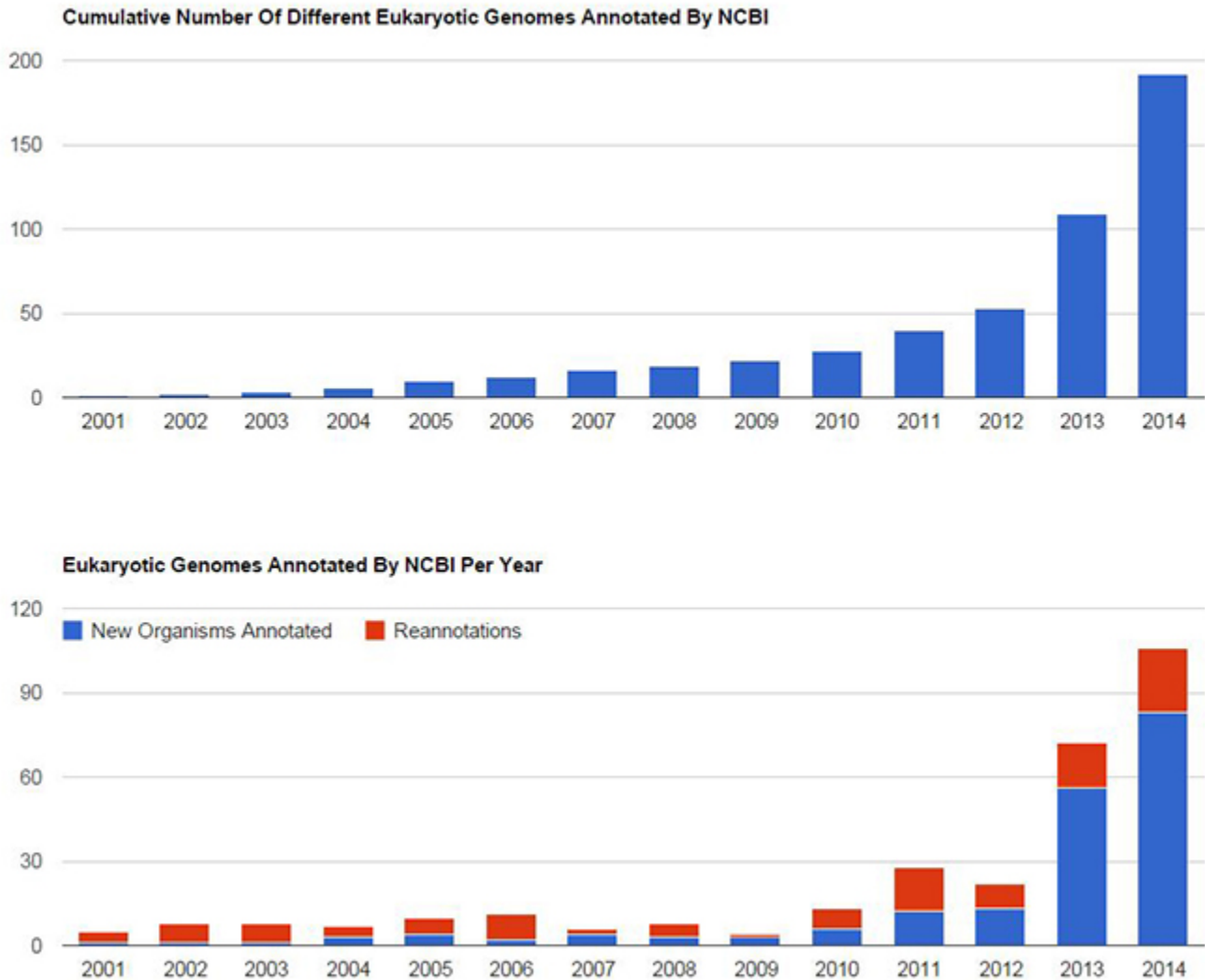
This webinar will stay at a basic level for sequence submissions; future webinars that illustrate more complex sequence submissions will be considered, depending on the feedback received from this presentation.

To register, visit the webinar [registration page](#). To see materials and videos from previous webinars, as well as descriptions of upcoming webinars, see the [NCBI Webinars page](#).

## NCBI E-Utilities webinar video now on YouTube

*Thursday, November 13, 2014*

October's webinar, "An Introduction to NCBI's E-Utilities, an NCBI API", is now on [YouTube](#) and has been added to the [NCBI Webinars playlist](#).



**Figure 1:** *Top:* Cumulative number of different eukaryotic genomes annotated by NCBI. *Bottom:* Eukaryotic genomes annotated by NCBI per year.

For more information about NCBI's webinars including descriptions of upcoming webinars and materials for past presentations, please see the [Webinars homepage](#).

## BLAST URL domain change in effect December 1

*Wednesday, November 12, 2014*

As [announced previously](#), BLAST searches sent to the [www.ncbi.nlm.nih.gov/blast](http://www.ncbi.nlm.nih.gov/blast) URL will not function as of December 1, 2014. The officially supported URL domain for BLAST searches at the NCBI is [blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov). Please update your bookmarks, links, and any scripts or applications.

## RefSeq release 68 available on FTP

*Friday, November 07, 2014*

The comprehensive RefSeq release 68 is now available on the [FTP site](#), with over 66 million records describing 46,968,574 proteins, 9,069,704 RNAs, and sequences from 49,312 distinct NCBI TaxIDs.

More details about RefSeq release 68 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the [accessions included in the release](#) and the [files installed](#) are available.

## **dbVar releases 1000 Genomes Phase 3 structural variants**

*Tuesday, November 04, 2014*

dbVar has released structural variation (SV) data generated by the [1000 Genomes Project Phase 3](#) as dbVar study [estd214](#). This large dataset contains SV from 2,500 subjects, and comprises nearly 63,000 variant regions and over 6 million calls, including insertions, deletions, copy number variants (CNVs), mobile element insertions, indels (deletion-insertions), and inversions. The data are available on assemblies GRCh37 (submitted) and GRCh38 (remapped). Genotypes are currently available in [VCF](#).

The data can be accessed from this [dbVar Study Page](#) and by [FTP](#).

## **dbVar releases copy number variation (CNV) data from developmental delay study cited in Nature Reviews Genetics**

*Monday, November 03, 2014*

dbVar recently released copy number variation (CNV) data from a study on dosage-sensitive genes ([PMID: 25217958](#)) that was highlighted in [Nature Reviews Genetics](#). In the study, CNV analysis was combined with protein-truncating single-nucleotide variation (SNV) and targeted resequencing to identify dosage-sensitive genes causing developmental delay.

CNV data from Coe et al. (2014) can be accessed at [dbVar](#), and the study itself can be found in [PubMed](#).

## NCBI News, October 2014

### BLAST+ 2.2.30 released

*Thursday, October 30, 2014*

A new version (2.2.30) of the stand-alone [BLAST executables](#) is now available, bringing several improvements to BLAST+. These improvements include tasks for BLASTX and TBLASTN (blastx-fast and tblastn-fast) that use longer words, as described in [Shiryev, Papadopoulos, Schaffer, and Agarwala \(2007\)](#), as well as support for composition-based statistics in RPS-BLAST. A number of bug fixes, including those for FASTA parsing, are also included.

The tarballs/installers are located on the [FTP site](#). LINUX, Windows, and MacOSX executables are available [here](#). The BLAST AMI at AWS will also be updated to 2.2.30 (see [this BLAST Help page](#) for information).

For more information, please see the full [release notes](#).

#### Related NCBI News stories:

- June 26, 2014: BLAST machine image hosted at Amazon Web Services (AWS)
- October 16, 2014: Amazon Web Services (AWS) Marketplace provides the easiest way to start an NCBI BLAST instance

### New Genome BLAST selector on the BLAST homepage

*Tuesday, October 28, 2014*

You can now easily find Genome-specific BLAST pages using the search box on the [BLAST homepage](#) under the “BLAST Assembled Genomes” section. This new feature allows you to quickly access and search BLAST databases for the genome of an organism of interest.

Simply start typing your organism name into the box and suggestions will appear. The autocomplete accepts species or strain-level eukaryotic and microbial names as well as metagenomic taxa (community and organism associated metagenomes).

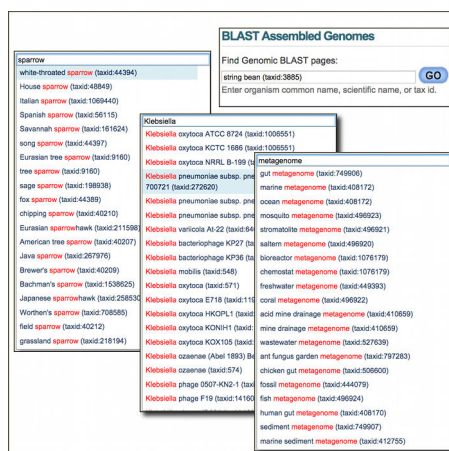
Once you select a suggestion, you will be taken to a BLAST page with the best (most complete, reference) genomic database preselected. In cases where there is no assembled genome sequence, the page will load with whole genome shotgun databases for the organism. If there is no specific genome-sequencing project, the page will load with default nucleotide database (nr/nt) limited to the organism of interest.

### Next NCBI webinar on November 5th

*Thursday, October 23, 2014*

On November 5<sup>th</sup>, NCBI will have a webinar entitled “Exploring and Downloading Sequences and Annotations for Genomes and Metagenomes at the NCBI”. This presentation will introduce you to how NCBI processes genome-level data and produces annotation through the prokaryotic and eukaryotic genome annotation pipelines and show you how to access and download these data from the NCBI site.

You will learn to find, browse, and download genome-level data for your organism of interest and for environmental and organismal metagenomes using the BioProject and Assembly resources. In addition to assembled and annotated data, you will see how to retrieve and download draft whole genome shotgun and read-level next-gen sequencing data from the Nucleotide and Sequence Read Archive (SRA) databases. You will



**Figure 1.** The genome BLAST autocomplete on the BLAST homepage (*top right*). The autocomplete provides matches to organism (common and scientific binomials) and strain names to find genomic datasets.

also see how to access results of precomputed analyses of genomes, as well as perform your own analyses of assembled and unassembled genomic data using NCBI's genome BLAST and SRA-BLAST services.

To register: <https://attendee.gotowebinar.com/register/7154056329796392706>

See materials and video from previous webinars and descriptions of upcoming webinars on the [NCBI Webinars page](#).

## GenBank release 204.0 is now available via FTP

Wednesday, October 22, 2014

Release 204.0 (10/20/2014) has 178,322,253 non-WGS, non-CON records containing 181,563,676,918 base pairs of sequence data. In addition, there are 196,049,974 WGS records containing 805,549,167,708 base pairs of sequence data.

During the 63 days between the close dates for GenBank Releases 203.0 and 204.0, the non-WGS/non-CON portion of GenBank grew by 15,840,696,543 base pairs and by 4,213,503 sequence records. During that same period, 532,480 records were updated; an average of 75,333 non-WGS/non-CON records were added and/or updated per day. Between releases 203.0 and 204.0, the WGS component of GenBank grew by 31,497,068,977 base pairs and by 6,969,555 sequence records.

The total number of sequence data files increased by 123 with this release. The divisions are as follows:

- BCT: 10 new files, now a total of 152
- CON: 8 new files, now a total of 286
- ENV: 2 new files, now a total of 76
- EST: 1 new file, now a total of 477
- INV: 84 new files, now a total of 124
- PAT: 4 new files, now a total of 214
- PLN: 3 new files, now a total of 89
- VRT: 11 new files, now a total of 44

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 680 GB (sequence files only). The ASN.1 data require approximately 557 GB.



More information about GenBank Release 204.0, including important changes included in this release, is available in the [release notes](#).

## dbSNP human Build 142 released

Friday, October 17, 2014

dbSNP human Build 142, based on the GRCh38 and GRCh37.p13 assemblies, is now available on the integrated NCBI Entrez system and through FTP. Build 142 provides 112 million Reference SNP (RS) clusters, including 51 million new RS created from 1000 Genomes Phase III variants as well as from other large sequencing projects. To see complete build statistics, visit the [dbSNP summary page](#). For more information on Build 142, please see this [dbSNP listserv announcement](#).

## Amazon Web Services (AWS) Marketplace provides the easiest way to start an NCBI BLAST instance


Thursday, October 16, 2014

BLAST instances can now be started from the Amazon Web Services (AWS) Marketplace. Using the Marketplace is the easiest way to start a BLAST instance at AWS. In addition, users who subscribe to the BLAST package will be notified when it is updated.

**awsmarketplace** Amazon Web Services Home

Sign in or Create a new account Your Account | Help | Sell on AWS Marketplace

Shop All Categories ▾



### NCBI BLAST

Sold by: NCBI

This BLAST AMI is a very exciting development as it allows users to perform sequence similarity searches without restriction they might encounter at a public website and without the work of setting up stand-alone BLAST. The AMI includes a FUSE client that automatically downloads the most popular BLAST databases from the NCBI, and users can still upload their own custom databases. The AMI allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified webpage.

**Customer Rating** Be the first to review this product

---

**Latest Version** 2014-09-30 v1

---

**Base Operating System** Linux/Unix, Ubuntu 12.04

---

**Delivery Method** 64-bit Amazon Machine Image (AMI) ([Learn more](#))

---

**Support** [See details below](#)

---

**AWS Services Required** Amazon EC2, Amazon EBS

---

**Highlights**

- This AMI is preconfigured with the latest BLAST+ release and has a simplified BLAST web page.
- This AMI includes a FUSE client that automatically downloads and caches popular NCBI databases such as nr, nt, swissprot, refseq, and PDB.
- This AMI supports a subset of the NCBI BLAST URL API allowing remote submission and formatting of searches.

You will have an opportunity to review your order before launching or being charged.

### Pricing Details

For region US East (Virginia)

**Hourly Fees**  
Total hourly fees will vary by instance type and EC2 region.

EC2 Instance Type	EC2 Usage	Software	Total
cc2.8xlarge	\$2.00/hr	\$0.00/hr	<b>\$2.00/hr</b>
cr1.8xlarge	\$3.50/hr	\$0.00/hr	<b>\$3.50/hr</b>
m3.medium	\$0.07/hr	\$0.00/hr	<b>\$0.07/hr</b>
m3.large	\$0.14/hr	\$0.00/hr	<b>\$0.14/hr</b>
m3.xlarge	\$0.28/hr	\$0.00/hr	<b>\$0.28/hr</b>
m3.2xlarge	\$0.56/hr	\$0.00/hr	<b>\$0.56/hr</b>
i2.xlarge	\$0.853/hr	\$0.00/hr	<b>\$0.853/hr</b>

Figure 1. NCBI BLAST on the AWS Marketplace.

As reported this summer, the BLAST instance at AWS is packaged as an AMI (Amazon Machine Image), which allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified web page. The BLAST AMI also includes a FUSE client that can download BLAST databases during the first search.

## Variation Reporter version 1.4 released

*Wednesday, October 15, 2014*

Variation Reporter has just been updated to version 1.4 and now supports querying and reporting on both GRCh37 and GRCh38. For more information about other improvements, see the Variation Reporter [release notes](#).

Variation Reporter is NCBI's tool for matching user-uploaded variant locations against known [dbSNP](#), [dbVar](#) and [ClinVar](#) data.

## Conserved Domain Database (CDD) version 3.12

*Tuesday, October 14, 2014*

Conserved Domain Database (CDD) version 3.12 is now available with 1526 new or updated NCBI-curated domains and 49,955 total domain models from CDD's database providers: Pfam, SMART, COG, TIGRFAMs, Protein Clusters, and the NCBI in-house curation project.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#).

## Updates to assembly alignments for NCBI Remap service

*Friday, October 10, 2014*

NCBI has updated its assembly alignment software (now version 1.7), which generates the alignments used for [Remap](#), NCBI's coordinate remapping service. The improvements include: better handling of alternate loci and fix patches, improved alignments in regions of copy number variation, better recognition of sequence regions that are unaltered between two versions of a WGS assembly, incorporation of fixes to BLAST including bugs affecting alignments around regions with multiple mismatches and indels, and assorted other quality improvements.

These changes improve coordinate remapping from one assembly to another, resulting in better accuracy when remapping features like SNPs or gene annotation to their inferred locations on a new assembly.

Most of the alignment sets available from the Remap service have been regenerated with the v1.7 software, including human GRCh37.p13 x GRCh38. Alignments are available for remapping between various assemblies for 36 taxa, including:

- human GRCh38 (GCF\_000001405.26)
- mouse GRCh38.p3 (GCF\_000001635.23)
- rat Rnor\_6.0 (GCF\_000001895.5)
- zebrafish GRCz10 (GCF\_000002035.5)

Remap can be used through the [web interface](#) and a [public API](#), and the assembly alignments are also available in GFF3 format from the [remap FTP site](#).

## New NCBI Insights blog: Sequence updates in human assembly GRCh38: improving gene annotation

*Thursday, October 09, 2014*

The latest blog post on the [NCBI Insights blog](#) continues the discussion of GRCh38. This time, the blog post focuses on how GRCh38 improved gene annotation.

## Zebrafish (*Danio rerio*) GRCz10 now annotated

*Wednesday, October 08, 2014*

Zebrafish (*Danio rerio*) GRCz10 is [annotated](#)! GRCz10 is an update to the Zv9 assembly, released by the Genome Reference Consortium (GRC), which now manages this reference genome assembly in addition to those for human and mouse. GRCz10 includes more than 1,000 new clone sequences and improvements to the order and orientation of assembly sequences.

RefSeq annotation of GRCz10 was produced by the [Eukaryotic Genome Annotation Pipeline](#). It is available in NCBI's sequence and [BLAST](#) databases, in [Gene](#), and is ready for [download](#). Gene displays annotation data for both Zv9 and GRCz10 to help transition to the new assembly.

GRCz10 annotation is based on transcript and protein evidence including alignments of nearly 2.3 billion RNA-Seq reads (169 billion bases) from 27 distinct BioSample accessions. A total of 30,741 genes and 63,217 transcripts were identified on GRCz10. This includes 14,442 genes (14,019 protein-coding) with known RefSeq transcripts (NM, NP, or NR accessions), and an additional 16,158 predicted genes (12,459 protein-coding) with model RefSeqs (XM, XP, or XR accessions). Note that predicted genes and model RefSeqs aren't included in some resources outside of NCBI. Detailed statistics of annotation results and input reagents are available in the [annotation report](#). NCBI has also annotated 16 other fish genomes, summarized in the [annotated genomes report](#).

For more information about the updates in GRCz10, please see the [GRC zebrafish page](#) or the [GRC blog post](#) on the new assembly. See other organisms that were recently annotated or are currently in the annotation pipeline on the [Eukaryotic Genome Annotation Pipeline status page](#).

## dbVar now accepts VCF submissions of structural variation data

*Friday, October 03, 2014*

dbVar, NCBI's database of genomic structural variation, now accepts submissions in the Variant Call Format (VCF) in addition to their other standard formats: Excel, Tab, and XML. [Instructions](#) for submitting in VCF can be found on the dbVar Home, Submission Guidelines, and Submission Templates pages.

dbVar VCF requirements are largely identical to those of the standard [1000 Genomes VCF spec v4.1](#). However, a few minor changes have been made and are detailed in the [documentation](#). VCF files must be accompanied by one or more files containing metadata using one of the standard formats listed above.

For more information, please visit the [dbVar homepage](#).

## New NCBI Insights blog post: NCBI's medical genetics resources

*Thursday, October 02, 2014*

The latest blog post on NCBI Insights, “[NCBI’s 3 Newest Medical Genetics Resources: GTR, MedGen and ClinVar](#)”, gives an overview of NCBI’s three medical genetics resources and outlines their content features. A more in-depth introduction is available in the [Medical Genetics Resources webinar](#) from June 2014.

## **NCBI webinar on E-Utilities October 15th**

*Wednesday, October 01, 2014*

On October 15<sup>th</sup>, NCBI will have a webinar entitled “An Introduction to NCBI’s E-Utilities, an NCBI API.” E-Utilities is a tool to assist programmers in accessing, searching and retrieving a wide variety of data from NCBI servers.

This presentation will introduce you to the Entrez Programming Utilities (E-Utilities), the public API for the NCBI Entrez system that includes 40 databases such as Pubmed, PMC, Gene, Genome, GEO and dbSNP. After covering the basic functions and URL syntax of the E-utilities, we will then demonstrate these functions using Entrez Direct, a set of UNIX command line programs that allow you to incorporate E-utility calls easily into simple shell scripts.

Click [here](#) to register.

## NCBI News, September 2014

### NCBI Sequence Viewer version 3.4 available

*Tuesday, September 30, 2014*

NCBI Sequence Viewer has recently been updated and now has improved visualization of graphs, sequence track and other text, as well as a reworked configuration dialog. A full list of new features, improvements and fixes is included in the [release notes](#).

Sequence Viewer provides a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.

### HIV-1, human interaction database updated

*Monday, September 29, 2014*

The HIV-1, human interaction database has been updated and is now on an improved [page](#). The improved interface includes help documentation and supports structured queries against [Gene](#), as well as browsing, filtering and downloading the protein and replication interaction data sets. The most recent data release (June 2014) includes 12,785 HIV-1, human protein-protein interactions for 3,142 human genes and 1,316 replication interactions for 1,250 human genes.

NCBI Resources How To

## Retroviruses

HIV-1 Human Interaction Database Browse About Help Publications Releases

### HIV-1 Human Interaction Database

- [About the database](#)
- [Help](#)
- [Publications](#)
- [Releases](#)

### Browse and Download Data

Protein and Replication Interactions

### Search NCBI Gene Records With HIV-1 Interaction Data

Searching human genes with all selected criteria. For HIV-1 genes, [click here](#).

Search criteria

Interaction type  Protein  Replication  Both

Gene Ontology (GO)

Protein domain name

Properties  Has phenotype  Has gene expression data  Has Homologene Cluster  Has >1 RefSeq transcript  Has biological pathways

Entrez Gene keywords

[Clear form](#)

**Figure 1.** The HIV-1 interactions database homepage.

The HIV-1, human interactions project collates published reports of two types of interactions: HIV-1, human protein interactions, and human gene knock-downs that affect virus replication which are reported as “replication interactions”.

# Virus Variation Resource pages for Ebolavirus, MERS coronavirus give quick and easy access to related sequences and other data

Friday, September 19, 2014

NCBI has created resource pages for [Ebolavirus](#) and [MERS coronavirus](#), giving users an easy way to find all sequences related to these pathogens. These pages aggregate links to virus data at NCBI and also provide important links out to other information at the CDC, WHO, and HealthMap.

**Ebolavirus Resource**

Ebolavirus human hemorrhagic fever/disease. Please see "Ebolavirus links" to the right for more information.

**Ebolavirus Resource Components**

The [Ebolavirus database](#) can be used to search and retrieve MERS genome and protein sequences based on standardized biological criteria.

The [Zaire ebolavirus reference genome](#) graphical display supports several interactive functions as described in the "How to use" section below.

**Ebolavirus database**  
 Search ebolavirus sequences  
 Help

**Other NCBI Ebolavirus Resources**  
 Ebolavirus publications  
 Ebolavirus genome browser  
 Ebolavirus taxonomy

**Ebolavirus links**  
 Health Map  
 CDC  
 WHO

**How to use the graphical viewer**

The [NCBI Graphical Sequence Viewer](#) displays sequence annotations using colored bars: **red bars** = gene features; **green bars** = coding regions; **black bars** = other sequence features such as mature peptides and conserved domains.

To learn more information about a given annotation feature, hover your mouse above it. A small pop-up window will appear that includes information about the feature, download options, and links to other NCBI databases.

Left click the tool icon to access a number of useful functions including genome sequence download, BLAST search and primer search.

Please, find detailed information about the use of this graphical display [here](#).

**Figure 1.** Ebolavirus Resource page. The main column gives a brief description of the resource and displays the NCBI Graphical Sequence Viewer. In the right column, from top to bottom, are links to: the [Ebolavirus database](#), other NCBI Ebolavirus resources, and links out to HealthMap, CDC, and WHO. The [MERS coronavirus resource page](#) has the same layout and links.

The Virus Variation resource pages all include a description of the resource components: the database, the reference genome graphical display and links to other resources, both external and within NCBI.

Dedicated Virus Variation databases for [Ebola virus](#) and [MERS coronavirus](#) have been developed. These databases allow searching for nucleotide and protein sequences by a variety of criteria including host, sequence patterns, region or country of isolation, and collection or release dates. The databases allow you to:

- Quickly find the sequences you need, through an intuitive search interface for all viral sequences using standardized protein/gene names and metadata
- Select the latest sequences based on date criteria or sorting of results
- Download sequences in many formats or find links to sequences in NCBI databases.

Visit the [Virus Variation homepage](#) to see resource pages for other viruses.

## Simplified FASTA headers included on new NCBI Genomes FTP site

*Wednesday, September 17, 2014*

Last month, a major revision of the [NCBI Genomes FTP site](#) was announced. In response to user feedback, a new format for FASTA headers of genome, protein and transcript records has been implemented. This new format is limited to records in the `/all/`, `/refseq/`, and `/genbank/` directories on the new Genomes FTP site and does not affect the Nucleotide database web FASTA displays.

Now, instead of "`>gi|xx|dbsrc|accession.version|description`", the new format is simply "`>accession.version description`".

For example, the header on the FASTA record for Homo sapiens chromosome 1 was previously:

```
>gi|568336023|gb|CM000663.2| Homo sapiens chromosome 1, GRCh38 reference primary assembly.
```

On the new Genomes FTP site, the header is now:

```
>CM000663.2 Homo sapiens chromosome 1, GRCh38 reference primary assembly.
```

NCBI has traditionally used a compound FASTA sequence identifier string in which multiple IDs were separated by “|” characters. This format provides more information, but requires that the individual sequence identifiers be parsed out of the compound string. The simpler sequence identifier string is identical to that used in the GFF annotation files on the genomes FTP site. Providing sequence and annotation files with matching sequence identifiers supports their use in commonly used RNA-Seq analysis packages and in other analysis pipelines that rely on simple string comparison to match sequence identifiers.

More information about the revised Genomes FTP site, including the new FASTA header format, is available on the [Genomes Download FAQ page](#).

## RefSeq release 67 available on FTP

*Thursday, September 11, 2014*

The full [RefSeq release 67](#) is now available on the FTP site with over 61 million records describing 45,166,402 proteins, 8,163,775 RNAs, and sequences from 41,913 different NCBI TaxIDs.

More details about the RefSeq release 67 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the [accessions included](#) in the release and the [files installed](#) are available.



**Figure 2.** Ebolavirus database. Users can get sequences by accession or browse by searching for a keyword, host, or region/country, among other options.

## Identical Protein Report Display option added to Protein database

*Tuesday, September 09, 2014*

A new display option has been added to the [Protein database](#) - the "Identical Protein Report". When viewing an individual record, this display allows you to access a list of all other identical proteins including those submitted as translations to GenBank, as well as RefSeq, UniProtKB/Swiss-Prot, PIR, PDB, and patented protein records.

[Display Settings:](#)  Identical Protein Report [Send to:](#)

## 60S ribosomal protein L23a [*Trypanosoma brucei*]

GenBank: AAX79509.1  
[GenPept](#) [FASTA](#) [Graphics](#)

RefSeq Selected Product: [XP\\_846140.1](#), 164 amino acids  
 Name: 60S ribosomal protein L23a [*Trypanosoma brucei brucei* strain 927/4 GUTat10.1]

Source	CDS Region in Nucleotide	Protein	Organism	Superkingdom
RefSeq	<a href="#">NC_007280.1:1362946-1363440+</a>	<a href="#">XP_846140.1</a>	<a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a>	Eukaryota
RefSeq	<a href="#">XM_841047.1:1-495+</a>	<a href="#">XP_846140.1</a>	<a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a>	Eukaryota
Swiss-Prot	N/A	<a href="#">P41165.1</a>	<a href="#">Trypanosoma brucei brucei</a>	Eukaryota
PDB	N/A	<a href="#">3ZF7_X</a>	<a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a>	Eukaryota
INSDC	<a href="#">L21172.1:43-539+</a>	<a href="#">AAC37186.1</a>	<a href="#">Trypanosoma brucei</a>	Eukaryota
INSDC	<a href="#">AC105378.10:65131-65625+</a>	<a href="#">AAX79509.1</a>	<a href="#">Trypanosoma brucei</a>	Eukaryota
INSDC	<a href="#">CP000070.1:1362946-1363440+</a>	<a href="#">AAZ12581.1</a>	<a href="#">Trypanosoma brucei brucei strain 927/4 GUTat10.1</a>	Eukaryota
INSDC	<a href="#">FN554970.1:1431278-1431772+</a>	<a href="#">CBH12694.1</a>	<a href="#">Trypanosoma brucei gambiense DAL972</a>	Eukaryota
Patent	N/A	<a href="#">AAQ39714.1</a>	Unknown	Unknown
Patent	N/A	<a href="#">ACC09504.1</a>	Unknown	Unknown

**Figure 1.** The “Identical Protein Report” display setting in the Protein database, showing identical proteins for 60S ribosomal protein L23 [*Trypanosoma brucei*].

As shown in Figure 1, the page title reflects the protein record from which you started. Beneath that, there is information on the suggested RefSeq preferred protein accession, protein length, and protein name. Identical proteins are presented in a tabular format that includes information on the database source (e.g., RefSeq, INSDC, etc.), the corresponding nucleotide CDS accession and location, the organism name, and the superkingdom. The displayed table can be downloaded for further use, and is also available through [Eutils](#).

The Identical Protein Report display setting provides important functions, such as:

- A mapping table between protein accessions and the nucleotide record(s) on which they are annotated, when relevant;
- For the RefSeq autonomous non-redundant protein dataset, a mapping table to the organisms to which the protein is relevant;
- And identification of highly conserved proteins when the identical protein sequence is found annotated on divergent species.

## NCBI News, August 2014

### Milestone: NCBI annotates 150th eukaryotic genome

Thursday, August 28, 2014

NCBI has now completed the genome annotation for 150 different organisms. The 150<sup>th</sup> organism is the Upper Galilee mountains blind mole rat (*Nannospalax galili*), a rodent of particular interest because of its resistance to cancer.

NCBI began annotating eukaryotic genomes in 2000, and now has complete genome annotations for 150 different organisms, including:

- 74 mammals,
- 39 other vertebrates,
- 21 invertebrates,
- and 16 plants.

Among these organisms, 40 were annotated for the first time in 2014. Data produced by the Eukaryotic Genome Annotation Pipeline is available in the Reference Sequences (RefSeq) collection, BLAST non-redundant and organism-specific databases, Gene database, and on the NCBI FTP site.

View genomes currently in progress and browse the list of all eukaryotes ever annotated by NCBI using the Eukaryotic Genome Annotation Pipeline. Need a public genome annotated? Make a request!

### The new NCBI Genomes FTP site is here!

Tuesday, August 26, 2014

NCBI has released a major revision of the genomes FTP site. The new FTP site structure provides a single entry point to access sequence and annotation content of both GenBank and RefSeq genomes data. The FTP site can be accessed directly for FTP, or from links provided in NCBI's Assembly database.

The initial release of the redesigned genomes FTP site adds three new directories, namely 'genbank', 'refseq', and 'all' to the existing ftp area – <ftp://ftp.ncbi.nlm.nih.gov/genomes/>. It includes >28,000 GenBank and >17,000 RefSeq assemblies ranging from archaea to human and provides a consistent core set of files for the sequence and annotation products. Additional file formats will be added in future updates.

The revised FTP site offers several advantages including:

- comprehensive provision of GenBank and RefSeq genomes data available in NCBI's Assembly database
- provision of a consistent core set of files including:
  - FASTA format for genomic sequences, accessioned transcript products, and accessioned protein products
  - GenBank/GenPept format for genomic, transcript, and protein records
  - GFF (version 3) format for annotated genomic records
  - Md5checksums for all files provided per assembly
- **consistent use of accession.version as the primary sequence ID for both GFF and FASTA files**; this facilitates the use of these data in some public domain RNAseq read mapping tools.

To give those with automated tools time to update, we plan to maintain the older content and structure of the preexisting /genomes/ FTP site in parallel with the new structure until March 1, 2015. The older content will be archived or deleted after that date. Please contact [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov) if you have concerns or questions about these changes.

More information on the initial release and documentation of file formats is available in the following FTP README files:

- [genomes](#)
- [GenBank & RefSeq](#)
- [assembly structure](#).

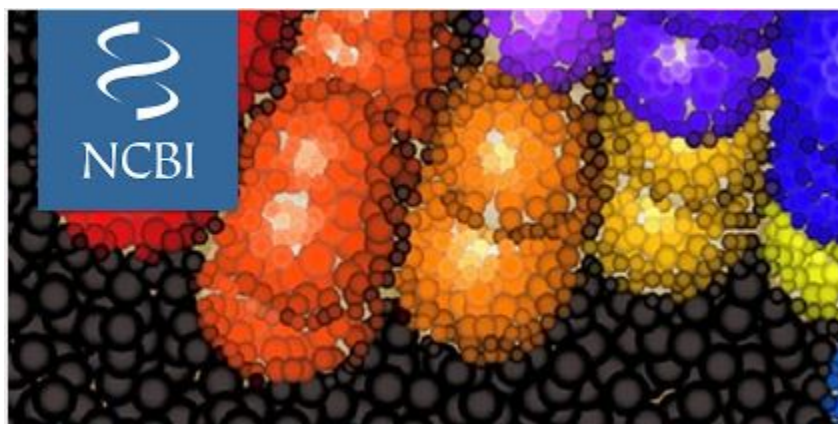
There is also a Genomes FTP FAQ available at [www.ncbi.nlm.nih.gov/genome/doc/ftpfaq](http://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq).

## New NCBI YouTube video: Downloading FASTA sequences in Sequence Viewer

Friday, August 22, 2014

The newest [video](#) on NCBI's YouTube channel is a quick tutorial on downloading FASTA sequences for certain gene features using NCBI's [Sequence Viewer](#).

This video is one of several on the [Sequence Viewer](#) playlist. Subscribe to the [NCBI YouTube channel](#) for notifications on all new videos and to see videos on My NCBI, Variation Viewer, E-Utilities and many more of the programs and services NCBI provides.



## Rat annotation release 105 now on Gene, FTP, sequence and BLAST databases

Friday, August 22, 2014

Rat (*Rattus norvegicus*) assemblies Rnor\_6.0 (GCF\_000001895.5, reference) and Rn\_Celera (GCF\_000002265.2) are annotated in release 105.

This annotation was produced by the [Eukaryotic Genome Annotation Pipeline](#) and is available in the sequence and [BLAST databases](#), in Gene, and on the [FTP site](#).

RNA-Seq data from 340 distinct BioSample accessions were aligned to help gene prediction. A total of 29,998 genes and 61,506 transcripts were identified on Rnor\_6.0. More statistics are available in the [annotation report](#).

See what other annotation runs are in progress on the [Eukaryotic Genome Annotation Pipeline status page](#).

## New NCBI Insights blog: How to comply with NIH Public Access Policy

Thursday, August 21, 2014

The newest blog post on [NCBI Insights](#), "Advice for NIH Grantees: How to Comply with NIH Public Access Policy" guides grantees through the compliance process, from determining whether the Public Access policy applies to their publication to tracking compliance status with My Bibliography.

## "BLAST in the Cloud!" is the newest video on the NCBI Webinars YouTube playlist

Wednesday, August 20, 2014

The video recording of July's NCBI webinar, "BLAST in the Cloud", is live on the NCBI Webinars YouTube playlist, complete with closed captioning.



The [NCBI Webinars playlist](#) contains video recordings of past webinars, and will be updated regularly. Each of the videos on the playlist is accessible from the [NCBI Webinars](#) page as well.

## GenBank release 203.0 is now available via FTP

Wednesday, August 20, 2014

Release 203.0 (8/16/2014) has 174,108,750 non-WGS, non-CON records containing 165,722,980,375 base pairs of sequence data. In addition, there are 189,080,419 WGS records containing 774,052,098,731 base pairs of sequence data.

During the 65 days between the close dates for GenBank Releases 202.0 and 203.0, the non-WGS/non-CON portion of GenBank grew by 3,900,134,732 base pairs and by 755,674 sequence records. During that same period, 403,182 records were updated; an average of 17,828 non-WGS/non-CON records were added and/or updated per day. Between releases 202.0 and 203.0, the WGS component of GenBank grew by 54,470,139,988 base pairs and by 13,301,355 sequence records.

The total number of sequence data files increased by 41 with this release. The divisions are as follows:

- BCT: 6 new files, now a total of 142
- CON: 11 new files, now a total of 278
- ENV: 1 new file, now a total of 74
- EST: 1 new file, now a total of 476
- INV: 1 new file, now a total of 40
- PAT: 1 new file, now a total of 210
- PLN: 16 new files, now a total of 86
- PRI: 1 new file, now a total of 48
- SYN: 1 new file, now a total of 8
- VRL: 1 new file, now a total of 32
- VRT: 1 new file, now a total of 33

For downloading purposes, please keep in mind that the uncompressed GenBank flatfiles are approximately 652 GB (sequence files only). The ASN.1 data require approximately 544 GB.

More information about GenBank Release 203.0 and coming changes are available in the [release notes](#).

## Genome Workbench 2.8.0 released

*Wednesday, August 20, 2014*

Genome Workbench 2.8.0 is available, as of August 18<sup>th</sup>. New features include exporting alignments to tab delimited format, and new flexible broadcasting between bio trees. For the full list of fixes, improvements and features, see the Genome Workbench [release notes](#).

## "NCBI's OSIRIS: Quality Assurance for DNA Forensic Profiling" webinar on September 17th

*Tuesday, August 19, 2014*

On September 17th, NCBI will host a webinar that will cover OSIRIS, an open-source forensics analysis program. **To sign up for this webinar, please go [here](#).**

Identification of people, animals and tissues by forensic, identification and stem cell transplant engraftment laboratories is typically done by analyzing PCR-amplified short tandem repeats (STRs). OSIRIS minimizes analysis time and increases accuracy by identifying artifacts. In addition, OSIRIS can process thousands of samples per day, and the program can also be used by medical and biological research laboratories to identify and validate tissue and cell lines.

The FBI has accepted OSIRIS as a validated expert system, and it is used by the U.S. Army Criminal Investigation Laboratory. OSIRIS can be downloaded [here](#).

To see past and upcoming webinars, please visit the [NCBI Webinars](#) page.

## UniVec build 8.0 now available for VecScreen searches and FTP

*Friday, August 01, 2014*

UniVec, NCBI's non-redundant database of vector sequences, has been updated to build 8.0, which enables searches run using NCBI's [VecScreen](#) tool to detect more of the foreign sequences introduced during the cloning or sequencing process. UniVec build 8.0 is also available via [FTP](#).

This build added 2 complete vector sequences and 257 adapter and primer sequences, including a large number of oligonucleotides used in next-generation sequencing protocols, bringing the total number of sequences represented in the UniVec database to 2,282.

UniVec is a non-redundant database of sequences commonly attached to cDNA or genomic DNA during the cloning process. UniVec primarily consists of the unique segments from a large number of vectors but also includes many linker, adapter and primer sequences. Redundant sub-sequences have been eliminated from the database to make searches more efficient and to simplify interpretation of the results. For more details, see the [UniVec](#) page.





## NCBI News, July 2014

### General Research Use collection streamlines access to patient-level data in dbGaP

*Tuesday, July 29, 2014*

In response to many requests from dbGaP users to simplify and streamline the data access request process while respecting patient consent, dbGaP staff have identified “General Research Use” individuals from different studies and created a collection that allows users to access data on these individuals through a single access request.

Most studies in dbGaP have a significant fraction of participants who consented for “General Research Use.” NIH recognizes that the consents for these study participants are essentially the same, even though the individuals participated in different studies. Therefore, NIH decided to create a streamlined process that would allow users to obtain data on the collection of the individuals who consented for “General Research Use” in one single request.

Investigators approved for access to the datasets within the Collection will have access to the data for the standard one-year approval period. While you may only wish to use data on some of the individuals for your research, approved users will have access to data on the full set of individuals. In addition, any new individuals in the “General Research Use” category will be automatically added to this collection; you will not need to make another request to make use of the data relating to these new individuals during your one-year approval period.

You can obtain access to this collection (currently 71 studies) through a single access request for “dbGaP Collection: Compilation of Individual-Level Genomic Data for General Research Use.” The datasets included in this collection have been designated as appropriate for general research use (GRU) by submitting institutions, which indicates that there are no further limitations on secondary research use beyond those outlined in the Genomic Data User Code of Conduct.

To help expedite the processing of requests for access to the Collection, the NIH review will be conducted by a central Data Access Committee. GRU-designated datasets will be added to the study only after the publication embargo for the original study has expired.

For more information, visit the [dbGaP study page](#).

### NCBI/CDC/FDA/USDA collaboration using whole genome sequencing (WGS) to improve food safety is honored with an HHSinnovates award

*Tuesday, July 22, 2014*

A collaborative project between NCBI and several other Federal and state partners to reduce the time and improve the accuracy of detecting foodborne pathogens by using whole genome sequencing (WGS) techniques received an HHSinnovates award on July 21.

The HHSinnovates program was initiated in 2010 to recognize new ideas and solutions developed by HHS employees and their collaborators. Six finalist teams were recognized at the awards ceremony. The WGS Food Safety Project, which also involved the Centers for Disease Control and Prevention (CDC), the Food and Drug Administration (FDA), the U.S. Department of Agriculture (USDA), and state public health laboratories, was one of three projects to be honored as “Secretary's Picks” by HHS Secretary Sylvia Mathews Burwell.

Studies Included in Collection										
Number of subjects per molecular data type										
Study	Consent Group	16s rRNA (NGS)	CNV Genotypes	RNA_Seq (NGS)	SNP Genotypes (Array)	SNP Genotypes (NGS)	SNP Genotypes (imputed)	Targeted Genome (NGS)	Whole Exome (NGS)	Whole Genome (NGS)
Total subjects		80	22034	2	35898	2866	2563	4949	3826	36
NIH Exome Sequencing of Familial Amyotrophic Lateral Sclerosis Project <a href="#">phs000101.v5.p1</a>	GRU	0	0	0	2914	247	0	0	247	0
Ischemic Stroke Genetics Study (ISGS) <a href="#">phs000102.v1.p1</a>	GRU	0	0	0	266	0	0	0	0	0
GWAS for Genetic Determinants of Bone Fragility <a href="#">phs000138.v2.p1</a>	GRU	0	0	0	1487	0	0	0	0	0
Genetics Consortium for Late Onset of Alzheimer's Disease (LOAD CIDR Project) <a href="#">phs000160.v1.p1</a>	GRU	0	0	0	1132	0	0	0	0	0
NIA - Late Onset Alzheimer's Disease and National Cell Repository for Alzheimer's Disease Family Study: Genome-Wide Association Study for Susceptibility Loci <a href="#">phs000168.v1.p1</a>	GRU	0	0	0	3007	0	0	0	0	0
Whole Genome Association Study of Visceral Adiposity in the HABC Study <a href="#">phs000169.v1.p1</a>	GRU	0	0	0	2801	0	0	0	0	0
International Standards for Cytogenomic Arrays <a href="#">phs000205.v5.p2</a>	GRU	0	22034	0	0	0	0	0	0	0

Figure 1. A sample of the studies included in the dbGaP General Research Use collection.

Presenting the award, HHS Deputy Secretary Bill Corr said: “Together all these folks engaged in a demonstration project to showcase the benefits of using whole genome sequencing for food surveillance and detection purposes. They showed that whole genome sequencing can produce faster detection of foodborne pathogens than the traditional method, helping us stop an outbreak of disease in its tracks, and for that we deeply appreciate your work.” The award went to the specific individuals leading the project in the various agencies; in the case of NCBI, Senior Scientist William Klimke, Ph.D., was honored for his work in heading NCBI’s part of the project.

WGS provides greater specificity than other techniques, such as the commonly used pulsed-field gel electrophoresis (PFGE), in identifying the DNA fingerprint of bacteria. It also can more rapidly determine whether isolates are related to a foodborne disease outbreak.

The demonstration project involves real-time sequencing of *Listeria monocytogenes* isolates from human DNA as well as the food supply chain. In the project, the whole genomes of isolates are sequenced and the sequencing data are sent to NCBI, which performs assembly, annotation, and analysis and then sends results back to CDC, FDA, USDA, and the labs.

Collaborative projects using WGS for other pathogens related to food safety are also underway.

## Major revision of the NCBI genomes FTP site this summer

Friday, July 18, 2014

Within the next two weeks, NCBI will make a major revision to the [genomes FTP site](#). This redesign will expand available content and facilitate data access through an organized, predictable directory hierarchy. The updated site will also provide greater support for downloading assembled genome sequences and/or corresponding annotation data. To give those with automated tools time to update, we plan to maintain the older content and structure of the preexisting /genomes/ FTP site in parallel with the new structure for six months.

The new FTP site structure provides a single entry point to access content representing either [GenBank](#) or [RefSeq](#) data. Advantages of the updated genomes FTP site include the comprehensive provision of data through a single process flow that is reliant on content in [NCBI's Assembly database](#) (which excludes viruses at this time), integration of quality assurance regression tests, and provision of a consistent core set of files for all organisms and assemblies.

Stay tuned to the NCBI News site and our social media accounts ([Facebook](#), [Twitter](#), [LinkedIn](#), [NCBI Announce listserv](#)) for more information about the changes to come and the official launch of the revamped genomes FTP site.

## NCBI webinar "Using the New NCBI Variation Viewer to Explore Human Genetic Variation" on August 13th

*Wednesday, July 16, 2014*

On August 13<sup>th</sup>, NCBI will host a webinar entitled "Using the New NCBI Variation Viewer to Explore Human Genetic Variation". This presentation will show you how to find human sequence variants by chromosome position, gene, disease names and database identifiers (RefSNP, Variant region IDs) using NCBI's new [Variation Viewer](#).

You will learn how to browse the genome, navigate by gene or exon, filter results by one or more categories including allele frequencies from [1000 Genomes](#) or [GO-ESP](#), and link to related information in NCBI's molecular databases and medical genetic resources such as [ClinVar](#), [MedGen](#), and [GTR](#). You will also be shown how to upload your own data to add to the display and download results.

Anyone who works with clinical or research variation data will find that the Variation Viewer provides a convenient and powerful way to access human variation data in a genomic context that is fully integrated with all other NCBI tools and databases.

To register, please go to this link: <https://attendee.gotowebinar.com/register/2762824590748330498>.

## RefSeq release 66 available on FTP site

*Thursday, July 10, 2014*

The full [RefSeq release 66](#) is now available with nearly 59 million records describing 43,671,159 proteins, 7,568,770 RNAs, and sequences from 41,263 different NCBI Taxons.

More details about the RefSeq release 66 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the accessions included in the [release](#) and the [files installed](#) are available.

## NCBI's latest YouTube video presents special features in SciENCv

*Monday, July 07, 2014*

NCBI's latest [YouTube video](#) focuses on special features in [SciENCv](#) (Science Experts Network Curriculum Vitae) that help users create, share, and maintain NIH Biosketch profiles for federal grant applications.

While this video centers on a specific case, anyone with a My NCBI account may use SciENCv. For more general information on SciENCv, click on these links:

- [SciENCv homepage](#)
- [SciENCv blog post on NCBI Insights](#)

## **"BLAST in the Cloud!" webinar on July 30th showcases NCBI-BLAST Amazon Machine Image**

*Tuesday, July 01, 2014*

As stated on June 26, web and standalone BLAST are now available on Amazon Web Services (AWS). On July 30, 2014, NCBI will offer a webinar entitled "BLAST in the Cloud". This presentation will show you how to log on to AWS and deploy the NCBI-BLAST Amazon Machine Image (AMI) quickly. The BLAST AMI includes the BLAST+ applications, a client that can download databases from the NCBI, a web application that implements a subset of the NCBI URL API, and a simplified BLAST search webpage. Prior knowledge of using web and standalone BLAST is required.

To register, please go to this link: <https://attendee.gotowebinar.com/register/8126572163773355778>.

## NCBI News, June 2014

### BLAST machine image (AMI) hosted at Amazon Web Services

Thursday, June 26, 2014

The NCBI now has a BLAST installation at [Amazon Web Services](#), as part of an effort to deliver services to users with new cloud technologies. The installation can be accessed as an Amazon Machine Image (AMI), which allows users to run stand-alone searches with the BLAST+ applications, submit searches through a subset of the NCBI-BLAST URL API, and perform searches with a simplified web page. The AMI also includes a FUSE client that can download BLAST databases during the first search.

For information about the AMI and links to documentation, please see [this page](#).

### Green monkey annotation release 100 now available

Monday, June 23, 2014

The [green monkey \(\*Chlorocebus sabaesus\*\) annotation](#) is now accessible in the Nucleotide, Protein sequence and Gene databases, searchable using [BLAST](#), and downloadable from the [FTP site](#).

*Chlorocebus sabaesus* annotation release 100, based on the sequence assembly Chlorocebus\_sabaesus 1.1 (GCF\_000409795.2) identifies a total of 29,648 genes. This annotation used 22 billion short reads, from 179 distinct BioSample accessions, available from the [Sequence Read Archive](#) to assist in gene prediction. This large amount of short reads allowed the identification of alternative variants for over 13,000 genes.

More statistics are available in the [Chlorocebus sabaesus Annotation Release 100 Report](#).

See what other annotation runs are in progress on the [Eukaryotic Genome Annotation Pipeline status page](#).

### NCBI's latest YouTube video explores Variation Viewer

Wednesday, June 18, 2014

The [most recent video from NCBI](#) demonstrates the basic functions of the [Variation Viewer](#), a tool for navigating variant data in dbSNP, dbVar and ClinVar in a genomic context.

### GenBank release 202.0 is now available via FTP

Tuesday, June 17, 2014

Release 202.0 (6/12/2014) has 173,353,076 non-WGS, non-CON records containing 161,822,845,643 base pairs of sequence data. In addition, there are 175,779,064 WGS records containing 719,581,958,743 base pairs of sequence data.

During the 60 days between the close dates for GenBank Releases 201.0 and 202.0, the non-WGS/non-CON portion of GenBank grew by 2,009,433,883 base pairs and by 1,608,590 sequence records. During that same period, 564,904 records were updated; an average of 36,225 non-WGS/non-CON records were added and/or updated per day. Between releases 201.0 and 202.0, the WGS component of GenBank grew by 98,566,526,306 base pairs and by 32,332,274 sequence records.

The total number of sequence data files increased by 43 with this release. The divisions are as follows:

- BCT: 3 new files, now a total of 136
- CON: 21 new files, now a total of 267

- ENV: 4 new files, now a total of 73
- EST: 1 less file, now a total of 475
- GSS: 2 new files, now a total of 287
- INV: 1 new file, now a total of 39
- PAT: 5 new files, now a total of 209
- PLN: 2 new files, now a total of 70
- TSA: 6 new files, now a total of 156

Note that the loss of one EST data file is *not* due to removal of EST sequences.

For downloading purposes, please keep in mind that the GenBank flatfiles are approximately 42 GB (sequence files only). The ASN.1 data are approximately 538 GB.

More information about GenBank Release 202.0 and coming changes are available in the [release notes](#).

## RefSeq model sequences can now be constructed from genomic and transcript sequences

*Friday, June 13, 2014*

Software version 6.0 of the [Eukaryotic Genome Annotation Pipeline](#) has recently been released. Starting with this release, RefSeq transcript and protein models, which have traditionally been constructed based on the genomic sequence alone, can now be constructed from a combination of the genomic sequence (upon which the model is called) and transcript sequence that compensates for small gaps in the genomic sequence.

This offers a significant improvement in completeness and quality for RefSeq model transcripts and proteins. Sequence records for such models contain the RefSeq attribute “*assembly gap*” and can be queried in Entrez with “*assembly\_gap[properties]*”.

For example, over 3,900 models benefited from this improvement in the recent [green anole annotation release 101](#). One model transcript, [XM\\_003230654.2](#), contains three exons which are derived from the component genomic contig [AAWZ02039332.1](#) of scaffold [NW\\_003342544.1](#) and is also extended at the 5-prime end beyond the end of the scaffold (e.g., into an assembly gap) based on the alignment of transcript [GAFK01002911.1](#). As a result, the transcript [XM\\_003230654.2](#) and protein [XP\\_003230702.2](#) are now complete, while the previous versions, [XM\\_003230654.1](#) and [XP\\_003230702.1](#), annotated in 2011, were partial.

A detailed description of the sequence records for gap-filled models is in the [Eukaryotic Genome Annotation Pipeline software release notes](#).

More information on the RefSeq project can be found [here](#), and more information on annotation runs in progress can be found on the [Eukaryotic Genome Annotation Pipeline status page](#).

## Genome Workbench 2.7.19 released

*Tuesday, June 10, 2014*

Genome Workbench 2.7.19 has been released. The update has several new features, including improved searching and case sensitivity in Text View. The [release notes](#) include more information on features, fixes and improvements.

## dbSNP human Build 141 now available

*Wednesday, June 04, 2014*

dbSNP human Build 141, based on the GRCh38 and GRCh37.p13 assemblies, is now available on the integrated [NCBI Entrez system](#) and through [FTP](#). Build 141 provides more than 260 million submitted SNP (ss) and over 62 million Reference SNP (rs) clusters. To see complete build statistics, visit the [dbSNP summary page](#). For more information on Build 141, including notes on downloading, policy revisions, and reporting on RefSNP, see [this dbSNP listserv announcement](#).

**Update:** In addition to the primary assembly unit of assembled chromosomes, the dbSNP annotation of GRCh38 and other Genome Reference Consortium (GRC) assemblies will also contain:

- Patch sequences: sequences provided as assembly updates outside of the normal release cycle;
- Alternate loci: sequences that provide an alternative representation of a locus found in a largely haploid assembly;
- PAR sequences: pseudo-autosomal region found on the X and Y chromosomes of mammals; and
- Unplaced sequences: sequences found in an assembly that are not associated with any chromosome.

For more detailed definitions of these sequences, visit the [GRC Assembly Terminology page](#).

Although interesting genetic variations may occur in these sequences, most researchers may never see them, due to current VCF file specification limits. VCF files are a file format used by the [1000 Genomes Project](#) to store genomic variation information. To support reporting on these additional, non-primary chromosome locations, NCBI has released companion files for clinical data, named with a "papu" (patch, alternate, PAR, unplaced) extension. The files are available on the dbSNP FTP sites for [GRCh38](#) and [GRCh37.p13](#). Please note that GRCh38 currently does not have PATCH sequences. VCF files will be updated when these sequences are released for GRCh38.

## New features simplify access to annotation information in NCBI's Gene

Tuesday, June 03, 2014

NCBI's [Gene resource](#) is pleased to announce several new features aimed at providing easier access to annotation information.

**First**, the "Genomic context" section for genes annotated using [NCBI's Eukaryotic Genome Annotation Pipeline](#), including human, mouse, and 130 other species, has been restructured to include a table that includes the Annotation Release, Assembly, and sequence Location. The table provides a convenient view of the location on the reference primary assembly.

**Genomic context** See TIF in Epigenomics

Location: 6p21.3

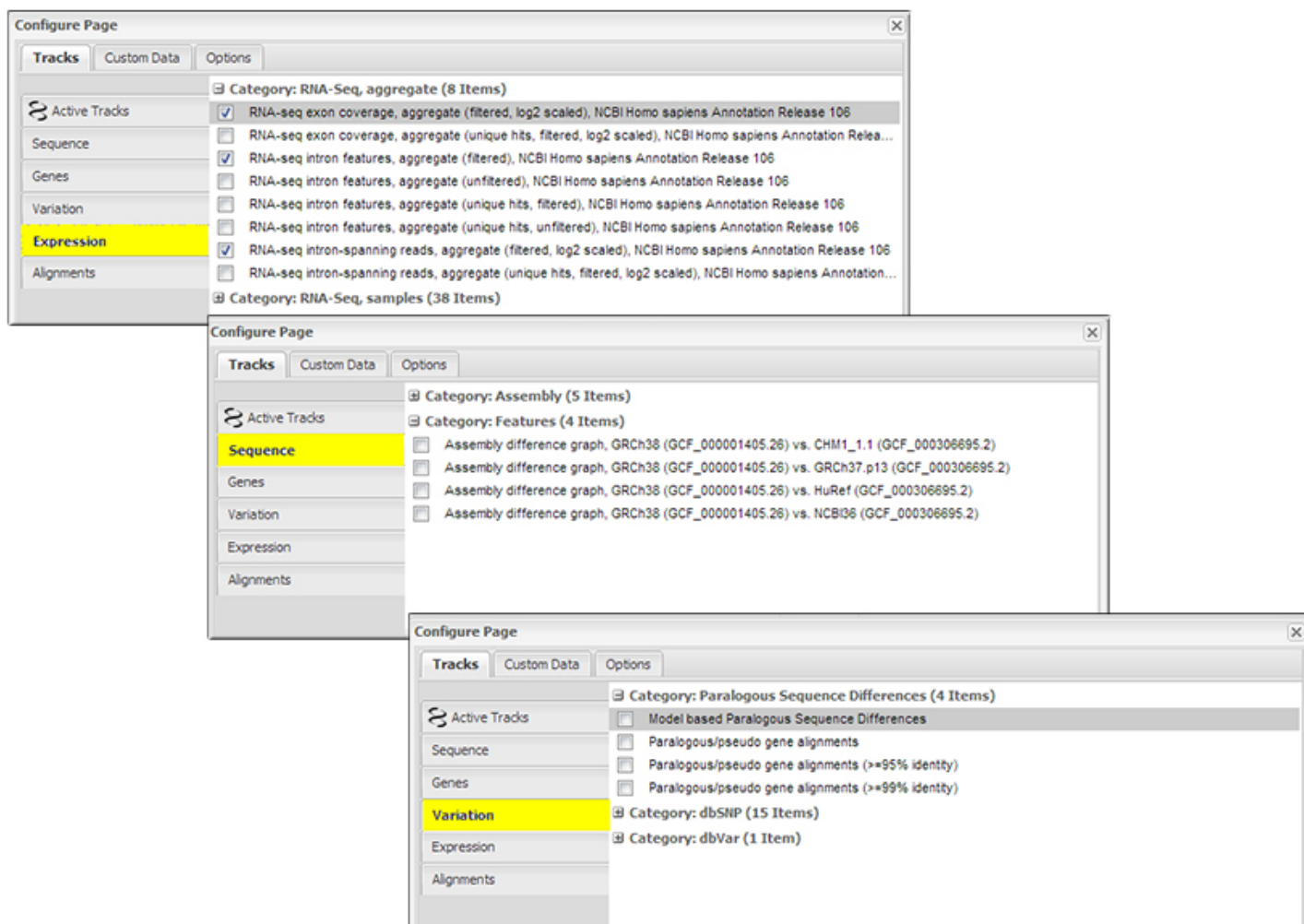
Annotation release	Status	Assembly	Chr	Location
106	current	GRCh38 ( <a href="#">GCF_000001405.25</a> )	6	NC_000006.12 (31575567..31578336)
105	previous assembly	GRCh37 p13 ( <a href="#">GCF_000001405.25</a> )	6	NC_000006.11 (31543344..31546113)

**Second**, to make it easier for users working with previous assembly versions, the same table mentioned above includes the sequence location from the last annotation of the previous assembly version. This feature is currently limited to human, where it displays the location on the GRCh37.p13 assembly. It will be expanded to more organisms with future assembly updates.

**Third**, the "Genomic regions, transcripts, and products" section also includes the last assembly in the pulldown menu, giving you access to the prior annotation in the graphical view and through the "Go to nucleotide:" links.

**Fourth**, one of the most exciting changes can be found in the Graphical View "Configure" page, accessed using the button in the upper right corner of the graphic. This interface now provides access to many more tracks than were previously available, including:

- Under the "Expression" tab, RNA-seq expression tracks computed for each individual BioSample that were aligned as part of the annotation process. These data can provide valuable information about differential expression in tissues or developmental stages. RNA-seq tracks are currently available for 65 taxa that have been annotated using NCBI's Eukaryotic Genome Annotation Pipeline.
- Under the "Sequence" tab, Assembly difference graphs that highlight differences between the GRCh38 and other human assemblies.
- Under the "Variation" tab, paralogous gene alignment tracks that show the alignment of paralogous gene features. These tracks are a useful view of how similar a gene is to any pseudogenes or other paralogs that are annotated in the genome.



**Fifth**, the Variation section for human genes includes links to the new Variation Viewer genome browser, using either the GRCh37.p13 or GRCh38 assemblies.

Please stay tuned as documentation is updated to reflect these new changes.



## NCBI News, May 2014

### BLAST URL domain changes to take effect December 1, 2014

*Thursday, May 22, 2014*

As of December 1, 2014, BLAST searches sent to the [www.ncbi.nlm.nih.gov/blast](http://www.ncbi.nlm.nih.gov/blast) URL will not function. The officially supported URL domain for BLAST searches at the NCBI is **blast.ncbi.nlm.nih.gov**. Please update your bookmarks, links and any scripts or applications.

### GTR/ClinVar/MedGen webinar on June 18 will explore NCBI resources

*Wednesday, May 21, 2014*

On June 18, 2014, NCBI will offer a webinar entitled “Introducing 3 NCBI Resources to Navigate Testing for Disease Linked Variants: MedGen, GTR and ClinVar”. This webinar will delve into the lifecycle of genetic testing and teach attendees how to navigate the NIH Genetic Testing Registry, ClinVar, and MedGen resources. These resources can be used to prepare for clinical cases, access detailed information about orderable genetic tests, interpret test results, and more.

To register, please go to [this link](#).

### RefSeq release 65 available on FTP site

*Tuesday, May 20, 2014*

The full RefSeq release 65 is now available with nearly 52 million records describing 38,633,935 proteins, 7,051,549 RNAs, and sequences from 36,335 different organisms.

As mentioned previously, several changes to the RefSeq release FTP site are now in place with this most recent release, including changes to directory names and file names. For more detail on these changes, please refer to [this announcement](#).

More details about the RefSeq release 65 are included in the [release statistics](#) and [release notes](#). In addition, reports indicating the [accessions included in the release](#) and the [files installed](#) are available.

### The NIH Genetic Testing Registry now has information on more than 3,500 cancer tests

*Tuesday, May 20, 2014*

Thanks to the efforts of testing laboratories that have voluntarily submitted information about their tests to NIH's Genetic Testing Registry (GTR), the database currently has information on more than 17,000 genetic tests, including about 3,500 related to cancer.

In late 2013, GTR expanded to include information about tests that evaluate somatic mutations (alterations in DNA that occur after conception). The response to that expansion was strong: laboratories have already submitted information on over 120 somatic tests for more than 200 conditions. Many of the somatic tests are for use in the area of cancer, including tests for prediction, prognosis, recurrence, and therapeutic management. To explore somatic tests in GTR, see [this link](#).

In addition to somatic tests, GTR includes information on tests for hereditary cancers. GTR has included these types of tests since its inception in 2012, and laboratories have submitted information on more than 3,500 of them to date. For example, GTR has information on 199 tests for *BRCA1* or *BRCA2* from laboratories across the world, including 59 tests offered by 18 US labs. As advanced DNA sequencing techniques permeate clinical testing, complex panels for cancer have grown sharply in GTR: 30 of the tests for *BRCA1* or *BRCA2* are panels that evaluate 5 or more genes.

Pharmacogenetic tests, some of which relate to cancer drug responses, are another category of tests included in GTR. For example, tests have been registered in GTR for responses to tamoxifen, irinotecan, fluorouracil and thioguanine.

For all types of tests and in all medical areas, GTR aims to provide detailed information, such as the purpose of the test, target populations, methods, what it measures, analytical validity, clinical validity, clinical utility, and ordering information. Details about the laboratories include location, contact information, certifications and licenses.

GTR also includes extensive information beyond that relating to individual tests and laboratories. GTR links to context-specific information about medical conditions, genes, sequence variation, test standards, practice guidelines, pharmacogenetic information, clinical trials, molecular resources, and consumer support sites.

The GTR staff is excited about the growing participation in and the use of the database, as well as about recently introduced improvements to its functionality, such as the new advanced search feature for tests that was introduced in March. Searches can quickly find tests based on somatic or germline targets, pharmacogenetic responses, next-generation sequencing methods, number of targets (for complex panels), required specimen types, laboratory location, and more.

To learn more about GTR, visit the [Genetic Testing Registry website](#), or see [this article](#) in *Nucleic Acids Research*. If you happen to be attending the American Society of Clinical Oncology annual meeting at McCormick Place in Chicago, check out the GTR poster on May 31, 1:15 p.m. to 5 p.m. (in the Tumor Biology section of the General Poster Sessions). GTR Director Wendy Rubinstein, M.D., Ph.D. will be at the meeting from May 31 to June 3 and would be happy to talk with you about how to use the GTR website or how your laboratory can participate. You can contact us in advance via [gtr@ncbi.nlm.nih.gov](mailto:gtr@ncbi.nlm.nih.gov) to arrange a time to meet.

## SciENcv 2.0 brings major improvements to My NCBI

*Friday, May 16, 2014*

Recent updates to SciENcv add several new useful features to the service.

Users can now:

- create multiple SciENcv profiles;
- download SciENcv profiles in PDF, Word, or XML format;
- add delegates to their SciENcv profiles;
- and add a mini profile in SciENcv to link to PubMed Commons.

The most recent [NLM Technical Bulletin](#) describes these changes in more detail, as will an upcoming [NCBI Insights](#) blog post.

SciENcv is a feature in My NCBI that helps users create online professional profiles that can be made public to share with others. SciENcv allows you to document education, employment, research activities, [ORCID iDs](#), publications, grants, and other professional contributions. A guide on how to navigate and use SciENcv is on the [Bookshelf](#).

## NCBI Sequence Viewer version 3.2 available

*Tuesday, May 06, 2014*

NCBI [Sequence Viewer](#) has recently been updated and now has support for multi-track upload and improved display of long track titles, as well as improved tooltips for variation features. A full list of new features, improvements and fixes is included in the [release notes](#).

Sequence Viewer provides a graphical view of sequences and color-coded annotations on regions of sequences stored in the Nucleotide and Protein databases.



## NCBI News, April 2014

### Coffee Break tutorial: The promise of PCSK9

*Thursday, April 24, 2014*

The latest [Coffee Break tutorial](#) explores PCSK9, an enzyme that plays a major regulatory role in cholesterol homeostasis, and the cholesterol-lowering drugs that target it. This Coffee Break also includes a video exploration of [ClinicalTrials.gov](#).

### New NCBI Insights Blog: Sequence updates in human genome assembly GRCh38

*Wednesday, April 23, 2014*

The latest [NCBI Insights blog post](#) explores just one of the many changes and improvements introduced by the newest human reference genome released in December 2013, [GRCh38](#).

### HomoloGene release 68 now available

*Tuesday, April 22, 2014*

HomoloGene release 68 is now available on [Homologene!](#) In this release, genome annotation was updated for 19 organisms, the number of HomoloGene groups increased to 44,233, and one organism, *Xenopus tropicalis*, was added. Release 68 is also available on the [FTP site](#).

HomoloGene is an NCBI resource that identifies and clusters homologous genes, transcripts and proteins for selected eukaryotes.

### Milestone: NCBI's Taxonomy database contains over 300,000 species with formal scientific names!

*Friday, April 18, 2014*

NCBI's Taxonomy database has now surpassed 300,000 individual records of species with formal scientific names. The majority of these represent eukaryotic organisms. While worldwide estimates of prokaryotic and viral species number in the millions or tens of millions, very few have been formally described, isolated, or are able to be cultured. However, the Taxonomy database contains listings for nearly all of the prokaryotes and viruses that have been described.

The [Taxonomy database](#), created in 1991, is a standard nomenclature and classification repository that includes organism names and taxonomic lineages for each of the sequences represented in the International Nucleotide Sequence Database Collaboration (INSDC).

### GenBank release 201.0 is now available via FTP

*Thursday, April 17, 2014*

Release 201.0 (4/13/2014) has 171,744,486 non-WGS, non-CON records containing 159,813,411,760 base pairs of sequence data. In addition, there are 143,446,790 WGS records containing 621,015,432,437 base pairs of sequence data.

During the 60 days between the close dates for GenBank Releases 200.0 and 201.0, the non-WGS/non-CON portion of GenBank grew by 1,869,618,589 base pairs and by 620,737 sequence records. During the same period, 2,046,345 records were updated; an average of 44,451 non-WGS/non-CON records per day were added and/or updated. Between releases 200.0 and 201.0, the WGS component of GenBank grew by 29,636,733,893 base pairs and by 3,720,995 sequence records.

The total number of sequence data files increased by 18 with this release. The divisions are as follows:

- BCT: 15 new files, now a total of 133
- CON: 4 new files, now a total of 246
- ENV: 2 new files, now a total of 69
- EST: 1 new file, now a total of 476
- GSS: 1 new file, now a total of 285
- MAM: 1 new files, now a total of 9
- PLN: 1 new file, now a total of 68
- VRL: 2 new files, now a total of 31

For downloading purposes, please keep in mind that the GenBank flatfiles are approximately 625 GB (sequence files only). ASN.1 data are approximately 527 GB.

More information about GenBank Release 201.0 and coming changes are available in the [release notes](#).

## Enterococci: From commensals to leading causes of drug resistant infection on Bookshelf

*Tuesday, April 15, 2014*

“*Enterococci: From commensals to leading causes of drug resistant infection*” (Michael S Gilmore, Don B Clewell, Yasuyoshi Ike, and Nathan Shankar, editors; Boston: Massachusetts Eye and Ear Infirmary; 2014), a comprehensive text aiming to advance the understanding of *Enterococci* is [free to access on the NCBI Bookshelf](#). This book has been compiled from peer-reviewed content contributed by leaders in the *Enterococcus* research community, and will be regularly updated on the Bookshelf.

*Enterococci* are an ancient and highly evolved genus of bacteria that were first described 115 years ago. They adapted to survive in a range of extreme environments, and thrive inside the gastrointestinal tracts of many species, including humans. In the human gut, commensal enterococci act in harmony with other microbes to do good gastric deeds, such as helping with digestion and crowding out harmful microbes.

However, *Enterococci* have continued to evolve to have more harmful roles. About 40 years ago, new traits in *E. faecalis* and *E. faecium* propelled these microbes to become the leading cause of multidrug-resistant, hospital-acquired infections. In addition, *Enterococci* can spread their new traits for antibiotic resistance to other pathogens, such as *Staphylococcus aureus*.

To date, there are over 40 distinct species of *Enterococcus*. Understanding the appearance of new strains of antibiotic-resistant enterococci, some of which feature new resistance mechanisms, is key to finding a solution to multidrug resistance.

## CCDS release 16 for mouse now public in Gene

*Monday, April 14, 2014*

The Consensus Coding Sequence (CCDS) update for *Mus musculus* annotation release 104 released last week is now reflected in [Gene](#). This update adds 803 new CCDS IDs and 97 genes into the mouse CCDS set. CCDS

release 16 includes a total of 23,880 CCDS IDs that correspond to 20,079 GeneIDs. For more information, visit the [CCDS homepage](#).

## **New on the Bookshelf: The Art and Politics of Science, a memoir by Dr. Harold Varmus**

*Thursday, April 10, 2014*

Dr. Harold Varmus's memoir, *The Art and Politics of Science* (W.W. Norton and Company, New York; 2009), is now freely available on the [NCBI Bookshelf](#).

In this book, he chronicles his path from a graduate student in English literature at Harvard to co-reipient of the Nobel Prize for cellular origin of retroviral oncogenes, to director of the National Institutes of Health, to President and CEO of Memorial Sloan-Kettering Cancer Center.

## **Coffee Break tutorial: Brown fat and obesity**

*Tuesday, April 01, 2014*

The latest [Coffee Break tutorial](#) discusses EHMT1, an enzyme responsible for brown fat production, and its possibility as a target for new obesity treatments.

The tutorial includes a discussion of recent research on brown fat cells, EHMT1<sup>adipo</sup> knockout mice, and a video exploration of ClinVar, PubMed, and Bookshelf.





## NCBI News, March 2014

### New NCBI YouTube video: Create custom databases for BLAST

Friday, March 28, 2014

In the newest NCBI video on YouTube, we show you how to create custom databases in BLAST. This video gives you a step-by-step tutorial on limiting your web BLAST searches to a customized set of sequences.

[YouTube](#)

### Come to the NCBI Discovery Workshops on May 6th & 7th!

Friday, March 28, 2014

The NCBI Discovery Workshops will be held on the NIH Campus on May 6 and 7. To get more information and to register, visit the [Discovery Workshops homepage](#).

The NCBI Discovery Workshops, a 2-day event, comprise of four workshops that will teach you how to use the NCBI Web resources more effectively. The May 2014 Workshops consist of four 2.5-hour hands-on sessions, with each session focusing on a different related group of NCBI tools and databases:

- Sequences, Genomes, and Maps
- Proteins, Domains, and Structures
- NCBI BLAST Services
- Human Variation and Disease Genes

Materials from all Discovery Workshops offerings are available from the [Education FTP directory](#).

### NCBI will attend the 2014 ACMG Annual Clinical Genetics Meeting

Thursday, March 20, 2014

NCBI staff will attend the [2014 ACMG Annual Clinical Genetics Meeting](#) in Nashville, TN on March 25-29.

At Booth #1105, you'll be able to do a variety of things including:

- Generate a differential diagnosis based on clinical features,
- Search the ACMG incidental findings gene set against your variant result,
- Join NIH's open access mission and submit data to GTR and ClinVar,
- And get hands-on help from staff.

For more details, see the [GTR page](#).

### NCBI requests feedback on proposed BLAST XML specification update

Monday, March 17, 2014

The BLAST development team is planning to update the BLAST XML specification in the Summer of 2014 and would like feedback from the user community on the proposed changes. This update is designed to improve consistency of the BLAST output with XML standards and implement new and useful elements. The [BLAST proposal](#) outlines these intended changes.

If you are a BLAST XML user, please provide feedback on the proposed changes using this web form. We thank you in advance for your input.

## RefSeq full release 64 out

*Friday, March 14, 2014*

The full RefSeq release 64 is now available with nearly 50 million records describing 37,818,139 proteins, 6,198,996 RNAs, and sequences from 33,693 different organisms.

Some important updates include the following:

**SNP annotation update:** A list of updated organisms and dbSNP annotation summary is available in the SNP RefSeq release notes folder on the FTP site ("[refseq63.snp.rpt](#)").

**Domain annotation update:** RefSeq domain and site features that are provided by the [Conserved Domain Database](#) were updated in conjunction with CDD release 3.11. For more information on release 3.11, see the NCBI News story from last month.

**New annotation for the updated human reference genome assembly, GRCh38:** The Genome Reference Consortium released a major update to the human reference genome assembly (GRCh38) in late December 2013. In January 2014, this updated assembly, plus two other human genome assemblies (HuRef and CHM1\_1.1), was annotated using NCBI's eukaryotic genome annotation pipeline which integrated information from curated RefSeqs, cDNAs, ESTs, protein alignments, and RNA-Seq data from the Human BodyMap2 project. Results for all three genomes are available as [NCBI Annotation release 106](#).

More details about the RefSeq release 64 is included in the [release statistics](#) and [release notes](#). In addition, reports indicating the accessions included [in the release](#) and [the files installed](#) are available.

## Orthologous genes and gene regions now accessible through Gene

*Wednesday, March 12, 2014*

Each Gene record now provides access to orthologous genes and regions in the "General gene information" section of the Gene record (Figure 1). In addition, complex loci in a particular species, such as the human immunoglobulin heavy locus, now have links to the corresponding individual members.

The "Orthologs from Annotation Pipeline" link under the Homology subsection of "General gene information" accesses the set of orthologs in selected vertebrate genomes using the method described in PMID: [PMC3882889](#). For example, this link from the zebrafish *abl1* gene record (Gene ID: [100000720](#)) or from any other member of this orthology group provides [80 orthologous gene records](#) from a wide range of vertebrate species (birds, mammals, turtles, fishes, and the coelacanth). These ortholog data are supplemental to those currently available from the HomoloGene resource also linked under the "Gene information: Homology" section of the Gene record. The Annotation Pipeline method is being improved to include more distantly related organisms in the future.

Region gene records are available for loci that are officially named and are composed of multiple parts or clusters of related genes. The "Related region members" section in a region gene record has a "Review record(s) in Gene" link that provides all genes that are components of the region. For example, the link from the human and mouse immunoglobulin H region records (Gene IDs: [3492](#) & [11507](#)) provide [182](#) and [215](#) records respectively.

The data for both gene orthology groups and gene regions are available in the [gene\\_group.gz](#) file in the Gene area of the NCBI FTP site. In the file, the terms “Ortholog”, “Region members”, and “Region parent” are used to report these new relationships.

## New dbGaP online system for registering studies and applying for data access introduces time-saving features

*Wednesday, March 12, 2014*

In an effort to reduce burden, NIH has developed an online system for researchers and their institutional officials to register studies, submit data, and access data in dbGaP. The online system introduces a number of time-saving features, such as automatically completing data fields from other sources, for example, using eRA Commons to provide the investigator’s name, institution, and Institutional Signing Official.

Tutorials on the online forms for study registration and data access are available on Youtube:

### **dbGaP: Complete a Study Registration**

YouTube

### **dbGaP: Controlled Access Data**

YouTube

### **dbGaP: Renew Authorized Access**

YouTube

### **dbGaP: Close Out a Controlled Access Project**

YouTube

Additional information can be found in the [NIH Guide Notice](#).

## New Sorting and Output Options for E-utilities

*Monday, March 10, 2014*

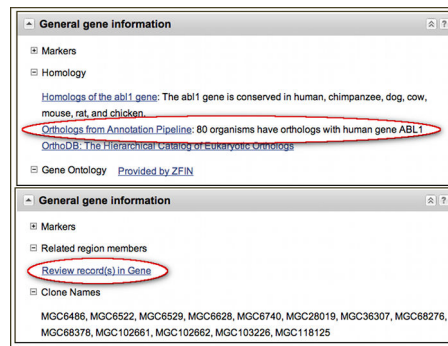
E-utilities, the API to the NCBI Entrez system, now offers two new options for data retrieval.

First, ESearch now offers a fully supported *&sort* parameter that determines the sort order of the UIDs returned. While each Entrez database has a default sort order, each also provides a variety of other sorting options found in the “Display Settings” menu at the top of a search result page. Any of these options may now be given to ESearch using *&sort*.

For example, to sort a PubMed retrieval by First Author, simply add “*&sort=first+author*” to your ESearch URL. If you are also supplying “*&usehistory=y*”, then the UIDs will also be sorted on the Entrez History so that when retrieved by ESummary or EFetch, the sort order will be retained in that output. Of particular interest is the new [relevance sort option](#) now available in PubMed and several other databases. To enable this for ESearch, use “*&sort=relevance*”.

Second, the output for EInfo, ESearch and ESummary are now available in the popular [JSON](#) format. To request data in JSON, simply append “*&retmode=json*” to the E-utility URL. Click on the following links for example outputs:

- [EInfo](#)



**Figure 1.** General gene information section of Gene records. *Top panel:* The Homology subsection of the zebrafish *abl1* record (Gene ID: 100000720) showing the link to “Orthologs from Annotation Pipeline” (circled in red) as well as the “Homologs of the *abl1* gene” link to HomoloGene. *Bottom panel:* mouse *Igh* region (Gene ID: 111507) showing the link to “Review record(s) in Gene” (circled in red).

- [ESearch](#)
- [ESummary](#)

Please see the [E-utilities documentation](#) for additional details.

PubMed   [RSS](#) [Save search](#) [Advanced](#) [Help](#)


**Display Settings:**  Summary, 20 per page **Sorted by Relevance** **Send to:**  **Filters:** [Manage Filters](#)

**Results: 1 to 20 of 24446** << First < Prev Page  of 1223 Next > Last >>

[Thunderclap headache: diagnostic considerations and neuroimaging features.](#)  
1. Mortimer AM, Bradley MD, Stoodley NG, Renowden SA.  
Clin Radiol. 2013 Mar;68(3):e101-13. doi: 10.1016/j.crad.2012.08.032. Epub 2012 Dec 11.  
Review.  
PMID: 23245274 [PubMed - indexed for MEDLINE]  
[Related citations](#)

[A case of cerebellar infarction presenting as thunderclap headache.](#)  
2. Jo YS, Choi JY, Han SD, Kim YD, Na SJ.  
Neurol Sci. 2012 Apr;33(2):321-3. doi: 10.1007/s10072-011-0673-6. Epub 2011 Jul 1.  
PMID: 21720896 [PubMed - indexed for MEDLINE]

**New feature**  
Results currently sorted by Relevance - **Sort by Recently Added**

**Results by year**  
  
[Download CSV](#)

Results sorted by Relevance.



## NCBI News, February 2014

### Genome Workbench Update 2.7.15 released

*Wednesday, February 26, 2014*

Genome Workbench 2.7.15 has been released. The update includes several new features like Multiple Alignment View, Active Objects Inspector, and binary packages for Linux OpenSUSE 13.1. The [release notes](#) include more information on features, fixes and improvements.

### New CDD Release v.3.11 includes recomputed PSSMs and more

*Wednesday, February 19, 2014*

Conserved Domain Database (CDD) version 3.11 is now available with 596 new or updated NCBI-curated and 49,641 total domain models. The new version now contains the most recent Pfam release 27.

Updates to the Conserved Domain Database include:

- Position-specific score matrices (PSSMs) have been recomputed for many models in CDD, and frequency tables have been added to the PSSMs;
- The search databases distributed as part of this release can now be used with the more recent versions of RPS-BLAST (BLAST release 2.2.28 and up) using composition-based scoring. This abolishes the need to mask out compositionally biased regions in query sequences;
- Domain annotation displays in CD-Search, BATCH CD-Search, and other services now all use a uniform display style. A new display option in CD-Search and BATCH CD-Search provides “standard” results, in addition to “concise” and “full” results. “Standard” results will provide, for each region on the query sequence, the best0-scoring domain model (if any) from each of CDD’s database providers (Pfam, SMART, COG, TIGRFAMs, Protein Clusters, and the NCBI in-house curation project), but will suppress redundancy from within a single provider's results list.

You can access CDD at the [Conserved Domains homepage](#) and find updated content on the [CDD FTP site](#).

### GenBank has milestone 200th release

*Tuesday, February 18, 2014*

[GenBank's 200th release](#) is now available through NCBI's Entrez and BLAST services.

Release 200.0 (2/12/2014) has 171,123,749 non-WGS, non-CON records containing 157,943,793,171 base pairs of sequence data. In addition, there are 139,725,795 WGS records containing 591,378,698,544 base pairs of sequence data.

During the 64 days between the close dates for GenBank Releases 199.0 and 200.0, the non-WGS/non-CON portion of GenBank grew by 1,713,261,609 base pairs and by 1,792,342 sequence records. During that same period, 4,979,722 records were updated. An average of 105,813 non-WGS/non-CON records were added and/or updated per day. Between releases 199.9 and 200.0, the WGS component of GenBank grew by 34,614,377,046 base pairs and by 5,907,225 sequence records.

The total number of sequence data files increased by 34 with this release. The divisions are as follows:

- BCT: 4 new files, now a total of 118 files
- CON: 11 new files, now a total of 242 files

- GSS: 5 new files, now a total of 284 files
- INV: 2 new files, now a total of 38 files
- PAT: 5 new files, now a total of 204 files
- PLN: 2 new files, now a total of 67 files
- PRI 1 new file, now a total of 47 files
- TSA 3 new files, now a total of 150 files
- VRT: 1 new file, now a total of 32 files.

For downloading purposes, please keep in mind that the GenBank flatfiles are approximately 625 GB (sequence files only). The ASN.1 data are approximately 522 GB.

More information about GenBank Release 200.0 and coming changes are available in the [release notes](#).

## Human genome annotation release 106 now available

*Tuesday, February 11, 2014*

The human (*Homo sapiens*) genome annotation has recently been updated to [annotation release 106](#). The data is now available in the Nucleotide, Protein and Gene databases, is searchable using [BLAST](#), and can be downloaded from the [FTP site](#).

The annotated assemblies include [GRCh38](#) (GCF\_000001405.26), which was published in December 2013 and represents a major, chromosome-coordinate changing update to the reference assembly (for more information, visit the [GRC](#)). Annotation release 106 also includes a re-annotation of the assemblies [CHM1\\_1.1](#) (GCF\_000306695.2) and [HuRef](#) (GCF\_000002125.1).

Some highlights of the GRCh38 annotation results (as compared to [annotation release 105](#) of the previous assembly, GRCh37.p13) include:

1. A total of 29,399 genes are predicted (an increase of 5.6%)
2. A total of 69,826 protein-coding transcripts are annotated (an increase of 3.4%)
3. The number of CDSs annotated as partial decreased from 96 to 56
4. The number of curated RefSeq transcripts with an alignment split across scaffolds decreased from 30 to 5
5. The number of protein coding genes found only on alternate loci and/or novel patches increased from 28 to 64.

There are significant improvements to gene annotation on the new GRCh38 assembly. For example:

- [SRGAP2](#): Previously split across scaffolds with an inversion
- [DPP6](#): Previously split across scaffolds
- [EPPK1](#): Previously internally partial
- [DOC2B](#): Previously 3' partial

See what other annotation runs are in progress on the [Eukaryotic Genome Annotation Pipeline status page](#).

## NCBI releases Entrez Direct, the Entrez utilities on the UNIX command line

*Thursday, February 06, 2014*

NCBI has just released Entrez Direct, a new software suite that enables users to use the UNIX command line to directly access NCBI databases, as well as to parse and format the data to create customized downloads.



For over a decade, researchers have been able to access Entrez search, retrieval and linking functions through the web interface on the [NCBI site](#) or through [Entrez Utilities \(E-Utilities\)](#), the public API for Entrez. Entrez Direct now brings Entrez to the UNIX command line by offering a set of UNIX executables that call the E-utilities directly and provide a variety of post-processing functions. Using these functions, arbitrary fields from complex record formats can be parsed and converted to simple tables suitable for importing into spreadsheets or external databases.

Entrez Direct is available as a simple [FTP download](#) and has extensive [documentation](#) on the NCBI web site. The package itself consists of a master Perl script and several UNIX shell scripts that serve as an interface to the Perl script. Therefore, Entrez Direct will run on UNIX and Macintosh computers that have the Perl language installed, and under the [Cygwin](#) UNIX-emulation environment on Windows PCs.

The [Entrez Direct documentation](#) provides several examples of how these programs may be used, and we have provided three additional examples below. Links are provided to the commands and output files.

### 1. Retrieve a set of PubMed abstracts

### 2. Download a table of start and stop coordinates for all genes on human chromosome 17 from all available assemblies.

First, “`esearch`” retrieves all current genes on human chromosome 17, and these results are passed to “`esummary`”. The resulting XML document summaries are passed to “`xtract`”, which parses the assembly and gene coordinate details and writes them to an output [file](#). The first few lines of this file are shown below.

### 3. Download a table of start and stop coordinates for all genes on human chromosome 17 from the *current* reference assembly.

We will be posting additional Entrez Direct examples and explanations on the [NCBI Insights](#) blog and in the [Entrez Direct documentation](#), and will continue to announce updates on the [utilities-announce](#) list.

For more information:

- [Entrez Direct documentation](#)
- [E-utilities documentation](#)

## Sequence Viewer updated to version 3.1

*Tuesday, February 04, 2014*

[NCBI Sequence Viewer](#) provides a graphical view of sequences and color-coded annotations on regions of sequence stored in the Nucleotide and Protein databases. Sequence Viewer has recently been updated and now has improved compatibility with Internet Explorer for users embedding Sequence Viewer on their web sites and improved performance and response time.

A full list of new features, improvements and fixes is included in the [release notes](#).

```
esearch -db pubmed -query "thyroid peroxidase genetics" | efetch
-format abstract > example1.out
```

This [command](#) searches PubMed with the query “thyroid peroxidase genetics” and then downloads the retrieved records as abstracts, which are then written to a local file.

```
esearch -db gene -query "17[chr] AND human[orgn] AND alive[prop]" |
esummary | xtract -pattern DocumentSummary -element Id -block
LocationHistType -pfx "\n" -element
AnnotationRelease,AssemblyAccVer,ChrAccVer,ChrStart,ChrStop >
example2.out
```

This [command](#) demonstrates the parsing function “xtract” that is unique to Entrez Direct.

```
7157
105   GCF_000001405.25   NC_000017.10   7590867   7571719
105   GCF_000002125.1   AC_000149.1   7484282   7465168
105   GCF_000306695.2   NC_018928.2   7600002   7580865
104   GCF_000001405.22   NC_000017.10   7590867   7571719
...
1636
105   GCF_000001405.25   NC_000017.10   61554421   61575740
105   GCF_000002125.1   AC_000149.1   56922780   56944099
105   GCF_000306695.2   NC_018928.2   61618549   61639868
104   GCF_000001405.22   NC_000017.10   61554421   61575740
...
```

This [file](#) contains data segmented for each gene and begins with a line containing the Gene ID, followed by lines with these columns: annotation release, assembly accession.version, chromosome accession.version, start coordinate, stop coordinate.

```
esearch -db gene -query "17[chr] AND human[orgn] AND alive[prop]" |
esummary | xtract -pattern DocumentSummary -element Id -block
LocationHistType -match "AssemblyAccVer:GCF_000001405.25" -pfx "\n"
-element AnnotationRelease,ChrAccVer,ChrStart,ChrStop > example3.out
```

This [command](#) is an extension of Example 2 in that it uses the “-match” option to limit the output to data from the current reference assembly (GCF\_000001405.25).

```
7157
105  NC_000017.10  7590867  7571719
1636
105  NC_000017.10  61554421  61575740
6532
105  NC_000017.10  28562985  28521336
672
105  NC_000017.10  41277499  41196311
...
```

The output [file](#) has the same format as that for Example 2, except that column 2 (assembly accession.version) is omitted.



## NCBI News, January 2014

### Human CCDS release 15 now available on web and FTP

*Monday, January 27, 2014*

The Consensus Coding Sequence (CCDS) update for Homo sapiens annotation release 105 is now available on the [CCDS website](#) and [FTP site](#). This release adds 349 new CCDS IDs to the human CCDS dataset and is based on comparative analysis of NCBI Homo sapiens annotation release 105 and [Ensembl release 74](#). The human CCDS dataset now includes 29,045 proteins that correspond to 18,683 genes.

This is the final CCDS update for human that is based on the human reference assembly GRCh37. The next CCDS update for human will be based on the updated assembly GRCh38 and is tentatively expected to be released in July 2014.

### RefSeq release 63 now available

*Tuesday, January 21, 2014*

The full RefSeq release 63 is now available with nearly 50 million records describing 37,371,278 proteins, 5,760,653 RNAs, and sequences from 33,485 different organisms.

Some important updates include the following:

**Directory name change:** The RefSeq release directory “microbial” will be removed. Two new directories, “archaea” and “bacteria” will be added. This change will appear in release 65 in May 2014.

**WGS process flow change:** WGS accessions will no longer be processed on a per-project (WGS prefix) basis. Instead, these accessions will be processed and packaged the same as non-WGS accessions. This will significantly reduce the number of files in the /complete/ and (new) /archaea/ and /bacteria/ directories. Therefore, there will no longer be a series of files named like “microbialNZ\_\*”. Instead, all WGS scaffolds will be found in concatenated files just like all other accession data. We will continue to provide a separate file for the WGS master records. This change will appear in release 65 in May 2014.

**Human Genome GRCh38 Annotation plans:** The Genome Reference Consortium released an updated assembly for the human reference genome (GRCh38) in late December 2013. NCBI annotation of the RefSeq copy of this assembly is currently in progress. We anticipate releasing annotation in early to mid-February and including it in RefSeq release 64 in March 2014.

More details about RefSeq release 63 is included in the [release statistics](#) and [release notes](#). In addition, reports indicating the accessions included in [the release](#) and [the files installed](#) are available.

### Taxonomy database now shows type material, sequences from type specimens and strains now labeled in Entrez

*Tuesday, January 21, 2014*

The naming, classification and identification of organisms traditionally relies on the concept of type material, which defines the representative examples (“name-bearing”) of a species. For larger organisms, the type material is often a preserved specimen in a museum drawer, but the type concept also extends to type bacterial strains as cultures deposited in a culture collection. Of course, modern taxonomy also relies on molecular sequence information to define species. In many cases, sequence information is available for type specimens and strains. Accordingly, the NCBI has started to curate type material from the Taxonomy database, and are using this data

to label sequences from type specimens or strains in the sequence databases. The figure below shows type material as it appears in the NCBI taxonomy entry and a sequence record for the recently described African monkey species, *Cercopithecus lomamiensis*.

Sequence from type material is particularly important because the species identification is virtually certain to be correct. The Entrez query "[sequence from type](#)"[filter] can be used to retrieve these sequence entries and can be used in combination with other queries as in the following examples.

#### By Organism

- [sequence from type](#) [filter] AND bacteria [Organism]
- [sequence from type](#) [filter] AND animals [Organism]

#### By Collection

- "[sequence from type](#)"[filter] AND collection cbs (type strains at the CBS culture collection)
- "[sequence from type](#)"[filter] AND collection mcz (type specimen at the Museum of Comparative Zoology)

#### By Author

- "[sequence from type](#)"[filter] AND hedges sb
- "[sequence from type](#)"[filter] AND Baldwin c

As shown in the figure below, "[sequence from type](#)"[filter] is also useful as an Entrez query to limit BLAST searches to reliably identified sequences, particularly when working with prokaryotes.

Stay tuned for more developments and added features coming to the Taxonomy database.

