NCBI News

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Department of Health and Human Services

Spring 2004

Transitioning from LocusLink to Entrez Gene

A gene-based view of annotated genomes is essential to capitalize on the increase in the sequencing and analysis of model genomes. The Entrez Gene database has been developed to supply key connections between maps, sequences, expression profiles, structure, function, homology data, and the scientific literature. Unique identifiers are assigned to genes with defining sequence, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are tracked, and functional information is added when available. Access Entrez Gene from the Entrez Home Page or directly at:

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

The Entrez Gene help document provides tips to ease the transition for LocusLink users to the current Entrez Gene database.

The default display format for Entrez Gene is the graphics display shown in Figure 1 for BMP7, which resembles the traditional view of a LocusLink record. The array of colored boxes at the head of LocusLink reports that provide links to gene-related resources is replaced by the "Links" menu in Gene, which includes additional links, such as those to Books, GEO, UniSTS, and Taxonomy. The Gene Transcripts and Products section is provided when a gene has been annotated on a genomic Reference Sequence

continued on page 6

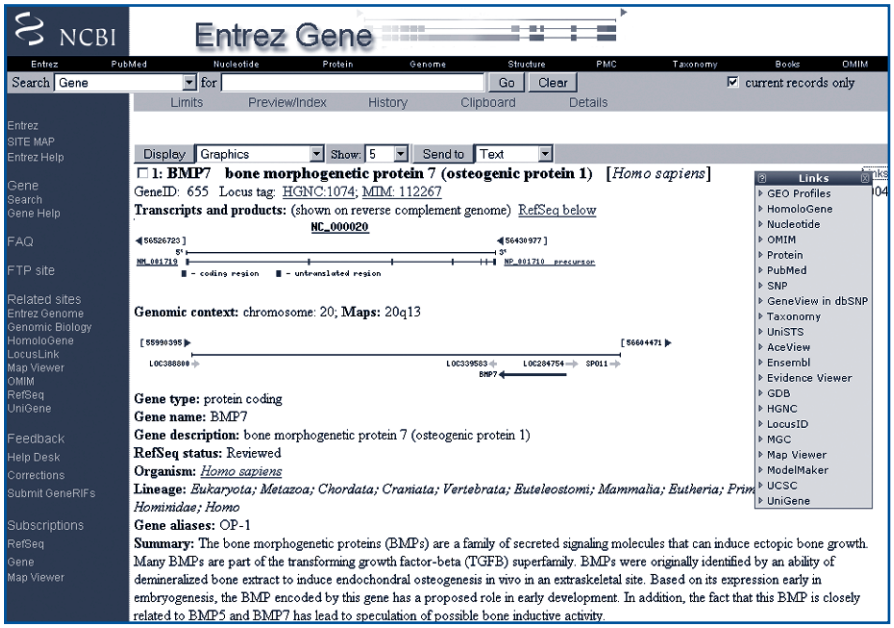

Figure 1. Entrez Gene display for human BMP7 showing links to over 20 related resources in the "Links" pulldown menu.

Cancer Chromosomes: a New Entrez Database

Three databases, the NCI/NCBI SKY (Spectral Karyotyping)/M-FISH (Multiplex-FISH) and CGH (Comparative Genomic Hybridization) Database, the NCI Mitelman Database of Chromosome Aberrations in Cancer, and the NCI Recurrent Chromosome Aberrations in Cancer databases are now integrated into NCBI's Entrez system as the "Cancer Chromosomes" database. Cancer Chromosomes supports searches for cytogenetic, clinical, or reference information using the flexible Entrez search and retrieval sys-

continued on page 3

In this issue

NCBI News

NCBI News is distributed four times a year. We welcome communication from users of NCBI databases and software and invite suggestions for articles in future issues. Send correspondence to *NCBI News* at the address below. To subscribe to NCBI News, send your name and address to either the street or E-mail address below.

NCBI News
National Library of Medicine
Bldg. 38A, Room 3S-308
8600 Rockville Pike
Bethesda, MD 20894
Phone: (301) 496-2475
Fax: (301) 480-9241
E-mail: info@ncbi.nlm.nih.gov

Editors
Dennis Benson
David Wheeler

Contributors
Susan Dombrowski
Scott McGinnis
Tao Tao

Writers
Vyvy Pham
David Wheeler

Editing and Production
Robert Yates

Graphic Design
Robert Yates

In 1988, Congress established the National Center for Biotechnology Information as part of the National Library of Medicine; its charge is to create information systems for molecular biology and genetics data and perform research in computational molecular biology.