## New NCBI Insights blog: NCBI Remap tool helps you transition to newest human reference genome assembly, GRCh38

*Thursday, January 16, 2014*

NCBI's Genome Remapping Service (NCBI Remap) allows you to map annotation data from one genomic assembly to another for a selected set of organisms. This may be particularly helpful in updating your annotations for the human reference genome assembly, which has recently been updated to the new version, GRCh38. The [newest blog post on NCBI Insights](#) describes how Remap works and how it allows you to analyze data in the context of the newest genome assembly. Finally, the blog post provides links to Remap-related documentation including an overview, FAQs, and a YouTube tutorial.

## Sequence Viewer PDF rendering available - YouTube video tutorial

*Tuesday, January 14, 2014*

NCBI has created a YouTube [video tutorial](#) showing you how to generate a PDF rendering of your Sequence Viewer display. The short clip takes you through the downloading process, and shows you what you can do with your file after creating a PDF.

## Genome Workbench Update 2.7.12

*Tuesday, January 14, 2014*

### Cercopithecus lomamiensis


**Taxonomy ID:** 1191211  
**Genbank common name:** lesula  
**Inherited blast name:** primates  
**Rank:** species  
**Genetic code:** [Translation table 1 \(Standard\)](#)  
**Mitochondrial genetic code:** [Translation table 2 \(Vertebrate Mitochondrial\)](#)  
**Other names:**  
 synonym: **Cercopithecus sp. ASB-2012**  
 authority: **Cercopithecus lomamiensis Hart et al. 2012**

type material: YPM MAM 14192  
 type material: YPM MAM 14191  
 type material: **YPM MAM 14189**  
 type material: **YPM MAM 14080**  
 type material: YPM 14192  
 type material: YPM 14191  
 type material: YPM 14189  
 type material: YPM 14080

Entrez records	
Database name	Direct links
Nucleotide	<a href="#">8</a>
Protein	<a href="#">4</a>
PubMed Central	<a href="#">1</a>
Taxonomy	<a href="#">1</a>

LOCUS	JN106060	4688 bp	DNA	linear	PRI 05-MAR-2013
DEFINITION	Cercopithecus lomamiensis isolate ME408 X chromosome intergenic region genomic sequence.				
ACCESSION	JN106060				
VERSION	JN106060.1 GI:387865320				
KEYWORDS	.				
SOURCE	Cercopithecus lomamiensis (lesula)				
ORGANISM	<a href="#">Cercopithecus lomamiensis</a> Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Cercopithecidae; Cercopithecinae; Cercopithecus.				
REFERENCE	1 (bases 1 to 4688)				
AUTHORS	Hart, J.A., Detwiler, K.M., Gilbert, C.C., Burrell, A.S., Fuller, J.L., Emetsu, M., Hart, T.B., Vosper, A., Sargis, E.J. and Tosi, A.J.				
TITLE	Lesula: A New Species of Cercopithecus Monkey Endemic to the Democratic Republic of Congo and Implications for Conservation of Congo's Central Basin				

**YPM Mammalogy - Online Catalog**



17 Jan 2014 15:49:31

[Terms of Use](#)

Items 1-1 of 1 matching items.

[New Search](#)

<b>YPM MAM 014080</b>	
Taxon Name.....	Cercopithecus lomamiensis J. Hart, Detwiler, Gilbert, Burrell, Fuller, Emetsu, T. Hart, Vosper, Sargis, and Tosi, 2012 - HOLOTYPE
Locality.....	Africa, Democratic Republic of Congo, Orientale Province, Tshopo District, Lohumonoko, shot. Elev. 470 m.
Latitude.....	-1.02237 24.42368
Collected.....	M. Emetsu, 12 Aug 2008.
Higher Ranks.....	Primates; Cercopithecidae
Common Name.....	Lesula
Other Attributes.....	TISSUE: SKIN; male male; adult adult; skin, fat tissue sample at at New York University; skeleton (skull only); TYPE: SKELETON

ORIGIN	1 tttcctttgc agggacatgg atggagtgg aagtcattat cctcagcaat ctaatgcagg 61 aattgaaac caaacaccac atgttctcac ttataaatgg gagctgaatt atgagaacac
--------	---

**Figure 1.** Type material information as it appears in the NCBI Taxonomy database (upper left) and nucleotide database (lower right) for *Cercopithecus lomamiensis*. Both records refer to the type material housed in the Yale Peabody Museum Mammalogy collection (lower left, YPM MAM 140180).

Genome Workbench 2.7.12 has been released. The update includes several new features like a faster Tree Renderer, redesigned data import and an improved GFF format reader. The [release notes](#) include more information on features, fixes and improvements.

**Choose Search Set**

**Database**  
 Representative genomes only  All genomes

**Organism**  
*Optional*  
enterobacteria (taxid:91347)  Exclude +  
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

**Entrez Query**  
*Optional*  
Sequence from type[Filter]  
Enter an Entrez query to limit search

**Figure 2.** The “Choose Search Set” section of the [Microbial Genomes BLAST](#) page. The optional Organism limit, enterobacteria, and the Entrez query ‘Sequence from type[Filter]’ restricts the search to sequences from enterobacteria type strains.

## VAST+ released: Find similar 3D structures for macromolecular complexes

Thursday, January 09, 2014

VAST+ is a new tool designed to identify macromolecules that have similar 3-dimensional structures with an emphasis on finding similar macromolecular complexes. The similarities are calculated using purely geometric criteria without regard to sequence similarity, and therefore can identify distant homologs.


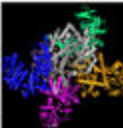
This new tool is built upon the original [Vector Alignment Search Tool \(VAST\)](#) and expands the capabilities of that program by taking into account the [biological unit](#) (“biounit”) of each structure, not just individual protein chains or their substructures.

A recent [publication](#) provides detailed information about the VAST+ algorithm. In addition, the extensive VAST+ help document includes a [comparison of original VAST and VAST+](#), as well as [examples of how can this tool can be used](#) to learn more about proteins.

*Please note that in order to view the 3D superpositions of similar biological units, you must install the most recent version of the NCBI molecular viewing software, [Cn3D 4.3.1](#).*



**Glutamate Dehydrogenase** query structure

MMDB ID: 11869 PDB ID: 1B26  
 Biological unit 1: hexameric  
 Source Organism: *Thermotoga maritima*  
 Number of protein: 6 (GLUTAMATE DEHYDROGENASE ▼)

---

**Similar Structures** Original VAST


▼ Display Filters 262 structures displayed PDB ID or MMDB ID Search Within Results

PDB ID	Description	Aligned Proteins	RMSD	Aligned Residues	Sequence Identity
1	1B3B Thermotoga Maritima Glutamate De...Mutant N97d...	6	0.30Å	2454	99%
2	2TMG Thermotoga Maritima Glutamate De...Mutant S128r...	6	0.59Å	2448	99%
3	1BVU Glutamate Dehydrog..from Thermococcus Litoralis	6	1.45Å	2454	53%
4	1GTM Structure of Glutamate Dehydrogenase	6	1.52Å	2454	55%

**Aligned Molecules** detailed view of match Matched Biological Unit


Query Biounit

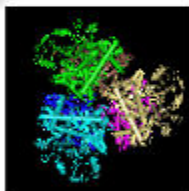
MMDB ID: 11869 (PDB ID: 1B26)



Matched Biounit

MMDB ID: 5314 (PDB ID: 1GTM)





MMDB ID: 5314 PDB ID: 1GTM  
 Biological unit 1: hexameric  
 Source Organism: *Pyrococcus furiosu...*  
 Number of proteins: 6  
 Number of chemicals: 12  
 Aligned residues: 2454  
 Sequence identity: 55%  
 Structure identity (RMSD): 1.52Å

5	2YFQ Crystal structure...Glutamate Dehy...Peptoniphilus...	6	1.53Å	2325	46%
6	3AOG Crystal structure...Glutamate Dehy...Thermus Ther...	6	1.59Å	2448	55%

**partial biounit matches**

34	1BXG Phenylalanine Dehydrogenase Structure In Tern...	2	4.07Å	662	27%
35	3VPX Crystal Structure Of Leucine Dehydrogenase...	2	4.40Å	637	23%

Figure. VAST+ search results for *Thermotoga maritima* Glutamate Dehydrogenase (PDB ID 1B26) with detailed view of a match to 1GTM.

### For more information:

- [VAST+ home page](#)
- [Publication in Nucleic Acids Research](#)
- [Vast+ examples](#)
- [Cn3D home page](#)

## NCBI Insights blog: A Librarian's Guide to NCBI - an intensive training course for medical librarians to be offered April 2014

Wednesday, January 08, 2014

The NCBI in partnership with the National Library of Medicine Training Center (NTC) will offer the Librarian's Guide to NCBI course on the NIH campus in April 2014. This will be the second presentation of the course; it was previously offered in the spring of 2013 (NCBI Insights April 11 and May 6, 2013). After the course, we will

post lecture slides and hands-on practical exercises on the education area of the NCBI FTP site and video tutorials of the course lectures will be available on the NCBI YouTube channel. [Materials](#) from the 2013 course are available, as well as lecture videos for the [expression module](#). More information, including prerequisites, is available in the [newest NCBI Insights blog post](#).

## Mouse genome annotation release 104 available

*Wednesday, January 08, 2014*

The mouse (*Mus musculus*) genome annotation has recently been updated to [annotation release 104](#) and is now available in the Nucleotide, Protein sequence and Gene databases, is searchable using [BLAST](#), and can be downloaded from the [FTP site](#).

Mouse annotation release 104, based on the sequence assemblies GRCm38.p2 ([GCF\\_000001635.22](#)) and Mm\_Celera ([GCF\\_000002165.2](#)), identifies a total of 35,389 genes, as well as 100,581 transcripts on GRCm38.p2. RNA-Seq data from 31 distinct BioSample accessions were aligned to assist in gene prediction. More statistics are available in the [annotation report](#).

See what other annotation runs are in progress on the [Eukaryotic Genome Annotation Pipeline status page](#).

## BLAST+ 2.2.29 now available

*Tuesday, January 07, 2014*

Stand-alone BLAST version 2.2.29+ is now available for download from the [FTP site](#). BLAST 2.2.29+ provides a number of important improvements and bug fixes. Some improvements include improved blastn batch query performance, source releases build optimized multi-thread binaries by default, and improved multithreading by better dividing the BLAST database among threads. The [BLAST Release notes](#) lists more upgrades and fixes.

## NCBI News, December 2013

### New human genome assembly (GRCh38) released!

*Tuesday, December 24, 2013*

On December 24th, the [Genome Reference Consortium](#) (GRC) submitted a new assembly for the human genome (GRCh38) to [GenBank](#). These data are now available in the Assembly database with accession [GCA\\_000001405.15](#) and are also available on the [FTP site](#). Please note the GRC provides these assemblies as unannotated sequences.

Now that the GRC sequences are available in GenBank, our [Reference Sequence \(RefSeq\)](#) Genome Annotation Group has downloaded these sequences and has begun processing them using our [eukaryotic annotation pipeline](#). The resulting human chromosome sequences will continue to have the RefSeq accessions [NC\\_000001-NC\\_000024](#), but their versions will increment as the update to the GRCh38 assembly includes a sequence change for all chromosomes. The process of annotating the human genome generally takes about 2 weeks. When this is complete, we will incorporate these sequences into various analysis and display tools, such as [human genome BLAST](#), [NCBI Remapping Service](#), and various genome viewers. Thus, at the end of this process each chromosome will be represented by both an unannotated sequence in GenBank (the original GRC data) and an annotated sequence in the RefSeq collection.

Please check back frequently for updates on the [NCBI News](#) and our social media sites ([NCBI Twitter Channel](#), [NCBI Facebook Page](#), [NCBI Announce RSS Feed](#), [NCBI Announce Email ListServ](#)) as this process unfolds.

In addition, we have a series of posts on the [NCBI Insights Blog](#) site on topics such as how NCBI processes genome annotations, a tip to remap annotations from older assemblies to GRCh38, and highlighting some loci that have changed significantly in the new assembly.

### Annotation reports now generated for recently annotated organisms

*Monday, December 23, 2013*

The NCBI Eukaryotic Genome Annotation Pipeline now publishes a report to accompany each new annotation. This report provides statistics on the annotation products, such as the number of genes, the number and length of coding and non-coding transcripts, and the number of transcripts per gene. It also presents statistics on the protein and transcript alignments that were used by the gene prediction process.

See the annotation reports generated for rat and potato:

[http://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Rattus\\_norvegicus/104](http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rattus_norvegicus/104)

[http://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Solanum\\_tuberosum/100](http://www.ncbi.nlm.nih.gov/genome/annotation_euk/Solanum_tuberosum/100)

Annotation reports for other recently annotated organisms can be accessed from the [Eukaryotic Genome Annotation Pipeline status page](#).

Learn more about how the eukaryotic genome annotation pipeline works: <http://www.ncbi.nlm.nih.gov/news/12-17-2013-new-handbook-chapters-genome-annotation-pipelines/>

### Meet PubMed Commons: The new comments forum in PubMed

*Thursday, December 19, 2013*

If you are one of the millions of people who visit PubMed today, be on the look-out for something different. On each abstract page, there's now a section called [PubMed Commons](#). It's a forum for scientific discussion on publications open to any authors in the world's largest biomedical literature database.

Read more at [PubMed Commons Blog](#).

## Rat genome annotation release 104

*Wednesday, December 18, 2013*

The rat (*Rattus norvegicus*) genome annotation has recently been updated to [annotation release 104](#) and is now available in the Nucleotide, Protein sequence and Gene databases, is searchable using [BLAST](#), and can be downloaded from the [FTP site](#).

Rat annotation release 104, based on the sequence assemblies Rnor\_5.0 (GCF\_000001895.4, reference) and Rn\_Celera (GCF\_000002265.2), identifies a total of 31,451 genes. In addition, 64,745 transcripts were identified on Rnor\_5.0. A new annotation pipeline step in this update is the alignment of RNA-Seq data from 85 distinct BioSample accessions to assist in gene prediction.

More statistics are available in the [Rat Annotation Release 104 Report](#).

See what other annotation runs are in progress on the [Eukaryotic genome annotation pipeline status page](#).

## New NCBI Handbook chapters: Eukaryotic and prokaryotic genome annotation pipelines

*Tuesday, December 17, 2013*

In order to increase the utility of genomic information, we provide gene annotation and other features on Reference Sequence (RefSeq) genome records. Genome annotation is a multi-step process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons, and other mobile elements.

Depending upon the genome, the identification of key genomic features and their locations on RefSeq genome records are provided by outside sources (the submitter's annotation copied from the GenBank genomic sequence records or curated annotation provided by a model organism database, like [FlyBase](#) or [WormBase](#)), or are generated by annotation pipelines developed at NCBI specifically for eukaryotic or for prokaryotic genomes.

An overview of each pipeline is available in our web documentation. In addition to web documentation of our [eukaryotic genome annotation pipeline](#) and [prokaryotic genome annotation process](#).

Our newest NCBI Handbook Chapters on the eukaryotic and prokaryotic annotation pipelines describe the processes in greater detail, including information on algorithms, history, annotation standards and special considerations like multiple annotation assemblies:

- [Eukaryotic Genome Annotation Handbook chapter](#)
- [Prokaryotic Genome Annotation Pipeline Handbook chapter](#)

We also provide [eukaryotic genome annotation policies](#) and the [status of genomes in the current pipeline](#), as well as information about [prokaryotic genome annotation standards](#).

## Sequence Viewer has been updated

*Tuesday, December 17, 2013*

NCBI [Sequence Viewer](#) provides a graphical view of sequences and color-coded annotations on regions of sequence stored in the Nucleotide and Protein databases. Sequence Viewer has recently been updated and now has better loading and management of uploaded custom tracks, improved naming of downloaded files including sequence ranges and file extensions, and easier embedding in external Web sites.

A full list of new features, improvements and fixes is available at: <http://www.ncbi.nlm.nih.gov/tools/sviewer/release-notes/>

## GenBank release 199 now available

*Monday, December 16, 2013*

[GenBank Release 199](#) is now available through NCBI's Entrez and BLAST services.

Release 199.0 (12/10/2013) has 169,331,407 non-WGS, non-CON records containing 156,230,531,562 base pairs of sequence data. In addition, there are 133,818,570 WGS records containing 556,764,321,498 base pairs of sequence data.

During the 54 days between the close dates for GenBank Releases 198.0 and 199.0, the non-WGS/non-CON portion of GenBank grew by 1,054,036,863 base pairs and by 996,011 sequence records. During the same period, 494,249 records were updated; an average of 27,597 non-WGS/non-CON records per day were added and/or updated. Between releases 198.0 and 199.0, the WGS component of GenBank grew by 20,922,153,757 base pairs and by 3,615,365 sequence records.

The total number of sequence data files increased by 18 with this release. The divisions are as follows:

- BCT: 2 new files, now a total of 114
- CON: 5 new files, now a total of 231
- ENV: 2 new files, now a total of 67
- EST: 1 new file, now a total of 475
- GSS: 1 new file, now a total of 279
- PAT: 2 new files, now a total of 199
- PLN: 1 new file, now a total of 65
- TSA: 2 new files, now a total of 147
- VRL: 2 new files, now a total of 29

For downloading purposes, please keep in mind that the GenBank flatfiles are approximately 618 GB (sequence files). ASN.1 data are approximately 508 GB.

**Change to Accession Format:** As mentioned in the GenBank Release 198.0 news story, CON-division WGS scaffolds will have new format accession numbers. For an example of the new accession format, please see Section 1.3.2 of the [GenBank release notes](#). We do not currently plan to update existing records with the new accession format.

## NCBI Video: Submitting manuscripts on NIHMS

*Thursday, December 05, 2013*

NCBI's latest [YouTube video](#) takes you through the manuscript submission process on the NIH Manuscript Submission System (NIHMS), step-by-step. NIHMS enables publishers, authors, and principal investigators to submit manuscripts for processing and archiving in PubMed Central.

## **PMCID - PMID - Manuscript ID - DOI Converter Upgraded**

*Tuesday, December 03, 2013*

We have upgraded the [PMCID - PMID - Manuscript ID - DOI Converter](#). The updated ID Converter API allows you to convert IDs for publications referenced in PubMed and PMC.

The ID Converter tool allows you to convert IDs for publications referenced in PubMed and PMC. You can also cross-reference Open Access NIH Manuscript Submission IDs (NIHMS) and Digital Object Identifiers (DOIs) often used by publishers. For example, these identifiers refer to the same publication:

PMCID: PMC3702208

PMID: 24288678

NIHMS: NIHMS518180

DOI: 10.1007/s00213-013-3057-1

This tool uses an underlying web service, which is also publicly available for those needing programmatic access to this data. For more information, see the [ID Converter API documentation](#).

## NCBI News, November 2013

### NCBI Insights blog post: Creating custom BLAST databases

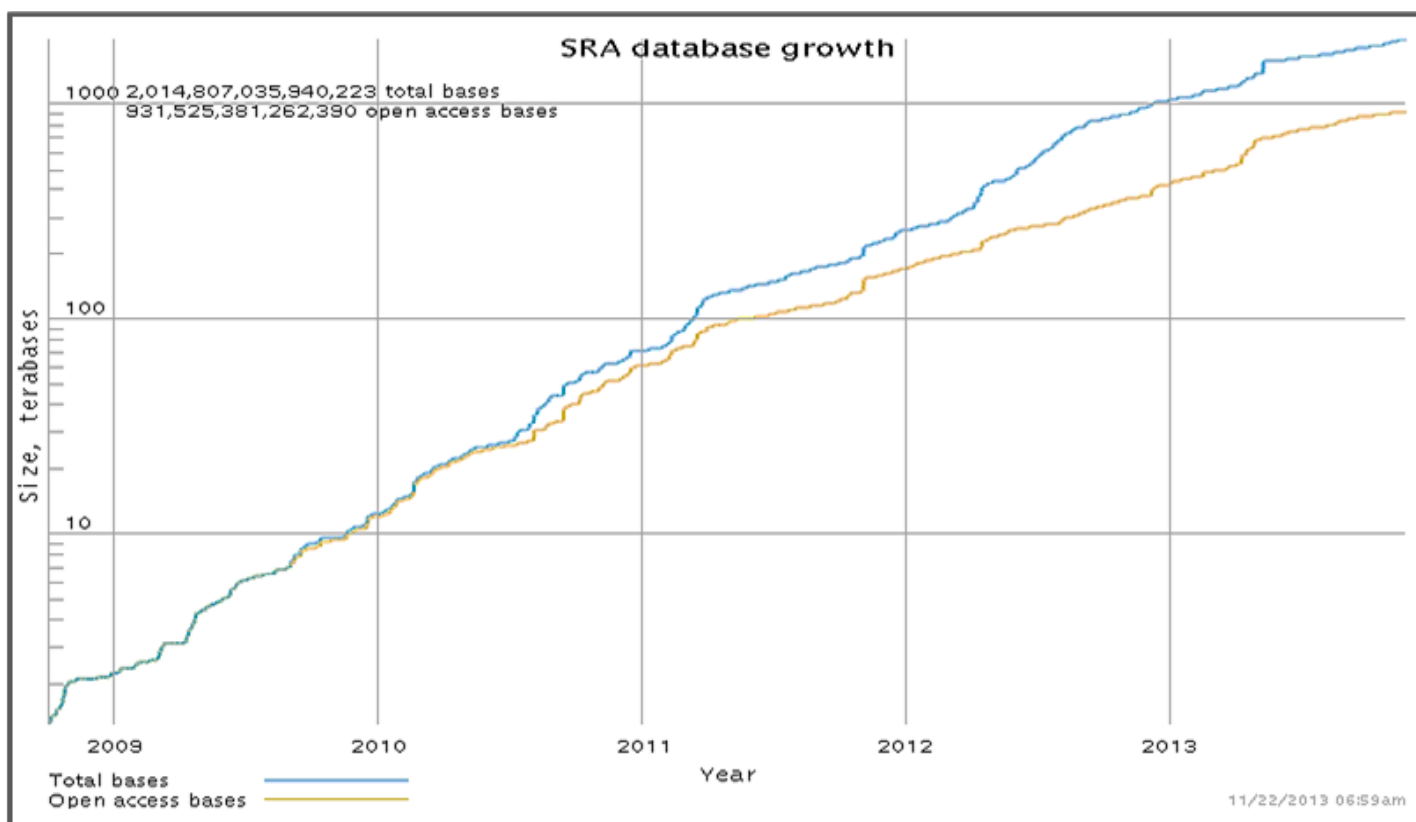
Tuesday, November 26, 2013

The latest [Quick Tip blog](#) focuses on creating custom BLAST databases, which is an easy way to speed up BLAST analysis. The blog post takes you through the process, from selecting the appropriate parent database to manipulating the filters and Entrez Query.

### SRA milestone: Over 2 petabases of sequence data

Monday, November 25, 2013

The [Short Read Archive \(SRA\)](#) now contains more than 2 petabases of high-throughput sequence data. One petabase of data is open access, while the rest are sequences from 40,000 individuals who have participated in human clinical studies catalogued in dbGaP.



### Planned change in bacterial strain-level information management

Thursday, November 21, 2013

Please be aware that there is an upcoming change (January 2014) in how NCBI manages organism strain information. Due to significant increases in the volume of strain-specific sequencing, we are changing our management of strain information.

Next generation sequencing has already changed the way microbial genomes are being used. The scope of microbial sequencing projects has shifted from a single isolate representing an organism to multi-isolate and multi-species projects representing microbial communities. Consequently, in the first nine months of 2013 the sequences of more than 6000 prokaryotic genomes were released by INSDC (DDBJ/ENA/GenBank).

NCBI is introducing several changes in prokaryotic genomes and related resources such as Assembly, BioProject, BioSample, and Taxonomy that will affect your submissions, data downloads, analysis tools, and parsers.

### Taxonomy

Assigning strain-level TaxID will be discontinued in January 2014 because curation of strain-level TaxIDs will not remain possible under such growth. However, the thousands of existing strain-level TaxIDs will remain, and we will continue to add informal strain-specific names for genomes from specimens that have not yet been identified to the species level, e.g. “*Rhizobium* sp. CCGE 510” and “*Micromonas* sp. RCC299”. The strain information will continue to be collected and displayed.

### BioSample

Submitters of genome sequences will be required to register sample meta-data in the BioSample database for each organism that they are sequencing. The BioSample submission will include the strain information and other metadata, such as culture collection and isolation information, as appropriate. The BioSample accession will be a link on the GenBank records, and the GenBank records themselves will display the strain in the source information.

### BioProject

Submitters of genome sequences are already required to register meta-data about the research project in the BioProject database. We no longer require a one-to-one relationship between a BioProject accession and a genome. Instead, a research effort examining multiple strains of a species or multiple species of drug-resistant bacteria, for example, could be registered as a single BioProject.

### Assembly

Each genome assembly is loaded to the Assembly database and assigned an Assembly accession. The Assembly accession is specific for a particular genome submission.

## What defines a genome?

A BioProject ID or accession cannot be used to define a single genome, since many may belong to a multi-isolate or multi-species project. Furthermore, a TaxID can no longer reliably define an individual genome since unique TaxIDs will not be assigned for individual strains and isolates. The collection of DNA sequences of an individual sample (isolate) will be represented by a BioSample accession and if raw sequence reads are assembled and submitted to GenBank they will get a unique Assembly accession. The Assembly accession is specific for a particular genome submission. For example, sequence data generated from a single sample (with a BioSample accession) could be assembled with two different algorithms and so have two sets of GenBank accessions, each with its own Assembly accession.

For example, BioProject [PRJNA203445](#) is a multi-species project with multiple strains and isolates of different food pathogens. Each isolate has its own BioSample accession and each assembled genome has its own Assembly accession. This BioProject includes an isolate of *Listeria monocytogenes* (TaxID 1639, strain R2-502) which was registered as BioSample SAMN02203126, and its genome is represented in GenBank records CP006595-CP006596, which are tracked as a group in the Assembly database under accession GCA\_000438585.



## FTP files

Genome text reports on the [FTP site](#) have been modified to include the BioSample and Assembly accessions. These two columns were added at the end of the tables to minimize problems for existing parsers. Initially, not all assemblies will have a BioSample accession because we are still in the process of back-filling BioSamples for genomes.

**These changes will occur in January 2014. We will be releasing more information as the date approaches.**

## Exploring next-gen sequencing experiments with SRA-BLAST

*Tuesday, November 19, 2013*

NCBI's [SRA-BLAST](#) has two new features that substantially improve its ability to explore the myriad of next-gen sequencing studies available from NCBI's Sequence Read Archive (SRA).

First, we have expanded SRA-BLAST to include technologies beyond 454, which means that more than 100,000 experiments are currently available through this service. Except for two types of data -- human data with [controlled access](#) that are only available through [dbGaP](#) and reads stored as alignment references ([cSRA format](#))-- experiments that can be searched through BLAST now include data from all current next-sequencing technologies that produce read lengths long enough (approximately 100 bases) for general BLAST searching.

The other new feature is that SRA-BLAST now offers two different ways of finding data sets to search. The BLAST service itself provides an autocomplete feature under "Choose Search Set" that finds matches to experiment, study and run accessions as well as text from experiment descriptions (Figure 1, top panel). You can now also use the Entrez SRA system to identify experiments of interest and load these as BLAST databases in SRA BLAST through the 'Send to' menu from the SRA search results (Figure 1, bottom panel).

## Example

There are many studies in SRA in which the experiments form a series with varying conditions. We can use SRA BLAST to examine changes in the data under these different conditions. For example, let's look at study [SRP001041](#), which contains metagenomic sequence data from a depth profile of the North Pacific subtropical gyre from station ALOHA with samples from 25, 75, 110 and 500 meters (SRX007372, SRX007369, SRX007370, and SRX007371). (See the [Hawaii Ocean Time Series website](#) for details on the sampling location and projects there.) We can easily find these experiments using the following search in SRA, which retrieves the four DNA-based experiments from the different depths.

```
SRP001041 AND dna data[Filter]
```

Using SRA BLAST we can profile the abundance of *Prochlorococcus*, a tiny prokaryotic photosynthetic organism that plays a large role in carbon cycling in the open ocean (reviewed in Partensky F, Hess WR, and Vault D, 1999, PMID: [98958](#)).

After selecting one of the four experiments, for example the one from 25 meters, we can load it as a BLAST database through the 'Send to' menu (Figure 2).

We can then use the coding region of one of the genes involved in *Prochlorococcus* photosynthesis to measure the abundance of these organisms at different depths. The photosystem I P700 chlorophyll a apoprotein A1 (*psaA*) region ([CP000551.1|:c1473975-1471672](#)) *Prochlorococcus marinus* str. AS9601 serves as a useful marker query. The following link will set up a BLAST search with the gene region against the 25-meter data.

[Set up SRA-BLAST](#)

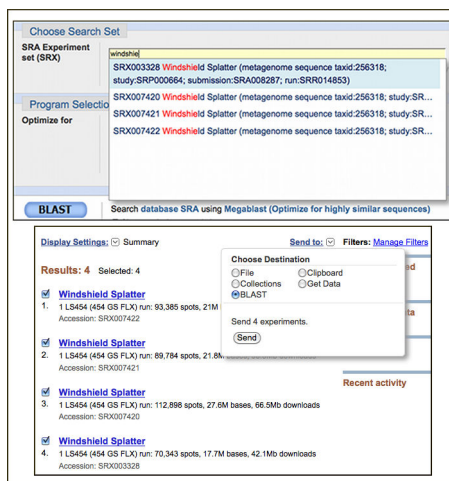


Figure 1. Two ways to find and select SRA experiments to search with the SRA BLAST service shown with the Windshield Splatter Metagenome experiments (Kosakovsky et al. 2009, PMID:2775585). Top panel. The autocomplete 'Choose Search Set' database selector matches SRA identifiers (experiment, study, submission, run) as well as text in the title. Bottom panel. The 'Send to' menu in Entrez SRA database can set selected search results as a BLAST database.

Figure 3 shows the BLAST results for the *psaA* query at different depths indicating the decline in abundance of the organisms with depths below 100 meters.

The SRA-BLAST service combined with the new 'Send to' feature in the Entrez SRA database provides a convenient and interesting way to explore the many datasets now in NCBI's Sequence Read Archive.

## NCBI Insights blog post: Saved Searches and E-mail Alerts

Monday, November 18, 2013

As part of the My NCBI service, PubMed and other Entrez databases allow users to save searches and then receive regular e-mail alerts about new records retrieved by that search. Please see the new [NCBI Insights blog post](#) for details about setting up these searches and alerts.

For more information, see the following:

- [PubMed Help: Saved Searches](#)
- [My NCBI Help: Saved Searches](#)

## RefSeq release 62 now available

Monday, November 18, 2013

The complete RefSeq release 62 is now available with nearly 50 million records describing more than 36,036,343 proteins and 5,178,509 transcripts from 31,646 different organisms. More details about RefSeq release 62 are in the [release statistics](#) and the [release notes](#).

## NCBI's Eukaryotic Annotation Pipeline has now annotated the genomes for 100 different organisms!

Tuesday, November 12, 2013

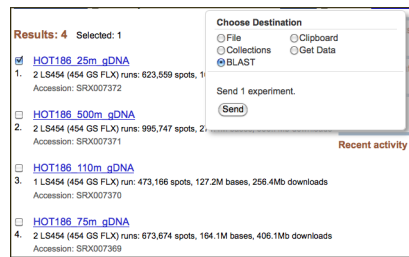


Figure 2. Entrez SRA summaries of four metagenomic experiments from a vertical profile of the North Pacific Ocean. One or more selected experiments may be set as BLAST databases for searching through the 'Send to' menu.

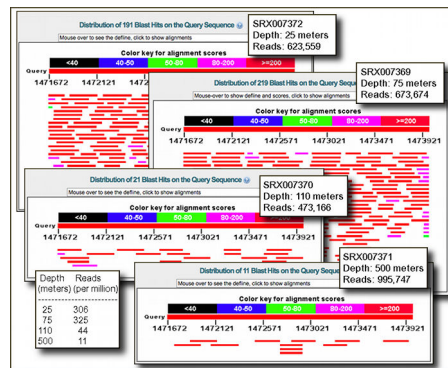


Figure 3. SRA-BLAST results showing graphic overviews for searches against the metagenomic data from differing depths in the North Pacific showing the sharp decrease in the abundance of the *Prochlorococcus* *psaA* sequence below 100 meters. The inset table at the bottom left shows read counts that match the *psaA* query sequence normalized by the number of reads in each experiment.

NCBI began annotating eukaryotic genomes in 2000. We have now completed the genome annotation for 100 different organisms, including 50 mammals, 25 other vertebrates, 17 invertebrates and 8 plants. Among these 100 organisms, 47 were annotated for the first time in 2013. The lucky 100th organism is the Chinese alligator (*Alligator sinensis*).

Recent improvements in the [Eukaryotic Genome Annotation Pipeline](#) have not only increased the throughput but have also improved the quality of the annotation produced. For example, incorporation of RNASeq data for use in gene prediction has permitted the annotation of organisms with little traditional transcript or protein sequence. View [annotation runs recently completed or in progress](#).

Data produced by the Eukaryotic Genome Annotation Pipeline is available in the [Reference Sequences \(RefSeq\)](#) collection, [BLAST](#) non-redundant and organism-specific databases, [Gene](#) database, and on the [NCBI FTP site](#).

**Need a public genome annotated? Make a request!**

## NCBI's 25th Anniversary and The Jim Gray eScience Award

Tuesday, November 05, 2013

November 2013 marks 25 years since the founding of the National Center for Biotechnology Information (NCBI).

In honor of NCBI's 25th anniversary, United States Senator Ben Cardin read a statement into the Congressional Record recognizing years of service in providing access to biomedical and genomic information to enhance the world's science and health.



In addition, on November 1st, an awards and recognition program was held to commemorate this occasion. At this event, Tony Hey, Ph.D., Vice President of Microsoft Research, presented NCBI Director David Lipman, M.D., with the [Jim Gray eScience Award](#) which recognizes outstanding contributions to the field of data-intensive computing in the pursuit of open, supportive, and collaborative research models.

For more information, see this [Microsoft Research Connections Blog post](#).

The NCBI awards program also featured presentations by:

- Michael M. Gottesman, M.D., NIH Deputy Director for Intramural Research - Introductory Remarks
- Donald A.B. Lindberg, M.D., Director of the NLM - Recollections on the origins of the NCBI
- Sir Richard J. Roberts, Ph.D, Chief Scientific Officer of New England Biolabs - Keynote Address: "A personal recollection of GenBank and NCBI"

## New SNP data available for several organisms!

Monday, November 04, 2013

New SNP data (build 139) is now available on the [web](#) and in [FTP files](#) for several organisms, including gorilla, horse, dog, sheep, rabbit, opossum, platypus, wild turkey, zebra finch, tomato, grape and aspergillus.

## Update on PubMed Commons' comments in the early pilot phase

Friday, November 01, 2013

[PubMed Commons](#) is a new system that enables researchers to share their opinions about scientific publications indexed in the PubMed database. Participation in PubMed Commons requires users with My NCBI accounts to join before they can view or add comments.

As of November 1, 2013, there are about 1,000 people signed up in the Commons and in just four days of public access the amount of comments on PubMed records doubled to over 200.

Approximately a third of the first ~200 comments included critique or pointed to other studies or reviews with the potential to change people's interpretations or conclusions. Some authors posted corrections or changed their own conclusions in the light of others' subsequent work. Authors also used PubMed Commons to update people



Tony Hey, Ph.D., Vice President of Microsoft Research, presents the Jim Gray eScience award to David Lipman, M.D., Director of the NCBI.



Michael Gottesman, M.D., NIH Deputy Director for Intramural Research, Sir Richard Roberts, Ph.D., Chief Scientific Officer of New England Biolabs and David Lipman, M.D., Director of the NCBI.

on their work – including links to databases that have moved, providing contextual information and backstories as well as new, relevant work.

Many PubMed Commons participants took the opportunity to add links to relevant papers and data, sometimes in the non-PubMed academic literature or data repositories – including complete datasets, data re-analyses, blog posts and full text pre-prints of the article.

Around half of the comments were principally discussion, developing lines of thoughts and raising or asking questions and there has already been some interesting back and forth between PubMed Commons participants interested in an issue and authors of the PubMed records.

**For more information, please see:**

[PubMed Commons Homepage](#)

NCBI Insights Blog Posts:

- ["PubMed Commons: A New Forum for Scientific Discourse"](#)
- ["Early Developments in the PubMed Commons Pilot"](#)
- ["Joining PubMed Commons: A Step-by-step Guide"](#)



## NCBI News, October 2013

### Human CCDS release 14 is now available in the Gene database

*Tuesday, October 29, 2013*

The Consensus Coding Sequence (CCDS) update for Homo sapiens annotation release 105 was released this week. The new CCDS data is available in the [CCDS web site](#) and [FTP site](#). In addition, this update is now reflected in [relevant Gene database records](#).

The [Consensus CDS \(CCDS\) project](#) is a collaborative effort (groups from NCBI, EBI, Sanger and UCSC) to identify a core set of consistently annotated, high quality human and mouse protein coding regions. For this update, the NCBI, Ensembl, and Sanger (Havana) annotations of the most updated human reference genome (GRCh38.p13 assembly, NCBI annotation release 105, [Ensembl annotation release 73](#)) were analyzed.

CCDS release 14 includes a total of 28,694 CCDS IDs that correspond to 18,673 GeneIDs. This update adds 978 new CCDS IDs, and adds 74 Genes into the human CCDS set.

### New NCBI Insights Blog Post: Joining PubMed Commons - A step-by-step guide

*Wednesday, October 23, 2013*

[PubMed Commons](#) is a new system that enables researchers to share their opinions about scientific publications indexed in the PubMed database. Participation in PubMed Commons requires users with My NCBI accounts to join before they can view or add comments. A new [NCBI Insights Blog post](#) describes how to join PubMed Commons.

**For more information, please see:**

[PubMed Commons Homepage](#)

["Joining PubMed Commons: A Step-by-step Guide"](#)

### GenBank Release 198.0 is Available

*Tuesday, October 22, 2013*

The new release for [GenBank](#) is now available via [FTP](#), as well as in the [Nucleotide database](#) and [BLAST services](#). *Please note* that delivery of this release missed the normal target date of October 15th due to a partial shutdown of the United States government which impacted NCBI operations. When the shutdown ended on October 17th, we expedited release processing and delivered 198.0 only a week later than usual. Our apologies for the delay.

Release 198.0 (10/17/2013) 168,335,396 non-WGS, non-CON records which were comprised of 155,176,494,699 basepairs of sequence data. In addition, there were 130,203,205 WGS records containing 535,842,167,741 basepairs of sequence data.

During the 60 days between the close dates for GenBank Releases 197.0 and 198.0, the non-WGS/non-CON portion of GenBank grew by 983,573,688 basepairs and by 1,039,556 sequence records and the WGS component of GenBank grew by 35,421,755,076 basepairs and by 5,391,185 sequence records.

The total number of sequence data files increased by 25 with this release, with the divisions that expanded in file number:

- BCT = 6 new files, now a total of 112

- CON = 11 new files, now a total of 226
- ENV = 3 new files, now a total of 65
- INV = 1 new file, now a total of 36
- GSS = 5 new files, now a total of 278
- PAT = 2 new files, now a total of 197
- PLN = 1 new file, now a total of 64
- VRL = 1 new file, now a total of 27

For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 613 GB (sequence files only).

**Upcoming Change:** As of the December 2013 GenBank release, new CON-division WGS scaffolds will have a new accession format.

Prior to this date, WGS scaffolds constructed from WGS contigs were labeled with a '2+6' accession number format with two leading alphabetic characters followed by six digits. For example, AABR00000000.

The new accession format for newly-processed records will mirror that of the underlying WGS contigs:

- 4 letter WGS project code
- 2 digit assembly-version number
- "S" (for 'scaffold')
- Six or seven digits

For example, AABR06S000001 and AABR06S112651.

We do not currently plan to update existing records with the new accession format, but only ones that are newly-processed beginning with the 199.0 GenBank release.

For additional release information, see the [Release Notes](#) and README files in individual directories.

## PubMed Commons is now live!

*Tuesday, October 22, 2013*

NCBI has released [PubMed Commons](#), currently in pilot phase, which is a new system that enables researchers to share their opinions about scientific publications indexed in the PubMed database. This is intended to be a forum for open and constructive criticism and discussion of scientific issues. A [new NCBI Insights Blog post](#) provides more information and explains how researchers can join in!

**For more information, please see:**

[PubMed Commons Homepage](#)

[NCBI Insights Blog post: "PubMed Commons - a new forum for scientific discourse"](#)

## NCBI Staff will be attending the ASHG 2013 Meeting

*Monday, October 21, 2013*

The [2013 National Meeting for the American Society for Human Genetics \(ASHG\)](#) will be held from October 22nd through the 26th in Boston, MA. NCBI Staff members will be displaying Posters, presenting a workshop and attending the NCBI Booth to answer questions, participate in community dialog, and gain feedback from users.



**Wednesday evening at 6pm in room 102 - Dr. Peter Cooper will be hosting a free workshop on "[Discovering Biological Data at the NCBI](#)".**

- *This workshop will show how to use the NCBI Entrez system to perform searches and find related molecular data starting with a list of reviewed human genes.*
- **Abstract:** The National Center for Biotechnology Information (NCBI) is the premier repository for biological information in the U.S. and is the primary archive for submitter-provided data through resources such as the Sequence Read Archive (SRA), GenBank, GEO, dbSNP, dbVar and dbGaP. Resources at NCBI use the Entrez system to search various databases and display records. This workshop will give a basic introduction to using the Entrez system to perform searches and find related data starting with a list of reviewed human genes. Specific tasks covered include finding reference sequences, mapping variations, identifying homologous genes, exploring expression studies, and using MyNCBI to save searches and manage data.

**Several staff members will be available at the [NCBI Booth #755](#):**

- Wednesday, October 23: 10:00 am – 6:00 pm
- Thursday, October 24: 10:00 am – 4:30 pm
- Friday, October 25: 10:00 am – 2:30 pm

**Some of the NCBI Posters that will be presented:**

Wednesday - Oct 23, 2013 11:30am-12:30pm

**1686W:** "Representation of Medical Variation at NCBI: ClinVar, Gene, and MedGen."

- D. Maglott, M. Landrum, J. Lee, W. Rubinstein, K. Katz, W. Jang, D. Hoffman, S. Chitipiralla, M. Ovetsky, J. Garner, R. Tully, L. Phan, D. Shao, R. Maiti, R. Villamarin, S. Gorelenkov, S. Sherry, D. M. Church

**740W:** "Pharmacogenetics at NCBI."

- A. J. Malheiro, W. Rubinstein, B. Kattman, J. Lee, D. Maglott, V. Hem, M. Ovetsky, G. Song, K. Katz, C. Wallin, R. Villamarin, J. Ostell

Thursday - Oct 24, 2013 10:30am-11:30am

**1441T:** "Web-based tools to support the clinical genetics lab."

- D. M. Church, L. Kalman, V. Ananiev, N. Bouk, C. Chen, A. Doubintchik, M. Halavi, M. Landrum, P. Meric, L. Phan, D. Shao, D. Slotta, J. Trow, M. Ward, D. R. Maglott

**1543T:** "Variation data services at NCBI: archives, tools, and curation for research and medicine."

- S. Sherry, K. Address, V. Ananiev, C. Chen, D. Church, M. Feolo, J. Garner, T. Hefferon, D. Hoffman, B. Holmes, M. Kholodov, A. Kitts, J. Lee, J. Lopez, D. Maglott, R. Maiti, L. Phan, G. Riley, W. Rubinstein, D. Rudnev, Y. Shao, E. Shekhtman, K. Sirotkin, D. Slotta, R. Tully, R. Villamarin-Salomon, Q. Wang, M. Ward, H. Zhang, C. Xiao

**2619T:** "ClinVar: Improving Access to Clinically Relevant Variants for the Research and Clinical Genomics Communities."

- M. J. Landrum, J. Lee, G. Riley, R. Tully, S. Chitipiralla, M. Halavi, D. Hoffman, J. B. Holmes, W. Jang, K. Katz, M. Ovetsky, A. Sethi, R. Villamarin, D. M. Church, W. S. Rubinstein, D. R. Maglott

Thursday - Oct 24, 2013 11:30am-12:30pm

**1546T:** "The database of Genotypes and Phenotypes: dbGaP."

- M. Feolo, R. Bagoutdinov, S. Dracheva, L. Hao, Y. Jin, M. Kimura, M. Lee, J. Mena, N. Popova, S. Pretel, N. Sharopova, S. Stefanov, A. Stine, A. Sturcke, K. T. Tryka, Z. Wang, M. Xu, L. Ziyabari, S. T. Sherry

**1594T:** "Change can be good: updating the human reference genome assembly."

- V. A. Schneider, P. Flicek, T. Graves, T. Hubbard, D. M. Church for the Genome Reference Consortium

**2628T:** "The NIH Genetic Testing Registry: 2013 status report on genetic testing."

- W. S. Rubinstein, B. L. Kattman, A. J. Malheiro, J. M. Lee, D. R. Maglott, V. Hem, M. Ovetsky, G. Song, C. Wallin, K. S. Katz, R. Villamarin-Salomon, C. Fomous, J. M. Ostell

## **Organism BLAST pages now use top-level RefSeq genomic records instead of scaffold records**

*Monday, October 21, 2013*

The organism BLAST pages are being updated to use top-level (chromosome + unplaced and unlocalized scaffolds) RefSeq genomic records instead of scaffold records. This change has also been made for the human and mouse G+T BLAST databases. Reporting hits in chromosome coordinates is more useful for public reporting and also makes it easier to relate the results to data on other sites.

For more information, see this BLAST News Story on the "[Update to organism BLAST databases](#)".

## NCBI News, September 2013

### Try the new My NCBI Feature: SciENcv

*Thursday, September 26, 2013*

The National Institutes of Health has [issued an invitation](#) to researchers to test the beta version of the Science Experts Network (SciENcv). [SciENcv](#) is a new feature available in [My NCBI](#) that helps users create an online professional profile that can be made public to share with others. The addition of this new feature complements [My NCBI's My Bibliography](#) which aids users in managing a list of their citations (journal articles, books/chapters, patents, presentations and meetings) which can be saved directly from PubMed or manually added using [My Bibliography](#) templates.

Currently, the beta version of [SciENcv](#) enables researchers to quickly assemble the information for and generate an NIH biographical sketch (biosketch). [My Bibliography](#) users and [eRA Commons](#) account holders who have linked their eRA account to [My NCBI](#) will find their [SciENcv](#) profile automatically populated with relevant data. This information can be modified and users can add information about their education, employment, research activities, publications, honors, research grants, and other professional contributions. In addition, an [ORCID® iD](#) can be included in the [SciENcv](#) profile.

Future plans for this project include the ability to generate alternative biographical sketch formats so that users can create and electronically submit their data to many federal agencies.

[SciENcv](#) is a cooperative project requested by the [Federal Demonstration Partnership](#). Seven federal science agencies formed an interagency workgroup to develop the concept:

- National Institutes of Health
- National Science Foundation
- Department of Defense
- Department of Energy
- Environmental Protection Agency
- U.S. Department of Agriculture
- and Smithsonian Institution

[SciENcv](#) is being built by the NIH National Center for Biotechnology Information (NCBI) under the direction of the workgroup. For additional details on the mission and guiding principles of the [SciENcv](#) project, please see the [project page](#).

**For more information, see:**

- [NLM Technical Bulletin SciENcv Article](#)
- [NIH Grants Notice - NOT-OD-13-114.html](#)
- [SciENcv Help](#)
- [My NCBI Help](#)
- [My Bibliography Help](#)

### Comments Requested: NIH genomic data sharing policy

*Friday, September 20, 2013*

The National Institutes of Health (NIH) is seeking public comments on the draft Genomic Data Sharing (GDS) Policy that promotes sharing, for research purposes, of human and non-human genomic data generated from NIH-supported and NIH-conducted research.

The draft GDS Policy describes the responsibilities of investigators and institutions for the submission of genomic data to the NIH and the use of controlled-access data obtained from dbGaP or other NIH databases. It also addresses issues and considerations pertaining to informed consent, data management and intellectual property.

To read the proposed policy and learn how to submit comments, please go to the [Federal Register NIH Draft GDS Policy page](#).

## RefSeq release 61 now available

*Wednesday, September 18, 2013*

The complete RefSeq release 61 contains 41,958,567 records, 33,139,114 proteins, 4,528,216 RNAs, and sequences from 29,414 different organisms. The new release reflects SNP Build 138 and includes human annotation 105 with new gene model splice variants from RNA-Seq data. For additional information on the new splice variants, see the [NCBI News item](#) about human annotation release 105. More details about RefSeq release 61 are in the [release statistics](#) and the [release notes](#).

## A new NCBI Insights post about the use of NCBI Data for scientific discovery

*Monday, September 16, 2013*

A new [NCBI Insights blog post](#) highlights how three research groups reused data from NCBI to make important discoveries.

Read about these case studies:

- "Identifying Common Genes and Networks in Multi-Organ Fibrosis" using GEO data
- "Predicting Adverse Drug Reactions Using Publicly Available PubChem BioAssay Data" from PubChem
- "Prediction of Susceptibility to Major Depression by a Model of Interactions of Multiple Functional Genetic Variants and Environmental Factors" with the help of dbGaP data

## New PubChem social media sites help keep users up-to-date!

*Thursday, September 05, 2013*

The [PubChem Project](#) has several new ways for users interested in the chemical and bioactivity resource to learn about announcements, updates and new tools.

These include:

- [PubChem Announcements page](#)
- [PubChem RSS feed](#)
- [PubChem Twitter channel](#)
- [PubChem Facebook page](#)
- [PubChem Google+ page](#)
- [PubChem Blog site](#)

In addition to being able to follow the new PubChem social media streams, a new "SHARE" button on the top of PubChem record pages allows users to share PubChem pages with friends and colleagues.

Read more about the Social Media Campaign and the new SHARE feature in the first [first PubChem Blog posting!](#)

## NCBI News, August 2013

### Human genome annotation release 105 with new splice variants

*Tuesday, August 27, 2013*

The NCBI recently finished a re-annotation of three complete (GRCh37.p13, CHM1\_1.1, and HuRef) assemblies and one partial (CRA\_TCAGchr7v2) assembly of the human genome. Annotated genomic, transcript, and protein records are available through the integrated Entrez system (Nucleotide, Protein, Gene) and may be downloaded through FTP or the Aspera protocol. The new annotation provides additional splice variants for many human genes. Genes may now be annotated with both known Reference Sequences (NM\_ style accessions) and gene models (XM\_ style accessions.) RefSeq models are generated from mRNA, protein, and RNAseq data. Twelve thousand genes are now annotated with both known and model RefSeqs on the GRCh37.p13 assembly, approximately doubling the number of splice variants represented.

This is NCBI's last full annotation of the GRCh37 assembly. The next full annotation release for human will include GRCh38. See the [Genome Reference Consortium](#) site for information on the upcoming human genome build.

### dbSNP Build 138, phase III, now available

*Thursday, August 22, 2013*

dbSNP build 138 phase III update is now available. This update includes data for mouse, *Arabidopsis thaliana*, honeybee, *C. elegans*, and rice. Build 138 provides more than 505 million submitted and 226 million reference variants for 131 species. To see complete build statistics visit the SNP [summary page](#). You may access build 138 SNP data through the integrated NCBI [Entrez system](#) and download data through FTP or Aspera protocol.

### Sequence Viewer 2.27: new features, improvements, and help documentation

*Wednesday, August 21, 2013*

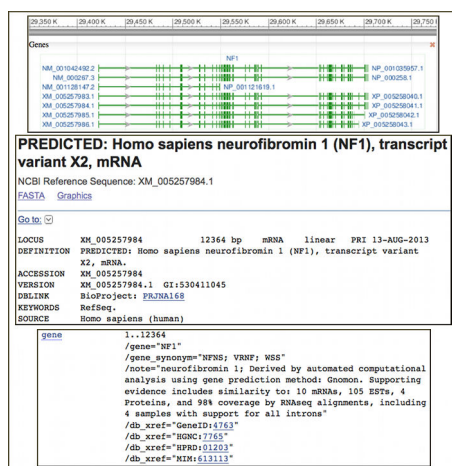
Sequence Viewer 2.27 ([Release Notes](#)) is now appearing on the NCBI site (nucleotide, protein, gene, SNP) and available for [embedding](#) in outside pages. Version 2.27 has important new features and improvements including new tiling path and contig (scaffold) tracks for assembled records, drag and drop reordering and one click removal of tracks within the graphical view. The Sequence Viewer also now has a reorganized and improved [help documentation site](#).

### > 10,000 tests now listed in the NIH Genetic Testing Registry

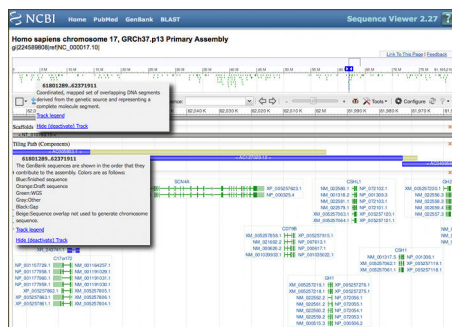
*Wednesday, August 21, 2013*

The NIH Genetic Testing Registry (GTR) is a free online resource that provides centralized access to comprehensive genetic test information voluntarily submitted by test providers. As of August 19, 2013, more than 10,000 tests for over 3,350 conditions have been submitted by laboratories.

For information on how to submit data to GTR, see the documentation describing the [GTR Submission Process](#).



New splice variants for the human NF1 gene (Gene ID: 4763). Top panel. A graphical view of chromosome 17 showing the seven splice variants for NF1, the last three are predictions based on mRNA, protein, and RNAseq data. Lower panels. A Nucleotide database view of splice variant X2 (XM\_005257984) showing the gene feature with supporting evidence.



Sequence viewer 2.27 showing a region of chromosome 17 near the SCN4A gene. The new Tiling Path and Scaffolds tracks show how the region is assembled. Tracks now can be dragged to new positions using the mouse and may be dismissed by clicking the red 'x' at the right hand margin.

## GenBank Release 197.0 is Available

Friday, August 16, 2013

The new release for GenBank is now available via FTP, as well as in the Nucleotide database and BLAST services.

Release 197.0 (08/15/2013) 167,295,840 non-WGS, non-CON records which were comprised of 154,192,921,011 basepairs of sequence data. In addition, there were 124,812,020 WGS records containing 500,420,412,665 basepairs of sequence data.

During the 60 days between the close dates for GenBank Releases 196.0 and 197.0, the non-WGS/non-CON portion of GenBank grew by 1,593,690,899 basepairs and by 1,555,676 sequence records and the WGS component of GenBank grew by 46,590,660,345 basepairs and by 12,323,984 sequence records.

The total number of sequence data files increased by 25 with this release, with the divisions that expanded in file number:

- BCT = 3 new files, now a total of 106
- CON = 7 new files, now a total of 215
- ENV = 1 new file, now a total of 62

- EST = 1 new file, now a total of 474
- GSS = 5 new files, now a total of 278
- PLN = 1 new file, now a total of 63
- PRI = 1 new file, now a total of 46
- TSA = 4 new file, now a total of 145
- VRL = 1 new file, now a total of 26
- VRT = 1 new file, now a total of 31

For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 607 GB (sequence files only).

For additional release information, see the [Release Notes](#) and README files in individual directories.





## NCBI News, July 2013

### Tenth Anniversary of RefSeq FTP Releases

Friday, July 26, 2013

The July 2013 RefSeq FTP release marks the 10th anniversary of RefSeq comprehensive FTP releases. We mark this occasion with a sincere "Thank you!" to the scientific community for continued interest and support, comments, and useful suggestions for improvements that have been made over the past years.



There has been significant total growth since the first release in June 2003! So, we thought you might be interested in seeing how much the RefSeq data has grown.

#### Growth in the number of accessions, by molecule type:

Type of Sequence	June 2003 (Release 1)	July 2013 (Release 60)	Percentage Growth over 10 years
Genomic	64,729	4,165,752	6,336%
RNA	211,803	4,243,209	1,903%
Protein	785,143	32,504,738	4,040%

#### Growth in the number of species, per node:

Taxonomic Node	June 2003 (Release 1)	July 2013 (Release 60)	Percentage Growth over 10 years
Complete	2005	28,560	1,324%
Fungi	27	785	2,807%
Invertebrates	80	1,121	1,310%
Microbes	334	20,213	5,952%
Mitochondria	417	3,793	810%
Plants	30	349	1,063%
Plasmids	36	1,501	4,069%
Plastids	31	359	1,058%
Protozoa	39	179	359%

Table continued from previous page.

Taxonomic Node	June 2003 (Release 1)	July 2013 (Release 60)	Percentage Growth over 10 years
Mammals	74	580	684%
Non-mammalian Vertebrates	206	1,796	772%
Viruses	1179	3,536	200%

## RefSeq Release 60 is Available for FTP

Friday, July 26, 2013

The complete RefSeq release 60 contains 40,913,699 records, 32,504,738 proteins, 4,243,209 RNAs, and sequences from 28,560 different organisms. See the [Release statistics file](#) or [Release notes](#) for more information.

### There are several important announcements for RefSeq release 60.

Selected announcements described below include:

- A new bacterial protein data model and accession series
- Suppression of some bacterial genomes
- Changes in annotation of human and vertebrate transcript records
- Policy change to allow a mixture of known and model accessions for eukaryotic genes

Please see the [release note announcement for RefSeq release 60](#) and documents in the [new announcement directory](#) for the full set of announcements with detailed information.

### Bacterial genomes, new protein data model and accession series (WP)

NCBI continues to expand the RefSeq bacterial genomes node to include ALL complete and draft genomes that meet minimum assembly and annotation quality criteria. This means that RefSeq will include more than one genome of the same strain which may be provided through strain population sampling or sequencing to monitor a disease outbreak. NCBI is in the process of re-annotating all bacterial genomes, with the exception of a small number for which annotation is provided by, or in collaboration with, another group (such as *E. coli* str. K12 substr. MG1655).

Due to the expanded scope of the RefSeq bacterial node, we anticipated a very large increase in the number of identical (redundant) proteins; therefore, we have introduced a new data model for bacterial proteins whereby we are providing a true non-redundant protein dataset associated with a new accession prefix, 'WP'. Details about the new data model with examples was announced between release cycles.

This release includes a new supplemental file providing mapping of WP accessions to tax\_id and species name, for the subset of WP accessions that are annotated on genomes of different species. For example, see WP\_000002243.1. The mapping file is available in the [release-catalog directory](#).

We strongly encourage you to read the [full announcement](#).

### Suppression of some bacterial genomes

Please note that some RefSeq bacterial genomes were recently suppressed. This includes unannotated genomes that had not been processed by NCBI's annotation pipeline yet and annotated genomes with identified

annotation quality issues. This has resulted in a net decrease in RefSeq bacterial genomic accessions in this release. Many of the suppressed accessions will be reinstated when annotation is provided.

## Changes in annotation of human and vertebrate transcript records

Recent changes to human and other vertebrate transcript records includes:

- removal of exon numbers
- expanded reporting of support evidence, in a structured comment with the header 'Evidence Data'
- (new) reporting gene and transcript attributes, in a structured comment with the header 'RefSeq Attributes'
- removal of mitochondrial localization information from the record DEFINITION line (moved to Attributes)

Please see the [detailed description of these changes](#).

## Policy change to allow a mixture of known and model accessions for eukaryotic genes

Previously, we did not allow a mixture of X\* series accessions (genome annotation models) and N\* series accessions (based on cDNA and curation) for a gene. We have changed this policy in order to provide increased annotation of splice variants. RefSeq models are calculated using cDNA, protein, and RNAseq data. There may be good support at the level of each exon pair; however, the long range exon combination represented in the model may not be fully supported and thus is less likely to be represented with a N\* series accession. For example, see Gene ID: [100306968](#).

## New NCBI Insights Post: "New Pandoravirus Sequences are Accessible in GenBank"

*Wednesday, July 24, 2013*

A new [NCBI Insights Blog post](#) provides information on a recent article that describes the discovery and characterization of two "giant" viruses that are proposed to comprise the first members of the "Pandoravirus" genus. The authors of this publication have submitted assembled and annotated genomes to NCBI, which are currently available in the Nucleotide database with the accessions [KC977571](#) and [KC977570](#).

### For more information see:

- [NCBI Insights Blog Post: "New Pandoravirus Sequences are Accessible in GenBank"](#)
- Philippe, et al. "Pandoraviruses: Amoeba Viruses with Genomes Up to 2.5 Mb Reaching That of Parasitic Eukaryotes." *Science*. 2013 Jul 19;341(6143):281-6. doi: 10.1126/science.1239181.

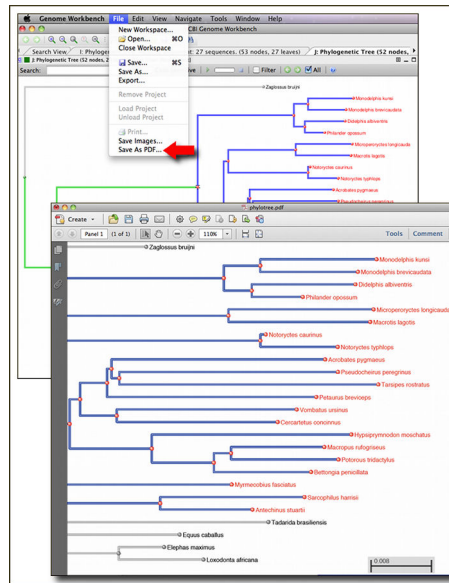
### Related NCBI Resources:

- [PubMed database: "Pandoravirus: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes" abstract](#)
- [GenBank](#)
- [Nucleotide database: \[KC977571\]\(#\) or \[KC977570\]\(#\) genome sequence records](#)
- [Protein database: protein sequences annotated on \[KC977571\]\(#\) or \[KC977570\]\(#\)](#)
- [BLAST](#)
- [CD-Search](#)

## Genome Workbench 7.6 with Publication Quality Graphics Export

*Monday, July 08, 2013*

The latest release 2.7.6 (2.7.5) of Genome Workbench, NCBI's standalone sequence analysis and annotation platform, now produces publication quality graphical output (PDF). A [new tutorial](#) shows how to use this helpful feature. The [release notes](#) have more information on this and other improvements.



Exporting a phylogenetic tree view as a PDF from Genome Workbench. The bottom panel shows the high quality PDF.



## NCBI News, June 2013

### Come to the NCBI Discovery Workshops on July 30 & 31!

*Friday, June 28, 2013*

Spaces are still available for the free, 2-day NCBI Discovery Workshops to be held on the NIH Campus on July 30 and 31, 2013. For more information and to register, visit the [Discovery Workshops homepage](#).

The NCBI Discovery Workshops comprise four workshops that will teach you how to use the NCBI Web resources more effectively. The July 2013 Workshops consist of four 2.5-hour hands-on sessions, with each session focusing on a different related group of NCBI tools and databases:

- Sequences, Genomes, and Maps
- Proteins, Domains, and Structures
- NCBI BLAST Services
- Human Variation and Disease Genes

Materials from all Discovery Workshops offerings are available from the [Education FTP directory](#).

### Upload and graphically compare your own data with NCBI Epigenomics tracks

*Wednesday, June 26, 2013*

Recently, a new “Upload Tracks” system has been added to the [NCBI Epigenomics](#) resource to allow users to view and compare their own data with information stored at NCBI.

The NCBI Epigenomics resource, a comprehensive public repository for whole-genome epigenetic datasets, contains information from a subset of data in the [Gene Expression Omnibus \(GEO\)](#), which has been subjected to additional review and annotation. Currently there are over 4200 viewable and downloadable datasets from over 1200 samples that have been isolated from five well-studied species.

From the NCBI Epigenomics homepage (Figure 1A), you can access the “Upload” page where your own datasets can be uploaded and displayed as tracks in the Epigenomics genome viewer.

Please note that the “Upload Tracks” feature requires logging into a [My NCBI Account](#) which facilitates the storing of information for future examination and also ensures that the data is only visible and viewable by the account holder. The uploaded information cannot be viewed, downloaded or used by any other user of the Epigenomics resource.

Once logged into NCBI, the Epigenomics “Upload” page (Figure 1B) contains fields for entering important metadata information as well as the dataset itself into the system.

Each user-uploaded dataset is listed in the “My Uploads” collection as an experiment (Figure 2A). This allows for uploaded data tracks to be selected for operations such as adding to user-created collections or viewing in concert with other database tracks in the customizable genome viewer interface (Figure 2B).

FOR MORE INFORMATION:

[Epigenomics Homepage & Epigenomics Upload Page](#)

- [Uploading Epigenomics DataSets - YouTube Video | HelpDoc](#)
- [Using the Epigenomics Genome Viewer - YouTube Video | HelpDoc](#)

[My NCBI - Sign In/Registration Page & YouTube Video](#)

## **SRA-BLAST has been updated with new features and performance enhancements**

*Tuesday, June 25, 2013*

SRA-BLAST has undergone a dramatic update, both in terms of user interface and search performance. These updates to SRA-BLAST make it an even more useful tool for searching through more than 700 trillion open-access bases currently housed within the [Sequence Read Archive \(SRA\)](#).

New features include:

- Targeted searching within one or more SRA Experiment sets (i.e., "SRX accessions"). Users may now search combined datasets of up to 2 billion individual reads.
- An "autocomplete" feature that will allow users to specify SRX accession, SRX title, organism scientific name, or tax id to help build the search set.
- Data obtained from Roche 454 and newer Illumina instruments (HiSeq and MiSeq).

## **Welcome to the NCBI News site!**

*Tuesday, June 25, 2013*

This is the place to get the latest information about NCBI, and feature stories about NCBI services and activities. The NCBI News site offers readers fast and integrated access to the most important news stories about announcements, changes, updates and improvements at NCBI.

This site replaces the [NCBI Newsletter](#), previously published on the NCBI Bookshelf. In comparison to the Bookshelf Newsletter, News stories will have a more rapid publication cycle enabling the release of announcements and updates as they become available.

In the "Follow us" portlet, the News site displays icons which link directly to NCBI's [RSS Feeds](#) and [Email ListServes](#) in a single, easily accessible place, as well as icons which link to the [NCBI Insights Blog](#) and NCBI's social media outlets on [Twitter](#), [Facebook](#), and [YouTube](#).

In addition to reading articles describing changes and updates to our website and data, the NCBI News site offers the ability to share interesting information with others. By clicking the "Share" button located at the top and bottom of each news story, readers can post the title of the article and a link directly to social media sites such as Facebook, Twitter, LinkedIn, WordPress, Reddit, Tumblr, Pinterest, StumbleUpon, and many others....

## **Dr. David Lipman Receives White House "Open Science" Champions of Change Award on Behalf of NCBI**

*Thursday, June 20, 2013*

[Dr. David Lipman](#), Director of the NLM's National Center for Biotechnology Information (NCBI), was among those honored by the White House on June 20 for their outstanding work in "promoting and using open scientific data and publications to accelerate progress and improve our world" as a [White House "Open Science" Champion of Change](#).



Figure 1. A) The Epigenomics homepage has links for lots of helpful information and tools including the “Upload Tracks” feature. B) The “Upload” page contains a form for the input of relevant information about the dataset. Required information (a) includes Track name, File type, Dataset from either in an uploadable file or a public URL, Species & Genome assembly to serve as the framework for the alignment, and Feature type - specific histone modifications (e.g. H3K4me3, H3K27me3), DNA methylation, chromatin accessibility and more. Additional optional metadata fields (b) are also available for the user to store information which can be used for quick comparisons with other samples in the system.

As Director of the NCBI, Dr. Lipman was honored for his leadership in making biomedical data and health information publicly and easily available to all, including scientists, medical professionals, patients, educators and students.

“I am truly honored that the White House has recognized our work in providing resources such as NCBI’s GenBank database of all publicly available DNA sequences and PubMed Central, an online archive of peer-reviewed biomedical sciences literature,” said Dr. Lipman. “The success of these databases and NCBI’s many other resources is a reflection of the hard work, dedication and talent of all those working at NCBI.”

Links: [White House Press Release](#) & [NLM Announcement](#)

## GenBank Release 196.0 is Available

Tuesday, June 18, 2013

The new release for GenBank is now available via <ftp.ncbi.nlm.nih.gov>, as well as in the [Nucleotide database](#) and [BLAST services](#).

Release 196.0 (06/13/2013) 165,740,164 non-WGS, non-CON records which were comprised of 152,599,230,112 basepairs of sequence data. In addition, there were 112,488,036 WGS records containing 453,829,752,320 basepairs of sequence data.

Figure 2. A) User-uploaded datasets are listed in the “My Uploads” collection and shown as independent experiments with supplied metadata, such as cell type, tissue type, differentiation state, etc. These are displayed in the filterable and sortable “Experiments” table. Using the check boxes, at left, to select tracks of interest and clicking “View on Genome” will open a window with the tracks in a customizable genome viewer. B) The uploaded data are shown at the top with user-provided Track names and directly comparable to selected Epigenomics experiment tracks, as well as other NCBI tracks containing Gene annotation, Genome-wide association study, Cited variant, and CpG island information.

Figure 2. A) User-uploaded datasets are listed in the “My Uploads” collection and shown as independent experiments with supplied metadata, such as cell type, tissue type, differentiation state, etc. These are displayed in the filterable and sortable “Experiments” table. Using the check boxes, at left, to select tracks of interest and clicking “View on Genome” will open a window with the tracks in a customizable genome viewer. B) The uploaded data are shown at the top with user-provided Track names and directly comparable to selected Epigenomics experiment tracks, as well as other NCBI tracks containing Gene annotation, Genome-wide association study, Cited variant, and CpG island information.

During the 63 days between the close dates for GenBank Releases 195.0 and 196.0, the non-WGS/non-CON portion of GenBank grew by 1,420,250,957 basepairs and by 1,603,433 sequence records. During that same period, 590,119 records were updated. An average of 34,818 non-WGS/non-CON records were added and/or updated per day. In addition, between releases 195.0 and 196.0, the WGS component of GenBank grew by 35,803,158,714 basepairs and by 1,978,722 sequence records.

The total number of sequence data files increased by 19 with this release, with the divisions that expanded in file number:

- BCT = 3 new files, now a total of 103
- CON = 3 new files, now a total of 208
- ENV = 1 new file, now a total of 61
- EST = 1 new file, now a total of 473
- GSS = 3 new files, now a total of 273
- INV = 1 new file, now a total of 35
- PAT = 5 new files, now a total of 195
- PLN = 1 new file, now a total of 62
- ROD = 1 new file, now a total of 31

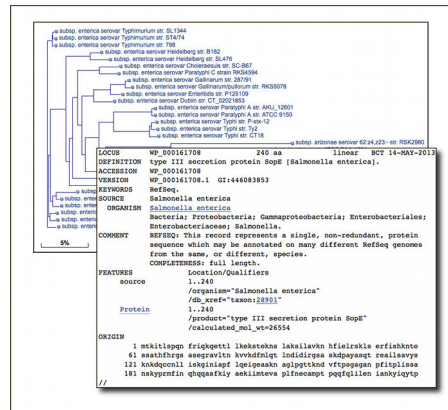
For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 600 GB (sequence files only).

For additional release information, see the [Release Notes](#) and README files in individual directories.

## **New RefSeq Bacterial Protein Products and Emerging RefSeq Data Model**

*Tuesday, June 11, 2013*

The NCBI Reference Sequence Project (RefSeq) project is now producing a non-redundant set of sequences to serve as annotation reagents for bacterial genomes. This is to help reduce and control redundancy in the protein databases while maintaining information content in response to high volume sequencing and annotation of multiple isolates. These new protein records begin with the accession prefix 'WP' and are used represent each unique bacterial sequence in the RefSeq data. These new proteins are independent of any particular bacterial genome and can be associated with more than one isolate, strain or species. Bacterial genomes will now be annotated using these WP proteins. Existing RefSeq bacterial sequences (YP and NP) accessions now point to the corresponding WP record, and WP records have replaced ZP accessions, which were formerly annotated on partially assembled whole genome shotgun genomes. Please see the [Reference Sequence Announcement](#) for further details and the plan for phasing in the implementation.



A non-redundant record ([WP\\_000161708](#)) representing the invasion-associated secreted protein SopE from *Salmonella enterica*.

## NCBI News, May 2013

### Need to Find Information about Genetic Tests? Try GTR!

*Monday, May 13, 2013*

A change in how people find information about genetic tests is imminent. On February 29, 2012, NIH's [Genetic Testing Registry \(GTR\)](#) was launched to provide access to a central repository for genetic testing information and to make it easier for clinicians to navigate the rapidly changing landscape of genetic tests. The GeneTests Laboratory Directory, long a source of information for clinicians, has been used by laboratories to seed information in GTR. NIH will no longer support the GeneTests website as of June 4, 2013. [GeneReviews](#) continues to be available through [NCBI's Bookshelf](#) and throughout GTR.

GTR is a free online resource that provides centralized access to comprehensive genetic test information voluntarily submitted by test providers. The entries listed in GTR include clinical and research tests for heritable mutations, including pharmacogenetic tests and tests using complex arrays and multiplex panels. GTR provides a wide range of information such as the test purpose and methods; the molecular, cytogenetic and biochemical test targets; evidence of clinical validity and clinical utility; ordering information; and laboratory credentials and contact information.

Currently there are over 3,700 registered tests for over 2300 conditions and 3300 genes in GTR.

Take a look at [NIH's Genetic Testing Registry](#) or watch a YouTube video to see how you can [Locate a Genetic Test in Under Three Minutes!](#)

### RefSeq Release 59 is Available for FTP

*Friday, May 03, 2013*

The [complete RefSeq release 59](#) contains 39,040,745 sequence records for 31,593,499 proteins, 3,579,371 RNAs, and sequences from 24,656 different organisms. Check out [RefSeq's homepage](#) to learn more about The Project and see the [Release statistics file](#) or [Release notes](#) for more information about this particular release.

### New YouTube Video: Complying with the NIH Public Access Policy with My Bibliography

*Thursday, May 02, 2013*

NIH-funded researchers are required to comply with the [NIH Public Access Policy](#). NCBI's [My Bibliography](#) was developed to assist scientists and their delegates in linking funding information with their citations. A new [NCBI YouTube video](#) about the use of My Bibliography for Public Access Compliance is available for more information and a demonstration.



## NCBI News, April 2013

### New publication: "BLAST: a more efficient report with usability improvements."

*Tuesday, April 30, 2013*

A new publication, "BLAST: a more efficient report with usability improvements," (PMID: [23609542](#)) is now available in [free full-text](#) from the Webserver Issue of Nucleic Acids Research. The paper describes the recent improvements in the NCBI BLAST Web output. These include more efficient loading of results, the ability to retrieve only the aligned regions, to display query-based or subject-based views of results in the graphical sequence viewer and to customize the descriptions table. A [factsheet](#) and a [video](#) on the NCBI YouTube channel provide a practical introduction to these features.

### "A Librarian's Guide to NCBI" Course was a Success!

*Monday, April 29, 2013*

Last week (April 15-19, 2013), NCBI in collaboration with the [National Library of Medicine](#) and the [National Network of Libraries of Medicine NLM Training Center](#) at the University of Utah presented "A Librarian's Guide to NCBI". This new course, which was highly rated by participants, was designed to prepare health science librarians for supporting and training patrons about NCBI molecular databases and tools at their own institutions.

As promised in last week's [NCBI Insights Blog post](#), the materials used in "A Librarian's Guide to NCBI" are now available for download and use for personal enlightenment or to supplement training in workshops or courses.

On a typical day the course offered two modules, each one focused on a different aspect of molecular data at NCBI and included a short lecture followed by an assessment quiz, instructor-led practical demonstrations, and individual practice problems. In addition to the modules, there were two discussion sessions reviewing library patron questions provided by participants, an open question and answer session with NCBI engineering branch supervisors, a tour of the National Library of Medicine, and a visit by NCBI Director David Lipman. An online forum for the librarian cohort is being developed for continued communications and support.

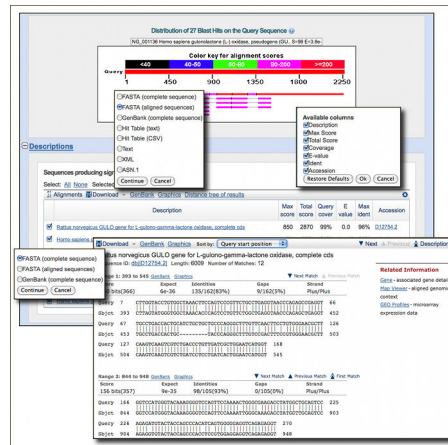
Based on strong participant evaluations and requests, we are planning to offer the Librarian's Guide at least once a year.

Check back on [NCBI's Education page](#) for future offerings of this and other NCBI courses.

In addition to their usage in "A Librarian's Guide to NCBI", the curricular materials were developed as separate stand-alone modules to be used by educators and bioinformatics trainers. These are now available for download on the [Librarian Course FTP site](#).

Modules were developed to explain and demonstrate related NCBI resources for use by researchers of broad biology-based disciplines. In each of the eight modules the following questions were answered:

- Why are the data generated?
- How are the data generated / determined / measured?
- How does the NCBI organize and represent the data?
- What tools are available at the NCBI to analyze / search the data?
- What experimental questions can be answered with the data?
- What are the caveats / limits of data interpretation?



The re-designed BLAST results showing many of the new options.

- What would library patrons want to do with the data?

Each module featured a 30-minute lecture followed by a brief assessment quiz with a discussion of the answers, instructor-led practical demonstrations, and individual practice problems, and topics covered were:

- [Molecular Biology Basics](#) - a review of molecular biology concepts focusing on biological information flow and the gene as a central theme and showed how the NCBI Gene database serves as a central access point for molecular data at NCBI.
- [Advanced Entrez Searching](#) - a demonstration of how to use the Entrez integrated database and search system to find relevant data using both basic and advanced interfaces and fielded searches. The module also demonstrated the importance of pre-compiled and pre-computed relationships for navigating within a database and laterally across the Entrez system.
- [NCBI BLAST](#) - a full-day introduction to sequence similarity searching using NCBI's Basic Local Alignment Search Tool (BLAST). This module covered the basics of sequence alignment algorithms, scoring matrices, and local alignment statistics and used practical protein and nucleotide search examples that highlighted features of the BLAST web service designed to give the most relevant results.
- [Sequences & Genomes](#) - an exploration of the essential role of nucleotide and protein sequence data in modern biological research and the Nucleotide database as the backbone of the NCBI molecular databases. The module explained how NCBI manages and processes sequence and other data associated with genomes and their annotation. Demonstrations and exercises showed how to identify the most up-to-date and well-annotated sequence.
- [Sequence Variation and its Consequences](#) - an examination of the many databases and tools at NCBI that provide access to variation data emphasizing the association between variation and disease risk. After describing the different types of genetic variation as well as the major study methods that produce these data, practical demonstrations and exercises demonstrated how to navigate the NCBI variation resources to find specific data and important attributes, such as geographic population, allele frequency, and disease association.
- [Gene Expression & Biological Pathways](#) - a review of NCBI databases and tools relevant to the study of gene expression. The module provided basic background on the importance of gene expression in various biological phenomena and high-throughput techniques for measuring expression. Practical demonstrations showed how to find and compare expression patterns of genes in different samples in microarray datasets and expression profiles, and how to map selected genes onto metabolic pathways.
- [Protein Structures](#) - an illustration of the usefulness and interconnectedness of NCBI protein structure databases and tools using the example DNA Topoisomerase II. The module covered basic concepts of structural biology and the importance of 3D structure information in understanding the normal functions



of proteins and abnormal functions that result in disease. Practical examples showed how to find available 3D structural data for a given protein sequence, detect functional domains within the sequence, view 3D structure data using Cn3D, and explore the relationship between protein sequence and structure data.

- [Drugs & Other Small Molecules](#) - a tour of NCBI's Chemical and Bioactivity Databases developed by The PubChem Project. The module explained and explored the data in and relationship between PubChem databases (Compound, Substance and BioAssay). Practical examples elucidated the types of data that are accessible from these resources, and provided case-study specific, guided demonstrations for finding information to answer important scientific questions.

Based on course feedback, we plan to expand the course materials to include a set of videos of the lectures and demonstrations to be produced for the [NCBI YouTube Channel](#) as well as a set of worked exercises suitable for classroom teaching.

Visit [NCBI's Education page](#) for links to these and other training materials.

## GenBank Release 195.0 is Available

*Tuesday, April 16, 2013*

The new release for [GenBank](#) is now available via <ftp.ncbi.nlm.nih.gov>, as well as in the [Nucleotide database](#) and [BLAST services](#).

In release 195.0 (04/11/2013), the total number of non-WGS, non-CON records was comprised of basepairs of sequence data. In addition, there were 164,136,731 WGS records containing 151,178,979,155 basepairs of sequence data.

During the 57 days between the close dates for GenBank Releases 194.0 and 195.0, the non-WGS/non-CON portion of GenBank grew by 1,037,624,297 basepairs and by 1,250,004 sequence records, with an average of 32,964 non-WGS/non-CON records added and/or updated per day. In addition, the WGS component of GenBank grew by 27,125,603,190 basepairs and by 7,408,023 sequence records.

The total number of sequence data files increased by 30 with this release, with the divisions that expanded in file number:

- BCT = 2 new files, now a total of 100
- CON = 8 new files, now a total of 205
- ENV = 1 new file, now a total of 60
- EST = 3 new files, now a total of 472
- PHG = 1 new file, now a total of 2
- PLN = 1 new file, now a total of 61
- TSA = 3 new files, now a total of 141
- VRL = 1 new file, now a total of 25

For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 594 GB (sequence files only).

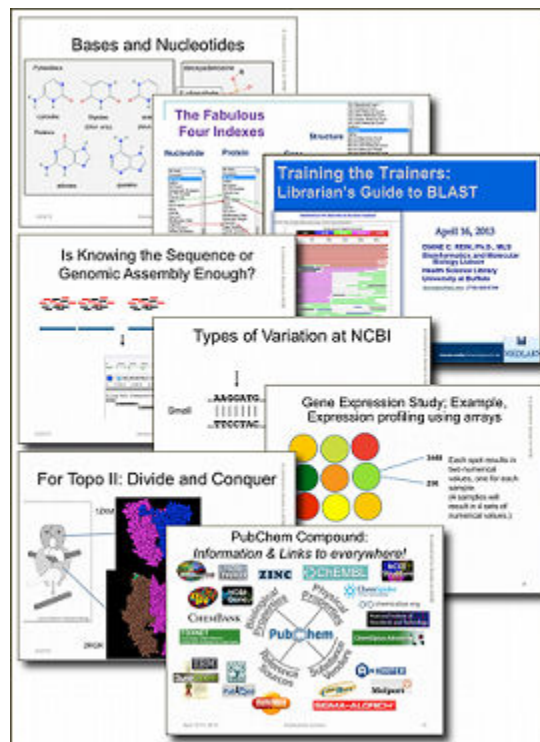
For additional release information, see the [Release Notes](#) and README files in individual directories.

## New Educational Initiative: A Librarian's Guide to NCBI

*Thursday, April 11, 2013*



Participants, instructors, and organizers in the first offering of “A Librarian’s Guide to NCBI” outside the National Library of Medicine including librarians from 21 universities, medical centers and research institutions representing 14 states. Instructors were NCBI Staff Members Peter Cooper, Bonnie Maidak, Wayne Matten, Majda Valjavec-Gratian, Eric Sayers and Rana Morris, as well as Diane Rein from the University at Buffalo.

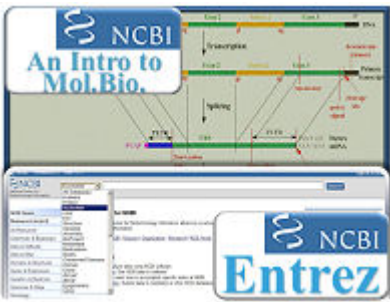


Sample slides from the eight modules of A Librarian’s Guide to NCBI. Complete PowerPoint files are available from the [Librarian Course FTP site](#).

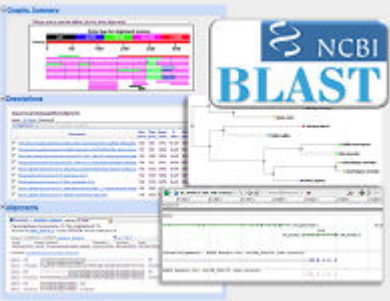
Next week NCBI will premiere [A Librarian’s Guide to NCBI](#), a new course aimed at teaching health science librarians about NCBI resources. For more information, a [new NCBI Insights Blog](#) introduces the course and updates on the course and the availability of the curricular materials will be publicized on [Twitter](#) and [Facebook](#).

### **A Librarian’s Guide to NCBI Course Modules**

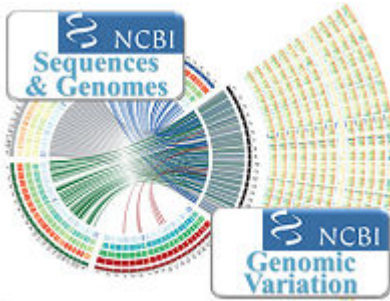
Day 1



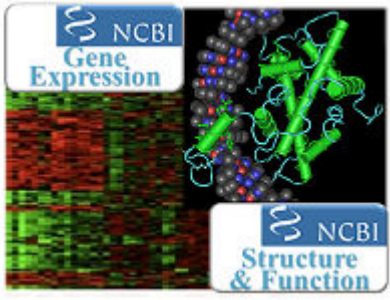
Day 2



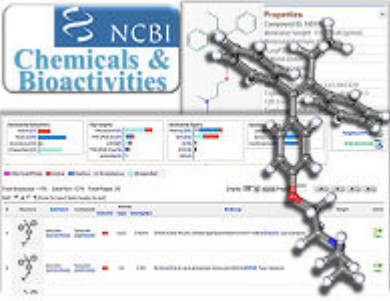
Day 3



Day 4



Day 5



The image displays a vertical sequence of five panels, each representing a day of a course. Each panel features a logo for a specific NCBI service and a representative screenshot of its interface. Day 1: 'An Intro to Mol. Bio.' logo and Entrez interface. Day 2: 'BLAST' logo and BLAST search results interface. Day 3: 'Sequences & Genomes' and 'Genomic Variation' logos with a circular genomic map. Day 4: 'Gene Expression' and 'Structure & Function' logos with a heatmap and a 3D protein structure. Day 5: 'Chemicals & Bioactivities' logo and a chemical structure interface.

## PubChem Releases New and Enhanced Webpage Widgets

Wednesday, April 10, 2013

New [PubChem Widgets](#) (Chemical Structure Carousel, Classification Listing and Autocomplete) have been developed for you to use in your own webpage. In addition, existing table-based Widgets (including Bioactivity, Patents and PubMed) have been enhanced with a Link/Embed button that allows you to open the widget in a PubChem page or embed the widget in your own page as an iframe.

- The [Structure Carousel](#) displays chemical structure thumbnail images along with names/synonyms, and will also show related annotations, when available, such as medication information, literature, patents, bioactivities, and 3D structures.
- The [Classification Listing](#) displays the classification, when available, of a PubChem Compound, Substance, or BioAssay. Current Classifications include MeSH, ChEBI, KEGG, LIPID MAPS, and Gene Ontology.
- The [Autocomplete Widget](#) is an embeddable tool that suggests a list of terms when you type input into a search field.

For complete documentation about all PubChem Widgets, see: [http://pubchem.ncbi.nlm.nih.gov/widget/docs/widget\\_help.html](http://pubchem.ncbi.nlm.nih.gov/widget/docs/widget_help.html)

## BLAST 2.2.28 now available

*Friday, April 05, 2013*

Stand-alone BLAST version 2.2.28+ is now available for download from the [FTP site](#). BLAST 2.2.28+ provides a number of important new features, improvements and some bug fixes. New features include composition-based statistics for Reverse PSI-BLAST (rpsblast), expanded options (query coverage, subject title, and taxonomy) for tabular output, and batch subsequence retrieval in blastdbcmd. Improvements include adaptive [BATCH\\_SIZE](#) resulting in more efficient searching, and incremental production of XML results. The [Blast Release Notes](#) have more details.

## Try it out! The New PubChem Upload Beta Site

*Friday, April 05, 2013*

A new beta version [PubChem Upload system](#) is available to try out. It features streamlined procedures for data submissions and updates to both the [PubChem Substance](#) and [BioAssay databases](#).

The new capabilities offered by PubChem Upload include:

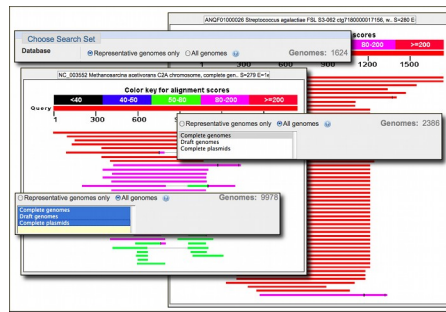
- Assay & Substance wizards to assist novice users
- Greatly improved UI speed using newer web technology (minimizing possible time-outs)
- Easy new user registration/easy upgrade
- Improved help with a tutorial and hints built into user interface
- Substance input in varied formats CID, SID, SMILES, etc.
- PubChem substance/assay templates for new submissions or for record updates
- Error display integrated with substance list displays
- Full editing and integration of assay data & description tables
- Expanded import/export for data description table spreadsheets

This system will eventually replace the original [Pubchem Deposition Gateway](#).

## New database options in Microbial Genomes BLAST: Representative Genomes

*Friday, April 05, 2013*

**Microbial Genomes BLAST** has new database options including 'Representative genomes', now the default database, and 'All genomes'. Representative genomes provide a smaller less redundant set of records for a given bacterial species. These representatives are selected by the research community and NCBI computational processes and are especially helpful for microbial species that are highly represented by genomes for numerous strains in NCBI databases, such as *Escherichia coli*. The 'All genomes' option offers the choice of Complete genomes, Draft genomes, or Complete plasmids. You can search these sets individually or in any combination. The microbial BLAST report also has a new 'Genome' link to the species page in Entrez Genome in the alignments section of the BLAST report. [Run a search.](#)





## NCBI News, March 2013

### New CDD Release v3.10 Includes an Updated PSSM Calculation

*Tuesday, March 26, 2013*

Conserved Domain Database (CDD) version 3.10 is now available with 1104 new or updated NCBI-curated and 48,034 total domain models. For more information, see the [CDD News page](#).

In this new release, position-specific score matrices (PSSMs) are now provided in an extended format. They contain 28 rows instead of 26, and also come with intermediate data in addition to the final scoring matrix. The latter will make it possible to directly generate search databases for the current version of RPS-BLAST, DELTA-BLAST, as well as an upcoming new version of RPS-BLAST that supports composition-corrected scoring.

*Please note:* PSSMs been re-computed for a large fraction of the models in CDD, which has slightly affected the resulting sequence annotations.

To use pre-computed PSSMs for formatting RPS-BLAST search databases, the "makeprofiledb" application must be installed. Details on how to run "makeprofiledb" can be found in the [CDD FTP README file](#).

### NCBI Presents Genetic Variation and Medical Resources at the ACMG 2013 Meeting

*Thursday, March 21, 2013*

NCBI staff members are presenting the NIH Genetic Testing Registry (GTR) and other NCBI clinical genetics resources, including MedGen and ClinVar, at the [American College of Medical Genetics and Genomics 2013 Annual Clinical Genetics Meeting](#) in Phoenix, Arizona.

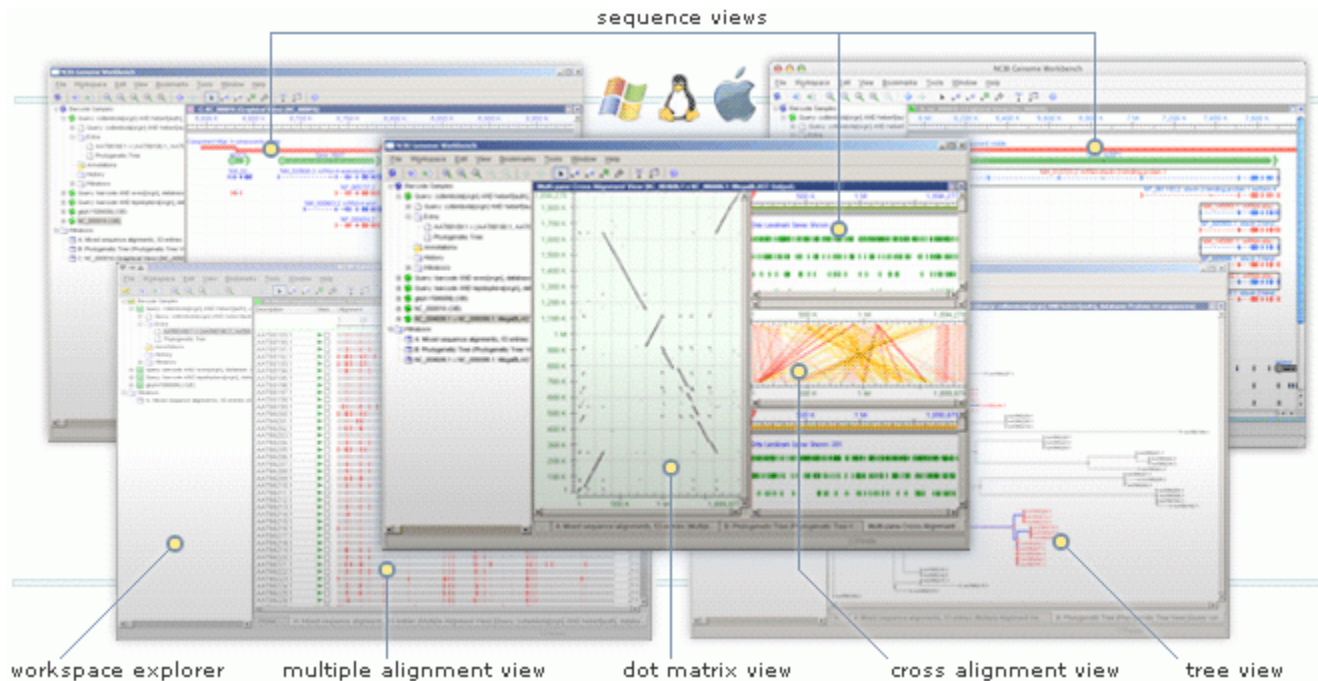
#### ACMG 2013 Presentations:

- Wednesday, March 20, 2013: [The NIH Genetic Testing Registry \(GTR\): Public Meeting on Transition Plan with GeneTests](#)
- Friday, March 22, 2013: [Medical Variation and Phenotypes at NCBI: GTR, MedGen, and ClinVar](#)

### Genome Workbench 2.7.0 Now Available

*Wednesday, March 20, 2013*

Version 2.7.0 of [Genome Workbench](#) is now available with many new features, improvements and some bug fixes.



## From the Genome Workbench Release Notes:

### New Features

- New, reimplemented SNP Table View ([SNP Table View Tutorial](#))
- Graphical Sequence View: implemented new persistent markers on sequences ([Sequence View Markers Tutorial](#))
- Implemented reader of 5-column feature table format
- Text View: Implemented search
- Genome Workbench project can now be opened as Table (Project Tree View, project context menu / Open Table View). New feature allows user to sort and search project content.

### Bug Fixes and Improvements

- Multiple Alignment View: added new column "Organism name" (taxonomic name)
- Multiple Alignment View: fixed issues with BLAST results with multiple queries
- BLAST Search: improved Blast DB selection from the most recently used list
- BLAST: improved parameters selection interface not to allow mutually exclusive options
- Graphical Sequence View: fixed crash with protein alignments
- Graphical Sequence View: fixed crash with visualization of pile-up graph for alignments
- Graphical Sequence View: improved 'Project features for aligned sequences' mode for alignment tracks
- Graphical Sequence View: fixed display of alignments with local ids
- Graphical Sequence View: restored defline as part of sequence track label, sequence track tooltip improved to be pinnable
- Graphical Sequence View: implement 5' end visual tagging (sequence track)
- Splign tool: fixed crash
- Table Import: improved recognition of DOS/Unix line breaks, various small issues fixed
- OpenGL support: resolved issues with some limited functionality graphics drivers, improved work with Remote Desktop



- Tree View: improvements and fixes for PDF printing
- Generic Table View: implemented support for Copy from cell function
- Generic Table View: added support for disabled rows
- Table Views (all): added pop-up tooltip to show full content of extra long cells
- VCF support: improved label generation for VCF imported variants
- Alignment Summary View: fixed number of bugs and issues
- Text View: implemented molecule separation in Flat File mode
- Create Gene Model Tool: added option to propagate ncRNA features
- Status Bar: added explicit warning, when relaxed molecule id comparison is used in broadcasting
- BAM support: fixed crash working with network paths and some incorrect path locations
- Open Dialog: number of small GUI cleanups and improvements, improvements of MacOS copy and paste
- Broadcast Options dialog: save settings for program restart

## RefSeq Release 58 is Available for FTP

*Friday, March 15, 2013*

The [complete RefSeq release 58](#) contains 36,938,203 sequence records from 34,169,407 records including 3,345,543 RNAs, and 30,489,893 proteins from 22,460 different organisms.

Check out [RefSeq's new homepage](#) to learn more about The Project and see the [Release statistics file](#) or [Release notes](#) for more information about this particular release.

## NCBI now provides interim GFF-formatted updates for human and mouse refseq annotations

*Tuesday, March 12, 2013*

The interim updates contain features projected from current RefSeq transcripts and curated genomic sequences and placed on the latest assembly version. The current RefSeqs include transcript variants that are new or have been updated since the last full annotation. The latest assembly version may include additional or updated genome patches compared to the assembly version used for the full annotation.

The [General Feature Format \(GFF\)](#)-formatted updates are available on the FTP site. For example, see the first interim update for the [human assembly GRCh37.p11](#): [ftp://ftp.ncbi.nlm.nih.gov/genomes/H\\_sapiens/GFF\\_interim](ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/GFF_interim).

## Genome Workbench is the Featured Resource in OpenHelix's "Tip of the Week"

*Thursday, March 07, 2013*

Thanks to OpenHelix's Blog site for featuring our Genome Workbench as their "Top of the Week"! They describe it as "a useful program ... (with) a great set of videos to introduce you to the workbench's functions and features."

For more information about Genome Workbench, check out its homepage and ample documentation at: <http://www.ncbi.nlm.nih.gov/tools/gbench/>

## **New Quick Tip on NCBI Insights Blog - How To Format Sequence Data For GenBank Submissions**

*Thursday, March 07, 2013*

A [new Quick Tip](#) on the NCBI Insights blog shows how to properly format FASTA files for submission to GenBank. The post gives step-by-step instructions for submitting single- and multiple-nucleotide sequences.

## NCBI News, February 2013

### New Science Feature on NCBI Insights - Transcriptome of Tasmanian devil and its transmissible cancer

Thursday, February 28, 2013

A new Science Feature is now on NCBI Insights Blog - "The Tasmanian Devil and Cancer as an Infectious Disease: Analysis of transcriptome data." Recent research about a strange and deadly infection that causes a transmissible cancer in the Tasmanian devil has provided considerable data and insights on the mechanism of pathology.

This post shows how to access the relevant literature and these RNA-Seq data through PubMed Central, the BioProjects, and Sequence Read Archive (SRA). It also demonstrates using BLAST and the SRA run browser to analyze expression levels of specific protein coding and miRNA genes in these data sets.

### New Quick Tip on NCBI Insights Blog - how to download bacterial genomes using the Entrez API

Wednesday, February 20, 2013

A new Quick Tip, "How to Download Bacterial Genomes Using the Entrez API" on the NCBI Insights blog shows how to download bacterial genomes programmatically for a list of species using the E-utilities, the application programming interface (API) to NCBI's Entrez system of databases.

This strategy takes advantage of NCBI's redesigned Genome database that links all genome sequences for a given species to one record, making it easy to obtain the desired sequences after finding the right Genome record. The post includes a demonstration script that is easy to adapt to download genomes of interest.

### GenBank Release 194.0 is Available

Tuesday, February 19, 2013

The new release for GenBank is now available via <ftp.ncbi.nlm.nih.gov>, as well as in the [Nucleotide database](#) and [BLAST services](#).

In release 194.0 (02/15/2013), the total number of non-WGS, non-CON records was comprised of basepairs of sequence data. In addition, there were 103,101,291 WGS records containing 390,900,990,416 basepairs of sequence data.

During the 63 days between the close dates for GenBank Releases 193.0 and 194.0, the non-WGS/non-CON portion of GenBank grew by 1,750,490,954 basepairs and by 1,746,402 sequence records, with an average of 45,507 non-WGS/non-CON records added and/or updated per day. In addition, the WGS component of GenBank grew by 34,898,067,578 basepairs and by 10,333,526 sequence records.

The total number of sequence data files increased by 45 with this release, with the divisions that expanded in file number:

- BCT = 4 new files, now a total of 98
- CON = 8 new files, now a total of 187
- ENV = 2 new files, now a total of 59
- GSS = 4 new files, now a total of 270
- INV = 1 new file, now a total of 34

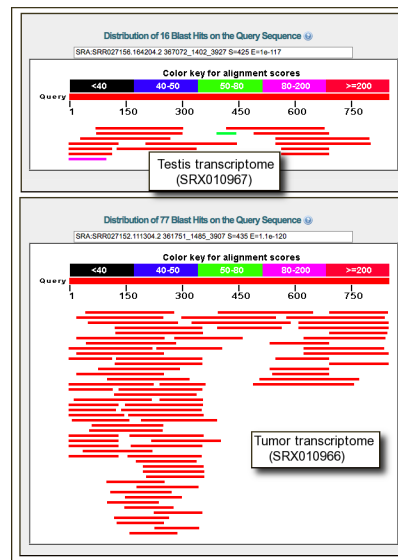


Figure 1. SRA BLAST results using the devil POMC transcript as a query against testis transcriptome data (top panel) and facial tumor transcriptome data (bottom panel). This gene is highly expressed in the tumor sample.

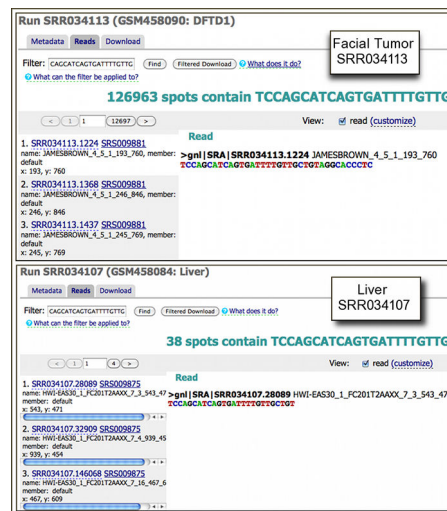


Figure 2. Filtering SRA runs for liver (top panel) and facial tumor (bottom panel) using the stem sequence of MIR338 brain-specific miRNA. High counts for this sequence in the tumor run is consistent with a neural origin for this tumor.

- PAT = 4 new files, now a total of 190
- TSA = 5 new files, now a total of 138
- VRT = 2 new files, now a total of 30

The total number of AUT (author name) index files increased by 4 with this release and is now composed of 110 files.

For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 587 GB (sequence files only) or 632GB (including the 'short directory', 'index' and \*.txt files) when uncompressed. The ASN.1 formatted datafiles are approximately 481 GB.

For additional release information, see the [Release Notes](#) and README files in individual directories.

## One change is noteworthy of mentioning here as described in the Release Notes sections 1.3.2 & 1.3.3:

With GenBank 193.0 experimental "Release Catalog" products were produced in a plan to eventually replace the old GenBank "index" files. GenBank release 193.0 was the last release for which the legacy index files will be provided.

Shortly, Release Catalogs and other supplemental files for Release 194.0 will be made available. In these files every GenBank sequence record will be represented by a 10-field, TAB-delimited row of data. *[Please note that when a value does not exist for one of these fields, the field in the catalog will be empty (eg, two sequential TAB characters can be present, with nothing between them.)]*

The new files will be made available in a new sub-directory of the GenBank FTP area: <ftp.ncbi.nlm.nih.gov/genbank/catalog>.

The data fields in the Release Catalog files are:

- Accession Number
- Accession.Version
- NCBI GI Identifier (if assigned)
- Molecule Type (dna, rna, mrna, etc)
- Sequence Length
- Organism Name
- NCBI Taxonomy Database Identifier
- Division Code
- BioProject Accession Number
- BioSample Accession Number

Here is an example of a row of Release Catalog data for the entry CP003933:

- CP003933 CP003933.1 429549985 dna 3618794 Sinorhizobium meliloti GR4 1235461 BCT PRJNA175860

In addition, new "PMID List" and "Gene List" TAB-delimited files will accompany the Release Catalog.

The format of the PubMed Identifier List file is:

- Accession1 Accession1.Version PMID-1,PMID-2,PMID-3,.....
- Accession2 Accession2.Version PMID-1,PMID-2,PMID-3,.....
- ....

The format of the Gene List file is:

- Accession1 Accession1.Version Gene-Symbol-1 Locus-Tag-1
- Accession1 Accession1.Version Gene-Symbol-2 Locus-Tag-2
- ....

We plan to provide the release catalog and accompanying lists via files that are specific to EST, GSS, and non-EST/GSS (everything else), however at least initially the catalogs and lists will not include the contig sequence records for WGS projects.

## NCBI Insights' First Quick Tip: How to find functional protein homologs using conserved domains

Tuesday, February 12, 2013

Protein researchers often want find homologous proteins in different species. In the first "Quick Tips & Tricks" post of the NCBI Insights Blog, "[Using Conserved Domains to Find Functional Homologs](#)" describes a step-by-step method for how to do this using curated functional domain information and links provided in the Conserved Domains Database (CDD).

## **New NCBI Insights Blog Explains the IE7 Warning**

*Tuesday, February 05, 2013*

For many months, people viewing NCBI webpages using the web browser Internet Explorer 7 have seen a warning on the top of their webpages. A new NCBI Insights Blog explains to users what the Internet Explorer 7 warning means.

The [NCBI Insights blog post](#) explains:

- Why the IE7 web browser is no longer supported.
- What "stop supporting this browser" means.
- What actually happened on the NCBI website on January 1, 2013.
- What this really means about for people using IE7 to view the NCBI website.

Links are provided for more information and to the most up-to-date web browsers.

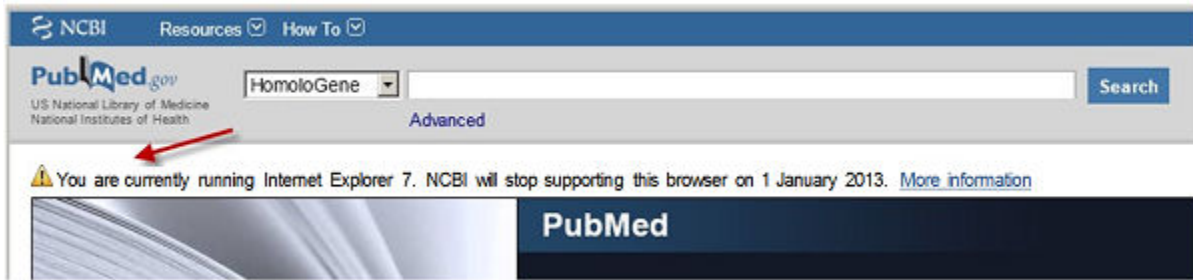
Click here to read the article: "[What does NCBI's Internet Explorer 7 warning mean?](#)".

## **PubReader Article View Now In Use By KoreaMed Synapse**

*Friday, February 01, 2013*

KoreaMed Synapse (<http://synapse.koreamed.org>), a digital archive and reference linking platform of Korean medical journals, is now using NCBI's new [PubReader](#) presentation style to display their full-text journal articles.

KoreaMed's database of over 120 journals now includes a blue 'PubReader' icon for each full-text article. NCBI [launched PubReader in December 2012](#) as a convenient new way to view full-text articles in PubMed Central on desktops as well as tablets and mobile devices. In tandem with the launch, NCBI made the code used to create PubReader freely available on GitHub at <https://github.com/NCBITools/PubReader>.



This message, seen by people viewing NCBI webpages using Internet Explorer 7, has caused some concern among some users about exactly what changed on January 1, 2013 and wondering whether or not they would still be able to access PubMed and other NCBI resources.





## NCBI News, January 2013

### First NCBI Blog Post Highlights New PubReader For PMC Articles

Thursday, January 31, 2013

The first post to the *NCBI Insights* blog focuses on the recently launched PubReader view for full-text articles in PubMed Central.

The PubReader makes full-text research papers not only more readable but also more portable. Any article that is available in full-text HTML in PubMed Central is viewable in the PubReader format. The PubReader code is freely available to developers via GitHub.

Read about it here: <http://ncbiinsights.ncbi.nlm.nih.gov/2013/01/29/new-pubreader-view-for-full-text-articles>.

### Now Available: NCBI Insights Blog!

Monday, January 28, 2013

The new *NCBI Insights Blog* was created to provide an insider's perspective, help our users better understand us and our resources, explore issues of scientific interest that drive our resource development, and demonstrate how you can use our resources to help enhance your research.



We will post articles in four categories:

- *NCBI Explained* - provides an insider's perspective on our resources and policies to help you better understand us and avoid some common misconceptions and misunderstandings.
- *What's New* - introduces our new and updated resources and include specific examples that demonstrate how you can use these to enhance your research.
- *Quick Tips & Tricks* - explains hows to perform specific tasks using our website. Selected topics will be chosen based on questions you have asked and suggestions you have provided.
- *Science Features* - explores current topics in science and demonstrate how you can find relevant data or resources on our website for further exploration.

This blog is a complement to our existing [education and outreach efforts](#), such as News and Social Media publicity, Webinar and Workshop training programs, and Help Desk user support.

Be sure to check in to the *NCBI Insights Blog* every week or so and let us know what you think!

### ~300,000 ChemAxon Structures are now in PubChem

Wednesday, January 23, 2013

Just over 300,000 structures from ChemAxon's [chemicalize.org](#) database are now available in PubChem, including approximately 62,000 novel structures. All structures have links back to [chemicalize.org](#) data pages which list predicted data including pKa, logP/D, names and identifiers, and much more.

The [announcement from ChemAxon](#) provides additional details.

## Genetic Testing Registry Records will list Molecular Pathology CPT Codes


*Thursday, January 17, 2013*

Under a recent agreement with the American Medical Association (AMA), the National Institute of Health's [Genetic Testing Registry \(GTR\)](#) now integrates the AMA's Current Procedural Terminology (CPT) codes for molecular pathology tests into records describing genetic test information by providers.

The addition of the [molecular pathology CPT codes](#), along with [SNOMED CT](#), [LOINC](#), and [IHTSDO](#) clinical terminology, further enhances GTR's interoperability with electronic medical records and laboratory information management systems.

## Come to the NCBI Discovery Workshops on February 4 & 5!

*Wednesday, January 16, 2013*

Spaces are still available for the free, 2-day [Discovery Workshops](#) to be held on the [NIH Campus](#). Each hands-on session emphasizes a different set of NCBI resources using specific examples to highlight important features of the resources and tools and to demonstrate how to accomplish common tasks. 

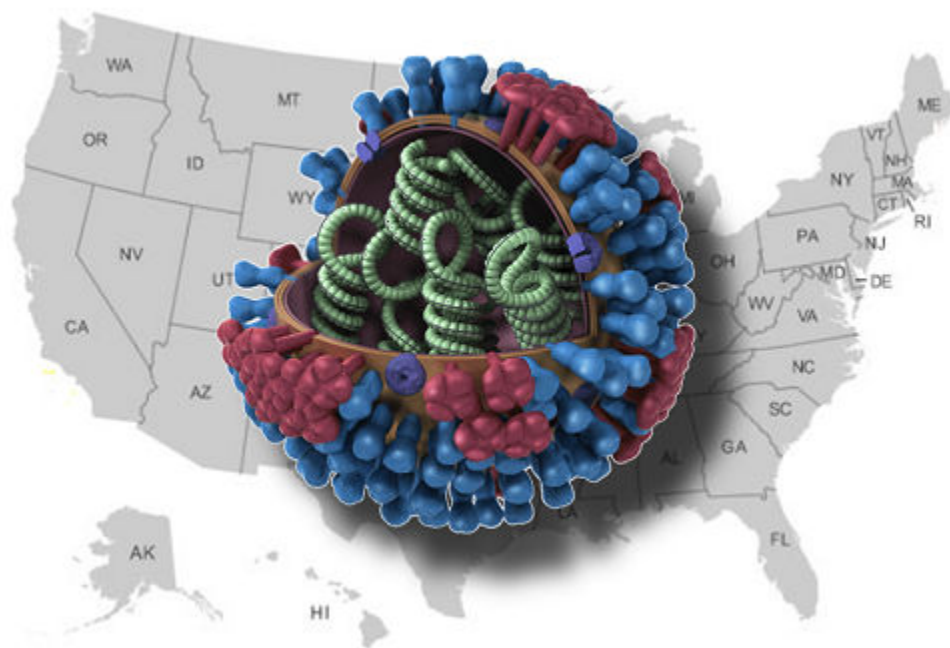
## Now in GenBank: Flu Sequences from the Current Influenza Season

*Tuesday, January 15, 2013*

The U.S. Naval Health Research Center, San Diego and the U.S. Air Force School of Aerospace Medicine's Epidemiology Laboratory Service recently sequenced 46 influenza A and B viruses and has released them in GenBank.

The sequences were collected through their global surveillance programs as recently as November 2012 and represent the only human flu sequences of the current season available in GenBank. The sequences have GenBank accession numbers [CY131965-CY131967](#) and [CY130158-CY130200](#).

According to CDC, the U.S. is seeing an early flu season this year; 47 states have reported widespread flu for the first week in 2013, and the State of New York has declared an influenza emergency. The timely release of influenza sequences is important for researchers trying to understand the viruses circulating in the US and is critical for vaccine development. Additional sequences are expected later this week. The US military's fast action in releasing these invaluable data sets an exceptional example of data sharing for the influenza research community.



## NIH Online Magazine features NCBI Researcher Teresa Przytycka

*Tuesday, January 15, 2013*

NIH's Intramural Research Program's online magazine, "Research in Action," has published a feature story on Teresa M. Przytycka, Ph.D. and her research team entitled "[Algorithms for Life.](#)"



Dr. Teresa Przytycka is a Senior Investigator and Head of the Algorithmic Methods in Computational and Systems Biology Section in NCBI's [Computational Biology Branch](#). For more information on her research and team, please see her [Lab's webpage](#) and a [YouTube video](#) made to accompany the NIH online magazine's article:



## RefSeq Release 57 is Available for FTP

*Monday, January 14, 2013*

The complete RefSeq release 57 contains 34,169,407 records including 3,267,605 RNAs, and 27,845,459 proteins from 21,415 different organisms.

See the [Release statistics file](#) or [Release notes](#) for more information.

Please note that this update includes information from [dbSNP Build 137](#), and includes incremental updates for [human records](#).

In addition, in the first quarter of 2013 the bacterial RefSeq collection has been expanded to include more microbial genomes that represent complete or draft assemblies from novel microbial isolates as well as clinical and population samples. As part of this expansion, bacterial RefSeq genomes will be re-annotated to increase consistency across this dataset.

## A New Eukaryotic Genome Annotation Status Page Keeps Researchers Informed

*Thursday, January 10, 2013*

Researchers often wonder "What's the status of a new build or annotation for my organism's genome?" You can now find out on the new [Eukaryotic Annotation Status Page](#).

## New Rat Genome Available in the MapViewer

*Wednesday, January 09, 2013*

The new *Rattus norvegicus* genome build (5.1) and annotation release (103) are now available in the [MapViewer](#). Also, check out the [Genome page](#) listing additional information and BioProjects for the [Norway Rat](#).

**Eukaryotic RefSeq Genome Annotation Status**

NCBI uses an automated pipeline to provide annotation on some RefSeq genome records (see more information about the annotation process). This page provides information about:

- Eukaryotic RefSeq genomes currently in the NCBI annotation pipeline
- Eukaryotic RefSeq genome annotations that were recently released

**Which genomes are annotated?**

Only genomes with assemblies that are published in the NCBI database are eligible for annotation. NCBI makes this selection based on several factors:

- NIH/NCBI priorities:** Mammals are a top priority.
- Assembly quality:** Assemblies with high N50 > 50,000 bases and/or a scaffold N50 > 50,000 bases are preferred.
- Community interest requests:** (See [Request a RefSeq genome](#))
- Biological, evolutionary, or economic importance:**
- Public availability of supporting transcript data:**
- Availability of gene annotation on the RefSeq record:** that annotates on- or off- the RefSeq record. Exceptions include:
  - NCBI always generates annotation for non-mammalian vertebrates and invertebrates.
  - NCBI annotation of plant genomes is available on the RefSeq record.

**Annotation runs in progress**

An annotation run is marked in progress until the data produced is available in the sequence databases, in [Gene](#) and on the [FTP](#) site.

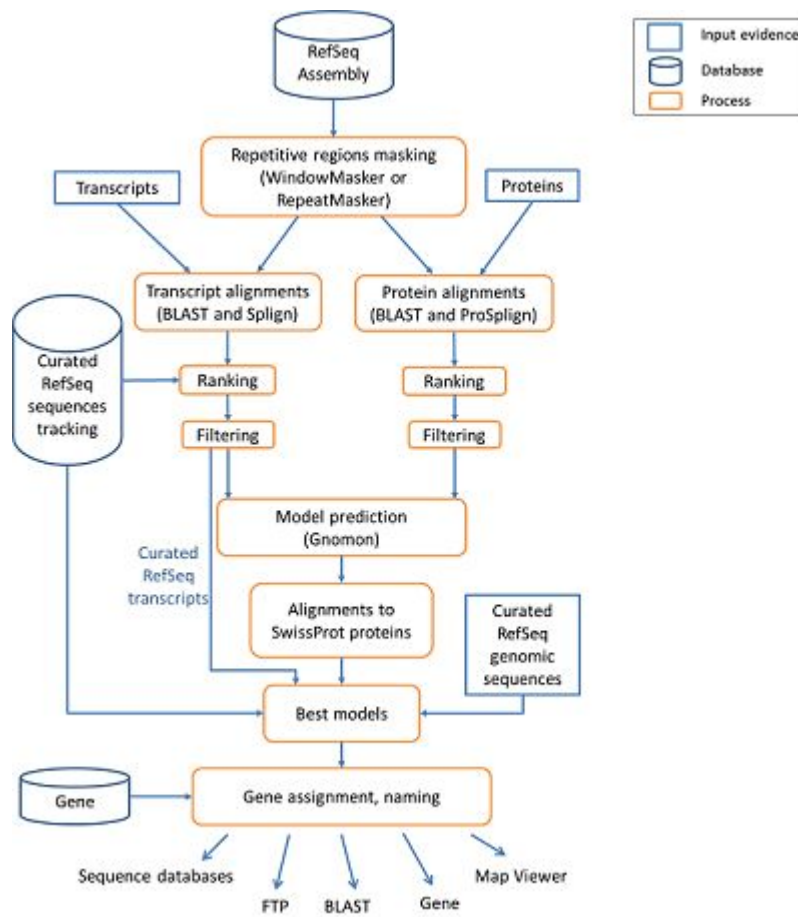
Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Status
<i>Domesticus leucogenus (red-throated white-eared nighthawk)</i>	Dom_1.0	101	2012-12-21	Automated processing in progress

**Recently completed annotation runs**

Annotation runs that were completed within the last year:

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links
<i>Oryza sativa (japanese medlar)</i>	ASM1367v1 (GCF_00013675.1)	100	2012-12-16	2012-12-21	<a href="#">FTP</a>
<i>Oryza sativa (javanica)</i>	Osir_v3.1 (GCF_000298735.1)	100	2012-11-22	2012-12-02	<a href="#">FTP</a>
<i>Gorilla gorilla (western gorilla)</i>	gorGor2.1 (GCF_000151805.1)	100	2012-11-20	2012-12-06	<a href="#">FTP</a>
<i>Felis catus (domestic cat)</i>	Felis_catus_5.2 (GCF_000161305.1)	100	2012-11-01	2012-11-07	<a href="#">FTP</a>

**The NCBI Eukaryotic Genome Annotation Status Page.** This new webpage explains the rationale for "Which genomes are annotated?", and lists "Recently completed annotation runs" and "Annotation runs in progress" along with links to the data and Assembly statistics.



**The NCBI Eukaryotic Genome Annotation Pipeline.** This process provides content for various NCBI resources including Nucleotide, Protein and Genome databases, the BLAST sequence similarity algorithm, and the MapViewer genome browser. For more information, see the [Eukaryotic Genome Annotation Process page](#).



## NCBI News, December 2012

### NCBI Introduces PubReader, a New View for Full-Text Articles

*Wednesday, December 19, 2012*

As announced recently in the National Library of Medicine's [Technical Bulletin](#), NCBI now offers PubReader – a new, reader-friendly display format for full-text articles in the PubMed Central (PMC) literature database.

Leveraging the features of the latest underlying web technologies (HTML5 and CSS3), PubReader addresses usability and readability challenges specific to viewing research articles on tablets and other small-screen devices. PubReader also works on desktop and laptop computers.

Any article that is available in full-text HTML in PMC is viewable in the PubReader format. Furthermore, the PubReader functions with any of the [latest browsers](#) without the need to download an app or any additional software.

Like a printed journal article, PubReader breaks a document into multiple columns and pages to improve readability and navigation. PubReader can expand a page to whatever fits your screen, with multiple columns on a desktop monitor or a single column page on a small tablet. It will even switch to two columns if you rotate the tablet to a landscape view. When you adjust the font size or change the size of the browser window, page boundaries and columns adjust automatically.

The PubReader presentation shown in Figure 1 offers a variety of common options for moving between pages. You can use the PageUp, PageDown, RightArrow, LeftArrow keys on a keyboard, tap or click in the right or left margin (Figure 1A), use finger swipes on a touch screen device, or use the progress bar (Figure 1B) at the bottom of the screen to jump across the page range.

There is an image strip (Figure 1C) at the bottom of the PubReader page with thumbnails of all figures and tables in the article, allowing you to pop up an earlier figure/table and then close it in an instant without losing your place in the article. This same feature works with inline figures, as well as tables and citations. You will discover that PubReader has a number of other features to improve your reading experience as you use it in PMC.

You can read more about the PubReader view on the PubReader [about page](#). You can try it directly with an example record (PMCID: [3396517](#)) or by clicking on the “PubReader” link for an article in a PMC search result list or in the article itself as shown in Figure 2.

The CSS and JavaScript code used to create the PubReader display are freely available from [NCBITools](#) on the public code repository GitHub. Anyone can use or adapt this code to display journal articles or other content that is structured as an HTML5 document.

### GenBank Release 193.0 is Available

*Monday, December 17, 2012*

The new release for [GenBank](#) is now available via <ftp.ncbi.nlm.nih.gov>, as well as in the [Nucleotide database](#) and [Blast services](#).

In release 193.0 (12/12/2012), the total number of non-WGS, non-CON records was 161,140,325 comprised of 148,390,863,904 basepairs of sequence data. In addition, there were 92,767,765 WGS records containing 356,002,922,838 basepairs of sequence data.

During the 65 days between the close dates for GenBank Releases 192.0 and 193.0, 830,113 records were updated with an average of 62,780 non-WGS/non-CON records added and/or updated per day.

Mapping the HLA-DO/HLA-DM complex by FRET and mutagenesis  
Proc Natl Acad Sci U S A. 2012 July 10; 109(28): 11276–11281.

Proceedings of the National Academy of Sciences of the United States of America  
National Academy of Sciences

**Mapping the HLA-DO/HLA-DM complex by FRET and mutagenesis**

Taejin Yoon, Henriette Macmillan, Elizabeth D. Mellins

Additional article information

**ABSTRACT**

HLA-DO (DO) is a nonclassical class II heterodimer that inhibits peptide exchange on DM, and influences DM localization to late endosomes and exosomes. DM acts as a chaperone for DO, and is required for its egress from the endoplasmic reticulum (ER). These reciprocal functions are based on direct DO/DM binding, but the topology of DO/DM complexes is not known, in part, because of technical limitations stemming from DO instability. We generated two variants of recombinant soluble DO with increased stability [zippered DO $\alpha$ P11A (szDOv) and chimeric sDO-Fc] and confirmed their conformational integrity and

**Fig. 2.** In vitro FRET with szDOv and sDM. (A) Locations of introduced free Cys on DO (blue circles) and natural free Cys on DM (red circle). The DO structure was simulated, based on HLA-DR1 (Protein Data Bank ID code 1A4D); DO $\alpha$  (cyan); DO $\beta$  (purple). ...

Page 1 of 24

www.ncbi.nlm.nih.gov/pmc/articles/PMC3396517/figure/fig02/

Figure 1. PubReader display of the first screen of PMC3396517 as seen on a standard desktop display. One of the figures in the image strip (C) in the document is selected popping up an enlarged version. A. Clicking the right margin advances to the next screen. Clicking on the icon at B toggles between the image strip (C) and a linear progress bar (not) shown.

Journal List • Proc Natl Acad Sci U S A • v 109(28); Jul 10, 2012 • PMC3396517

PNAS  
Proceedings of the National Academy of Sciences of the United States of America

Proc Natl Acad Sci U S A. 2012 July 10; 109(28): 11276–11281.  
Published online 2012 June 25. doi: 10.1073/pnas.1113965109  
Immunology

PMCID: PMC3396517

**Mapping the HLA-DO/HLA-DM complex by FRET and mutagenesis**

Taejin Yoon,<sup>1,2</sup> Henriette Macmillan,<sup>2</sup> Sarah E. Mortimer,<sup>2</sup> Wei Jiang,<sup>2</sup> Cornelia H. Rinderknecht,<sup>1,2</sup> Lawrence J. Stern,<sup>1</sup> and Elizabeth D. Mellins<sup>1,2,3</sup>

Author Information • Copyright and License Information ►

**ABSTRACT**

HLA-DO (DO) is a nonclassical class II heterodimer that inhibits peptide exchange on DM, HLA-DM (DM), and influences DM localization to late endosomes and exosomes. DM acts as a chaperone for DO, and is required for its egress from the endoplasmic reticulum (ER). These reciprocal functions are based on direct DO/DM binding, but the topology of DO/DM complexes is not known, in part, because of technical limitations stemming from DO instability. We generated two variants of recombinant soluble DO with increased stability [zippered DO $\alpha$ P11A (szDOv) and chimeric sDO-Fc] and confirmed their conformational integrity and

**Formats:**  
Abstract | Article | **PubReader** | ePub (beta) | PDF (1.2M)

**Related citations in PubMed**

A point mutation in the groove of HLA-DO allows egress from the endoplasmic reticulum lumen [Proc Natl Acad Sci U S A. 2005]  
Secondary structure composition and pH-dependent conformational changes of soluble recombinant HLA-DM [J Biol Chem. 1998]  
"Chemical analogues" of HLA-DM can induce a peptide-receptive state in HLA-DR molecules. [J Biol Chem. 2004]  
The role of H2-O and HLA-DO in major histocompatibility complex class II-restricted antigen processing [Immunol Rev. 1999]  
How HLA-DM affects the peptide repertoire bound to HLA-DR [Immunity. 1997]

Mapping the HLA-DO/HLA-DM complex by FRET and mutagenesis  
Taejin Yoon, Henriette Macmillan, Sarah E. Mortimer, Wei Jiang, Cornelia H. Rinderknecht, Lawrence J. Stern, Elizabeth D. Mellins  
Proc Natl Acad Sci U S A. 2012 July 10; 109(28): 11276–11281. Published online 2012 June 25.  
doi: 10.1073/pnas.1113965109  
PMCID: PMC3396517  
Article | **PubReader** | PDF-1.2M | Supplementary Material

Figure 2. Accessing the PubReader from standard record views in PMC. The link is in the Formats section in the full PMC article (Top Panel) and under the summary in search results (Bottom Panel).

- The non-WGS/non-CON portion of GenBank grew by almost 3 billion basepairs and by more than 3.25 million sequence records
- The WGS component of GenBank grew by over 22 billion basepairs and by more than 6.3 million sequence records



The total number of sequence data files increased by 45 with this release, with the divisions that expanded in file number:

- BCT = 3 new files, now a total of 94
- CON = 3 new files, now a total of 179
- EST = 3 new files, now a total of 469
- GSS = 6 new files, now a total of 166
- PAT = 4 new files, now a total of 186
- PLN = 1 new file, now a total of 60
- ROD = 1 new file, now a total of 30
- TSA = 21 new files, now a total of 133
- VRL = 2 new files, now a total of 24
- VRT = 1 new file, now a total of 28

The total number of AUT (author name) index files increased by 4 with this release and is now composed of 106 files.

For downloading purposes, please keep in mind that these GenBank flatfiles are roughly 579 GB (sequence files only) or 624 GB (including the 'short directory', 'index' and the \*.txt files) when uncompressed. The ASN.1 formatted datafiles are approximately 474 GB.

For additional release information, see the [Release Notes](#) and README files in individual directories.

## Now in PubChem: 8+ million Patented Chemicals from the SureChem Database

*Monday, December 10, 2012*

The [SureChem patent chemistry database](#) has deposited chemistry information for their complete collection of US, EP and WO full text patents from the 1976 to the present in [PubChem](#).

More than [8 million structures](#) are now available in the [PubChem Substance database](#). Over 4 million of these structures are new to the PubChem Compound database which greatly expands public access to novel medicinal chemistry. For more information about each substance, users can click on the External ID link listed on the PubChem Substance records to go to the corresponding record SureChem's website.

## SRA Surpasses a PetaBase of Sequence Data

*Tuesday, December 04, 2012*

The [SRA database](#) now contains more than a PetaBase ( $1 \times 10^{15}$  bases) of sequence data, including more than 390 TeraBases of open-access data from INSDC partners NCBI, EBI and DDBJ and more than 610 TeraBases of controlled-access human clinical sequence data.

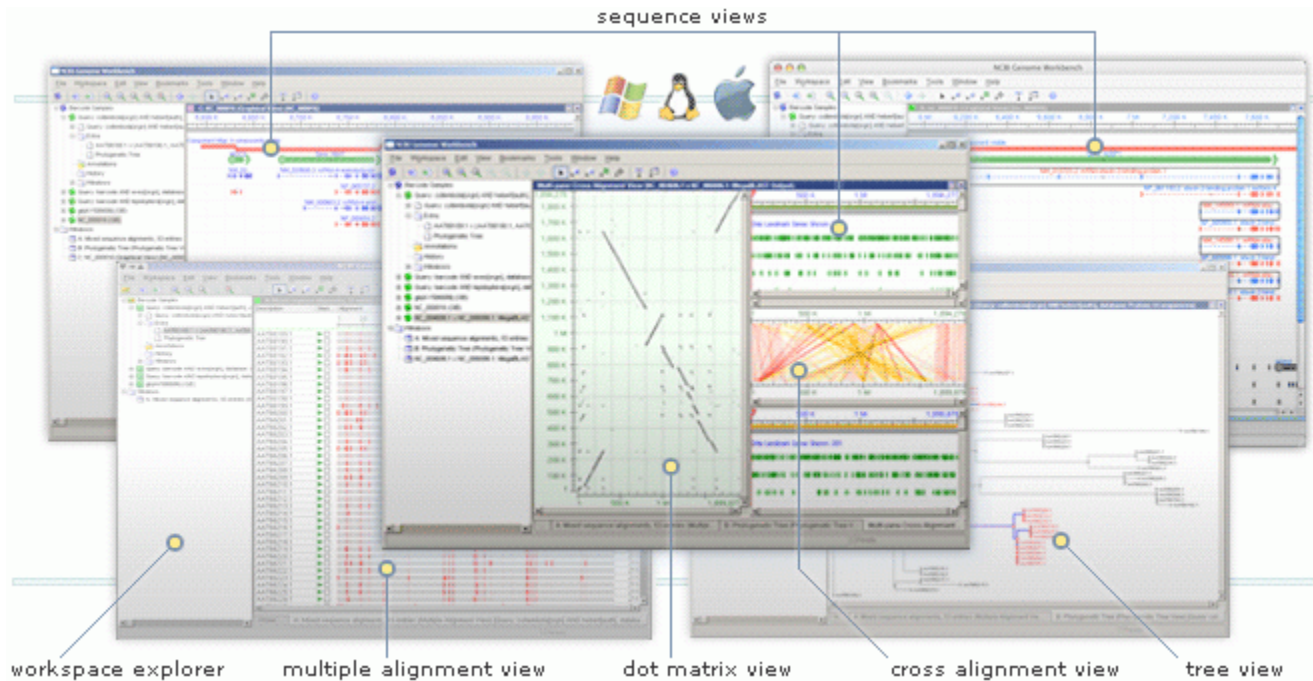


## NCBI News, November 2012

### New version of Genome Workbench is Available

Monday, November 19, 2012

Version 2.6.5 of [Genome Workbench](#) has been released with some new features and many improvements and bug fixes.



### From the [Genome Workbench Release Notes](#):

#### New Features

- Add Tools Quick Launch items to context menu
- Graphical Sequence View: implemented separation graphs into sub-tracks based on meta-information title

#### Bug Fixes and Improvements

- Updated VCF and GFF readers, fixed issues with checking sequence ids
- Connection issues resolved for users running international versions of Windows or MacOS
- BLAST tool dialog freeze for a long time loading list of available BLAST databases
- BLAST RID load: fixed to automatically import both alignments and local sequences
- MUSCLE tool integration: fixed alignment coordinate remapping(shift error)
- Graphical Sequence View: fixed display of tracks with alt-loci alignments
- Graphical Sequence View: fixed fixed alignment score coloration for SRA
- Graphical Sequence View: tooltips improved alignment statistics and positions
- Graphical Sequence View: fixed indels rendering in protein-to-genomic alignments
- Graphical Sequence View: fixed empty tooltip lines (sequence)
- Multiple Alignment View: fixed tooltips to be displayed only over sequence part
- Linux: improved README for users building custom Genome Workbench
- Search View: fixed failures handling multiple sequence ids

- Search View: Kozak patterns updated
- BAM/cSRA support: fixed Open File masking for MacOS and Linux
- Text View: fixed crash in Flat File mode
- Tree View: fixed view title after tree sort operation
- MacOS: fixed opening projects from network paths/mappings/mountpoints
- MacOS: fixed issues with pasting into Open View dialog
- HTTP connection component: fixed "User-Agent:" self-identification to better conform to standards
- Improved logging and Feedback reporting not to collect potentially sensitive information

## RefSeq Release 56 is Available for FTP

*Wednesday, November 14, 2012*

The [complete RefSeq release 56](#) contains 23,892,460 records including 2,729,041 genomic sequences, 3,185,652 RNAs, and 17,977,767 proteins from 18,512 different organisms.

See the [Release statistics file](#) or [Release notes](#) for more information.

Please note that this update includes information from [dbSNP Build 137](#), and includes incremental updates for [human records](#).

In addition, in the first quarter of 2013 the bacterial RefSeq collection will expand to include more genomes that represent complete or draft assemblies from novel microbial isolates as well as clinical and population samples. As part of this expansion, bacterial RefSeq genomes will be re-annotated to increase consistency across this dataset.

## New CDD release Available & now Mirrors TIGRFAM v13

*Monday, November 05, 2012*

Version 3.09 of the [Conserved Domain Database \(CDD\)](#) is now available. 46,629 conserved domain models are indexed for searching at NCBI and include a mirror of [TIGRFAM v13](#).

The new data are incorporated in the [CDD website](#) and [Conserved Domain BLAST Search](#) services. More detailed statistics are available from the [CDD News page](#). CDD matrices and other information can be downloaded from the [FTP site](#).

## NCBI will be Presenting and Exhibiting at ASHG 2012

*Monday, November 05, 2012*

NCBI Staff members will present, display posters, run a workshop and exhibit at the [ASHG 2012 Annual Meeting](#) in the [Moscone Center](#), San Francisco, CA which takes place from Tuesday, November 6th through Saturday, November 10th, 2012.

Wednesday, November 7 through Friday, November 9, 2012 from 10am-4:30pm:

- Come visit us at the NCBI Booth (#224) in the Exhibit Hall, Lower Level South
- *NCBI staff will be available to answer questions, listen to your suggestions, and offer live demonstrations with NCBI tools and databases.*

Presentations on Tuesday, November 6, 2012

- 9-9:15am: "Managing information about human phenotype at NCBI" in the "Getting Ready for The Human Phenome Project" Satellite Meeting - San Francisco Marriott Marquis, Section 7, Yerba Buena Ballroom, Lower Level B2
- 1-4pm: "Getting the Most from the Human Genome: Understanding Updates and Improvements in the Reference Assembly" in the "Genome Reference Consortium Workshop" Moscone Center, Room: 236/238, East Mezzanine Level South

#### Workshop on Wednesday, November 7, 2012

- 12:15-2:15pm: " Workshop: Discovering Biological Data at NCBI" in the Moscone Center, Room: 304/306, Esplanade Level South
- *This workshop provides an introduction to using the Entrez system to perform searches and find related data starting with a list of reviewed human genes. Specific tasks covered include finding reference sequences, mapping variations, identifying homologous genes, exploring expression studies, and using MyNCBI to save searches and manage data.*

#### Presentations on Saturday, November 10, 2012

- 9:40-11:40am: " Improving the accuracy of variant identification" & "Introducing ClinVar" in the "Centralizing the Deposition and Curation of Human Mutations" Section, Moscone Center, Room: 132, Lower Level North

For the full list of NCBI's events, see the ["NCBI at ASHG 2012" Schedule](#) page.



## NCBI News, October 2012

### Human CCDS Release 11 Issued

*Wednesday, October 31, 2012*

This update adds 1138 new and 2 reinstated records which add sequences for 93 genes to the Human Consensus CoDing Sequences (CCDS) dataset. CCDS Release 11 includes a total of 27,511 CCDS IDs that correspond to 18,535 Gene IDs.

The NCBI, Ensembl, and Sanger (Havana) annotation of the human reference genome (assembly GCF\_000001405.21, NCBI annotation release 103, Ensembl annotation release 68) was analyzed to identify additional coding sequences (CDS) that are consistently annotated.

CCDS data is available in the [CCDS web site](#) and [FTP site](#) and will become available in the collaborators' genome and/or gene browser web sites according to each browser's update cycle. See the [Releases & Statistics report](#) for details.

### NCBI's Genetic Testing Registry at AMP's Annual Genomic Medicine Meeting

*Wednesday, October 24, 2012*

The Head of the NIH Genetic Testing Registry is presenting and is available for questions at the [Association for Molecular Pathology's Annual Meeting on Genomic Medicine](#) in Long Beach, CA from October 25-27, 2012.

Tuesday, October 23rd: GeT-RM Panel Meeting 2012 at 3-5pm in the Naples Ballroom

- CDC/NCBI Clinical NextGen Sequencing Reference Material Project Open Forum
- Demonstration of NCBI's GeT-RM browser and preliminary data analysis

Wednesday, October 24th: Open Forum at 8am-6pm in the Sienna room

- Come by and get help registering your lab with the NIH Genetic Testing Registry!

### NCBI Genetic Counselors at the National Society for Genetic Counselors' Meeting

*Monday, October 22, 2012*

NCBI Genetic Counselors are presenting at the [National Society for Genetic Counselors' 2012 Annual Education Conference](#) in Boston, MA from October 24-27, 2012.

Plenary Session: "Clinical Genetic Resources at NCBI"

Thursday, October 25th @ 4:30-5:15pm in Room 302/304/306

- Discover how to become a power user of the GTR with MyNCBI and other navigation tips.
- Demonstrate how to find a test by a variant, and identify the variant's clinical significance.
- List at least four databases at NCBI of use to genetics professionals.

Learn more about the Genetic Testing Registry at two Open Forums:

- Thursday, October 25 @ 7-10pm in Room 208: "Get help to register your lab and tests"
- Friday, October 26 @ 7-10pm in Room 205: "Learn to navigate the Site"

## New dbSNP Release for Mouse and Cow

Tuesday, October 16, 2012

A new dbSNP build 137 has been released for Mouse and Cow with data indexed in Entrez, available by FTP, and mapped to genomic assemblies.

New Mouse information:

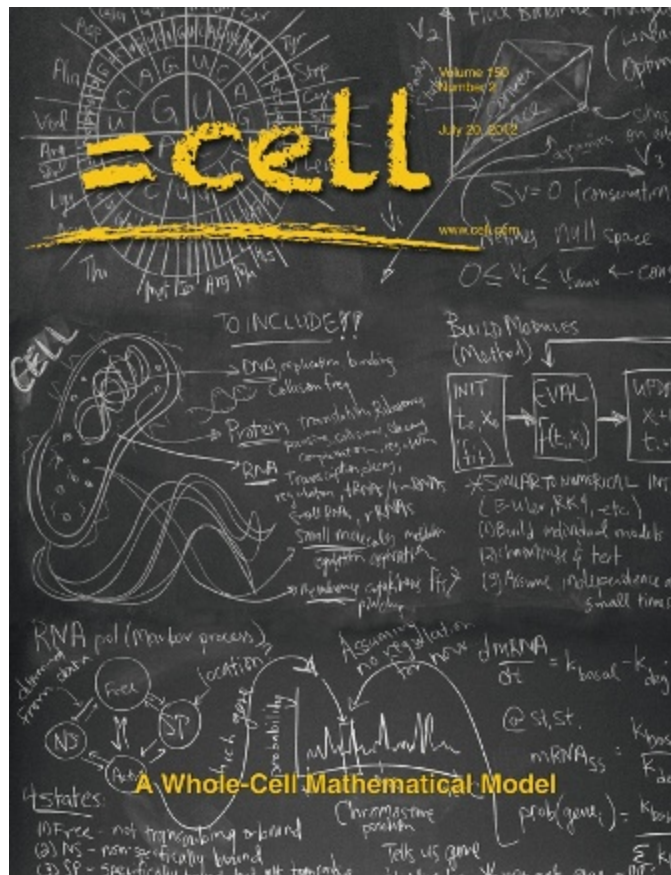
- [Mapped Assemblies on the FTP site](#): GRCm38 (GCF\_000001635.20) and Mm\_Celera (GCF\_000002165.2)
- New RS as shown in Entrez: 54562956
- Total RS as shown in Entrez: 70036850

New Cow information:

- [Mapped Assemblies on the FTP site](#): Btau\_4.6.1 (GCF\_000003205.5) and Bos\_taurus\_UMD\_3.1 (GCF\_000003055.4)
- New RS as shown in Entrez: 5522517
- Total RS as shown in Entrez: 26539698

## NLM-Funded Investigator Creates First Complete Computerized Simulation of an Organism

Wednesday, October 10, 2012



A team led by Markus Covert, PhD, a bioengineering professor at Stanford, used data compiled from more than 900 scientific papers to construct the first complete computer model of an organism, *Mycoplasma genitalium*. Dr.



Covert was the recipient of a 2009 [NIH Pioneer Award](#), jointly sponsored by the National Library of Medicine (NLM) and the [NIH Common Fund](#).

This ground-breaking achievement in computational biology earned the cover of the July 20, 2012 issue of *Cell* and represents a "transforming approach to answering questions about fundamental biological processes," according to James M. Anderson, MD, PhD, director of NIH's Division of Program Coordination, Planning and Strategic Initiatives. The single-cell bacterium was chosen for its relative simplicity; it has 525 genes, compared to *E. coli*'s 4,288. The model runs on 128 computers and:

- Describes the life cycle of a single cell from the level of individual molecules and their interactions
- Accounts for the specific function of every annotated gene product
- Accurately predicts a wide range of observable behaviors

In a [July 20, 2012 article in \*The New York Times\*](#), Covert described the object-oriented approach his team used to design the 28 separate modules that represent *M. genitalium*'s biological processes: "The major modeling insight we had a few years ago was to break up the functionality of the cell into subgroups, which we could model individually, each with its own mathematics, and then to integrate these submodels together into a whole."

This approach uses more than 1,900 parameters observed experimentally and reported in 900+ articles, and integrates them in a manner that enables understanding and provides direction for real-world experiments. "If you use a model to guide your experiments, you're going to discover things faster. We've shown that time and time again," said Covert. NIH's Anderson is clear on the significance: "[Comprehensive computer models of entire cells have the potential to advance our understanding of cellular function and, ultimately, to inform new approaches for the diagnosis and treatment of disease.](#)"

#### Reference:

Karr JR, Sanghvi JC, Macklin DN, Gutschow MV, Jacobs JM, Bolival B Jr, Assad-Garcia N, Glass JI, Covert MW. A whole-cell computational model predicts phenotype from genotype. *Cell*. 2012 Jul 20;150(2):389-401. Available from: <http://www.sciencedirect.com/science/article/pii/S0092867412007763> PubMed PMID: 22817898

To learn more about the NLM Extramural Programs (grants) Division: <http://www.nlm.nih.gov/ep/ep-overview.html>

## Try the new PubChem Classification Browser!

Tuesday, October 09, 2012

Use the new [PubChem Classification Browser](#) to search for PubChem data using a specific classification term from MeSH, Gene Ontology, KEGG, ChEBI or LIPID MAPS.

The [Help Document](#) includes [examples](#) of how the PubChem Classification Browser can be used to find PubChem Compounds classified as methotrexate as well as for identifying PubChem BioAssays that have tested protein targets involved in DNA repair.

## New PubChem Widgets: embed PubChem tables in your own web pages!

Monday, October 01, 2012

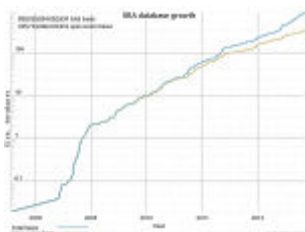
Show [PubChem](#) data your way using the new PubChem Widgets. Display concise tables of patents, bioactivities, and literature from PubChem on your web pages. See the [\(online help document\)](#) for details and examples.



## NCBI News, September 2012

### How has the SRA grown!

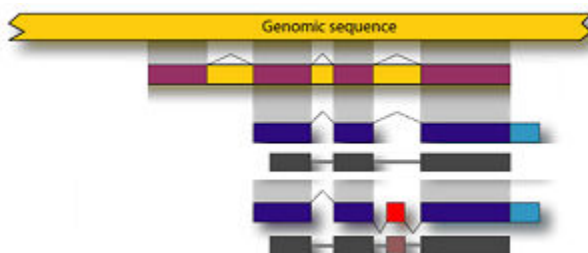
Monday, September 24, 2012



The Sequence Read Archive (SRA) now contains >859,000,000,000,000 bases with >335.7 open-access TeraBases available!

### RefSeq Release 55 is out!

Thursday, September 20, 2012



RefSeq Release 55 is available through Entrez, BLAST, and the [RefSeq FTP site](#). The current release includes 23,207,572 sequence records from 17,994 different organisms. The [Release Notes](#) provide more detailed information. Visit the [Reference Sequence Help Book](#) for more information on NCBI Reference Sequences.

### Now in PubChem: >6 million chemicals from SCRIADB with links to USPTO patents

Thursday, September 20, 2012

More than 6 million structures from SCRIADB are now available in (PubChem. Extracted from the complex work units of more than 300,000 USPTO patents between the years of 2001-2012, these PubChem Substance records link to both SCRIADB and the USPTO websites.

Click [here](#) to retrieve the SCRIADB records in PubChem Substance.

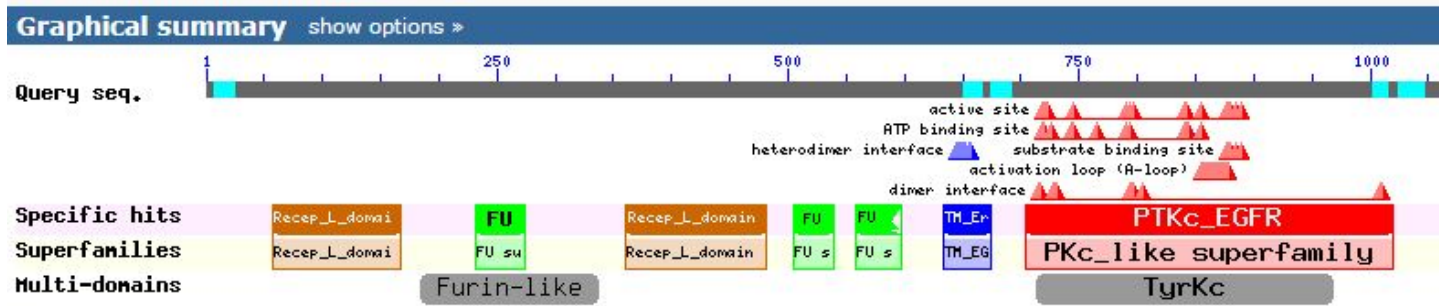
A [full text article](#) describing the SCRIADB project is available in PubMed Central.

### New CDD release contains 239 new or updated NCBI-curated domains

Wednesday, September 19, 2012

## Conserved domains on [gi|29725609|ref|NP\_005219]

epidermal growth factor receptor isoform a precursor [Homo sapiens]



Version 3.08 of the [Conserved Domain Database \(CDD\)](#) is now available. 46,507 conserved domain models have been indexed for searching with 239 as new or updated NCBI-curated domains. More detailed statistics are available from the [CDD News page](#).

CDD data has been incorporated in the Entrez and BLAST search services and CDD matrices and other information can be downloaded from the [CDD FTP site](#).

## PubChem databases and services are now HTTPS compatible

*Monday, September 17, 2012*

PubChem, NCBI's small molecule database and search system, is now Hypertext Transfer Protocol Secure (HTTPS) compatible, allowing encryption in interactions with the site.

## PubChem's PUG REST 1.0 is now available!

*Thursday, September 13, 2012*

PubChem Power User Gateway (PUG) REST interface 1.0 version is now available and replaces the April beta release!

This service allows access to PubChem data and services through simple HTTP requests, for use in scripts, javascript embedded in web pages, and 3rd party applications, without the overhead of XML and SOAP envelopes that are required for other versions of PUG. Unlike other PubChem services, PUG REST can deliver customized output or specific data elements without having to download the full record and can accept a variety of input types such as chemical name, InChIKey, SMILES in addition to numeric identifiers. The [PUG REST help document](#) provides more information on formulation HTTP requests and the available functions. A new [PUG REST tutorial](#) provides examples of PUG REST.

## The BLAST+ User Manual has been revised & updated

*Thursday, September 13, 2012*

The BLAST+ manual available on the [NCBI Bookshelf](#) has been extensively revised and updated. [Appendix C](#) now shows all available options for the different BLAST+ programs.

## Stand-alone BLAST has been updated

*Wednesday, September 12, 2012*

Stand-alone BLAST version 2.2.27+ is now available for download from the [BLAST Executables FTP site](#). This new version contains a number of important improvements and some bug fixes. Improvements include more

accurate composition-based statistics for translating searches (blastx), the ability to run remote searches with deltablast, reduced memory usage by blastn with short queries, and improved gap placement by blastn in megablast mode. See the [Blast News](#) for more detail.

## PubChem reaches milestones on its 8th BDay!

*Wednesday, September 12, 2012*



PubChem now has more than 100 million live records in PubChem Substance, and nearly 200 million bioactivity outcomes PubChem BioAssay. More than 200 data submitters have submitted data to the PubChem project since its launch on September 16, 2004.

## A new version of Genome Workbench is available

*Thursday, September 06, 2012*

A new version (2.6.0) of NCBI's [Genome Workbench](#) is now available. Genome Workbench is a standalone sequence viewer, annotation, and analysis platform. This version has many new features, improvements, and a few bug fixes that are described in the [Release Notes](#).

## NCBI is now using Genome Annotation Release numbers

*Tuesday, September 04, 2012*

The NCBI eukaryotic genome annotation pipeline is now using Annotation Release numbers to decrease confusion regarding the independent notions of a genome assembly and its annotation.

Annotation Release numbers:

- are integer values that increment each time the genome annotation is updated.
- have initial values starting at 100 or higher. are incremented independently for each organism.
- are used for the set of annotations calculated on one or more genome assemblies.

Please see the documentation describing [The NCBI eukaryotic Genome Annotation Process](#) page for more information.



## NCBI News, July 2012

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: July 12, 2012; Updated: July 12, 2012.

### Registration now open for NCBI Discovery Workshops September 4-5 at NLM

Registration is now open for the two-day Discovery Workshops to be offered on September 4 -5, on the NIH campus in Bethesda, Maryland. The course is free and is open to anyone interested in NCBI resources. The workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. For more information see the [Discovery Workshops page](#), which also includes a [registration link](#).

### 1000 Genomes Dataset Browser

The new [1000 Genomes Browser](#) shows variants, genotypes, and supporting sequence read alignments produced by the [1000 Genomes project](#). The genotype data are based on the Phase 1, March 2012 set and the variation (NCBI SNP) data are from SNP build 135. The 1000 Genomes browser is accessible from the new [NCBI Variation page](#) (Figure 1) that also provides links to other NCBI variation resources including [SNP](#), [dbVar](#), [dbGaP](#), the [Variation Reporter](#), [Clinical Remap](#), and the [Phenotype Genotype Integrator](#). The graphical portion of the genome browser is based on the NCBI graphical sequence viewer (GSV) and offers the familiar features and functions of the GSV. A table of genotypes organized by 1000 Genome populations is shown under the sequence viewer. A summary genotype frequency shows in the table for each population. The population sections can be expanded to show individual level genotypes.

The browser initially opens with an expanded view of human chromosome1 (Figure 1, *Top panel*). The search function on the right-hand side of the browser allows searches for RefSNP accession numbers, gene names, and chromosome positions. The search results show a list of matching items. Clicking on one of these items jumps the display to that position in the genome browser. The browser also allows scrolling either through the Scroll Region arrows on the table of genotypes or through the navigation controls in the graphical portion.

NCBI SNP records that are included in the 1000 Genomes data also link directly to the corresponding position in the browser. Figure 2 shows the region containing a SNP ([rs8176058](#)) in the Kell blood group antigen gene ([KEL](#)). For any region, the individual next-generation sequencing reads can be selected and loaded into the graphical browser through the expandable Subjects dialog box (Figure 2, *Bottom panel, left*). The Subjects dialog allows selecting individuals by population as well as filtering by the characteristics of the next-generation data. A portion of the aligned exome-sequencing reads for the heterozygous Toscan individual, [NA20507](#), is shown on the left-hand side of the bottom panel of Figure 2.

The data from the 1000 genomes project are representative of the increasing importance and presence of “big data” at the NCBI. Currently these data and associated metadata are stored in many different databases at the NCBI including the [Sequence Read Archive \(SRA\)](#), [SNP](#), [BioSample](#), and [BioProject](#). The 1000 Genomes Browser provides a simple and powerful single interface to complex and very large sets of data and metadata that comprise the 1000 Genomes project.

---

**Author Affiliations:** 1 NCBI; Email: [cooper@ncbi.nlm.nih.gov](mailto:cooper@ncbi.nlm.nih.gov). 2 NCBI; Email: [morrisrc@ncbi.nlm.nih.gov](mailto:morrisrc@ncbi.nlm.nih.gov).

✉ Corresponding author.

**Variation** Search NCBI  Search

**Variation**  
Access NCBI's variation resources

[www.ncbi.nlm.nih.gov/variation/](http://www.ncbi.nlm.nih.gov/variation/)

Getting Started	Variation Tools	Variation Databases
<a href="#">How to submit variants: dbSNP</a>	<a href="#">Variation Reporter</a>	<a href="#">dbSNP</a>
<a href="#">How to submit variants: dbVar</a>	<a href="#">Clinical Remap</a>	<a href="#">dbVar</a>
<a href="#">How to submit your clinical data</a>	<a href="#">Phenotype-Genotype Integrator</a>	<a href="#">dbGaP</a>
<a href="#">FAQ</a>	<a href="#">1000 Genomes Browser</a>	<a href="#">ClinVar</a>

Assembly	Genome Build	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh37.p5	37.3	Z	142655008	NT_007914.15	3250631	Rev	G	Fwd	<a href="#">view</a>	remap
GRCh37.p5 (Patches)	37.3	Z	NA	NW_003571040.1	1180668	NA	G	NA	<a href="#">view</a>	remap
reference	36.3	Z	142365130	NT_007914.14	3310213	Rev	G	Fwd	<a href="#">view</a>	blast
Celera	36.3	Z	137491868	NW_923651.1	956473	Rev	G	Fwd	<a href="#">view</a>	blast
HuRef	36.3	Z	136993297	NW_001839073.1	3293750	Rev	G	Fwd	<a href="#">view</a>	blast
CRA_TCAGchr7v2	36.3	Z	142056880	NT_079596.2	42092448	Rev	G	Fwd	<a href="#">view</a>	blast

**SNP: rs8176058**

**Homo sapiens: GRCh37.p5 Chr 1 (NC\_000001.10)**

Go  Enter a location, gene name or phenotype

Search Results:

Name	Type	Chr	Location
KEL	GENE	7	142,638,201 - 142,659,503
NM_000420.2	TRANSCRIPT	7	142,638,201 - 142,659,503
BC003135.1	TRANSCRIPT	7	142,638,201 - 142,659,503
KEL	STS	7	142,654,928 - 142,655,051
KEL	STS	7	142,654,928 - 142,655,051

Figure 1. Access to the 1000 Genomes browser. *Top panel:* The Variation Portal that serves as a gateway to a variety of NCBI variation resources including the 1000 Genomes Browser, red arrow. *Center panel:* NCBI SNP records for Reference SNPs included in the 1000 Genomes data link directly to the location in the 1000 Genomes Browser. *Bottom panel:* The 1000 Genomes Browser search interface. Useful search terms include gene names, SNP accessions, and chromosomal positions.



**Homo sapiens: GRCh37.p5 Chr 7 (NC\_000007.13): 142.65M - 142.66M**

Association Results

NCBI Genes

GRCh37 genome-wide recombination rate from Phase 2 HapMap estimated from phase 2 HapMap

1000 Genomes (all submissions)

dbSNP submissions not present in 1000 Genomes

**rs8176058**

SNP rs8176058; Alleles: A/C/G  
total range: NC\_000007.13 (142,638,201..142,659,503)  
total length: 21,303  
strand: minus  
MIN: 613

142,655 K **rs8176058** 142,655,010

recombination rate from Phase 2 HapMap estimated from phase 2 HapMap

in 1000 Genomes

rs8176058

rs61729031

rs200430183

**Subjects**

Platform:  ILLUMINA,  LS454,  SOLID

Aligner:  BFAST,  BWA,  MOSAIK,  SSAHA2

Alignment Type:  exome,  high\_coverage,  low\_coverage

Tracks in view

Sample	Bio Sample	Population	Platform	Aligner	Alignment Type
NA20507	SR5001675	TSI	SOLID	BFAST	exome (-)

Available Tracks

Sample	Bio Sample	Population	Platform	Aligner	Alignment Type
NA20504	SR5001672	TSI	SOLID	BFAST	exome (+)
NA20505	SR5001673	TSI	SOLID	BFAST	exome (+)
NA20506	SR5001674	TSI	SOLID	BFAST	exome (+)
NA20508	SR5001676	TSI	SOLID	BFAST	exome (+)
NA20509	SR5001677	TSI	SOLID	BFAST	exome (+)
NA20510	SR5001678	TSI	SOLID	BFAST	exome (+)
NA20512	SR5001680	TSI	SOLID	BFAST	exome (+)
NA20513	SR5001681	TSI	SOLID	BFAST	exome (+)

Figure 2. The 1000 Genomes Browser showing views for the SNP **rs8176058**, a polymorphism in the Kell blood group antigen protein, the product of the *KEL* gene. *Top panel*: Initial overview on genotypes for populations showing overall frequencies. The major allele is in bold font. Sections expand to show individual level genotypes. Clicking the red arrow opens a dialog box (*lower left panel*) that allows selecting and loading next generation sequence read alignments from individuals into the browser. The lower right panel shows some of the exome sequencing reads from a heterozygous individual (**NA20507**) from the Toscan population (TSI) aligned at the position of **rs8176058**.

## PubMed News

### PubMed Send to Citation Manager and Favorites

PubMed now offers the ability to download citations for use in citation manager software such as Endnote, RefWorks or other bibliography program through the "Send to" menu. The [PubMed Technical Bulletin](#) has more details on using this feature.

Abstracts in PubMed also now include a "Save items" section that will provide easy way to add items of interest to a My NCBI collection. If you are signed in to My NCBI clicking the "Favorite" button adds the citation to a new My NCBI collection, Favorites. You can add multiple items to My Collections, including Favorites, in My NCBI through the "Send to" menu in the upper right of search result displays. For more information on My NCBI and [My Collections](#) please visit [My NCBI Help](#) on the NCBI Bookshelf.

The screenshot illustrates the process of saving PubMed search results to a My NCBI collection. It is divided into three main sections:

- Search Results:** Shows two search results. The first result is "Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Nucleic Acids Res. 2012 Jan;40(Database issue):D48-53. Epub 2011 Dec 5. PMID: 22144687 [PubMed - In process] Free PMC Article". The second result is "The 2012 Nucleic Acids Research Database Issue and the Database Collection. Galperin MY, Fernández-Suárez XM. Nucleic Acids Res. 2012 Jan;40(Database issue):D1-8. Epub 2011 Dec 5. PMID: 22144685 [PubMed - In process] Free PMC Article".
- Send to Menu:** A dropdown menu is open, showing options: File, Collections (selected), Order, and Citation manager. Below these options is a button labeled "Add 11 items." and another button labeled "Add to Collections". A red arrow points to the "Add to Collections" button.
- My NCBI — Collections Dialog:** A dialog box is open, showing "11 items from PubMed". It asks "What would you like to do?" with options: "Create new collection" and "Append to an existing collection" (selected). Below this is a "Choose a collection:" dropdown menu with "Favorites" selected. A "Save" button is at the bottom, with a red arrow pointing to it. Below the "Save" button is the text "Or cancel and return to your selections."

The second part of the screenshot shows a specific search result for "The Sequence Read Archive: explosive growth of sequencing data." with the following details:

- Title:** The Sequence Read Archive: explosive growth of sequencing data.
- Author:** Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration.
- Abstract:** New generation sequencing platforms are producing data with significantly higher throughput and lower cost. A portion of this capacity is devoted to individual and community scientific projects. As these projects reach publication, raw sequencing datasets are submitted into the primary next-generation sequence data archive, the Sequence Read Archive (SRA). Archiving experimental data is the key to the

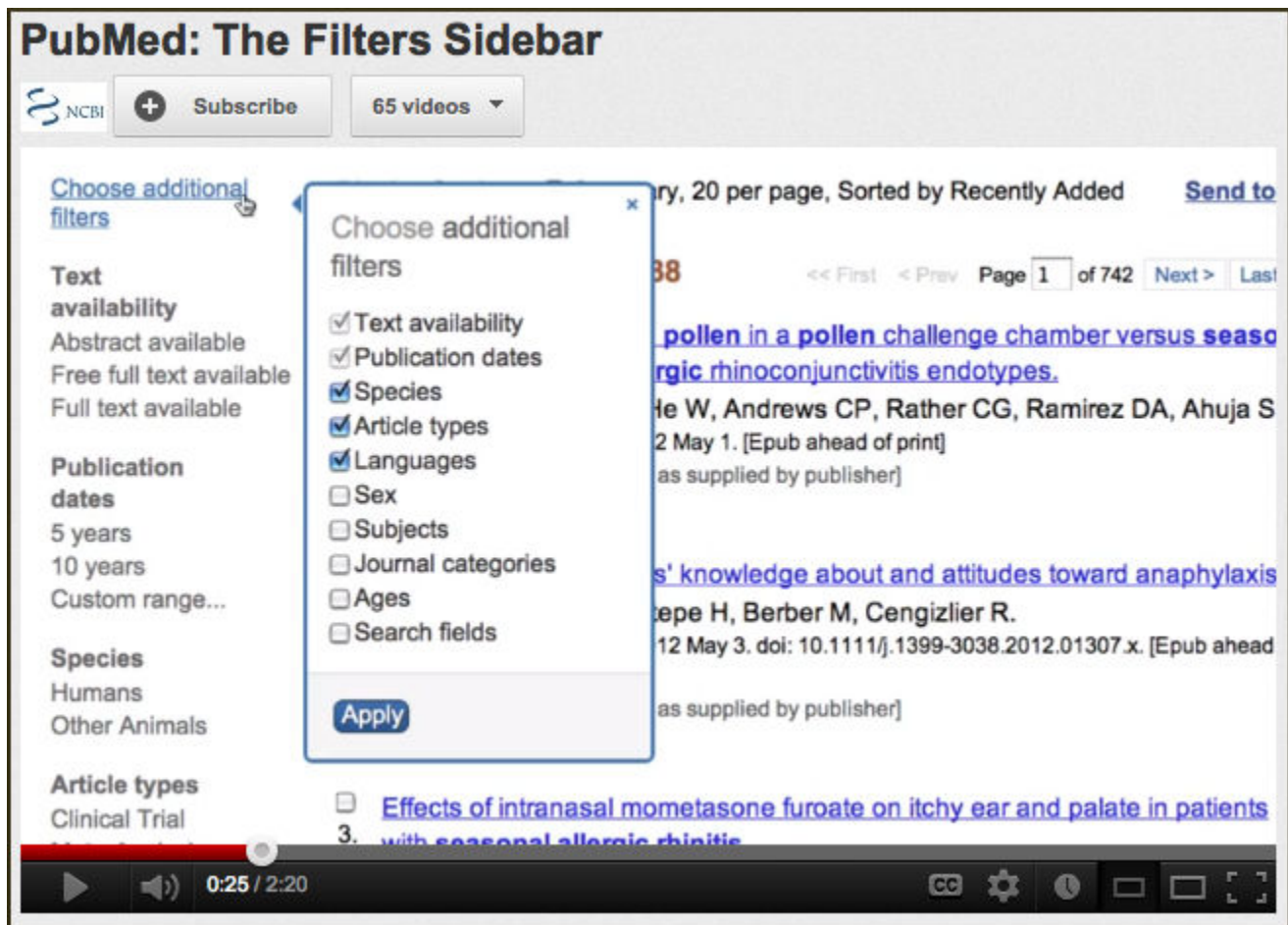
Below the abstract, there is a "Save items" section with a dropdown menu showing "Favorite" (selected) and "Favorites". A red arrow points to the "Favorite" option. Below the dropdown are buttons for "Create collection...", "Manage collections...", and "The sequence read archive. [Nucleic Acids Res. 2011]".

The third part of the screenshot shows the "My NCBI — Collections - Favorites" page. It displays the following information:

- Collection Name:** My NCBI — Collections - Favorites
- Settings:** This collection is private (make it public) | Edit settings for this collection | Save collection to a text file | Save collection to a csv file
- Display Settings:** Sort by Author
- Select:** All, None 11 items selected | Delete | View
- Item 1:** BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, Kimelman M, Pruitt KD, Resenchuk S, Tatusova T, Yaschenko E, Ostell J. Nucleic Acids Res. 2012 Jan;40(Database issue):D57-63. Epub 2011 Dec 1. PubMed [citation] PMID: 22139929 PMCID: PMC3245069

## PubMed Filter Sidebar

PubMed now has a Filter Sidebar in the PubMed results. The useful features of the popular Limits page have been made more visible by placing them in this Filter Sidebar and should make it easier to refine PubMed search results. For more information, please see the [NLM Technical Bulletin](#). A new [video](#) on NCBI's [YouTube Channel](#) also demonstrates this useful new addition to PubMed searching.



## BLAST News

### New Microbial Genomes BLAST Service

A new [microbial BLAST service](#) is now live. The service is easier to use and has the familiar format and features of the standard BLAST services at NCBI including the ability to select of taxonomic categories using an auto-complete "Organism" box and to include or exclude multiple taxonomic categories. Other standard features of the BLAST pages such as "Edit and Resubmit" and the ability to optimize for a specific search are also included. For nucleotide databases the search sets have also been divided into Complete and Draft genomes.

### Article on Primer-BLAST Published

An article describing Primer-BLAST, NCBI's PCR primer designing service, is now available in *BMC Bioinformatics*.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. 2012 Jun 18;13(1):134. PubMed PMID: 22708584.

## BLAST Programming Interface: End of OLD BLAST=true option

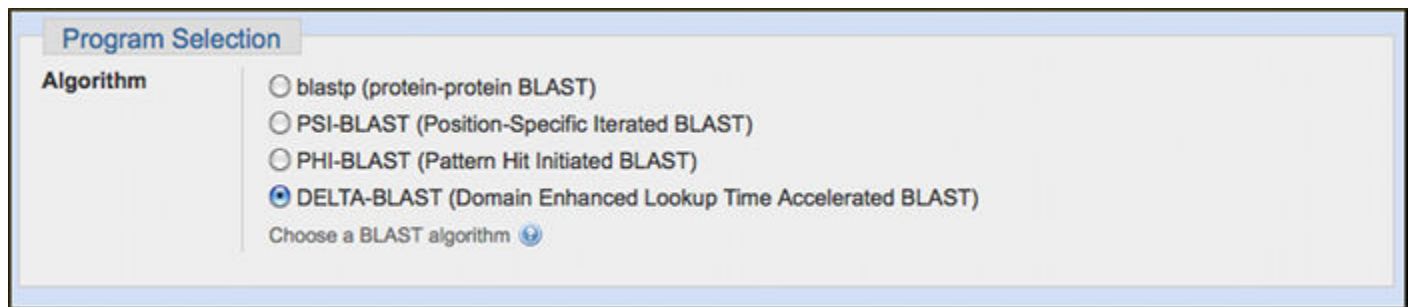
Beginning Sept. 10, 2012, the BLAST service will ignore the OLD\_BLAST parameter in posted URLs. We are removing this old and little-used option to prepare for upcoming enhancements to the BLAST service later this year. Setting OLD\_BLAST=true produces an older version of the BLAST HTML results that a few people have used for automated processing (parsing) of results. NCBI BLAST supports a number of different and more stable parsable formats. These include XML, tabular reports and ASN.1. For more details, please see [BLAST Developer Information](#) and links on that page.

## DELTA-BLAST Service and Article

As described in the [April 2012 NCBI News Domain Enhanced Lookup Time Accelerated BLAST \(DELTA-BLAST\)](#) included in the BLAST 2.2.26+ release, offers a more sensitive protein-protein BLAST search by performing a position specific score matrix search using results from an initial conserved domain search. A paper in Biology Direct describes the DELTA-BLAST algorithm and discusses its enhanced sensitivity compared to other methods.


Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. Domain enhanced lookup time accelerated BLAST. *Biol Direct*. 2012 Apr 17;7(1):12. PubMed PMID: [22510480](#).

The protein BLAST web service offers [DELTA-BLAST](#) as a Protein BLAST program selection option on the Basic Protein BLAST service.



The image shows a screenshot of the 'Program Selection' section of the NCBI BLAST web interface. It features a list of algorithms with radio buttons for selection. The 'DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)' option is selected, indicated by a blue dot in the radio button. Below the list is a text prompt 'Choose a BLAST algorithm' with a help icon.

Algorithm	Selection
<input type="radio"/> blastp (protein-protein BLAST)	Not selected
<input type="radio"/> PSI-BLAST (Position-Specific Iterated BLAST)	Not selected
<input type="radio"/> PHI-BLAST (Pattern Hit Initiated BLAST)	Not selected
<input checked="" type="radio"/> DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)	Selected

Choose a BLAST algorithm 

DELTA-BLAST improves the sensitivity and selectivity of most protein searches that have strong conserved domain results. Figure 3 shows the differing alignments, scores, and expect values for the same match with standard protein blast, blastp (RID: [0E4F72U7013](#), Figure 3, *Top panel*) and DELTA-BLAST (RID: [0E5RMA6X016](#), Figure 3, *Bottom panel*). Both of these searches use the human hemoglobin subunit beta protein (NP\_000509) as a query against bacterial sequences from the NCBI RefSeq protein database. Standard protein BLAST finds a globin protein (YP\_485375) from the purple non-sulfur bacterium *Rhodopseudomonas palustris* HaA2 with an expect value of  $1 \times 10^{-4}$ . In these results the same expect value is found for some non-globin sequences including an aspartate kinase, an amino peptidase, and a succinate-semialdehyde dehydrogenase. In addition the blastp alignment does not match the conserved histidine (position 93 in NP\_000509) that is part of the heme wbinding site in the human hemoglobin domain and structure (Figure 3, *Middle panel*). The blastp alignment also inserts a gap in the conserved alpha helix in this region. In contrast, DELTA-BLAST finds the same protein but with a much better expect value of  $3 \times 10^{-27}$ , thus easily segregating the hit from the non-globin proteins found in the blastp search. Moreover, the alignment now corresponds to the globin conserved domain, matching the conserved histidine and preserving the secondary structure block.

DELTA-BLAST is an important new addition that extends the capabilities of the NCBI BLAST service and produces more accurate alignments and more discriminating statistics by using conserved domain information in the initial search.

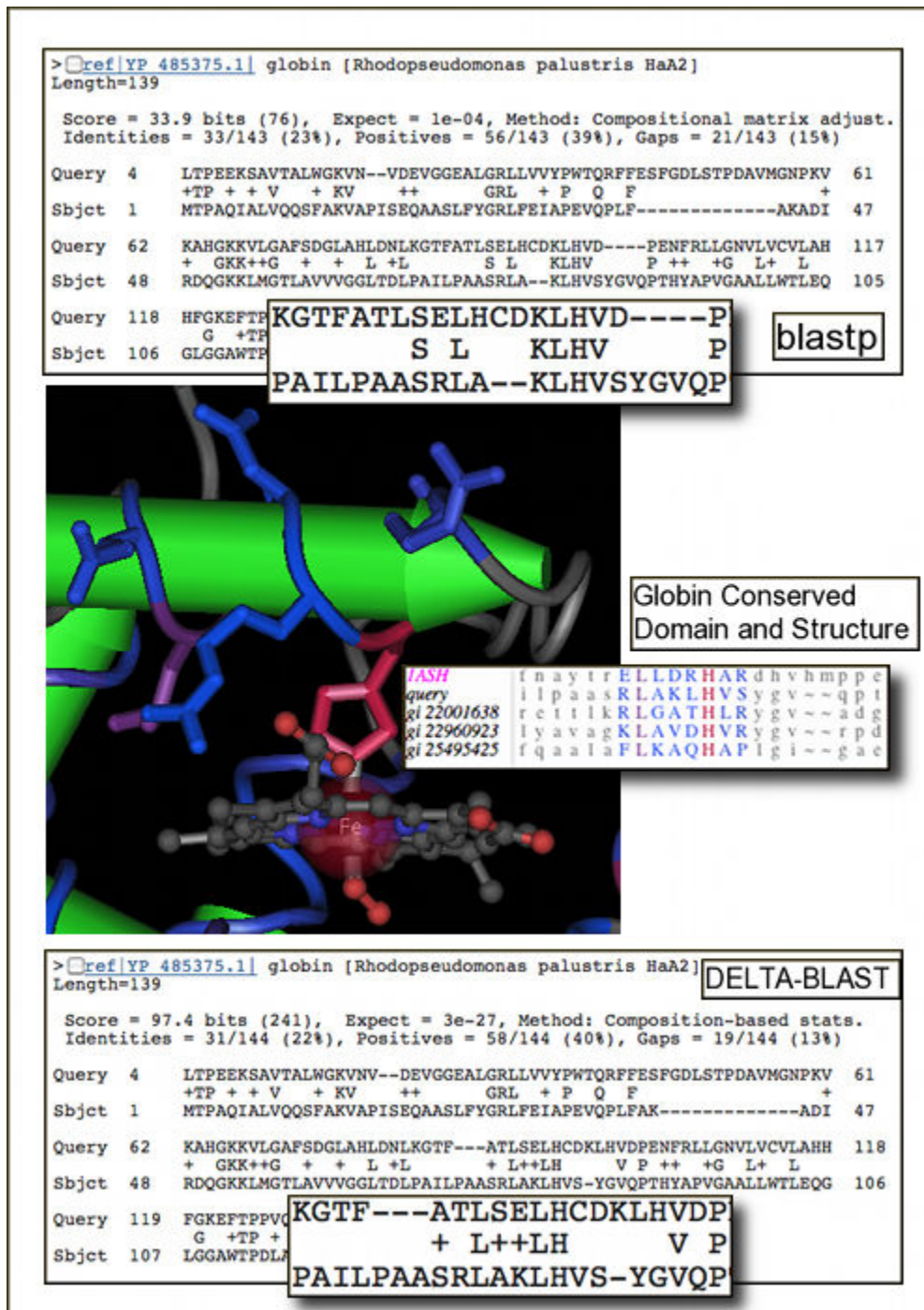


Figure 3. Comparison of standard blastp and DELTA-BLAST statistics and alignments. The match found between the human hemoglobin beta (NP\_000509) a globin (YP\_485375) from the purple non-sulfur bacterium *Rhodopseudomonas palustris HaA2* is shown. *Top panel*: Protein blast results (RID: 0E4F72U7013). In the blastp alignment the conserved histidine at position 92 in the human protein is not aligned with a corresponding histidine in the bacterial sequence, and gaps are inserted into the conserved alpha helix in this region. *Middle panel*: Partial conserved domain alignment and structure for globin (cd01040) showing the conserved histidine (red H) residue and alpha helix (colored block). *Bottom panel*: DELTA-BLAST alignment (RID: 0E5RMA6X016). The DELTA-BLAST result gives a much more significant expect value and more accurate alignment for the globin domain accurately aligning the conserved histidine and preserving the alpha helix.

## New HomoloGene Build: Rhesus macaque now included

HomoloGene, the NCBI resource that identifies and clusters homologous genes, transcripts and proteins for selected eukaryotes, has a new build (Build 66). With this build, HomoloGene for the first time includes genes and sequences for the Rhesus monkey (*Macaca mulatta*). The new build also includes updated annotations for human, chimpanzee, dog, cow, mouse, rat, chicken, zebrafish, fruitfly, yeast, arabidopsis, and rice. HomoloGene data are available from the NCBI [FTP](#) site.

1: HomoloGene:41. Gene conserved in Euteleostomi

Genes	Proteins
<i>Genes identified as putative homologs of one another during the construction of HomoloGene.</i>	<i>Proteins used in sequence comparisons and their conserved domain architectures.</i>
<input type="checkbox"/> BRCA2, <i>H.sapiens</i> breast cancer 2, early onset	<input type="checkbox"/> NP_000050.2 3418 aa
<input type="checkbox"/> BRCA2, <i>P.troglodytes</i> breast cancer 2, early onset	<input type="checkbox"/> XP_509619.2 3418 aa
<input type="checkbox"/> BRCA2, <i>M.mulatta</i> breast cancer 2, early onset	<input type="checkbox"/> XP_001118184.2 3364 aa
<input type="checkbox"/> BRCA2, <i>C.lupus</i> breast cancer 2, early onset	<input type="checkbox"/> NP_001006654.2 3446 aa
<input type="checkbox"/> BRCA2, <i>B.taurus</i> breast cancer 2, early onset	<input type="checkbox"/> XP_002691853.1 3427 aa
<input type="checkbox"/> Brca2, <i>M.musculus</i> breast cancer 2	<input type="checkbox"/> NP_033895.2 3329 aa
<input type="checkbox"/> Brca2, <i>R.norvegicus</i> breast cancer 2	<input type="checkbox"/> NP_113730.1 3343 aa
<input type="checkbox"/> BRCA2, <i>G.gallus</i> breast cancer 2, early onset	<input type="checkbox"/> NP_989607.2 3397 aa
<input type="checkbox"/> brca2, <i>D.rerio</i> breast cancer 2, early onset	<input type="checkbox"/> NP_001103864.2 2874 aa

## Microbial Genomes Update

Ninety-two finished microbial (archaeal and bacterial) complete genome sequences were released for 90 microbial strains (7 archaea and 83 bacteria) from April 2012 through June 2012. These include three complete plasmid sequences and 89 chromosome sequences. The original sequence data files submitted to the International Sequence Database Collaboration (INSDC) are available in the [Bacteria](#) directory in the genomes area of the GenBank FTP site. RefSeq versions were released for a selected set of 391 of the complete INSDC microbial genome sequences for 387 microbial strains during the same period. These are available from the [/genomes/Bacteria](#) directory on the FTP site.

In addition, data from 754 microbial whole genome-shotgun (WGS) sequencing projects were added to the INSDC during this period. The original submitted files are available in the [Bacteria\\_DRAFT](#) directory in the GenBank genomes area. RefSeq provisional versions of 84 WGS microbial projects were released in the [/genomes/Bacteria\\_DRAFT](#) area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

## GenBank News

GenBank release 190 is available through the NCBI web and [FTP](#) sites. The current release incorporates data available as of June 15, 2012 and, with the whole-genome shotgun portion, contains 428,920,607,871 bases from 236,206,989 sequence records. [Release notes](#) describe the current state of data and upcoming changes. The [GenBank page](#) provides more information on the database content and scope as well as submission information.

## RefSeq News

RefSeq Release 54 is available through Entrez, BLAST, and from the [RefSeq FTP](#) area. The current release includes 21.9 million Reference Sequence records from 17,605 different species or strains. The RefSeq [release notes](#) provide more detailed information.

## GRC Plans New Human Genome Build and Requests Input

The [Genome Reference Consortium \(GRC\)](#), which produces assemblies that are the basis for NCBI Reference assemblies for human, mouse, and zebrafish, is planning a new build of the human genome (GRCh38) for summer of 2013. Anyone who has questions, concerns, or input, may submit these on the [GRC contact form](#). The [GRC blog](#) provides insights into the complexities and the process of updating, correcting, and representing the human genome.

## NCBI Now Offers IPv6 Access

The NCBI website now supports the new six-byte Internet Protocol addresses (IPv6) for HTTP access as well as data downloads using FTP, Aspera, and RSync. The [World IPv6 Launch](#) site has additional information on the transition to IPv6.

## Keeping Up with NCBI

Seventeen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List [summary page](#). Subscribe to the [NCBI Announce list](#) to receive updates on the NCBI News.

Twenty-six [RSS feeds](#) are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's [Facebook](#) page and [Twitter feed](#) also provide updates on NCBI resources.

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.





## NCBI News, April 2012

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: March 30, 2012; Updated: March 30, 2012.

### NCBI Discovery Workshops May 15-16 at NLM: Seats still available

NCBI will present a two-day workshop May 15 and 16, 2012, on the NIH campus in Bethesda, Maryland. The course is free and is open to anyone interested in NCBI resources. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. These workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The [Discovery Workshops page](#) has more details and a link to register for the course.

### Assembly: a Companion to the Genome Database

The new [NCBI Assembly database](#) provides statistics, update history and links to sequences for eukaryotic genome assemblies including assemblies for previous genome builds. Assemblies of interest can be found either by text searches on the main assembly page or through the [assembly browser](#) that provides easy access by organism. Assemblies are also linked through the Genome database main page or from a Genome record for a eukaryotic species as shown in Figure 1. Each assembly is assigned an accession and a version that unambiguously identifies the sequences in a particular version of an assembly. The database contains the placement of each scaffold in the assembly along with the name and sequence accession and version for each chromosome and scaffold. The database also organizes and provides assembly descriptive items such as assembly names and synonyms, as well as statistical reports including scaffold counts and weighted scaffold and contig length medians (N50). Figure 2 shows the page for the latest mouse genome assembly (GRCm38). This page provides access to the primary assembly and alternate loci sequences and statistics. The [Assembly Help](#) documentation provides more detailed information on using the Assembly Database.

### New Videos on NCBI's YouTube Channel

Eleven new tutorial videos have been added to the NCBI YouTube channels in the past few months. To make topics of interest easier to find, the Tutorials playlist now provides special playlists for certain resources. The channel now features separate tutorial playlists for Genome Workbench ([7 videos](#)), Sequence Viewer ([4 videos](#)), My NCBI ([4 videos](#)), Genetic Testing Registry ([2 videos](#)) and General ([22 videos](#)).

Five of the recent videos are in the General playlist and include tutorials on using the new Advanced Search Builder in PubMed ([video, advanced search page](#)), an overview of RefSeqGene reference standard records for selected human genes ([video, resource page](#)), an introduction to the E-utilities, the programming interface to the Entrez system ([video, E-Utilities Help Manual](#)), a demonstration of the highlight sequence features tool in sequence databases ([video, NCBI News](#)), and a video on how to use Genome Remapping Tool ([video, tool page](#)) that can map coordinates of genes and other markers from one genome build to another.

Three of the new videos are about My NCBI, the service that allows registered users to customize their experience and to save and share results, searches, and preferences through their accounts. New titles in the My NCBI playlist are [My Bibliography](#), [Save Searches and Set E-mail Alerts](#), and [Save Search Results in Collections](#).

Assembly





[Browse by organism](#)

## Assembly

Genome assembly organization and additional information.

### Using Assembly

- [Assembly Help](#)
- [Browse by Organism](#)
- [NCBI Ass](#)
- [Assembly](#)

### Submitting an Assembly

- [Submission Information](#)
- [Submission FAQ](#)

### Related Resources

- [Genome](#)
- [Genome Reference Consortium](#)

## Mus musculus (house mouse)

The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced

Lineage: Eukaryota[1408]; Metazoa[572]; Chordata[237]; Craniata[232]; Vertebrata[231]; Euteleostomi[226]; Mammalia[111]; Eutheria[107]; Euarchontoglires[57]; Glires[26]; Rodentia[24]; Sciurognathi[20]; Muroidea[14]; Muridae[4]; Murinae[4]; Mus[2]; Mus[2]; Mus musculus[1]

The mouse is one of the major organisms for modeling human disease and comparative genome analysis. There are over 450 inbred strains of mice, providing a wealth of different genotypes and phenotypes for genetic and other studies. In addition, thousands of spontaneous, radiation- or chemically-induced, and transgenic mutants provide potential models [More...](#)

**Organism Overview** [See also: Genome list](#) [Organelle List](#)

Chromosomes

Click on chromosome name to open MapViewer

Assembly and Annotation

Default assembly

5 other assemblies are available

Assembly Name	MGSCv37
Last sequence update	23-Feb-2012

### Assembly information by organism

Results: 1 to 6 of 6 << First < Prev Page 1 of 1 Next > Last >>

Organism	Name	Submitter	Genome representation	Assembly level	Version status	Default status
Mus musculus	GRCm38 <small>UCSC Name: mm10</small>	Mouse Genome Sequencing Consortium	complete	Chromosome	latest	Not default
Mus musculus	ASM216v1	Celera Genomics	complete	Chromosome	latest	Not default
Mus musculus	Mm_Celera	Celera Genomics	complete	Chromosome	latest	Not default
Mus musculus	MmusALLPATHS1	Broad Institute	complete	Scaffold	latest	Not default
Mus musculus	MmusSOAP1	Broad Institute	complete	Scaffold	latest	Not default
Mus musculus	ASM18119v1	Genome Reference Consortium	complete	Contig	latest	Not default

Figure 1. Accessing the Assembly database. *Top panel.* The Assembly main page with the search box and access to the Assembly Browser (Browse by Organism, red circle). *Middle panel* the mouse genome overview with showing information for the Default assembly (MGSCv37) with a link to all the assemblies. *Bottom panel.* The Assembly Browser showing the six latest assemblies for the mouse.

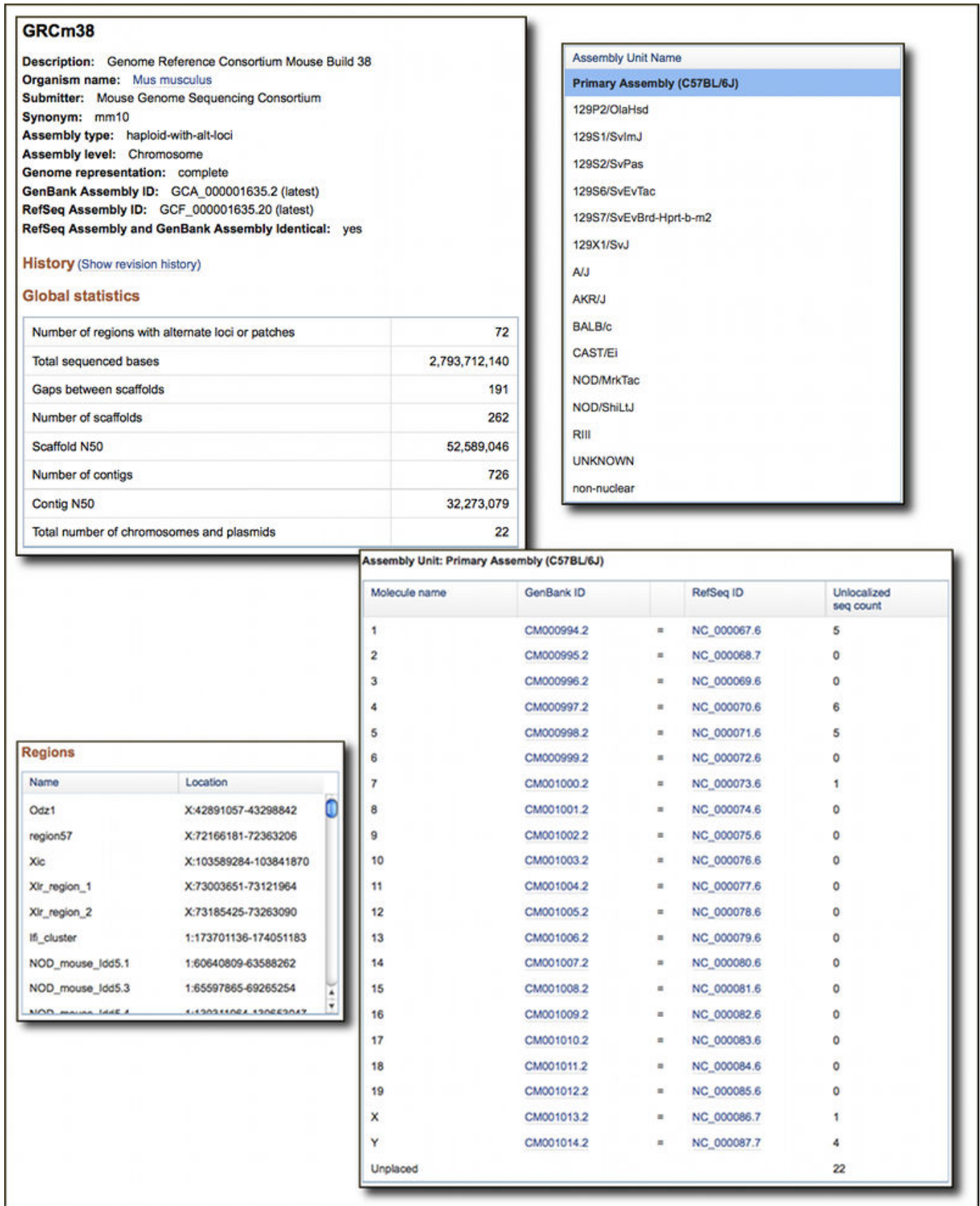


Figure 2. Aspects of the mouse GRCm38 assembly. *Top left panel.* General Assembly Definition showing names, synonyms, and assembly identifiers. *Top right panel.* Assembly units including the primary assembly for C57BL/6J and alternate loci for other mouse strains. These alternate loci are also available by region (*Lower left panel*). *Lower right panel.* Listing of the molecules and corresponding nucleotide accessions making up the selected assembly unit, the C57BL/6J primary assembly in this case. Detailed statistics are available for each molecule in the assembly in a separate tab (not shown).

One new [video](#) demonstrating how load a genome into Genome Workbench was recently added to the playlist for Genome Workbench, NCBI's standalone sequence analysis and annotation platform.

Most recently a new playlist was created for two tutorials ([GTR: Homepage and Basic Search Functions](#) and [GTR: Locate a Test in Under Three Minutes](#)) featuring the newly launched Genetic Testing Registry, a repository of information about available genetic tests. Additional information about the GTR is provided in the following section of this newsletter.

The image displays two overlapping screenshots of YouTube video thumbnails. The top thumbnail is titled "GTR: Homepage and Basic Search Functions" and shows a search bar with "ehlers" entered, yielding results for "Ehlers-Danlos Syndrome Type IV" and "Ehlers-Danlos Syndrome, Classic Type". A watermark "www.youtube.com/ncbinlm" is overlaid on this thumbnail. The bottom thumbnail is titled "GTR: Locate a Test in Under Three Minutes" and shows a search for "CN073359[DISCU]" with results for "Multiple endocrine neoplasia, type 2".

## The Genetic Testing Registry: Finding Genetic Tests and Related Information

The NCBI recently released [the Genetic Testing Registry \(GTR\)](#). This new resource is a voluntary registry of genetic tests and laboratories with detailed information about the tests and their providers. The initial scope of GTR includes single gene tests for Mendelian disorders, as well as arrays, panels and pharmacogenetic tests. The registry includes detailed information about the purpose of the test, methodology, analytical and clinical validity,

and information on clinical usefulness. GTR provides access to information from the [GeneReviews](#) book on the NCBI Bookshelf – peer reviewed descriptions of genetic diseases and information on genetics tests and NCBI molecular databases such as [Gene](#). GTR is a central hub for information about genetic conditions and also provides context-specific links to a variety of resources, including practice guidelines, published literature, and genetic information. As mentioned in the previous section of this newsletter, [two new videos on the NCBI YouTube channel](#) provide quick introductions to the GTR. The original [NIH press release](#) has more information about the GTR.

**GTR: GENETIC TESTING REGISTRY**

All GTR Tests Conditions/Phenotypes Genes Labs GeneReviews

SCID Search All GTR

Adenosine Deaminase-Deficient Severe C...  
 RS-SCID: Severe combined immunodefici...  
 SCID due to absent class II HLA antigens  
 SCIDX1: X-linked severe combined immun...

ditions/phenotypes, genes, and labs.

**IMPORTANT NOTE:** NIH does not independently verify information submitted to the GTR; it relies on submitters to provide information that is accurate and not misleading. NIH makes no endorsements of tests or laboratories listed in the GTR. GTR is not a substitute for medical advice. **Patients and consumers** with specific questions about a genetic test should contact a health care provider or a genetics professional.

## BLAST News

### BLAST 2.26+ Release

The latest version of the C++ build of BLAST+ (2.2.26) is now available from the [BLAST FTP area](#) and is running on the NCBI [BLAST Web service](#). This new BLAST+ release contains a number of important changes and improvements including the three listed below.

**Domain Enhanced Lookup Time Accelerated BLAST (DELTA-BLAST)** is a new BLAST algorithm that can be more sensitive than standard protein-protein BLAST. DELTA-BLAST identifies conserved domains in the query sequence using Reverse PSI BLAST and then uses this information to construct a Position Specific Score Matrix (PSSM) then performs a PSSM search against the BLAST protein database. DELTA-BLAST can be invoked on the Protein-protein BLAST Web Service by selecting the DELTA-BLAST radio button in the “Program Selection” area of the submission form. The standalone BLAST package has DELTA-BLAST as a separate program (deltablast). Running DELTA-BLAST locally requires a special version of CDD database (cdd\_delta) available from the [BLAST db directory](#) on the FTP site.

A new **Finite Size Correction** has been added to the to the blastp algorithm to improve the accuracy of BLAST statistics (Expect values). The new finite size correction especially improves statistics for matches for short query or short database sequences.

Standalone BLAST now contains the program **makeprofiledb**, a C++ coded replacement for the NCBI C toolkit program `formatrpsdb`. `Makeprofiledb` can generate search sets for RPS-BLAST, including the specialized data needed by DELTA-BLAST.

## Final version of C-toolkit BLAST Package

Version 2.2.26 is the final version of NCBI C language toolkit BLAST. The source code for these applications will no longer be developed, but will continue to be available. Users of these legacy programs should migrate to the BLAST+ applications that are being actively developed. The [BLAST Command Line Applications User Manual](#) provides help on transitioning to the BLAST+ applications.

## Netblast (blastcl3) Service Discontinued: Replaced by remote Option in BLAST+

The Netblast client (`blastcl3`) that has provided batch search access to the NCBI Web BLAST service will be discontinued in the near future. The BLAST+ applications replace and improve upon the functions provided by `blastcl3`. `Blastcl3` users should switch to BLAST+ as soon as possible. Locally installed BLAST+ applications can perform remote searches using the NCBI Web service when the 'remote' option is included on the command line. The BLAST+ remote service has a number of advantages over the `blastcl3` application. `Blastcl3` requires a persistent connection during the entire search, can only submit one query at a time, and is unable to return the BLAST Request ID (RID) used in the search. The BLAST+ remote service can submit multiple queries (from FASTA input) at once, poll for the results using the BLAST RID, and also print the RID in the BLAST report. Using the BLAST RID, it is possible to reformat the search locally with the `blast_formatter` application, reformat the search at the NCBI web site, or use analysis tools such as the BLAST treeview or the taxonomy report.

## Changes in the BLAST Database List on the NCBI Web Services

A new **microbial 16S ribosomal** RNA sequence database is now available on nucleotide-nucleotide BLAST search page. This database contains Archaeal and Bacterial 16S sequences from the [Archaeal 16S Ribosomal RNA](#) and [Bacterial 16S Ribosomal RNA Targeted Loci](#) Projects. This database should be helpful in classifying unknown microbial 16S sequences from a wide range of sources.

Sequences from **environmental samples** formerly available in the `env_nr` and the `env_nt` databases are now available in the Metagenomic proteins database and, for nucleotide sequences, through the Whole Genome Shotgun Contigs (WGS) database by selecting "metagenomes (taxid: 408169)" as an Organism limit.

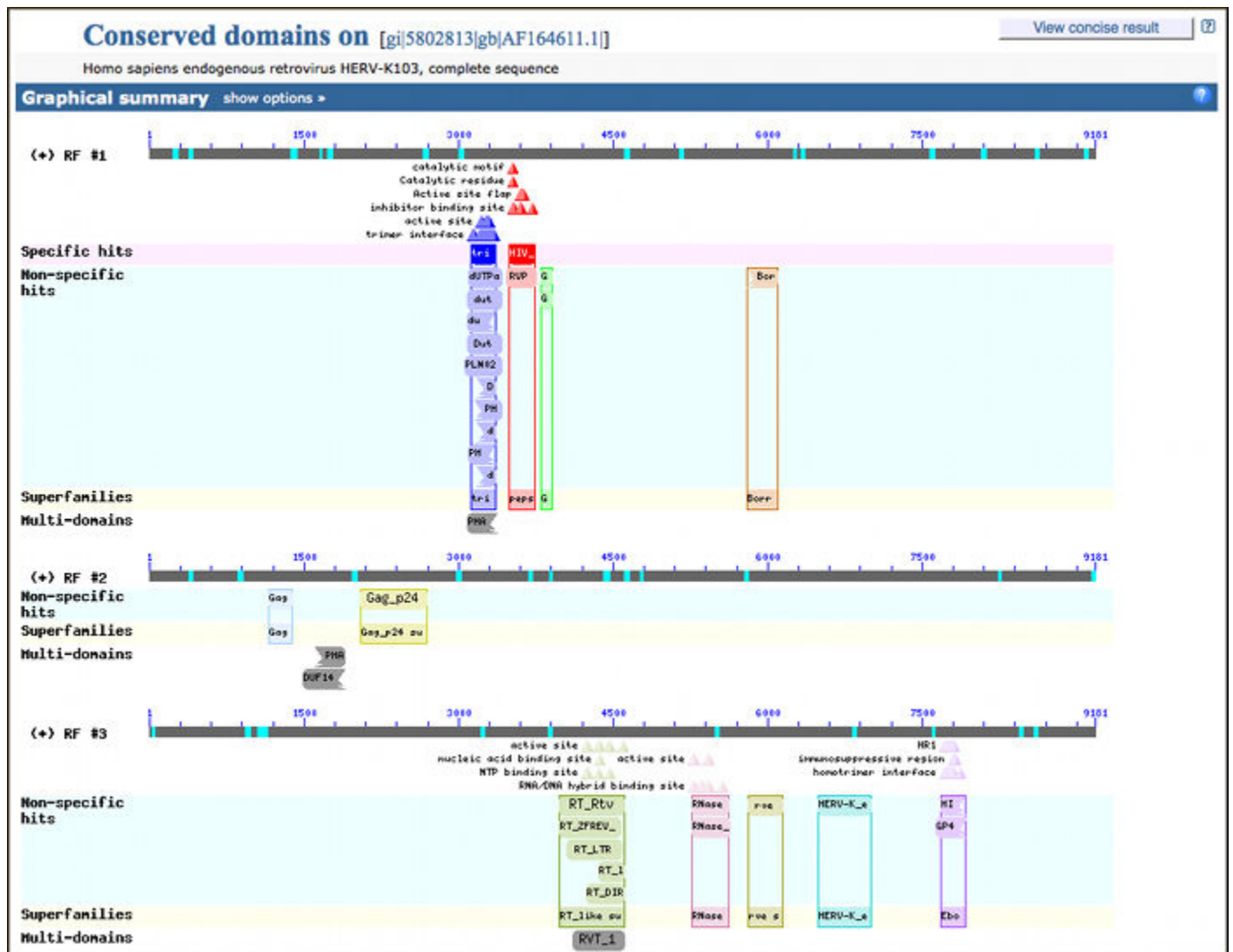
The following image shows the selections needed on the BLAST submission form to search these three new or modified databases.

The image displays three sequential screenshots of the NCBI search interface, illustrating the process of selecting a search set. Each screenshot is titled "Choose Search Set".

- Top Screenshot:** Shows the "Database" section with three radio buttons: "Human genomic + transcript", "Mouse genomic + transcript", and "Others (nr etc.):". The "Others (nr etc.):" option is selected. Below it, a dropdown menu is open, showing "16S ribosomal RNA sequences (Bacteria and Archaea)".
- Middle Screenshot:** Shows the "Database" section with a dropdown menu open, displaying "Metagenomic proteins(env\_nr)".
- Bottom Screenshot:** Shows the "Database" section with the "Others (nr etc.):" option selected. Below it, a dropdown menu is open, showing "Whole-genome shotgun contigs (wgs)". Below the database selection, there is an "Organism" section with a text input field containing "metagenomes (taxid:408169)", an "Exclude" checkbox, and a "+" button. A note below reads: "Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown."

## CDD Results Now Shown for Translated BLAST (blastx) Searches

Conserved Domain Search results are now provided for all translated (blastx) searches with query sequences shorter than 10,000 bases. Conserved domain searches are performed with all six reading frames of the query sequence and results are reported for each frame that has matches. This is very useful for helping to characterize coding regions on genomic regions as shown immediately below from the results for a blastx search with a human endogenous retrovirus ([AF164611](#)).



## Remap and Variation Reporter: Two New Services for Mapping Locations onto Genome Builds

The Genome Remapping Service (Remap) and the Variation Reporter are related tools that find locations on current and past genome builds.

The Remap tool translates or projects the coordinates of genes, variants (SNPs), and other sequence-based markers from one genome assembly (build) to another for human, mouse, rat, zebrafish and sea urchin (*Strongylocentrotus purpuratus*). It also includes a Clinical Remap version that performs coordinate remapping between genome assemblies and the reference standard RefSeqGene records. Figure 3 and Figure 4 show the submission and results for the Remap service. Locations to be projected can be in a variety of common genome annotation formats such as UCSC Browser Extensible Data (BED) format, Gene Transfer Format (GTF), Generic Feature Format (GFF and GFF3), Human Genome Variation Society (HGVS) nomenclature, and Genome Variation Format (GVF) among others. When projection of features is successful, the service reports the new locations with the submitted annotations in the selected format for downloading and also provides output in a format suitable for loading into Genome Workbench, the NCBI's standalone sequence analysis and annotation platform. A programming interface (API) is also available for the Remap service. A demonstration PERL script (remap\_api.pl) that accesses the service is available from the NCBI FTP site.



The Variation Reporter, shown in Figure 5, takes a set of locations in a human genome assembly and identifies known human variations (NCBI Reference SNPs) at those positions. This service is particularly helpful for identifying experimentally or clinically determined variants. Like the Remap service, the Variation Reporter accepts a variety of genome annotation formats – HGVS, GVF and BED. The results provide the location of the variants in the selected build and important information about any identified known variants including the dbSNP ID, the known allele, and, if available, clinical information, minor allele frequency, links to literature, and functional consequences. The results also provide the genomic context by displaying the mapped locations in the graphic sequence viewer (Figure 5, bottom panel). The Remap Service and the Variation Reporter are useful for interconverting annotations between genome builds and mapping and identifying experimentally determined variants.

## NCBI Aspera Download Site Available for NCBI Databases and Tools

An [Aspera protocol download site](#) is available as an alternative to FTP for all NCBI downloads. The Aspera protocol provides a much faster transfer rate and is most important for downloading very large data sets such as those from next-generation sequencing studies, but can be used to improve download performance for any public NCBI data files or software packages. The Aspera protocol site requires the free AsperaConnect client application available from [Aspera Connect](#). The [Aspera Transfer Guide](#), available on the [NCBI Bookshelf](#), provides additional information on using the fast download site.

## 1000 Genomes Project Data Now on Amazon Cloud Service

As announced in the recent [NIH press release](#), data from the 1000 Genomes project - the world's largest set of data on human genetic variation produced by the international 1000 Genomes Project — are now publicly available on the [Amazon Web Services \(AWS\) cloud](#). 1000 genomes data may also be downloaded from the NCBI though [FTP](#) or through the [Aspera protocol site](#).

## Microbial Genomes Update

One hundred ninety nine finished microbial (archaeal and bacterial) genomes were released from November 2011 through March 2012. The original sequence data files submitted to the International Sequence Database Collaboration (INSDC) are available in the [Bacteria directory](#) in the genomes area of the GenBank FTP site. RefSeq provisional versions were released for a selected set of 118 of the complete INSDC microbial genomes during the same period. These are available from the [/genomes/Bacteria](#) directory on the FTP site.

In addition, data from 1,135 microbial whole genome-shotgun (WGS) sequencing projects were added to the INSDC during this period. The original submitted files are available in the [Bacteria\\_DRAFT](#) directory in the GenBank genomes area. RefSeq provisional versions of 210 WGS microbial projects were released in the [/genomes/Bacteria\\_DRAFT](#) area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

## NCBI Articles in Nucleic Acids Research Database Issue

The [Nucleic Acids Research 2011 Database Issue](#) contains 10 articles about NCBI resources, tools, and databases including BioAssay, SRA, GEO, BioProject / BioSample, Taxonomy Epigenomics, MMDB (Structure), RefSeq and GenBank. Free full-text articles from the database issue are available from PubMed Central and the publisher's site and are linked to the [summaries](#) and [abstracts](#) in PubMed.

**Panel A: Assembly-Assembly**

Assembly-Assembly Clinical Remap A

**Genome Information**

Source Organism \*   
Start typing to get a list of available organisms

Source Assembly \*  
 HuRef  
 GRCh37.p5  
 NCBI36 (hg18)  
 NCBI35 (hg17)

Target Assembly \*  
 HuRef  
 NCBI36 (hg18)  
 NCBI35 (hg17)  
 NCBI34 (hg16)

Alignments performed: January 16, 2012

**First Pass**

GRCh37.p5 Coverage:  
 NCBI36 (hg18) Cover:  
 Percent Identity: 0.999

**Remapping Options**

Minimum ratio of bases that  
 Maximum ratio for differenc  
 Allow multiple locations to  
 Merge Fragments:

**Panel B: Clinical Remap**

Assembly-Assembly Clinical Remap B

**Genome Information**

Available only for human

I have data on \*  
 GRCh37 (hg19)  
 GRCh37.p7  
 NCBI36 (hg18)  
 RefSeqGene

I want to map data to \*  
 RefSeqGene

**Remapping Options**

**Define RefSeqGenes**

Map to any available RefSeqGene sequence  
 Map only to the RefSeqGenes I specify

**Define Transcripts/Proteins**

Provide locations on NMs/NPs associated with RefSeqGenes  
 Provide locations on NMs/NPs even if there is no RefSeqGene

Not all regions of the genome have RefSeqGenes. You can choose to get data for any available RefSeqGene or only specific ones. To request a RefSeqGene for a gene click [here](#)

**Panel C: Data**

Data C

Input format:  Output format:

Upload a file:

OR

Paste data here:

```
Chr19 57742377 57746915 AURKC
Chr19 1086578 1095391 POLR2E
Chr19 1205798 1228434 STK11
Chr19 45754550 45808541 MARK4
```

You can paste multiple lines into the text area

**Panel D: Output Data**

```
chr19 . Variation 196079 196079 . + . ID=rs4046282;gbkey=Variation
chr19 . Variation 196107 196107 . + . ID=rs3866749;gbkey=Variation
chr19 . Variation 196158 196162 . + . ID=rs4046286;gbkey=Variation
chr19 . Variation 196182 196182 . + . ID=rs3965607;gbkey=Variation
chr19 . Variation 486839 486839 . + . ID=rs16990554;gbkey=Variation
```

Figure 3. Submission forms for the Genome Remapping Service. **A.** Genome Remap set to map a set of locations from human build 37 to build 36. **B.** The Clinical Remap tab set to map a set of locations from build 37 to RefSeqGene records. **C.** BED format for gene position shown in the data text area for the Genome Remap. **D.** Data in GFF3 format showing the positions of variations to be projected on to RefSeqGene records in Clinical Remap.

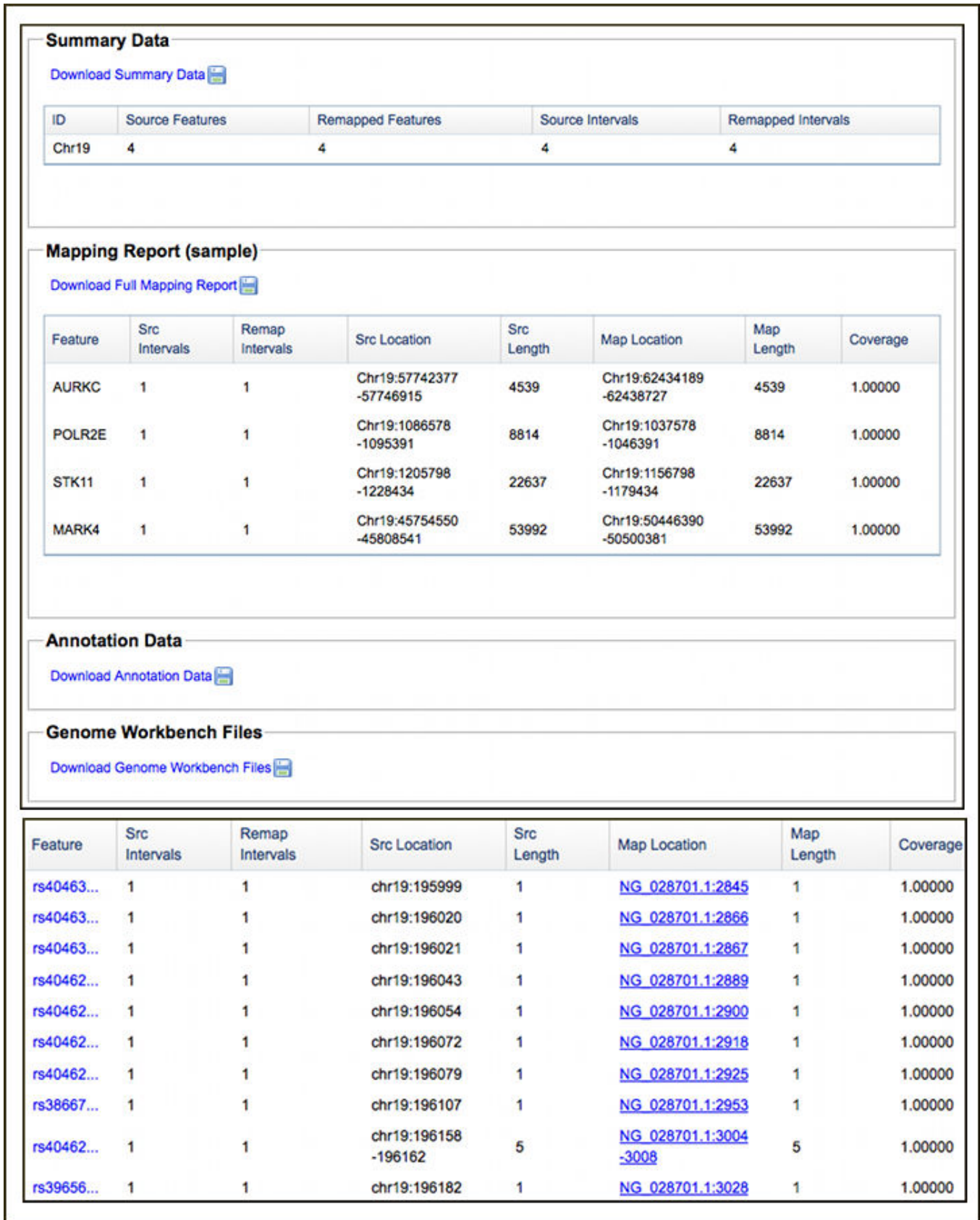


Figure 4. Output from the Remap service. *Top panel.* Results of projecting gene locations from human build 36 onto build 37. The output provides downloadable results in the form of spreadsheets (Mapping Report and Annotation Data). Annotation data are also available in a format that can be loaded into NCBI's [Genome Workbench](#), a standalone sequence analysis and annotation platform. *Bottom panel.* Mapping Report from Clinical Remap showing the projection of variations onto RefSeqGene records. The Clinical Remap Service also produces the Summary Data, Annotation Data and Genome Workbench files.

**Genome Information**

Please select the organism and assembly upon which your variants are annotated.

Select Organism \*

Select Assembly \*   
 GRCh37 (hg19) patch release  
 NCBI36 (hg18)

**Data**

Input format:

Upload a file:

OR

Paste data here: 

```
NC_000019:g.46201890C>T
NC_000007:g.150883955T>A
NC_000019:g.45411941T>C
NC_000007:g.113528849T>C
NC_000011:g.66828674A>G
```

You can paste multiple variations into the area.

Click on the value in the Submitted Loc column to show it in Sequence Viewer. [Download Report](#)

Submitted Id	Submitted Loc	Reported Allele	Cytoband	NCBI Id	Somatic Observed?	GMAF	Clinical Information	PubMed	Consequences
NC_000019:g.46201890C>T	<a href="#">NC_000019:g.46201890-46201890</a>	NC_000019:g.46201890C>T	19q13.3	<a href="#">rs151211223</a>	no				non_synonymous_codon
NC_000007:g.150883955T>A	<a href="#">NC_000007.13:150883955-150883955</a>	NC_000007.13:g.150883955T>A	7q36	<a href="#">rs151344617</a>	no			<a href="#">1</a>	non_synonymous_codon
NC_000007:g.150883955T>A	<a href="#">NC_000007.13:150883955-150883955</a>	NC_000007.13:g.150883955T>A	7q36	<a href="#">rs151344617</a>	no			<a href="#">1</a>	intron_variant
NC_000019:g.45411941T>C	<a href="#">NC_000019:g.45411941-45411941</a>	NC_000019:g.45411941T>C	19q13.3	<a href="#">rs429358</a>	yes	C: 0.154		<a href="#">67</a>	non_synonymous_codon
NC_000007:g.113528849T>C	<a href="#">NC_000007.13:113528849-113528849</a>		7q31.1						
NC_000011:g.66828674A>G	<a href="#">NC_000011:g.66828674-66828674</a>		11q13.2						

NC\_000019.9: 46M\_46M (108b+) - Find on Sequence:

46,201,840 46,201,850 46,201,860 46,201,870 46,201,880 46,201,890 46,201,900 46,201,910 46,201,920 46,201,930 46,201,940

SNP

rs186227446 rs141204641 rs151211223

Genes

rs106227446 rs141204641

QPCTL  
 total range: NC\_000019.9 (46,195,741..46,207,248)  
 total length: 11,508  
 strand: plus

Links & Tools  
 View GeneID: [58814 \(QPCTL\)](#)  
 View HGNC: [25952](#)  
 View HPRD: [28612](#)

Clinical Variants

Cited Variants

Figure 5. The Variation Reporter submission form and results. Top panel. Submission form maps locations of variations onto human genome builds. The input data in this case are variations in Human Genome Variation Society (HGVS) notation. Bottom panel. Results of mapping the variations onto build 37. The first four of the six variations map to NCBI Reference SNP locations. The corresponding identifiers and other information from dbSNP is shown for each of these. The location and genomic context for each mapped location is available for each of the mapped locations in the graphical sequence viewer. Clicking the linked location (red arrow) loads that marker and surrounding region in the sequence viewer.

## GenBank News

GenBank release 189 is available through Entrez, BLAST and from the [GenBank FTP](#) area. The current release incorporates data available as of April 15, 2011 and, with the whole-genome shotgun portion, contains 411,959,832,946 bases from 232,729,719 sequence records. [Release notes](#) describe the current state of data and upcoming changes. The [GenBank page](#) provides more information on the database content and scope as well as submission information.

## RefSeq News

### RefSeq Release 52

RefSeq Release 52 is available through Entrez, BLAST, and from the [RefSeq FTP area](#). The current release includes 20.2 million Reference Sequence records from 16,923 different species or strains. The [RefSeq release notes](#) provide more detailed information.

### RefSeq Genome Annotation Files in GFF3 Format

NCBI now offers Reference Sequence (RefSeq) genome annotation files in the latest [Generic Feature Format \(GFF3\)](#) specification (1.20). RefSeq genome data can be downloaded from the [genomes area](#) of the NCBI FTP site. GFF3 files are in the GFF directory within each organism directory. Currently GFF3 files are available for the NCBI annotations of the latest assemblies for [human](#), [cow](#), [dog](#), [chicken](#), and many others.

## Keeping Up with NCBI

Seventeen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the [Announcement List summary page](#). Subscribe to the [NCBI Announce list](#) to receive updates on the NCBI News.

Twenty-five [RSS feeds](#) are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's [Facebook](#) page and [Twitter feed](#) also provide updates on NCBI resources.

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, November 2011

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: November 16, 2011; Updated: November 16, 2011.

### Phase One Rollout of the New Genome Site

A completely redesigned Genome site, [www.ncbi.nlm.nih.gov/genome](http://www.ncbi.nlm.nih.gov/genome), is now available. Major improvements include a more natural organization at the level of the organism for prokaryotic, eukaryotic, and viral genomes. Reports include information about the availability of nuclear or prokaryotic primary genomes as well as organelles and plasmids. The new Genome resource provides a summary view of the data from all genome-scale projects including genome maps, assemblies, annotation, and transcriptomes. Genome collects data from primary data resources and provides links to more detailed information. While not new with this release, it is worth noting that the Genome interface has been upgraded to the new NCBI standard with the new search bar, Limits and Advanced Search pages, and NCBI footer. Moreover, search results and record views in Genome are discovery-oriented and feature the Discovery Column with analysis tools and easy access to related data. Figure 1 shows sample pages from the Genome resource and highlights these new features. An [information page](#) accompanying the release provides additional details and help with transitioning to the new service.

The new Genome site is much easier to navigate and provides rapid access to all genome data for a particular organism. The new site will continue to improve as additional displays and features are added in phases. A feature article in the next NCBI News will provide more detailed coverage of Genome with examples illustrating the power of the new system.

### Note: Changes affecting genome identifiers

Because of the reorganization to a natural classification system, older genome identifiers are no longer valid. Typically these genome identifiers were not exposed in the previous system and were used mainly for programmatic access. To aid in the transition to the new system, a [file](#) that maps previous Genome identifiers to the identifier for the genome sequence is available on the FTP site. The sequence identifiers can be used to retrieve the genome sequence from the Nucleotide database.

### New BLAST videos on NCBI's YouTube channel


Two new instructional videos about BLAST statistics are on the NCBI YouTube channel: [An explanation of BLAST E-values \(pt.1\)](#), and [Answers to a few E-value FAQs \(pt.2\)](#). These BLAST videos should help in interpreting BLAST output and designing more effective BLAST search strategies. The BLAST videos join a growing collection of 54 videos on the [NCBI YouTube channel](#).

The image displays three overlapping screenshots of the NCBI Genome browser interface. The top-left screenshot shows the 'Genome information for human (Homo sapiens)' page, featuring a chromosome map, assembly statistics, and a table of related bio-projects. The top-right screenshot shows the 'Genome information for Staphylococcus aureus' page, including a sub-species tree and a table of genome projects. The bottom-left screenshot shows search results for the query 'plants', listing species like Zea mays, Oryza sativa, and Physcomitrella patens. The bottom-right screenshot shows the 'Genome information for Staphylococcus aureus subsp. aureus MRSA252' page, which includes a table of genomic features and a graphical sequence viewer of the chromosome.

Figure 1. Sample Genome pages. *Top panel, right*: Species-level page for human showing the right hand Discovery Column. *Top panel, left*: Species-level page for the bacterium *Staphylococcus aureus*. The subspecies tree has links to strain-level pages. *Bottom panel, right*: Strain-level page for the antibiotic resistant *S. aureus* MRSA252 showing the imbedded graphical sequence viewer display of the chromosome. *Bottom panel, left*: Sample search results for the query “plants” showing the updated search result page.



## BLAST Results: Expect Values, part 1



54 videos
▾

Subscribe

www.youtube.com/ncbinlm

Alignments

Select All   [Get selected sequences](#)   [Distance tree of results](#)   [Multiple alignment](#)

```


> ref|NP\_001008976.1 | UGM apolipoprotein A-I [Pan troglodytes]
Length=100
GENE ID: 449498 APOA2 | apolipoprotein A-I [Pan troglodytes]
(10 or fewer PubMed links)
Score = 177 bits (448), Expect = 2e-61, Method: Compositional matrix adjust.
Identities = 97/100 (97%), Positives = 100/100 (100%), Gaps = 0/100 (0%)

Query 1  MKLLAATVLLLITCSLEGALVRRQAKEPCVESLVSQYFQTVTDYGKDLMEKVKSPELQAE  60
          MKLLAATVLLLITCSLEGALVRRQAKEPCV++LVSQYFQTVTDYGKDLMEKVKSPELQAE
Sbjct 1  MKLLAATVLLLITCSLEGALVRRQAKEPCVDNLVSQYFQTVTDYGKDLMEKVKSPELQAE  60

Query 61  AKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQPATO  100
          AKSYFEKSKEQLTPLIKKAGTELVNFLSYFVELGTQPATO
                    
```

Homology?

## BLAST Results: Expect Values, part 2



54 videos
▾

Subscribe

### Question #3:

What is an E-value of 0.0?

E-value < 1e-179

2:55 / 3:08

CC ■ 360p

## Entrez Utility Changes: New EFetch Version and New alternative ESummary XML

An updated EFetch Entrez Utility (version 2.0) is now in production. EFetch retrieves records from the NCBI databases by unique identifier. Additions include support for the BioSample , BioProjects, and SRA databases as

well as defined default values for retrieval mode (retmode) and retrieval type (rettype). Updated retmode and rettype values are given in the [table](#) in the [Entrez Programming Utilities Help Manual](#).

An alternative XML record is now available from the ESummary Entrez Utility. The content in the new record is unique to each Entrez database and has additional content not available in the traditional ESummary record. The new XML can be requested by including &version=2.0 in the ESummary URL. The traditional ESummary record will continue to be supported and will be returned without the version parameter. The [traditional](#) and [version 2.0](#) ESummary for NM\_000240 (gi=33469954) from the nuccore (nucleotide) database are shown below. The [release notes](#) provide details on changes. The [Entrez Programming Utilities Help Manual](#) has complete information on EFetch, ESummary, and the other EUtility programs.

Traditional XML	ESummary XML for Nucleotide NM_000240
<pre> &lt;eSummaryResult&gt;   --&lt;DocSum&gt;     &lt;Id&gt;33469954&lt;/Id&gt;     &lt;Item Name="Caption" Type="String"&gt;NM_000240&lt;/Item&gt;     --&lt;Item Name="Title" Type="String"&gt;       Homo sapiens monoamine oxidase A (MAOA), nuclear gene encoding mitochondrial protein, mRNA     &lt;/Item&gt;     &lt;Item Name="Extra" Type="String"&gt;gi33469954[refNM_000240.2][33469954]&lt;/Item&gt;     &lt;Item Name="Gi" Type="Integer"&gt;33469954&lt;/Item&gt;     &lt;Item Name="CreateDate" Type="String"&gt;1999/04/01&lt;/Item&gt;     &lt;Item Name="UpdateDate" Type="String"&gt;2011/10/23&lt;/Item&gt;     &lt;Item Name="Flags" Type="Integer"&gt;512&lt;/Item&gt;     &lt;Item Name="TaxId" Type="Integer"&gt;9606&lt;/Item&gt;     &lt;Item Name="Length" Type="Integer"&gt;4090&lt;/Item&gt;     &lt;Item Name="Status" Type="String"&gt;live&lt;/Item&gt;     &lt;Item Name="ReplacedBy" Type="String"/&gt;     &lt;Item Name="Comment" Type="String"&gt;&lt;/Item&gt;   &lt;/DocSum&gt; &lt;/eSummaryResult&gt; </pre>	<pre> --&lt;Statistics&gt;   &lt;Stat type="Length" count="4090"/&gt;   &lt;Stat type="Length" subtype="literal" count="4090"/&gt;   &lt;Stat type="all" count="6"/&gt;   &lt;Stat type="blob_size" count="16469"/&gt;   &lt;Stat type="cdregion" count="1"/&gt;   &lt;Stat type="cdregion" subtype="CDS" count="1"/&gt;   &lt;Stat type="gene" count="1"/&gt;   &lt;Stat type="gene" subtype="Gene" count="1"/&gt;   &lt;Stat type="imp" count="3"/&gt;   &lt;Stat type="imp" subtype="polyA_signal" count="1"/&gt;   &lt;Stat type="imp" subtype="polyA_site" count="2"/&gt;   &lt;Stat type="org" count="1"/&gt;   &lt;Stat type="pub" count="10"/&gt;   &lt;Stat type="pub" subtype="PubMed" count="5"/&gt;   &lt;Stat type="pub" subtype="PubMed/Gene-ref" count="5"/&gt;   &lt;Stat source="CDS" type="all" count="13"/&gt;   &lt;Stat source="CDS" type="prot" count="1"/&gt;   &lt;Stat source="CDS" type="region" count="2"/&gt;   &lt;Stat source="CDS" type="region" subtype="Region" count="2"/&gt;   &lt;Stat source="CDS" type="site" count="10"/&gt;   &lt;Stat source="CDS" type="site" subtype="Site" count="10"/&gt;   &lt;Stat source="CDS/CDD" type="all" count="6"/&gt;   &lt;Stat source="CDS/CDD" type="region" count="6"/&gt;   &lt;Stat source="CDS/CDD" type="region" subtype="Region" count="6"/&gt;   &lt;Stat source="CDS/SNP" type="all" count="30"/&gt;   &lt;Stat source="CDS/SNP" type="imp" count="30"/&gt;   &lt;Stat source="CDS/SNP" type="imp" subtype="variation" count="30"/&gt;   &lt;Stat source="Exon" type="all" count="15"/&gt;   &lt;Stat source="Exon" type="evidence" count="15"/&gt;   &lt;Stat source="Exon" type="imp" count="15"/&gt;   &lt;Stat source="Exon" type="imp" subtype="exon" count="15"/&gt;   &lt;Stat source="SNP" type="all" count="42"/&gt;   &lt;Stat source="SNP" type="imp" count="42"/&gt;   &lt;Stat source="SNP" type="imp" subtype="variation" count="42"/&gt;   &lt;Stat source="STS" type="all" count="9"/&gt;   &lt;Stat source="STS" type="imp" count="9"/&gt;   &lt;Stat source="STS" type="imp" subtype="STS" count="9"/&gt;   &lt;Stat source="all" type="Length" count="4090"/&gt;   &lt;Stat source="all" type="all" count="121"/&gt;   &lt;Stat source="all" type="blob_size" count="16469"/&gt;   &lt;Stat source="all" type="cdregion" count="1"/&gt;   &lt;Stat source="all" type="evidence" count="15"/&gt;   &lt;Stat source="all" type="gene" count="1"/&gt;   &lt;Stat source="all" type="imp" count="99"/&gt;   &lt;Stat source="all" type="org" count="1"/&gt;   &lt;Stat source="all" type="prox" count="1"/&gt;   &lt;Stat source="all" type="pub" count="10"/&gt;   &lt;Stat source="all" type="region" count="8"/&gt;   &lt;Stat source="all" type="site" count="10"/&gt; &lt;/Statistics&gt; </pre>
Version 2.0 XML	ESummary XML for Nucleotide NM_000240
<pre> &lt;eSummaryResult&gt;   --&lt;DocumentSummarySet status="OK"&gt;     --&lt;DocumentSummary uid="33469954"&gt;       &lt;Caption&gt;NM_000240&lt;/Caption&gt;       --&lt;Title&gt;         Homo sapiens monoamine oxidase A (MAOA), nuclear gene encoding mitochondrial protein, mRNA       &lt;/Title&gt;       &lt;Extra&gt;gi33469954[refNM_000240.2]&lt;/Extra&gt;       &lt;Gi&gt;33469954&lt;/Gi&gt;       &lt;CreateDate&gt;1999/04/01&lt;/CreateDate&gt;       &lt;UpdateDate&gt;2011/10/23&lt;/UpdateDate&gt;       &lt;Flags&gt;512&lt;/Flags&gt;       &lt;TaxId&gt;9606&lt;/TaxId&gt;       &lt;Slen&gt;4090&lt;/Slen&gt;       &lt;Biomol&gt;mRNA&lt;/Biomol&gt;       &lt;MolType&gt;rna&lt;/MolType&gt;       &lt;Topology&gt;linear&lt;/Topology&gt;       &lt;SourceDb&gt;refseq&lt;/SourceDb&gt;       &lt;SegSetSize&gt;0&lt;/SegSetSize&gt;       &lt;ProjectId&gt;0&lt;/ProjectId&gt;       &lt;Genome&gt;genomic&lt;/Genome&gt;       &lt;SubType&gt;chromosomeimap&lt;/SubType&gt;       &lt;SubName&gt;XIXp11.3&lt;/SubName&gt;       &lt;AssemblyG&gt;14165523&lt;/AssemblyG&gt;       &lt;AssemblyAcc&gt;X60819.1&lt;/AssemblyAcc&gt;       &lt;Tech&gt;       &lt;Completeness&gt;has-right&lt;/Completeness&gt;       &lt;GeneticCode&gt;1&lt;/GeneticCode&gt;       &lt;Strand&gt;       &lt;Organism&gt;Homo sapiens&lt;/Organism&gt;       +&lt;Statistics&gt;&lt;/Statistics&gt;       &lt;AccessionVersion&gt;NM_000240.2&lt;/AccessionVersion&gt;       &lt;Properties na="1"&gt;1&lt;/Properties&gt;       &lt;Comment&gt;       &lt;OSLT indexed="yes"&gt;NM_000240.2&lt;/OSLT&gt;       &lt;IdGiClass mol="2" repr="2" gi_state="10" sat="4" sat_key="59045920" owner="20" sat_name="NCBI" o       &lt;/DocumentSummary&gt;     &lt;/DocumentSummarySet&gt;   &lt;/eSummaryResult&gt; </pre>	<pre> --&lt;Statistics&gt;   &lt;Stat type="Length" count="4090"/&gt;   &lt;Stat type="Length" subtype="literal" count="4090"/&gt;   &lt;Stat type="all" count="6"/&gt;   &lt;Stat type="blob_size" count="16469"/&gt;   &lt;Stat type="cdregion" count="1"/&gt;   &lt;Stat type="cdregion" subtype="CDS" count="1"/&gt;   &lt;Stat type="gene" count="1"/&gt;   &lt;Stat type="gene" subtype="Gene" count="1"/&gt;   &lt;Stat type="imp" count="3"/&gt;   &lt;Stat type="imp" subtype="polyA_signal" count="1"/&gt;   &lt;Stat type="imp" subtype="polyA_site" count="2"/&gt;   &lt;Stat type="org" count="1"/&gt;   &lt;Stat type="pub" count="10"/&gt;   &lt;Stat type="pub" subtype="PubMed" count="5"/&gt;   &lt;Stat type="pub" subtype="PubMed/Gene-ref" count="5"/&gt;   &lt;Stat source="CDS" type="all" count="13"/&gt;   &lt;Stat source="CDS" type="prot" count="1"/&gt;   &lt;Stat source="CDS" type="region" count="2"/&gt;   &lt;Stat source="CDS" type="region" subtype="Region" count="2"/&gt;   &lt;Stat source="CDS" type="site" count="10"/&gt;   &lt;Stat source="CDS" type="site" subtype="Site" count="10"/&gt;   &lt;Stat source="CDS/CDD" type="all" count="6"/&gt;   &lt;Stat source="CDS/CDD" type="region" count="6"/&gt;   &lt;Stat source="CDS/CDD" type="region" subtype="Region" count="6"/&gt;   &lt;Stat source="CDS/SNP" type="all" count="30"/&gt;   &lt;Stat source="CDS/SNP" type="imp" count="30"/&gt;   &lt;Stat source="CDS/SNP" type="imp" subtype="variation" count="30"/&gt;   &lt;Stat source="Exon" type="all" count="15"/&gt;   &lt;Stat source="Exon" type="evidence" count="15"/&gt;   &lt;Stat source="Exon" type="imp" count="15"/&gt;   &lt;Stat source="Exon" type="imp" subtype="exon" count="15"/&gt;   &lt;Stat source="SNP" type="all" count="42"/&gt;   &lt;Stat source="SNP" type="imp" count="42"/&gt;   &lt;Stat source="SNP" type="imp" subtype="variation" count="42"/&gt;   &lt;Stat source="STS" type="all" count="9"/&gt;   &lt;Stat source="STS" type="imp" count="9"/&gt;   &lt;Stat source="STS" type="imp" subtype="STS" count="9"/&gt;   &lt;Stat source="all" type="Length" count="4090"/&gt;   &lt;Stat source="all" type="all" count="121"/&gt;   &lt;Stat source="all" type="blob_size" count="16469"/&gt;   &lt;Stat source="all" type="cdregion" count="1"/&gt;   &lt;Stat source="all" type="evidence" count="15"/&gt;   &lt;Stat source="all" type="gene" count="1"/&gt;   &lt;Stat source="all" type="imp" count="99"/&gt;   &lt;Stat source="all" type="org" count="1"/&gt;   &lt;Stat source="all" type="prox" count="1"/&gt;   &lt;Stat source="all" type="pub" count="10"/&gt;   &lt;Stat source="all" type="region" count="8"/&gt;   &lt;Stat source="all" type="site" count="10"/&gt; &lt;/Statistics&gt; </pre>

## Highlight Features Link Now on Sequence Records

A Highlight Features link now appears in the Analyze this Sequence section of protein and nucleotide sequence records displayed at the NCBI site. This link activates the new Feature Highlight function described in the

August 2011 NCBI News. Clicking the link opens the Feature Highlight Bar and highlights the first coding sequence (CDS) feature or the first linked feature if no CDS feature is present.

As pointed out in the original NCBI News article, the Highlight Features function is helpful in visualizing the extent and location of such important features as genes, coding regions, exons, and mRNAs in nucleotide sequences and conserved domains, modification sites, and interaction sites in protein sequences. This function joins the links to BLAST, Primer-BLAST, and Conserved Domain searches as well as the Find-in-Sequence pattern finder as on-the-fly analysis capabilities in the NCBI sequence databases. The image below shows the link and the activated CDS highlight on the ReSeqGene record for the human monoamine oxidase gene (NG\_008957).

**Homo sapiens monoamine oxidase A (MAOA), RefSeqGene on chromosome X**

NCBI Reference Sequence: NG\_008957.1

FASTA Graphics

Go to: [dropdown]

LOCUS NG\_008957 97660 bp  
 DEFINITION Homo sapiens monoamine oxidase X.  
 ACCESSION NG\_008957  
 VERSION NG\_008957.1 GI:212549708  
 KEYWORDS RefSeqGene.  
 SOURCE Homo sapiens (human)  
 ORGANISM [Homo sapiens](#)  
 Eukaryota; Metazoa; Chordata; C  
 Mammalia; Eutheria; Euarchontog  
 Catarrhini; Hominidae; Homo.  
 COMMENT REVIEWED [REFSEQ](#): This record has  
 reference sequence was derived  
[BX530072.B](#) and [BX537148.1](#).  
 This sequence is a reference standard in the [RefSeqGene](#) project.

Change region shown [dropdown]  
 Customize view [dropdown]

Analyze this sequence [dropdown]  
 Run BLAST  
 Pick Primers  
**Highlight Sequence Features**  
 Find in this Sequence

Articles about the MAOA gene [dropdown]  
<sup>3</sup>H kinetic isotope effects and pH dependence of catalysis as mechanistic probe [Biochemistry. 2011]  
 Association of the MAOA promoter uVNTR polymorphism with suicide [BMC Med Genet. 2011]  
 Search for association between suicide and 5-HTT, MAOA and DAT [Arch Med Sadowe] Kryminol. 2010]

92581 accctgccca ccttcccaag taactctgtg taacctcttg gttcccttga aggggtgatcc  
 92641 gtaaacccgt gggcaggatt ttctttgggg gcacagagac tgccacaaag tggagcggct  
 92701 acatggaag ggcagttgag gctggagaac gacagactag ggagtaagc aggaaagccc  
 92761 aggctctctc cctccagagt cacggcaacg tttttggcat ctggcttctg tagttcttga  
 92821 cactgataga atctgtatgt ccatttctct gccctcact catggggctc  
 92881 agcagggcct tgaatctgta gaaactatac agcctctttt cataataacc  
 92941 cttcaggtc taaatggtc tcgggaaggt gaccgagaaa gatatctgg  
 93001 tgaatcaang gtaagtttgg tgactctggg cactatctct ccttagacc  
 93061 ataaactca catctccctt ctctagcct cggatttaat tatagatgc  
 93121 gggctccat gcattgatct tgcagtgtt ttgtctctct tgtcagcat  
 93181 tcatgatctg tttctctca tetaggacgt tcacgggta gaaatcacc  
 93241 ggaagggaac ctgccctctg tttctggcct gctgaagtc attggattc  
 93301 aactgccctg gggtttctg tgtacaata caagctcctg ccacggctc  
 93361 tcttatgctc tctgctcact ggtttcaat accaccaaga ggaataat  
 93421 aaggctgtgt cattgggcca tgtttaagtg tactggattt aactacctt  
 93481 caatcattgt taaagtaaaa acaattcaaa gaatcaccta attaatttc  
 93541 gctccatctt atttgcagt gtagatcaac tcatgttaat tgatagaat  
 93601 atcaactttc gaaattcaca aagttaaacg tgatgtgctc atcagaaaac  
 93661 cctgttttta ttcccttcaa tgcaaaatac atgatgattt cagaaacaa  
 93721 ttctgtctgt ggaggtggag taggtgaagg ccagcctgt aactgtcct  
 93781 taggcaatgg tgaactgtca ttacagagcc tagaggctca cagcctcct  
 93841 cctccacttt ggatcaggaa atagtaaagg aaagcagtgt tgggggtag  
 93901 ccctcagacc agaatgggga catcttctgt tctgctgctc caggaatct

join(5182..5254,32353..32447,42130..42267,60711..60815,  
 61544..61635,77012..77153,80080..80229,80533..80692,  
 81538..81634,85066..85119,89520..89577,90789..90886,  
 92633..92744,92948..93010,93206..93352)  
 /gene="MAOA"  
 /EC\_number=" 1.4.3.4 "  
 /note="MAO-A; monoamine oxidase type A"  
 /codon\_start=1  
 /product="amine oxidase [flavin-containing] A"  
 /protein\_id=" NP\_000231.1 "  
 /db\_xref="GI:4557735"  
 /db\_xref="CCDS: CCDS14260.1 "  
 /db\_xref="GeneID: 4128 "  
 /db\_xref="HGNC: 6833 "  
 /db\_xref="MIM: 309850 "

CDS Feature 1 of 1 NG\_008957 : 15 segments

Details [dropdown] Display: FASTA GenBank Help [X]

## New BLAST 16S Prokaryotic Ribosomal RNA Database

A prokaryotic 16S ribosomal RNA database is now available through the database pull-down list on the main nucleotide BLAST service. The 16S database contains both bacterial and archaeal sequences from two RefSeq Targeted Loci projects (BioProjects PRJNA33175 and PRJNA33317). These data represent near full-length 16S

ribosomal RNA sequences from more than 250 archaeal and 7200 bacterial strains. The 16S BLAST database is useful for identifying or establishing the taxonomic affinities of unknown bacterial 16S sequences such as those from environmental or organismal samples or metagenomes. Figure 2 shows how the database can be used to partially classify a 16S sequence (JF340503) obtained from a concrete sewer biofilm (PubMed: 21981064, PopSet: 330372088). The top panel of the figure shows the basic nucleotide BLAST form with the 16S database selected. The center panel shows the BLAST results (RID: CUR81JZY012). The results indicate that the query sequence has the closest affinity to the acetobacteriaceae, particularly *Acidocella facilis*. The [BLAST Distance Tree](#), also shown in the figure provides a useful way to see the results of the analysis at a glance.

The pre-formatted 16S microbial database is also available in the [BLAST db FTP directory](#) as the file [16SMicrobial.tar.gz](#).

## New Phenotype-Genotype Integrator (PheGenI)

The [Phenotype-Genotype Integrator \(PheGenI\)](#) is a new service that integrates genome-wide association study (GWAS) catalog data from NHGRI with molecular and literature databases at the NCBI. PheGenI takes chromosome location, gene, SNP, or phenotype as input and provides annotated tables of SNPs, genes, association results, and gene expression data. A new [tutorial video](#) on YouTube demonstrates how to use PheGenI.

## Eukaryotic Genome Builds and Updates

Twelve new genome assemblies with annotations have recently been released at the NCBI. Nine of the new builds are genomes that make their first appearance (build 1.1) at NCBI. Highlights include the first genome for a sponge (*Amphimedon queenslandica*), the first for a reptile – the green anole (*Anolis carolinensis*), the first for a perciform fish – the Nile tilapia (*Oreochromis niloticus*), and two new rodent genomes – the guinea pig (*Cavia porcellus*) and the Chinese hamster CHO-K1 cell line (*Cricetulus griseus*). In addition updated annotations for six more genomes are also available including human build 37.3 described in the next section. A complete list of new builds and updates is given below. The NCBI [BioProject](#), [Genome](#), [Gene](#), [Nucleotide](#), [Protein](#), [BLAST](#) and [Map Viewer](#) services provide access to these data. The assemblies and annotations may also be downloaded from the [genomes area](#) of the FTP site.

### First NCBI Builds (build 1.1)

Sponge (*Amphimedon queenslandica*) [[BioProject](#), [Map Viewer](#)]

Buff-tailed bumblebee (*Bombus terrestris*) [[BioProject](#), [Map Viewer](#)]

Nile tilapia (*Oreochromis niloticus*) [[BioProject](#), [Map Viewer](#)]

Domestic turkey (*Meleagris gallopavo*) [[BioProject](#), [Map Viewer](#)]

Green Anole (*Anolis carolinensis*) [[BioProject](#), [Map Viewer](#)]

Guinea pig (*Cavia porcellus*) [[BioProject](#), [Map Viewer](#)]

African savannah elephant (*Loxodonta africana*) [[BioProject](#), [Map Viewer](#)]

White-faced gibbon (*Nomascus leucogenys*) [[BioProject](#), [Map Viewer](#)]

Chinese hamster (CHO-K1 cell line) (*Cricetulus griseus*) [[BioProject](#), [Map Viewer](#)]

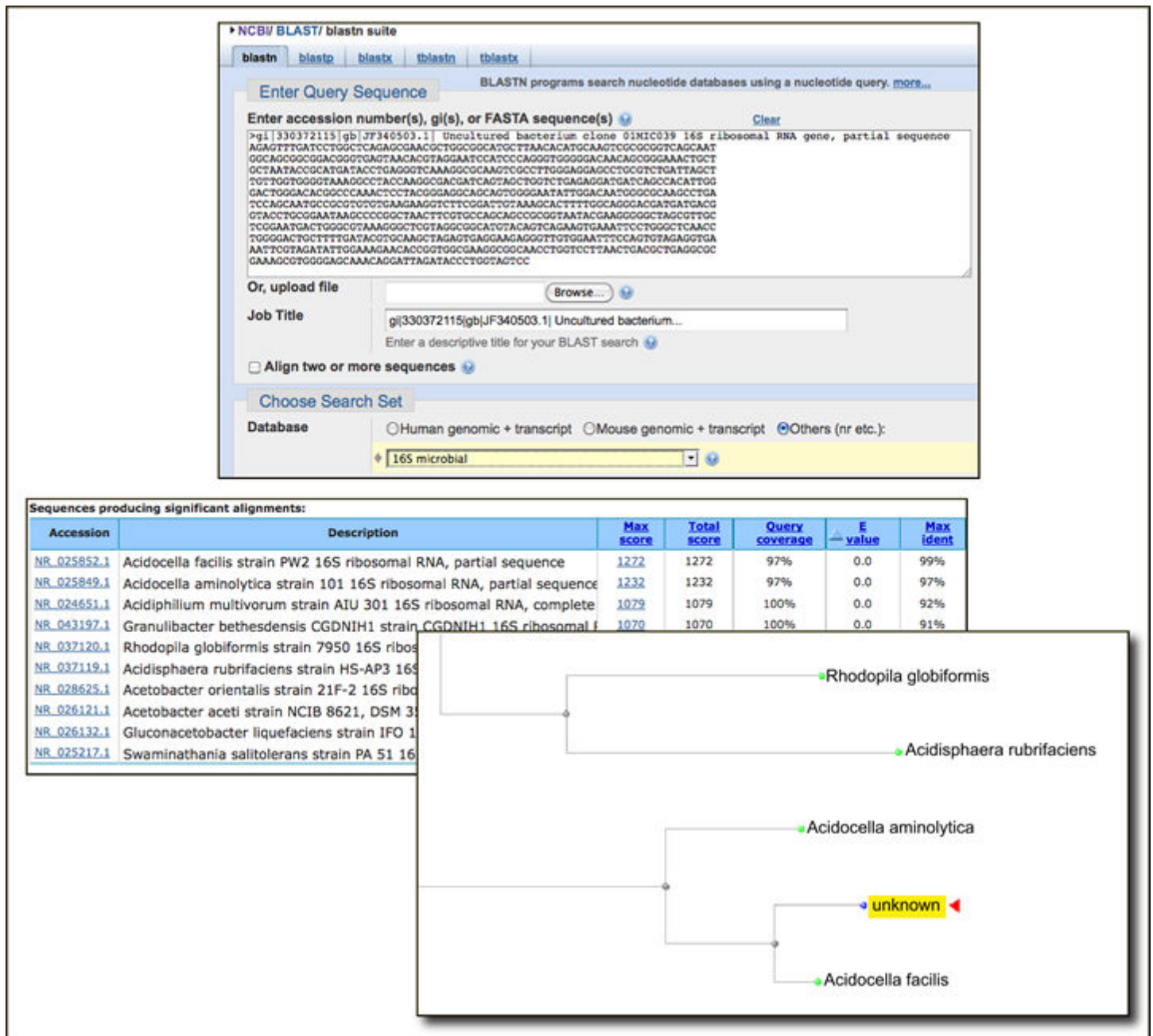


Figure 2. Using the NCBI nucleotide BLAST service with the new 16S microbial rRNA database. *Top panel.* The nucleotide BLAST search form with the 16S microbial database selected. The query sequence (JF340503) is a 16S sequence obtained from an environmental biofilm (PopSet: 330372088). *Center panel.* BLAST results (RID: CUR81/ZY012). The best match is to the 16S ribosomal RNA sequence (NR\_025852) from *Acidocella facilis* strain PW2. The linked BLAST Distance Tree of the results (bottom panel) shows the placement within *Acidocella* at a glance.

## New Builds

Honeybee (*Apis mellifera*), build 5.1 [BioProject, Map Viewer]

Pea aphid (*Acyrtosiphon pisum*), build 2.1 [BioProject, Map Viewer]

Zebrafish (*Danio rerio*), build 5.1 [BioProject, Map Viewer]

Chimpanzee (*Pan troglodytes*), build 3.1 [BioProject, Map Viewer]

## Updated Annotations

Fruit fly (*Drosophila melanogaster*), build 9.4 [[BioProject](#), [Map Viewer](#)]

Horse (*Equus caballus*), EquCab2.0 [[BioProject](#), [Map Viewer](#)]

Dog (*Canis lupus familiaris*), build 2.2 [[BioProject](#), [Map Viewer](#)]

Duck-billed platypus (*Ornithorhynchus anatinus*), build 1.2 [[BioProject](#), [Map Viewer](#)]

Thale cress (*Arabidopsis thaliana*), TAIR 10 [[BioProject](#), [Map Viewer](#)]

Human (*Homo sapiens*), build 37.3 [[BioProject](#), [Map Viewer](#)]

## Human Genome Update

The NCBI human genome annotation has been updated to version 37.3 and is now available in the [Map Viewer](#), the Entrez system and [human genome BLAST](#). The [build statistics](#) have more information on the contents of the release. The update includes the [Genome Reference Consortium](#) sequence patches from [patch 5](#). The patches are currently available as separate sequences from the chromosome assemblies. Patches that correct problems in the current assembly (fix patches) will be incorporated in the next complete genome assembly (build 38).

## Microbial Genomes Update

Fifty-eight finished microbial (archaeal and bacterial) genomes were released during September and October 2011. The original sequence data files submitted to the International Sequence Database Collaboration (INSDC) are available in the [Bacteria directory](#) in the genomes area of the GenBank FTP site. RefSeq provisional versions were released for a selected set of 32 of the complete INSDC microbial genomes during the same period. These are available from the [/genomes/Bacteria](#) directory on the FTP site.

In addition, data from 425 microbial whole genome-shotgun (WGS) sequencing projects were added to the INSDC during this period. The original submitted files are available in the [Bacteria\\_DRAFT](#) directory in the GenBank genomes area. RefSeq provisional versions of 89 WGS microbial projects were released in the [/genomes/Bacteria\\_DRAFT](#) area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

## GenBank News

GenBank release 186 is available through Entrez, BLAST and from the [GenBank FTP](#) area. The current release incorporates data available as of Oct 13, 2011 and, with the whole-genome shotgun portion, contains 350,733,781,429 bases from 212,788,863 sequence records. [Release notes](#) describe the current state of data and upcoming changes.

## RefSeq News

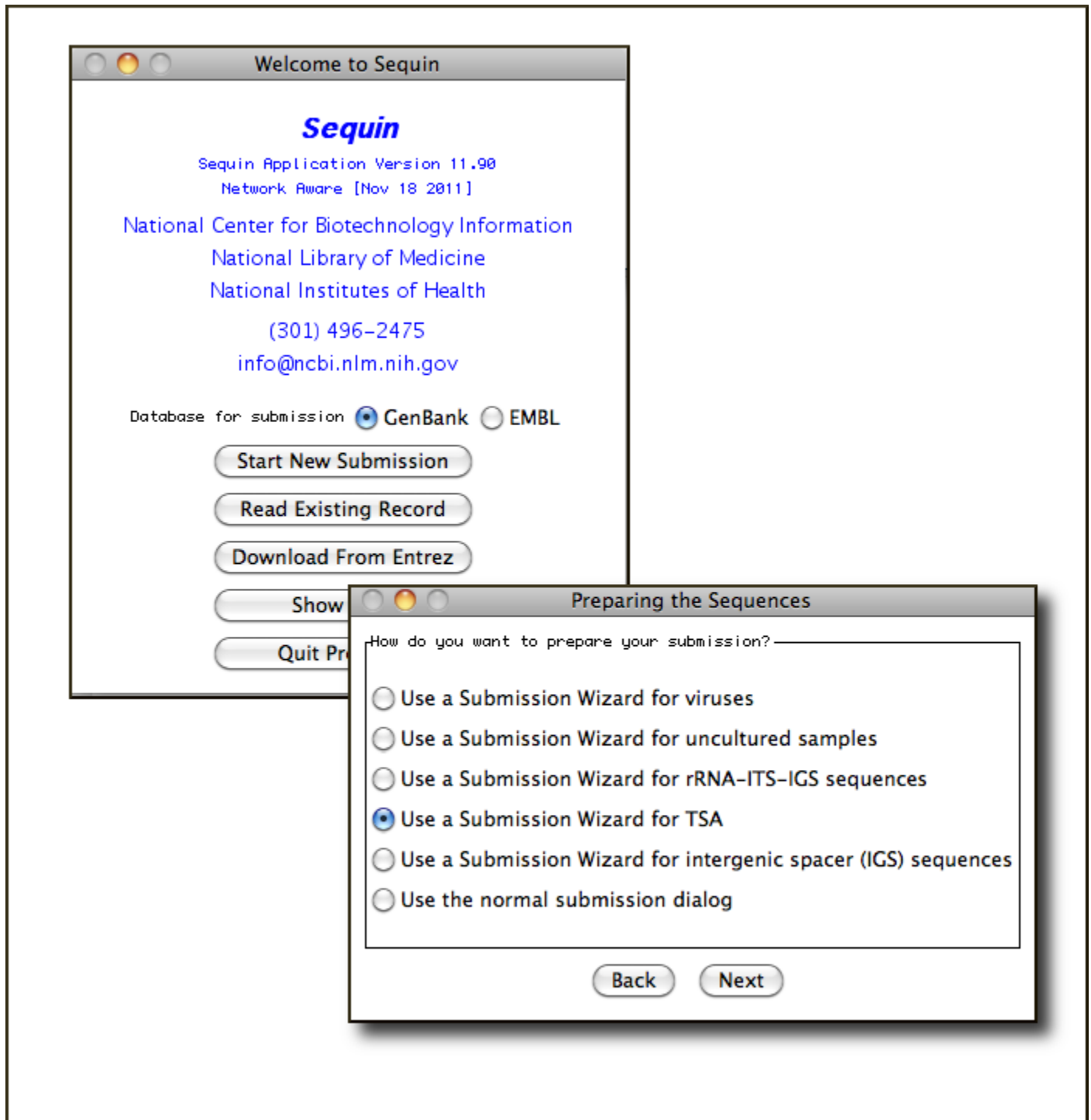
RefSeq Release 50 is available through Entrez, BLAST, and from the [RefSeq FTP area](#). The current release includes 18.8 million Reference Sequence records from 16,392 different species or strains. The RefSeq [release notes](#) provide more detailed information.

## Conserved Domain Database Update

Version 3.01 of the Conserved Domain Database is now available. The new release contains 298 new or updated NCBI-curated domain models. More detailed statistics are available from the [CDD News page](#). CDD matrices and other information can be downloaded from the [FTP site](#). CDD data are incorporated in the Entrez and BLAST search services at the NCBI Website.

## Sequin Now with Transcriptome Shotgun Assembly and Internal Transcribed Spacer Sequence Submission Wizards

A new version (11.90) of Sequin, the NCBI's standalone submission preparation software, is now available for [download](#). Packages are available for Linux, Unix, Windows, and Mac OSX systems. Improvements include new Submissions Wizards for Transcriptome Shotgun Assemblies (TSA) and ribosomal RNA intergenic spacer sequences (ITS); a Sequencing Method Page for information about the sequencing technology and assembly methods; a Sequence Deletion Tool for removing sequences from the submission; and updated feature and qualifier wizards complying with the latest INSDC Feature Documentation. The [Sequin page](#) has more information, a [Quick guide](#), [FAQs](#), and extensive [help documentation](#) on using Sequin to prepare submissions.



## NCBI C++ Toolkit Major Release

NCBI C++ Toolkit v7.0.0 is now available from the [FTP site](#). The [release notes](#) describe the highlights and contents of this release. The Toolkit contains C++ language sources of NCBI software that can be used to build standalone BLAST, Sequin, Cn3D, and other NCBI tools and utilities. The [NCBI C++ Toolkit Book](#) has in-depth information on working with the toolkit and provides access to source browsers and other useful resources.



## Announce Lists and RSS Feeds

Seventeen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the [Announcement List summary page](#). Subscribe to the [NCBI Announce list](#) to receive updates on the NCBI News.

Twenty-one [RSS feeds](#) are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's [Facebook](#) page and [Twitter feed](#) also provide updates on NCBI resources.

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, August 2011

Peter Cooper, Ph. D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: August 31, 2011; Updated: September 1, 2011.

### NCBI Discovery Workshops September 27-28 at NLM: Seats still available

NCBI will present a two-day workshop September 27 and 28, 2011, on the NIH campus in Bethesda, Maryland. The course is free and is open to anyone interested in NCBI resources. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. These workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The [Discovery Workshops page](#) has more details and a link to register for the course.

### Feature Highlight Now Available in Sequence Databases

A new way to highlight annotated sequence features, named sites, and regions listed in the FEATURES table is now available in the Entrez sequence databases. The new tool is helpful in visualizing the extent and location for such important features as genes, coding regions, exons, and mRNAs in nucleotide sequences and conserved domains, modification sites, and interaction sites in protein sequences.

#### Activating the Feature Highlight

Figure 1 shows an example of the new feature highlight for an [exon](#) in the MAOA gene ([NG\\_008957](#)). Clicking on any of the hyperlinked items in the left-hand column of the FEATURES table section of a sequence record displayed in the Entrez Nucleotide or Protein services highlights the corresponding region on the sequence. Single segment features -- for example, an exon (Figure 1), or multiple segment features -- for example, mRNA alignments on genomic DNA (Figure 2), can be highlighted. The highlighted segments are displayed with white residue letters and a brown background. The Highlight Bar also opens at the bottom of the page that provides additional information and controls.

#### The Feature Highlight Bar

The Feature Highlight Bar provides details about the highlighted region, controls for navigating additional features on the record, and has links to display the highlighted regions as separate sequences for downloading and further analysis. The Details box that is open by default on the Highlight Bar shows the detailed annotation from the FEATURES table for the now highlighted region. The Details box can be collapsed if desired by clicking on the Details link. Clicking the link again re-opens the box. The number of highlighted segments is shown at the right of the sequence accession in the Highlight Bar. In the example in Figure 2, opposite strand features are indicated with the notation "minus strand" to the right of the number of segments on the Highlight Bar.

#### Navigating Using the Feature Highlight Bar

If there is more than one feature of the same type, the navigational arrows on the bar allow jumping to the next, previous, first, and last instances of that feature. The Feature pull-down list at the left-hand side of the bar allows selecting other available feature types. The highlight moves to the next available instance of the selected feature type. The Feature link returns the display to the corresponding position in the FEATURES table of the record.

FEATURES	Location/Qualifiers
<a href="#">source</a>	1..97660 /organism="Homo sapiens" /mol_type="genomic DNA" /db_xref="taxon:9606" /chromosome="X" /map="Xp11.3"
<a href="#">STS</a>	3835..4158 /standard_name="PM130047P4" /db_xref="UniSTS:270611"
<a href="#">gene</a>	5001..95660 /gene="MAOA" /note="monoamine oxidase A" /db_xref="GeneID:4128" /db_xref="HGNC:6833" /db_xref="MIM:309850"
<a href="#">mRNA</a>	join(5001..5254,32353..32447,42130..42267,60711..60815, 61544..61635,77012..77153,80080..80229,80533..80692, 81538..81634,85066..85119,89520..89577,90789..90886, 92633..92744,92948..93010,93206..95660) /gene="MAOA" /product="monoamine oxidase A" /transcript_id="NM_000240.2" /db_xref="GI:33469954" /db_xref="GeneID:4128" /db_xref="HGNC:6833" /db_xref="MIM:309850"
<a href="#">exon</a>	5001..5254 /gene="MAOA" /inference="alignment:Splice:1.39.8" /number=1
<a href="#">CDS</a>	join(5182..5254,32353..32447,42130..42267,60711..60815, 61544..61635,77012..77153,80080..80229,80533..80692, 81538..81634,85066..85119,89520..89577,90789..90886, 92633..92744,92948..93010,93206..93352)

```

4801 caggcgteta cccccacctc agtgccctgac actccgcggg gttcaataca agaacctcct
4861 gcaccocagta atccttttcca gctgccgaca caaggacatt ctaaacctaa taactctcgc
4921 cgagtgtcag tacaagggtc gcgcccgctc tcagtgccca gctccccccg ggtatcagct
4981 gaaacatcag ctccgccctt gggcgctccc ggagtatcag caaaagggtt cgccccgccc
5041 acagtgcccg gtccccccg ggtatcaaaa gaaggatcgg ctccgccccc gggctccccg
5101 ggggagtga tagaagggtc cttcccacc cttgccgtcc ccaactcctgt gcctacgacc
5161 caggagcgtg tcagccaag catggagaat caagagaagg cgagtatcgc gggccacatg
5221 ttcgacgtag tcgtgatcgg aggtggcatt tcaggtcagt gtggaccgtt
5281 gggggacctt gggcagtgag gggtagggga acctacagta gctcttgtg
5341 tcctcatgc atgcgagagt gtagttagc catggcttg ccccatatc
5401 gagtgggggt tgtgccagtt ttgctggtgg tgtgactgg ggagggcaga
5461 tactactact actattaat actaatattt aattagctct tgcgtgca
5521 cactttacgt ggattttctc agtctcaac agtctctgta ggtgggaac
5581 cactttttaa cccccccgca actgagctat gggacttga actgactat

```

5001..5254  
/gene="MAOA"  
/inference="alignment:Splice:1.39.8"  
/number=1

[exon](#) Feature 1 of 15 NG\_008957: 1 segment [Details](#) [FASTA](#) [GenBank](#) [Help](#)

- [CDS](#)
- [exon](#)
- [gene](#)
- [mRNA](#)
- [STS](#)

Figure 1. Feature highlighting shown for an exon feature on the NCBI RefSeq Gene record for Monoamine Oxygenase A (NG\_008957). Clicking on the exon link in the left-hand column of the FEATURES table activates highlight and opens the Highlight Feature Bar at the bottom of the page. Other feature types can be highlighted by selecting them from the Feature pull-down list. Clicking the Feature link returns to the FEATURES table of the record. The number of features of the selected type is shown – 15 in the case of exon features for NG\_008957. Clicking the navigational arrows allows jumping to the next or previous feature of a given type. The details box, which may be closed if desired, re-states the range and qualifiers for the highlighted feature. The FASTA and GenBank links display the highlighted region as a separate view available for copying, downloading, or submitting for further analysis.

## Displaying Highlighted Regions as Separate Sequences

The FASTA and GenBank links on the right-hand side of the bar present the highlighted sub-sequence in these formats in the Nucleotide or Protein Entrez system and provide a simple means to display and download the corresponding sequence or to forward it to the available analysis tools: BLAST, Primer-BLAST, Find in this Sequence, and Identify Conserved Domains (protein only). As shown in Figure 2, the sequence displayed in

The screenshot shows the NCBI Sequence Viewer interface. The main window is split into two panels. The top panel (Back panel) displays a DNA sequence with several segments highlighted in red. The bottom panel (Front panel) shows the FASTA format sequence for the selected region, which is the complementary strand of the highlighted mRNA feature. The FASTA sequence is as follows:

```

>(gi|18121563:c26962-26820, c25342-25247, c24184-24134, c23582-23481,
c21044-20680) Human DNA sequence from clone RP13-377G1 on chromosome Xp11.22-11.3
Contains the AKAP4 gene for A kinase (PRKA) anchor protein 4 and the 5' end of the
CCNB3 gene for cyclin B3, complete sequence
AGTCTGGTCCAAACAGCTGACAGGGGTGGCAGCCAACTGCAGGTGCCCAAGAAGTTCAGTTC
CATCTAAAGGGGGCACATCCCTTCTGGGTGTCACGTTTCAGCCAAACATCTAAAAGAAGTTCATCATC
AAGATGTCTGATATTTGACTGGTTCAGCAGCCACAGGGGTGTGCAAGGTAGATCTTACAACCCAG
AAGGACACCAAGATCAGGACCGAAAGTATATGCTTTGTCGATGTCTCCACCTCAATGTAGAAGATAA
AGATTACAAGGATGCTGCTAGTCCAGCTCAGAAGGCAACTTAAACCTGGGAAGTCTGGAAGAAAAAGAG
ATTATCGTATCAAGGACACTGAGAAGAAAGACCAGCTAAGTCTTCTTTTATAGACAGAGGGAT
CTGTATGCCCTTTCAAACAGCTCCCTCTGATCCTGTAAGTCTCCTCAACTGGCTTCTCAGTATCTCCA
GAAGATGCCCTGGGTTTCCAACATGCACTGAGCCCTCAACCTGTAAGTAAACATAAAGTAGGAGAC
ACAGAGGGCGAATATCACAGAGCATCTCTGAGAAGTCTACAGTGTCTATCCGATCAAGTGAACATAG
ATTATTTGATGAACAGACCTCAAACCTACGCTTAGAAATGACAGCAGCTAAAACACCAACATAATCA
AAGTCTTCAGTCTCCAGCCAAACCTCTAGCACTCAGAGAGCAGTATTTCCTCC
    
```

The interface also includes navigation controls at the bottom, such as 'CDS', 'Feature', '6 of 11', 'AL663119 : 5 segments (minus strand)', 'Details', and 'Display: FASTA GenBank Help'.

Figure 2. Highlighting and displaying a multi-segmented mRNA feature on the minus strand of a BAC clone sequence (AL663119). Back panel. Highlighting a splice variant of AKAP4 gene that is the sixth mRNA feature on the record. There are 5 highlighted segments on the minus strand of the record. Clicking the FASTA link displays the corresponding region shown in the Front panel. The complementary strand is shown automatically giving the same sequence as the mRNA.

FASTA format is the appropriate strand for the feature, in this case the complementary or minus strand of the record.

### Summary

The new ability to highlight features in sequence records complements the Find-in-Sequence tool described in the September 2010 NCBI News and adds powerful new visualization and search options to the NCBI sequence database.

### New videos on NCBI's YouTube channel

Three new videos are available on NCBI's YouTube Channel. Two instructional videos show how to display the six-frame translations of a DNA sequence in the graphical sequence viewer on the web (Sequence Viewer: Six Frame Translations) and in the standalone Genome Workbench annotation and analysis tool (Genome Workbench: Six Frame Translations).

The PMC 10th Anniversary video, celebrating the ten years of the PubMed Central online public access full-text database, now joins the two other anniversary celebrations: the NCBI 20th anniversary video and the collection of talks from the GenBank 25th Anniversary.



## Updated Genome Workbench (v2.4.0)

An update to NCBI's [Genome Workbench](#) (v2.4.0) is now available. Genome Workbench is a standalone sequence viewer, annotation and analysis platform. The new version has many new features, improvements, and a few bug fixes that are described in the [release notes](#). The latest Genome Workbench pre-compile packages and source code are available from the [download page](#).

## Conserved Domain Database updated (v2.31)

Version 2.31 of the [Conserved Domain Database](#) (CDD) is now available. The new release contains 292 new or updated NCBI-curated domain models and now includes domains from [SMART](#) version 6. The CDD data are searchable in the Entrez and [BLAST](#) services at the NCBI Website and are available for download from the [FTP site](#).

## Microbial Genomes Update

One hundred fifty-five finished microbial (archaeal and bacterial) genomes were released during June, July and August 2011. The original sequence data files submitted to International Sequence Database Collaboration (INSDC) are available in the [Bacteria directory](#) in the genomes area of the GenBank FTP site. RefSeq provisional versions were released for a selected set of 86 of the complete INSDC microbial genomes during the same period. These are available from the [/genomes/Bacteria](#) directory on the FTP site.

In addition, 317 microbial whole genome-shotgun (WGS) sequencing projects were added to the INSDC during this period. The original submitted files are available in the [Bacteria\\_DRAFT](#) directory in the GenBank genomes

area. RefSeq provisional versions of 86 WGS microbial projects were released in the [/genomes/Bacteria\\_DRAFT](#) area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

## GenBank News

GenBank release 185 is available through the NCBI web and [FTP](#) sites. The current release incorporates data available as of August 14, 2011 and, with the whole-genome shotgun portion, contains 338,987,064,933 bases from 207,281,745 sequence records. [Release notes](#) describe the current state of data and upcoming changes.

## RefSeq News

RefSeq Release 48 is available through Entrez, BLAST, and the [RefSeq FTP site](#). The current release includes 18.2 million Reference Sequence records from 12,235 different organisms. The RefSeq [release notes](#) provide more detailed information.

## NCBI will no longer archive new sequencing data from The Cancer Genome Atlas (TCGA)

NCBI will no longer archive new sequencing data from [The Cancer Genome Atlas \(TCGA\)](#) in the Sequence Read Archive. The release of dbGaP's TCGA study phs000178 version 5 on August 15, 2011 constitutes the up-to-date and final compendium of files available at NCBI.

## The Growth of PubChem

PubChem now contains over 30,000,000 chemically unique compounds with over 500,000 bioassays. Research labs, institutes, organizations, and companies have submitted over 85,000,000 substances over the past seven years. Steady growth in content and usage is expected to continue. The [PubChem news page](#) has more details.

## New Simple Object Access Protocol (SOAP)-based BLAST service

A new Simple Object Access Protocol (SOAP)-based service is available. The SOAP interface can be used to develop applications that interact with the NCBI BLAST web service to submit searches and retrieve results. [Documentation](#) and links to the [Web Service Definition Language \(WSDL\)](#) and sample clients are available on the NCBI Bookshelf.

## NCBI at the ICHG/ASHG Meeting in Montreal: Workshop on Medical Genetics

NCBI scientists will present a special workshop at the combined [International Congress of Human Genetics and American Society for Human Genetics meeting](#) at Montreal Convention Center on October 12th at 12:30 P.M. The workshop entitled, “[Genetics and Medicine at the National Center for Biotechnology Information \(NCBI\)](#)”, will provide information on genome-scale resources for medical genetic genetics available at the NCBI including finding and downloading data, analysis, and management of data sets. NCBI will also staff an exhibit booth at the meeting.

## Announce Lists and RSS Feeds

Seventeen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the [Announcement List summary page](#). Subscribe to the [NCBI Announce list](#) to receive updates on the NCBI News.

Twenty-one [RSS feeds](#) are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's [Facebook](#) page and [Twitter feed](#) also provide updates on NCBI resources.

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, June 2011

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: May 13, 2011; Updated: June 30, 2011.

### Featured Resource: Re-designed PopSet

NCBI's PopSet database of related sequences and alignments from phylogenetic, population, mutation, and ecosystem studies has been completely redesigned and now features an embedded graphical alignment and better integration of related data from other PopSets and other Entrez databases. The new pages also include on-the-fly analysis with BLAST and Tree View.

### The PopSet Record View

The PopSet record view is now fully integrated with the updated Entrez system and can be addressed simply with the PopSet database name and the identifier as shown below.

<http://www.ncbi.nlm.nih.gov/popset/298351991>

The record display shown in Figure 1 consists of up to three sections: the study details showing the article reporting the current set; a list of the sequence records in the set; and, when available, the submitted alignment displayed in the embedded Graphical Sequence Viewer (GSV), now also appearing in Entrez Gene and SNP record views. The PopSet embedded alignment view shows the alignment portion of the full GSV display of the master or top sequence in the multiple-alignment. Clicking on the “Open full-view” link opens the GSV nucleotide view of the top sequence showing the detailed alignment tracks.

As in the other Entrez databases, the “Display Setting” menu controls the format of the records displayed; the “Send to” menu manages saving data, shown in Figure 2. Display options are similar to those available for the Nucleotide database and include the standard sequence formats such as FASTA and GenBank. The sequence record formats are presented within the PopSet display rather than by linking to the sequence database.

The “Send to” menu can send data to the Entrez clipboard, Collections in a My NCBI account, or to a file on the local computer. The file saving format options include the standard sequence formats, popular multiple alignment formats – FASTA plus gap, CLUSTAL, Nexus, and Phylip – are also available making the alignments easy to use for local analysis.

### Improved PopSet-PopSet Connections

PopSet now features more explicit connections between PopSets associated with the same study. As always, following the link from a PubMed record retrieves all PopSets for molecules used in the study. In the previous version of PopSet, however, it was not easy to navigate from one PopSet to others that are part of the same study. The PopSet-PopSet link now provides rapid access to related PopSets. The related PopSets also are listed “Other data sets from this study” in the right-hand Discovery Column of the full record. Figure 3 shows the items in the Discovery Column and the corresponding related data in PopSet and PubMed.

### Analysis Tools: BLAST and Tree View

For PopSets with fewer than 100 sequences, analysis tools are available at the top of the right-hand Discovery Column (Figure 3). These allow generating or re-generating an alignment with BLAST or, if a submitted alignment is present, displaying a distance tree (Tree View) based on the alignment. Figure 4 shows the results of

**Carnivora apolipoprotein B (APOB) gene, partial cds.**  
 PopSet: 298351991  
 GenBank FASTA

**Study Details**  
**Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences.**  
 Eizirik, E., Murphy, W.J., Koepfli, K.P., Johnson, W.E., Dragoo, J.W., Wayne, R.K. and O'Brien, S.J. (2010) Mol. Phylogenet. Evol. 56:(1)49-63  
 PMID: 20138220 [Citation](#)

**Sequences in this data set**

Description	Mar...	Seq. S...	First	Alignment	Last	Seq. End	Seq. L...
GU930905.1 Ursus americanus ap...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930904.1 Ailuropoda melanole...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930903.1 Ailurus fulgens apoli...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930902.1 Odobenus rosmarus			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930901.1 Mirounga angustirost...			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930900.1 Mydaus marchei apo...			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930899.1 Conepatus leuconot...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930898.1 Spilogale putorius ap...			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930897.1 Mephitis mephitis ap...			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930896.1 Urocyon cinereoarge...			252	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	280	932	932
GU930895.1 Nyctereutes procyon...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930894.1 Genetta genetta apo...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930893.1 Civettictis civetta ap...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930892.1 Fossa fossana apoli...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930891.1 Rhynchogale melleri...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930890.1 Ichneumia albicauda...			239	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	267	519	919
GU930889.1 Helogale parvula ap...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930888.1 Suricata suricatta ap...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930887.1 Panthera onca apoli...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930886.1 Leopardus pardalis a...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930885.1 Lynx lynx apolipopro...			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933
GU930884.1 Felis catus apolipoprotein B (APOB) gene, partial cds			253	A R A T T C C C T T T T T A T G T A R A A G A T T T C C A G G	281	933	933

**Alignment**  
 253 - 281 (29 bases shown)

Open Full View

PubMed  
Taxonomy

Figure 1. The new PopSet record display showing the Study Details, the Sequences list, and the submitted Alignment for a phylogenetic set (PopSet: 298351991) of apolipoprotein B sequences from mammals. The Study Details shows the title of the study with a link to the citation in PubMed and the full-text in PMC (not shown) when available. The list of sequences provides the sequence titles and a link to each record in the Nucleotide database. The submitted Alignment is displayed in the embedded Graphical Sequence Viewer.

the BLAST and Tree View tools for a phylogenetic study set that has a submitted alignment. The link to run BLAST is especially useful in cases where the set does not contain a submitted alignment, for example PopSet: 338197537. In such cases the Tree View can be invoked after running the BLAST alignment through the “Distance tree of results” link on the BLAST output.

## Summary

The NCBI PopSet database has been fully updated to the new Entrez system and includes new record displays and better access to related information. These improvements will make the growing collection of PopSets easier to access, download, and analyze.

## New My NCBI Interface

My NCBI now has customizable modules making it even easier to manage your NCBI preferences, collections, bibliographies, saved searches, and more. A [video](#) highlighting the new homepage and features is on the NCBI YouTube Channel.

**Display Settings:**  PopSet **Send to:**

**Format**

- Summary
- PopSet
- GenBank
- FASTA
- FASTA (text)
- ASN.1
- Revision History
- Accession List
- GI List

**Choose Destination**

- File
- Clipboard
- Collections
- Analysis Tool

Download 1 items.

**Format**

- CLUSTAL**
- PopSet
- GenBank
- FASTA
- ASN.1
- XML
- INSDSeq.XML
- TinySeq.XML
- Feature Table
- FASTA plus Gap
- CLUSTAL
- Nexus
- Phylip

**Sequences in this data set**

Accession	Species
<a href="#">GU930905.1</a>	Ursus americanus
<a href="#">GU930904.1</a>	Alluropoda melanocephala
<a href="#">GU930903.1</a>	Alluropoda fulgens apoda
<a href="#">GU930902.1</a>	Odobenus rosmarus
<a href="#">GU930901.1</a>	Mirounga angustirostris
<a href="#">GU930900.1</a>	Mydaus marchei

```

CLUSTAL W (1.83) multiple sequence alignment
gi|298351991      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298351993      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298351995      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298351997      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298351999      -GCTGTAACAATGAAAACAGCATTGAGGCCATGTAGGAATAAATGGAGAAGCCAATCT
gi|298352001      -GCTGTAACAATGAAAACAGCATTGAGGCCATGTAGGAATAAATGGAGAAGCCAATCT
gi|298352003      -GCTGTAACAATGAAAACAGCATTGAGGCCATGTAGGAATAAATGGAGAAGCCAATCT
gi|298352005      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352007      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352009      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352011      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352013      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352015      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352017      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352019      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352021      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352023      -----AGAGCAGTGTGGGGCCACATAGGAATAAATGGAGAAGCCAATCT
gi|298352025      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352027      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352029      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352031      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
gi|298352033      TGCTGGAACAATGAGAACAGCATTGAGGCCACGTAGGAATAAATGGAGAAGCCAATCT
    
```

Figure 2. “Display Settings” (upper left) and “Send to” (upper right) menus for the new PopSet record display. PopSet retains its own separate sequence record formats (FASTA, GenBank, ASN.1). These are displayed within the PopSet database rather than in the sequence databases. Download options for PopSets with alignments include popular multiple-alignment formats such as CLUSTAL (lower panel).

The screenshot displays the NCBI Discovery column interface. On the left, the 'Analyze this data set' section includes 'Run BLAST alignment' and 'Tree View'. Below it, 'Article reporting this data set' provides a PubMed citation: 'Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear genes [Mol Phylogenet Evol. 2010]'. The 'Other data sets from this study' section lists related PopSet records such as 'Laurasiatheria RASA2 gene, partial sequence.' and 'Carnivora recombination activation protein 2 (RAG2) gene, partial cds.'. The 'Related information' section includes links for 'PopSet', 'Nucleotide', 'Protein', 'PubMed', and 'Taxonomy'.

The center image shows a list of 13 PopSet records, each with a checkbox, a link to the gene, and details about the phylogenetic study (e.g., '1. Laurasiatheria RASA2 gene, partial sequence. phylogenetic study, 34 aligned sequences').

The right-hand image is a preview of a PubMed article titled 'Pattern and timing of diversification of the mammalian order Carnivora inferred from multiple nuclear gene sequences.' by Eizirik E, Murphy WJ, Koepfli KP, Johnson WE, Dragoo JW, Wayne RK, O'Brien SJ. The abstract discusses the phylogenetic relationships among carnivoran families and the construction of a molecular timescale for the evolution of this mammalian order.

Figure 3. The Discovery column (left-hand image) for a PopSet record showing related PopSets (center image) and result of following the link to PubMed (right-hand image). The Discovery Column has Analysis Tools, a database ad for PubMed showing the article title, an ad for related PopSets (“Other datasets from this study”), and the traditional Links menu – now shown as Related information. Following the “See all ...” link in the related PopSets ad or the PopSet link produces the results shown in the center image, PopSet for other sequence targets reported in the linked PubMed citation. Following the linked article title in the PubMed ad or the PubMed link in the Related Information section retrieves the citation.

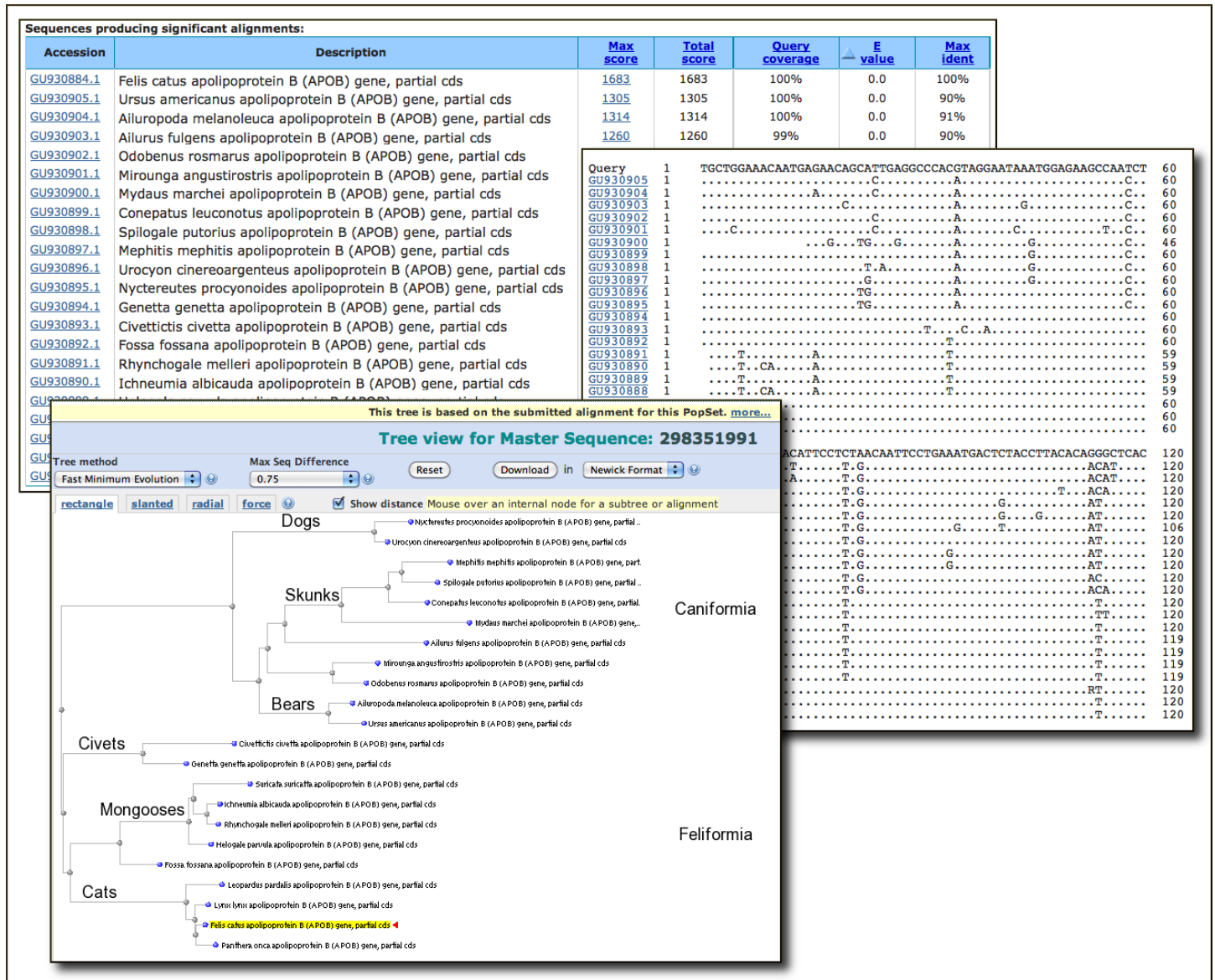


Figure 4. Results of Analysis Tools links “Run BLAST” and “Tree View” from PopSet: 298351991. The BLAST search is implemented using the first sequence as a query against the remaining members of the PopSet. The results are presented in BLAST flat query anchored format with identities shown as dots (upper images). The Tree View link uses submitted alignment in the BLAST Tree View service (lower image). In this case the result shows a molecular phylogeny for the sequences in the set, supporting major groups of the mammalian order carnivora. Names of groups were added manually and are not produced by the software.

The screenshot displays the My NCBI user interface. At the top left is the My NCBI logo. The main area is divided into several panels:

- Search NCBI databases:** A search box with "PubMed" selected and a "Search" button. A hint below states: "Hint: clicking the 'Search' button without any terms listed in the search box will transport you to that database's homepage."
- My Bibliography:** Shows "Your bibliography contains 11 items." and "Your bibliography is private." It lists "Most recently added citations:" with a list of articles including Cooper PS, Lipshultz D, etc., and Sayers EW, Barrett T, etc.
- Recent Activity:** A table showing search activity:
 

Time	Database	Type	Search Query
5:15 PM	Protein	search	zebrafish[Organism] AND
5:12 PM	PubMed	search	koonin[au] AND pubmed
- Saved Searches:** A table listing saved searches:
 

Search Name	What's New	Last Searched
popset_pubmed[Filter] AND alignment_present[Filter] AND popset_popset[...]	23	2 days ago
blast AND sequence_analysis[Mesh] AND computational_biology[mesh]	0	2 days ago
- Collections:** A table listing collections:
 

Collection Name	Items	Privacy	Type
My Bibliography	11	Private	Standard
Other Citations	63	Private	Standard

Overlaid on the bottom right is a YouTube video player titled "My NCBI: New Home Page". The video content shows the new My NCBI home page with the text "My NCBI Home Page" and "Tour the new My NCBI home page". The video player shows 48 videos, a subscribe button, and 3,017 likes.

## Transcriptome Shotgun Assembly (TSA) Database Available for BLAST

The Transcriptome Shotgun Assembly (TSA) BLAST database is now available from the database list for the main NCBI BLAST services. TSA is an archive of computationally assembled mRNA sequences from primary data such as Expressed Sequence Tag (EST) and raw sequence reads. These sequences were previously a part of the BLAST nucleotide nr (nt) database but have been moved because of their increasing numbers and special characteristics. The [TSA page](#) has more information on the nature and sources of TSA sequences.

## New Attributes for Human Variants in dbSNP

New attributes related to allele origin, clinical significance, and population genetics are available in dbSNP. These attributes allow searching and filtering of human variations for the characteristics listed below.

1. **Allele Origin:** Summarizes the reported origin(s) of the variant allele asserted by each submitter for the submitted SNP (ss). Current values are germline, somatic, and unknown. Additional attributes will be added in the future including not-tested, tested-inconclusive, and other.
2. **Clinical significance:** Reports potential health impact of the allele. Possible values:
  - unknown
  - untested
  - non-pathogenic
  - probable-non-pathogenic
  - probable-pathogenic
  - pathogenic
  - drug response
  - histocompatibility
  - other
3. **Global minor allele frequency (MAF):** Shows the minor allele frequency for each RefSNP included in a default global population. Since this is being provided to distinguish common polymorphism from rare variants, the MAF is actually the second most frequent allele value. For example, if there are 3 alleles with frequencies of 0.50, 0.49, and 0.01, the MAF will be reported as 0.49. The current default global population is 1000Genome phase 1 genotype data from 629 worldwide individuals, released in the 08-04-2010 dataset.
4. **Suspect:** Variation suspected to be false positive due to various artifacts.

These new attributes are shown in the images below for the rs429358 [Cluster Report](#) and [Document Summary](#).

Reference SNP(refSNP) Cluster Report: rs429358		** With probable-pathogenic allele <a href="#">[detail]</a> **	
RefSNP	Allele		
Organism: human ( <a href="#">Homo sapiens</a> )	<b>Variation Class:</b> SNP: single nucleotide variation		
Molecule Type: Genomic	<b>RefSNP Alleles:</b> C/T		
Created/Updated in build: 80/132	<b>Allele Origin:</b> T:Germline C:Germline		
Map to Genome Build: <a href="#">37.1</a>	<b>Ancestral Allele:</b> C		
<b>Validation Status:</b>	<b>Clinical Source:</b> VarView  OMIM		
<b>Citation:</b> <a href="#">PubMed</a>	<b>Clinical Significance:</b> With probable-pathogenic allele <a href="#">[detail]</a>		
	<b>MAF/MinorAlleleCount:</b> C=0.076/96		
	<b>MAF Source:</b> 1000 Genomes		

[rs429358](#) [*Homo sapiens*]

GGCTGGGCGCGGACATGGAGGACGTG [C/T] GCGGCCCGCTGGTGCAGTACCGCGG

19 MapView VarView PubMed GeneView SeqView Protein 3D OMIM

Allele Origin: T-Germline C-Germline  
MAF/MinorAlleleCount: C=0.0763/96  
Clinical Significance: probable-pathogenic

Please see the [online help](#) for more information and more examples.

## Updated BLAST Genome Search Pages

The genome-specific BLAST pages linked to the top of the NCBI [BLAST homepage](#) and accessible from the [Map Viewer homepage](#) now use the standard BLAST form with genome specific databases. This change eliminates the older separate interface and provides the full functionality of the standard BLAST interface including the ability to adjust all algorithm parameters, the capability to edit and re-submit searches, to sort descriptions and alignments in the output, and the full range of formatting and downloading options.

## NLM Contest: Show off your Apps! Invitation to Submit Applications that Work with NLM Biomedical Data

The National Library of Medicine (NLM) is challenging people to create innovative software applications that use the Library's vast collection of biomedical data. The purpose of this contest is to foster the development of innovative software applications that will further NLM's mission of aiding the dissemination and exchange of scientific and other information pertinent to medicine and public health. Winners will be recognized at an awards ceremony at the National Library of Medicine and links to their application will be publicized on NLM Web sites. The NLM "Show Off Your Apps" Challenge is open to individuals over the age of 18, teams of individuals, and organizations in the United States. Eligible software applications must make use of NLM's vast collection of biomedical data including downloadable data sets, application programming interfaces, and/or software tools. The [challenge.gov website](#) has detailed information on the contest.

Applications should be submitted to the [challenge.gov](#) site by August 31, 2011.

## New Videos on NCBI's YouTube Channel

In addition to the video introducing the new My NCBI mentioned above, four other instructional videos recently became available on NCBI's YouTube channel:

- [Saving search results in My NCBI Collections](#)
- [Loading sequences and adjusting graphical views in NCBI's Genome Workbench](#)
- [Requesting permission to use controlled access data in dbGaP](#)
- [Using BLAST from NCBI's graphical sequence viewer](#)

## The Sequence Read and Trace Archive Databases to Continue

Recently, NCBI announced that the Sequence Read Archive (SRA) and Trace Archive repositories would be discontinued due to budget constraints ([NCBI News, March 2011](#)). However, with the commitment of interim funding and a plan for future support developed in collaboration with other NIH Institutes and NIH grantees, NCBI will now continue to accept submissions and maintain the Sequence Read Archive (SRA) and Trace Archive repositories for high-throughput sequence data. These repositories will now focus on high-throughput data that support other kinds of data at the NCBI including:

- RNA-Seq, ChIP-Seq, and epigenomic data that are submitted to GEO
- Genomic and Transcriptomic assemblies that are submitted to GenBank
- 16S ribosomal RNA data associated with metagenomics that are submitted to GenBank

The [full announcement](#) on the NCBI site has more details.



## BLAST 2.2.25+ Release and New Set-up Instructions

Stand-alone BLAST+ (v2.2.25) is now on the [FTP site](#). Improvements include hard-masking of databases, faster formatting of databases using makeblastdb, XML and best hit options for Blast2Sequences, multiple query psiblast, selection of any master sequence in psiblast with multiple alignment input, and query and subject length in tabular output. The [BLAST News](#) has more detailed information on changes. Detailed set-up instructions for standalone BLAST are now a part of the [BLAST User Manual](#) on the NCBI Bookshelf.

## Microbial Genomes Update

One hundred thirty five finished microbial genomes were released between March 1 and May 31, 2011. The original sequence data files submitted to GenBank/EMBL/DDBJ are available in the [Bacteria](#) directory in the /genbank/genomes area of the GenBank FTP site. One hundred twelve RefSeq provisional versions were made from a selected set of finished genomes. These are available from the [/genomes/Bacteria](#) directory on the FTP site.

In addition, 305 microbial whole genome shotgun-sequencing projects were added to GenBank during this period. The original submitted files are available in the [Bacteria\\_DRAFT](#) directory in the GenBank genomes area. RefSeq provisional versions of 84 of these projects are available in the [/genomes/Bacteria\\_DRAFT](#) area of the FTP site.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated Entrez search and retrieval system and the BLAST sequence similarity search service.

## RefSeq News

RefSeq Release 47 is available through Entrez, BLAST, and the [RefSeq FTP site](#). The current release includes 17.6 million sequence records from 12,000 organisms. [Release notes](#) provide more detailed information.

## GenBank News

GenBank release 183 is available through the NCBI web and [FTP sites](#). The current release incorporates data available as of Apr 11, 2011 and, with the whole-genome shotgun portion, contains 317,952,894,329 bases from 198,156,212 sequence records. [Release notes](#) describe the current state of data and upcoming changes.

## NCBI Discovery Workshops at Washington University: July 26-27, 2011

NCBI will present a two-day workshop on July 26 and 27th, at Washington University in St. Louis, Missouri. The course is free and is open to anyone interested in NCBI resources. The workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. The [Discovery Workshops page](#) has more information.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the [Announcement List summary page](#). Subscribe to the [NCBI Announce list](#) to receive updates on the NCBI News.

Twelve [RSS feeds](#) are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

NCBI's [Facebook](#) page and [Twitter](#) feed also provide updates on NCBI resources.

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, March 2011

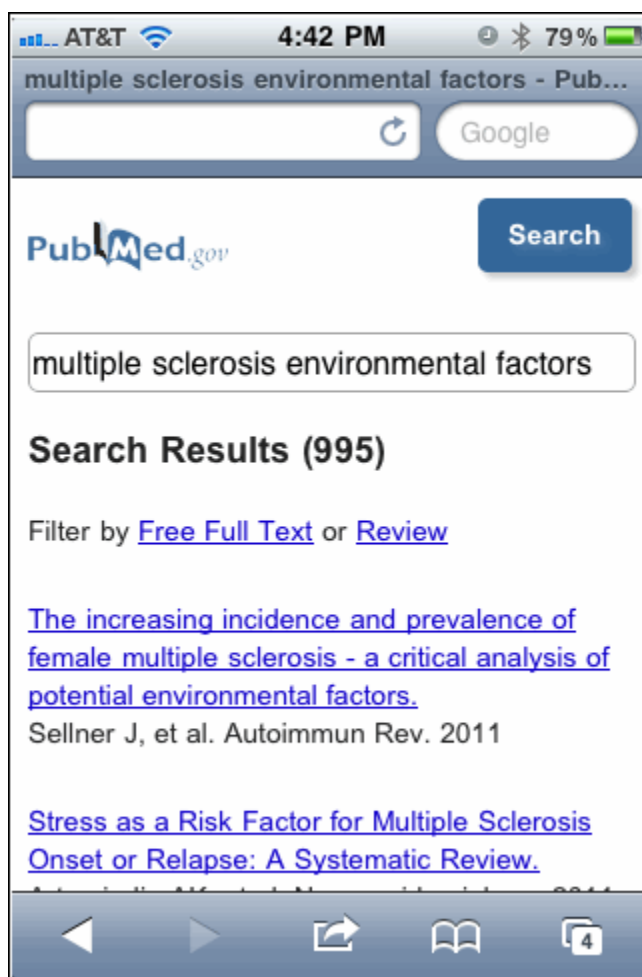
Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: March 18, 2011; Updated: March 18, 2011.

### PubMed Interface for Mobile Devices Now Available

PubMed Mobile Beta is now live. This is a special lightweight web interface that makes PubMed faster to load and easier to use on smart phones and other mobile devices. The PubMed site will eventually automatically detect browsers on mobile devices and will load the new interface. For now, direct access is available at the following address.

<http://www.ncbi.nlm.nih.gov/m/pubmed/>



The recent PubMed Technical Bulletin provides additional information.

### NCBI Bookshelf Updated to the New Entrez Design

The NCBI Bookshelf is now fully updated to the discovery-oriented Entrez design that first appeared in PubMed in 2009. Bookshelf has a [new homepage](#), results displays, [search limits](#), and an [Advanced search page](#). A new

[browser](#) helps to quickly find titles of interest. The [NLM Technical Bulletin](#) explains the new features in more detail.

## New Organism Builds in UniGene

Two new organisms have builds in UniGene: the important plant pathogen *Pythium ultimum*, and the mountain pine beetle (*Dendroctonus ponderosae*), a serious forest pest.

*Pythium ultimum* ([build information](#), [6,663 clusters](#), [FTP](#)) is the cause of damping-off disease in seedlings and rot in stored root vegetables. *Pythium ultimum* is a member of the Oomycetes, a fungi-like group of organisms that also contains the *Phytophthora* causative agents of potato (late) blight.

The mountain pine beetle ([build information](#), [6,783 clusters](#), [FTP](#)) is a bark beetle that is a destructive pest of pine forests in the western United States. Analysis of this beetle's transcriptome should provide important insights into pheromone biosynthesis and potentially specific control mechanisms for these pests.



Root rot caused by *Pythium* (left); the mountain pine beetle (right).

## NCBI YouTube Video Update

A new video on NCBI's YouTube Channel shows how to display Primer-BLAST search results on the Graphical Sequence Viewer (GSV), a feature of the GSV described in the [October 2011 NCBI News](#). This Video joins nine others that are included in the [Tutorials Playlist](#). These videos provide brief instructions on useful features of the NCBI Web services.

**NCBI YouTube Channel**  
NCBINLM's Channel

[Subscribe](#) [All](#) [Uploads](#) [Favorites](#) [Playlists](#)

**Primer-BLAST**

Input PCR template: NC\_020302.3 (Primer design beta globin region (HBB)), and hemoglobin, beta (HBB), and hemoglobin, delta (HBD), and hemoglobin, gamma 1 (HBG1), and hemoglobin, gamma A (HBG1), and hemoglobin, gamma C (HGC), and HbA2 gene on chromosome 11

Range: 27196..29452  
Specificity of primers: primer specificity was not determined as specificity checking option was not selected

**Graphical view of primer pairs**

36,857 - 28,789 (2,888 bases shown, positive strand)

**Detailed primer reports**

Primer pair 1	Sequence (5'>3')	Stand on template	Length	Start	Stop	Tm	GC%
Forward primer	TGGTCTGGCAATTCCTCTGG	Plus	20	28113	28132	59.97	65.00%
Reverse primer	AAGGCCAAGCCCAATCCCGCA	Minus	20	28221	28203	60.03	65.00%
Product length	111						
Primer pair 2	Sequence (5'>3')	Stand on template	Length	Start	Stop	Tm	GC%
Forward primer	AGGGGGCAAGAACTCCGGCA	Plus	20	27618	27638	60.11	65.00%
Reverse primer	CCAGGGCAATCCCGAAGCA	Minus	20	28190	28111	59.97	65.00%
Product length	193						
Primer pair 3	Sequence (5'>3')	Stand on template	Length	Start	Stop	Tm	GC%
Forward primer	CCCTCAAGTGGAGGCTGGC	Plus	20	28074	28093	59.97	70.00%
Reverse primer	AGGCCAAGCCCAATCCCGCA	Minus	20	28221	28201	59.91	65.00%
Product length	147						
Primer pair 4	Sequence (5'>3')	Stand on template	Length	Start	Stop	Tm	GC%
Forward primer	CCCTCAAGTGGAGGCTGGC	Plus	20	28074	28093	59.97	70.00%
Reverse primer	AGGCCAAGCCCAATCCCGCA	Minus	20	28221	28201	59.91	65.00%
Product length	147						

When Primer BLAST returns, we see the results presented graphically.

**Sequence Viewer: Using Primer-BLAST**  
From: NCBINLM | Mar 3, 2011 | 317 views

A quick demonstration of how to initiate Primer-BLAST from within the graphics, or sequence viewer, display in nucleotide records at NCBI.

[View comments, related videos, and more](#)

**Tutorials**  
Short video tutorials about using NCBI resources. [More Info](#)

- Save Searches and Set E-mail Alerts**  
NCBINLM - 16,798 views | 1:45
- Sequence Viewer: Using Primer-BLAST**  
NCBINLM - 317 views | 2:06
- Sequence Viewer: Configure your view**  
NCBINLM - 559 views | 2:41
- Epigenomics: Using the Sample Browser**  
NCBINLM - 517 views | 2:26
- Epigenomics: How to Download Data**  
NCBINLM - 502 views | 1:57
- Epigenomics: How to View Track Data**  
NCBINLM - 608 views | 2:33
- Find in This Sequence**  
NCBINLM - 8,622 views

Links to relevant NCBI YouTube Videos now appear in the Discovery Column in Entrez record views or search results under certain circumstances. Specific video ads appear for only a few weeks at a time. These videos provide brief tutorials on how to use relevant features of the current page and promote new or underused but useful aspects of the service. The first of these video ad campaigns appeared in the sequence databases to introduce the Find-in-Sequence feature on sequence records described in the [September 2010 NCBI News](#). The most recent video ad appeared on PubMed search results to promote using My NCBI to set up custom searches and e-mail alerts. The video portlet that is embedded in the page allows the video to be played in position, helpful for providing instruction within the current view. Alternatively the larger version of the video may be played on YouTube by clicking the “[See larger video at YouTube](#)” link.

Display Settings:  Summary, 200 per page, Sorted by Pub Date Send to:

**Results: 1 to 200 of 6116** << First < Prev Page 1 of 31 Next > Last >>

[Effect of bisphenol-A on the expression of selected genes involved in cell cycle and apoptosis in the OVCAR-3 cell line.](#)  
 1. Ptak A, Wróbel A, Gregoraszczyk EL.  
 Toxicol Lett. 2011 Apr 10;202(1):30-5. Epub 2011 Jan 26.  
 PMID: 21277958 [PubMed - in process]  
[Related citations](#)

[Fate of bisphenol A during treatment with the litter-decomposing fungi Stropharia rugosoannulata and Stropharia coronilla.](#)  
 2. Kabiersch G, Rajasärkkä J, Ullrich R, Tuomela M, Hofrichter M, Virta M, Hatakka A, Steffen K.  
 Chemosphere. 2011 Apr;83(3):226-32. Epub 2011 Feb 3.  
 PMID: 21295326 [PubMed - in process]  
[Related citations](#)

[Removal capacity and pathways of phenolic endocrine disruptors in an estuarine wetland of natural reed bed.](#)  
 3. Yang L, Li Z, Zou L, Gao H.  
 Chemosphere. 2011 Apr;83(3):233-9. Epub 2011 Jan 26.  
 PMID: 21269659 [PubMed - in process]

[Filter your results:](#)  
 All (6116)  
[Free Full Text \(892\)](#)  
[Review \(209\)](#)  
[Manage Filters](#)

**E-mail Alert Video Tutorial**



See larger video at YouTube  
 See all NCBI YouTube video channel videos

Additional video ads will appear in many of the Entrez databases in the future.

## RefSeq News

RefSeq Release 46 is available through Entrez, BLAST, and the [RefSeq FTP site](#). The current release includes nearly 17 million sequence records from 11,734 organisms. The [release notes](#) provide more detailed information.

## GenBank News

GenBank release 182 is available through the NCBI web and [FTP sites](#). The current release incorporates data available as of Feb 15, 2010 and contains 124,277,818,310 bases from 132,015,054 sequence records. [Release notes](#) describe the current state of data and upcoming changes.

## Microbial Genomes Update

Sixty-three finished microbial genomes were released during January and February 2011. The original sequence data files submitted to GenBank/EMBL/DDBJ are available in the [Bacteria directory](#) in the genomes area of the GenBank FTP site. Fifty-nine [RefSeq provisional versions](#) were made from a selected set of finished genomes.

In addition, 184 microbial whole genome shotgun-sequencing projects were added to GenBank during this period. The original submitted files are available in the [Bacteria\\_DRAFT directory](#) in the GenBank genomes area. [RefSeq provisional versions](#) of 55 of these projects are also available.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated [Entrez](#) search and retrieval system.

## Mouse Genome Annotation Release (build 37.2) and Updated Mouse Consensus Coding Sequence (CCDS) Data

A new mouse genome annotation (v.37.2) is now available in [Entrez](#), the [Map Viewer](#), BLAST, and on the [FTP site](#). This includes the reference C57BL/6 strain genome (MGSCv37) and Celera assemblies as well as alternate

regional haplotype assemblies (ALT\_LOCI) designated for 12 [other strains](#). Figure 1 highlights mouse genome and the alternate locus for the beta globin region. The mouse [Consensus Coding Sequences \(CCDS\)](#) have also been updated to include new coding sequences that are consistently annotated by NCBI, the [European Bioinformatics Institute](#), [Wellcome Trust Sanger Institute](#), and the [University of California, Santa Cruz](#). This update adds of 4,561 new CCDS records and 2,685 Genes into the mouse CCDS set. Mouse build 37.2 includes a total of 22,187 CCDS records that correspond to 19,509 GeneIDs. The [statistics](#) report has more details. The mouse CCDS update can be downloaded from the [NCBI FTP site](#).

## HomoloGene Release 65 Now Available

The new HomoloGene release 65 includes updated annotations for human, zebrafish, fruitfly, *C. elegans*, and *Arabidopsis thaliana*. Clusters also feature an updated related UniGene section as shown in Figure 2 that groups linked UniGene records by organism. The [HomoloGene homepage](#) has the latest statistics.

## Genome Workbench Version 2.2.2 Release

Genome Workbench is NCBI's standalone genome analysis and annotation tool. The new Genome Workbench release has several important enhancements including: integration of WindowMasker, improved displays of tree views, better network support for clients behind firewalls, improved support for international symbols, new status reporting for background work, and updated documentation. The [Genome Workbench homepage](#) has more information and links to extensive help documentation and access to download the compiled program for several versions of Windows, Mac, and Linux operating systems as well as the C++ source code for building Genome Workbench on all platforms.

## NCBI Responds to a Report of Contamination in the Sequence Databases

A report by Longo *et al.* in the on-line journal *PLoS One* provides examples of cross-species contamination in DNA sequence data. The reported contamination is almost entirely in high-throughput, low-coverage sequences and is removed by filters in the NCBI genome pipeline for finished genomes. The [full response](#) is available at the NCBI website.

## NCBI Discontinues the Short Read Archive, Trace Archive, and Peptidome

Due to budget constraints, NCBI has discontinued the Peptidome repository for protein/peptide mass spectroscopy data and will be discontinuing the Sequence Read Archive (SRA) and Trace Archive repositories for high-throughput sequence data. Data for these projects will continue to be available on the NCBI FTP site for the foreseeable future. The [full announcement](#) with additional details and information on the future disposition of data associated with these projects is available on the NCBI website.

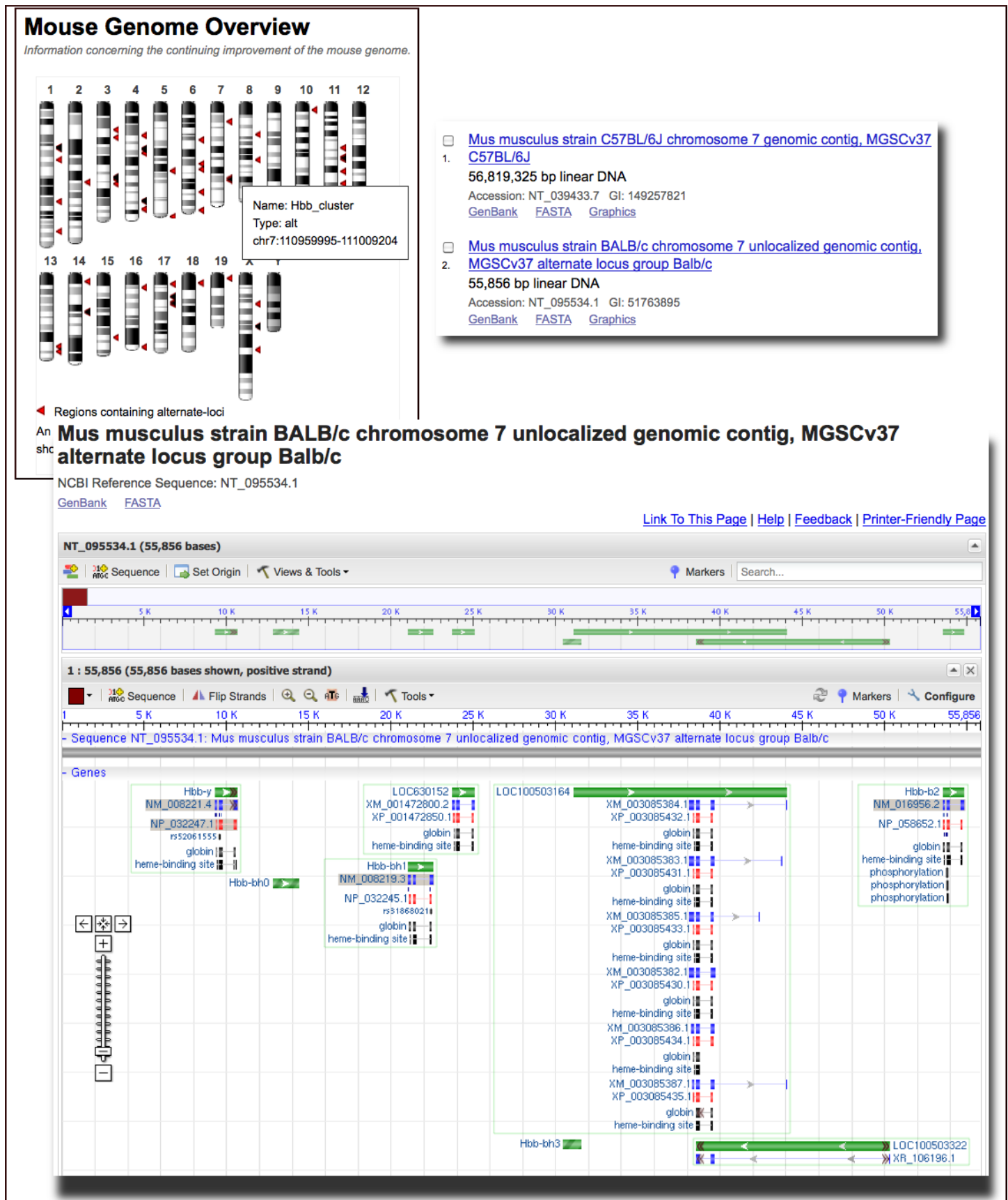


Figure 1. Mouse Genome v37.2. *Upper left panel*, mouse genome overview from the Genome Reference Consortium page showing the regions with alternate loci for other strains. *Upper right panel*, nucleotide summaries for the reference C57BL/6J (NT\_039433.7) and alternate locus for the BALB/c strain (NT\_095534.1). Both cover the beta globin region on chromosome 7. *Lower panel*, graphical sequence view of the beta globin alternate locus for the BALB/c strain.



1: HomoloGene:55465. Gene conserved in Euteleostomi
[Download](#) , [Links](#)

**Genes**  
*Genes identified as putative homologs of one another during the construction of HomoloGene.*

- [CFTR, \*H.sapiens\*](#)  
cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
- [CFTR, \*P.trogodytes\*](#)  
cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
- [CFTR, \*C.lupus\*](#)  
cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
- [CFTR, \*B.taurus\*](#)  
cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
- [Cftr, \*M.musculus\*](#)  
cystic fibrosis transmembrane conductance regulator homolog
- [Cftr, \*R.norvegicus\*](#)  
cystic fibrosis transmembrane homolog (human)
- [cftr, \*D.erio\*](#)  
cystic fibrosis transmembrane ATP-binding cassette (sub-fam

**Proteins**  
*Proteins used in sequence comparisons and their conserved domain architectures.*

- [NP\\_000483.3](#)  
1480 aa
- [XP\\_519330.2](#)  
1480 aa
- [NP\\_001007144.1](#)  
1483 aa
- [NP\\_776443.1](#)  
1481 aa
- [NP\\_066388.1](#)  
1476 aa

**UniGene**  
*Links to UniGene entries found by comparing the homologous proteins against the transcript database.*

**Chordata**

- [Mammalia - mammals \(12 transcripts, 12 species\)](#)
- [Actinopterygii - bony fishes \(7 transcripts, 6 species\)](#)
- [Amphibia - amphibians \(4 transcripts, 2 species\)](#)
- [Aves - birds \(1 transcript, 1 species\)](#)

Items 1 - 7 of 7 One page.

- 1:** [Cystic fibrosis transmembrane conductance regulator II](#)  
LOC100136366, *Salmo salar*  
Ssa.12: 2 sequences.
- 2:** [Cystic fibrosis transmembrane conductance regulator I](#)  
LOC100136364, *Salmo salar*  
Ssa.13: 5 sequences.
- 3:** [Transcribed locus, strongly similar to NP\\_001117006.1 cystic fibrosis transmembrane conductance regulator II \[Salmo salar\]](#)  
*Oncorhynchus mykiss*  
Omy.20273: 3 sequences.
- 4:** [Cystic fibrosis transmembrane conductance regulator \(CFTR\)](#)  
*Fundulus heteroclitus*  
Fhe.2641: 1 sequences.
- 5:** [Cystic fibrosis transmembrane conductance regulator, ATP-binding cassette \(sub-family C, member 7\)](#)  
*cftr, Danio rerio*  
Dr.74099: 7 sequences.  
Order cDNA clone
- 6:** [Cystic fibrosis transmembrane conductance regulator](#)  
*cftr, Takifugu rubripes*  
Tru.3674: 1 sequences.
- 7:** [Transcribed locus, moderately similar to NP\\_001117006.1 cystic fibrosis transmembrane conductance regulator II \[Salmo salar\]](#)  
*Oryzias latipes*  
Ola.17068: 5 sequences.

Figure 2. HomoloGene Cluster 55465 containing vertebrate CFTR homologs. *Top panel*, gene and sequence cluster. *Center panel*, new related UniGene section with organism summaries. *Bottom panel*, related UniGene clusters for bony fishes.



## NCBI News, January 2011

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: January 10, 2011; Updated: January 28, 2011.

### NCBI Discovery Workshops: Feb 15-16, 2011

NCBI will present a two-day workshop on February 15-16, on the NIH campus in Bethesda, Maryland. The course is free and is open to anyone interested in NCBI resources. The workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI website. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. For more information see the [Discovery Workshop](#) page, which also includes a registration link.

### Updated Resources for Genomic Libraries and Clones

NCBI has updated resources for finding genomic libraries and genomic clones from genome sequencing projects for a large number of organisms. The new [CloneDB](#) (Figure 1, *Top panel*) replaces the Clone Registry as the resource for finding descriptions, sources, and detailed statistics on available genomic libraries for a large number of organisms.

The new [Library Browser](#) (Figure 1, *Bottom panel*) allows filtering by organism, vector type, distributors, and number of associated database end or insert sequences. The linked [Clone Finder](#) (Figure 2), now available for human, mouse, rat, cow, horse, pig, and zebra finch, quickly identifies clones that span regions on assembled genomes.

The [Clone Finder](#) locates clones by chromosomal position or by features such as genes, SNPs, markers, or transcript sequence accession number. Clones may also be found in regions bounded by any two markers (Figure 2, *Middle panel*). The initial query may be refined to specific mapping data sets, population sources, and libraries. The graphical display in Clone Finder shows features annotated on the genome including assembled contigs, their components, genes, and aligned transcripts (Figure 2, *Bottom panel*).

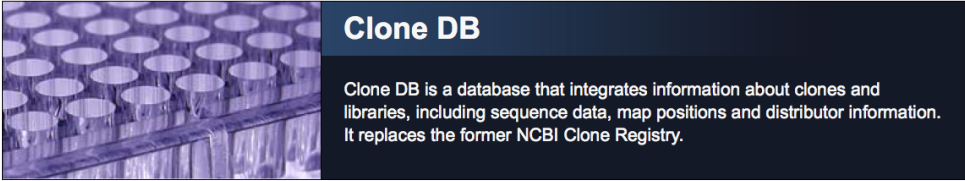
Together CloneDB's [Library browser](#) and the [Clone Finder](#) provide essential access to these important molecular reagents.

### New Gene-BioSystems Links Highlight the Gene in the Biological Pathway

Links from [BioSystems](#) are now fully listed in a separate section of Gene records (Figure 3, *Top panel*). Each of these new links point to a specific pathway or system and leads to the record in BioSystems with the specific gene highlighted (Figure 3, *Bottom panel*). For pathways imported from the [Kyoto Encyclopedia of Genes and Genomes \(KEGG\)](#) the BioSystems record provides the option to highlight the record in the large pathway diagram at the KEGG site (Figure 3, *Middle panel*). A feature article in the [July 2009 NCBI News](#) describes the BioSystems resource in detail.

## CloneDB

Connecting the lab with the genome



**Clone DB**

Clone DB is a database that integrates information about clones and libraries, including sequence data, map positions and distributor information. It replaces the former NCBI Clone Registry.

### Getting Started

- [An overview of Clone DB](#)
- [FAQ](#)
- [News and Announcements](#)

### Tools

- [Genomic clone library browser](#)
- [Clone DB Distributors](#)
- [Clone DB FTP site](#)
- [Clone Finder](#)
- [Search Old Clone Registry](#)

### Related Resources

- [NCBI MapViewer](#)
- [NHGRI Structural Variation Project](#)
- [Human BAC Resource](#)
- [CCAP Clones](#)

### Clone DB: Genomic Clone Library Browser

This browser provides a listing of genomic clone libraries with records in Clone DB. Data filters are provided to assist with navigation of the table.

- [Library Distributors](#): Contact information for distributors in this library browser
- [FAQ](#): Frequently asked questions about this library browser

**Filters currently applied**

[undo] (organism=Homo sapiens)

[undo] (vector=BAC)

[undo] (distributor=CHORI)

Source: NCB

Genomic Clone Library Browser						
Library Name	Library Abbreviation	Vector types	Distributors	Total clones	Total end sequences	Total insert sequences
<a href="#">RPCI human BAC library 11</a>	RP11	BAC	Empire Genomics imaGenes CHORI Invitrogen	292,384	391,211	32,849
<a href="#">The CHORI-17 BAC Library from a hydatidiform (haploid) mole</a>	CH17	BAC	CHORI	164,737	325,659	246
<a href="#">RPCI human BAC library 13</a>	RP13	BAC	Empire Genomics CHORI	2,778	0	1,128
<a href="#">CHORI-507 Human BAC Library</a>	CH507	BAC	CHORI			
<a href="#">CHORI 15 Human female BAC library</a>	CH15	BAC	CHORI			
<a href="#">CHORI-16 Human sheared BAC library</a>	CH16	BAC	CHORI			
<a href="#">CHORI 502 Human MHC BAC Library</a>	CH502	BAC	CHORI			
<a href="#">CHORI 501 Human MHC BAC Library</a>	CH501	BAC	CHORI			
<a href="#">CHORI 14 Human male BAC library</a>	CH14	BAC	CHORI Empire Genomics			

**Filter by End Sequence**

>=100

>=1000

>=10,000

>=100,000

>=1,000,000

Source: NCB

**Filter by Insert Sequence**

>=100

>=500

>=1000

>=5000

>=10,000

Source: NCB

#### RPCI human BAC library 11

##### Library Summary

<b>Library Name:</b>	RPCI human BAC library 11
<b>Library Abbreviation:</b>	RP11
<b>Organism:</b>	Homo sapiens
<b>Distributors:</b>	CHORI, Empire Genomics, imaGenes, Invitrogen
<b>Vector type(s):</b>	BAC
<b># clones Clone DB:</b>	292,384
<b># end sequences Clone DB:</b>	391,211
<b># insert sequences Clone DB:</b>	32,849
<b># clones with both ends sequenced:</b>	132,872

##### Library Details

DNA Source

Library segment	Sex	Organ
ALL	male	blood

Figure 1. The CloneDB homepage (Top panel) and the associated Genomic Clone Library Browser (Bottom panel). The Library Browser provides Filters to narrow down the selected libraries. The display has the *Homo sapiens* organism filter and BAC vector filters applied. The inset shows the record for the RP11 BAC library with links to Distributors.

**CloneFinder Home**

Clone Finder is a tool developed to facilitate the identification of clones within given genomic regions.

**Related Resources**

- NCBI Home
- NCBI Clone
- NCBI Web Search
- NCBI Site map
- Genome Browser agreement
- Genome Biology
- Taxonomy
- Entrez (Global Query)
- BLAST
- Map Viewer

**Vertebrates**

**Mammals**

**Primates**

Scientific name	Common name	Build
<i>Homo sapiens</i>	human	Build 37.2 Build 36.3

**Rodents**

Scientific name	Common name	Build
<i>Mus musculus</i>	laboratory mouse	Build 37.1
<i>Rattus norvegicus</i>	rat	RGSC v3.4

**Other Mammals**

Scientific name	Common name	Build
<i>Bos taurus</i>	cattle	5.2 Btau 4.0
<i>Equus caballus</i>	horse	EquCab2.0
<i>Sus scrofa</i>	pig	Scrofa9.2 Scrofa5

**Other Vertebrates**

Scientific name	Common name	Build
<i>Gallus gallus</i>	chicken	Build 2.1
<i>Taeniopygia guttata</i>	zebra finch	Build 1.1

**Invertebrates**

**Insects**

Scientific name	Common name	Build
<i>Apis mellifera</i>	honey bee	Amel_4.0

---

**Clone Search**

**Homo sapiens Clone Finder** Build: **Build 37.2 (current)** [Change Build](#)

**Specify Region**

Search by Position | **Search by Feature**

Region	Feature type	Feature name
Chromosome: <b>19</b>	From: <b>Gene</b>	<b>BCAM</b>
Assembly: <b>GRCh37.p2</b>	To: <b>Gene</b>	<b>SFRS16</b>

[Go](#)

**Select Region**

Assembly	Chromosome	Begin	End	Length
GRCh37.p2	19	45,312,338	45,594,782	282,445

Select placement ranges to include

Begin	End	Length
<input checked="" type="checkbox"/> 45,312,338	45,324,678	12,341
<input checked="" type="checkbox"/> 45,582,518	45,594,782	12,265
<input type="checkbox"/> 49,298,319	49,314,320	16,002

[Find Clones](#)

**Set Data display filters**

- Dataset selection
- DNA Source
- Population Selection
- Library Selection  Check all  Clear all

**BAC vectors**

<input type="checkbox"/> CH17	<input type="checkbox"/> CTA	<input type="checkbox"/> CTB	<input type="checkbox"/> CTC	<input type="checkbox"/> CTD
<input type="checkbox"/> GS1	<input checked="" type="checkbox"/> RP11	<input type="checkbox"/> RP13		

---

**Homo sapiens CloneFinder** Build **37.2**

Assembly GRCh37.p2 - Primary Assembly [Reference Complete][Assembly GCF\_000001405.14][Assembly Unit GCF\_000001305.13] [Back to search](#)

**Chromosome 19 NC\_000019.9** 45,312,338-45,574,214 bp

**Data Summary** Download: [Image](#) [Excel](#)

45,312,338 45,350 K 45,400 K 45,450 K 45,500 K 45,574,214

Contigs: NT\_011109.16

Components: AC011481.4 AC011489.6

Genes on sequence: BCAM PWR12 CTFP1M RPL16 SFRS16

Transcripts on sequence: NM\_001042724.1 NM\_002656.2 NM\_001294.1 NM\_006509.2 NM\_007056.2

Ensembl gene annotations: BCAM PWR12 CTFP1M RPL16 SFRS16

Ensembl transcript: RP11-1147O10

Feature RP11-1147O10  
Type Clone  
Description(s) Library: RP11  
Chrom 19  
Chr Pos 45,339,777 - 45,492,883  
Contig NT\_011109.16  
Contig Pos 17,607,995 - 17,761,101  
Span 153,107

Clone Library: R

Clone Ends  
AQ776185.1 span: 407  
AQ823421.1 span: 307

Concordant  Discordant

Figure 2. The Clone Finder tool. *Top panel:* The Clone Finder homepage with access to clones for a number of genomes in Map Viewer. *Middle panel:* Clone Finder for Homo sapiens Build 37.2 set to find BAC clones from the RP11 library for the region between the genes BCAM and SFRS16. *Bottom panel:* Results shown in the browser with the clone RP11-1147O10 selected showing position information. BAC end sequences are listed and linked to the corresponding Genome Survey Sequence (GSS) records.

**Pathways from BioSystems**

[Angiotensin receptor Tie2-mediated signaling, organism-specific biosystem](#) (from Pathway Interaction Database)

[Cell surface interactions at the vascular wall, organism-specific biosystem](#) (from REACTOME)

[Class A/1 \(Rhodopsin-like receptors\), organism-specific biosystem](#) (from REACTOME)

[Common Pathway, organism-specific biosystem](#) (from REACTOME)

[Complement and coagulation cascades, organism-specific biosystem](#) (from KEGG)

[Complement and coagulation cascades, conserved biosystem](#) (from KEGG)

[Diabetes pathways, organism-specific biosystem](#) (from REACTOME)

[FOXA2 and FOXA3 transcription factor networks, organism-specific biosystem](#) (from Pathway Interaction Database)

[Formation of Fibrin Clot \(Clotting Cascade\), organism-specific biosystem](#) (from REACTOME)

[Formation of Platelet plug, organism-specific biosystem](#) (from REACTOME)

[G alpha \(q\) signalling events, organism-specific biosystem](#) (from REACTOME)

[GPCR downstream signaling, organism-specific biosystem](#) (from REACTOME)

[GPCR ligand binding, organism-specific biosystem](#) (from REACTOME)

[Gamma-carboxylation of protein precursors, organism-specific biosystem](#) (from REACTOME)

[Gamma-carboxylation, transport, and amino-terminal cleavage of proteins, organism-specific biosystem](#) (from REACTOME)

[Hemostasis, organism-specific biosystem](#) (from REACTOME)

**COMPLEMENT AND COAGULATION CASCADES**

Genes	Proteins	Small Molecules	Related BioSystems	Citations	Comments
View or save all or selected records in Entrez Gene   Clear Selections   <b>Highlight Selected Records in Source Database</b>					
<input checked="" type="checkbox"/>	2147	F2	2147	coagulation factor II (thrombin)	
<input type="checkbox"/>	2	A2M	2	alpha-2-macroglobulin	
<input type="checkbox"/>	462	SERPINC1	462	serpin peptidase inhibitor, clade C (antithrombin), member 1	
<input type="checkbox"/>	623	BDKRB1	623	bradykinin receptor B1	
<input type="checkbox"/>	624	BDKRB2	624	bradykinin receptor B2	
<input type="checkbox"/>	629	CFB	629	complement factor B	
<input type="checkbox"/>	710	SERPING1	710	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1	
<input type="checkbox"/>	712	C1QA	712	complement component 1, q subcomponent, A chain	
<input type="checkbox"/>	713	C1QB	713	complement component 1, q subcomponent, B chain	
<input type="checkbox"/>	714	C1QC	714	complement component 1, q subcomponent, C chain	

Page 1 of 7 | Displaying Genes 1 - 10 of 69

Figure 3. Linking to BioSystems from the Gene record for F2 coagulation factor II (thrombin) (Gene ID: 2147). *Top panel:* Explicit links to Biosystems from the Gene records. *Bottom panel:* BioSystems record for Complement and coagulation cascades, organism-specific biosystem with Gene ID 2147 selected (bsid83073). Clicking the Highlight link opens the pathway diagram at the KEGG site with F2 highlighted. *Middle panel:* A portion of the KEGG pathway showing the Coagulation cascade.

## New Organisms in UniGene

Five new organisms have builds in UniGene: the hydrozoan (Cnidaria) *Clytia haemisphaerica*; the perigord truffle, *Tuber melanosporum*; the English or Truffle Oak, *Quercus robur*; the two-spotted spider mite, *Tetranychus urticae*; and the salmon louse, *Lepeophtheirus salmonis*.

*Clytia haemisphaerica* ([Che build information](#), 4,637 clusters, FTP) is a marine hydrozoan in the phylum Cnidaria, the phylum that also contains jellyfish (scyphozoa) and corals (anthozoa). Unlike *Hydra*, the other hydrozoan in UniGene, *Clytia* has a free-swimming medusa stage. Gene and genomic information from *Clytia* has the potential to provide important insights on the evolution of animal body plans.

The perigord truffle ([Tme build information](#), 7,543 clusters, FTP), an ascomycete fungus and the source of the gastronomically highly prized black truffle, and the truffle oak ([Qro build information](#), 7,170 clusters, FTP) are two organisms linked in a symbiotic mycorrhizal association. UniGene sets from these two organisms should support studies of genes involved in the evolution, function, and maintenance of symbiosis.

Two parasitic arthropods of economic importance also join UniGene. The two-spotted spider mite ([Tur build information](#), 7,177 clusters, FTP) is a significant pest of ornamental and horticultural plants. The salmon louse ([Lsl build information](#), 9,363 clusters, FTP) is an ectoparasitic copepod parasite that can cause significant mortality in farmed and wild salmon. These sets may prove helpful in understanding the biology of parasitism and provide targets for control of these pests.

## NCBI Databases in Nucleic Acids Research Database Issue

The Nucleic Acids Research 2011 Database Issue contains nine articles about NCBI resources, tools, and databases including Gene, GEO, Epigenomics, CDD and GenBank. Free full-text articles from the database issue are available from PubMed Central and the publisher's site and are linked to the [summaries](#) and [abstracts](#) in PubMed.

## dbSNP BLAST Pages Updated

The [dbSNP BLAST page](#) has an updated submission form and output format. The new pages have improved organism selection, chromosome specific database selection, and many of the convenient features of the other BLAST services.

## New Mammalian Genomes at NCBI

Updated genome annotations for the rat ([build 4.2](#)), cow ([build 5.2](#)), and a new pig assembly ([build 2.1](#)) are now available for searching and viewing in Entrez, BLAST, the Map Viewer, and for downloading from [genomes area](#) of the FTP site.

## Microbial Genomes Update

Sixty-five finished microbial genomes were released during November and December 2010. The original sequence data files submitted to GenBank/EMBL/DDJB are available in the [Bacteria directory](#) in the genomes area of the GenBank FTP site. [RefSeq provisional versions](#) were made for a selected set of 46 these genomes.

In addition, 100 microbial whole genome shotgun-sequencing projects were added to GenBank during this period. The original submitted files are available in the [Bacteria\\_DRAFT directory](#) in the GenBank genomes area. [RefSeq provisional versions](#) of 64 of these projects are also available.

All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated [Entrez](#) search and retrieval system.

## New Video on NCBI's YouTube Channel

A [new video](#) that shows how use My NCBI to save searches and set up automated E-mail alerts for new results is now available on [NCBI's YouTube channel](#).

## GenBank News

GenBank release 181 is available through the NCBI web and [FTP](#) sites. The current release incorporates data available as of Dec 15, 2010 and contains 122,082,812,719 bases from 129,902,276 sequence records. [Release notes](#) describe the current state of data and upcoming changes.

## RefSeq News

RefSeq Release 45 is now available through the Entrez system and can be downloaded from the [FTP site](#). This full release incorporates genomic, transcript, and protein data available as of January 7, 2011 and includes 16,748,646 records from 11,536 different species and strains. The [release notes](#) describe changes since the last release. New in this release is the inclusion of additional features present on the corresponding UniProt/Swiss-Prot record for a subset of RefSeq proteins. These new features are indicated with a Note that identifies the source accession number. An example from [NP\\_080213.3](#) is shown below.

```
Site          147
              /site_type="phosphorylation"
              /experiment="experimental evidence, no
              additional details recorded"
              /note="Phosphoserine; propagated from
              UniProtKB/Swiss-Prot(Q9D0F4.1)"
```

The [RefSeq Homepage](#) has more information on the RefSeq project.

## Journals Database Now a Part of NLM Catalog

The NCBI Journals Database is now part of NCBI [NLM Catalog](#). The NLM Catalog contains the detailed MEDLINE indexing information for the journals in PubMed and other NCBI databases and will maintain the functions of the Journals database.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html).

To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook and Twitter: [www.twitter.com/NCBI](http://www.twitter.com/NCBI).

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, October 2010

Peter Cooper, Ph.D.<sup>1</sup> and Rana Morris, Ph.D.<sup>2</sup>

Created: November 10, 2010; Updated: December 29, 2010.

### New Databases and Tools

#### Instructional Videos on the NCBI YouTube Channel

The NCBI YouTube channel now has ten instructional videos that demonstrate how to use NCBI tools and databases. All ten are available through the [Video Tutorials playlist](#). Topics include several [How-to tasks](#) from the NCBI Homepage, using the new Find-in-Sequence feature in the sequence databases, browsing the new Epigenomics resource, and using Genome Workbench.

The screenshot shows the NCBI YouTube channel interface. The main video player displays the title "Obtain Genomic Sequence" with the subtitle "For and Surrounding a Gene" and the NCBI logo. Below the video player are options for "Info", "Favorite", "Share", "Playlists", and "Flag". The video description reads: "Obtain Genomic Sequence for a gene. From: NCBINLM | October 24, 2010 | 127 views. Update, Obtain Genomic Sequence for and surrounding a gene, short version." To the right of the video player is a "How to use NCBI" playlist containing ten videos:

- Epigenomics: Using the Sample Browser (NCBINLM - 59 views, 2:26)
- Epigenomics: How to Download Data (NCBINLM - 10 views, 1:57)
- Epigenomics: How to View Track Data (NCBINLM - 35 views, 2:33)
- Find in This Sequence (NCBINLM - 196 views, 2:17)
- Genome Workbench: Search and View a (NCBINLM - 142 views, 2:18)
- Genome Workbench: Phylogenetic Trees (NCBINLM - 65 views, 2:49)

#### The RefSeqGene Project: a Stable Resource for Gene Annotations

NCBI produces RefSeqGene as a subset of NCBI's Reference Sequence (RefSeq) project. These genomic sequences are intended to serve as reference standards for well-characterized genes and offer stable genomic platforms with a permanent identifier and a core content that does not change. This provides a standard system for numbering exons and introns and defining the coordinates of variations and other features. Each RefSeqGene record offers gene-specific sequences for each gene as well as upstream and downstream flanking

regions. All records are experimentally well-supported, come from a single genomic clone, and represent the allele present in the Genome Reference Consortium reference genome whenever possible. The RefSeqGene project is an active member of the [Locus Reference Genomic \(LRG\)](#) collaboration that accepts submission of variants, nominations for target genes, and provides input on curation.

Entrez [nucleotide](#) system incorporates RefSeqGene records where they may be selected by adding the term `refseqgene[ keyword ]` to any query. RefSeqGene records are also linked from their corresponding RefSeq transcript, genomic clone, protein sequence and gene records. The [RefSeqGene homepage](#) has documentation, links to related resources and tools – including the RefSeqGene BLAST service , described below – and a [list of available RefSeq gene records](#) by gene symbol. The RefSeqGene data are available for download from the [RefSeq area](#) of the FTP site.

## RefSeqGene BLAST Service Now Available, Preview of Enhancements to BLAST

The NCBI BLAST web service now includes a specialized BLAST service that searches the NCBI RefSeqGene records. RefSeqGene BLAST also offers a preview of changes in format coming to the main BLAST services. These changes include a two-line format in the Descriptions section allowing more of the sequence title to be displayed. The second line has separate links to display formats in the sequence databases (GenBank, FASTA). Additional links to related data such as corresponding Gene records will be added to this second line in the future. Also on the second line is a link to display BLAST output in the graphical sequence viewer. This latter option allows the BLAST hits to be displayed in the context of the biological features annotated on the database record and provides a powerful new way to look at BLAST results.

### Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)

- Make specific primers with [Primer3](#)
- Search [trace archives](#)
- Find [conserved domains](#) in
- Find sequences with similar [protein](#)
- Search sequences that have [similar](#)
- Search [immunoglobulins](#) (IgBLAST)
- Search for [SNPs](#) (snp)
- Screen sequence for [vector](#)
- [Align](#) two (or more) sequen
- Search [protein](#) or [nucleoti](#)
- Search SRA [transcript and](#)
- Constraint Based Protein M
- Needleman [Wunsch Globa](#)
- Search [RefSeqGene](#)

Description
Homo sapiens monoamine oxidase A (MAOA), RefSeqGene on chromosome X NG_008957.1 <a href="#">GenBank</a> <a href="#">FASTA</a> <a href="#">Graphics</a>
Homo sapiens monoamine oxidase B (MAOB), RefSeqGene on chromosome X NG_008723.1 <a href="#">GenBank</a> <a href="#">FASTA</a> <a href="#">Graphics</a>

Subject: ref|NG\_008723.1| Length: 122865 Sort by: [E value](#) [Score](#) [Percent identity](#) [Query start position](#) [Subject start position](#)

ref|NG\_008723.1| Homo sapiens monoamine oxidase B (MAOB), RefSeqGene on chromosome X  
[GenBank](#) [FASTA](#) [Graphics](#) [Download Sequence](#)

Score	Expect	Identities	Gaps	Strand
140 bits(154)	3e-31	94/105(90%)	0/105(0%)	Plus/Plus

```

Query 763 TTACATCCAGTACACTATGAAGAGAAGAACTGGTGTGAGGACAGTACTCTGGGGCTGC 822
Sbjct 112197 TTGCAGCCAGTGCATTATGAAGAAAAGAACTGGTGTGAGGACAGTACTCTGGGGCTGC 112256

Query 823 TACACGGCGTACTTCCCTCCTGGGATCATGACTCAATATGGAAGG 867
Sbjct 112257 TACACAACTTATTTCCCCCTGGGATCCTGACTCAATATGGAAGG 112301
                    
```

NG\_008723.1 (122,865 bases)

## Microbial Genomes

Twenty-one finished microbial genomes were released during September 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>. In addition 25 microbial whole genome shotgun sequencing projects were added. Original submitted files are available in [ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria\\_DRAFT/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria_DRAFT/) and RefSeq provisional versions are in [ftp://ftp.ncbi.nih.gov/genomes/Bacteria\\_DRAFT/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria_DRAFT/). All GenBank and RefSeq microbial genomes are incorporated in the NCBI integrated [Entrez](#) search and retrieval system.

## GenBank News

GenBank release 180.0 is available through the NCBI web and [FTP](#) sites. The current release includes information available as of October 15, 2010. [Release notes](#) describe the current state of data and upcoming changes.

## Updates and Enhancements

### Updated Gene Pages

The [Entrez Gene](#) database has moved to the updated interface introduced in PubMed last year. The new format has better controls and options for displaying and downloading records. More importantly, Gene record displays have several improvements including easier navigation, customizable display options, and better integration with closely related data in the NCBI system. Integration with the NCBI Reference Sequence transcripts, proteins, and genomic assemblies, as well as sequence variations from the dbSNP has been greatly improved through the graphical sequence viewer embedded in the gene record, as shown in the accompanying image.

Display Settings:  Full Report [Send to:](#)

## TH tyrosine hydroxylase [ *Homo sapiens* ]

Gene ID: 7054, updated on 1-Nov-2010

**Summary**

**Official Symbol** TH provided by [HGNC](#)

**Official Full Name** tyrosine hydroxylase provided by [HGNC](#)

**Primary source** [HGNC:11782](#)

**See related** [Ensembl:ENSG00000180176](#); [HPRD:01865](#); [MIM:191290](#)

**Gene type** protein coding

**RefSeq status** REVIEWED

**Organism** [Homo sapiens](#)

**Lineage** Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo

**Also known as** TYH; DYT14; DYT5b; TH

**Summary** The protein encoded by this gene is involved in the conversion of tyrosine to dopamine. It is the rate-limiting enzyme in the synthesis of catecholamines, hence plays a key role in the physiology of adrenergic neurons. Mutations in this gene have been associated with autosomal recessive Segawa syndrome. Alternatively spliced transcript variants encoding different isoforms have been noted for this gene. [provided by RefSeq]

**Genomic regions, transcripts, and products**

(minus strand) Go to [reference sequence details](#)

-2,194,294 : -2,183,900 (10,395 bases shown, negative strand) [Open Full View](#)

assembly - Sequence [nc\\_000011.9: Homo sapiens chromosome 11, GRCh37.p2 primary](#)

- NCBI genes

TH

NM\_199292.2

NP\_954986.2

NM\_000360.3

NP\_000351.2

NM\_199293.2

NP\_954987.2

BLAST Genome-specific: [NC\\_000011.9 \(2,185,159..2,193,035\)](#)

BLAST Genome-specific: [NM\\_199292.2 \(2,185,159..2,193,035\)](#)

BLAST Genomic: [NC\\_000011.9 \(2,185,159..2,193,035\)](#)

BLAST mRNA: [NM\\_199292.2](#)

FASTA View: [NC\\_000011.9 \(2,185,159..2,193,035\)](#)

FASTA View: [NM\\_199292.2](#)

GenBank View: [NC\\_000011.9 \(2,185,159..2,193,035\)](#)

GenBank View: [NM\\_199292.2](#)

Graphical View: [NM\\_199292.2](#)

View GeneID: 7054; Gene Symbol: TH

View HGNC: 11782

View HPRD: 01865

View MIM: 191290

Set Sequence Origin At Position

Zoom In

Zoom Out

Zoom On Range

Zoom To Sequence

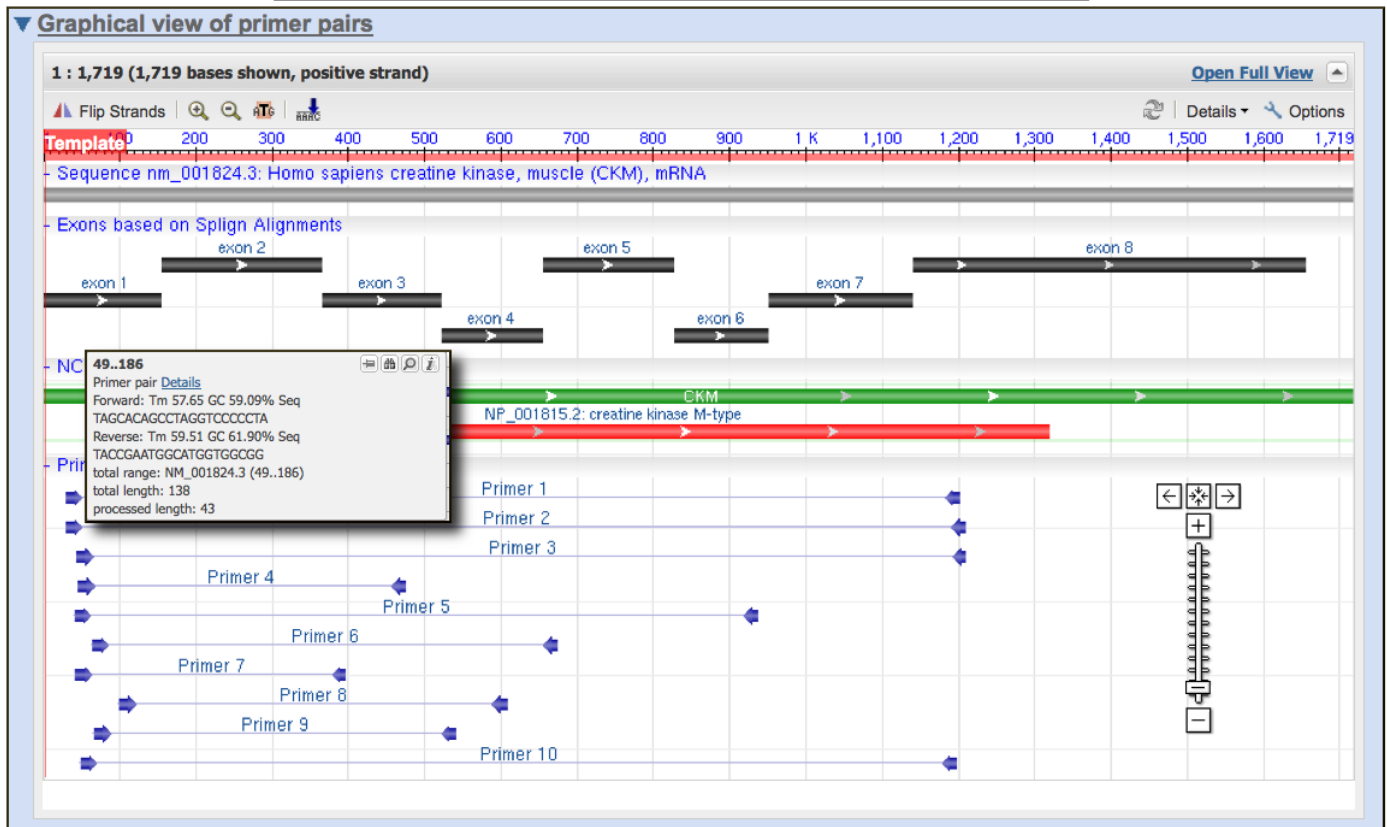
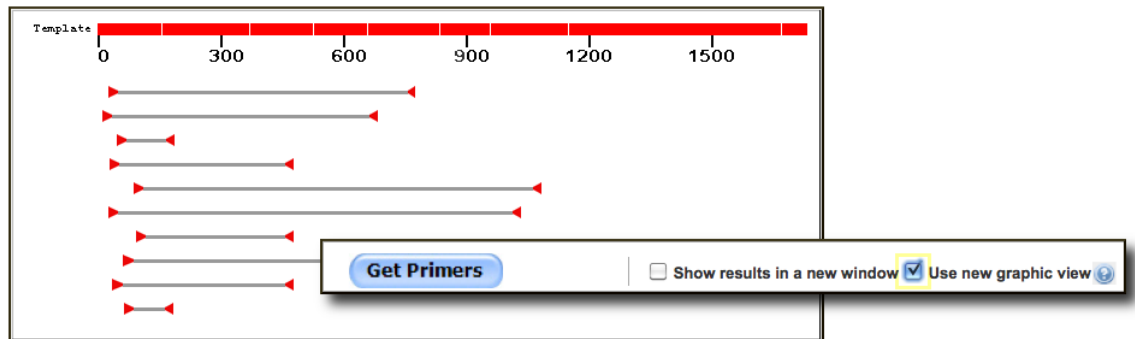
Set Sequence Origin At Feature

Properties

Views & Tools

## Graphic Display in Primer-BLAST

Primer BLAST results now offer an alternative to the standard graphic that allows primer alignments to be displayed in the Graphical Sequence viewer. This new option is activated by the *Use new graphic view* option on the web form. Results in the graphical sequence viewer display the primer binding sites in the context of the biological features of the sequence such as the locations of exons, introns, coding sequences, and untranslated regions making it easier to assess the usefulness of particular primer pairs.



## Genome Workbench 2.2.0

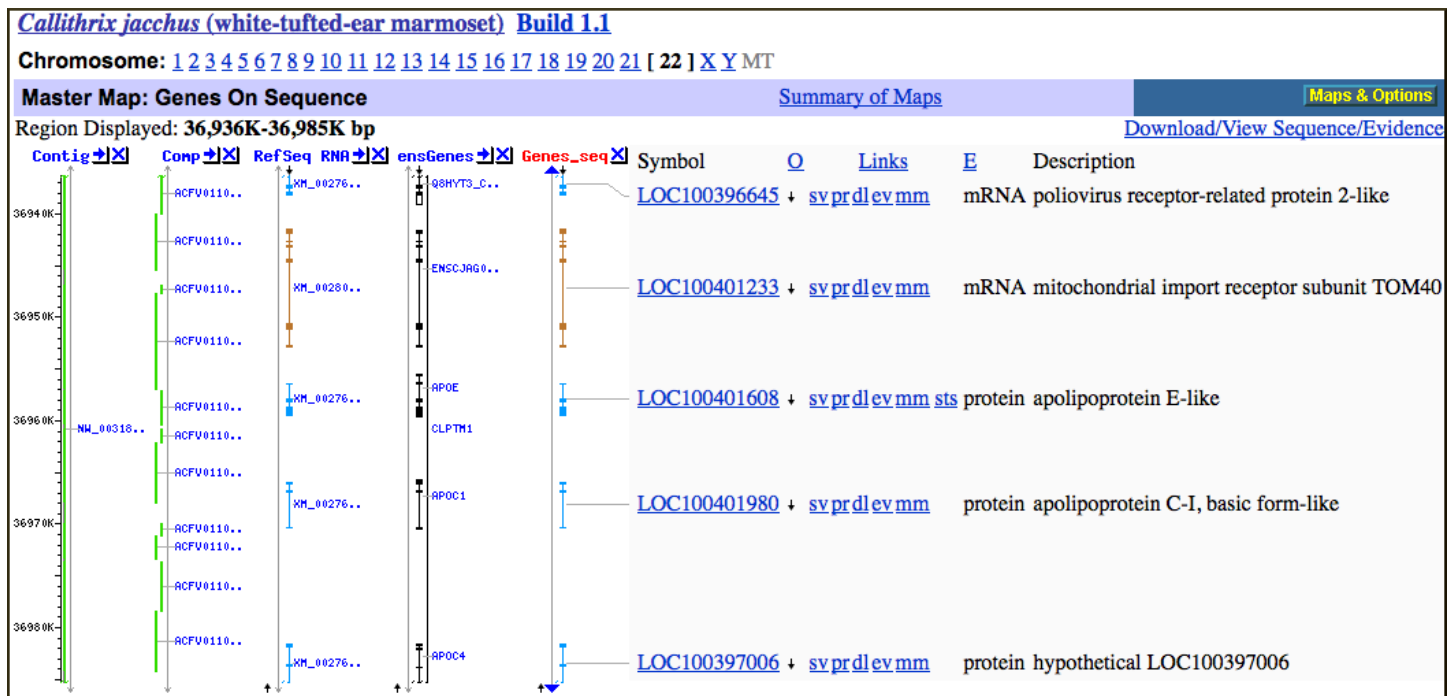
A new release of the NCBI Genome Workbench – the downloadable graphical sequence analysis, annotation, and display platform – is now available. This latest version, 2.2.20, includes several bug fixes described in the [release notes](#). More information about Genome Workbench along with a link to instructions for downloading and installing the program are provided on the [homepage](#).

## Reference Human Genome, GRCh37, Updated to Patch 2

The reference human genome has been updated to patch 2 (GRCh37.p2). The new build is available in [Entrez](#), the [Map Viewer](#), [BLAST](#), and on the [FTP site](#). This update includes 70 patches. Patches are updated regional assemblies that either provide additional alternate assemblies for alleles that are not adequately represented in the current genome (Novel patches) or correct assembly errors in the current build (Fix patches). GRCh37.p2 comprises 52 novel patches and 18 fix patches. Fix patches be incorporated the next major genome build changing the tiling path while novel patches will be incorporated as alternate loci.

## Giant Panda and Marmoset Added to Map Viewer

The NCBI Map Viewer and genomes FTP area (<ftp://ftp.ncbi.nih.gov/genomes/>) have new genome assemblies and their corresponding annotations for the giant panda (*Ailuropoda melanoleuca*, build 1.1), and the marmoset (*Callithrix jacchus*, build 1.1) build as well as updated annotations for the rhesus monkey (*Macaca mulatta*, build 1.2) and Sumatran orangutan (*Pongo abelii*, build 1.2). The giant panda genome reported in by the January 21, 2010 issue of Nature is a whole genome shotgun assembly produced from next generation (Illumina GA) sequencing reads with approximately 60x genome coverage. Contigs are not placed on chromosomes. However, the Map Viewer provides graphical displays of the contigs, genes, gene models, and aligned carnivore expressed sequences from GenBank and RefSeq. The NCBI marmoset genome and annotation is based on the whole genome shotgun assembly released by Washington University Genome Sequencing Center and the Baylor College of Medicine Human Genome Sequencing Center as *Callithrix jacchus*-3.2 in March 2009. The marmoset genome comprises 22 autosomes plus X and Y sex chromosomes with mapped genes, human and marmoset expressed sequences, and STS markers. Genomic BLAST services for both [panda](#) and [marmoset](#) allow similarity search results to be displayed in genomic context in the panda and marmoset sequence maps.



## RefSeq

RefSeq Release 44 is now available through the Entrez system and can be downloaded from the [FTP site](#). This full release incorporates genomic, transcript, and protein data available as of November 7, 2010 and includes 16,421,261 records from 11,354 different species and strains. The [release notes](#) describe changes since the last release. The [RefSeq Homepage](#) has more information on the RefSeq project.

## Changes affecting E-utilities

### GEO Database Name changes: geo to geopfiles

Recently the name of the [GEO Profiles](#) database in E-utilities changed from 'geo' to 'geopfiles'. While the old name (db=geo) will still function for a time, requests should be changed to use the new name (db=geopfiles). ELink users should be aware that all linknames including 'geo' will no longer function. Instead, these names

should include 'geoprofiles' rather than 'geo'. For example, the linkname of links from Gene to GEO Profiles is now linkname=gene\_geoprofiles.

## **Retirement of the Journals Database: journal data has been moved to nlmcatalog**

The NCBI Journals Database will be retired in mid-December, approximately on December 13, 2010. The NCBI NLM Catalog will contain the detailed MEDLINE indexing information for the journals in PubMed and other NCBI databases. ESearch URLs for db=journals will automatically map to db=nlmcatalog. ESummary and EFetch will retrieve NLM Catalog XML. The [NLM Technical Bulletin](#) has more details on this change.

## **New DTDs**

The PubMed E-Utility DTDs will be updated for 2011 in mid-December, approximately on December 13, 2010. The new DTDs are available from the following pages.

[http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed\\_110101.dtd](http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed_110101.dtd)

[http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/nlmmedlinecitationset\\_110101.dtd](http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/nlmmedlinecitationset_110101.dtd)

## **SOAP Update**

The NCBI E-Utility/SOAP Web site has been updated and includes a new WSDL and examples on usage. Please consult the [EUtility/SOAP homepage](#) for more information.

## **Announce Lists and RSS Feeds**

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook and Twitter: [twitter.com/NCBI](http://twitter.com/NCBI).

Send comments and questions about NCBI resources to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, September 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: September 27, 2010.

### New Databases and Tools

#### Find-in-Sequence now available in nucleotide and protein databases

The new Find-in-Sequence search finds and highlights subsequences, patterns, or motifs in nucleotide or protein sequences displayed in the Entrez system. In nucleotide sequences, sub-sequences or patterns of interest may include restriction enzyme recognition sites, transcription factor binding sites and other promoter elements, poly-adenylation signals, start and stop codons, and many others. In proteins, these may include potential cleavage locations, sites of potential post-translational modification, motifs representing active sites, cofactor binding pockets, or other structural or functional signatures. The link to open the Find-in-Sequence search box appears in the Analysis Tools portlet in the right-hand Discovery column of nucleotide and protein record views. Find-in-sequence works with single and multiple sequence displays with any format that shows the sequence (GenBank, GenPept, FASTA). The search box that opens at the bottom of the display has the familiar look and feel of the find in page function of the Web browser showing all matches and providing means to jump to the next or previous match. However, unlike the simple find function, Find-in-Sequence allows matching across spaces and line breaks in the formatted sequence and can use standard nucleotide and protein ambiguity codes as well as [Prosite patterns](#) for protein sequences. Find-in-Sequence provides rapid mapping of custom features onto nucleotide and protein sequences and is a powerful addition to the suite of analysis tools now available directly from sequence records in the Entrez system.

A [video](#) demonstrating this feature is now available on the NCBI YouTube channel.

Display Settings: GenBank Send: Change region shown

## Homo sapiens prolactin (PRL), transcript variant 1, mRNA

NCBI Reference Sequence: NM\_000948.4  
[FASTA](#) [Graphics](#)

Go to:

LOCUS NM\_000948 1359 bp mRNA linear PRI 20-SEP-2010  
 DEFINITION Homo sapiens prolactin (PRL), transcript variant 1, mRNA.  
 ACCESSION NM\_000948  
 VERSION NM\_000948.4 GI:254281326

[brown, isoform A \[Drosophila melanogaster\]](#)

1. NCBI Reference Sequence: NP\_523824.1  
[GenPept](#) [Graphics](#) [Related Sequences](#) [Identical Proteins](#)

>gi|17647219|ref|NP\_523824.1| brown, isoform A [Drosophila melanogaster]  
 MQESGGSSGGPSPSLCLEWKQLNYVDPQEQSNYSFWNECRKKRELRLIQDASGHMKTGDLIALIGGSA  
 TLLAAISQRLRGNLTGDDVVLNMGAMERHQMTRISSFLPQFEINVKTFAYEHLVFMHFHMRRTK  
 AEKRQRVADLLAVGLRDAHTRIQQLSGGERKRLSLAEELITDPIFLFCDEPTTGLDPSAYSVIKTLR  
 DSFETPSGESSASGSGSKSIEMEVVAESHELQTMRELPAVGLSNSPNG  
 LFTHIILMDGGRIVYQGRTEQAQKFFPTDLGYELPINCNPADPYLKTLDLADKE  
 LYSGSWLLARSYSGDYLVKHVQNFKKIRWIVQVYLLMVRPMTEDLRNIRSGL  
 TGGLTQRTVQDVGGSIFPLSNEMIFPFSYGVYIFPAALPIIRREVGGTY  
 KGYVFLSVIYASIIYTRGFLLYLSMGFLMSLSAVAAGVGVFLSLSLFSDEK  
 YNNVDVTPGLKYLSPFYSNEALMKFWIDIDINDCPVNEHDPCKTGVVEV  
 LVVVAVIFHIVSFGLVRRYIHRSGY

[Homo sapiens](#)  
 NCBI Reference Sequence: NP\_003733.2  
[Related Sequences](#) [Identical Proteins](#)

NP\_003733.2| bile salt export pump [Homo sapiens]  
 FESDKSYNDDKRSRLQDEKKGDDVVRVGFQVLRFRSSSTDIWLMFVGSICAF  
 YIDYDVELQELQIPGKACVNNITVWNSLQNMNTNGTRCGLNIESEMIK  
 ICFWVIAAARQIQKNRKFYFRIRMRMEIGWFDNCNSVGLNTRFSDINKIN  
 GFLGFFRGNKLTLLVIISVSPILIGIAATIGLSVSKFTDELKAYAKAGVV  
 EVERYEKNLVFAQRWGIKGIWGFPTGFVWCLIFLCYALAFNYGSTLVLD  
 ALNLGNASPCLEAFATGRAAATSIFETIDRKPFIIDCMSEDEGKLDRIKGEI  
 DLNLMVIKPGENTALVGPSSGAGKSTALQLIQRFPDCEGMVTVDGHDIRSLN  
 TIAENIRYGREDAIMEDIVQAAKEANAYNFIIMDLPOQFDTLVGGGGQMS  
 LLLDMATSALDNESEAMVQEVLSKIQHGHITIIISVAHRLSTVRAADTIIGPE  
 YFTLVTLQSQGNQALNEEDIKDATEDDMILARTFSRGSYQDLSRASIRQRSK  
 TYEEDRDKDIPVQEEVPEPVRRLKFSAPWEPYMLVGSVGAAVNGTVTP  
 DIALSPQISGQYSPDMLRQRSQINGVCLLFVANGCVSLFTQFLQGYAFKSGELLTKRLKFGFRAML  
 GQDIAPDDLRLNSPGALTTRLATDASVQVGAAGSQIGMIVNSFTNVTVMAMIIAFSPWKLSLVILCFPPF  
 LALSGATQTRMLTGPASRDKQALEMVGQITNEALSNIPTVAGIGKRRFIEALETELEKPPKTAIQKANI  
 YGFCFAFQACIMFIANSASYRYGGYLSNEGLHFSYVFRVIVSAVLSATALGRAFSYTPSYAKAKISAAR  
 FFQLLRQPPISVYNTAGEKWDNFQKIDFVDCKFTYPSRPDSQVLNGLSVISPGQTLAFVGGSSCGKKS  
 TETATTPEYVDRDCKMTCINDEYVUNHPTDENTCTICGDSPEACTMNTYVYVDRYKPTDMDUTA

ORIGIN  
 1 atcettatc tatatctctt ggtatttagt gtaaaaaatt taaaatcttt acctagcaat  
 61 cttgaggaag aaacttgata actgataata catgagattt ttacctaatg gaaatataat  
 121 cctatatatt caacaaactt tagaagaata agataaaatt taaagtaaat gactctctgta  
 181 gttttataga tctctcaaac caatctagtc tcagatctca ccttccatcat ttctctcatt  
 241 tctctttggc ctaattaatc aaaaatcttc ctagaatggt cattctctgc cagtatgtct  
 301 tcttgaatat gaataagaaa taaaatacca tttgatggtt gaaatattgg gggtaatctc  
 361 aatgacgaaa atagatgacc aggaaaaggg aaacgaatgc ctgattcatt atattcattga  
 421 agatatcaaa ggtttataaa gccaatatct gggaaagaga aaaccgtgag acttccagat  
 481 cttctctggt gaagtgtggt tcttgcacac atcaagaac tcaacatcaa aggatgccaa  
 541 tggaaaagggt cctctctgct getgetggtt tcaaacctgc tectgtgcca gagegtggcc  
 601 cctctgcccc tetgtccggg cggggctgcc cgatgccagg tgacctcteg agacctgttt  
 661 gaccgcggcg tegtctctgc ccactacatc cataacctct cctcagaatg gttcagcgaa  
 721 ttcgataaac ggtataccca tggccggggg ttcattacca aggccatcaa cagctgccac  
 781 acttctctccc ttgccacccc cgaagacaag gagcaagccc aacagatgaa tcaaaaagac  
 841 tttctgagcc tgatagtcag catattgca tcttggaaatg agcctctgta tcatctggtc  
 901 acggaagtac gtggtatgca agaagccccg gaggctatcc tatccaagac tgtagagatt  
 961 gaggagcaaa ccaaacggct tctagagggc atggagctga tagtcagcca ggttcatcct  
 1021 gaaaccaaag aaatgagat ctacctgttc tggtegggac ttccatcctt gcagatggct  
 1081 gatgaagagt ctgcctcttc tgcattatc aaacctgtcc actgctctag cagggattea  
 1141 cataaaatcg acaattatct caagctcctg aagtgcgcaa tcatccacaa caacaactgc  
 1201 taagcccaca tccatttcat ctatttctga gaaggtcctt aatgatccgt tccattgcaa  
 1261 gcttctttta gttgatctc ttttgaatcc atgcttgggt gtaaccagtc tctctttaa  
 1321 aaataaaaac tgactcctta gagacatcaa aatctaaaa

//

atg Find 12 of 25 NM\_000948 : 520-522

[AG]-x(4)-G-K-[ST] Find 1 of 3 NP\_523824 : 66-73

## Recent NCBI Publication on dbVar

The NCBI Database of Genomic Structural Variation (dbVar) announced in the February 2010 issue of the NCBI News was described in a recent correspondence article in *Nature Genetics*.

Church DM, Lappalainen I, Sneddon TP, Hinton J, Maguire M, Lopez J, Garner J, Paschall J, DiCuccio M, Yaschenko E, Scherer SW, Feuk L, Flicek P. **Public data archives for genomic structural variation.** *Nat Genet.* 42, 813-814 (2010). PMID: 20877315

The publication and resource are also announced in a recent [NIH press release](#).

dbVAR provides information on large scale (> 1Kb) variations in genome sequences and includes copy number variants and other deletions and insertions.

## Epigenomics Resource

A new Epigenomics resource ([www.ncbi.nlm.nih.gov/epigenomics](http://www.ncbi.nlm.nih.gov/epigenomics)) is part of the Entrez search and retrieval system. Epigenomics is a new field of research with the goal of understanding how different cell types and lineages acquire distinct patterns of gene expression. The Epigenomics database contains results of genome-wide studies on modifications of chromatin (histone modification, DNA methylation, DNAase footprinting) in various cell types. These studies assay programmable changes that affect gene expression (epigenetics). A Sample Browser provides access to experimental data. Data can be filtered based on biological attributes such as cell type, differentiation stage, health status, and others. Data may be displayed on the genome using either the UCSC genome browser or NCBI graphical sequence viewer as shown below.

**Term Filter** Containing word(s):  Filter

**Attribute Filter** ▼

Species	Cell Type	Lab
All	centroblast	All
Arabidopsis thaliana	centrocyte	Biotech Research and Innovation Centre/
Caenorhabditis elegans	embryonic kidney cell	Center for Genomic Regulation
Drosophila melanogaster	embryonic stem cell	Cold Spring Harbor Laboratory
Homo sapiens	epithelial cell	Dana-Farber Cancer Institute
Mus musculus	erythrocyte progenitor cell	Laboratory of Molecular Immunology/Nati
	fibroblast	University of Michigan

**Samples**

View on Genome Download

Sample ID	Species	Cell Type	Tissue Type	Cell Line	Cell Population	Differentiation State
<input checked="" type="checkbox"/> ESM000113	Homo sapiens	epithelial cell	mammary gland	MDA-MB-231		
<input type="checkbox"/> ESM000262	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000263	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000264	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000330	Homo sapiens	epithelial cell		SH-SY5Y		
<input type="checkbox"/> ESM000331	Homo sapiens	epithelial cell		HeLa S3		
<input type="checkbox"/> ESM000366	Homo sapiens	epithelial cell		VCaP		
<input type="checkbox"/> ESM000367	Homo sapiens	epithelial cell		VCaP		
<input type="checkbox"/> ESM000368	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000369	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000370	Homo sapiens	epithelial cell		VCaP		
<input type="checkbox"/> ESM000371	Homo sapiens	epithelial cell		LNCaP		
<input type="checkbox"/> ESM000372	Homo sapiens	epithelial cell		LNCaP		

NC\_000006.11 (171,115,067 bases)

Sequence Set Origin Views & Tools Markers Search...

151,089,060 : 151,445,269 (356,210 bases shown, positive strand)

Sequence Flip Strands Tools

151,100 K 151,150 K 151,200 K 151,250 K 151,300 K 151,350 K 151,400 K 15

- Sequence NC\_000006.11: Homo sapiens chromosome 6, GRCh37 pr

a: H3K36me2, Max:188 / Min:0 - Human breast cancer epithelial cell, MDA-MB-231 cell line, mammar

a: H3K36me3, Max:145 / Min:0 - Human breast cancer epithelial cell, MDA-MB-231 cell line, mammar

- Genes

**150465793..152068537**  
 name: NA000000603.1  
**Title:** Human breast cancer epithelial cell, MDA-MB-231 cell line, mammary gland, adenocarcinoma : H3K36me3  
**Comment:** High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing Submitted By: Cold Spring Harbor Laboratory Global DNA methylation is examined by bisulfite-seq in both human primary and cancer cell lines. H3K4me2 and H3K36me3 are also examined.

You are here: NCBI > DNA & RNA > Nucleotide Database Write to the Help Desk

## Microbial Genomes

Eleven finished microbial genomes were released during September 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

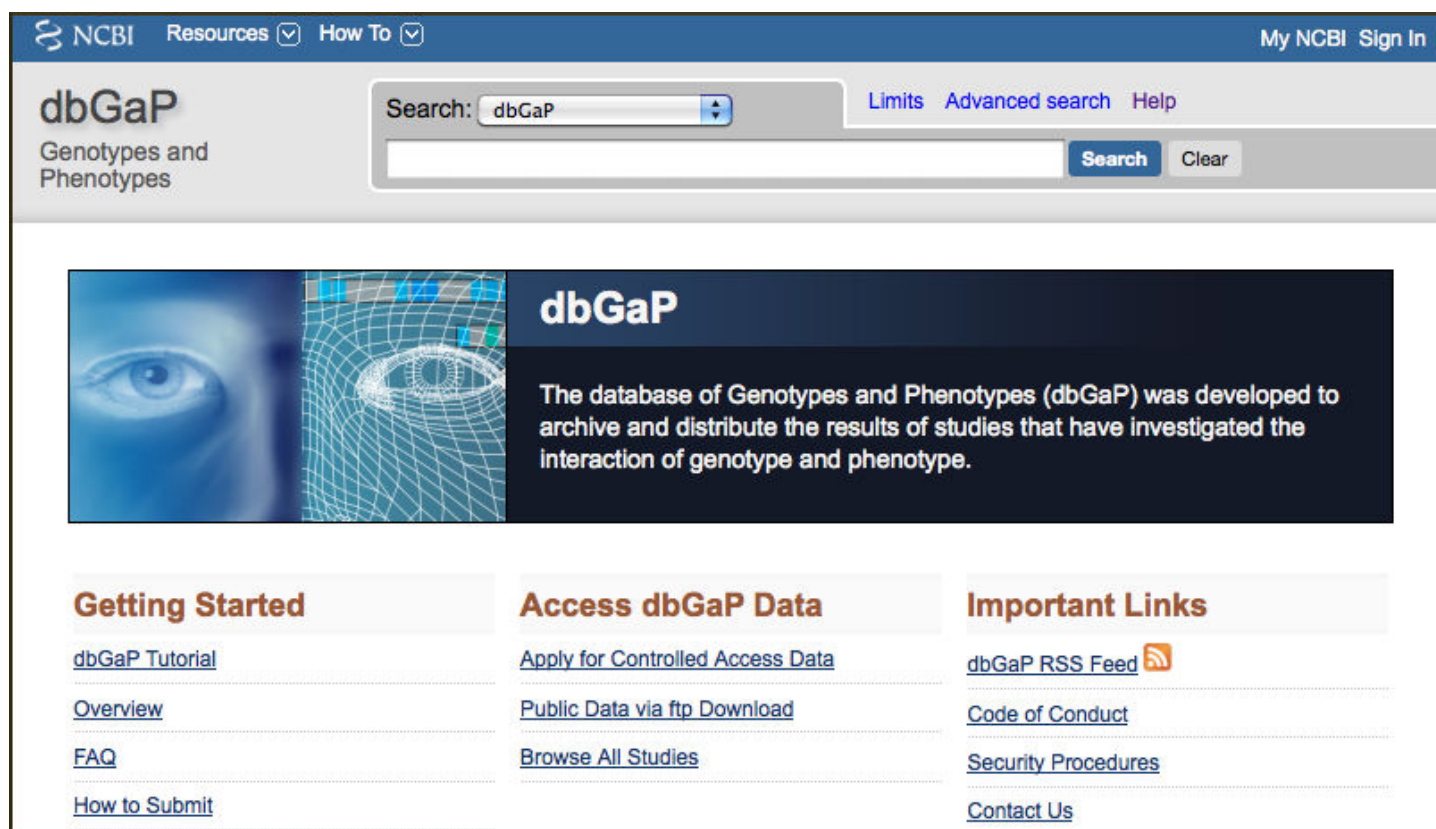
## GenBank News

GenBank release 179.0 is available via web and FTP. The current release includes information available as of August 16, 2010. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### dbGaP

The Entrez dbGaP resource has migrated to the streamlined discovery-oriented design that has been in service in PubMed for nearly a year, described fully in the [November 2009](#) issue of the NCBI News. Included in the re-design is a new home page ([www.ncbi.nlm.nih.gov/gap](http://www.ncbi.nlm.nih.gov/gap)), Limits and Advanced Search page. The image of the human face on the homepage, shown immediately below, is a composite of faces from approximately 10,000 people.



NCBI Resources How To My NCBI Sign In

dbGaP Genotypes and Phenotypes

Search: dbGaP Limits Advanced search Help

Search Clear

**dbGaP**

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the results of studies that have investigated the interaction of genotype and phenotype.

**Getting Started**

- [dbGaP Tutorial](#)
- [Overview](#)
- [FAQ](#)
- [How to Submit](#)

**Access dbGaP Data**

- [Apply for Controlled Access Data](#)
- [Public Data via ftp Download](#)
- [Browse All Studies](#)

**Important Links**

- [dbGaP RSS Feed](#)
- [Code of Conduct](#)
- [Security Procedures](#)
- [Contact Us](#)

### PubMed DTDs

PubMed E-Utility 2011 DTDs will be updated in mid-December, approximately on December 13. The new DTDs can be found:

[http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed\\_110101.dtd](http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/pubmed_110101.dtd)

## RefSeq

RefSeq Release 43 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of September 16. It includes 11,223,078 records from 10,854 different species and strains. Changes since the last release are described in the release notes (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release43.txt>). More information on the RefSeq project is available on the RefSeq Homepage ([www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/)).

## Entrez Gene Record Status

Entrez Gene ([www.ncbi.nlm.nih.gov/gene](http://www.ncbi.nlm.nih.gov/gene)) has two new properties field terms to identify the status of records that are no longer current (alive). Replaced records, retrieved with the search term “replaced”[Properties], are those that have been made secondary to (merged with) another gene record while discontinued records, retrieved with “discontinued”[Properties], are no longer current but not identified with another record. As before, current records can be retrieved with the term “alive”[properties] or by using “current only”[Filter].

## NCBI Workshop at ASHG Annual Meeting

NCBI scientists will present a special workshop at the American Society for Human Genetics meeting on November 4 at 7:15 p.m. The workshop entitled, “[A Practical Guide to Genome-Scale Data at NCBI](#)”, will provide information on genome-scale resources available at NCBI including finding and downloading data, analysis, and management of data sets. NCBI will also staff an exhibit booth at the meeting.

### Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook and Twitter: [twitter.com/NCBI](https://twitter.com/NCBI).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, August 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: August 31, 2010.

### New Databases and Tools

#### NCBI Publication

Cooper P, Lipshultz D, Matten WT, McGinnis SD, Pechous S, Romiti ML, Tao T, Valjavec-Gratian M, Sayers EW. **Education resources of the National Center for Biotechnology Information.** *Brief Bioinform.* 2010 Jun 22 [Epub ahead of print]. PMID: 20570844

#### Microbial Genomes

Twenty finished microbial genomes were released during August 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

### GenBank News

GenBank release 179.0 is available via web and FTP. The current release includes information available as of August 16, 2010. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

### Updates and Enhancements

#### PubMed

In July 2010, PubMed hit a milestone by adding its 20 millionth citation. Ten million PubMed citations have links to the full-text version of the article.

#### Genome Workbench

Genome Workbench version 2.1.2 was released with improvements and fixes. New features include support for 64-bit Windows 7 allowing researchers to study bigger, more complicated sequences; revamped network setup to be more tolerant of restrictive firewalls; more robust BLAST implementation; and new rendering for variation tracks. In addition, the application has been tuned to reduce memory consumption. Complete release details can be found in the Release Notes: [www.ncbi.nlm.nih.gov/projects/gbench/release-notes.html](http://www.ncbi.nlm.nih.gov/projects/gbench/release-notes.html)

#### BLAST

BLAST executable version 2.2.4 was released with new features as well as bug fixes. For example, the BLAST Archive format was introduced to permit reformatting of stand-alone BLAST searches with the blast-formatter. For a complete list of fixes and features, see the [BLAST News](#) page and follow links to download the application.

### Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on

the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook and Twitter: [twitter.com/NCBI](https://twitter.com/NCBI).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, July 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: July 31, 2010.

### Featured Resource: Updated Entrez Sequence Database Interfaces

The Entrez sequence databases (Nucleotide, Protein, GSS, and EST) have recently completed migration to the streamlined discovery-oriented design that has been in service in PubMed for nearly a year, described fully in the November 2009 issue of the NCBI News. The sequence database re-design includes new homepages, a simpler interface, and new options for downloading, displaying sequences, and connecting to related data.

#### New homepages

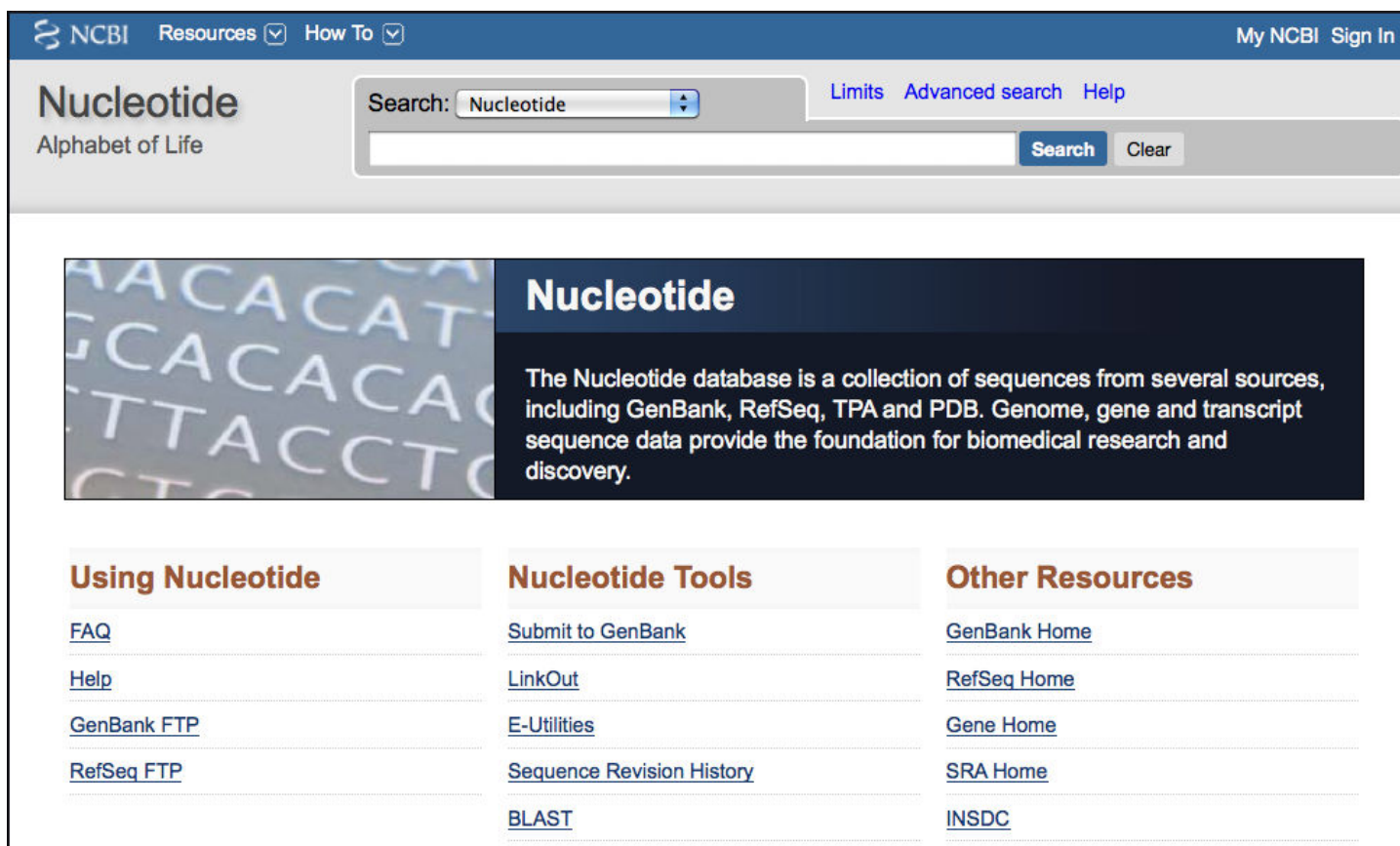
The sequence database homepages now have a simplified design with three columns of links with access to information about using the resource, tools for submitting data, searching and analysis, and other related resources at the NCBI site. As in the PubMed and the Site Guide (NCBI Homepage), the new sequence database pages have the new header including the search bar with access to all NCBI Entrez databases and the *Resources* and *How To* pull-down lists to aid navigation and to access to practical task-oriented help and the footer that provides rapid navigation to all major areas of the NCBI site. Figure 1 shows the new Entrez nucleotide homepage.

#### Improvements to the search interface

The new interface is simplified eliminating the four control tabs of the previous version: *Limits*, *Preview/Index*, *History*, *Clipboard*, and *Details*. These functions are still available but are now easier to access and use. A redesigned *Limits* page is linked at the upper-right of the *Search Box* that appears on all pages. The new *Advanced search* page also linked above the *Search Box* combines the functions of the old *Preview/Index* and *History* tabs. The *Search details* providing the translation of the query by the Entrez engine is linked on the *Advanced search* page but is also shown in the right-hand *Discovery* column as in the new PubMed. The *Clipboard*, when populated, is now accessible as a link at the top of the *Discovery* column as described below.

#### Using the *Advanced search* page

Figure 2 shows the *Advanced search* page for Entrez protein. The page functions as an independent search interface that allows formulation of complex queries. The *Search builder* and *Search history* sections replace the previous *Preview/Index* page and facilitate the construction of more precise queries. The pull-down list in the *Search Builder* shows all of the fields indexed for a particular database. The *Show Index* link opens an alphabetical list of terms for the selected field. When a term is entered in the *Search Builder*, the index will open to the closest match in the index. The *Add to Search Box* button puts the field-restricted queries into the *Search Box*. These may be run using the *Search* button or may be added to the *Search History* using the *Preview* button. Entries in the *Search History* may be combined to give very precise results. The example in Figure 2 combines searches for frogs (#5), RefSeq proteins (#2), and prolactin (#3) to obtain the prolactin protein records for *Xenopus laevis*, NP\_001086486, and *Xenopus (Siluriana) tropicalis*, NP\_001093699).



NCBI Resources How To My NCBI Sign In

**Nucleotide**  
Alphabet of Life

Search: Nucleotide Limits Advanced search Help

Search Clear

**Nucleotide**

The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.

**Using Nucleotide**

- [FAQ](#)
- [Help](#)
- [GenBank FTP](#)
- [RefSeq FTP](#)

**Nucleotide Tools**

- [Submit to GenBank](#)
- [LinkOut](#)
- [E-Utilities](#)
- [Sequence Revision History](#)
- [BLAST](#)

**Other Resources**

- [GenBank Home](#)
- [RefSeq Home](#)
- [Gene Home](#)
- [SRA Home](#)
- [INSDC](#)

**Figure 1. The new nucleotide homepage with access to related resources.** The new sequence database homepages include the NCBI header and footer (not shown) that provide easy navigation to other parts of the site and links to task-oriented help documentation.

## Search results and other multiple record pages

Search result pages showing document summaries and all other multiple records views now incorporate fully the features of the PubMed redesign. Figure 3 shows the new search results for a protein search – prolactin[protein name] AND (Birds OR Mammals). The document summaries are presented in a new format with the record title first and hyperlinked to the record view. The summary now also shows the length of the sequence. The sequence identifiers, accession version and GI number, are listed below the summary. The old links menu has been removed from the document summaries in search results. However links to retrieve related sequences (similar by BLAST) are shown under the summary by default in the nucleotide and protein databases. The protein summaries also include a link to identical sequences. All links available for each record can be displayed if desired though individual settings in a My NCBI account. See the [My NCBI Help](#) manual for more information on customizing these preferences.

Multiple-record displays in one of the full-record formats such as GenBank or FASTA have the same set of controls in the new style as the summaries. The only additional feature is the *Customize view* control, described for single sequence views in the March 2009 NCBI News. In the case of multiple records, this control allows toggling the reverse complement of the displayed sequences.

## New items in the right-hand Discovery column

The search filters that formerly were a series of tabs are now implemented as a set of links at the upper part of the right-hand Discovery column. In the case of the results shown in Figure 3, the RefSeq filter has been clicked filtering the output to show only the six Reference Sequences. Clicking the plus sign (+) at the right of the

**Protein Advanced Search** [« Back to Protein](#)

**Search Box** [Limits](#) [Details](#) [Help](#)  
 ((#2) AND #5) AND #3 [Search](#) [Preview](#) [Clear](#)

**Search Builder**  
 All Fields [AND](#) [Add to Search Box](#)  
[Show Index](#)

[Search Builder Instructions](#)

**Search History**

Search	Most Recent Queries	Time	Result
#16	Search ((#2) AND #5) AND #3	17:02:27	<a href="#">2</a>
#15	Search creatine kinase[Protein Name]	14:19:27	<a href="#">149</a>
#14	Search creatine kinase	14:13:08	<a href="#">986</a>
#5	Search frogs	14:01:43	<a href="#">106353</a>
#4	Search (#2) AND #3	13:58:27	<a href="#">10</a>
#3	Search "srcdb refseq"[Properties]	13:57:38	<a href="#">10929822</a>
#2	Search "prolactin"[Protein Name]	13:56:37	<a href="#">360</a>
#1	Search guinea pig[organism]	13:54:22	<a href="#">1692</a>
#0	protein clipboard	13:58:26	<a href="#">10</a>

[Less History](#) [Clear History](#)

Protein Name creatine kinase [AND](#) [Add to Search Box](#)

creatine kinase (149)  
 creatine kinase 1 (5)  
 creatine kinase 1 2 (1)  
 creatine kinase 2 (4)  
 creatine kinase 2 2 (1)  
 creatine kinase 28aa (1)  
 creatine kinase 3 (1)  
 creatine kinase 4 (1)  
 creatine kinase a (2)  
 creatine kinase b (11)

[Show Index](#)  
[Previous 200](#)  
[Next 200](#)  
[Close Index List](#)

**Figure 2. Entrez protein Advanced search page.** The Search Builder function and Search History replace the former Preview/Index page. The pull-down list shows all indexed fields for the database. The Show Index link (green arrow) expands the display to show all terms indexed for a particular field. The bottom panel shows the index for the Protein Name field matching the search term creatine kinase. Entries from the Search Builder and the Search History can be combined in the Search Box to construct complex queries. Clicking on the numbered entries in the Search History provides Options for combining searches, removing History entries, loading results, showing queries, and saving the search in My NCBI. Combining #2, #5 and #3 finds the two *Xenopus* RefSeq proteins for prolactin.

selected filter will add the filter as a term to the search. The Search details, previously a control tab that shows how the query is interpreted or mapped to Entrez controlled vocabularies is now exposed in the Discovery column. The Search details at the bottom right of Figure 3 show how the terms Birds and Mammals were expanded, translated, and mapped to the organism field (NCBI Taxonomy) to generate more accurate results. The former Clipboard tab now appears as a link above the right-hand column only if the NCBI Clipboard contains items (Figure 3, upper right).

Also in the Discovery column are two new items, *Analyze these sequences* and *Find related data* that provide access to analysis tools and pre-computed relationships – true Discovery components. *Analyze these sequences*, available for displays containing 20 or fewer records, provides direct access to the NCBI BLAST service from all sequence databases and, for multiple protein records, the NCBI multiple-alignment tool COBALT, described in the May 2009 NCBI News. This is a convenient interface to COBALT since it allows direct submission after using

Display Settings: Summary, 20 per page, Sorted by Default order      Send to:      Clipboard: 265 items

### Results: 6

- [prolactin precursor \[Oryctolagus cuniculus\]](#)  
1. 227 aa protein  
NP\_001076144.1 GI:130504953  
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)
- [prolactin \[Ovis aries\]](#)  
2. 240 aa protein  
NP\_001009306.1 GI:57164329  
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)
- [prolactin precursor \[Equus caballus\]](#)  
3. 229 aa protein  
NP\_001075365.1 GI:126352562  
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)
- [prolactin precursor \[Bos taurus\]](#)  
4. 229 aa protein  
NP\_776378.2 GI:46810277  
[Related Sequences](#) [Identical Proteins](#) [Item in clipboard](#)
- [prolactin precursor \[Homo sapiens\]](#)  
5. 227 aa protein  
NP\_001157030.1 GI:254675133  
[Related Sequences](#) [Identical Proteins](#)
- [prolactin precursor \[Gallus gallus\]](#)  
6. 229 aa protein  
NP\_990797.1 GI:49169789  
[Related Sequences](#) [Identical Proteins](#)

Display Settings: Summary, 20 per page, Sorted by Default order      Send to:

Prolactin[protein name] AND (Birds OR Mammals)

**Find related data**

Database: Nucleotide

Option: **Encoding mRNA**

Link from: Nucleotide, Mature Peptide, Order cDNA Clone, **Encoding mRNA**, WGS Project

Find it

**Filter your results:**

[All \(265\)](#)

Bacteria (0)

[Related Structures \(198\)](#)

**RefSeq (6)**

[Manage Filters](#)

---

**Top Organisms [Tree]**

Oryctolagus cuniculus (1)

Homo sapiens (1)

Bos taurus (1)

Ovis aries (1)

Equus caballus (1)

All other taxa (1)

[More...](#)

---

**Analyze these sequences**

Run BLAST

Align sequences with COBALT

---

**Find related data**

Database: Select

[Find items](#)

---

**Search details**

```

prolactin[protein name] AND
(("Aves"[Organism] OR
Birds[All Fields]) OR
("Mammalia"[Organism] OR
Mammals[All Fields]))

```

[Search](#)      [See more...](#)

**Figure 3.** Entrez protein search results for the query `Prolactin[protein name] AND (Birds OR Mammals)` showing the new summary style. The Discovery column now has the search filters, *Search details*, the analysis tool BLAST and COBALT, and the *Find related data* device that displays linked records related to the present set of results. The inset shows the Encoding mRNA selection that retrieves the corresponding RefSeq mRNA records for these proteins. The *Search details* show how the terms Birds and Mammals were mapped to the controlled vocabulary of the indexed Organism field.

Entrez to collect the desired input set. For example the set of homologous prolactin proteins from mammals and birds in Figure 3 can now be easily aligned using COBALT.

The links shown in the *Find related data* list previously were presented on the Display pull-down list at the top of the results. The new location in the Discovery column makes these links easier to find and use. These provide access to the wealth of pre-computed and pre-compiled relationships that make Entrez a powerful discovery system. In many cases there may be more than one kind of link to another database. The inset in Figure 3 shows

the multiple links to the nucleotide database available from the protein data. Selecting “Encoding mRNA” and clicking *Find items* links to the six corresponding RefSeq mRNA records.

### **Display Settings and Send to menus**

The *Display settings* and *Send to* menus are now accessed through links at the left (top and bottom) and right (top and bottom) respectively of single- and multiple-record displays (Figure 3). The upper panel of Figure 4 shows the expanded *Display settings* menu that provides the ability to select any of the record formats available for the database, to modify the number of records displayed, and to sort by publication, release date, accession, or organism. As in PubMed, My NCBI now allows the default settings for these options in the sequence database to be set to different values if desired. This new option in My NCBI is described in more detail in the Updates and Enhancements section of the current NCBI News. The *Send to* menu provides options for saving items to the NCBI clipboard, Collections in My NCBI, or to download sequence records in various formats to a local file (Figure 4, lower left). With the new mechanism for saving records it is no longer necessary to display the record in the desired format before downloading.

### **Nucleotide Send menu and Coding Sequence Download**

For nucleotide records the *Send* menu allows downloading either the complete record or all annotated coding sequence regions as FASTA formatted sequences directly from the parent records. The feature allows downloading either the nucleotide sequences or the corresponding protein translation. This often-requested feature works with annotated CDS features on any nucleotide record from simple open reading frames on mRNA sequences to complex multi-exon genes on mammalian genomic regions. Each downloaded CDS has its own structured title that includes a unique identifier incorporating the parent sequence accessions, gene symbol, protein product, reading frame, protein identifier, and location on parent sequence. This CDS download function can be used to quickly create a local database of sequences for analysis.

## **Summary**

The updated Entrez interface now implemented in the sequence databases provides a streamlined and less complex search interface as well as improved consistency of form and function across the molecular and literature resources making the NCBI site easier to use. New options for displaying and downloading records especially the ability to download coding regions are important improvements to the utility and flexibility of the molecular biology Web services at the NCBI. The presence of analysis tools and improved access to related data on search results and record displays increase the power of Entrez as a system for scientific discovery.

## **New Databases and Tools**

### **Bibliography Management**

My NCBI will replace eRA Commons for grantee bibliography management. The *My Bibliography* function in My NCBI will link with Commons to allow scientists to maintain and manage a list of their publications including journal articles, book chapters, meeting abstracts, talks and presentations, patents, and other materials. Journal articles can be those found in PubMed, journals not indexed or not yet appearing in PubMed, or manuscripts submitted to NIHMS. Citations may not be entered manually into eRA Commons at this point, and users must use My Bibliography to manage their bibliographies. For more information, please see the [NIH Announcement](#).

### **NCBI Discovery Workshops**

NCBI will present a two-day workshop on September 29-30, on the NIH campus in Bethesda, MD. The workshops provide hands-on experience exploring practical examples using tools and databases on the NCBI

The image shows a series of overlapping windows from a web application. At the top is a 'Display Settings:' dropdown menu. Below it is a settings panel with three columns: 'Format', 'Items per page', and 'Sort by'. The 'Format' column has radio buttons for Summary (selected), GenPept, GenPept (full), FASTA, FASTA (text), ASN.1, Revision History, Accession List, and GI List. The 'Items per page' column has radio buttons for 5, 10, 20 (selected), 50, 100, and 200. The 'Sort by' column has radio buttons for Default order, Accession, Date Modified (selected), Date Released, Organism Name, and Taxonomy ID. An 'Apply' button is at the bottom right of this panel. Below the settings panel are two 'Send to:' dropdown menus. The left one is open, showing a 'Choose Destination' section with radio buttons for File (selected) and Clipboard, and a 'Collections' option. Below this is a list of formats: GenPept (highlighted), GenPept (full), FASTA, ASN.1, XML, INSDSeq XML, TinySeq XML, Feature Table, Accession List, and GI List. A 'Download 6 items.' message and a 'Format' dropdown (set to GenPept) are also visible, along with a 'Create File' button. The right 'Send to:' dropdown is also open, showing 'Complete Record' (selected) and 'Coding Sequences' options. Below this is another 'Choose Destination' section with radio buttons for File and Clipboard, and a 'Collections' option. A 'Download features.' message and a 'Format' dropdown (set to FASTA Nucleotide) are visible, with a 'Create File' button and a dropdown menu showing 'FASTA Nucleotide' (highlighted) and 'FASTA Protein' options.

**Figure 4.** *Display settings and Send to menu for the Protein database.* These menus have equivalent options in all sequence databases for multiple and single record displays. The *Send* menu that replaces the *Send to* menu for nucleotide records with annotated coding regions has an option to download either the complete record or coding regions in FASTA nucleotide or protein.

website. The four workshops are Sequences, Genomes, and Maps; Proteins, Domains and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. For more information see the [Discovery Workshop](#) page, which also includes a registration link.

## Microbial Genomes

Twenty-one finished microbial genomes were released in July 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 178.0 is available through the NCBI web and FTP sites. The current release includes information available as of June 11, 2010. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### My NCBI now offers sequence database preferences

My NCBI allows Preferences to be set for record format and result display settings for the Entrez sequence databases (Nucleotide, Protein, EST and GSS). Record formats can be changed from the default (GenBank, GenPept, EST, GSS) to FASTA or graphics for Nucleotide and Protein or to GenBank or FASTA for EST or GSS records. All available options on the *Display settings* menu in the sequence databases can also be changed from the default settings. The relevant preferences dialogs from My NCBI are shown below.

The screenshot shows the My NCBI interface. At the top, there is a navigation bar with links for Home, PubMed, GenBank, and BLAST, along with a user profile for 'pscooper' and options for Sign Out and My NCBI. Below this is the 'My NCBI' logo and a subtitle: 'A division of the National Library of Medicine at the National Institutes of Health'.

The main content area is divided into several preference sections:

- Nucleotide Preferences:** Record Display Format (GenBank), Result Display Settings (200, Date Modified).
- Protein Preferences:** Record Display Format (GenPept).
- GSS Preferences:** Record Display Format, Result Display Settings.
- EST Preferences:** Record Display Format, Result Display Settings.

Two pop-up windows are overlaid on the Protein Preferences section:

**Result Display Settings for Protein (top):** Set the default values for sorting order and number of items to be displayed per page in Protein:
 

- Default items per page:** Radio buttons for 5, 10, 20 (selected), 50, 100, 200.
- Default sort by:** Radio buttons for Default Order (selected), Date Modified, Organism Name, Accession, Date Released, Taxonomy ID.
- A green **Save** button is at the bottom left.

**Record Display Settings for Protein (bottom):** Set the default values for displaying record in Protein:
 

- Default record format:** Radio buttons for GenPept (selected), FASTA, Graphics.
- A green **Save** button is at the bottom left.
- At the bottom right, it says 'Or cancel and return to the [preferences page](#)'.

## RefSeq

RefSeq Release 42 is available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of July 21, 2010. It includes 15,038,858 records from 10,728 different organisms. Changes since the last release can be found in the release notes (<ftp.ncbi.nih.gov/refseq/release/release-notes/RefSeq-release42.txt>)

## New FTP file for Gene

Entrez Gene calculates matches between NCBI and Ensembl annotations and reports the matches in the Entrez gene Full Report display, the "matches Ensembl" index property, and the gene2ensembl FTP file. A new FTP file, README\_ensembl, will soon be added to provide a summary of species whose annotations have been compared, including release and assembly information, and the date when the comparison was last performed. A complete description of the file is on the ftp site: <ftp.ncbi.nih.gov/gene/DATA/README>

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on



the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook and Twitter: [twitter.com/NCBI](https://twitter.com/NCBI).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, June 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: June 23, 2010.

### New Databases and Tools

#### Selected Structures

A new Selected Structures filter is available with Entrez Structure (MMDB) search results. This feature provides a way of sorting results into subsets based on characteristics of the retrieved records. Available subsets are keyed to protein domain families, source organisms, the presence of specific molecular complexes, and the presence of links to literature from PubMed or PubMed Central. The Selected Structures filter links appear in a box in the upper right corner of the Structure search result pages. Clicking on any of the linked numbers produces a subset of the results with the listed property. A portion of the results of a search for “p53 tumor suppressor” in the Structure database with the Selected Structures feature is shown below. Clicking on the filter for those records with protein-DNA complexes (green arrow) narrows the results to the four structures that have bound DNA (lower panel).

Items 1 - 58 of 58 One page.

1: 2GEQ [Related Structures, Literature, Domains, Chemicals, Other Links](#)



Crystal Structure Of A P53 Core Dimer Bound To Dna [TranscriptionDNA]  
 Taxonomy: [Mus musculus](#), [synthetic construct](#)  
 Proteins: 2; Nucleic acids: 2 (DNA); Chemicals: 2  
 modified: 2009/09/04; MMDB ID: 76523

2: 3GLU [Related Structures, Literature, Domains, Chemicals, Other Links](#)



Crystal Structure Of Human Sirt3 With Acecs2 Peptide [HydrolaseHYDROLASE  
 REGULATOR, EC: 3.5.1.-]  
 Taxonomy: [Homo sapiens](#)  
 Proteins: 2; Chemicals: 2  
 modified: 2009/09/17; MMDB ID: 73353

3: One page.

1: 2GEQ [Related Structures, Literature, Domains, Chemicals, Other Links](#)



Crystal Structure Of A P53 Core Dimer Bound To Dna [TranscriptionDNA]  
 Taxonomy: [Mus musculus](#), [synthetic construct](#)  
 Proteins: 2; Nucleic acids: 2 (DNA); Chemicals: 2  
 modified: 2009/09/04; MMDB ID: 76523

2: 1BF5 [Related Structures, Literature, Domains, Other Links](#)



Tyrosine Phosphorylated Stat-1DNA COMPLEX [Gene RegulationDNA]  
 Taxonomy: [Homo sapiens](#), [synthetic construct](#)  
 Proteins: 1; Nucleic acids: 2 (DNA)  
 modified: 2009/07/16; MMDB ID: 71645

3: 1TUP [Related Structures, Literature, Domains, Chemicals, Other Links](#)



Tumor Suppressor P53 Complexed With Dna [Antitumor ProteinDNA]  
 Taxonomy: [Homo sapiens](#), [synthetic construct](#)  
 Proteins: 3; Nucleic acids: 2 (DNA); Chemicals: 1  
 modified: 2007/10/13; MMDB ID: 51571

4: 1TSR [Related Structures, Literature, Domains, Chemicals, Other Links](#)



P53 Core Domain In Complex With Dna [Antitumor ProteinDNA]  
 Taxonomy: [Homo sapiens](#), [synthetic construct](#)  
 Proteins: 3; Nucleic acids: 2 (DNA); Chemicals: 1  
 modified: 2007/10/13; MMDB ID: 51561

**Selected Structures** Structure Count

Protein Domain Families

Families

- P53: P53 DNA-binding domain 42
- SIRT1 20
- RPA1N 16
- HECTc 3
- ANK 2
- ... All 8 Families 1

Superfamilies

- P53: P53 DNA-binding domain 57
- ... All 8 Families 20

Complexes

- Protein-Protein 47
- Protein-DNA 4 →
- Protein-Chemical 39

Literature

- PubMed 57
- PMC 23
- Taxonomy 58
- Homo sapiens, (human), species, primates 37
- Synthetic construct, species, other sequences 14
- Saccharomyces cerevisiae, (baker's yeast), species, ascomycetes 11
- Danio rerio, (zebrafish), species, bony fishes 3
- Mus musculus, (house mouse), species, rodents 2
- ... All 10 Organisms

See the [structure help documentation](#) to read more about Selected Structures.

## New Nucleotide and Protein Pages

The [Nucleotide](#) and [Protein](#) pages have been updated with a new format and display settings. One important new feature is the ability to download coding sequence (CDS) regions as FASTA formatted sequences directly from the parent records. CDS downloads can be accessed using the new “Send” menu in the upper right of single or multiple record displays. See the [note](#) on the NCBI Facebook page for additional details.

## Taxonomy Links to Wikipedia

Over 52,000 NCBI Taxonomy pages now contain links to Wikipedia articles through the External Information Resources section (LinkOut). To see an example, visit the taxonomy record for *Saimiri oerstedii*, the Central American Squirrel monkey. You can follow the link to the [Wikipedia article](#) about this species. These new LinkOuts are provided through the [iPhylo project](#).

## Microbial Genomes

Twenty-one finished microbial genomes were released in June 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 178.0 is available through the web and FTP sites. The current release includes information available as of June 11, 2010. The Release Notes provide more details: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### Genome Workbench

Genome Workbench version 2.1.0 is available with numerous fixes and enhancements, giving researchers an even more powerful tool. Enhancements include an improved tree viewer used for phylogenetic tree analysis, integration of the MUSCLE alignment tool, and numerous updates to graphical views. [Release notes](#) for this version provide details on new features and bug fixes.

Genome Workbench is available for Windows, Linux, and Mac OS X and can be downloaded from the homepage: [www.ncbi.nlm.nih.gov/projects/gbench/](http://www.ncbi.nlm.nih.gov/projects/gbench/).

### PubMed Central Turns Ten

PubMed Central celebrated its 10<sup>th</sup> Anniversary this year. Created as a free, online archive in 2000, it has grown to contain two million articles from 37,000 journals, over 600 of which are full deposit journals. A complete history has been compiled in the May-June issue of the [NLM Technical Bulletin](#). NCBI celebrated the anniversary with a short symposium with speakers both from the NCBI and the outside community in May 2010.

### PubMed

Recent improvements to PubMed include an updated Advanced Search page that provides a better-integrated and more intuitive mechanism for constructing complex searches and the ability to select what data are sent through the E-mail option on the Send menu. The improved Advanced Search page and changes to the E-mail option are described in more detail in articles [e9](#) and [e10](#) in the May-June 2010 issue of the [NLM Technical Bulletin](#).

### MyNCBI

My NCBI now allows sign-in by partner organization accounts. The improvement is described in more detail in the recent [NLM Technical Bulletin](#).

### YouTube

Four How-to Tutorial videos have been added to NCBI's YouTube channel: [www.youtube.com/ncbinlm](http://www.youtube.com/ncbinlm). The following are the currently available instructional videos:

- [Download a custom set of records.](#)
- [Obtain genomic sequence for and near a gene \(short and extended versions\).](#)
- [Retrieve all sequences for an organism.](#)

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook ([www.facebook.com/ncbi.nlm](http://www.facebook.com/ncbi.nlm)) and Twitter ([twitter.com/NCBI](http://twitter.com/NCBI)).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, May 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: May 20, 2010.

### New Databases and Tools

#### NCBI Education

The NCBI [Education web page](#) has been updated with new information and a new format. The page provides a central point of access for help documents, teaching materials, news outlets, and other educational resources. In addition, NCBI has created new courses and workshops that will be offered on a limited basis.

Discovery Workshops show participants how to use the NCBI Web resources more effectively. Each workshop consists of four 2.5-hour hands-on sessions, with each session emphasizing a different group of NCBI tools and databases. Areas of study will include: Sequences, Genomes, and Maps; Proteins, Domains, and Structures; NCBI BLAST Services; and Human Variation and Disease Genes. Each year, NCBI will hold three workshops on the NIH campus and four at universities, colleges, or government research centers located in four of the eight regions of the National Network of Libraries of Medicine. For more information, see the [Discovery Workshop page](#).

Live Webinars will also be offered as short (30-60 minute) instruction modules. NCBI instructors will demonstrate effective use of various NCBI tools and resources with emphasis on recent updates and changes. Current webinar topics include: NCBI Overview; What's New at NCBI; NCBI BLAST Updates; and Genome Updates. Please see the [NCBI Webinar web page](#) for more information.

#### Microbial Genomes

Twenty-three finished microbial genomes were released during May 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

### GenBank News

GenBank release 177.0 is available via web and FTP. The current release includes information available as of April 14, 2010. Release notes are available on the on the NCBI ftp site: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>

### Updates and Enhancements

#### RefSeq

RefSeq Release 41 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of May 9, 2010. This release reached a milestone of over 10 million proteins, contained in 14,472,060 records from 10,567 different organisms.

## YouTube

NCBI now has a YouTube channel containing interviews and presentations from GenBank's 25<sup>th</sup> anniversary. The most recent post is NCBI's 20<sup>th</sup> anniversary video. Please see: <http://www.youtube.com/ncbinlm>.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also stay updated on NCBI's resources on Facebook: [www.facebook.com/ncbi.nlm](http://www.facebook.com/ncbi.nlm) and Twitter: [twitter.com/NCBI](https://twitter.com/NCBI).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, April 2010

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: April 21, 2010; Updated: May 5, 2010.

### New Databases and Tools

#### NIH Roadmap Epigenomics Project

The NIH Roadmap Epigenomics Mapping Consortium has the goal of producing a public resource of human epigenomic data to catalyze basic biological and disease-oriented research. Data from the project are now being deposited in the Gene Expression Omnibus (GEO) database at NCBI. The Epigenomics data listings page provided by GEO also allows for simple data downloading as well as data visualization in NCBI's Sequence Viewer. To view the project data listings page along with information on data access policies please see: [www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/](http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/)

#### Microbial Genomes

Twenty-five finished microbial genomes were released during April 2010. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

### GenBank News

GenBank release 177.0 is available on the NCBI web and FTP sites. The current release includes information available as of April 14, 2010. The FTP site also provides the GenBank release notes describing the current release and upcoming changes: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

### Updates and Enhancements

#### OMIM's New Look

The record display for the [Online Mendelian Inheritance in Man](#) (OMIM) database was recently updated. These changes are an aspect of the migration of OMIM to the updated display and retrieval system in Entrez. The text on the OMIM records has been re-styled to make it consistent with the rest of the Entrez system. The blue sidebar has been eliminated, and the Table of Contents, with embedded links for each section of the record, has been moved to the right-hand column. The Links menu is now displayed in a fully expanded format in this column with links to related information in other databases such as PubMed and Entrez Gene, as well as tools such as Map Viewer. OMIM also now works with the simplified URL format available in many other Entrez databases that allows direct access to a record using the appropriate identifier, in this case the MIM number. For example, the OMIM records for Breast Cancer 1 Gene, BRCA1 (MIM: 113705) can be accessed through the URL: [www.ncbi.nlm.nih.gov/omim/113705](http://www.ncbi.nlm.nih.gov/omim/113705).

#### Entrez Gene

Entrez Gene has added new links to the links menu on the right-hand side of records that provide easy access to gene-specific RefSeq sequences. The *RefSeq Proteins* link retrieves the RefSeq protein records specific to the gene being displayed. Another new link, *RefSeq RNAs*, retrieves the corresponding RefSeq transcript records. For a

subset of human genes, the *RefSeqGene* link is available that retrieves the curated RefSeqGene record spanning the annotated genomic region containing the gene.

## PubMed

PubMed now includes citations from specific online books and book chapters on the NCBI Bookshelf. Citations from *GeneReviews* and *Essentials of Glycobiology* are currently included with more books coming. The new book citations are provided by a link to the bookshelf labeled Books and Documents. In order to accommodate the book citations, the PubMed Related Articles link has been renamed Related Citations. The March-April 2010 issue of the *NLM Technical Bulletin* presents more information on the new book citations: [www.nlm.nih.gov/pubs/techbull/ma10/ma10\\_pm\\_books.html](http://www.nlm.nih.gov/pubs/techbull/ma10/ma10_pm_books.html).

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

The NCBI Facebook (<http://www.facebook.com/ncbi.nlm>) and Twitter (<http://twitter.com/NCBI>) accounts provide another source for NCBI news, updates, and events.

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, March 2010

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, MS<sup>2</sup>

Created: March 23, 2010; Updated: March 31, 2010.

### New Databases and Tools

#### Genetic Testing Registry

NIH has created the Genetic Testing Registry (GTR) that will provide access to information about the availability, validity, and usefulness of genetic tests. The NIH Office of the Director is overseeing the project. NCBI is responsible for developing the registry, expected to be fully ready in 2011. A press release with more information about GTR ([www.nih.gov/news/health/mar2010/od-18.htm](http://www.nih.gov/news/health/mar2010/od-18.htm)) and a preliminary GTR Website ([www.ncbi.nlm.nih.gov/gtr/](http://www.ncbi.nlm.nih.gov/gtr/)) are now available.

#### NCBI on Facebook and Twitter

NCBI updates and news are now on Facebook at “National Center for Biotechnology Information” and on Twitter at [twitter.com/NCBI](http://twitter.com/NCBI).

#### Bookshelf

Eighteen new reports are now on the NCBI Bookshelf in the collection of NIH-funded reports from the National Academies. To view these and other books go to: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

#### Microbial Genomes

Twenty-two finished microbial genomes were added to the NCBI databases during February. The original sequence data files submitted to GenBank/EMBL/DDBJ are on the FTP site: [ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

#### GenBank News

GenBank release 176.0 is available on the NCBI Web Service and FTP site. The current release incorporates sequence data as of February 19, 2010. Release notes containing detailed information are available on the FTP site: [ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt](ftp://ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt).

### Updates and Enhancements

#### My NCBI Update

PubMed search results can now be customized using My NCBI through the new Result Display Settings option. For detailed information, see the NLM Technical Bulletin article: [www.nlm.nih.gov/pubs/techbull/ma10/ma10\\_pm\\_results.html](http://www.nlm.nih.gov/pubs/techbull/ma10/ma10_pm_results.html)

## E-utility Usage Policy

The requirement announced in December 2009 that all E-utilities requests must contain non-null values for both the “&email” and “&tool” parameters has been relaxed somewhat for limited use of the service. The revised policy is described in detail in the E-utilities [help manual](#) on the Bookshelf.

## BLAST Release 2.2.23

BLAST version 2.2.23 of the stand-alone application is now available. Users are encouraged to move to the BLAST+ applications available at <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>. The BLAST+ programs supplant the legacy C toolkit BLAST package and will be the focus of future development. Specific enhancements for this release can be found on the [BLAST News](#) page.

## RefSeq

RefSeq Release 40 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of March 7, 2010. It includes 13,853,798 records from 10,291 different species and strains. Changes since the last release can be found in the release notes (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release40.txt>). More information on the RefSeq project is available on the RefSeq Homepage ([www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/)).

## Exhibits

NCBI will exhibit at the American Association for Cancer Research annual meeting held April 17-21 in Washington, D.C.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). For information on receiving updates about the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html).

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/).

Users can also keep up-to-date on changes, improvements and enhancements of NCBI resources on Facebook and Twitter: [twitter.com/NCBI](http://twitter.com/NCBI).

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, February 2010

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, MS<sup>2</sup>

Created: March 4, 2010; Updated: March 11, 2010.

### New Databases and Tools

#### dbVar

The newly released Database of Genomic Structural Variation (dbVar) contains data and analyses from studies on large-scale genomic variation and provides associations of defined variants with phenotype information. The [dbVar homepage](#) provides additional documentation including an overview of structural variation, a Frequently Asked Questions list, and information on submitting data. A dbVar [RSS feed](#) has also been established to provide announcements and updates.

#### NLM Press Release

NCBI scientists have assisted in the design of a computational model to accurately detect signals of regulatory elements responsible for development of the human heart and its maintenance of function. An [NIH Press Release](#) provides more information on this project.

#### Bookshelf

New books added to the Bookshelf include: *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*, *Familial Cancer Syndromes*, and *UMLS Reference Manual*. To view these and other books see: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

#### E-Utilities Documentation

The [Entrez Programming Utilities](#) (E-Utilities) Help documentation has been added to the NCBI Bookshelf. This help document has been split into chapters for better organization and is now fully integrated with the Entrez search and retrieval system as a part of the Bookshelf database.

#### Microbial Genomes

Twenty-two finished microbial genomes were added to the NCBI databases during February. The original sequence data files submitted to GenBank/EMBL/DDBJ are on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

#### GenBank News

GenBank release 176.0 is available on the NCBI Web Service and FTP site. The current release incorporates sequence data as of February 19, 2010. Release notes containing detailed information are available on the FTP site: <ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt>.

## Updates and Enhancements

### BLAST

The NCBI BLAST development team reports on improvements in the updated C++ BLAST software in the December issue of *BMC Bioinformatics*:

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009. Dec 15;10:421. PMID: [20003500](https://pubmed.ncbi.nlm.nih.gov/20003500/); PMCID: [PMC2803857](https://pubmed.ncbi.nlm.nih.gov/PMC2803857/).

The article illustrates the new BLAST+ suites' improved user interface, enhanced performance for long sequences, and better integration with the NCBI Web BLAST service.

### Exhibits

NCBI will have an exhibit booth at the American Association for Cancer Research annual meeting held April 17-21 in Washington, D.C.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, January 2010

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, MS<sup>2</sup>

Created: February 23, 2010; Updated: March 1, 2010.

### New Databases and Tools

#### Bookshelf

New books added to the Bookshelf include *Advancing the Nation's Health Needs: NIH Research Training Programs*, *The Threat of Pandemic Influenza: Are We Ready?*, *Asbestos: Selected Cancers*, and *Implications of Nanotechnology for Environmental Health Research*. View these and other books on the Bookshelf Web service ([www.ncbi.nlm.nih.gov/books/](http://www.ncbi.nlm.nih.gov/books/)).

#### NCBI Publications

Six NCBI-related articles appear in the recent Database issue of *Nucleic Acids Research*.

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. **GenBank**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D46-51. Epub 2009 Nov 12. PMID: [19910366](#)
2. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, Bryant SH. **The NCBI BioSystems database**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D492-6. Epub 2009 Oct 23. PMID: [19854944](#)
3. Ji L, Barrett T, Ayanbule O, Troup DB, Rudnev D, Muertter RN, Tomashevsky M, Soboleva A, Slotta DJ. **NCBI Peptidome: a new repository for mass spectrometry proteomics data**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D731-5. Epub 2009 Nov 26. PMID: [19942688](#)
4. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, John Wilbur W, Yaschenko E, Ye J. **Database Resources of the National Center for Biotechnology Information**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D5-16. Epub 2009 Nov 12. PMID: [19910364](#)
5. Shumway M, Cochrane G, Sugawara H. **Archiving next generation sequencing data**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D870-1. Epub 2009 Dec 3. PMID: [19965774](#)
6. Wang Y, Bolton E, Dracheva S, Karapetyan K, Shoemaker BA, Suzek TO, Wang J, Xiao J, Zhang J, Bryant SH. **An overview of the PubChem BioAssay resource**. *Nucleic Acids Res.* 2010 Jan;38(Database issue):D255-66. Epub 2009 Nov 19. PMID: [19933261](#)

#### Microbial Genomes

Twenty-two finished microbial genomes were added to the NCBI databases in January, 2010. The original sequence data files submitted to GenBank/EMBL/DDJB are on the FTP site ([ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/)). The RefSeq provisional versions of these genomes are also available ([ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/)).

## GenBank News

GenBank release 175.0 is available on the NCBI Web Service and FTP site. The current release incorporates sequence data as of December 15, 2009. A new GenBank release will be available in February. Release notes containing detailed information are available on the FTP site: <ftp.ncbi.nlm.nih.gov/genbank/gbrel.txt>.

## Updates and Enhancements

### RefSeq

RefSeq Release 39 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of January 23, 2010. It includes 13,656,433 records from 10,171 different species and strains. Changes since the last release can be found in the release notes (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release39.txt>). More information on the RefSeq project is available on the RefSeq Homepage ([www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/)).

### My NCBI

The My Bibliography portion of My NCBI now helps manage publication compliance with the NIH Public Access Policy. The 2010 January-February issue of the NLM Technical Bulletin includes an [article](#) that explains this new feature.

### COBALT Improvements

The COBALT multiple sequence alignment service produces downloadable multiple-alignment output in the popular text formats FASTA plus gaps, Clustal, Phylip, Nexus, and ASN.1. This feature greatly increases the versatility of COBALT by allowing export of results to alignment editors and other multiple sequence alignment tools. The [BLAST News](#) page has more information.

### UniVec Database

A new version (build 5.2) of UniVec, NCBI's non-redundant database of vector sequences, is now available on the VecScreen Web service ([www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html](http://www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html)).

The number of sequences in the database increased by three percent with this new build. The newly added sequences include 28 complete vector sequences, and 23 adaptor, primer and multiple-cloning site sequences.

The vector BLAST database, available from the BLAST area of the FTP site (<ftp.ncbi.nih.gov/blast/db/>), has also been updated to contain full-length versions of all sequences from GenBank that were used to build UniVec 5.2.

### Exhibits

NCBI will have an exhibit booth at the American Association for the Advancement of Science annual meeting held February 18-22 in San Diego, CA.

## Announce Lists and RSS Feeds

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page ([www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)



Twelve RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Send comments and questions about NCBI resources to: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or call 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, December 2009

Peter Cooper, PhD and Dawn Lipshultz

Created: December 9, 2009; Updated: December 9, 2009.

### New Databases and Tools

#### PubMed Central

The PubMed Central URL has been redesigned to allow for easier usability as well as consistency across the NCBI site. The previous site, [www.pubmedcentral.gov](http://www.pubmedcentral.gov), has been moved into the NCBI domain as a sub-site. The PMC home page can be accessed at [www.ncbi.nlm.nih.gov/pmc/](http://www.ncbi.nlm.nih.gov/pmc/).

PMC also now provides a “Preview” Table of Contents for embargoed journal issues for which at least one article has been released in PMC. The embargoed articles will be released according to the date reflected in the table of contents.

#### Microbial Genomes

Six finished microbial genomes were added to the NCBI databases between November 25 and December 31, 2009. The original sequence data files submitted to GenBank/EMBL/DDBJ are on the FTP site: [ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nlm.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nlm.nih.gov/genomes/Bacteria](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria).

### GenBank News

GenBank release 175.0 is available on the NCBI Web Service and FTP site. The current release incorporates sequence data as of December 15, 2009. Release notes ([gbrel.txt](#)) describing details of the release and upcoming changes are in the GenBank FTP directory.

### Updates and Enhancements

#### Entrez Utilities Policy Change

Effective June 1, 2010, all E-Utility requests must contain non-null values for both `&tool` and `&email` parameters. Standard usage policies are described on the following page: [http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils\\_help.html#UserSystemRequirements](http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html#UserSystemRequirements). If you have any concerns or questions about the policies, please contact [eutilities@ncbi.nlm.nih.gov](mailto:eutilities@ncbi.nlm.nih.gov).

#### My Bibliography

The My NCBI My Bibliography tool has been enhanced to allow users to add citations from books, meetings, presentations, patents, and articles not found in PubMed. The About My NCBI page provides information on how to use the tool to the best of its ability: [www.ncbi.nlm.nih.gov/sites/myncbi/about/](http://www.ncbi.nlm.nih.gov/sites/myncbi/about/).

#### Exhibits

NCBI will exhibit at the American Association for the Advancement of Science Conference on February 18-22 in San Diego, California.

## Announce Lists and RSS Feeds

NCBI offers eighteen topic-specific mailing lists that provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, Structures, Conserved Domains, BioSystems and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, November 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: November 23, 2009.

### Featured Resource: New Discovery-oriented PubMed and NCBI Homepage

A new and improved interface to the PubMed search and retrieval system is now in service at the NCBI site.

[www.ncbi.nlm.nih.gov/pubmed/](http://www.ncbi.nlm.nih.gov/pubmed/)

Accompanying the new PubMed is a completely re-designed NCBI Homepage.

[www.ncbi.nlm.nih.gov/guide/](http://www.ncbi.nlm.nih.gov/guide/)

Both pages feature a standard search bar with menus to access all NCBI resources, a list of instructions for completing common tasks (“How to...”), and a universal NCBI footer, soon to be on all NCBI pages, that provides quick links to NCBI resources.

The new PubMed is more streamlined than before with the more popular options easier to find along with access to related resources presented in a more obvious way. All previous functionality is still present for advanced searching, with options combined in a more unified and logical way bringing related tasks together and separating those that are different. The new NCBI Homepage is designed to serve as a Site Guide with a listing of all NCBI resources classified by topic and directions for common tasks readily accessible.

### The NCBI Site Guide

The NCBI Homepage is now the NCBI Site Guide as shown in Figure 1. This page is designed to provide rapid access to major areas of the NCBI Web site and to provide help and guidance for selecting the most appropriate databases, tools, and other resources for the task at hand. The central section of the Site Guide provides a list of 15 categories of associated NCBI resources: Literature, DNA & RNA, Proteins, Sequence Analysis, Genes & Expression, Genomes, Maps & Markers, Domains & Structures, Genetics & Medicine, Taxonomy, Data & Software, Training & Tutorials, Homology, Small Molecules, and Variation. Each one of these categories expands to a list of relevant databases and other features of the NCBI site grouped into Databases, Tools, Downloads and Submissions. For example, the DNA & RNA group expands when clicked to an organized list of 12 databases, five analysis tools, four avenues for downloading data, and two submission pathways. Each item in the Resources list has a brief description and a main heading link that leads directly to the relevant page. NCBI resources that are particularly relevant, new, or otherwise important are highlighted in these lists. These Featured Resources are also available in a section of the new universal NCBI footer, described below. In addition to the Featured Resources, a separate section on the right-hand-side of the Site Guide lists Popular Resources. The most commonly accessed resources at the NCBI site based on usage statistics are listed in this section. These provide rapid shortcuts to databases such as PubMed and Gene and tools such as BLAST.

### Common Elements on New NCBI Pages

Both the new PubMed and the new Site Guide have a new search bar and footer area that will aid navigation shown in Figure 2. These two features will be standard on all NCBI pages in the near future. The search bar at the top of the page has the traditional database pull-down list providing access to all NCBI Entrez databases. This bar also has the new “Resources” and “How To” pull-down lists to aid navigation and to access to practical

**Resources**

NCBI Home

All Resources (A-Z)

Literature

DNA & RNA

Proteins

Sequence Analysis

Genes & Expression

Genomes

Maps & Markers

Domains & Structures

Genetics & Medicine

Taxonomy

Data & Software

Training & Tutorials

Homology

Small Molecules

Variation

**Genotype and Phenotype**

Data from Genome Wide Association studies that links genes and diseases. See study variables, protocols, and analysis.

1 2 3

**How To...**

- Obtain the full text of an article
- Retrieve all sequences for an organism or taxon
- Find a homolog for a gene in another organism
- Find genes associated with a phenotype or disease
- Design PCR primers and check them for specificity
- Find the function of a gene or gene product
- Find syntenic regions between the genomes of two organisms

**Popular Resources**

- PubMed
- PubMed Central
- Bookshelf
- BLAST
- Gene
- Nucleotide
- Protein
- GEO
- Conserved Domains
- Structure
- PubChem

**NCBI News**

NCBI News - September 05 Oct 2009 2009  
The September 2009 issue of the NCBI News is available ...

NCBI News - August 19 Aug 2009 2009  
The August 2009 issue of the NCBI News is available online ...

NCBI News - July 2009 17 Jul 2009  
The July 2009 issue of the NCBI News is now available online...

**DNA & RNA**

**Resources** **How To**

**DATABASES**

**BioSystems**

Database that groups biomedical molecules, and sequence data relationships.

**Database of Expressed Sequence Tags**

A division of GenBank that contains reads of cDNA (transcript) sequences searched directly through the Nucleotide database.

**Database of Genome Survey Sequences**

A division of GenBank that contains reads of genomic DNA. dbGSS through the Nucleotide GSS Database.

**GenBank**

The NIH genetic sequence data collection of all publicly available sequences. GenBank is part of the International Nucleotide Sequence Database Collaboration.

**Quick Links**

BLAST (Basic Local Alignment Search Tool)

**DNA & RNA**

**Resources** **How To**

- Download a large, custom set of records from NCBI
- View/download features around an object or between two objects on a chromosome
- Link from an object on a map to another resource
- Obtain a genomic DNA clone for a gene
- Retrieve all sequences for an organism or taxon
- Find a curated version of a sequence record (NCBI Reference Sequence)
- Find transcript sequences for a gene
- Design PCR primers and check them for specificity
- Save a text search and/or receive regular search results by e-mail

**How To: Retrieve all sequences for an organism or taxon**

Starting with an organism or taxon name

- Search the **Taxonomy** database with the organism name. Accepted common names usually work at all taxonomic levels. Use the scientific name or formal name if no results are obtained with the common name.
- Click on the desired taxon name in the results. For terminal taxa - generally subspecies, species, or strains - this link leads directly to the summary page. For higher taxa this link will lead to the Taxonomy Browser showing the lower taxa contained within the higher taxon.
- If necessary, click on the desired taxon link in the Taxonomy Browser to reach the summary page.
- The number of records in each database are linked in the Entrez records table on the taxon summary page. Click the linked number of records in the table to retrieve all records from the chosen sequence database (Nucleotide, Nucleotide EST, Nucleotide GSS, Protein).

**Figure 1.** The new NCBI Site Guide that is now the Homepage featuring Resource categories and How To lists. Selecting a category (DNA & RNA) brings up an alphabetized list of Databases, Tools, Downloads and Submissions resources that pertain to that category. The How To tabs present a list of instructions for completing common tasks for the resource such as “How to retrieve all sequences for an organism” shown here.

task-oriented help. The Resources are the 15 categories of resources from the Site Guide, described above. Items in this list expand when selected to provide rapid access to the NCBI featured resources in that category plus a link to retrieve the entire category list in the new Site Guide. The “How To” pull-down links to the practical step-by-step instructions for common tasks. These are the same “How To” directions available from the Site Guide.

The new footer provides rapid navigation to all major areas of the NCBI site. At the top of the footer is a chain of links indicating the current page and its location in the Resource categories and providing “breadcrumbs” leading up the hierarchy. The four columns of links include the Resources, Featured and Popular categories from the Site Guide as well as help documents under “Getting Started” and NCBI news, background, and contact information under “NCBI Information.”

## Using the New PubMed Interface

The new PubMed interface has a more basic format without the various tabs that were present on the previous version. The functions provided by the tabs including Limits, Preview/Index, and History are now available as part of the Advanced Search page linked at the top of the search form. Popular links that were previously available on the blue-side bar in the old system are now organized into three columns below the search bar on the PubMed Homepage: Using PubMed containing help documents; PubMed Tools with the Single and Batch Citation Matchers, Clinical Queries, and Topic-Specific Queries; and More Resources with links to the Related MeSH and Journals databases, Clinical Trials, and the Entrez programming utilities (E-utilities). Table 1 provides a map of features in the old PubMed interface to their equivalents in the new interface.

**Table 1. Mapping of old PubMed features to the equivalent features in the new PubMed interface.**

<b>Old Feature</b>	<b>Corresponding New Feature</b>
Limits Tab	Advanced Search: Limit by ... section
Preview/index Tab	Advanced Search: Index of Fields and Field Values
History Tab	Advanced Search: Search History
Details Tab	Advanced Search: Details Link. Also in Discovery Column: Search Details in search results.
Clipboard Tab	Discovery Column: Clipboard Link at the top of the Discovery Column (Appears only if there are items in the Clipboard)
Display pull-down list, formats	Display Settings menu
Display pull-down list, links	Discovery Column: Related Data pull-down list
Show (number of records displayed)	Display Settings menu: Items per Page
Sort By pull-down list	Display Settings menu: Sort by
Send to pull-down list (File, Clipboard, Collections, E-mail, Order)	Send to Link (Send to text and printer are not on this menu and are available as separate features.)
Send to text	Display Settings: Summary(text), Abstract(text), MEDLINE, XML
Send to printer	No longer a separate feature. Web browser's print function produces formatted output with no graphics and no Discovery Column items.
Send to RSS feed	RSS link above the PubMed search box
Filter Tabs (Free Full Text, Reviews)	Discovery Column: Filter your results
Blue side-bar links: Entrez/PubMed; PubMed Services; Related Resources.	Bottom center of PubMed page. Categorical columns of links: Using PubMed; PubMed Tools; More Resources.

The image displays two panels from the NCBI website. The top panel shows the search bar with a pull-down menu for 'PubMed' and an auto-suggest list of search terms. The bottom panel shows the footer with five columns of categorized links: GETTING STARTED, RESOURCES, POPULAR, FEATURED, and NCBI INFORMATION.

**Top Panel: Search Bar and Auto-suggest List**

Search: PubMed

Advanced search Help

genome wide association A

genome wide association analysis

genome wide association asthma

genome wide association autism

genome wide association alzheimer

genome wide association and diabetes

association and obesity

association alcohol

association and stroke

wide association analysis

association and alzheimer

**Left Panel: Navigation Menu**

- All Resources
- Literature
- DNA & RNA**
  - BankIt
  - BLAST
  - GenBank
  - Genome Workbench
  - Influenza Virus
  - Nucleotide Database
  - PopSet
  - Reference Sequence (RefSeq)
  - Sequence Read Archive (SRA)
  - Trace Archive
  - All DNA & RNA Resources...
- Proteins
- Sequence Analysis
- Genes & Expression
- Genomes
- Maps & Markers
- Domains & Structures
- Genetics & Medicine
- Taxonomy
- Data & Software
- Training & Tutorials
- Homology
- Small Molecules
- Variation

**Bottom Panel: Footer**

You are here: NCBI > Literature > PubMed

Help Desk

GETTING STARTED	RESOURCES	POPULAR	FEATURED	NCBI INFORMATION
Site Map	Literature	PubMed	GenBank	About NCBI
NCBI Help Manual	DNA & RNA	PubMed Central	Reference Sequences	Research at NCBI
NCBI Handbook	Proteins	Bookshelf	Map Viewer	NCBI Newsletter
Training & Tutorials	Sequence Analysis	BLAST	Genome Projects	NCBI FTP Site
	Genes & Expression	Gene	Human Genome	Contact Us
	Genomes	Nucleotide	Mouse Genome	
	Maps & Markers	Protein	Influenza Virus	
	Domains & Structures	GEO	Primer-BLAST	
	Genetics & Medicine	Conserved Domains	Short Read Archive	
	Taxonomy	Structure		
	Data & Software	PubChem		
	Training & Tutorials			
	Homology			
	Small Molecules			
	Variation			

**Figure 2. The NCBI search bar (top panel) and footer (bottom panel) from the new PubMed pages.** The search bar and footer, presently on the site guide and PubMed pages, will be standard on all NCBI Web pages. The search bar provides access to all databases through the pull down list. The Advanced Search link accesses the Index, Limits, Details and History features. The auto-suggest query term list is currently available only in PubMed. The “Resources” and “How To” menus link to the categories and instructions from the Site Guide. The NCBI footer provides rapid navigation to all major areas of the NCBI site through the five columns of categorized links.



## Example: Finding genome wide association studies on late-onset Alzheimer Disease

A search for genome wide association studies for late-onset Alzheimer disease is a useful demonstration of the new PubMed interface and features. Typing “genome wide association Al ...” in the PubMed search box begins the search. As the query is typed, suggested queries appear below the search (Figure 2, *top panel*). These auto-complete suggestions are taken from recent productive queries from PubMed visitors that match the current query. Selecting the query “genome wide association study alzheimer” retrieves the set of 60 results shown in Figure 3. Like the PubMed Homepage, this new results summary page is simpler than the previous version, lacking the numerous tabs, display options, and other devices along the top of the page. Display and save options are now incorporated into the “Display settings” and “Send to” pop-up menus. These menus are present on both the Summary, shown in Figure 3, and the Abstract displays shown in Figure 4. The “Display Settings” menu allows for selection of any standard PubMed format, for changing the number of records displayed, and altering the sort order of the records. The previous option to send to text is automatically invoked by choosing any format other than Summary or Abstract (MEDLINE, XML, PMID List, Summary(Text), Abstract(Text)). The new “Send to” menu provides various ways of saving records for later use by sending them to a local file (File), the collections in MyNCBI (Collections), the LoansomeDoc ordering system (Order), the NCBI clipboard (Clipboard), or to an e-mail account (E-mail). The File and E-mail options allow for the selection of format and sorting order before sending.

In addition to standard Discovery items such as Title search and the PubMed Central Ad, the right-hand Discovery Column on the summary display contains three new items that provide advanced functions: Filter links, Related data, and Search details. The Filter links provide filtered or limited results and supersedes the filter tabs in the old PubMed. Default filters show review articles or articles with free full-text. A filter may be added to the current search by clicking on the link to apply the filter, clicking the plus sign that appears to append the filter to the search, and running the search with the new terms. In the current example, applying the Free Full Text and Review Filters in succession finds one review article with full-text in PubMed Central. Custom filters may be added through a MyNCBI account by following the “Manage filters” link.

The Related Data feature links to related items in the Entrez system for the entire set of articles displayed. These relationships may be based on computed similarity as with PubMed related articles or based on known linkages as when an article reports a nucleotide or protein sequence. For instance, there are 12 Gene records that cite members of the full-text-filtered set of articles in the current example.

The Search details feature shows useful information about query translation and mapping to the Medical Subject Headings (MeSH). The query here maps to the MeSH terms “alzheimer disease” and “genome wide association study”. More precise results may be obtained by editing the query translation so that only the indexed MeSH terms are searched. The following query results in a more relevant set of 23 articles:

```
genome wide association study[MeSH Terms] AND alzheimer disease[MeSH Terms]
```

Clicking on any of the titles in the set of results displays the new Abstract view of the PubMed record shown in Figure 4. This new format combines the formerly separate Web displays of Abstract and Abstract Plus formats and replaces the Citation format by including an expandable list that contains information from the Citation format: Publication Types, MeSH Terms, and Substances. The plain-text Abstract format is still available through the Display Settings menu described above for the search summary page.

## Summary

The new NCBI Site Guide and PubMed interface are designed to be more intuitive and less complicated. These improvements are part of the ongoing NCBI Discovery Initiative: making the NCBI interfaces easier to use and exposing relevant related resources. The changes in the Homepage and PubMed herald changes coming to all

The screenshot shows the PubMed search results interface with several callout boxes highlighting new features:

- Format, Items per page, and Sort by:** A callout box at the top left shows options for 'Summary', 'Abstract', 'MEDLINE', etc., and 'Items per page' (5, 10, 20, 50, 100, 200). The 'Sort by' options include 'Recently Added', 'Pub Date', 'First Author', 'Last Author', 'Journal', and 'Title'.
- Choose Destination:** A callout box at the top right shows options for 'File', 'Clipboard', 'Collections', 'E-mail', and 'Order', with an 'Apply' button.
- Display Settings:** A callout box at the top center shows 'Display Settings: Summary, 20 per page, Sorted by Recently Added' and 'Send to:'. Below this, the main results area shows 'Results: 1 to 20 of 60' and a list of search results.
- Filter your results:** A callout box on the right side of the results area shows a 'Filter your results:' menu with options: 'All (60)', 'Review (5)', and 'Free Full Text (19)'. A 'Manage Filters' link is also visible.
- Titles with your search terms:** A callout box on the right side shows a list of titles related to the search terms, such as 'Genome-wide association study implicates a chromosome 12 ris'.
- 16 free full-text articles in PubMed Central:** A callout box on the right side shows a list of 16 free full-text articles in PubMed Central, such as 'Genes within the serotonergic system are differentially expressed'.
- Find related data:** A callout box at the bottom left shows a 'Find related data' menu with 'Database: Gene' and 'Option: Gene Links'. It includes the text: 'Gene records that cite the current articles. Citations in Gene are added manually by NCBI or imported from outside public resources.' and a 'Find items' button.
- Search details:** A callout box at the bottom right shows the search query: '("genome"[MeSH Terms] OR "genome"[All Fields]) AND wide[All Fields] AND ("association"[MeSH Terms] OR "association"[All Fields]) AND ("alzheimer disease"[MeSH Terms] OR "alzheimer disease"[All Fields])' and a 'Search' button.

**Figure 3. The new PubMed results page.** The Display Settings menu manages formatting, number displayed and sorting order. The Send to menu manages destinations for the results. The right-hand Discovery Column now contains Filter links, a Find Related data menu that provides access to related items – previously on the “Display” pull-down list in the old PubMed, and the Search details that show term mappings and translations.

Display Settings:  Abstract
Send to:

Am J Hum Genet. 2008 Nov;83(5):623-32. Epub 2008 Oct 30.

### Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE.

Bertram L, Lange C, Mullin K, Parkinson M, Hsiao M, Hogan MF, Schjeide BM, Hooli B, Divito J, Ionita I, Jiang H, Laird N, Moscarillo T, Ohlsen KL, Elliott K, Wang X, Hu-Lince D, Ryder M, Murphy A, Wagner SL, Blacker D, Becker KD, Tanzi RE.

Genetics and Aging Research Unit, Mass General Institute for Neurodegenerative Disease (MIND), Department of Neurology, Massachusetts General Hospital, Charlestown, MA 02129, USA.

Alzheimer's disease (AD) is a gene... have been established to either cau... PSEN2(1-4)) or to increase suscept... late-onset AD is as high as 80%, (3... date. We performed a genome-wide... polymorphisms (SNPs) on a large (... self-reported European descent. We... significant genome-wide association... onset age. One of these signals (p... reflects APOE-epsilon4, which map... in three additional independent AD... almost 900 families. Two of these S... (combined p values 0.007 and 0.00... strongest association signal also sh... generated in an independent sampl... Although the precise identity of the... compelling evidence for the existen... APOE-epsilon4, primarily acts as a...

PMID: 18976728 [PubMed - indexed fo...]

Publication Types, MeSH Terms, Substances

**Publication Types:**

- [Research Support, N.I.H., Extramural](#)
- [Research Support, Non-U.S. Gov't](#)

**MeSH Terms:**

- [Age of Onset](#)
- [Algorithms](#)
- [Alleles](#)
- [Alzheimer Disease/genetics\\*](#)
- [Apolipoproteins E/genetics\\*](#)
- [Bayes Theorem](#)
- [Case-Control Studies](#)
- [Chromosomes, Human, Pair 14](#)
- [European Continental Ancestry Group](#)
- [Genetic Markers](#)
- [Genetic Predisposition to Disease\\*](#)
- [Genome-Wide Association Study\\*](#)
- [Humans](#)
- [Linear Models](#)
- [Linkage Disequilibrium](#)
- [Pedigree](#)
- [Polymorphism, Single Nucleotide](#)

**Substances:**

- [Apolipoproteins E](#)
- [Genetic Markers](#)

LinkOut - more resources

**Full Text Sources:**

- [Elsevier Science](#)
- [OhioLINK Electronic Journal Center](#)
- [PubMed Central](#)
- [Swets Information Services](#)
- [UK PubMed Central](#)

**Other Literature Sources:**

- [COS Scholar Universe](#)

**Medical:**

- [Genetics Home Reference](#)
- [Alzheimer's Disease - MedlinePlus Health Information](#)

**Related articles**

- ▶ A high-density whole-genome association study reveals that APOE is the [J Clin Psychiatry. 2007]
- ▶ SNPing away at complex diseases: analysis of single-nucleotide polym[Am J Hum Genet. 2000]
- ▶ APOE and other loci affect age-at-onset in [Am J Med Genet B Neuropsychiatr Genet. 2005]
- ▶ **Review** Dancing in the dark? The status of late-onset Alzheimer's dise [J Mol Neurosci. 2001]
- ▶ **Review** The current status of Alzheimer's disease genetics: what do [Pharmacol Res. 2004]

» See reviews... | » See all...

**Cited by 1 PubMed Central article**

- ▶ Genetics of Alzheimer's disease: recent advances. [Genome Med. 2009]

**All links from this record**

- ▶ Related Articles
- ▶ Gene
- ▶ Gene (GeneRIF)
- ▶ HomoloGene
- ▶ Nucleotide
- ▶ Nucleotide (RefSeq)
- ▶ Nucleotide (Weighted)
- ▶ OMIM (calculated)
- ▶ Protein (RefSeq)
- ▶ Protein (Weighted)
- ▶ References for this PMC Article
- ▶ SNP (Cited)
- ▶ Taxonomy via GenBank
- ▶ UniGene
- ▶ Protein
- ▶ SNP
- ▶ GEO Profiles
- ▶ Free in PMC
- ▶ Cited in PMC

**Figure 4. The new Abstract format display in PubMed.** The Abstract format combines the previous Abstract Plus and Citation formats by providing expandable sections with access to Publication Types, MeSH Terms, Substances, and LinkOut items.

NCBI interfaces that will produce a more consistent, effective, and powerfully integrated set of databases and tools.

## New Databases and Tools

### Bookshelf

New books added to the Bookshelf include: *Comparative Oncology*, *Preterm Birth: Causes, Consequences, and Prevention*, *Advancing Nuclear Medicine Through Innovation*, and various *Drug Class Reviews* titles. To browse any of these books go to [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books)

### Microbial Genomes

Thirty-three finished microbial genomes were added to the NCBI databases between October 29 and November 24, 2009. The original sequence data files submitted to GenBank/EMBL/DDBJ are on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 174.0 is on the NCBI Web Service and FTP site. The current release incorporates sequence data as of October 16, 2009. Release notes with detailed information are on the FTP site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### RefSeq

RefSeq Release 38 is now available through the Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release>). This full release incorporates genomic, transcript, and protein data available as of November 7, 2009. It includes 13,436,447 records from 9,115 different species and strains. Changes since the last release can be found in the release notes (<ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release38.txt>). More information on the RefSeq project is available on the RefSeq Homepage: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/).

### Entrez Gene

New features have been added to Entrez Gene displays. A 'Recent Activity' display now appears in each Gene record page in addition to the summary page. The Additional Links section now includes a 'Gene LinkOut' subsection that displays relevant external gene links that have been submitted by external databases. To receive Gene-related announcements, sign up for the 'gene-announce' mailing list.

### Exhibits

NCBI will have an exhibit booth at the [American Society for Cell Biology](#) annual meeting in San Diego, CA, held December 5-9, 2009.

### PubMed E-Utilities

PubMed 2010 DTDs will go into effect on December 14. The 2010 DTDs are available from the Entrez DTD page: [eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/index.shtml](http://eutils.ncbi.nlm.nih.gov/corehtml/query/DTD/index.shtml). Specific DTD changes are noted in the Revision Notes section near the top of each DTD. Additional information is available from the Announcement to the

NLM Data Licensees 2010 DTD and XML Changes; File Distribution Schedule Changes: [www.nlm.nih.gov/bsd/licensee/announce/2009.html#d09\\_17](http://www.nlm.nih.gov/bsd/licensee/announce/2009.html#d09_17).

## Announce Lists and RSS Feeds

Topic-specific mailing lists provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The Announcement List summary page describes the various lists and how to subscribe:

[www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html).

The NCBI Announce mailing list sends notices on *NCBI News* updates and important changes at the NCBI site.

[www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now produced by NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

[www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI through electronic mail, [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, October 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: October 30, 2009.

### New Databases and Tools

#### New NCBI Homepage

The newly designed NCBI Homepage is more streamlined with resources, tools, and services divided into organized groups. The “All Resources” link provides a list of all NCBI resources while the subject lists below provide resources according to those headings. The “How To” link goes to lists of short, FAQ-like questions and answers on many topics. There are also How To help links within each subject category. The default “All Databases” search in the Search box performs a global query search on all Entrez Databases. A featured resource article in the next NCBI News issue will describe the New Homepage in detail.

#### New PubMed Homepage and Search Service

The PubMed service has a new design and provides new functionality. The PubMed Redesign article in the NLM Technical Bulletin outlines changes and describes how to find previously used functionalities.

[www.nlm.nih.gov/pubs/techbull/so09/so09\\_pm\\_redesign.html](http://www.nlm.nih.gov/pubs/techbull/so09/so09_pm_redesign.html)

A featured resource article in the next NCBI News issue will describe improvements and enhancements in detail.

#### Microbial Genomes

Twelve finished microbial genomes were added to the NCBI databases between September 14 and October 28. The original sequence data files submitted to GenBank/EMBL/DDBJ are on the FTP site: [ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

#### GenBank News

GenBank release 174.0 is now on the NCBI Web Service and FTP site. The current release incorporates sequence data as of October 16, 2009. Release notes with detailed information are on the FTP site: [ftp.ncbi.nih.gov/genbank/gbrel.txt](ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

### Updates and Enhancements

#### PubMed Central

Two new pages have been added to the PMC site. The “New in PMC” page provides updates on PMC developments and features.

[www.ncbi.nlm.nih.gov/pmc/about/new\\_in\\_pmc.html](http://www.ncbi.nlm.nih.gov/pmc/about/new_in_pmc.html)

The “Public Access and PMC” page provides information on the relationship between PMC and the NIH Public Access Policy.

[www.ncbi.nlm.nih.gov/pmc/about/public-access-info.html](http://www.ncbi.nlm.nih.gov/pmc/about/public-access-info.html)

## BLAST

BLAST version 2.2.22 is now available for download from the BLAST section of the NCBI FTP site.

<ftp.ncbi.nih.gov/blast/executables/>

BLAST 2.2.22 includes BLAST+ command-line applications. The BLAST News page provides an extensive list of improvements and BLAST+ application changes.

[blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastNews#1](blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastNews#1)

## Publications

A new article summarizing Mammalian Gene Collection project will soon be published in *Genome Research*. The article is available electronically ahead of the print publication:

The MGC Project Team. "The completion of the Mammalian Gene Collection (MGC)." *Genome Research*. 2009 Oct 28 [Epub ahead of print]. PMID: 19767417.

## Announce Lists and RSS Feeds

Three new mailing lists are available for updates and changes to NCBI resources: NCBI Structures, Conserved Domains, and BioSystems. Topic-specific mailing lists provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The Announcement List summary page describes the various lists and how to subscribe:

[www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html).

The NCBI Announce mailing list sends notices on *NCBI News* updates and important announcements.

[www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now produced by NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce.

[www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI through electronic mail, [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, September 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: September 14, 2009.

### Featured Resource: The Genome Reference Consortium Human Genome Build 37 now Available

In August the NCBI released the annotation of build 37 of the human genome. This build includes new sequence and assembly provided by the Genome Reference Consortium (GRC). The GRC is a collaboration of the Wellcome Trust Sanger Center, the Washington University Genome Center, the European Bioinformatics Institute and the NCBI. The goal of the GRC is to correct misassembled regions, to close remaining gaps, and to provide alternate assemblies of structurally variant positions (loci) in the genome. Build 37, also known as GRCh37, includes updates for all human chromosomes, closes 25 sequence gaps, corrects over 150 problems in build 36, and adds nine alternate loci.

The GRC page at NCBI provides additional details about this new assembly.

[www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/](http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/)

The NCBI Website provides easy access for searching and exploring the sequences and annotations of the new and improved primary reference genome and alternate loci through the Entrez system, the graphical sequence viewer, the Map Viewer, and the NCBI Web BLAST services.

### GRCh37 Sequences at NCBI

The GRCh37 assembly includes the assembled human chromosomes, some unlocalized and unplaced sequence, and alternate assemblies for structurally variable regions in the genome. The primary assembly chromosome sequences are available under accession numbers CM000663 through CM000686. These are assemblies of the 22 autosomes plus the X and Y chromosomes. The nine alternate assemblies are for the following regions: the UDP glucuronosyltransferase 2, polypeptide B17 gene (UGT2B17) on chromosome 4 (accession GL000257); the Major Histocompatibility Complex (MHC) on chromosome 6 (accessions GL000250 through GL000256); and the microtubule-associated protein tau (MAPT) gene on chromosome 17 (accession GL000258).

The NCBI genome annotation pipeline has created a corresponding set of 31 reference sequences (RefSeqs) that provide the locations of genes and other features on the GRCh37 reference assembly and alternate loci. Table 1 shows the correspondence between the RefSeq and GenBank records for GRCh37.

**Table 1. Correspondence of GenBank, RefSeq accession numbers, and assembled sequences for the GRCh37 reference genome.**

GenBank Accession	RefSeq Accession	Description
CM000663	NC_000001	Chromosome 1
CM000664	NC_000002	Chromosome 2
CM000665	NC_000003	Chromosome 3
CM000666	NC_000004	Chromosome 4
CM000667	NC_000005	Chromosome 5
CM000668	NC_000006	Chromosome 6

Table 1 continued from previous page.

GenBank Accession	RefSeq Accession	Description
CM000669	NC_000007	Chromosome 7
CM000670	NC_000008	Chromosome 8
CM000671	NC_000009	Chromosome 9
CM000672	NC_000010	Chromosome 10
CM000673	NC_000011	Chromosome 11
CM000674	NC_000012	Chromosome 12
CM000675	NC_000013	Chromosome 13
CM000676	NC_000014	Chromosome 14
CM000677	NC_000015	Chromosome 15
CM000678	NC_000016	Chromosome 16
CM000679	NC_000017	Chromosome 17
CM000680	NC_000018	Chromosome 18
CM000681	NC_000019	Chromosome 19
CM000682	NC_000020	Chromosome 20
CM000683	NC_000021	Chromosome 21
CM000684	NC_000022	Chromosome 22
CM000685	NC_000023	Chromosome X
CM000686	NC_000024	Chromosome Y
GL000250	NT_167244	MHC Region (ALT_REF_LOCI_1)
GL000251	NT_113891	MHC Region (ALT_REF_LOCI_2)
GL000252	NT_167245	MHC Region (ALT_REF_LOCI_3)
GL000253	NT_167246	MHC Region (ALT_REF_LOCI_4)
GL000254	NT_167247	MHC Region (ALT_REF_LOCI_5)
GL000255	NT_167248	MHC Region (ALT_REF_LOCI_6)
GL000256	NT_167249	MHC Region (ALT_REF_LOCI_7)
GL000257	NT_167250	UGT2B17 Region (ALT_REF_LOCI_8)
GL000258	NT_167251	MAPT Region (ALT_REF_LOCI_9)

## Retrieving and Viewing GRCh37 at NCBI

GRCh37 sequences and annotations are easily retrieved and viewed in the Entrez system and the NCBI Map Viewer. A search for GRCh37[Title] in the Entrez nucleotide database ([www.ncbi.nlm.nih.gov/sites/entrez?db=nucore](http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucore)) collects all 564 records associated with the current build. Restricting to reference sequences (ReSeq) using the filter tab limits the results to the 282 processed RefSeq versions of chromosomes, contigs, and alternate loci that include the annotations of biological features. Figure 1 (top panel) shows the traditional GenBank view of the GRCh37 chromosome 4 (NC\_000004) in the Entrez system. This abbreviated view can be adjusted with controls on the page to add biological features and sequence. However, the large number of features and long sequence make this an awkward way to browse the data. The graphical sequence viewer, offered as the “Graphics report” link at the top of the GenBank view, provides a better alternative for exploring

the chromosome record and its features. Following the Graphics report link and searching for the UGT2B17 as a marker results in the display of the region surrounding the UGT2B17 gene on chromosome 4 as shown in the bottom panel of Figure 1. The graphical viewer provides details of gene position structure and orientation, alignments of transcripts and proteins, and the ability to display SNPs and other markers. Each annotated gene or transcript in the graphical view has links to sequence display formats, other databases such as Gene, and the ability to run a BLAST search with the annotated genomic, transcript, or protein sequence (Figure 2).

The NCBI Map Viewer is another useful way to view aspects of the genome build. The human genome map viewer is accessible from the Map Viewer Homepage:

[www.ncbi.nlm.nih.gov/mapview/](http://www.ncbi.nlm.nih.gov/mapview/)

All genes, transcripts, and proteins associated with the genome have links to the build in the Map Viewer from the corresponding records in the Entrez system. In addition, the NCBI Web BLAST service can link results of searches against human genome plus transcript database as well as those from the separate human genome BLAST service directly into the Map Viewer. This BLAST search option can be used to highlight improvements in the human genome build as shown in the following example.

### **Example: Exploring Changes in Chromosome 4 in Build 37**

As mentioned previously the GRCh37 assembly closed 25 gaps in the previous build (build 36) of the human genome. One such gap is in the region surrounding the UGT2B17 gene on chromosome 4. In build 36, this region appears to contain a partial duplication surrounding a gap. Since the human genome BLAST service and the Map Viewer allow searches against both GRCh37 and build 36, changes in the structure of this region between the two builds are easily demonstrated. Using the genomic region corresponding to the transmembrane serine protease 11E (TMPRSS11E) as a query in human genome BLAST ([NC\\_000004](#), [bases 69313167-69363322](#)) shows the apparent duplication in build 36. This search is set-up directly from the TMPRSS11E gene in the graphical viewer by following the genome specific BLAST link from the Views and Tools pop-up menu (Figure 2, top panel, left). The results against build 36 show two near-perfect matches for the TMPRSS11E genomic region on different contigs flanking an apparent gap (Figure 2, top panel, right). This highlights an apparent duplication – but an incomplete one since the upper contig contains the UGT2B17 gene while the lower contig appears to lack this gene. This structure (duplication and gap) is known to be an artifact caused by the incorporation of two different alleles, one of which is a null allele for UGT2B17, into the build 36 genome (1). The current build solves this problem by incorporating the UGT2B17 containing allele into the primary reference genome and providing a separate record, ALT\_REF\_LOCI\_8 ([NT\\_167250](#)), for the null allele. The structure of the new reference assembly and the alternate allele are easily demonstrated in the same way as for build 36 by a human genome BLAST search against build 37 (Figure 2, bottom panel).

### **Summary**

The genome reference consortium (GRC) build 37 provides a more accurate and improved representation of the human genome by correcting errors, closing gaps, and providing alternate representations of structurally variant regions. The GRC itself, a collaboration among sequencing centers and bioinformatics resource and analysis centers such as the NCBI, will continue to provide the most up to date and accurate sequence and annotation for the reference human genome as additional data and analysis alter the view of the genome. The NCBI Website will continue to offer improved and more powerful visualization and analysis tools for investigating the human genome.

Format: [GenBank](#) [FASTA](#) [Graphics](#) [More Formats](#) [Download](#) [Save](#) [Links](#)

★ Try the [Graphics report](#) for a more informative view of the biological features.

NCBI Reference Sequence: NC\_000004.11

### Homo sapiens chromosome 4, GRCh37 primary reference assembly

[Comment](#) [Features](#)

LOCUS NC\_000004 191154276 bp DNA linear CON 10-JUN-2009  
 DEFINITION Homo sapiens chromosome 4, GRCh37 primary reference assembly.  
 ACCESSION NC\_000004 GPC\_000000028  
 VERSION NC\_000004.11 GI:224589816  
 DBLINK Project:168  
 KEYWORDS  
 SOURCE Homo sapiens (human)  
 ORGANISM [Homo sapiens](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Homo.

**Change Region Shown** ▾

**Customize View** ▲

Abbreviated view  
 Customize

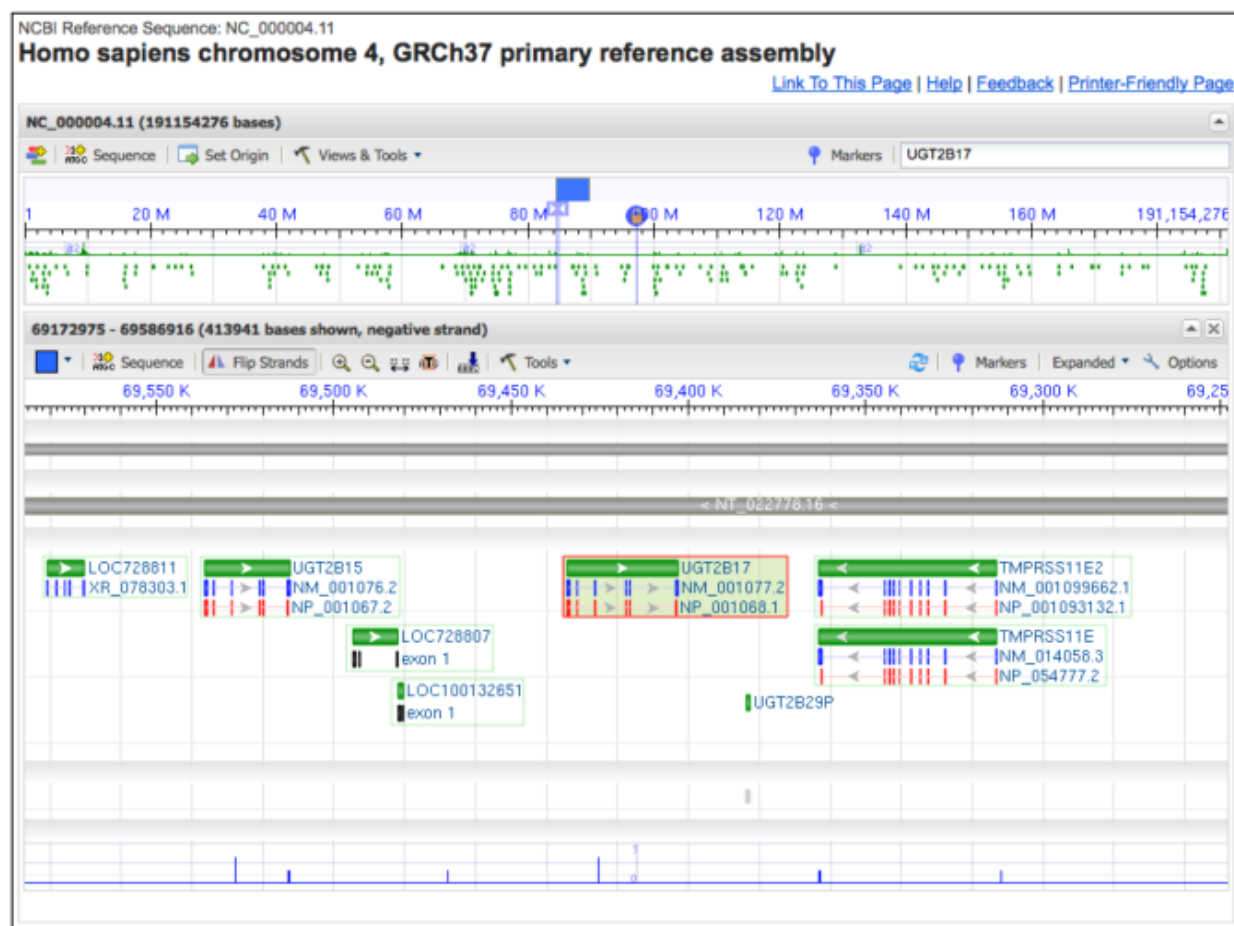
**Basic Features**

Default features  
 Gene, RNA, and CDS features only

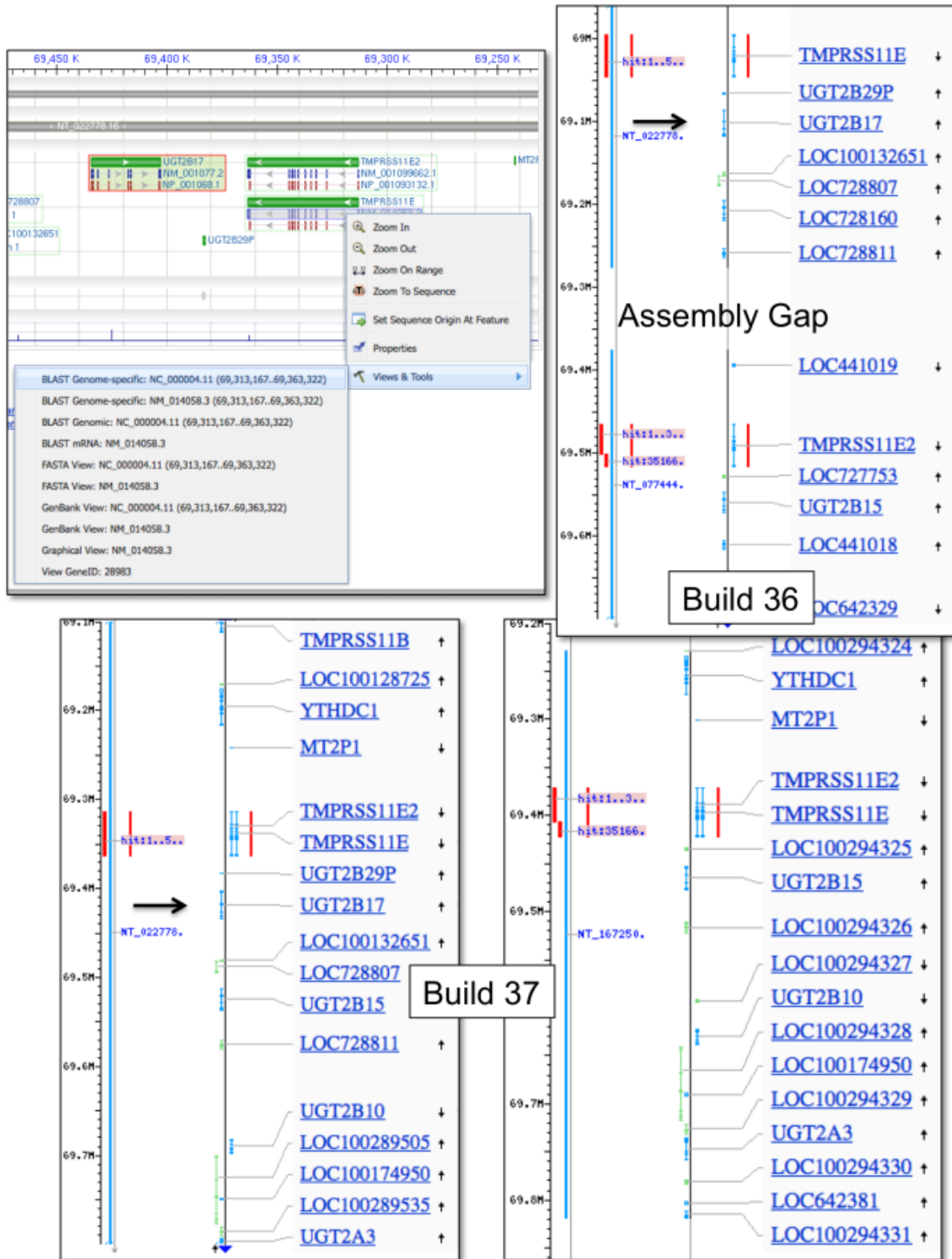
**Sequence display options**

Show sequence  
 Show minus strand

[Update View](#)



**Figure 1. Chromosome 4 record from the GRCh37 primary reference assembly.** *Top panel.* The GenBank record display in Entrez showing the controls that allow changing features and sequence options. The “Graphics report” option at the top of the page provides access to the graphical sequence viewer. *Bottom panel.* The UGT2B17 region of chromosome 4 in the graphical sequence viewer. The alternate locus for this region is the null allele for UGT2B17.



**Figure 2.** Structure of the UGT2B17 region on chromosome 4 in build 36 and the GRCh37 build (build 37) as demonstrated by Map Viewer displays of human genome BLAST results. *Top panel, left.* Human genome BLAST search set-up from the "Views and Tools" feature on the TMPRSS11E gene in the graphical viewer. *Top panel, right.* Human genome BLAST results, build 36, highlighting (red) the apparent duplication of the TMPRSS11E gene. The UGT2B17 gene (black arrow) is on the contig above the gap in the assembly. *Bottom panel, left.* TMPRSS11E BLAST results on the primary reference assembly showing the single result and the UGT2B17 gene. *Bottom panel, right.* BLAST results on the alternate locus for the UGT2B17 null allele. The apparent duplication and gap are resolved in GRCh37.

## Reference

1. Xue Y Sun. Adaptive evolution of UGT2B17 copy-number. Adaptive evolution of UGT2B17 copy-number. 2008;83(3):337–46. PubMed PMID: 18760392.

## New Databases and Tools

### New NCBI Homepage

A new NCBI Homepage is available for beta testing during the next two months. The new look is cleaner and better organized than the current page. New features include a “How To” section for answers to common questions and links to resource lists. The new page is available for testing at the following URL:

<http://preview.ncbi.nlm.nih.gov/guide/>

Feedback is appreciated and encouraged. Please send feedback to [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)

### PubMed Redesign

PubMed has also undergone reconstruction and is available for testing for a two week period. Many changes have been made that make search and retrieval easier and more comprehensive. The new design is quite different than the old but incorporates all of the new features that have been added over the past year such as Recent Activity, Ads, and Sensors. Please test the site and provide feedback on your experience.

<http://preview.ncbi.nlm.nih.gov/pubmed>

The National Library of Medicine Technical Bulletin provides a guide for making the transition to the new PubMed interface:

[www.nlm.nih.gov/pubs/techbull/so09/so09\\_pm\\_redesign.html](http://www.nlm.nih.gov/pubs/techbull/so09/so09_pm_redesign.html)

### Rapid Research Notes

Rapid Research Notes (RRN) is a new resource that contains articles published online for immediate communication. The H1N1 outbreak prompted the development of RRN, but future collections will consist of other biomedical information as well. See the RRN homepage ([www.ncbi.nlm.nih.gov/rrn/](http://www.ncbi.nlm.nih.gov/rrn/)) and the “About” page ([www.ncbi.nlm.nih.gov/rrn/about/index.html](http://www.ncbi.nlm.nih.gov/rrn/about/index.html)) for more information.

### Microbial Genomes

Sixty-four finished microbial genomes were released during the dates July 1 - September 14. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp://ncbi.nlm.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 173.0 is now on the NCBI Web and FTP sites. The current release includes data available as of August 21, 2009. The release notes provide detailed information and statistics on the release: <ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### RefSeq

RefSeq Release 37 is now part of the NCBI Entrez system and can be downloaded from the FTP site (<ftp.ncbi.nlm.nih.gov/refseq/release/>). This full release incorporates genomic, transcript, and protein data available as of September 3, 2009. It includes 12,941,750 records from 9,005 different species and strains. Changes since the previous release can be found in the release notes (<ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release37.txt>). More information on the RefSeq project is available on the RefSeq Homepage: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/).

### dbSNP

Complete data for the dbSNP Bovine build 130 are now part of the NCBI Entrez system and can be downloaded from the dbSNP FTP site. More detailed genome build information is available on the dbSNP page: [www.ncbi.nlm.nih.gov/SNP/snp\\_summary.cgi](http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi).

### Exhibits

NCBI will have an exhibit booth at the [American Society of Human Genetics annual meeting](#) in Honolulu, Hawaii, held October 20-24, 2009. Staff will present a tutorial, “The NCBI Discovery System: Integrated Access to Literature, Sequences, Genomes and Molecular Structures” on Wednesday, October 21 at 11:30 a.m. in the Convention Center (room 307).

## Announce Lists and RSS Feeds

Three new mailing lists are available for updates and changes to NCBI resources. The new announce lists are: NCBI Structures, Conserved Domains, and BioSystems.

Eighteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). For instructions on how to receive updates on the *NCBI News*, please visit: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.





## NCBI News, August 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: July 27, 2009.

### Featured Resource: The NCBI Short Read Archive (SRA) of Next-Generation Sequencing Data

The NCBI now maintains the Short Read Archive (SRA) ([www.ncbi.nlm.nih.gov/Traces/sra/](http://www.ncbi.nlm.nih.gov/Traces/sra/)) as a repository for data from sequencing projects that use the new massively parallel sequencing technologies, often called next-generation sequencing. These methods can generate hundreds of megabases to gigabases of data in a single instrument run, millions of times the output of a standard Sanger sequencing instrument. Applications of these technologies include sequencing of new genomes, re-sequencing of targeted genomic regions, sequencing complete genomes of multiple individuals to mine for variations, transcriptome sequencing to sample splice variants and expression levels, environmental samples and other metagenome sequencing, and chromatin DNA binding protein analysis. SRA provides the ability to search and display aspects of SRA project data through the SRA homepage (Figure 1, top panel), and the Entrez system (Figure 1, bottom panel). The SRA site also provides direct access to download data through the Aspera Connect ([www.aspera.com](http://www.aspera.com)) client that offers much faster transfers than traditional ftp. A recently added BLAST service allows searches against the transcriptome sequencing studies from the SRA data.

The Short Read Archive will become quite important as next-generation sequencing technologies continue to improve and become even less expensive. The power and capabilities of the SRA site will expand to provide better and more powerful options for searching and connecting these data to other resources.

### Next-Generation Sequencing Technologies

SRA accepts and presents data from all current next-generation sequencing platforms including 454 (Roche), Illumina, SOLiD (Applied Biosystems), HeliScope, and Complete Genomics. While these systems use different approaches to isolate and amplify the target molecules and to generate sequence, all rely on extreme miniaturization of the system components, simultaneous reactions in parallel in a flow cell, light-based detection of in the sequencing reactions, and image analysis to acquire sequence information from multiple reactions at once. These methods yield huge numbers of short sequence reads from a single instrument run. Individual read lengths vary from around 25 bases to more than 400 bases depending on the platform. Data can include sequence, quality scores, color values, and intensity graphs depending on the platform involved.

### Data in SRA

#### Data Concepts

Data in the SRA are classified into a hierarchy of Studies, Experiments, Samples, and their corresponding Runs. Studies have an overall goal and may be comprised of several Experiments. An Experiment describes specifically what was sequenced and the method used. It includes information about the source of the DNA, the Sample, the sequencing platform, and the processing of the data. Each Experiment is made up of one or more instrument Runs. A Run contains the results or reads from each spot in the instrument run. In the future, some data will also have an associated Analysis. These Analyses may include assemblies of the short reads into genomic or transcript contigs and alignment to existing genomes or alignments with SRA data. Records at each level have unique accession identifiers with a specific three letter prefix that indicates the type of record: ERP or SRP for

**Short Read Archive**

[Main](#) [Browse](#) [Search](#) [Download](#) [Submit](#) [Documentation](#) [Software](#)

[Announcements](#) [Provisional SRA Tracking](#) [History](#) [About](#)

The Short Read Archive (SRA) stores raw sequencing data from the "next" generation of sequencing platforms including Roche 454 GS System<sup>®</sup>, Illumina Genome Analyzer<sup>®</sup>, Applied Biosystems SOLiD<sup>®</sup> System, Helicos Heliscope<sup>®</sup>, Complete Genomics<sup>®</sup>, and others.

Current capabilities include:

- [Run Browser](#)
- [Study/Sample/Experiment/Analysis](#) browsers
- [Download facility](#)
- [Search SRA \(using Entrez\)](#)
- [Interactive submissions facility](#)
- [Automated submissions](#)

See [Sequence Read Archive Overview](#) for more information.

**NCBI Short Read Archive** [My NCBI](#) [\[Sign In\]](#) [\[Register\]](#)

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search  for    [Save Search](#)

[Limits](#) [Preview/Index](#) [History](#) [Clipboard](#) [Details](#)

Display  Show

All: 5929

Items 1 - 20 of 5929 Page  of 297 [Next](#)

1: [SRX006820](#) 454 sequencing of Thermoanaerobacter ethanolicus CCSD1 random whole genome shotgun library. [Links](#)

*Submitter:* JGI, PGF [Download data for this experiment SRX006820](#)

*Study:* Thermoanaerobacter ethanolicus CCSD1 Whole Genome Sequencing Project (SRP000977) • [Summary](#) • [Genome Project](#) • [All experiments](#)

*Sample:* [Thermoanaerobacter ethanolicus CCSD1 \(SRS004232\)](#)

*Instrument:* 454 GS FLX

#	Run	# of Spots	# of Bases
<b>Total: 1 run, 68,714 spots, 30.1M bases</b>			
1.	<a href="#">SRR019245</a>	68,714	30.1M

2: [SRX006819](#) 454 sequencing of Thermoanaerobacter ethanolicus CCSD1 random whole genome shotgun library. [Links](#)

*Submitter:* JGI, PGF [Download data for this experiment SRX006819](#)

*Study:* Thermoanaerobacter ethanolicus CCSD1 Whole Genome Sequencing Project (SRP000977) • [Summary](#) • [Genome Project](#) • [All experiments](#)

*Sample:* [Thermoanaerobacter ethanolicus CCSD1 \(SRS004232\)](#)

*Instrument:* 454 GS FLX

#	Run	# of Spots	# of Bases
<b>Total: 1 run, 592,348 spots, 339.7M bases</b>			
1.	<a href="#">SRR019244</a>	592,348	339.7M

3: [SRX006817](#) 454 sequencing of Ferroglobus placidus DSM 10642 random whole genome shotgun library. [Links](#)

*Submitter:* JGI, PGF [Download data for this experiment SRX006817](#)

*Study:* Ferroglobus placidus DSM 10642 Whole Genome Sequencing Project (SRP000975) • [Summary](#) • [Genome Project](#) • [All experiments](#)

*Sample:* [Ferroglobus placidus DSM 10642 \(SRS004230\)](#)

*Instrument:* 454 GS FLX

#	Run	# of Spots	# of Bases
<b>Total: 1 run, 99,256 spots, 43.7M bases</b>			
1.	<a href="#">SRR019238</a>	99,256	43.7M

**Figure 1. Short Read Archive Web access.** *Top panel.* The SRA homepage has access to the SRA browser as well as documentation, and a link to SRA submissions through tabs at the top of the page. *Bottom panel.* Entrez allows searches of SRA Experiment records. These link to the parent Study, and Runs in the SRA browser. Other Experiments for the same Study and Sample are linked to each record. See the text for details on the Study, Sample, Experiment and Run records.

Studies, SRS for samples, SRX for Experiments, and SRR for Runs. Figure 2 shows Study ([SRP000095](#), top panel), Experiment ([SRX000113](#), middle panel, and [SRX000114](#)), and Run ([SRR000416](#), bottom panel) records for the 454 sequencing of James Watson's genome by Cold Spring Harbor Laboratory. Study and Run records are displayed in the SRA browser. The corresponding Experiment records are displayed in the NCBI Entrez system as described in the next section.

## Searching and Viewing SRA Data in the SRA Browser and Entrez

Studies, Runs, and their associated Samples can be viewed and browsed through the SRA browser link on the SRA homepage.

[www.ncbi.nlm.nih.gov/Traces/sra](http://www.ncbi.nlm.nih.gov/Traces/sra)

Experiment records are available for searching in the Entrez SRA database.

[www.ncbi.nlm.nih.gov/sites/entrez?db=sra](http://www.ncbi.nlm.nih.gov/sites/entrez?db=sra)

As with other Entrez databases using field limits in search queries produce more precise results. The organism field is useful, as in all NCBI molecular databases, for finding experiments involving a particular taxon. The properties field is helpful for finding specific types of SRA studies. For example, the following query finds all human genomic resequencing Experiments – 984 at the time of this writing.

```
human[organism] AND study type resequencing[Properties] AND biomol genomic[Properties]
```

All of the available fields and their indexed terms can be browsed through the Preview/Index tab on the SRA Entrez search page.

The record for the Study associated with an Experiment, all Experiments for the Study, and Experiments that used the same sample are easily retrieved through links on the Entrez SRA Experiment record (Figure 2, middle panel). SRA Experiment records in Entrez are integrated with data from other Entrez databases. Links to PubMed, GEO datasets, Genome Projects, Nucleotide, and Taxonomy are currently available for the Experiment records. Currently 6,240 Experiments are available from 806 Studies.

## SRA BLAST Service

In addition to text searches of the SRA experiments through Entrez, NCBI also offers a nucleotide BLAST service for sequence similarity searching of 454 sequencing reads for transcriptome studies. This service is accessible from the “Specialized BLAST” section of the BLAST Homepage.

<http://blast.ncbi.nlm.nih.gov>

Databases are labeled by taxon. Currently there are transcriptome reads for 31 species and two metagenome data sets.

## Downloading SRA Data

SRA data can be downloaded through the “Download” tab on the SRA homepage or through the Download link that is present on Study, Sample, and Experiment records (Figure 2). Because data for SRA projects often exceed 10 Gigabytes, traditional ftp may be too slow to download data effectively. To avoid this problem, SRA download links use the fasp<sup>tm</sup> protocol developed by Aspera to transfer data. This protocol is more efficient and stable than traditional ftp. The free Aspera Connect Web browser plug-in, available from the company's Website, is required to download SRA data.

[www.asperasoft.com](http://www.asperasoft.com)

**Short Read Archive**

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST

Study Sample Run Browser Entrez SRA Experiments Entrez Pubmed Entrez GEO DataSets Entrez Genome Project Entrez WGS Project Entrez Taxonomy

## SRP000095 James Watson's Personal Genome Sequence

Study Type: Whole Genome Sequencing [Download fastq for entire study](#)

Submission: SRA000065 by CSHL on 2007-06-05T14:57:00Z

Abstract: James D. Watson's personal genome was sequenced at 6X coverage using 454 Life Sciences Technology.

Description:

Properties: INSDC Project id: 28335 External Link: [James Watson's Personal Genome Sequence \(home page at CSHL\)](#)

### Experiments

Show RUNs for each experiment

Accession	Spots	Bases
Total: 2	76.5M	20.3G
<a href="#">SRX000113</a>	1.2M	316.3M
<a href="#">SRX000114</a>	75.4M	20.0G

1: [SRX000113](#) 454 sequencing of Human James D. Watson genomic fragment library [Links](#)

**Experiment design:** James D. Watson whole genome shotgun library sequenced on 454 FLX. [Download data for this experiment SRX000113](#)

**Submission:** SRA000065 by CSHL

**Study Summary:** James Watson's Personal Genome Sequence (SRP000095) • [Study](#) • [All experiments](#)

Total: 8 runs, 1.2M spots, 316.3M bases

#	Run	# of Spots	# of Bases
1.	<a href="#">SRR000416</a>	167,212	45.7M
2.	<a href="#">SRR000440</a>	142,006	38.9M
3.	<a href="#">SRR000445</a>	109,498	30.4M
4.	<a href="#">SRR000481</a>	132,411	36M
5.	<a href="#">SRR000509</a>	165,661	45M
6.	<a href="#">SRR000533</a>	155,113	42.6M
7.	<a href="#">SRR000549</a>	101,757	27.9M
8.	<a href="#">SRR000550</a>	181,001	49.8M

**Project:** Project Jim

**Abstract:** James D. Watson's personal genome was sequenced at 6X coverage using 454 Life Sciences Technology.

**External link:** [James Watson's Personal Genome Sequence \(home page at CSHL\)](#)

**Center:** CSHL

**Center Project:** Project Jim

**Sample:** Nuclear genome isolate of James D. Watson. ([SRS000284](#)) [\(less...\)](#)

**Organism:** James D. Watson

**Library:** L1 [\(less...\)](#)

**Strategy:** WGS

**Source:** GENOMIC

**Selection:** RANDOM

**Layout:** SINGLE

**Construction protocol:** None provided.

**Platform:** LS454 [\(less...\)](#)

**Instrument model:** GS FLX

**Processing:**

**Base calls:** Base Space, 454Basecaller

**Quality score:** 454Basecaller, 64x1

**Spot descriptor:**

1 4

---

**Experiment: SRX000113**  
James D. Watson whole genome shotgun library sequenced on 454 FLX.

**Run:**

Accession:

Alias: DTEY4P1

Instrument model: 454 GS FLX

Date of run: 2005-12-07T14:54:15Z

Run center: 454MSC

**Statistics:**

Number of spots: 167212 [Show Plate image](#)

Number of reads: 334424

**Other:**

Study: James Watson's Personal Genome Sequence

Design: James D. Watson whole genome shotgun library sequenced on 454 FLX.

Platform: LS454

Sample: James D. Watson

Library Name: L1

Library Strategy: WGS

Library Source: GENOMIC

Library Selection: RANDOM

Library Layout: SINGLE

Library Construction Protocol: None provided.

Find spots:  X:  Y:    View:  reads (customize)  signals  intensity graph

[What can the filter be applied to?](#)

**Reads (separated)**

1. [SRR000416.1](#)  
name: DTEY4P101C0MOB  
plate:DTEY4P1, region:1, x:1120, y:825

2. [SRR000416.2](#)  
name: DTEY4P101AKHL5  
plate:DTEY4P1, region:1, x:116, y:923

3. [SRR000416.3](#)  
name: DTEY4P101DX4Y  
plate:DTEY4P1, region:1, x:1500, y:144

```
>gnl|SRA|SRR000416.1.1 DTEY4P101C0MOB Technical Read (Adapter)
tcag

>gnl|SRA|SRR000416.1.2 DTEY4P101C0MOB Application Read (Forward)
ACATGCTACTTGATTGTGTTCTGGTAGATGAAAGGATTAGATATCAAAGTTAAATTC
CTACTTGAGATTTTCAGAAATGTTCTCACTCCAAGTTTGAACCTTCTGGCTGCTTATTC
ACCATCTCTCCCTCACTATATTTGCTGAGCCAGCCTTGGCCCGTAGGAGTCTTCTCTAGG
AGTATTAGACAAGTTGGCATTTGATAAATTTCTGTCCTAACCAACCCCGTCTGAGACACGC
AACAGGGGATAGGCAAGGCAC
```

**Figure 2. SRA Study, Experiment, and Run records.** *Top panel.* The Study record (SRP000095) for 454 sequencing of James Watson's personal genome shown in the SRA browser. The record has links to display the two corresponding Experiments (*right arrow*) or to download the entire study (*diskette icon*). *Middle panel.* An experiment record (SRX000113) for James Watson's personal genome displayed in the Entrez SRA database with links to Reads (*right arrow*). *Bottom panel.* A Run (SRR000416) showing data for a single read (SRR000416.1) of the 16,772 reads from experiment SRX000065 shown in the SRA Run browser. The application read is the sequence determined for this spot in a single instrument run. The technical read is a four base tag specific to the platform. A signals table and intensity graph (not shown) that indicate light intensity for each base in the pyrosequencing reaction is also available for each 454 read.

Once installed Aspera Connect will launch to transfer data from SRA whenever a download link is clicked. SRA offers standard FASTA and the convenient and portable fastq format for download. The fastq format is ASCII text that includes the sequence plus the ASCII encoded quality scores.

## Submitting Data to SRA

SRA provides an interactive web-based interface for submissions that requires only a brief registration prior to submission. The Submissions tab on the SRA homepage accesses the registration and login page for SRA submissions (Figure 1, top panel). SRA also offers an automated submission pipeline for centers making multiple submissions. Detailed information on submitting to SRA is available in the SRA Submission Guidelines document.

[www.ncbi.nlm.nih.gov/Traces/sra/static/SRA\\_Submission\\_Guidelines.pdf](http://www.ncbi.nlm.nih.gov/Traces/sra/static/SRA_Submission_Guidelines.pdf)

## Summary

SRA data are rapidly dominating all other sequence data. Already the number of DNA bases available in SRA exceeds the number of bases in GenBank. In fact the output of a single important project, the 1000 genomes project ([www.1000genomes.org](http://www.1000genomes.org)), will produce more than 25 times the number of bases that are currently in GenBank by the time the project is completed. The NCBI and SRA will continue to support submission, retrieval, and analyses of these increasingly challenging and complex sequencing data. Means of displaying data, analyses, and integration of SRA data with other molecular databases will continue to improve making the SRA data a prominent part of the discovery system at the NCBI.

## New Databases and Tools

### Human Genome Build 37.1

Human genome build 37.1, the new Human Genome Reference Consortium assembly and annotation, is now displayed in the NCBI Entrez system and the NCBI Map Viewer site.

[www.ncbi.nlm.nih.gov/mapview/map\\_search.cgi?taxid=9606](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606)

## GenBank News

GenBank release 172.0 is incorporated into the NCBI and FTP sites (<ftp.ncbi.nih.gov/genbank/>). The current release includes data available as of June 10, 2009. Release notes ([gbrel.txt](#)) describing details of the release and upcoming changes are in the GenBank FTP directory.

NCBI is considering discontinuing the index files; affected users are encouraged to review the discussion of this change in the release notes and provide comments to the GenBank group.

## Updates and Enhancements

### HomoloGene

HomoloGene release 64 includes updated annotations for *Homo sapiens* (NCBI release 37.1), *Caenorhabditis elegans* (WS190, NCBI release 8.1), *Anopheles gambiae* (AgamP3.3, NCBI release 3.1), *Arabidopsis thaliana* (NCBI release 8.1), *Bos taurus* (NCBI release 3.1), and *Magnaporthe grisea* (NCBI release 3.1). The HomoloGene homepage has additional details.

[www.ncbi.nlm.nih.gov/homologene](http://www.ncbi.nlm.nih.gov/homologene)

## RefSeq

RefSeq Release 36, now available through NCBI Entrez and FTP (<ftp.ncbi.nlm.nih.gov/refseq/release/>) incorporates genomic, transcript, and protein data available as of July 2, 2009. It includes 12,141,825 records from 8,665 different species and strains. Changes since the previous release are described in the [notes](#) in the RefSeq FTP directory.

## BLAST

With the new BLAST 2.2.21 release, the BLAST+ command-line applications, written with the NCBI C++ toolbox, are now the major supported version of BLAST. The BLAST+ applications have a number of advantages over the older applications that include working more robustly with long sequences and database masking. The BLAST+ applications were described in the [January 2009](#) NCBI News. The FTP directory contains a complete user manual for the BLAST+ package.

[ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/user\\_manual.pdf](ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/user_manual.pdf)

## Influenza Virus Resource

The Influenza Virus Resource has an option for viewing “Sequences from Pandemic (H1N1) 2009 virus only” on the database search page.

[www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi?go=1](http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi?go=1)

The page also offers an option to exclude these sequences from search results if desired.

## PubMed Central

Are you interested in new titles added to PubMed Central? If so, the PMC RSS feed provides all new article titles as well as titles of newly scanned articles from archives.

[www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, PubChem, LinkOut, HomoloGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, July 2009

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: July 2, 2009; Updated: July 13, 2009.

### Featured Resource: The BioSystems Database of Biological Pathways

NCBI BioSystems is a new database that collects information on interacting sets of biomolecules involved in metabolic and signaling pathways, disease states, and other biological processes. BioSystems currently contains biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) and the EcoCyc (*Escherichia coli* K-12 MG1655) subset of the BioCyc databases and is designed to accommodate other data in the future. BioSystems is fully integrated with other databases in the Entrez system with links to related literature, genes, protein sequences, structures, chemical data, and to related BioSystems. Along with links to related data at the NCBI site, each BioSystem record provides links to detailed diagrams and annotations for individual pathways on the Web sites of the source databases. BioSystems adds an important new aspect to the NCBI system by linking many different kinds of molecular records in biochemical pathways and providing means to compare these pathways across organisms.

#### Searching BioSystems

BioSystems is available as part of the NCBI Entrez system and can be searched directly from the database page:

[www.ncbi.nlm.nih.gov/BioSystems/](http://www.ncbi.nlm.nih.gov/BioSystems/)

As with other Entrez databases, field-restricted queries in BioSystems give more precise search results. Many of the fields that work in the other Entrez molecular databases such as Organism and Title also are useful in BioSystems. All of the available fields can be browsed through the Preview/Index tab on the search page. The Limits tab provides a simple way to limit to certain types of records through pre-set fields. For example, the search can be limited to a specific source database, currently KEGG or EcoCyc. There is also a checkbox that allows restricting to organism-specific BioSystems, those with links to molecular records for a single species, or conserved BioSystems that group together orthologous organism specific pathways derived from KEGG reference pathways. Limiting to Conserved BioSystems provides more concise results and access to an overall summary of the pathway in all organisms. For example, a search with the phrase “purine metabolism” retrieves 3,509 records while limiting to conserved BioSystems or clicking the Conserved BioSystems filter tab provides a more concise set of only 26 records (Figure 1). As another example, the organism-specific record for purine metabolism in mouse is found directly with the following field-restricted query.

```
purine metabolism[Title] AND mouse[Organism]
```

#### BioSystems Records: Photosynthesis

The BioSystems database currently contains over 95 thousand pathways from the KEGG database including 255 conserved BioSystems and 286 *Escherichia coli* K-12 pathways from the EcoCyc portion of the BioCyc database. Figure 2 shows the BioSystems organism-specific record for photosynthesis from *Arabidopsis thaliana* (BioSystems ID 4001). The record has a description of the process from the source database, a thumbnail graphic that links to the larger diagram in the KEGG database, and a Tabbed Table of components and other aspects of the pathway (Genes, Proteins, Small Molecules, Related BioSystems, Citations and Comments). These components link to records in the NCBI Gene, Protein, PubChem, BioSystems, and PubMed databases.

The screenshot shows the BioSystems search interface. At the top left is the NCBI logo, and at the top right is the BioSystems logo. A search bar contains the text 'BioSystems' and 'for purine metabolism'. Below the search bar are buttons for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. The 'Display' section shows 'Summary' selected, 'Show 100', and 'Sort By'. A summary bar indicates 'All: 3509', 'Conserved BioSystems: 26', and 'Organism Specific BioSystems: 3483'. The results are on page 1 of 36. Four results are listed:

- 1: bsid215** Related BioSystems, Literature, Sequences, Small Molecules, Other Links
  - purine deoxyribonucleosides degradation**
  - Type: organism-specific biosystem
  - Description: E. coli can use all three naturally occurring purine deoxyribonucleosides (deoadenosine, deoxyguanosine, and deoxyinosine) as total sources of carbon and energy. All are cleaved by the same deoD-encoded phosphorylase, yielding the corresponding bases (adenine, guanine and hypoxanthine) and D-deoxyribose-1-phosphate,...
  - Organism: [Escherichia coli K-12](#)
  - Source: [BIOCYC \[ECO\\_PWY0-1297\]](#)
- 2: bsid870** Related BioSystems, Literature, Sequences, Small Molecules, Other Links
  - Purine metabolism**
  - Type: organism-specific biosystem
  - Organism: [Saccharomyces cerevisiae](#)
  - Source: [KEGG \[sce00230\]](#)
- 3: bsid307** Related BioSystems, Literature, Sequences, Small Molecules, Other Links
  - Purine metabolism**
  - Type: conserved biosystem
  - Source: [KEGG \[ko00230\]](#)
- 4: bsid44978** Related BioSystems, Literature, Sequences, Small Molecules, Other Links
  - Purine metabolism**
  - Type: organism-specific biosystem
  - Organism: [Bartonella bacilliformis KC583](#)
  - Source: [KEGG \[bbk00230\]](#)

**Figure 1.** Search results in BioSystems using purine metabolism as a query. Both EcoCyc (BioCyc) and KEGG source records are found. The Conserved BioSystems filter tab at the top of the results selects KEGG reference pathways that summarize orthologous pathways for large numbers of organisms.



UID:bsid4001 [Related BioSystems, Literature, Sequences, Small Molecules, Other Links](#)

## Photosynthesis

**Type :** organism-specific biosystem

**Description :** Photosynthesis in green plants and specialized bacteria is the process of utilizing light energy to synthesize organic compounds from carbon dioxide and water. It consists of the light dependent part (light reaction) and the light independent part (dark reaction, carbon fixation). The light reaction takes place in thylakoid, a membrane-bound compartment inside chloroplasts and cyanobacteria. The light energy is used by photosystems I and II to generate proton motive force and reducing power (NADPH or NADH). The proton motive force is used by ATP synthase to generate ATP, essentially in the same [more...](#)

**Organism:** [Arabidopsis thaliana](#)

**Source :** [KEGG \[ath00195\]](#)

**Genes** | **Proteins** | **Small Molecules** | **Related BioSystems** | **Citations** | **Comments**

Only top 50 records shown, click to view and/or save records in Entrez Gene [Highlight Selected Records in Source Database](#)

Gene ID	External ID	Name
<a href="#">838400</a>	AT1G03130	PSAD-2 (photosystem I subunit D-2); PSAD-2
<a href="#">839449</a>	AT1G03600	photosystem II family protein; AT1G03600
<a href="#">837178</a>	AT1G06680	PSBP-1 (PHOTOSYSTEM II SUBUNIT P-1); poly(U) binding; PSBP-1
<a href="#">837639</a>	AT1G10960	ATFD1 (FERREDOXIN 1); 2 iron, 2 sulfur cluster binding / electron carrier/ iron-sulfur cluster binding; ATFD1
<a href="#">837974</a>	AT1G14150	oxygen evolving enhancer 3 (PsbQ) family protein; AT1G14150
<a href="#">838139</a>	AT1G15700	ATPC2; enzyme regulator; ATPC2
<a href="#">838591</a>	AT1G20020	FNR2 (FERREDOXIN-NADP(+)-OXIDOREDUCTASE 2); NADPH dehydrogenase/ oxidoreductase/ poly(U) binding; FNR2
<a href="#">838622</a>	AT1G20340	DRT112; copper ion binding / electron carrier; DRT112
<a href="#">839918</a>	AT1G30380	PSAK (photosystem I subunit K); PSAK
<a href="#">839930</a>	AT1G30510	ATRFNR2 (ROOT FNR 2); FAD binding / NADP or NADPH binding / electron carrier/ ferredoxin-NADP+ reductase/ oxidoreductase; ATRFNR2
<a href="#">840021</a>	AT1G31330	PSAF (photosystem I subunit F); PSAF
<a href="#">840149</a>	AT1G32550	ferredoxin family protein; AT1G32550

Only top 50 records shown, click to view and/or save records in Entrez Gene

**Figure 2.** The full BioSystems record for photosynthesis for *Arabidopsis thaliana*. The description and thumbnail image are from the KEGG database and link to the KEGG site. The Tabbed Table at the bottom provides access to corresponding Gene, Protein, Small Molecule (PubChem), Related BioSystem, and Citations (PubMed) records for components of the pathway at NCBI

## Pathway Diagram

The full [photosynthesis diagram](#) is available at the KEGG site and is linked to the thumbnail image in the BioSystems record. The cartoon of the chloroplast thylakoid membrane at the KEGG site shows the multi-subunit protein complexes of the photosynthetic system (photosystems I and II, cytochrome b6/f, electron transport system, and the f-type ATPase). The boxes in the tables below the cartoon represent the genes or protein components of the system. Each of the green-filled boxes represents one or more *Arabidopsis* genes. The unfilled boxes represent genes or functions that are not known from *Arabidopsis*. In this case these missing components (PsbU, PsbV, PsbX, Psb28-2, PsaM, PsaX, PetL, and PetM) are specific to certain cyanobacterial systems and are available in the corresponding Conserved BioSystem or reference pathway.

## Tabbed Table of BioSystem Components

The Tabbed Table provides access to the components and other information linked to the pathway in the NCBI databases. Tabs include Genes, Proteins, Small Molecules, Related BioSystems, and Citations. The *Arabidopsis* genes and proteins corresponding to the filled boxes in the pathway diagram are listed in the Genes and Proteins tabbed sections of the table. The Small Molecule tab provides access to substrates, inhibitors, cofactors, and other non-protein entities involved in pathways from the NCBI PubChem database. Clicking on any entry will link to the corresponding Gene, Protein, or PubChem record. The link at the top of the table retrieves all records in the active tab. For example, clicking the top link in the Genes tab easily retrieves all 73 *Arabidopsis* genes involved in the photosynthetic pathway; clicking this link in the Small Molecules tab retrieves cofactors and substrates from PubChem (Figure 3). The Related BioSystems tab expands to display three types of related BioSystems: Linked BioSystems, Similar BioSystems, and Conserved BioSystems. Linked BioSystems are pathways that interact with the current pathway often providing substrates or receiving products of the current pathway. In this case, these are the antennal proteins light collecting system of photosynthesis (BioSystems ID 4002) and carbon fixation in photosynthetic organisms (BioSystems ID 4065). Similar BioSystems are pathways that share at least one identical protein sequence from the same source organism. Oxidative phosphorylation (BioSystems ID 4000) is the Similar Pathway to photosynthesis because it shares the subunits of the f-type ATPase. Finally, the Conserved BioSystem (BioSystems ID 340) is the reference pathway record for photosynthesis that gathers all orthologous pathways. The Conserved BioSystem includes the cyanobacterial components represented by unfilled boxes in the organism-specific pathway at the KEGG site. The remaining populated tab for the photosynthesis record, Citations, provides literature citations from the source database with links to the corresponding records in PubMed.

## Links Menus: Related Data

In addition to the linked data available in the Tabbed Table, each BioSystems record has related data available as Links menus at the upper right of the summary or full BioSystems record (Figure 4). These are the Related BioSystems, Literature, Sequences, Small Molecules, and Other Links menus. The BioSystems, Literature, and Small Molecules menus access the same related data available in the Tabbed Table. Additional related data in the Sequences menu that are not available in the Tabbed Table are HomoloGene, Protein Clusters, and Conserved Domain records that are linked from the proteins in the pathway.

The HomoloGene and Protein Clusters related data identify homologs from selected eukaryotic and microbial genomes and allow comparisons of pathways across taxa. The linked Conserved Domains can give additional structural and functional information about the proteins in the pathway. The Other Links menu connects the BioSystem to the PubChem BioAssay database for small molecules in the pathway that have been classified as active in one of the assays. There are also links to the NCBI taxonomy database for organism-specific BioSystems, and to the Structure database if any of the proteins in the pathway are linked to a three-dimensional structure. The following example shows how to use these related data to find structures for the photosynthetic complexes starting from the *Arabidopsis* organism-specific record.

**Top Panel: PubChem Records**

Structure	Substance ID	Compound ID	External ID	Name
	3304	5957	C00002	Adenosine triphosphate; Triphosphaden; Triphosaden; Myotriphos
	3307	5884	C00005	NADPH; TPNH; NChemBio.2007.9-comp16; CHEBI:16474
	3308	5886	C00006	Codehydrase II; NADP; NADP+; Spectrum_001530

**Center Panel: Gene and PubChem Records**

All: 73   Current Only: 73   Genes Genomes: 73   SNP GeneView: 0

Items 1 - 73 of 73

- 1: psbA**  
photosystem II protein D1 [*Arabidopsis thaliana*]  
Other Aliases: ArthCp002  
Genomic context: Chloroplast  
Annotation: NC\_000932.1 (383..1444, complement)  
GeneID: 844802
- 2: psbK**  
PSII K protein [*Arabidopsis thaliana*]  
Other Aliases: ArthCp005  
Other Designations: photosystem II protein K  
Genomic context: Chloroplast  
Annotation: NC\_000932.1 (7017..7202)  
GeneID: 844795
- 3: psbI**  
photosystem II protein I [*Arabidopsis thaliana*]  
Other Aliases: ArthCp006  
Genomic context: Chloroplast  
Annotation: NC\_000932.1 (7583..7693)  
GeneID: 844794

**Right Panel: Detailed PubChem Records**

Items 1 - 11 of 11   One page.

- 1: CID: 5884**  
NADPH; TPNH; NChemBio.2007.9-comp16 ...  
IUPAC: [[[(2R,3R,4R,5R)-5-(6-aminopurin-9-yl)-3-hydroxy-4-phosphonoxyoxolan-2-yl]methoxy-hydroxyphosphoryl] [(2R,3S,4R,5R)-5-(3-carbamoyl-4H-pyridin-1-yl)-3,4-dihydroxyoxolan-2-yl]methyl hydrogen phosphate  
MW: 745.420883 g/mol | MF: C<sub>21</sub>H<sub>30</sub>N<sub>7</sub>O<sub>17</sub>P<sub>3</sub>
- 2: CID: 5886**  
Codehydrase II; NADP; NADP+ ...  
IUPAC: [[[(2R,3R,4R,5R)-5-(6-aminopurin-9-yl)-3-hydroxy-4-phosphonoxyoxolan-2-yl]methoxy-hydroxyphosphoryl] [(2R,3S,4R,5R)-5-(3-carbamoylpyridin-1-ium-1-yl)-3,4-dihydroxyoxolan-2-yl]methyl hydrogen phosphate  
MW: 744.412943 g/mol | MF: C<sub>21</sub>H<sub>29</sub>N<sub>7</sub>O<sub>17</sub>P<sub>3</sub><sup>+</sup>
- 3: CID: 5957**  
Adenosine triphosphate; Triphosphaden; Triphosaden ...  
IUPAC: [[[(2R,3S,4R,5R)-5-(6-aminopurin-9-yl)-3,4-dihydroxyoxolan-2-yl]methoxy-hydroxyphosphoryl] phosphono hydrogen phosphate  
MW: 507.181023 g/mol | MF: C<sub>10</sub>H<sub>16</sub>N<sub>5</sub>O<sub>13</sub>P<sub>3</sub>  
Tested in BioAssays: All: 15, Active: 2; BioActivity Analysis

**Bottom Panel: Related BioSystems**

Conserved BioSystems   **Linked BioSystems**   Similar BioSystems

Click to view and/or save records in Entrez BioSystems

Biosystem ID	Source	External Accession	Name
4002	KEGG	ath00196	Photosynthesis - antenna proteins
4065	KEGG	ath00710	Carbon fixation in photosynthetic organisms

**Figure 3. The Tabbed Table and linked records from the *Arabidopsis* photosynthesis BioSystem.** *Top panel.* Tabbed Table with the Small Molecules tab selected showing PubChem records linked to the pathway. *Center panel.* Linked Gene and PubChem records in the corresponding Entrez databases. All genes and cofactors are available in Entrez. *Bottom panel.* The Related BioSystems tab with the Linked BioSystems tab selected showing the two systems or pathways, antenna proteins and carbon fixation, that connect to photosynthesis.

The figure displays three overlapping screenshots of the NCBI BioSystems interface for the record 'Photosynthesis' (bsid4001).

- Top Screenshot:** Shows the 'Sequences' menu open with options: Conserved Domains, Genes, Homologene, Protein Clusters, and Proteins.
- Middle Screenshot:** Shows the 'Links' menu open with options: Conserved BioSystems, Linked BioSystems, and Similar BioSystems. A red arrow points to the 'Conserved BioSystems' link.
- Bottom Screenshot:** Shows the 'Other Links' menu open with options: BioAssays via Actives and Structures. A red arrow points to the 'Structures' link.

Below the menu navigation, a list of protein structures is shown:

ID	Structure Name	EC	Proteins	Chemicals	MMDB ID
1: 2ZT9	Crystal Structure Of The Cytochrome B6f Complex From Nostoc Sp. Pcc 7120	1.10.99.1	8	7	69519
2: 2AXT	Crystal Structure Of Photosystem Ii From Thermosynechococcus Elongatus		36	15	36695
3: 2B50	Ferredoxin-Nadp Reductase	1.18.1.2	2	2	35698
4: 1Q90	Structure Of The Cytochrome B6f (Plastohydroquinone : Plastocyanin Oxidoreductase) From Chlamydomonas Reinhardtii	1.10.99.1	9	8	25730
5: 1IZL	Crystal Structure Of Photosystem Ii		28	7	21630

**Figure 4. The Links menus from BioSystems records.** The Sequences menu has Conserved Domains, HomoloGene, and Protein Clusters links in addition to the Proteins that are also in the Tabbed Table. Following the Conserved BioSystems link (left arrow) from the Arabidopsis record allows access through the Other Links menu to protein structures (right arrow) from the reference pathway for photosynthesis.

## Example: Using Related Data to Find Three-Dimensional Structures

The links to data available through the Tabbed Table and Links menus can be used to find important related data such as homologous pathways, genes, and proteins in other species as well additional structural and functional information. For example, using the Tabbed Table and the Links menus, it is easy to find the three-dimensional structures of some of the protein complexes involved in photosynthesis. As mentioned previously, structure records are linked to the BioSystems pathway through the “Other Links” menu. However, the *Arabidopsis* photosynthesis pathway has no direct structure links because none of the photosynthesis proteins with structures are from this species. Following the Related BioSystems link to the Conserved BioSystem gives access to proteins from orthologous pathways in all species including those with structure records. The Conserved BioSystem for photosynthesis (BioSystems ID 340) has links to five structures of photosynthetic protein complexes through the Other Links menu (Figure 4). Four of these structures are from cyanobacteria, and one is from the green alga, *Chlamydomonas reinhardtii*.

## Summary

The BioSystems database adds a new dimension to NCBI resources by connecting molecular, bioassay, and literature data to biological pathways and processes. This enables searches that for the first time can find all gene, protein records, and small molecules in a pathway for a particular organism. Moreover, integration with other NCBI databases allows comparisons of pathways across taxa and provides access to other data on structure and function of the biomolecules involved. Expanding the coverage of the Entrez system to pathways provides new connections among databases that should greatly increase the power of Entrez as a discovery system.

## New Databases and Tools

### New BankIt Submission Tool

A new version of the BankIt sequence submission tool is available for testing. The new tool will eventually replace current version of BankIt after the test period. Please see the New BankIt page for more information: [www.ncbi.nlm.nih.gov/WebSub/?tool=genbank](http://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank)

### Bookshelf

The book, *Familial Cancer Syndromes* has been added to the NCBI Bookshelf. To browse this book go to [www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=famcan](http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=famcan)

### Microbial Genomes

Twenty-one finished microbial genomes were released between May 29 and July 6, 2009. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: [ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

## GenBank News

GenBank release 172.0 is available on the NCBI Web and FTP sites. The current release includes information available as of June 10, 2009. Release notes are available on the on the NCBI ftp site: [ftp.ncbi.nih.gov/genbank/gbrel.txt](ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

NCBI is considering ceasing support for the index files; affected users are encouraged to review the discussion of this change in the release notes and provide comments to the GenBank group.

## Updates and Enhancements

### PubChem

The American Library Association has selected PubChem as one of the Best Free Reference Web Sites 2009. More PubChem announcements are available at: [pubchem.ncbi.nlm.nih.gov/pcnews.html](http://pubchem.ncbi.nlm.nih.gov/pcnews.html) and as an RSS feed.

### BLAST

As previewed in the [May 2009 NCBI News](#), COBALT multiple-sequence alignments can now be generated from protein BLAST results by clicking on the “Multiple Alignment” link. A direct submission form for generating protein multiple alignments using COBALT is also available in the Specialized BLAST section of the [BLAST Homepage](#). The [BLAST News](#) page provides additional details about COBALT. The BLAST homepage also provides a “Tip of the Day” for more efficient use of the BLAST tool.

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, PubChem, LinkOut, HomoloGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, June 2009

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, MS<sup>2</sup>

Created: June 3, 2009; Updated: June 10, 2009.

### Featured Resource: An Expanded Set of Discovery Components in the Entrez System

Several new features of the NCBI Entrez Web service are aspects of the ongoing Discovery Initiative described in the February and March 2009 Issues of the *NCBI News*. These new discovery components in the literature and sequence databases make the most relevant and interesting results more obvious and readily accessible.

There are three main categories of discovery components that now appear: Sensors, Database Ads, and Analysis Tools.

A sensor detects certain types of search terms and provides access to potentially more relevant results. For PubMed, new sensors include a Citation Sensor that is activated when someone searches with a literature citation and an Accession Sensor that provides a direct link to the sequence databases when someone searches with an NCBI sequence identifier. A variable type of sensor, the Hot Topic Sensor, also appears in PubMed. This new sensor that was inspired by the rapidly changing state of data for H1N1 influenza virus during the current outbreak appears for searches relevant to the recently added H1N1 viral sequences but in the future will be tailored to respond to other topical issues. The new more precise Gene Sensor that debuted in the PubMed database in January is now available in the protein and nucleotide databases.

A Database Ad promotes related information in other databases that may be more useful or may provide unexpected connections. New Database Ads in PubMed highlight the full-text PubMed Central database. The PubMed Central Ad that appears with PubMed results displays articles that are also available in full-text in the PubMed Central database. In the Abstract Plus View, the ads link to articles in PubMed Central that cite the PubMed record. A new Structure Ad appears in both the PubMed and sequence databases for articles that report a 3-D structure or for sequences derived from structure records. Viral Genome Resources Ads for influenza, dengue, SARS, and retroviruses such as HIV now appear in the sequence databases on sequence records of viral origin.

Analysis Tools that provide on-the-fly analysis are important components of the discovery initiative. Sequence analysis tools available for sequence records now include a direct link that will perform a BLAST search with the sequence as well as a link to run a conserved domain search for protein records. These new links accompany the direct link to design primers that has already been present on nucleotide records for several months.

All of these new discovery components are designed to help researchers find the most relevant information in the NCBI databases in the fewest mouse clicks.

### Sensors in Entrez

As mentioned above, new sensors in Entrez include the Citation Sensor, the Accession Sensor, the Hot Topic Sensor for the H1N1 influenza virus, and the new Gene Sensor for sequence databases.

The new Citation Sensor automatically returns results from the PubMed Citation Matcher when it detects a query resembling a literature citation in a PubMed Search. Citation queries often retrieve irrelevant results when

entered as a general PubMed search. The Citation Matcher service, now available as a part of the PubMed Advanced interface, is designed specifically for matching literature citations with PubMed records:

[www.ncbi.nlm.nih.gov/pubmed/advanced](http://www.ncbi.nlm.nih.gov/pubmed/advanced)

The Citation Sensor makes the power of the Citation Matcher more widely available. A minimal citation query would normally include an author name and a publication year or a journal name and publication year. For example, a search with “Lander 2001 Nature” quickly finds the *Nature* publication on the human genome sequence (Initial sequencing and analysis of the human genome sequence) as one of three articles found by the Citation Sensor (Figure 1, top panel). In comparison, the direct PubMed search retrieves 14 records, 11 of which are not from the journal *Nature*.

The Accession Sensor in PubMed is designed to provide relevant results when a PubMed search contains a sequence accession number. While GenBank sequence accession numbers reported in PubMed articles will find the source publication when used directly as a PubMed query, many accessions have no corresponding publication. Derivative sequence records such as NCBI Reference Sequences are often not associated directly with any PubMed records. Also, in many cases the goal of searching with accession identifiers is to find the sequence record itself and not the publication. In all of the above situations the accession sensor is quite useful in providing relevant results.

The middle panel of Figure 1 shows the results obtained in PubMed searching with a GenBank accession for the human dopamine D2 receptor (DRD2) mRNA (X51362). The search retrieves two PubMed citations that reference the accession as expected. The citation sensor in this case provides a convenient means to directly retrieve the sequence record without performing a separate search or following a link from one of the publications. The bottom panel of Figure 1 shows the results obtained using the corresponding NCBI Reference Sequence accession identifier for the DRD2 mRNA (NM\_000795). There are no results found in PubMed since the RefSeq identifier is not cited in any publications or included in the abstract. However, the accession sensor provides access directly to the correct sequence record.

Another kind of sensor, the Hot Topic sensor, now appears in PubMed in response to increased searches related to the recent H1N1 influenza outbreak. In its present form, the sensor appears at the top of the right hand discovery column when it detects search terms that indicate interest in the H1N1 influenza sequences, and provides a link to the specialized H1N1 Influenza page described in the May, 2009 NCBI News (Figure 2, top panel). The Hot Topic Sensor will be deployed in different formats in response to current events in order to provide easy access to topical results.

The Gene Sensor that has been active in PubMed for several months is now in the protein and nucleotide databases. As in PubMed, the Gene Sensor is triggered by a gene symbol in a search. The older sequence database gene search feature remains active and will still return results from the gene database when the search does not trigger the Gene Sensor. The middle panel of Figure 2 shows the Gene Sensor triggered in the nucleotide database by a search with the mammalian gene symbol AFM. The sensor allows retrieval of relevant gene records with access to nucleotide and protein sequences while the direct nucleotide results contain large numbers of irrelevant matches. The gene search results triggered by a search with “afamin” shown in the bottom panel of Figure 2 also provide a better set of results than the direct nucleotide search.

## Database Ads

Two new Database Ads for the full-text PubMed Central database appear in PubMed. A link appears in all PubMed search results (Figure 2, top panel) displaying all articles that are also available in PubMed Central. Another ad for PubMed Central appears in the Abstract Plus record view and links to articles also in PubMed Central that cite the current article (Figure 3, top panel). This not only provides rapid access to full-text articles, but also offers another mechanism to expand the search to potentially related articles. As PubMed Central



**All: 14** **Review: 4** ✕

**We found 3 articles in Nature 2001 by Lander:**

[Linkage disequilibrium in the human genome.](#) Reich DE et al. *Nature*. (2001)

[A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.](#) Sachidanandam R et al. *Nature*. (2001)

[Initial sequencing and analysis of the human genome.](#) Lander ES et al. *Nature*. (2001)

Items 1 - 14 of 14

**1: [Automation, parallelism, and robotics for proteomics.](#)**  
 Alterovitz G, Liu J, Chow J, Ramoni MF.  
 Proteomics. 2006 Jul;6(14):4016-22. Review.  
 PMID: 16786489 [PubMed - indexed for MEDLINE]  
[Related Articles](#)

**Are you looking for a sequence?**  
 Result for term [X51362](#) found in the Nucleotide database

▶ [Human mRNA for dopamine D2 receptor \[Homo sapiens\]](#)

Items 1 - 2 of 2 One page.

**1: [Sequence specific binding of cytosolic proteins to a 12 nucleotide sequence in the 5' untranslated region of FMR1 mRNA.](#)**  
 Iber H.  
 Biochim Biophys Acta. 1996 Dec 11;1309(3):167-73.  
 PMID: 8982249 [PubMed - indexed for MEDLINE]  
[Related Articles](#)

**2: [Human retina D2 receptor cDNAs have multiple polyadenylation sites and differ from a pituitary clone at the 5' non-coding region.](#)**  
 Robakis NK, Mohamadi M, Fu DY, Sambamurti K, Refolo LM.  
 Nucleic Acids Res. 1990 Mar 11;18(5):1299. No abstract available.  
 PMID: 2138729 [PubMed - indexed for MEDLINE]  
[Related Articles](#) [Free article in PMC](#) | [at journal site](#)

**The following term was not found in PubMed: NM\_000795.**  
 See [Details](#). No items found.

**Are you looking for a sequence?**  
 Result for term [NM\\_000795](#) found in the Nucleotide database

▶ [Homo sapiens dopamine receptor D2 \(DRD2\), transcript variant 1, mRNA \[Homo sapiens\]](#)

**Figure 1. Citation Sensor and Accession Sensor in PubMed.** *Top panel.* A search with “Lander 2001 Nature” showing the Citation Sensor. The Citation Sensor shows a more relevant set of results including the paper reporting the human genome sequence. *Center panel.* Accession Sensor triggered by a search with GenBank Accession number X51362. The sensor provides a direct link to the sequence record. The PubMed results contain the two papers linked to the nucleotide record. *Bottom panel.* Accession Sensor triggered by a search with Reference Sequence accession number NM\_000795. The sequence record has no linked articles in PubMed, but the sensor provides a direct link to the record in the nucleotide database.

Items 1 - 20 of 2463 Page 1 of 124 Next

**2009 H1N1 Flu Sequences**  
See the latest influenza A (H1N1) sequences from the 2009 outbreak.

**Titles with your search terms**

- ▶ Evolutionarily conserved protein sequences of influe [PLoS ONE. 2007]
- ▶ Nuclear import of influenza A viral ribonucleoprotein comple [Viro J. 2007]
- ▶ Heterologous influenza vRNA segments with identical n [Viro J. 2008]

» See more...

**1:** [Genetic diversity of swine influenza viruses isolated from pigs during 2000 to 2005 in Thailand.](#)  
Takemae N, Parchariyanon S, Damrongwatanapokin S, Uchida Y, Ruttanapumma R, Watanabe C, Yamaguchi S, Saito T.  
Influenza Other Respi Viruses. 2008 Sep;2(5):181-9.  
PMID: 19453423 [PubMed - in process]  
[Related Articles](#)

**2:** [Construction and cellular immune response induction of HA-based alphavirus replicon vaccines against human-avian influenza \(H5N1\).](#)  
Yang SG, Wo JE, Li MW, Mi FF, Yu CB, Lv GL, Cao HC, Lu HF, Wang BH, Zhu H, Li LJ.  
Vaccine. 2009 May 15. [Epub ahead of print]  
PMID: 19450640 [PubMed - as supplied by publisher]  
[Related Articles](#)

**3:** [Non-infectious plasmid engineered to simulate multiple viral threat agents.](#)  
Carrera M, Sagripanti JL.  
J Virol Methods. 2009 Jul;159(1):29-33. Epub 2009 Mar 3.  
PMID: 19442841 [PubMed - in process]  
[Related Articles](#)

**729 free full-text articles in PubMed Central**

- ▶ Inferring stabilizing mutations from protein phylog [PLoS Comput Biol. 2009]
- ▶ Virus variation resources at the National Center f [BMC Microbiol. 2009]
- ▶ Panorama phylogenetic diversity and distribution of Type A [PLoS ONE. 2009]

» See all (729)...

Gene Information

**AFM** afamin [Homo sapiens]

This gene is a member of the albumin gene family, which is comprised of four genes that localize to Location: 4q11-q13 | Also known as: ALB2; ALBA; ALF; MGC125338; MGC125339

▶ **afm** in [Homo sapiens](#) | [Mus musculus](#) | [Rattus norvegicus](#) | [All 6 Gene records](#) Gene

Items 1 - 212 of 212 One page.

**1:** [DS995480](#) Reports Links  
Collinsella stercoris DSM 13279 Scfld7 genomic scaffold, whole

Gene Information

This search in Gene shows [12 results](#), including:

[AFM](#) (*Homo sapiens*): afamin  
[Afm](#) (*Mus musculus*): afamin  
[Afm](#) (*Rattus norvegicus*): afamin Gene

Items 1 - 67 of 67 One page.

**1:** [NW\\_001838914](#) Reports Links  
Homo sapiens chromosome 4 genomic contig. alternate assembly

**Figure 2. Hot Topic Sensor, PubMed Central ad in PubMed, and Gene Sensor in PubMed.** *Top panel.* PubMed results for a search with “influenza A” showing the Hot Topic Sensor link to the Flu sequences at the top of the right-hand column (boxed in red) and an ad for the 729 articles in the results that have free full-text in PubMed Central at the bottom (boxed in red). *Middle panel.* New Gene Sensor display in the nucleotide database triggered by a search with the mammalian gene symbol “AFM” linking to more relevant results in Gene. *Bottom panel.* Older Gene search results in nucleotide triggered by a search with the gene product name “afamin”. The top three results in Gene may be more relevant than the corresponding nucleotide results.

continues to expand the number of citations, it may also provide a useful measure of the significance of a particular article.

A Structure Ad now appears in both the PubMed and sequence database record views (Figure 3). This ad features a thumbnail image of 3-D molecular structures reported in the PubMed article or linked directly to the sequence record. The image is linked to the corresponding record in the structure database. From here the structure may be displayed and manipulated in NCBI's Cn3D structure viewer. In the sequence databases, records for influenza, dengue viruses, SARS, and retroviruses like HIV now display an ad for the taxon-specific viral genome resources area of the NCBI Web site. An example of the ad is shown in the bottom panel of Figure 3 for an influenza virus sequence. The viral resources pages have collections of viral sequences, genotyping and other specialized tools that virus researchers may find more useful than those within the general Entrez.

## Analysis Tools

Direct links to sequence analysis tools in sequence records provide a means to instantly generate sequence-specific reagents through Primer-BLAST and update the annotation on all nucleotide and protein records through the ability to perform a live BLAST or conserved domain database search (Figure 3, bottom panel). Up to 20% of NCBI BLAST searches use NCBI database identifiers or copy-pasted NCBI formatted sequences as queries; the direct link to BLAST now makes it much easier to perform BLAST searches with NCBI database records.

## Summary

New Discovery components in the NCBI System – Sensors, Database Ads, and Analysis tools – make the Entrez system more powerful and easier to use by providing context sensitive results that traverse traditional database boundaries. These components not only make it possible to find relevant information in fewer steps but also help make more obvious unanticipated connections that are often essential to scientific discovery.

## New Databases and Tools

### BioSystems

NCBI BioSystems is a new database designed to aggregate biosystems information from collaborating public databases. BioSystems is a centralized repository of data that connects the biosystem records with associated literature, molecular, and chemical data throughout the Entrez system and facilitates computation on biosystems data. The NCBI BioSystems database currently contains biological pathways from the KEGG and BioCyc databases and is designed to accommodate other types of biosystems. Detailed diagrams and annotations for individual biosystems are available on the Web sites of the source databases. Links to Biosystems are now available from records in the NCBI Gene, HomoloGene, OMIM, and Protein Clusters databases. For more information, please see the BioSystems homepage: [www.ncbi.nlm.nih.gov/biosystems/](http://www.ncbi.nlm.nih.gov/biosystems/)

### Genome Resources

NCBI's Genome Resource pages provide a comprehensive guide for a specific organism including links to NCBI resources as well as outside groups and consortia. New genome resource pages are available for the Pea Aphid (*Acyrtosiphon pisum*) and goat (*Capra hircus*). Links can be found under the "Organism-Specific" section of the Genomic Biology page: [www.ncbi.nlm.nih.gov/Genomes/](http://www.ncbi.nlm.nih.gov/Genomes/).

### Microbial Genomes

Twenty-one finished microbial genomes were released between April 30 and May 28. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp://ncbi.nih.gov/genbank/genomes/>

1: [Nature](#). 2006 Nov 16;444(7117):378-82. nature [Links](#)

**Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors.**

[Yamada S](#), [Suzuki Y](#), [Suzuki T](#), [Le MQ](#), [Nidom CA](#), [Sakai-Tagawa Y](#), [Muramoto Y](#), [Ito M](#), [Kiso M](#), [Horimoto T](#), [Shinya K](#), [Sawada T](#), [Kiso M](#), [Usui T](#), [Murata T](#), [Lin Y](#), [Hay A](#), [Haire LF](#), [Stevens DJ](#), [Russell RJ](#), [Gamblin SJ](#), [Skehel JJ](#), [Kawaoka Y](#).

Division of Virology, Department of Microbiology and Immunology, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan.

H5N1 influenza A viruses have spread to numerous countries in Asia, Europe and Africa, infecting not only large numbers of poultry, but also an increasing number of humans, often with lethal effects. Human and avian influenza A viruses differ in their recognition of host cell receptors: the former preferentially recognize receptors with saccharides terminating in sialic acid- $\alpha$ 2,6-galactose (SA $\alpha$ 2,6Gal), whereas the latter prefer those ending in SA $\alpha$ 2,3Gal (refs 3-6). A conversion from SA $\alpha$ 2,3Gal to SA $\alpha$ 2,6Gal recognition is thought to be one of the changes that must occur before avian influenza viruses can replicate efficiently in humans and acquire the potential to cause a pandemic. By identifying mutations in the receptor-binding haemagglutinin (HA) molecule that would enable avian H5N1 viruses to recognize human-type host cell receptors, it may be possible to predict (and thus to increase preparedness for) the emergence of pandemic viruses. Here we show that some H5N1 viruses isolated from humans can bind to both human and avian receptors, in contrast to those isolated from chickens and ducks, which recognize the avian receptors exclusively. Mutations at positions 182 and 192 independently convert the HAs of H5N1 viruses known to recognize the avian receptor to ones that recognize the human receptor. Analysis of the crystal structure of the HA from an H5N1 virus used in our genetic experiments shows that the locations of these amino acids in the HA molecule are compatible with an effect on receptor binding. The amino acid changes that we identify might serve as molecular markers for assessing the pandemic potential of H5N1 field isolates.

PMID: 17108965 [PubMed - indexed for MEDLINE]

**Related articles**

- An avian influenza H5N1 virus that binds to a human-type receptor. [J Virol. 2007]
- Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus. [Science. 2006]
- Positive selection at the receptor-binding site of haemagglutinin H5 in viral sequence [J Gen Virol. 2008]
- Review** Evolving complexities of influenza virus and its receptors. [Trends Microbiol. 2008]
- Review** [Influenza virus receptors in the human airway] [Jirusu. 2006]

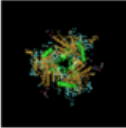
» See reviews... | » See all...

**Cited by 35 PubMed Central articles**

- Conserved amino acid markers from past influenza pandemic strains. [BMC Microbiol. 2009]
- Highly pathogenic avian influenza virus subtype H5N1 in Africa: a comprehensive phyloger [PLoS ONE. 2009]
- Molecular characterization of highly pathogenic H5N1 avian influenza A viruses isolated fr [PLoS ONE. 2009]

» See all...

**Structures reported by this article**

 **Influenza Virus (Vn1194) H5 Ha**  
PDB: 2IBX  
Source: Influenza A virus  
Method: X-Ray Diffraction |  
Resolution: 2.8 Å

PDB: 2IBXF

**Chain F, Influenza Virus (Vn1194) H5 Ha** Change Region Shown   
 Customize View

[Comment](#) [Features](#) [Sequence](#)

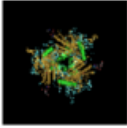
LOCUS	2IBX_F	160 aa	linear	VRL 24-SEP-2008
DEFINITION	Chain F, Influenza Virus (Vn1194) H5 Ha.			
ACCESSION	2IBX_F			
VERSION	2IBX_F GI:119390086			
DBSOURCE	pdb: molecule 2IBX, chain 70, release Aug 27, 2007; deposition: Sep 12, 2006; class: VirusVIRAL PROTEIN; source: Mol_id: 1; Organism_scientific: Influenza A Virus; Organism_common: Virus; Strain: H5n1 (Vn1194); Gene: Ha; Mol_id: 2; Organism_scientific: Influenza A Virus; Organism_common: Virus; Strain: H5n1 (Vn1194); Gene: Ha; Exp. method: X-Ray Diffraction.			
KEYWORDS	.			
SOURCE	Influenza A virus			
ORGANISM	Influenza A virus Viruses; ssRNA negative-strand viruses; Orthomyxoviridae; Influenzavirus A.			
REFERENCE	1 (residues 1 to 160)			
AUTHORS	Yamada,S., Suzuki,Y., Suzuki,T., Le,M.Q., Nidom,C.A., Sakai-Tagawa,Y., Muramoto,Y., Ito,M., Kiso,M., Horimoto,T., Shinya,K., Sawada,T., Kiso,M., Usui,T., Murata,T., Lin,Y., Hay,A., Haire,L.F., Stevens,D.J., Russell,R.J., Gamblin,S.J., Skehel,J.J. and Kawaoka,Y.			
TITLE	Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors			
JOURNAL	Nature 444 (7117), 378-382 (2006)			
PUBMED	<a href="#">17108965</a>			
REFERENCE	2 (residues 1 to 160)			
AUTHORS	Yamada,S., Russell,R.J., Gamblin,S.J., Skehel,J.J. and Kawaoka,Y.			
TITLE	Direct Submission			
JOURNAL	Submitted (12-SEP-2006)			
COMMENT	SEQRES.			

**Sequence Analysis Tools**

[BLAST Sequence](#)  
Find regions of similarity between this sequence and other sequences using BLAST.

[Conserved Domains](#)  
View conserved domains detected in this protein sequence using CD-search.

**Protein 3D Structure**

 **Influenza Virus (Vn1194) H5 Ha**  
PDB: 2IBX  
Source: Influenza A virus  
Method: X-Ray Diffraction  
Resolution: 2.8 Å

**Influenza Viral Resource**  
Flu-related NCBI resources including sequences, alignments, phylogeny and literature.

**Figure 3. PubMed Abstract Plus and Protein GenPept view showing database ads and analysis tools.** *Top panel.* Abstract Plus for an article reporting the 3-D structure of influenza haemagglutinin. The record has an ad for the 35 PubMed Central articles that cite the current article and an ad for the corresponding structures in the NCBI Structure database (both boxed in red). *Bottom panel.* A protein record for one of the influenza haemagglutinin chains. The protein record has BLAST and Conserved Domains Analysis Tools as well as a database ad for the Influenza Resources area of the Web site with access to all flu sequences and specific analytical tools.

*Bacteria*/. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 171.0 is available via web and FTP. The current release includes information available as of April 10, 2009. Release notes are available on the on the NCBI ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

NCBI is considering ceasing support for index files. Affected GenBank users are encouraged to read that section of the release notes and provide feedback to the GenBank group.

## Updates and Enhancements

### RefSeq

RefSeq Release 35 is now available via Entrez and FTP. This full release incorporates genomic, transcript, and protein data available as of May 4, 2009. It includes 10,993,891 records from 8,393 different species and strains. Changes since the previous release can be found in the release notes on the FTP site. The RefSeq website is: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/). The FTP site is: <ftp.ncbi.nlm.nih.gov/refseq/release>.

### dbSNP

Complete data for the dbSNP Human build 130 are available on the FTP site and for searching on the web. More detailed genome build information is available on the dbSNP page: [www.ncbi.nlm.nih.gov/SNP/snp\\_summary.cgi](http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi).

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are available which provide email announcements about changes and updates to NCBI resources including dbGaP, BLAST, GenBank, and Sequin. The various lists are described on the Announcement List summary page: [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html). To receive updates on the *NCBI News*, please see: [www.ncbi.nlm.nih.gov/About/news/announce\\_submit.html](http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, May 2009

Peter Cooper, PhD<sup>1</sup> and Dawn Lipshultz, MS<sup>2</sup>

Created: April 24, 2009.

### Featured Data: 2009 H1N1 Influenza Sequences

NCBI is the repository for the 2009 influenza virus sequences from the global H1N1 outbreak and is making every effort to make the sequences available as soon as possible. You can access the recent flu sequences and retrieve them individually from a special influenza virus resource page that is updated daily:

<http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html>

#### GenBank sequences from 2009 H1N1 influenza outbreak

All submitted influenza sequences are available in GenBank as soon as they are processed. The 2009 H1N1 influenza virus sequences are listed on this page and are available for BLAST searching [here](#), and are also available in the [NCBI Influenza Virus Sequence Database](#), and can be retrieved with sequences from other influenza viruses for further analyses using tools integrated to the database.

**The following 2009 H1N1 influenza virus sequences were submitted to NCBI and are available in GenBank:**

*May 06, 2009, 102 submitted by CDC:*

	PB2	PB1	PA	HA	NP	NA	MP	NS
<b>Influenza A virus</b> (A/Arizona/01/2009(H1N1))				GQ117067	GQ117063	GQ117064	GQ117066	GQ117065
<b>Influenza A virus</b> (A/Arizona/02/2009(H1N1))	GQ117076	GQ117075		GQ117079	GQ117074	GQ117077	GQ117078	
<b>Influenza A virus</b> (A/California/04/2009(H1N1))				GQ117044				
<b>Influenza A virus</b> (A/California/14/2009(H1N1))	GQ117035	GQ117034	GQ117037	GQ117040	GQ117033	GQ117036	GQ117039	GQ117038

#### H1N1 Flu Info

U.S. Info ›  
 Things You Can Do ›  
 Plan & Prepare ›  
 International Info ›

[HHS.gov](http://HHS.gov)    [CDC.gov](http://CDC.gov)


[Add This To Your Web Site!](#)

### Using Flu Database Query Builder to Download Sequences in FASTA Format

An easy way to get these sequences all at once is through the query builder on the influenza virus database search page:


[www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi](http://www.ncbi.nlm.nih.gov/genomes/FLU/Database/select.cgi)

On this page you can select the characteristics of the flu sequences of interest and then retrieve them or perform additional analyses such as multiple alignments and phylogenetic tree construction.



## Influenza Virus Resource

Information, Search and Analysis



HOME SEARCH SITE MAP
Flu home
**Database**
Genome Set
Alignment
Tree
BLAST
Annotation
FTP
Help
Contact us

**Main Page>>Database**

**What are you looking for?** Select one name each from the lists provided, and/or fill in the boxes. Multiple queries can be built by clicking the "Add to Query Builder" button every time a new query is made, and queries in any combination from the Query Builder can be selected to get sequences in the database. An advanced search tool is available [here](#).

show:  Protein sequence  Coding region  Nucleotide sequence

Virus Species	Host	Country/Region	Segment/Protein	Date Range	
<input type="text" value="any"/> Influenzavirus A Influenzavirus B Influenzavirus C	<input type="text" value="Ferret"/> Giant anteater Human Leopard	<input type="text" value="any"/> Africa Asia Europe	<input type="text" value="any"/> 1 (PB2) 2 (PB1) 3 (PA)	year month day From: <input type="text" value="2009"/> <input type="text" value="04"/> <input type="text" value=""/> To: <input type="text" value="2009"/> <input type="text" value="05"/> <input type="text" value=""/>	<a href="#">Help</a>

Subtype Min. length Max. length  
    [Search by a string](#) [Help](#)

Full-length sequences only [Help](#)
 Remove identical sequences [Help](#)
 Sequences from the FLU project only [Help](#)
 Include Lab strains [Help](#)

**Query Builder**

	Virus species	Host	Country/Region	Protein	Subtype	Date	Length	Key word	Full-length	Remove identical sequences	NIAID project	Include Lab strains	Number of sequences
<input checked="" type="checkbox"/>													

To get all of the recent H1N1 nucleotide sequences, select the nucleotide radio button on the “show” line of the form, set the species to “Influenzavirus A”, the host to “human”, the Country/Region to “any”, the collection date range to “2009/04:2009/05”, and the subtype to “H1N1”. Set the viral segment to “any” to get all segments, or select a specific segment or protein if desired. Click the “Get sequences” button to list the matching sequences. This search retrieves 304 records at the time of writing (May 7, 2009).

To download sequences in FASTA format choose the type of sequence desired – protein, coding region, nucleotide – from the “Select FASTA sequence for download” device at the top of the page. The file download dialog box appears that allows saving all sequences in FASTA format to a local file.



**Main Page>>Database**  
 Select/de-select sequences from the list below. Click on an action button to proceed.

Show Query Builder

Ordered by the following fields

Reorder sequences Add your own sequences

Do multiple alignment Build a tree - Select FASTA sequences to download - - Select accession list to download -

<input checked="" type="checkbox"/>	accession	length	host	segment	protein name	304 nucleotide sequences	Age	Gender
<input checked="" type="checkbox"/>	CY039527	1721	Human	4 (HA)	Influenza A virus (A/Netherlands			
<input checked="" type="checkbox"/>	CY039528	1441	Human	6 (NA) H1N1	Netherlands 2009/04/29	Influenza A virus (A/Netherlands		

## Using Batch Entrez to Download GenBank and Other Formats

Batch Entrez can be used to get the full records (GenBank, XML, or ASN.1) instead of the FASTA format for the flu sequences. To do this, download the list of accession numbers from the Flu database directly from the query builder results obtained above by selecting the desired sequence (protein or nucleotide) from the “Select accession list to download” device at the top of the results page. Save the list to a local file. Then upload these using the batch Entrez service to obtain the records as follows. Access the batch Entrez page.

[www.ncbi.nlm.nih.gov/sites/batchentrez](http://www.ncbi.nlm.nih.gov/sites/batchentrez)

Click the “Browse” button at the top of the batch Entrez page and point to the file containing the downloaded list of accessions. Click the “Retrieve” button then the link in the results to retrieve the influenza virus records in the Entrez nucleotide database. Once the records are in the Entrez Nucleotide service you can use the features of Entrez such as History and Preview/Index to refine your results if desired. See the Entrez help documentation for details.

[www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez](http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpentrez)

Click on individual records to in the results to view them in GenBank format or use the “Display” pull-down list to choose the format of interest and show all records. The records will be displayed 20 per page. Download the entire set by choosing the File option from the “Send to” pull-down list at the top of the first page of records.

## Obtaining the eight diagnostic records

The World Health Organization has identified the eight segments of the earliest H1N1 isolate from California as diagnostic sequences for the new influenza virus strain.

[www.who.int/csr/disease/swineflu/swineflu\\_genesequences\\_20090425.pdf](http://www.who.int/csr/disease/swineflu/swineflu_genesequences_20090425.pdf)

These sequences correspond to GenBank accession numbers FJ966079-FJ966086 available in the Entrez nucleotide service.

[www.ncbi.nlm.nih.gov/sites/entrez?db=nucore](http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucore)

Use the following query to retrieve these eight segments directly.

FJ966079:FJ966086[Accession]

## Flu Sequence Updates

NCBI is expecting new data on a daily basis as the outbreak continues. Check the flu virus pages listed above for breaking news and new sequences.

## Featured Resource: Protein Multiple Alignment Tool Web Service

NCBI will soon offer a Web multiple protein alignment service that uses the Constraint-Based Multiple Alignment Tool (COBALT)(1). COBALT can align a set of provided sequences or can be run as an extension to a Web BLAST search performing a multiple alignment on the set of protein sequences collected from the original BLAST search. The Web implementation of COBALT uses information from pairwise protein BLAST (blastp) scores, Conserved Domain Database results, and Prosite pattern matches as constraints in an initial pairwise alignment that is followed by a progressive multiple sequence alignment. The results from COBALT can be used with the BLAST treeview service to generate a phylogenetic tree from the multiple alignment. An often-requested service, the addition of a multiple alignment greatly enhances the suite of sequence analysis tools available at the NCBI and provides a new and powerful extension to BLAST.

## Running COBALT on a Set of Sequences

The Web interface to COBALT will be available from the BLAST homepage or directly from the following URL:

[www.ncbi.nlm.nih.gov/tools/cobalt/](http://www.ncbi.nlm.nih.gov/tools/cobalt/)

The basic COBALT interface shown in Figure 1 A has the advanced parameters available through a link that expands the form. Advanced parameters include gap open and extend penalties and constraint and clustering parameters. The default values for the constraint and clustering parameters have been optimized to give the best alignment without undue sacrifice of speed. Generally, altering the default constraints will degrade performance. Detailed information on any of the advanced parameters is available through help documentation linked to each option.

COBALT accepts protein sequences in FASTA format or NCBI identifiers as input. Figure 1 B shows a portion of the COBALT alignment obtained using nine of the protein sequences from the NCBI HomoloGene cluster for ATP citrate lyase (HomoloGene ID 854). The alignment contains an anomalous predicted protein from chimpanzee (XP\_511495) that is mis-spliced because of missing data in the genome. This sequence creates large gaps in the multiple alignment that interrupt the conserved Citryl-Coa lyase domain. One very useful feature of the Web interface to COBALT is the ability to edit the sequence set and perform another alignment with the modified set (Figure 1 C). In this case, un-checking the box next to the chimpanzee sequence and re-submitting the set produces the improved alignment shown in Figure 1 D.

## Running COBALT from Web BLAST Results

COBALT can also be run from the results of any Web protein BLAST search by clicking the “Multiple Alignment” link in the “Other reports” line on the BLAST results. This provides an easy way to collect homologs in a set of species and align them for phylogenetic or other comparative study. The most useful sets of sequences for these purposes come from searches with well-defined taxon-restricted databases. For example, a BLAST search with the human prolactin reference sequence (NP\_000939) can collect growth hormone family members from bony fishes for building a multiple alignment and a gene tree. A BLAST search using the following settings finds 13 full-length growth hormone homologs from four species of fish: Database = Reference proteins (refseq\_protein); Organism = bony fishes; Entrez query = srcdb refseq known[properties]; Expect threshold = 1e-6. The Entrez query limit eliminates proteins based entirely on gene predictions. The Expect threshold helps restrict the set to only closely related proteins and can be adjusted after expanding the “Algorithm parameters” of the BLAST form. A COBALT alignment can be generated by clicking the “Multiple alignment” link in the “Other

**COBALT** Constraint-based Multiple Alignment Tool

Home Recent Results Help My NCBI [Sign In] [Register]

COBALT computes a multiple sequence alignment using conserved domain and local sequence similarity information. [more...](#)

Enter Query Sequences [Reset page](#)

Enter accession, gi, or FASTA sequence  Clear

```
ref NP_001087.2
ref XP_511495.2
ref NP_598798.1
ref NP_001025711.1
ref NP_001002649.1
ref NP_523755.1
ref NP_506267.1
```

Or, upload FASTA file  Browse...

Job Title

**Align**  Show results in a new window

[Advanced parameters](#)

A

<input checked="" type="checkbox"/>	<a href="#">NP_001087</a>	980	INNPDMRVQILKDYVR---QHFPATPLLDYALEVEKITTSK-----	1017
<input checked="" type="checkbox"/>	<a href="#">XP_511495</a>	1179	INNPDMRVQILKDYVR---QHFPATPLLDYALEVEKITTSKMFEFALHPCFDENFPVQDKSQYPIPSGCPKPKLWFLLRLL	1255
<input checked="" type="checkbox"/>	<a href="#">NP_598798</a>	970	INNPDMRVQILKDFVK---QHFPATPLLDYALEVEKITTSK-----	PREDICTED: ATP citrate lyase [Pan troglodytes]
<input checked="" type="checkbox"/>	<a href="#">NP_001025711</a>	980	INNPDMRVQILKDYVK---QHFPATPLLDYALEVEKITTSK-----	1008
<input checked="" type="checkbox"/>	<a href="#">NP_001002649</a>	971	INNPDMRVQILKDFVK---QHFPATQLLDYALEVEKITTSK-----	1008
<input checked="" type="checkbox"/>	<a href="#">NP_523755</a>	966	INNPDMRVQILKDFVK---QHFPATQLLDYALEVEKITTSK-----	1003
<input checked="" type="checkbox"/>	<a href="#">NP_506267</a>	976	INNPDMRVQILKDFVK---QHFPATQLLDYALEVEKITTSK-----	1016

B

Phylogenetic Tree Edit and Resubmit Alignment parameters

**My Cobalt Results - Cobalt RID 00X37ER8212 (7 seqs)**

Descriptions  Select All **Re-align**

Legend for links to other resources: [U](#) UniGene [G](#) GEO [G](#) Gene [S](#) Structure [M](#) Map Viewer

Accession	Description	Links
<input checked="" type="checkbox"/> <a href="#">NP_001087.2</a>	ATP citrate lyase isoform 1 [Homo sapiens] >sp P53396.3 ACLY_HUMAN RecName: Full=ATP-citrate sy	<a href="#">UG</a>
<input type="checkbox"/> <a href="#">XP_511495.2</a>	PREDICTED: ATP citrate lyase [Pan troglodytes]	<a href="#">UG</a>
<input checked="" type="checkbox"/> <a href="#">NP_598798.1</a>	ATP citrate lyase [Mus musculus] >sp Q91V92.1 ACLY_MOUSE RecName: Full=ATP-citrate synthase; A	<a href="#">UG</a>
<input checked="" type="checkbox"/> <a href="#">NP_001025711.1</a>	ATP citrate lyase [Gallus gallus] >emb CAH65182.1  hypothetical protein [Gallus gallus]	<a href="#">UG</a>
<input checked="" type="checkbox"/> <a href="#">NP_001002649.1</a>	ATP citrate lyase [Danio rerio] >gb AAH76484.1  ATP citrate lyase [Danio rerio]	<a href="#">UG</a>
<input checked="" type="checkbox"/> <a href="#">NP_523755.1</a>	ATP citrate lyase, isoform A [Drosophila melanogaster] >ref NP_725514.1  ATP citrate lyase, isoform B [D	<a href="#">UG</a>
<input checked="" type="checkbox"/> <a href="#">NP_506267.1</a>	hypothetical protein B0365.1 [Caenorhabditis elegans] >emb CAB02690.1  C. elegans protein B0365.1, f	<a href="#">UG</a>

C

<input checked="" type="checkbox"/>	<a href="#">NP_001087</a>	944	KMFSKAFDSGII PMEFVNKMKKEGKLMIGIGHRVKSI NNPD MRVQILKDYV---RQHFPATPLLDYALEVEKITTSKKPN	1020
<input checked="" type="checkbox"/>	<a href="#">NP_598798</a>	934	KMFSKAFDSGII PMEFVNKMKKEGKLMIGIGHRVKSI NNPD MRVQILKDFV---RQHFPATPLLDYALEVEKITTSKKPN	1010
<input checked="" type="checkbox"/>	<a href="#">NP_001025711</a>	944	KMFSKAFDSGII PMEFVNKMKKEGKLMIGIGHRVKSI NNPD MRVQILKDYV---RQHFPATPLLDYALEVEKITTSKKPN	1020
<input checked="" type="checkbox"/>	<a href="#">NP_001002649</a>	935	KQFSKAFDSGMLPMEFVNKMKKDGKLMIGIGHRVKSI NNPD MRVQILKDFV---RQHFPATQLLDYALEVEKITTSKKPN	1011
<input checked="" type="checkbox"/>	<a href="#">NP_523755</a>	930	RQFSEAYDTNLHPMEFVNKMRKEGKLLIGIGHRVKSI NNPDV RVKIIKEFVL---ENFPACPLLKYALEVEKITTNKKPI	1003
<input checked="" type="checkbox"/>	<a href="#">NP_506267</a>	940	RQFSEAFDQGWSPNQFVGMERKRGRTHIMIGIGHRVKSI NNPDK RVEILKRFALNKKEFAQETPLLDYALEVEKITTAKKPI	1016

D

**Figure 1. COBALT interface and multiple alignments.** A. The basic COBALT interface with seven NCBI accessions from the Homologene cluster for ATP citrate lyase (Homologene ID 854) B. A portion of the COBALT multiple alignment showing the aberrant chimpanzee gene model, second line, XP\_511495. C. The sequences realigned after de-selecting the chimpanzee sequence. D. The multiple alignment without the chimpanzee sequence.

reports” line at the top of the Descriptions or Alignments section of the BLAST results (Figure 2, top panel). The COBALT results appear in a new browser window when ready (Figure 2, middle panel).

## Generating a Phylogenetic Tree from COBALT Results

A phylogenetic tree can be generated from any set of COBALT results by clicking the “Phylogenetic tree” link at the top (Figure 2, middle panel). The tree is generated using the Treeview (NCBI News, Summer 2006) feature of the BLAST Web service. This tree is calculated from a global multiple sequence alignment and is therefore more accurate than the Distance tree that can be generated from the BLAST results. The tree view display has options for redrawing the tree in different format, recalculating the distance metrics, downloading the tree in text formats, and displaying and realigning sequences from any node. The tree generated from a multiple alignment of the fish growth hormone family members collected by the BLAST search with the human prolactin precursor is shown on the bottom panel of Figure 2. The tree shows three distinct subfamilies: growth hormone, prolactin, and the somatolactins. The latter are pituitary hormones apparently found only in fishes(2).

## Retrieving Previous COBALT Results

COBALT results from the Web service are stored at NCBI and are available for later retrieval in the same way as ordinary BLAST results. Recent COBALT results are available through the “Recent Results” tab at top of the COBALT submission or results pages. Like BLAST results COBALT results may be retrieved up 36 hours from the time of the search using the Request Identifier (RID) that uniquely identifies each set of results.

## Summary and Future Directions

The COBALT multiple protein alignment tool expands the suite of sequence analysis tools available at the NCBI and provides a single pathway now for collecting related sequences using BLAST and then performing a rapid and accurate multiple alignment. Moreover for the first time multiple alignments can be used directly at the NCBI to generate and display phylogenetic trees making the NCBI Website a comprehensive resource for analyzing protein relationships. Upcoming improvements to the COBALT tool include the ability to re-format and download alignments in various standard formats such as FASTA plus gap. This will allow COBALT multiple alignments to be imported into other multiple alignment programs and editors.

## New Databases and Tools

### H1N1 Influenza Resources

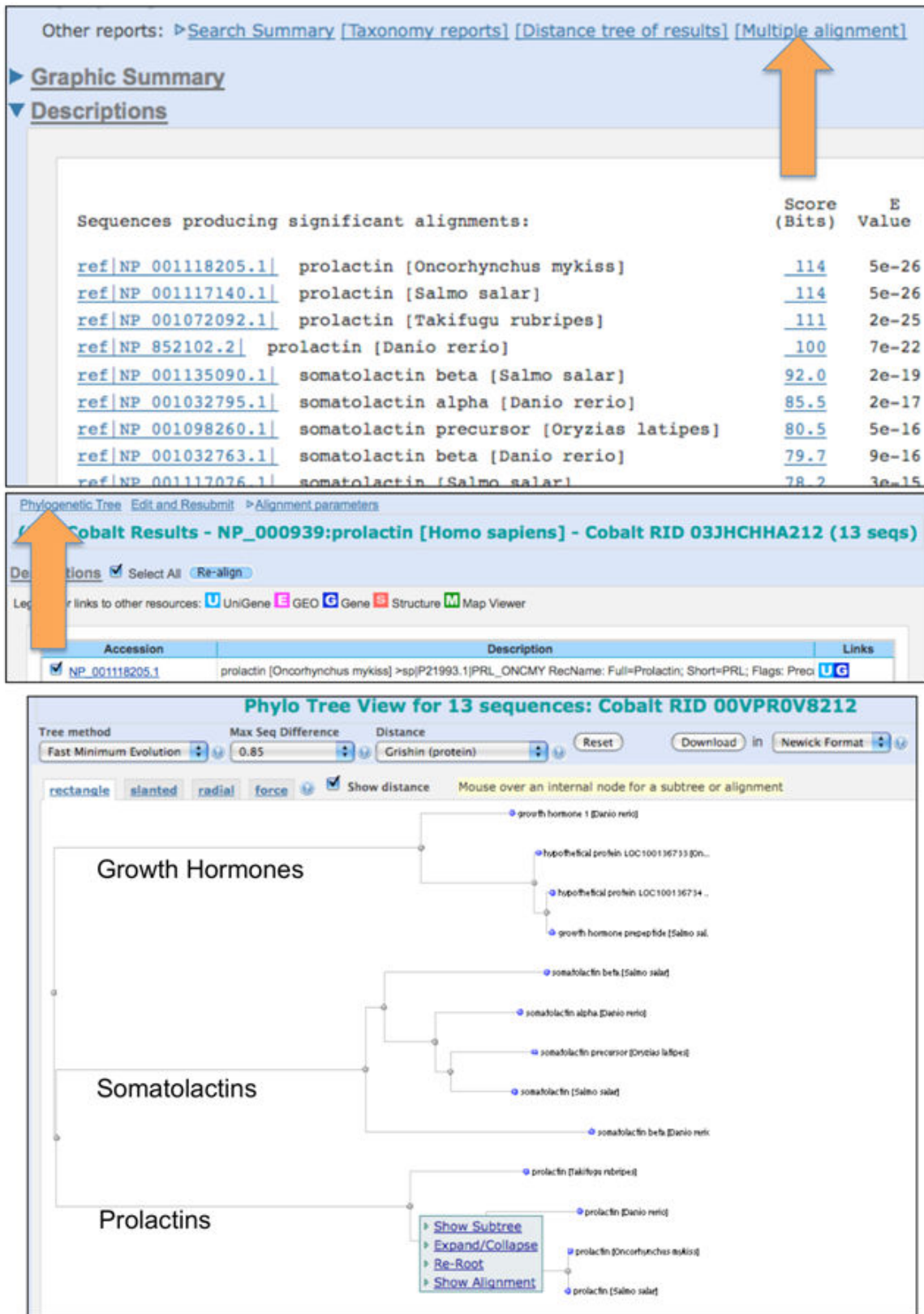
NCBI has various resources available as described above. The Influenza Virus Resource has 34 H1N1 influenza sequences listed on the following page: [www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html](http://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html). PubMed contains recently added [literature citations related to](#) the new strain of influenza.

### Peptide Data Resource

Peptidome is a new public repository that archives and distributes tandem mass spectrometry peptide and protein identification data. Web-based interfaces are available to browse and explore studies, peptides, and proteins. For more information see the Peptidome web page: [www.ncbi.nlm.nih.gov/projects/peptidome/](http://www.ncbi.nlm.nih.gov/projects/peptidome/).

### Genome Build

Build 1 of the *Vitis vinifera* (wine grape) genome is available in the Genomes database and on the NCBI Map Viewer. The Map Viewer page is: [http://www.ncbi.nlm.nih.gov/mapview/map\\_search.cgi?taxid=29760](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=29760)



**Figure 2.** BLAST, COBALT results, and phylogenetic tree of growth hormone family members for NCBI RefSeq proteins from teleost fishes. The sequences were collected by a BLAST search limited to NCBI known RefSeqs (refseq\_protein database, srcdb\_refseq\_known Entrez limit) with bony fishes as an organism limit (top panel). The query sequence was the human prolactin precursor (NP\_000939). A multiple alignment was generated from the BLAST results. The phylogenetic tree shown (bottom panel) was based on a second COBALT alignment (middle panel) using only the fish sequences after de-selecting the human sequence on the original alignment and re-aligning. The tree shows three distinct subfamilies: growth hormone, prolactin, and the somatolactins, a group of pituitary hormones specific to fishes.

## Bookshelf

The Bookshelf has added a new chapter to the *NCBI Help Manual*, GaP FAQ Archive. The Bookshelf website URL is: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books)

## Microbial Genomes

Fifteen finished microbial genomes were released between March 24-April 29. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: [ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

## GenBank News

GenBank release 171.0 is available via web and FTP. The current release includes information available as of April 10, 2009. Release notes are available on the on the NCBI ftp site: [ftp.ncbi.nih.gov/genbank/gbrel.txt](ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

NCBI is considering ceasing support for index files, so affected users are encouraged to read that section of the release notes and provide feedback to the GenBank group.

## Updates and Enhancements

### SRA Transcript BLAST

SRA transcript sequences are now searchable through a specialized BLAST page. All transcript sequences derived from 454 sequencing are available from NCBI's SRA database. To perform a search, go to [the SRA BLAST page](#).

### GEO DataSet Browser

A new GEO DataSet Browser is available for browsing the curated gene expression DataSets. The new tool is located: [www.ncbi.nlm.nih.gov/sites/GDSbrowser](http://www.ncbi.nlm.nih.gov/sites/GDSbrowser).

### Sequence Analysis Tools Links in Entrez Sequence Databases

A new Sequence Analysis Tools section is available on the right hand Discovery column of Nucleotide and Protein records in the Entrez system. Sequence Analysis Tools contains links to the BLAST service for both protein and nucleotide sequences. Nucleotide records also link to Primer BLAST service. Both of these links load the currently viewed sequence in the submission area of the tool ready to perform a BLAST or Primer BLAST search. In addition protein records have a link to pre-computed conserved domain results. These new links make it easy to perform sequence analysis on the fly from any sequence record.

## Exhibits

NCBI will be exhibiting at the American Society for Microbiology's 190<sup>th</sup> General Meeting on May 17-21 in Philadelphia, Pennsylvania.

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## References

1. Papadopoulos J, Agarwala R. COBALT: constraint-based alignment tool for multiple protein sequences. COBALT: constraint-based alignment tool for multiple protein sequences. 2007;23(9) PubMed PMID: 17332019.
2. Kaneko T. Cell biology of somatolactin. Cell biology of somatolactin. 1996;169:1–24. PubMed PMID: 8843651.





## NCBI News, April 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: March 12, 2009.

### Featured Resource: PubChem Now Offers 3-D Small Molecule Structures and a New Conformer Viewer (Pc3D)

NCBI's PubChem now features calculated three-dimensional conformers (3-D conformers) for a large proportion of the PubChem compound database (17 million records, 88%). In addition, conformers are clustered to provide a list of similar 3-D conformers. These similar conformers provide a more relevant and expanded set of compounds with potentially similar biological and pharmacological activity. PubChem also provides a viewer for small molecule conformers and alignments that produces an animated view from displays in the Web service.

The new standalone viewer, Pc3D, is the small-molecule equivalent of the now standard Cn3D viewer for macromolecular structures and provides additional features for visualizing the details of small molecule structures and alignments. These powerful new aspects of PubChem greatly expand the potential of PubChem as a chemical informatics resource for rational biological inhibitor and drug design.

#### Conformers and similar conformers

Three-dimensional conformers and similarities in PubChem are calculated using the commercial software packages Omega and OEShape (OpenEye Scientific Software, Santa Fe, New Mexico)<sup>1,2</sup>. Conformers are provided for all molecules in PubChem that are not too large or complex. A single low-energy conformer is presented in the PubChem Web display and is used for the calculation of similar conformers. All PubChem data including conformers may be accessed through the NCBI Entrez system.

[www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound](http://www.ncbi.nlm.nih.gov/sites/entrez?db=pccompound)

These 3-D conformers are also available for download from the FTP site:

[ftp.ncbi.nlm.nih.gov/pubchem/Compound\\_3D/](ftp://ncbi.nlm.nih.gov/pubchem/Compound_3D/)

Conformer data is provided in NCBI Abstract Syntax Notation (ASN), Extensible Markup Language (XML), and the Structures Data File (SDF) format on both the Web service and the FTP site.

Similarity measurements incorporate both shape (Tanimoto shape) and feature similarity (Tanimoto color). Features include the presence of ring systems, hydrophobic entities, positive and negative ionizable groups, and hydrogen bond donors and acceptors. Similar conformers are defined as those with greater than 80% shape similarity and more than 50% feature similarity.

Similar conformers are available from the "Related Structures" link on the PubChem summaries in search results or from the Compound Information section of a PubChem record. The "View Conformers" link will load the aligned conformers directly into the Web-based viewer. Looking at similar conformers rather than the traditional 2-D similar compounds often can lead to additional related molecules with interesting properties. For example, the opiate morphine (Compound ID 5288826) currently has 1,297 similar compounds and 1,418 similar conformers in PubChem. The sets are quite different, however, with only 514 of the compounds common to both sets. The 3-D set in this case has distinct links to BioAssay and literature data. For other compounds

there may be distinct links to protein structures as well. Thus examining the similar conformers can expand related compounds to a more diverse set of molecules.

## Viewing small molecule conformers and alignments

Figure 1 shows a PubChem Compound record for the stimulant methylphenidate (Compound ID 4158). The small 2-D structural formula is the default graphic on the right-hand-side of the structure summary under the “Structure and Quick Link Bar”. Clicking on the “3D” tab produces the small 3-D graphic shown in the figure. The 3-D viewers are easily launched from this graphic.

### Single conformers

A pop-up menu activated by clicking the 3-D graphic allows the conformer to be displayed in the Web-based viewer or in the standalone Pc3D viewer once it is installed. The standalone viewer is available for Windows, Linux and Mac OSX operating systems and may be downloaded and installed by following the instructions in the online Pc3D manual.

<http://pubchem.ncbi.nlm.nih.gov/pc3d/>

The top panel of figure 2 shows the single methylphenidate conformer displayed in the Web-based viewer. The conformer in the viewer is animated and rotates slightly about the three axes to provide perspective on the shape. Speed, zoom, rotation controls, and a toggle for hydrogen atoms on the left-hand-side of the viewer provide a means of changing the display. The Pc3D icon and link (“View in Pc3D”) within the Web-based viewer will download the structure and automatically display it in the installed Pc3D viewer. Pc3D has more flexible and sophisticated rendering options for conformers. The bottom panel of figure 2 shows methylphenidate rendered in a space-filling format, one of several display options available in Pc3D. The Options and Commands menus modify the display of the conformer. The online manual for Pc3D linked above has detailed instructions for using the various features of the program.

### Aligned conformers

Aligned conformers load directly into the Web-based viewer from the “View Conformers” link in the PubChem record as mentioned above. These alignments can also be loaded into an active view by clicking on the “Similar Conformers” link on the left-hand-side of the viewer. Pairwise conformer alignments are displayed one-at-a-time in the viewer. The alignment with the most similar conformer is shown first. Other alignments can be selected using the arrows at the top of the viewer to scroll through the alignments or by typing the number of the alignment in the collapsible search box beneath the arrows. The corresponding alignment can be loaded into the standalone viewer by clicking the “View in Pc3D” link. The alignment between methylphenidate and a pyrrolidine analog with inversion of configuration about the central carbon atom is shown in figure 3 for both the Web-based viewer (top panel) and Pc3D (bottom panel). (Note: The PubChem record for methylphenidate (CID 4158) does not specify the stereochemistry about the two chiral centers. In such cases, only a single enantiomer of the lowest energy structure is displayed and used for similarity calculations.)

## Summary

The availability of 3-D conformers and similarities in PubChem greatly expands the utility of the PubChem resource by providing visualization tools and extends the universe of similar compounds. These now include not only simple derivatives but also additional compounds with similar volume, steric and charge features. Analyses using these features will help investigators understand the nature and causes of biological activity and provide new candidates for drug and inhibitor design.

**Methylphenidate - Compound Summary** (CID 4158)

A central nervous system stimulant used most commonly in the treatment of attention-deficit disorders in children and for narcolepsy. Its mechanisms appear to be similar to those of DEXTROAMPHETAMINE.

**Table of Contents**

- Drug and Chemical Information
  - Medication Information
  - Pharmacological Action
  - Pharmacological Classification
  - Chemical Classification
  - Safety and Toxicology
  - Literature Links
  - Literature Mining
- BioActivity Results
- Synonyms
- Properties
- Descriptors
- Compound Information
- Substance Information
  - Category
- Exports

**Structure & Quick Link Bar**

2D | 3D

**Links**

- Web-based 3D Viewer [3D Viewer Download](#)
- Pc3D Viewer Application

**Compound ID 4158**

Molecular Weight	233.30616 [g/mol]	<a href="#">?</a>
Molecular Formula	C <sub>14</sub> H <sub>19</sub> NO <sub>2</sub>	<a href="#">?</a>
XLogP3	0.2	<a href="#">?</a>
H-Bond Donor	1	<a href="#">?</a>
H-Bond Acceptor	3	<a href="#">?</a>

**Drug and Chemical Information:** (Total:1) [?](#)

**Figure 1.** The PubChem compound summary page for methylphenidate. Clicking the “3D” tab on the “Structure & Quick Link Bar” changes the display to a 3-D conformer. A pop-up menu provides links to display the conformer in the Web-based viewer or the standalone Pc3D viewer.

## References

- Fontaine F, Bolton E, Borodina Y, Bryant SH. (2007) Fast 3D shape screening of large chemical databases through alignment-recycling. *Chem Cent J*. Jun 6;1:12.
- Borodina YV, Bolton E, Fontaine F, Bryant SH. (2007) Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space. *J Chem Inf Model*. Jul-Aug;47(4):1428-37.

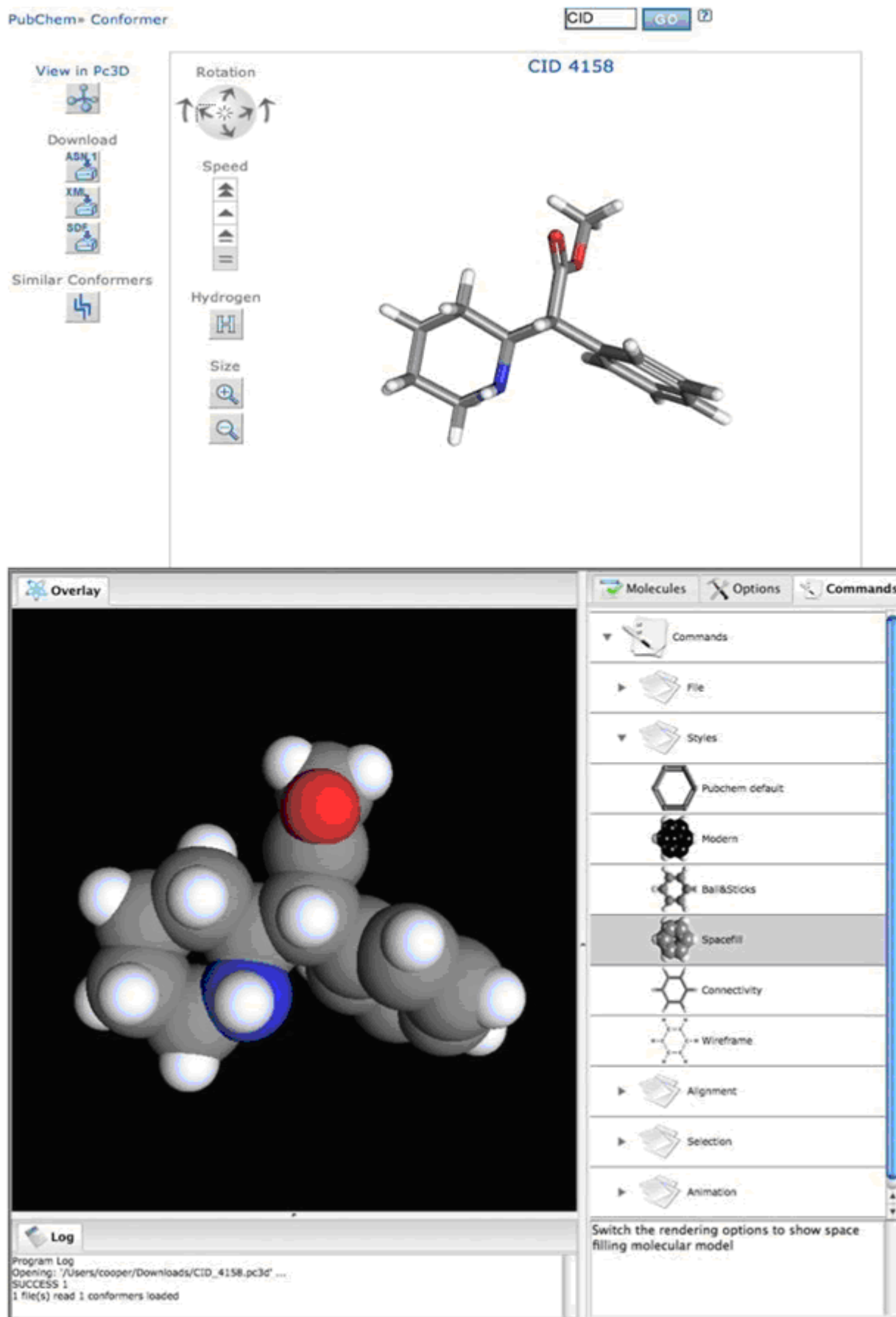
## New Databases and Tools

### Genome Build

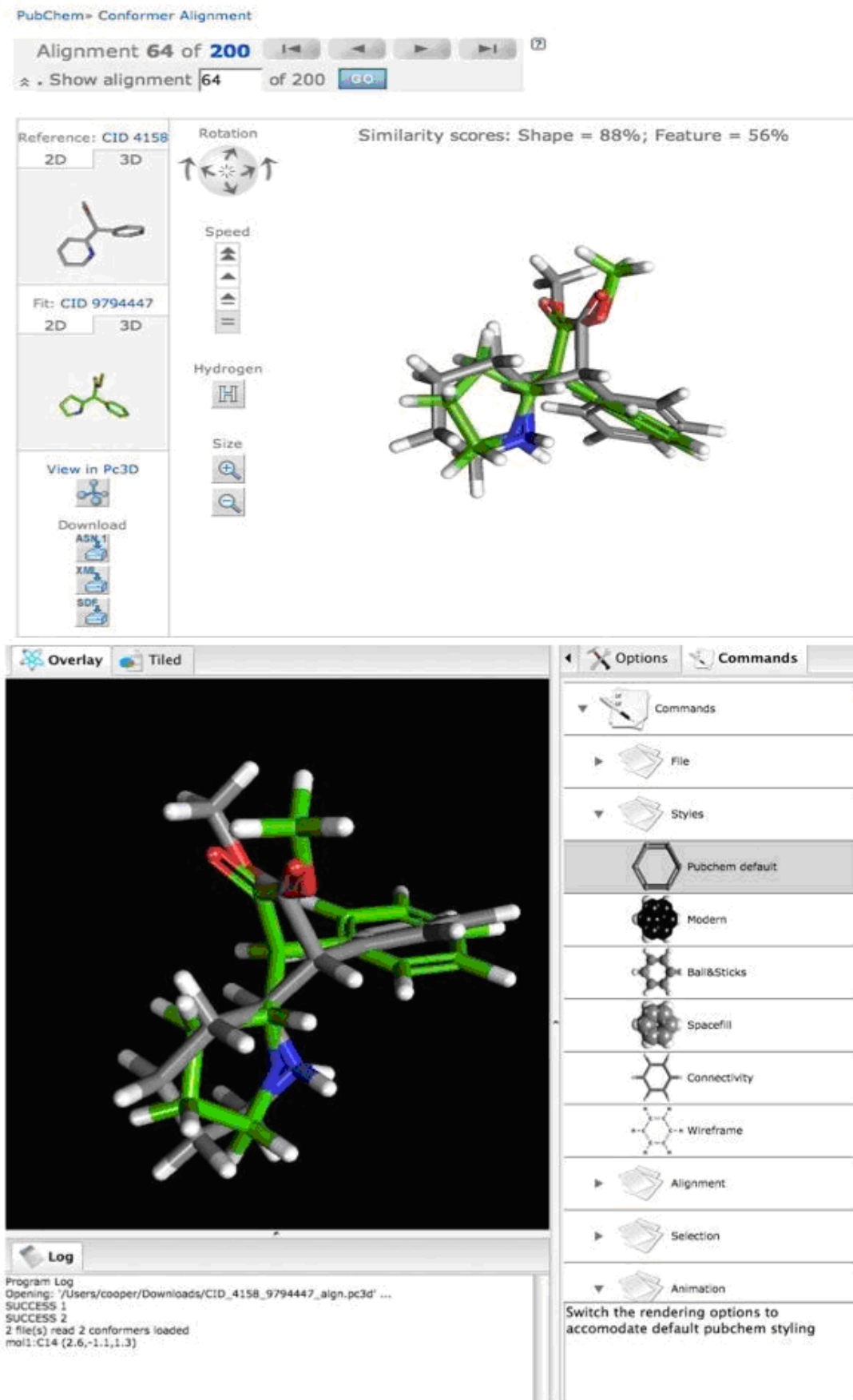
Build 1.1 of the genomes *Hydra magnapapillata* and *Taeniopygia guttata* (zebra finch) are available in the Genomes database and on the NCBI Map Viewer. The Map Viewer page is: <http://www.ncbi.nlm.nih.gov/mapview/>

### Bookshelf

The Bookshelf has added four new books: *Electrochemical Methods for Neuroscience*, *Frontiers in Neuroscience*, *Indwelling Neural Implants: Strategies for Contending with the InVivo Environment*, and *The National Academies Collection: Reports Funded by National Institutes of Health*. The Bookshelf website URL is: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books)



**Figure 2.** The lowest energy methylphenidate conformer rendered in the Web-based viewer (top panel) and in Pc3D. The style in Pc3D was changed to spacefill in this rendering.



**Figure 3.** Alignment between a methylphenidate conformer and a conformer for a pyrrolidine analog that has inverted stereochemistry about the central carbon. Top panel, rendered in the Web-based viewer. Bottom panel, rendered in Pc3D.

## Microbial Genomes

Nine finished microbial genomes were released between February 7 and March 24. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 170.0 is available from the NCBI Web and FTP sites. The current release includes information available as of February 13, 2009. With this release, the new DBLINK linetype is now valid for GenBank sequence records, and it will begin to appear in GenBank Update files, soon after GenBank 170.0 is made available. Release notes are on the on the ftp site: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

## Updates and Enhancements

### NCBI News

The *NCBI News* is now available in PDF format. There is a link for “Printable Copy” in the top right corner of each issue. Clicking that link will allow users to read and print a PDF version of the *News*.

### RefSeq

RefSeq Release 34 is now available via Entrez and FTP. This full release incorporates genomic, transcript, and protein data available as of March 6, 2009. It includes 10,021,870 records from 8,054 different species and strains. The RefSeq website is: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/). The FTP site is: <ftp.ncbi.nlm.nih.gov/refseq/release>. Changes since the previous release can be found in the release notes on the FTP site.

### PubMed

Two new features have been added to the PubMed Abstract Plus Display to enhance discovery within the NCBI databases. Ads now appear in the right-hand discovery column, which provide additional links to related data. One new ad provides a link to the structure database if a structure is reported in the article, and a second ad provides links to PubMed Central articles that have cited the PubMed article being viewed.

### Map Viewer

A new feature has been added to the Map Viewer homepage that links directly to a region on a chromosome. Clicking the “R” icon next to an organism name opens a box that provides options for chromosome number, assembly, and coordinate range. This feature links directly to the entered position on the given genome. The Map Viewer website is: [www.ncbi.nlm.nih.gov/mapview/](http://www.ncbi.nlm.nih.gov/mapview/).

### Exhibits

NCBI will be exhibiting at the American Society for Microbiology’s 190<sup>th</sup> General Meeting on May 17-21 in Philadelphia, Pennsylvania.

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.





## NCBI News, March 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: February 12, 2009.

### Featured Resource: The New Entrez Sequence View has an Emphasis on Discovery

NCBI now has an updated sequence view for nucleotide and protein records that provides better display options, links to analysis tools, and an emphasis on the discovery of related information in other NCBI databases (Figure 1). The new view retains the standard format of the sequence but has more obvious links to display FASTA, other formats, and specific subregions and features. The right-hand discovery column features direct links to analysis tools, and, most importantly, explicit links to valuable related information such as better-annotated sequences, and more information-rich databases such as Gene, HomoloGene, and PubMed.

#### New display links, regions and feature options

The format row at the top of the new view has direct links to display the sequence in FASTA format and to display the record and features in the NCBI Graphical Viewer (NCBI News, Aug 2008). Other less specialized reports such as ASN.1 and XML are available through the separate “More formats” list. A “Download” link to save the records in various formats to a local file and a “Save” link to store the records in the NCBI Clipboard or through MyNCBI in “My Collections” for later use are located at the right-hand end of this row along with the traditional “Links” menu that provides links to related information in other databases. Several of the more useful items from the “Links” menu are listed explicitly as Discovery column items described in the “Discovery Links” section below. The expandable display controls, “Change Region Shown” and “Customize View”, are located below the Links menu on the right-hand-side. The “Change Region Shown” control provides a convenient mechanism for displaying specific regions of the sequence. The “Customize View” control specifies the number and kinds of annotated features and the DNA strand that is shown.

#### Direct entry to analysis tools

The NCBI Discovery column begins just below the display controls with direct access to sequence analysis tools. The NCBI primer designing tool, Primer-BLAST (NCBI News, Nov 2008), is linked here now. Additional tools will be added in later versions of the viewer including a direct link to run a BLAST database search. In all cases these analysis tools are set in advance to provide the most relevant and up-to-date results appropriate to the context. For example, following the “Pick Primers” link from a human mRNA sequence in the viewer loads the Primer-BLAST form already set up to perform a specificity check against the appropriate background database, the human genome transcripts in this case. Moreover if the search is run with an older GenBank sequence as the query, the equivalent NCBI Reference Sequence (RefSeq) will be substituted to improve specificity checking. Direct access to analysis tools will streamline many visits to the NCBI Website by providing one-click, live access to the computing power of the NCBI.

#### Discovery Links

As shown in the figure, the links in the Discovery column expose several highly relevant and useful sets of related information. These include selected relevant articles in PubMed (“Articles about KLK6”), links to corresponding mRNA and protein Reference Sequences, a link to the corresponding Gene record (“More about the KLK6 gene”), and links to homologs in other species provided by HomoloGene (“Homologs of KLK6”). For

Format: [GenBank](#) [FASTA](#) [Graphics](#) [More Formats](#)[Download](#) [Save](#) [Links](#)

GenBank: AF013988.1

**Homo sapiens serine protease mRNA, complete cds**[Features](#) [Sequence](#)

LOCUS AF013988 1451 bp mRNA linear PRI 20-MAY-2008  
 DEFINITION Homo sapiens serine protease mRNA, complete cds.  
 ACCESSION AF013988  
 VERSION AF013988.1 GI:2318114  
 KEYWORDS .  
 SOURCE Homo sapiens (human)  
 ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;  
 Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;  
 Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 1451)  
 AUTHORS Little,S.P., Dixon,E.P., Norris,F., Buckley,W., Becker,G.W.,  
 Johnson,M., Dobbins,J.R., Wyrick,T., Miller,J.R., MacKellar,W.,  
 Hepburn,D., Corvalan,J., McClure,D., Liu,X., Stephenson,D.,  
 Clemens,J. and Johnstone,E.M.  
 TITLE Zyme, a novel and potentially amyloidogenic enzyme cDNA isolated  
 from Alzheimer's disease brain  
 JOURNAL J. Biol. Chem. 272 (40), 25135-25142 (1997)  
 PUBMED [9312124](#)

REFERENCE 2 (bases 1 to 1451)  
 AUTHORS Little,S.P., Johnstone,E.M. and Norris,F.  
 TITLE Direct Submission  
 JOURNAL Submitted (15-JUL-1997) CNS Division, Eli Lilly and Company, Lilly  
 Corporate Center, Indianapolis, IN 46285, USA

FEATURES  
 source Location/Qualifiers  
 1..1451  
 /organism="Homo sapiens"  
 /mol\_type="mRNA"  
 /strain="human"  
 /db\_xref="taxon:9606"  
 /chromosome="19"  
 /map="19q13.3"  
 /tissue\_type="Alzheimer's disease brain tissue"  
 147..881  
 /note="Zyme; protease bears homology to Kallikrein class  
 and can be localized to microvessels and microglia;  
 chymotrypsin-like"  
 /codon\_start=1  
 /product="serine protease"  
 /protein\_id="[AAB66483.1](#)"  
 /db\_xref="GI:2318115"  
 /translation="MKKLMVVLSLIAAAWAEQNKLVHGGPCDKTSHPYQAALYTSQH  
 LLCGGVLIHPLWVLTAAHCKKPNLQVFLGKHNLRQESSQEQSSVVRVAVHPDYDAAS  
 HDQDIMLLRLARPAKLSELIQPLERDCSANTTSCHILGWGKTADGDFPDTIQCAI  
 HLVSREECEHAYPGQITQNLMLCAGDEKYGKDCSQGDSGGPLVCGDHLRGLVSWGNIIP  
 GSKEKPGVYTNVCRYTNWIKTIQAK"

ORIGIN  
 1 gtcgaccac gcgctccgct ggctggctcg ctctctctct gggacacaga ggtcggcagg  
 61 cagcacacag agggacctac gggcagctgt tccttcccc gactcaagaa tccccggagg  
 121 cccggaggcc tgcagcagga gcgcccatga agaagctgat ggtgggtctg agtctgattg  
 181 ctgcagcctg ggcagaggag cagaataagt tgggtgatgg cggaccctgc gacaagacat  
 241 ctcaccctca ccaagctgcc ctctacacct cggggcaact gctctgtggt ggggtcctta  
 301 tccatccact tggggtctct acagctgccc actgcaaaaa accgaatott caggctcttc  
 361 tggggaagca taacctctcg caaagggaga gttcccagga gcagagttct gttgtccggg  
 421 ctgtgatcca cctgactat gatgocgcca gccatgacca ggacatcatg ctggtgoccc  
 481 tggcacgccc agccaaactc tctgaactca tccagccctc tcccctggag agggactgct  
 541 cagccaacac caccagctgc cacatcctgg gctggggcaa gacagcagat ggtgatttcc  
 601 ctgacaccat ccagttgtca tacatccacc tgggttcccg tgaggagtgt gagcatgct  
 661 accctggcca gatcacccag aacatggtgt gtgctgggga tgagaagtac ggaaggatt  
 721 cctgcccagg tgattctggg ggtccgctgg tatgtggaga ccacctccga ggccttgtgt  
 781 catggggtaa catcccctgt ggatcaaaag agaagccagg agtctacacc aacgtctgca  
 841 gatacacgaa ctggatccaa aaaaccattc agggcaagtg accctgacat gtgacatcta  
 901 cctcccgacc taccacocca ctggctggtt ccagaacgtc tctcacctag accttgctc  
 961 cctcctctct ctgcccagct ctgaccctga tgccttaataa accagcagac gtgagggtcc  
 1021 tgattctccc tggttttacc ccagctccat ccttgatca ctggggagga cgtgatgagt  
 1081 gaggacttgg gtcctcggtc ttacccccac cactaaagga atacagaaa atcccctcta  
 1141 ggcattctct ctcccacacc cttccacacg tttgatttct tctctgacag gccacgccc  
 1201 gtgtctggaa tcccagctcc gctgcttact gtgoggtgccc ccttgggatg tacctttctt  
 1261 cactgcagat ttctcaacctg taagatgaag ataaggatga tacagtctcc ataaggcagt  
 1321 ggctggttga aagatttaag gtttcacacc tatgacatac atggaatagc acctgggcca  
 1381 ccatgcactc aataaagaat gaattttatt atgaaaaaaa aaaaaaaaaa aaaaaaaaaa  
 1441 agggcggccc c

//

Change Region Shown

Customize View

**Pick Primers**

Design and test primers for this sequence using Primer-BLAST.

**Articles about KLK6**

- ▶ Prognostic value of kallikrein-related peptidase 6 protein expr [Cancer Sci. 2008]
- ▶ Co-expression of KLK6 and KLK10 as prognostic factors for su [Br J Cancer. 2008]
- ▶ Kallikrein 6 is a mediator of K-RAS-dependent migra [Biol Chem. 2008] » See all...

**Reference sequences**

- ▶ mRNA
- ▶ Protein

**More about the KLK6 gene**

Kallikreins are a subgroup of serine proteases having diverse physiological functions. Growing evidence suggests that many kallikreins are i...

Also Known As: Bssp, Kik7, MGC9355, NE...

**Homologs of KLK6**

The KLK6 gene is conserved in dog, cow, mouse, and rat.

**Order cDNA Clone**

The NIH MGC Collection contains a sequence-verified cDNA clone for KLK6.

Recent Activity

All links from this record

- ▶ Gene
- ▶ Gene Genotype
- ▶ GeneView in dbSNP
- ▶ Probe
- ▶ Protein
- ▶ PubMed
- ▶ PubMed (Weighted)
- ▶ Taxonomy
- ▶ Related Sequences
- ▶ Map Viewer
- ▶ OMIM
- ▶ GEO Profiles
- ▶ SNP

**Figure 1.** The new Entrez sequence view of a GenBank record for a human mRNA (AF013988) submitted in 1997. Links to alternate display formats are at the top of the record. Expandable display controls are at the top of the right-hand Discovery Column. One-click submission to design primers is available through the "Pick Primers" link. The items evident in the Discovery Column provide an instant update for the molecular biology, nomenclature, and literature relevant to this gene and its products.

the record shown in the figure, these Discovery column items automatically update the biology and nomenclature for this older GenBank record, identifying it at a glance as a transcript of the human kallikrein-related peptidase gene, KLK6, and providing in a single click the Reference Sequences for the three of known splice variants of this gene. Each of these has additional biological annotations directly on the record plus its own discovery column. A link to “Order cDNA Clone” provides access to molecular reagents for this transcript. The “Articles about ...” link is an enhanced set of human-reviewed references about KLK6 combining the citations from the Online Mendelian Inheritance in Man (OMIM) article for this gene with the linked articles from the NCBI Gene record. These constitute an essential set of literature about the biological roles of the KLK6 gene and its products. Finally, the “More about the KLK6 gene” link provides direct access to the Gene record, a gateway to the human genome and all molecular biology information about the KLK6 gene.

## Summary

The new Entrez sequence provides intuitive display controls, direct access to live analysis, and to the rich pre-compiled information available through Entrez Gene, OMIM and Homologene. The current version and future improvements move the sequence databases towards a condition where even older sequences become self-annotating and are automatically updated through the analysis performed at the NCBI evident in the Discovery column. These enhancements should make the NCBI Entrez system a more efficient experience for visitors and easier to use as a Discovery system.

## New Databases and Tools

### Genome Build

Build 1.1 of *Hydra magnapapillata* is available in the Genomes database and on the NCBI Map Viewer. The Map Viewer page is: [www.ncbi.nlm.nih.gov/mapview/map\\_search.cgi?taxid=6085](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=6085)

### Microbial Genomes

Twenty-eight finished microbial genomes were released between January 14 and February 6. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FPT site: [ftp.ncbi.nih.gov/genbank/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/). The RefSeq provisional versions of these genomes are also available: [ftp.ncbi.nih.gov/genomes/Bacteria/](ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

## GenBank News

GenBank release 170.0 is available via web and FTP. The current release includes information available as of February 13, 2009. With this release, the new DBLINK linetype is now legal for GenBank sequence records, and it will begin to appear in GenBank Update files, soon after GenBank 170.0 is made available. Release notes are on the on the ftp site: [ftp.ncbi.nih.gov/genbank/gbrel.txt](ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt)

## Updates and Enhancements

### Clone Finder

The Clone Finder tool, designed to identify genomic clones on a genome assembly, has been updated with an improved search interface that makes it easier to search by feature. Clone Finder also offers the ability to filter clones on results pages using feature information, and more informative pop-up menus on clone features. The performance of graphical view is improved and now integrates the library table view. The tool will now allow users to download all clones, or only clones from a given library in Excel format.

Documentation on Clone Finder can be found on the following Web page: [www.ncbi.nlm.nih.gov/projects/mapview/static/clonefinder\\_documentation.shtml](http://www.ncbi.nlm.nih.gov/projects/mapview/static/clonefinder_documentation.shtml)

## UniVec

UniVec database build 5.1 is now available. UniVec is a non-redundant database of vector sequences used in conjunction with the VecScreen tool to identify vector sequence contamination in nucleotide sequences. The number of sequences in UniVec has increased by 2% for build 5.1. The vector BLAST database has also been updated to contain full-length versions of all sequences from GenBank that were used in the current UniVec build.

## PubMed

The PubMed Summary page now displays information about free articles from publishers. The new information is in addition to the PubMed Central links that appear for full-text PMC articles. For more information, see the *NLM Technical Bulletin* article: [www.nlm.nih.gov/pubs/techbull/jf09/jf09\\_pm\\_free\\_article.html](http://www.nlm.nih.gov/pubs/techbull/jf09/jf09_pm_free_article.html). The current issue of the *Technical Bulletin* also has an informative article about shared settings in the My NCBI tool.

## Exhibits

NCBI will have an exhibit booth at the Experimental Biology Annual Meeting on April 18-22 in New Orleans, Louisiana.

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, February 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: January 20, 2009; Updated: February 4, 2009.

### Featured Resource: Improvements to NCBI Services Promote Discovery

In an effort to make the full potential of the NCBI Web services and underlying databases more available to users, the NCBI has begun a long-term project to improve the relevance and usefulness of search results. This effort is called the Discovery Initiative. A primary goal of the Discovery Initiative is to promote the discovery of previously hidden relationships in the large amount of pre-calculated similarity data and pre-compiled links between different molecular and literature databases available at the NCBI. Changes in search interfaces and result displays are being phased in gradually and include the appearance of database ads, alternative search suggestions, and various sensors that will bring to the surface results in other databases that may be more relevant to the search. Particular Discovery components will appear in a context-specific manner ultimately producing the most relevant result possible. To help with this effort NCBI is also designing Web interfaces and links so that the effectiveness and popularity of these changes can be measured and studied. In turn, the results of these studies will be used to improve NCBI's services even more.

#### Discovery Components in PubMed

Many Discovery-related changes can already be seen in the PubMed database, and more will be coming soon. Current Discovery components appearing on PubMed results pages may include Related Queries, Title Searches, a Review tab, the Gene Sensor, and Recent Activity as shown in Figure 1. The Abstract Plus page view may include the top five Related Articles with review articles highlighted, Patient Drug Information and Recent Activity (Figure 2).

The Related Queries component shown under "Also try ..." in the search results is a completely new kind of feature offering suggested queries from the most popular PubMed queries that contain the current search term. Using these suggested queries may provide more precise results than the current search.

Certain Discovery components improve and make more obvious previously existing pathways that are powerful but may have been cryptic before. For instance, the results available for 'Title Searches' and the 'Review' tab have been available by field-limited searches (term[Title], term AND review[Publication Type]). But, despite its usefulness, only a small minority of PubMed searchers use fielded searching.

Likewise, exposing the top five related articles the Abstract Plus view provides a more obvious alternative to the Related Articles available in the Links menu. Related Articles has been removed from the summary view of the search results, but is still available in the Abstract Plus record view. The new feature that highlights review articles here is a popular enhancement called 'Recent Activity' that provides a gateway to the broader literature relevant to a particular field. The Recent Activity component partially replaces the functions of the History tab allowing navigation to previous searches and record views in PubMed. Unlike the History tab however, Recent Activity provides access to searches and record views in other Entrez databases.

Other Discovery components such as the Gene Sensor, triggered when a gene symbol is used in a search, show results from other databases that may be more directly relevant than those from the current database. In some cases these sensors may provide useful results when the current database does not. For example, searching with the human gene symbol STK40 produces no results in PubMed itself. The Gene Sensor, however, reports results

for the search of STK40 in the Gene database and provides a link to the human, mouse, rat, and the complete search results in Gene. These gene records have links to a variety of molecular data including mRNA and genomic sequences, genomic regions and maps, expression information, homologs in other species, and in the case of the human gene, a literature citation when PubMed found none.

## Upcoming Changes

Many new Discovery-related changes will be appearing over the next several months. Discovery components will be released initially to a fraction of users so that NCBI can measure the effect and popularity of changes to the system. Some changes will appear first in PubMed and will be implemented in other databases as appropriate. The Recent Activity component, for example, was recently ported to some of the molecular databases after a test period in PubMed. In some cases, components will be database-specific such as the Taxonomy report that appears in the right-hand column of Entrez sequence database search results. In all cases the goal of these changes is to improve the usability and quality of results obtained from NCBI services.

## Summary

The NCBI Website is the premier portal to biomedical literature and molecular biology data. Interconnecting the literature and these data is an enormously rich set of similarity and linkage relationships where previously unknown connections are waiting to be uncovered. While access to these connections has always been possible, it's clear that many visitors are not enjoying the full potential of the system. Recent and upcoming changes to the NCBI Web experience will help expose these connections and promote discovery of the most biologically significant records and relationships in the NCBI databases.

## New Databases and Tools

### Bookshelf

The Bookshelf has added three new books: *BLAST: Command Line Applications User Manual*, *The Intolerable Burden of Malaria: A New Look at the Numbers*, and *The Intolerable Burden of Malaria III: Progress and Perspectives*. Books can be found at: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

### Genome Build

Build 1.1 of *Physomitrella patens* (moss) is available in the Genomes database and on the NCBI Map Viewer. The Map Viewer page for this organism is: [www.ncbi.nlm.nih.gov/mapview/map\\_search.cgi?taxid=3218](http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3218)

### Microbial Genomes

Twenty-six finished microbial genomes were released between December 19 and January 14. The original sequence data files submitted to GenBank/EMBL/DDBJ are available on the FTP site: <ftp://ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are also available: <ftp://ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 169.0 is available via web and FTP that includes information as of December 11, 2008. A new release will be available in February 2009. A new linetype, DBLINK, will be implemented in GenBank files beginning with the February 2009 release. More information can be found in Section 1.4.1 of the GenBank Release Notes. Release notes can be found on the ftp site: <ftp://ncbi.nih.gov/genbank/gbrel.txt>

All: 128 Review: 11

Gene Information

**TPH1** tryptophan hydroxylase 1 [Homo sapiens]

This gene encodes a member of the pterin-dependent aromatic acid hydroxylase family. The encoded protein catalyzes the f... [More](#)

Location: 11p15.3-p14

▶ [tph1 in Homo sapiens](#) | [Mus musculus](#) | [Rattus norvegicus](#) | [All 12 Gene records](#) Gene

Items 1 - 20 of 128 Page | 1 | of 7 Next

1: [Dopamine-melatonin neurons in the avian hypothalamus and their role as photoperiodic clocks.](#)  
 El Halawani ME, Kang SW, Leclerc B, Kosonsiriluk S, Chaiseha Y.  
 Gen Comp Endocrinol. 2008 Dec 11. [Epub ahead of print]  
 PMID: 19114045 [PubMed - as supplied by publisher]  
[Related Articles](#)

2: [Resequencing of serotonin-related genes and association of tagging SNPs to citalopram response.](#)  
 Peters EJ, Slager SL, Jenkins GD, Reinalda MS, Garriock HA, Shyn SI, Kraft JB, McGrath PJ, Hamilton SP.  
 Pharmacogenet Genomics. 2009 Jan;19(1):1-10.  
 PMID: 19077664 [PubMed - as supplied by publisher]  
[Related Articles](#)

3: [Lrp5 controls bone formation by inhibiting serotonin synthesis in the duodenum.](#)  
 Yadav VK, Ryu JH, Suda N, Tanaka KF, Gingrich JA, Schütz G, Glorieux FH, Chiang CY, Zajac JD, Insogna KL, Mann JJ, Hen R, Ducey P, Karsenty G.  
 Cell. 2008 Nov 28;135(5):825-37.  
 PMID: 19041748 [PubMed - indexed for MEDLINE]  
[Related Articles](#)

4: [Serotonin genes and gene-gene interactions in borderline personality disorder in a matched case-control study.](#)  
 Ni X, Chan D, Chan K, McMain S, Kennedy JL.  
 Prog Neuropsychopharmacol Biol Psychiatry. 2008 Nov 12. [Epub ahead of print]  
 PMID: 19032968 [PubMed - as supplied by publisher]  
[Related Articles](#)

**Also try:**

- ▶ [tph1](#) tph2
- ▶ [tph1](#) knockout
- ▶ [tph1](#) gene
- ▶ [tph1](#) polymorphism
- ▶ [tph1](#) depression

**Titles with your search terms**

- ▶ No association of **TPH1** 218A/C polymorphism with treatment response and ir [Neuropsychobiology. 2007]
- ▶ Stress upregulates **TPH1** but not **TPH2** mRNA in the rat dorsal raphe nucleus: ide [Cell Mol Neurobiol. 2008]
- ▶ **TPH2** and **TPH1**: association of variants and interactions with heroin addiction. [Behav Genet. 2008] » See all...

**Recent Activity** Turn Off Clear

- 🔍 [TPH1](#) (128)
- 📄 The medical treatment of obsessive-compulsive disorder and anxiety.
- 🔍 [clomipramine](#) (3295) PubMed
- 📄 Maylandia zebra M...[gci:193992698]
- 🔍 [\(Cichlidae\) AND "Maylandi...](#) (105438) Nucleotide

**Figure 1.** The new PubMed display for document summaries for a search with TPH1 (tryptophan hydroxylase 1). The gene symbol TPH1 triggers the Gene Sensor that appears at the top of the display. The Gene Sensor provides a direct link to the human gene record that has associated molecular data and a curated set of literature citations. Direct links to genes of the same name in rat, mouse, and a full search in the Gene database are also available. The right-hand discovery column shows the top five related searches listed under “Also try ...”, the result set obtained by a title search with the term highlighted in the title, and “Recent Activity”, the last five page views or searches performed. The Review tab at the top of the results displays only the review articles when selected.

## Updates and Enhancements

### RefSeq

RefSeq Release 33 is now available. This full release incorporates genomic, transcript, and protein data available as of January 16, 2009 and includes 10,325,282 records from 7,773 different organisms. The RefSeq website is: <http://www.ncbi.nlm.nih.gov/RefSeq/>. RefSeq data are also available through FTP: <ftp://ftp.ncbi.nih.gov/refseq/release/>.

### Short Read Archive

The Short Read Archive, or SRA, is now available for searching in the Entrez system. “SRA” has been added as a choice in the Entrez search pulldown menu.

### PubMed

PubMed has added a Recent Activity feature to the PubMed Abstract display page. This feature displays recent searches performed and the number of results found for those searches. The Recent Activity box also appears in some other Entrez database results pages with the databases being displayed within the box. This feature can be turned off if unwanted.

The PubMed and MeSH databases have been updated with the 2009 MeSH vocabulary.

1: [PLoS ONE](#). 2008;3(10):e3301. Epub 2008 Oct 15.

Open access to full text on  
**PLoS one**

**FREE** full text article  
in PubMed Central

Links

**Genetic disruption of both tryptophan hydroxylase genes dramatically reduces serotonin and affects behavior in models sensitive to antidepressants.**

**Savelieva KV, Zhao S, Pogorelov VM, Rajan I, Yang Q, Cullinan E, Lanthorn TH.**

Lexicon Pharmaceuticals Incorporated, The Woodlands, TX, USA. ksavelieva@lexpharma.com

The neurotransmitter serotonin (5-HT) plays an important role in both the peripheral and central nervous systems. The biosynthesis of serotonin is regulated by two rate-limiting enzymes, tryptophan hydroxylase-1 and -2 (TPH1 and TPH2). We used a gene-targeting approach to generate mice with selective and complete elimination of the two known TPH isoforms. This resulted in dramatically reduced central 5-HT levels in Tph2 knockout (TPH2KO) and Tph1/Tph2 double knockout (DKO) mice; and substantially reduced peripheral 5-HT levels in DKO, but not TPH2KO mice. Therefore, differential expression of the two isoforms of TPH was reflected in corresponding depletion of 5-HT content in the brain and periphery. Surprisingly, despite the prominent and evolutionarily ancient role that 5-HT plays in both vertebrate and invertebrate physiology, none of these mutations resulted in an overt phenotype. TPH2KO and DKO mice were viable and normal in appearance. Behavioral alterations in assays with predictive validity for antidepressants were among the very few phenotypes uncovered. These behavioral changes were subtle in the TPH2KO mice; they were enhanced in the DKO mice. Herein, we confirm findings from prior descriptions of TPH1 knockout mice and present the first reported phenotypic evaluations of Tph2 and Tph1/Tph2 knockout mice. The behavioral effects observed in the TPH2 KO and DKO mice strongly confirm the role of 5-HT and its synthetic enzymes in the etiology and treatment of affective disorders.

PMID: 18923670 [PubMed - indexed for MEDLINE]

PMCID: PMC2565062

**Related Articles**

- ▶ Late developmental stage-specific role of tryptophan hydroxylase 1 in brain serotonin levels. [J Neurosci. 2006]
- ▶ Tryptophan hydroxylase 1 knockout and tryptophan hydroxylase 2 polymo [Am J Physiol Lung Cell Mol Physiol. 2007]
- ▶ Deficiency of brain 5-HT synthesis but serotonergic neuron formation in Tph2 knockout mice. [J Neural Transm. 2008]
- ▶ **Review** [Abnormal cardiac activity in mice in the absence of peripheral serotonin synthesis] [J Soc Biol. 2004]
- ▶ **Review** Developmental role of tryptophan hydroxylase in the nervous system. [Mol Neurobiol. 2007]

» See Reviews... | » See All...

**Recent Activity**

[Turn Off](#) [Clear](#)

- Genetic disruption of both tryptophan hydroxylase genes dramatically reduces serotonin and...
- Crystal structure of tryptophan hydroxylase with bound amino acid substrate.
- Related Reviews for PubMe... (41) PubMed
- Deficiency of brain 5-HT synthesis but serotonergic neuron formation in Tph2 knockout mice...
- Modulation of peripheral serotonin levels by novel tryptophan hydroxylase inhibitors for t...

**Figure 2. The Links menus from BioSystems records.** Discovery components in the right hand column are the pre-computed Related Articles and Recent Activity. The Related Articles component highlights review articles and provides a link to all 100 related items in PubMed and a separate link to the ten related reviews.

## HomoloGene

HomoloGene now offers multiple sequence alignments (MSAs) generated using MUSCLE. MSAs of genes and their homologs can be viewed by choosing a HomoloGene cluster, and either clicking on the "Show Multiple Alignment" link, or by choosing "Multiple Alignment" from the display menu. The HomoloGene Web resource is: [www.ncbi.nlm.nih.gov/sites/entrez?db=homologene](http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene).

## BLAST

Binaries for BLAST 2.2.19 are available from the ftp site: <ftp://ncbi.nlm.nih.gov/blast/executables/LATEST/>. Some changes include: the BLASTDB environment variable now supports multiple database search paths; a smaller protein table is available to improve performance; and 'formatrpsdb' supports creation of databases larger than two gigabytes.

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## NCBI News, January 2009

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: December 15, 2008.

### Featured Resource: BLAST+, All New BLAST Available on Web Service and for Download

The all-new NCBI BLAST+, built with the NCBI C++ toolkit, is now live on the NCBI Web service and is available for download from the BLAST area of the ftp site. BLAST+ features improved integration between the Web service and standalone package including the ability to save and use search strategies developed on the Web in a local installation. Also featured is a logical correspondence between the local binaries and the functions apparent on the Web (blastn, blastp, tblastn, tblastx, psiblast, and rpsblast). The new binaries offer improved performance, and new, more flexible formatting options. Another improvement is that the separate network BLAST client and BLAST 2 Sequences software are no longer needed as they are incorporated as options in the program binaries.

#### Web Service

Important changes to the Web service include the incorporation of the BLAST 2 Sequences service on the main BLAST submission forms (described in the December 2008 issue of NCBI News), changes to the output format, and new options for downloading results and search strategies.

Figure 1 shows the new output for a multi-sequence protein BLAST search. Results for searches with multiple sequences are easier to read in the new format as the results are displayed one-at-a-time with the results from each query accessible from the “Results for ...” pull-down list. In addition, sequences with no hits can be quickly identified in the pull-down list by grayed-out titles.

The new output format features expandable / collapsible sections of the BLAST results (Graphic Summary, Conserved Domains, Descriptions, and Alignments) that allow selecting only the most relevant sections of a particular output for display. There are also expandable Formatting and Download options sections on the output page. These new options are more convenient than the previous separate Formatting page, and the download options now provide a direct way to save BLAST results as text, structured formats (XML or ASN.1), or as the hit table tabular format. In addition, the hit table is now available in a comma separated value (CSV) format that can be directly opened with standard spreadsheet applications.

#### Standalone BLAST Package

The BLAST+ package is available for a variety of computer platforms from the NCBI ftp site:

<ftp.ncbi.nih.gov/blast/executables/blast+/LATEST>

There are a number of important differences between BLAST+ and the traditional standalone BLAST package. The most apparent is the separation of the blastall program into individual binaries by BLAST search function and the replacement of other traditional programs. Table 1 presents a partial correspondence between the old and the new package programs.

The new programs also feature long-form command line options instead of the traditional single letter options, making the options easier to remember and providing more flexibility for additional options in later releases. The box below shows how the same search – blastn against the NCBI nucleotide database (nt) – would be

8 sequences (gi|151536252|gb|EV394895.1|EV394895...

Results for: 4:|cl|6865 gi|151536249|gb|EV394892.1|EV394892 EST\_sfon\_evi\_753977 sfonevi mixed\_tissue Salvelinus fontinalis ...(678bp) ▾

Query ID Description

1:|cl|6862 gi|151536252|gb|EV394895.1|EV394895 EST\_sfon\_evi\_754362 sfonevi mixed\_tissue Salvelinus fontinalis ...(791bp)

2:|cl|6863 gi|151536251|gb|EV394894.1|EV394894 EST\_sfon\_evi\_753978 sfonevi mixed\_tissue Salvelinus fontinalis ...(921bp)

3:|cl|6864 gi|151536250|gb|EV394893.1|EV394893 EST\_sfon\_evi\_754361 sfonevi mixed\_tissue Salvelinus fontinalis ...(678bp)

4:|cl|6865 gi|151536249|gb|EV394892.1|EV394892 EST\_sfon\_evi\_753977 sfonevi mixed\_tissue Salvelinus fontinalis ...(678bp)

\*5:|cl|6866 gi|151536248|gb|EV394891.1|EV394891 EST\_sfon\_evi\_754360 sfonevi mixed\_tissue Salvelinus fontinalis ...(791bp)

\*6:|cl|6867 gi|151536247|gb|EV394890.1|EV394890 EST\_sfon\_evi\_753976 sfonevi mixed\_tissue Salvelinus fontinalis ...(791bp)

7:|cl|6868 gi|151536246|gb|EV394889.1|EV394889 EST\_sfon\_evi\_754359 sfonevi mixed\_tissue Salvelinus fontinalis ...(841bp)

8:|cl|6869 gi|151536245|gb|EV394888.1|EV394888 EST\_sfon\_evi\_753975 sfonevi mixed\_tissue Salvelinus fontinalis ...(927bp)

▼ Graphic Summary

Distribution of 10 Blast Hits on the Query Sequence

NP\_955830 RNA terminal phosphate cyclase domain 1 [Danio rerio] S=176 E=2.4e-43

Color key for alignment scores

Query

0 100 200 300 400 500 600

<40 40-50 50-80 80-200 >=200

▼ Descriptions

▼ Alignments  Select All [Get selected sequences](#)

```
>|ref|NP_955830.1| UG RNA terminal phosphate cyclase domain 1 [Danio rerio]
|gb|AAH46087.1| G RNA terminal phosphate cyclase domain 1 [Danio rerio]
|gb|AAQ97842.1| G RTC domain containing 1 [Danio rerio]
Length=363

GENE_ID: 321129_rtccl | RNA terminal phosphate cyclase domain 1 [Danio rerio]
(10 or fewer PubMed links)

Score = 176 bits (447), Expect = 2e-43
Identities = 86/98 (87%), Positives = 92/98 (93%), Gaps = 1/98 (1%)
Frame = +2

Query 35  QGVYADKVGFEAAEMLLRNIRHNGCVDFLQDQLILFMALANGTSRIRTPVTLHTQTAI 214
+GVYADKVG EAAEMLLRNIRHNGCVDFLQDQLI+FMALANGTSR+RTGP+TLHTQTAI
Sbjct 265  KGVYADKVGIEAAEMLLRNIRHNGCVDFLQDQLIIFMALANGTSRMRTPITLHTQTAI 324

Query 215  HVAEQLTQAKFTVTKAEDENASNDTYITECQGVGTNP 328
HVAEQLT AKF ++KAEDENA NDTYI ECQGVG TNP
Sbjct 325  HVAEQLTNAKFAISKAEDENA-NDTYIIECQGVGATNP 361
```

**Figure 1.** The new Web BLAST output for a multiple sequence blastx search showing the pull-down list for the query sequences (upper panel) and collapsible sections (lower panel). Sequence titles with no hits are grayed-out in the pull-down list. The Descriptions section of the output is collapsed in this view.

written in traditional BLAST and BLAST+. The BLAST+ command line closely parallels the way the same search would be conducted on the Web service: select the nucleotide form from the Basic BLAST section of the BLAST Homepage – the equivalent of using the blastn binary, choose blastn from the Program Selection portion of the submission form – the equivalent of using the task blastn.

Traditional BLAST:

```
blastall -i input.seq -d nt -p blastn -e 0.001 -o output
```

## BLAST+

```
blastn -query input.seq -db nt -task blastn -evalue 0.001 -out output
```

**Table 1. A partial list of corresponding programs in traditional BLAST and the new BLAST+.**

BLAST+	Traditional BLAST program
blastn -task blastn	blastall -p blastn
blastp	blastall -p blastp
blastx	blastall -p blastx
tblastn	blastall -p tblastn
tblastx	blastall -p tblastx
blastn -task megablast	megablast
psiblast	blastpgp
makeblastdb	formatdb
blastdbcmd	fastacmd

## New Options

New standalone options for BLAST+ include tasks, search strategies, and custom output formats.

The task option manages the different aspects of nucleotide searches for the blastn binary (blastn, megablast, discontinuous megablast) but also offers pre-set conditions for short sequence searches (`-task blastn-short`, `-task blasp-short`).

Another of the completely new features in the BLAST+ package is the ability to import and export search strategies as mentioned above for Web BLAST. In the standalone package this is managed through the `-export_search_strategy` and `-import_search_strategy` command line switches. This allows reproducible exchange of search parameters between the NCBI Web service and the local installation. For greater flexibility, a different query and database can also be specified when using a saved search strategy.

Custom outputs are available now for the BLAST tabular and comma-separated formats. These offer a large range of combinations of statistics, locations, and identifiers. Custom output formats are managed through format specifiers passed to the `-outfmt` option in the search program. More detailed information on these and other options for any of the standalone program is available by using the `-help` option on the command line of the program.

## Replacement of BLAST 2 Sequences and the Netblast client

The BLAST+ package now includes the ability to compare two or more sequences to each other in each of the search programs. Using the `-subject` option instead of the `-db` will cause any of the search programs to behave as a BLAST 2 sequences program. This eliminates the need for the BLAST 2 sequences utility (bl2seq) included in the traditional BLAST package. As with the other changes this now makes the standalone package more similar to the Web version where the BLAST 2 sequences option is available for all Basic BLAST searches.

The BLAST+ package can also function as a network client to the NCBI Web BLAST service by specifying the `-remote` option on the command line of one of the search programs. This offers a more powerful option for submitting small to medium sized batches of sequences than the older netblast client (blastcl3) available as a separate distribution on the NCBI ftp site.

## Other Improvements and Additional Help

Other important improvements in BLAST+ include performance enhancements using query splitting, the ability to use pre-indexed databases for megablast as well as new abilities and options for the BLAST database applications `makeblastdb` and `blastdbcmd`. For additional information on any of these improvements and help on using and installing BLAST+ see the *BLAST Command Line Applications User Manual* available on the NCBI Bookshelf.

[www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpblast&part=CmdLineAppsManual](http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helpblast&part=CmdLineAppsManual)

## Summary

The new BLAST+ suite of programs offers more flexible options, enhanced performance, and improved integration with the NCBI Web BLAST services compared with the C toolkit BLAST. BLAST+ will continue to evolve and improve as the primary sequence similarity search tool at the NCBI and the most widely used similarity search tool in the world.

## New Databases and Tools

### Bookshelf

The Bookshelf has added two new books: *Animal Models of Cognitive Impairment* and *Baculovirus Molecular Biology*. Books can be found at: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

### Microbial Genomes

Sixteen finished microbial genomes were released between November 17 and December 17. The original sequence data files submitted to GenBank/EMBL/DDBJ can be found at: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are available via FTP at: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 169.0 is available via web and FTP. Release 169.0 includes information as of December 11, 2008.

A new linetype, DBLINK, will be implemented in GenBank files beginning with the February 2009 release. More information can be found in Section 1.4.1 of the GenBank Release Notes. Also, GenBank 'index' files are now provided without EST content, and without most GSS content. See Section 1.3.12 of the release notes for further details. Release notes can be found on the ftp site at: <ftp.ncbi.nih.gov/genbank/gbrel.txt>

NCBI is considering ceasing support for the index files. Affected users are encouraged to review Section 1.3.2 of the release notes and provide feedback to the GenBank newsgroup or the NCBI service desk at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov).

## Updates and Enhancements

### PubMed DocSum

The “Gene Sensor” is a new feature seen on PubMed results pages that provides definitive information for a gene. If a PubMed search is performed with a Gene Symbol, a Gene Information box appears on the results page with a brief description of the gene along with a link to the Entrez Gene database.

The PubMed Document Summary (DocSum) page has also been updated to provide more emphasis on an article’s title rather than its authors. These new features are part of an effort to improve the quality of search results and promote the discovery of new information.

### Sequin Version 9.0

Sequin version 9.00 for Macintosh, PC/Windows, and Unix computers is available from the NCBI ftp site: <ftp.ncbi.nih.gov/sequin>. Feature and qualifier dialogs have been updated to comply with the latest version of the International Sequence Database Collaboration (INSDC) Feature Documentation.

### PubMed Entrez Utilities

Updated 2009 NCBI PubMed Document Type Definitions (DTDs) are available from the Entrez DTD page at: <eutils.ncbi.nlm.nih.gov/entrez/query/DTD/index.shtml>. DTD changes for the 2009 production year are noted in the Revision Notes section near the top of each DTD.

### HomoloGene

HomoloGene release 62 is available via web and ftp. This new release includes an improved approach for predicting putative paralogs. The HomoloGene website is: [www.ncbi.nlm.nih.gov/homologene/](http://www.ncbi.nlm.nih.gov/homologene/). The website has a “Tip of the Day” feature that provides information on designing more successful searches.

### Entrez Gene

A few changes will be seen in the Entrez Gene database in the coming weeks. An option to sort by chromosome will present gene records first alphabetically by organism name, then numerically by chromosome, and finally numerically by the start position on the chromosome. This chromosome information will now appear in the Document Summaries returned by searches in Gene.

Different gene-to-gene relationships will soon be reported including a “Related functional gene” heading and link within the Entrez Gene full report. The functional gene will have a link to related pseudogenes in the “General gene information” section.

A new preferred symbol field contains the preferred symbol for a gene. The alias for this field is [PREF].

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, December 2008

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: November 19, 2008.

### Featured Resource: BLAST 2 Sequences Is Now Part of the Main BLAST Web Service

The former BLAST utility program Blast 2 Sequences is now available on all main BLAST Web forms. This program eliminates the need for the separate service formerly linked as a specialized BLAST page. More importantly, the service now provides the full functionality of the Web BLAST engine, the BLAST formatter, and the ability compare more than one sequence in a single search.

#### Accessing and Using the Service

A Blast 2 Sequences checkbox is now available on all BLAST forms linked to the Basic BLAST section of the BLAST homepage ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)). The old Blast 2 Sequences link on the BLAST Homepage will link directly to the Basic Nucleotide form already set up for Blast 2 Sequences. On the BLAST submission page, checking the Blast 2 Sequences checkbox changes the form so that the database selection portion is eliminated, and a new text area appears for entering target sequences (Figure 1). The newly added program tabs allow rapid selection of a different program if desired (blastp, blastn, blastx, tblastn, tblastx). Sequences in FASTA format or database accession numbers may be entered in either of the two text areas. Sets of sequences may also be uploaded to the service from files on disk. The ability to enter multiple sequences makes it possible to search a small custom database on the web or to perform an all-against-all comparison of a small number of sequences.

#### Output Format

The Blast 2 Sequences output provides a completely redesigned dot plot of the alignment that features a higher resolution display than the old Blast 2 Sequences service. The dot-plot may be collapsed if desired. The dot-plot is useful in many cases for identifying repeated domains in proteins or insertions, inversions, and translocations in nucleotide sequences. Figure 2 shows the dot-plot of the alignment of two bacterial genomes with large-scale relative rearrangements. The rearrangements are easily visible in the plot.

All formatting options available for the main Web BLAST service are also available for Blast 2 Sequences. These options include alternative alignment views such as pairwise, query anchored with mismatch and identity highlighting, and downloadable structured formats such as ASN.1, XML, and hit table. The ability to display results as a distance tree using the BLAST Treeview link is particularly useful when the Blast 2 Sequences option is used with a small custom database as shown in Figure 3 for selected vertebrate aromatic amino acid hydroxylases.

#### Summary

The Blast 2 Sequences utility is now fully integrated into the main NCBI BLAST service, providing the full power of the Web BLAST service along with flexible formatting options for comparing two sequences. This feature also extends the capability of Blast 2 Sequences to make it a “Blast several Sequences” service. These expanded options and abilities make Blast 2 Sequences a new, powerful, and flexible sequence analysis tool.

The image shows the NCBI Basic nucleotide BLAST form. At the top, there are tabs for different BLAST programs: **blastn**, **blastp**, **blastx**, **tblastn**, and **tblastx**. Below the tabs, a header reads "BLASTN programs search nucleotide subjects using a nucleotide query. [more...](#)".

The form is divided into two main sections: "Enter Query Sequence" and "Enter Subject Sequence".

**Enter Query Sequence:** This section contains a large text input field for "Enter accession number, gi, or FASTA sequence" with a "Clear" button. To the right, there is a "Query subrange" section with "From" and "To" input fields. Below the text field, there is an "Or, upload file" section with a "Browse..." button. A "Job Title" field is also present with the instruction "Enter a descriptive title for your BLAST search". A checkbox labeled "Blast 2 sequences" is checked.

**Enter Subject Sequence:** This section is identical in layout to the query section, with a text input field for "Enter accession number, gi, or FASTA sequence", a "Clear" button, a "Subject subrange" section with "From" and "To" fields, and an "Or, upload file" section with a "Browse..." button.

**Figure 1.** The new **Blast 2 Sequences** option on the Basic nucleotide BLAST form. Either one or many sequences can be entered into the two text areas or uploaded from files. Rapid access to other BLAST programs is available through the tabs at the top of the form.

## New Databases and Tools

### PMID : PMCID Converter

A converter is available to translate ID numbers for articles found in both PubMed and PubMed Central. The converter will convert IDs from PubMed to PMC and vice versa. It can be found at: <http://www.ncbi.nlm.nih.gov/sites/pmctopmid>

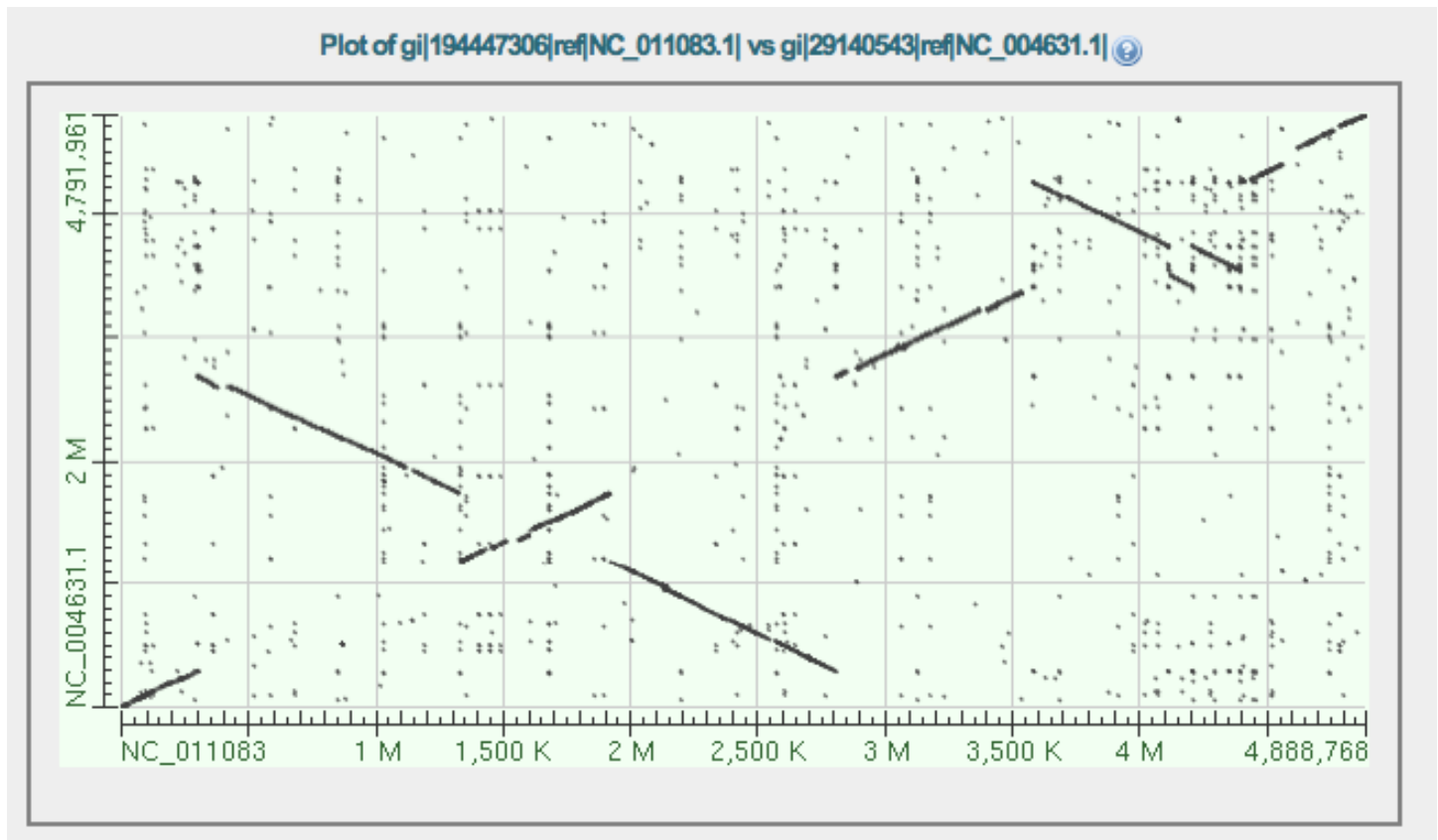
### Bookshelf

The Bookshelf has added two new books entitled: *The Epilepsies: Seizures, Syndromes, and Management* and *Essentials of Glycobiology*. Books can be found at: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

### Genome Resource Guide

An Aphid Genome Resource page is available at: [www.ncbi.nlm.nih.gov/projects/genome/guide/aphid/](http://www.ncbi.nlm.nih.gov/projects/genome/guide/aphid/). This page provides a gateway to aphid resources both within NCBI and the outside scientific community.





**Figure 2.** The expanded Dot Matrix view from Blast 2 Sequences showing the alignment of two *Salmonella enterica* subsp. *enterica* genome sequences (serovar Heidelberg str. SL476, accession NC\_001083 and serovar Typhi Ty2, accession NC\_004631). Three large-scale relative inversions and a smaller translocation are apparent as cross diagonal matches. Shared repetitive sequences appear as a characteristic set of off- diagonal column and row matches.

## Microbial Genomes

Fifteen finished microbial genomes were released from October 10-November 17. The original sequence data files submitted to GenBank/EMBL/DDBJ can be found at: <ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are available via FTP at: <ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

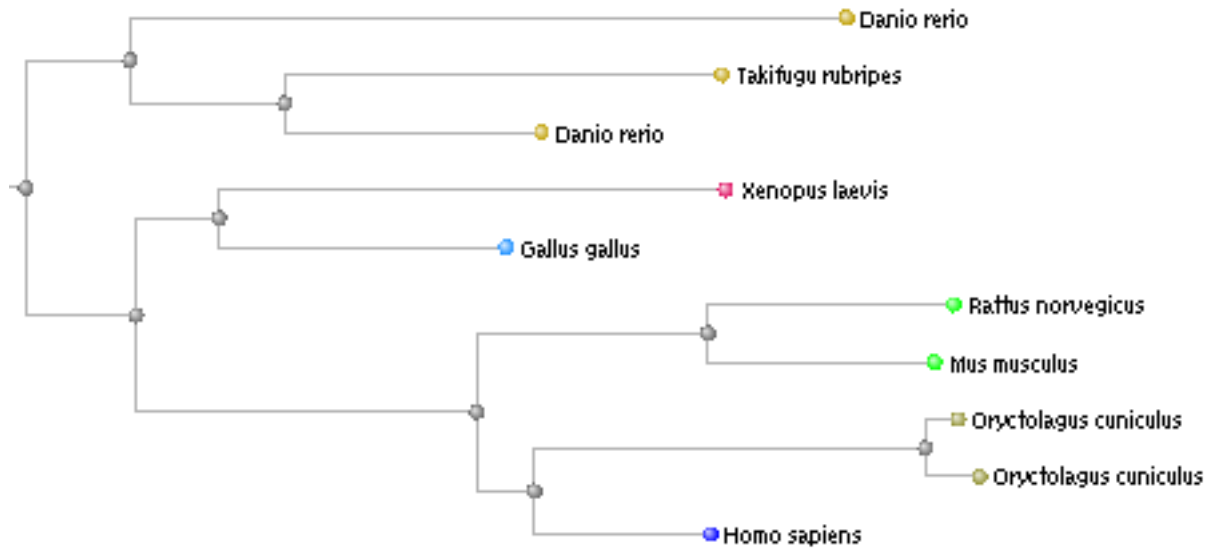
GenBank release 168.0 is available via web and FTP that includes information as of October 27, 2008. The database increased by 19.84 Gigabases since the last release, 167.0. This is a milestone for the largest growth between single releases.

## Updates and Enhancements

### RefSeq

RefSeq Release 32 is available via web and FTP. This full release incorporates genomic, transcript, and protein data available as of November 7, 2008 and includes 9,145,702 records from 5,513 different organisms. The RefSeq website is: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/).

RefSeq records for microbial projects with the prefix 'NZ' will now include two style formats where there may be 2 or 4 alphabetic characters following the underscore. These new accession formats have been designed to



**Figure 3.** A portion of the BLAST Treeview generated from the Blast 2 Sequences results for the alignment of human phenylalanine hydroxylase (accession NP\_00068) with a set of 34 other vertebrate aromatic amino acid hydroxylases. The portion shown here contains the tryptophan hydroxylase 1 homologs from human (*Homo sapiens*), NP\_004170; rabbit (*Oryctolagus cuniculus*), NP\_001093425 and NP\_001075741; mouse (*Mus musculus*), NP\_033440; rat (*Rattus norvegicus*), NP\_001094104; chicken (*Gallus gallus*), NP\_990287; *Xenopus laevis*, NP\_001080923; zebrafish (*Danio rerio*), NP\_001001843 and NP\_840091; and pufferfish (*Takifugu rubripes*), NP\_001027848.

support duplicating WGS microbial scaffolds and complete genomic models when a project is a mixture of contigs and scaffolds.

## Genome Assembly

Genome annotation for *Anopheles gambiae* build AgamP3.3 and *Caenorhabditis elegans* build WS190 were released in October. For more annotation updates and links to Genome Resource pages see: [www.ncbi.nlm.nih.gov/Genomes/](http://www.ncbi.nlm.nih.gov/Genomes/).

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. Please see: [www.ncbi.nlm.nih.gov/feed/](http://www.ncbi.nlm.nih.gov/feed/)

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.

## NCBI News, November 2008

Dawn Lipshultz, M.S.<sup>1</sup> and Peter Cooper, Ph.D.<sup>2</sup>

Created: October 15, 2008; Updated: October 28, 2008.

### Featured Resource: Primer-BLAST—NCBI's Primer Designer and Specificity Checker

NCBI now offers Primer-BLAST, a Web service for designing target specific oligonucleotide primers for use in polymerase chain reaction (PCR) protocols. Primer-BLAST may be accessed from the NCBI BLAST Homepage or through the direct URL:

<http://www.ncbi.nlm.nih.gov/tools/primer-blast/>

This service integrates primer designing code from the popular Primer3<sup>1</sup> software with a specificity check that uses a custom BLAST search. Primer-BLAST eliminates the need to design primers at another site and then perform a search with those primers using the main NCBI BLAST service to check specificity. Moreover, Primer-BLAST can design primers that amplify only a particular splice variant of a gene – an important feature for PCR protocols measuring tissue specific expression. Primer-BLAST has a number of additional enhancements that make it better at picking specific primers than Primer3 and NCBI BLAST used separately.

#### Primer-BLAST Input

The Primer-BLAST interface contains elements of both the Primer3 tool and the NCBI BLAST Web service. The submission form is divided into three sections with commonly adjusted settings relevant to the **target template**, **the primers**, and the background **specificity check**. As with other BLAST services, the more specialized options are available under an expandable section at the bottom of the submission form, called “Advanced parameters” in Primer-BLAST.

#### Template

The target template sequence or accession number to be analyzed is placed in the large text area in the **PCR Template** section of the form. Specific binding regions on the template may be specified using the range boxes. With just the target template, Primer-BLAST will design primer pairs that are specific to the target template and unique in the target database.

#### Primers

One or both primers may be supplied in the **Primer Parameters** section. The most important and commonly modified primer-specific parameters such as desired product size, T<sub>m</sub> ranges and T<sub>m</sub> difference may be modified here as well. Primers may accompany a template, but may also be used without a template. With only a primer pair, Primer-BLAST identifies potential templates in the target database. In this case, the results are more precise than a specificity check with ordinary BLAST and have additional PCR condition parameters such as melting temperature (T<sub>m</sub>) provided in the results.

#### Specificity

Parameters that may be adjusted in the **Specificity Checking** section of the form affect the nature of the background database and the level of stringency for avoiding nonspecific matches. The database specificity check does not simply find BLAST matches but includes adjustable mismatch cutoffs that focus on the 3' end of the

binding site to more effectively avoid mispriming. The most informative results are obtained by selecting the target database and organism that best match the template sequence source. The organism selection is managed by typing the name of the organism of interest in the “Organism” box. An auto-complete feature will find the closest match in the NCBI taxonomy database.

Four BLAST nucleotide databases are available for searching: RefSeq mRNA, Genome (selected reference assemblies), Genome (all chromosomes), and nr (the standard non-redundant database). The first two databases will give the most precise results as these are the best annotated and analyzed sets of sequences. In particular, if an NCBI mRNA Reference Sequence is used as a template with the RefSeq mRNA database, Primer-BLAST will design primers specific to the single splice variant represented by the reference sequence. The well characterized genome sequences of human, chimpanzee, mouse, rat, cow, dog, chicken, zebrafish, fruit fly, honeybee, *Arabidopsis*, and rice are available as the selected reference assemblies. These selected reference assemblies will produce results that are less redundant and easier to interpret when using genomic templates from these organisms. The remaining Genome database (all chromosomes) contains all chromosome Reference Sequences (NC\_ accessions) including the selected reference assemblies, as well as their alternative assemblies, and chromosome records for other organisms, most notably microbial genome sequences. Finally, the nr database is available for the widest coverage of organisms.

## Advanced Parameters

The **Advanced Parameters** section contains numerous adjustable settings affecting primer binding and efficacy from the Primer3 program. Two additional features that are only available in Primer-BLAST are options to avoid making primers that match repetitive or regions with biased base composition (low complexity regions) and an option to avoid regions that contain reported nucleotide polymorphisms (SNPs) from NCBI’s SNP database. As with Primer3, Primer-BLAST also has an option to design an internal hybridization probe for each product that can be used in various product detection assays.

## Example

Using Primer-BLAST to design primers to distinguish the two transcripts of the human uracil-DNA glycosylase genes (UNG, GeneID: 7374) demonstrates the utility and precision of Primer BLAST (Figure 1). The UNG products code for enzymes involved in removing misincorporated uracil during DNA replication. The longer UNG1 transcript (NM\_003362 ) codes for a mitochondrial enzyme while the shorter UNG2 transcript (NM\_080911) that lacks the portion coding for the mitochondrial targeting sequence apparently functions in the nuclear genome of replicating cells. The top panel of Figure 1 shows the Primer-BLAST results that are specific to the UNG2 splice variant. In this case the NCBI reference sequence was used as a template with RefSeq mRNA database limited to human. By default the service designs only UNG2 specific primers with the reference sequence as the template. These conditions may be relaxed by checking the option for “Allow primer to amplify mRNA splice variants.” The bottom panel of Figure 1 shows that under these conditions new primers are found that will amplify both transcripts from the UNG gene. Expanding the database coverage provides a means for checking primers in additional species. The results using the nr database predict that many of the primer pairs designed against the human UNG transcripts should also amplify transcripts from other primate species (not shown).

## Summary

Primer-BLAST provides a new more powerful and more comprehensive primer design tool by combining oligonucleotide primer design with specific genome and transcriptome level searches, and will save time and money by eliminating some of the trial and error of PCR primer design.

▶ [NCBI/Primer-BLAST: results](#) [more...](#)

**Input PCR template** [NM\\_080911.1](#) Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA  
**Range** 1 - 2053  
**Specificity of primers** Primer pairs are specific to input template as no other targets were found in selected database: NCBI Transcript Reference Sequences (Organism limited to Homo sapiens)

▼ **Summary of primer pairs**

▼ **Detailed primer reports**

**Primer pair 1**

	Sequence (5'→3')	Strand on template	Length	Start	Stop	Tm	GC%
<b>Forward primer</b>	GCCAGAAGACGCTCTACTCC	Plus	20	78	97	59.19	60.00%
<b>Reverse primer</b>	AGGTGAAGACTTGGTGTGGG	Minus	20	479	460	60.00	55.00%
<b>Product length</b>	402						

**Products on intended target**  
 >[NM\\_080911.1](#) Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```

product length = 402
Forward primer 1  GCCAGAAGACGCTCTACTCC  20
Template       78  ..... 97

Reverse primer 1  AGGTGAAGACTTGGTGTGGG  20
Template      479  ..... 460
    
```

**Primer pair 1**

	Sequence (5'→3')	Strand on template	Length	Start	Stop	Tm	GC%
<b>Forward primer</b>	GCCTTGTTTTCTTGTCTCTGG	Plus	20	813	832	59.99	50.00%
<b>Reverse primer</b>	TTTACCATAAAGGCAAGGG	Minus	20	1311	1292	59.93	45.00%
<b>Product length</b>	499						

**Products on intended target**

>[NM\\_080911.1](#) Homo sapiens uracil-DNA glycosylase (UNG), transcript variant 2, mRNA

```

product length = 499
Forward primer 1  GCCTTGTTTTCTTGTCTCTGG  20
Template       813  ..... 832

Reverse primer 1  TTTACCATAAAGGCAAGGG  20
Template      1311  ..... 1292
    
```

**Products on allowed transcript variants**

>[NM\\_003362.2](#) Homo sapiens uracil-DNA glycosylase (UNG), nuclear gene encoding mitochondrial protein, transcript variant 1, mRNA

```

product length = 499
Forward primer 1  GCCTTGTTTTCTTGTCTCTGG  20
Template       841  ..... 860

Reverse primer 1  TTTACCATAAAGGCAAGGG  20
Template      1339  ..... 1320
    
```

**Figure 1. Primer-BLAST results for UNG transcript variant 2.** The NCBI Reference sequence NM\_080911 was used as a template. **Top panel:** Primers specific to the single splice variant are reported by default with the mRNA RefSeq database limited to human sequences. **Bottom panel:** Primers that amplify both splice variants are found with the option to allow splice variants.

## Reference

1. Steve Rozen and Helen J. Skaletsky (2000) Primer3 on the WWW for general users and for biologist programmers. **In:** Krawetz S, Misener S (eds) *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press, Totowa, NJ, pp 365-386

## New Databases and Tools

### BLAST 2 Sequences

The NCBI BLAST web pages have a new option to align a query against a set of target sequences, rather than a BLAST database. This option allows you to align your query to one or more subject sequences and still use the standard BLAST web interface to optimize your search and change algorithm parameters. Each search is assigned a "Request ID" (RID) and is also listed under the "Recent Results" tab that you can access from the BLAST home page. The results are formatted as a standard BLAST report, except a "Dot Matrix view" (a "dot-plot" like graphic of the alignments) is available in the new report design if only one subject sequence was searched. Links to BLAST web pages can be found at: [blast.ncbi.nlm.nih.gov/BLAST.cgi](http://blast.ncbi.nlm.nih.gov/BLAST.cgi) Step-by-step instructions can be found at: [blast.ncbi.nlm.nih.gov/docs/align\\_seqs.pdf](http://blast.ncbi.nlm.nih.gov/docs/align_seqs.pdf)

### My NCBI

My NCBI version 2.0 was released in September with new features and updated functions. New features include: account and username options; navigation, preferences, and filter improvements; My Saved Searches tool; and My Collections tool. Another new feature, My Bibliography, allows authors to search and collect citations for their publications. For a full description of new and improved features, please see the *NLM Technical Bulletin* article at: [www.nlm.nih.gov/pubs/techbull/so08/so08\\_myncbi\\_redesign.html](http://www.nlm.nih.gov/pubs/techbull/so08/so08_myncbi_redesign.html) . My NCBI is located at: [www.ncbi.nlm.nih.gov/sites/myncbi/](http://www.ncbi.nlm.nih.gov/sites/myncbi/)

### BarSTool

The Barcode of Life project is one in which DNA barcode sequences are being collected from vouchered specimens of all biological species. The idea is to make it easier to identify biological specimens via a short gene sequence from a standardized position in the genome. NCBI has created a submission tool, Barcode Submission Tool (BarSTool), to submit barcode sequences to GenBank. For more information about BarSTool or the Barcode of Life project see: [www.ncbi.nlm.nih.gov/Genbank/barcode.html](http://www.ncbi.nlm.nih.gov/Genbank/barcode.html).

### Bookshelf

The Bookshelf has added two new books entitled: *Hepatitis C Viruses: Genomes and Molecular Biology* and *NCBI News*. Books can be found at: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

### Microbial Genomes

Thirty-two finished microbial genomes were released (August 15-October 10). The original sequence data files submitted to GenBank/EMBL/DDBJ can be found at: <ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are available via FTP at: <ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 167.0 is available via web and FTP that includes information as of August 19, 2008.

## Updates and Enhancements

### Discovery PubMed Pages

PubMed Summary results pages have begun showing related information from other resources on the right side of the page. Drug Sensor, an NCBI tool that detects drug names in a search, was released in August. As of September, additional resources will be shown to a percentage of users to aid in the discovery process.

### RefSeq

RefSeq Release 31 is available via web and FTP. This full release incorporates genomic, transcript, and protein data available as of August 30, 2008 and includes 9,145,702 records, 5,859,648 proteins, and sequences from 5,513 different organisms.

The RefSeq website is: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/).

### Genome Assembly

Genome annotation for *Ciona intestinalis* build 1.1 and *Arabidopsis thaliana* build 8.1 were released in September. For more annotation updates and links to Genome Resource pages see: [www.ncbi.nlm.nih.gov/Genomes/](http://www.ncbi.nlm.nih.gov/Genomes/).

## Announce Lists and RSS Feeds

Fifteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. <http://www.ncbi.nlm.nih.gov/feed/>

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.





## NCBI News, August 2008

Peter Cooper, Ph.D.<sup>1</sup> and Dawn Lipshultz, M.S.<sup>2</sup>

Created: August 26, 2008; Updated: November 25, 2008.

### Featured Resource: New Graphical Sequence Viewer

NCBI recently released a new sequence viewer that allows smooth zooming and browsing of records in a horizontal graphical format. Multiple horizontal tracks can display overlapping features such as assembly details, genes, transcripts, coding regions, protein products and polymorphisms. A multi-panel display also can show views of varying scales from hundreds of megabases at the chromosome level to a single gene view, or all the way down to the individual bases of the sequence. It also produces a text sequence display suitable for copying and pasting sequence to other applications. The sequence viewer is available as the “Graphics” display for nucleotide and protein sequences, the “Sequence Viewer” link from Gene and polymorphism (SNP) records, and from the Genes map in the NCBI Map Viewer.

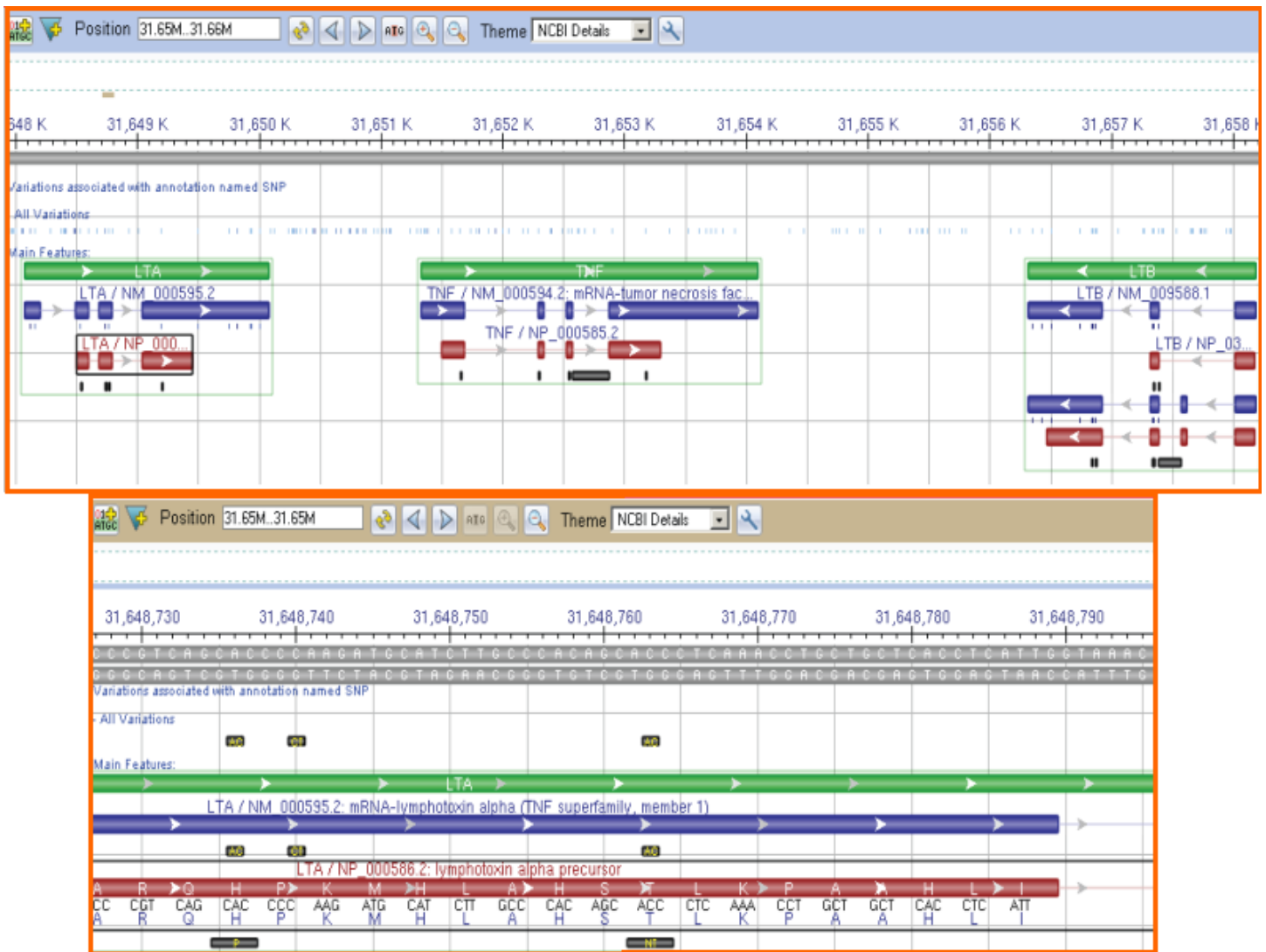
The default display in the sequence viewer consists of two panels. The top panel provides an overview of the entire sequence. This overview panel has a search feature that allows you to find features on the sequence and navigate to them. The wrench icon present on all graphical panels provides access to the configuration controls to change the color scheme, graphical format, and add content layers. The second default panel shows a graphical view (Figure 1) of the region highlighted in the overview. You can add additional graphical views by clicking the icon with the colored bars in the uppermost left of the overview. On each graphical panel a set of icons at the top provides navigational controls including the ability to scroll left or right along the sequence (arrows) and to zoom in or out (plus and minus magnifying glasses.) The “ATG” button in the center is a shortcut that zooms directly to the individual bases of the sequence. You can generate a text view of the region displayed in a graphical panel by clicking the “ATGC” icon in the upper left-hand-corner. The text view is color coded to match the features displayed in the corresponding graphics (genes, coding regions, exons, introns.)

The viewer is particularly well suited to displaying large and complex sequences with many features. The figure shows a region of human chromosome 6 displayed in two panels in the sequence viewer. This region contains three members of the tumor necrosis factor family, lymphotoxin alpha (LTA), tumor necrosis factor (TNF) and lymphotoxin beta (LTB). The lower panel has been configured to display polymorphisms and shows the nucleic acid, corresponding protein sequence and codons. The sequence viewer also features a simplified URL structure that makes it convenient to link to specific views. For example, to see human chromosome 6 in the viewer simply load the following URL in a Web browser:

[www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC\\_000006.10](http://www.ncbi.nlm.nih.gov/projects/sviewer/?id=NC_000006.10)

You can quickly find the tumor necrosis factor gene family members by searching for “lymphotoxin alpha” in the search box in the top panel and following the link to LTA.

The NCBI sequence viewer is powerful new platform for displaying complex sequence features in an easily navigable system. This viewer will continue to evolve and is destined to assume a more central role as a general sequence display and navigation tool on the NCBI Web site.



**Figure 1.** Two detailed graphical panels from the new NCBI sequence viewer showing the tumor necrosis factor gene family members on human chromosome 6 (NC\_000006.10). The lower panel is the zoomed-in display on the second exon of the lymphotoxin alpha gene showing nucleic acid sequence, protein sequence, codons, and polymorphisms.

## New Databases and Tools

### Primer-BLAST

The new Primer-BLAST tool helps users design more effective and gene-specific PCR primers. Primer-BLAST will design primers for an input template target and will also check the specificity of input primers. The tool uses code from the highly regarded Primer3 to help design primers, then uses a specialized BLAST search to eliminate primers that match non-target sequences. To get started with Primer-BLAST go to: [www.ncbi.nlm.nih.gov/tools/primer-blast/](http://www.ncbi.nlm.nih.gov/tools/primer-blast/).

### Bookshelf

The Bookshelf has added three new books entitled: *PubMed Clinical Q&A*, *Drug Class Reviews*, and *Advances in Patient Safety: From Research to Implementation*. Books can be found at: [www.ncbi.nlm.nih.gov/sites/entrez?db=Books](http://www.ncbi.nlm.nih.gov/sites/entrez?db=Books).

## Microbial Genomes

39 finished microbial genomes were released (June 12-August 13). The original sequence data files submitted to GenBank/EMBL/DDBJ can be found at: <ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/>. The RefSeq provisional versions of these genomes are available via FTP at: <ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>.

## GenBank News

GenBank release 166.0 is available via web and FTP that includes information as of June 11, 2008. GenBank has reached a milestone of 200 billion basepairs.

## Updates and Enhancements

### CDTree

CDTree, NCBI's versatile protein domain hierarchy viewer, has released version 3.1. The new version is now supported on Apple computers running OSX and higher. Other new features include an annotation matrix viewer, multi-CD operations, sequence tree coloring for greater flexibility, and higher efficiency in working with large hierarchies. Find more information at: [www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml](http://www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml).

### Clone Finder

The Clone Finder tool has been updated and supports mouse build 37, with more organisms coming shortly. Supported organisms will have a clone finder icon on the Map Viewer home page. Improvements to Clone Finder include: the addition of new libraries; improved searching and download mechanisms; and additional annotation. The mouse Clone Finder page is located at: [www.ncbi.nlm.nih.gov/projects/mapview/mvhome/mvclone.cgi?taxid=10090](http://www.ncbi.nlm.nih.gov/projects/mapview/mvhome/mvclone.cgi?taxid=10090).

### RefSeq

RefSeq Release 30 is available via web and FTP. This full release incorporates genomic, transcript, and protein data available as of July 7, 2008 and includes 8,572,852 records, 5,590,368 proteins, and sequences from 5,395 organisms. The RefSeq homepage is: [www.ncbi.nlm.nih.gov/RefSeq/](http://www.ncbi.nlm.nih.gov/RefSeq/).

## Genome Assembly

An updated genomic assembly of the red flour beetle, version 2.1, was released in June. Zebrafish Zv7, pig Sscrofa5, and horse EquCab2.0, were released in July. Cow build 4.0 and pea aphid build 1.0 were released in August. For more annotation updates and links to Genome Resource pages see: [www.ncbi.nlm.nih.gov/Genomes/](http://www.ncbi.nlm.nih.gov/Genomes/).

## Announce Lists and RSS Feeds

Fourteen topic-specific mailing lists are described on the Announcement List summary page. Announce lists provide email announcements about changes and updates to NCBI resources. [http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email\\_lists.html](http://www.ncbi.nlm.nih.gov/Sitemap/Summary/email_lists.html)

Seven RSS feeds are now available from NCBI including news on PubMed, PubMed Central, NCBI Bookshelf, LinkOut, HomoloGene, UniGene, and NCBI Announce. <http://www.ncbi.nlm.nih.gov/feed/>

Comments and questions about NCBI resources may be sent to NCBI at: [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov), or by calling 301-496-2475 between the hours of 8:30 a.m. and 5:30 p.m. EST, Monday through Friday.



## Archive Issues, 1994-2008

Online and PDF versions of previous are available by clicking on the links below:

### January 2008

dbGaP; Web BLAST Interface Redesign; Genome Builds and Map Viewer; New Protein Cluster Database; GenBank Release 164; RefSeq Release 27; BLAST Lab ([PDF](#))

### Fall/Winter 2006/07

PubMed Abstract Plus; CDTree & Cn3D; Whole Genome Shotgun; New BLAST View; New Genome Builds; UniGene; RefSeq Release 22; GenBank Release 158; NCBI Courses; Submissions Corner; PubChem Grows to 15MM ([HTML](#), [PDF](#))

### Summer 2006

Influenza Database and Tools; Trace Archives at 1 Billion; Entrez Nucleotide Split Database; Third Party Annotation Database; RefSeq Release 18; 1918 Killer Flu Virus; UniGene; GenBank Release 155; Mammoths and Moas at NCBI; Recent NCBI Pubs; NCBI Papers Most Cited; NCBI Courses; BLAST Lab; Genome Builds and Map Viewer. ([HTML](#), [PDF](#))

### November 2005

OMSSA; Probe Database Debut; New Structure Link from Protein; BLAST Download Update; New Microbial Genomes; Nucleotide Database Splits; NCBI 4-Pack Course; RefSeq Release 14; New Organisms in UniGene; GenBank Passes 100 Gigabases; New BLAST Formatter; Splign Alignment Tool; GenBank Release 150; New Genome Builds; Submission Corner ([HTML](#), [PDF](#))

### May 2005

GENSAT; My NCBI; Influenza Virus Resource; NCBI Toolkit; New Microbial Genomes; Iceman Preserved; RefSeq Updates; RefSeq Release; UniGene; GenBank Release 147; New Genome Build; CCDS Database; NCBI Courses; PubMed Corrects Spelling; BLAST Lab; LocusLink Retired ([HTML](#), [PDF](#))

### Summer/Fall 2004

Entrez E-Utilities; PubChem; GenePlot; New NLM Catalog; New Genome Builds; New Microbial Genomes in GenBank; Whole Genome Shotgun Project Page; New Format Option Web BLAST; Trace Archive Grows; New Organisms in UniGene; RefSeq Version 8; Submissions Corner; Predicted Records; GenBank Release 144; BLAST 2.2.10 Released; Recent Publications ([HTML](#), [PDF](#))

### Spring 2004

Transitioning from LocusLink to Entrez Gene; New Cancer Chromosomes Database; HomoloGene's New Look; BLAST Link (BLink); Debut of the HCT Database; 350KB Sequence Length Limit Removed; New Eukaryotic Genomes; Environmental Samples; HIV Protein-Interaction Database; e-PCR and Reverse e-PCR; New Organisms in UniGene; RefSeq Accession Numbers Get Longer; FieldGuidePlus Course; RefSeq Release 6; GenBank Release 142; Entrez Tools 'Hot Spot'; BLAST Lab; New Microbial Genomes; Entrez Quiz ([HTML](#), [PDF](#))

## Fall/Winter 2003

Entrez Query Goes "Global"; Register Your Genome Project; New Genome Builds; Entrez Gene Database; Recent Publications; New Microbial Genomes; KOGs and COGs in CDD; Submission Corner; GenBank Release 139; UniGene Adds Four; RefSeq Version 3 ([HTML](#), [PDF](#))

## Summer 2003

Reference Human Genome; SARS Coronavirus Resource; Gene Expression Omnibus (GEO); Major Histocompatibility Complex (dbMHC); RefSeq Release 1; GenBank Release 137; New Microbial Genomes; Sequence Revision History; BLAST Lab ([HTML](#), [PDF](#))

## Spring 2003

Field Guide to GenBank; Human Reference Sequence; UniGene Expands; Rat Genome Assembly; Taxonomy Browser; Search the NCBI Web Site; Recent Publications; New Genomes; Entrez Quiz; Submission Corner; GenBank Updates; GenBank Release 135 ([HTML](#), [PDF](#))

## Fall/Winter 2002

Plasmodium falciparum Genome; Third Party Annotation Database; Map Viewers; Structure Similarities; PubMed Central; The NCBI Handbook; BLAST Lab; New Microbial Genomes ([HTML](#), [PDF](#))

## Summer 2002

Cn3D 4.0; SNP Population Grows; Viral Reference Sequences; New Genomes in GenBank; Mouse Genome Resources; Mosquito Genome; Sequin MacroSend; BLAST version 2.2.4 ([HTML](#), [PDF](#))

## Spring 2002

Model Maker; Virus Reference Sequences; New MapViewer Displays; Mouse Genome BLAST; Organism-Specific BLAST; ProtEST; Trace Archive Expands; Find Out "About NCBI"; New FTP Hierarchy; Barbara Rapp Leaves NCBI ([HTML](#), [PDF](#))

## Winter 2001

COGs Update; Plant Genomes; LinkOut; Investigator Profile: Stephen Altschul; GenBank News; Expanded Bookshelf; BLAST Enhancements ([HTML](#), [PDF](#))

## Fall 2001

Using TaxPlot to Compare Genomes; New RefSeq Accession Numbers for Curated Genomic Regions; GenBank News; DART Targets Protein Domains; Evidence Viewer Facilitates Analysis of NCBI Human Gene Models ([HTML](#), [PDF](#))

## Spring 2001

Human Genome Map Viewer; Investigator Profile: Eugene V. Koonin; Mouse Genome Resources; UniSTS Integrates Markers; GenBank Mirror Sites; New BLAST Features ([HTML](#), [PDF](#))

## Fall / Winter 2000

The Human Genome Sequence; BLink Enhances Entrez Exploration; Human Gene Nomenclature; Standalone BLAST Additions; Mirror FTP Site for GenBank ([HTML](#), [PDF](#))

## Summer 2000

Conserved Domain Database; Enhanced Taxonomy Access; New Human-Mouse Homology Map; Gene Expression Omnibus; GenBank Adds a Pair of Pathogens; Protein Molecular Weight in Entrez; OMIM in Entrez; Web Server Software for BLAST; PSI-BLAST 2.1; Address Change for FTP Server ([HTML](#), [PDF](#))

## Spring 2000

Dazzling Graphics with Cn3D 3.0; BLAST Offers Taxonomic Views; HomoloGene: Clusters of Clusters; Fly Genome Deposited in GenBank; Drosophila Finds New Home Page; Human Genome Map Viewer ([HTML](#), [PDF](#))

## Winter 2000

Entrez Genomes; IgBLAST; BLAST 1.4; PubMed Central; Mitochondria Energize RefSeq; PSI-BLAST Profiles; Textbooks Linked to PubMed; BankIt 3.0; Mouse and Rat in LocusLink; Malaria Menace Mapped ([HTML](#), [PDF](#))

## Fall 1999

Enhanced Entrez; New Home Page; Cn3D 2.5; News Briefs; GenBank Contig Division; VecScreen; BLAST Lab ([HTML](#), [PDF](#))

## Summer 1999

Decade of Data; LocusLink; Qblast; Cn3D 2.5 Released; RefSeq; Exhibits and Workshops; Coffee Break ([HTML](#), [PDF](#))

## Spring 1999

Human Genome Resource; Sequences from Times Past; Sequin 2.90; SAGEmap; Profile: PSI-BLAST Team; BLAST Lab; Protocol for EST Submissions ([HTML](#), [PDF](#))

## Winter 1999

Submitting Alignments; Compound Accession Number; HIV-1 Subtyping Tool; Changes to BLAST Output; UniGene for Rat; BLAST Lab: Standalone BLAST; Profile: CASP3 Winners ([HTML](#), [PDF](#))

## November 1998

GeneMap '98; Cn3d 2.0; PHI-BLAST; Submitting Identical Sequences from Multiple Sources; Authorin No Longer Accepted; Genes and Disease Web Site; Flexibility Added to BLAST; dbSNP; NCBI Marks 10th Anniversary ([HTML](#), [PDF](#))

## July 1998

Complete Genomes in GenBank; Unfinished Microbial Genomes; Tracking Human Sequencing; BLAST 2 Sequences; GeneMap '98; New Network Clients; Malaria Genetics Web Site ([HTML](#), [PDF](#))

## February 1998

BLAST Version 2.0; COGs; GenBank Submissions; High Throughput Sequencing; GenBank Reaches One Billion ([HTML](#), [PDF](#))

## August 1997

PubMed Launched; Using Sequin; Structure Neighbors; ORF Finder; Electronic PCR; CGAP Revolutionizes Research ([HTML](#), [PDF](#))

## August 1996

See in 3D: New Entrez Release; UniGene Collection; Entrez CD-ROM Discontinued; QUERY E-mail Server; Human/Mouse Homology Map; Images in OMIM; Genome Survey Sequences; Sequin Quick Guide; New BLAST Services ([HTML](#), [PDF](#))

## March 1996

OMIM Database Available; Entrez: Graphical Views of the Genomes Division; Sequin Pre-Release; Entrez in Secure Environments; BLAST Service Update; Entrez CD-ROM Discontinued ([HTML](#), [PDF](#))

## September 1995

GenBank Enters Megabase Era; Entrez Takes Graphical View; GenBank Taxonomy; BankIt Submissions Mount ([HTML](#), [PDF](#))

## March 1995

GenBank Offers BankIt; GenBank Receives Merck Sequences; Entrez on the Net; Molecular Modeling Database; GenBank Fellows ([HTML](#), [PDF](#))

## August 1994

Mosaic Adds dbEST/dbSTS; Entrez CD-ROM's Third Year; Repository CD-ROM Service Ends; GenBank Fellows; Entrez Replaces NCBI-Sequences; GenBank Flat File Expands; Profile: Dan Graur; Special: GenBank Services ([PDF](#))

## February 1994

GenBank's Taxonomy; Access NCBI Through WWW; New STS Database, Division; CD-ROM Entrez Expands; GenBank: Focus on Quality; NCBI's Board; NCBI Services ([PDF](#))