NIH Publication No. 04-3272

ISSN 1060-8788
ISSN 1098-8408 (Online Version)

HomoloGene: An Entrez Database with a New Look

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes. The genomes represented in the recent Build 36 of HomoloGene include *H. sapiens, M.musculus, R.norvegicus , D. melanogaster , A. gambiae, C. elegans , S. pombe, S. cerevisiae , N. crassa, M. grisea, A. thaliana,* and *P. falciparum.*

NCBI has adopted a new HomoloGene build procedure which is guided by the taxonomic tree, relies on conserved gene order and measures of DNA similarity among closely related species, while making use of protein similarity for more distantly related organisms. The new computational procedure greatly increases the reliability of the computed homologous gene sets and the resulting HomoloGene entries now include paralogs in addition to orthologs. For more details or to search the database, see the Homologene home page at:

www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene

New Search Strategies Supported

Because HomoloGene is now an Entrez database, it can be queried using an assortment of fielded terms combined with boolean operators. Among the fields unique to HomoloGene is the "Ancestor" field which refers to the taxonomic group of the last common ancestor of the species represented in a HomoloGene entry. Using the "Ancestor" field it is possible to limit a search to genes conserved in one of 9 ancestral groups: *Sordariomycetes* (147,550 entries), *Eukaryota* (2,759 entries), *Fungi/Metazoa* (33,154 entries), *Bilateria* (33,213 entries), *Coelomata* (33,316 entries), *Mammalia* (9,172 entries), *Ascomycota* (1,083 entries), *Insecta* (1,689 entries), *Rodentia* (1,587 entries).

New Views of the Data

HomoloGene reports include homology and phenotype information drawn from Online Mendelian Inheritance in Man (OMIM), Mouse Genome Informatics (MGI), Zebrafish Information Network (ZFIN), *Saccharomyces* Genome Database (SGD), Clusters of Orthologous Groups (COG), and FlyBase. A "Pairwise Scores" display gives a table of pairwise statistics for members of a Homologene group that includes percent amino acid and nucleotide identities, the Jukes-Cantor genetic distance parameter, D, the ratio of non-synonymous to synonymous amino acid substitutions (Ka/Ks) for predicted proteins, and the ratio of nucleotide identities within non-coding regions of the transcript to those within coding regions (Knr/Knc).

—DW

New HomoloGene FTP File Formats

The Homologene data is available by FTP where the data for each build is contained in two files; "homologene.data" and "homologene.xml.gz". Follow the "FTP site" link in the sidebar on the Homologene home page to download the files.

homologene.data

homologene.data is a tab delimited file containing, from left to right:

- HomoloGene group id
- Taxonomy ID
- gene ID
- gene symbol
- geninfo identifier (gi) of the protein product of the gene
- accession number of the protein product of the gene

homologene.xml.gz

homologene.xml.gz is a compressed file that contains a complete XML version of the HomoloGene build and includes the information available on the public webpage. The Homologene XML DTD is available in the archive "homologene.dtd.tar" at the top level of the ftp site.

The old HomoloGene FTP files of the formats used in "hmlg.ftp" and "hmlg.trip.ftp" will be discontinued after a transition period. During the transition, a new set of codes, reflecting the new build procedure, will be used in these files to indicate the nature of the evidence for homology: b - reciprocal best, B - reciprocal best in a self-consistent triplet, m - similarity between sequences that do not give reciprocal best hits.

Cancer Chromosomes
continued from page 1

tem. Search tips are provided in the Help document at:

www.ncbi.nlm.nih.gov/entrez/query/SkyCgh/help.html

Search "Cancer Chromosomes" from the database pulldown menu on the NCBI home page or navigate to the "Cancer Chromosomes" page for advanced searches via the link on the Entrez home page at:

www.ncbi.nlm.nih.gov/Entrez

Three search formats are offered on the Entrez Chromosomes home page: a conventional Entrez Query, a Quick/Simple Search, and an Advanced Search. The Entrez Query is performed using the search box at the top of the page, and, as with other Entrez databases, searches may be combined using term limits and Boolean expressions. The Simple Search, available via a link in the sidebar on the Cancer Chromosomes page, offers a set of menus from which one may select search terms to indicate a disease site or diagnosis. These terms can be combined with specifications for a particular chromosomal location and anomaly. The Advanced Search form, also linked from the sidebar, is arranged similarly. This form contains three main sections, labeled Cytogenetic, Clinical, and Reference, which offer a combination of forms and menus of search terms to help in the construction of complex queries. Diagnostic terms vary among databases; the SKY/M-FISH database uses ICD-3-O terms, whereas the Mitelman and Recurrent databases use a different system. The menus include all ICD-O-3 terms entered into the database to date and all terms used in the Mitelman and Recurrent databases. Descriptions of the sections and terms indexed are given in the Help document.

Searches based on case information, such as diagnosis and disease site, return a "case-based report" that lists all cases matching the query terms. Searches based on underlying cytogenetic features are displayed as a "clone/cell report" in which each clone or cell-line is listed separately.

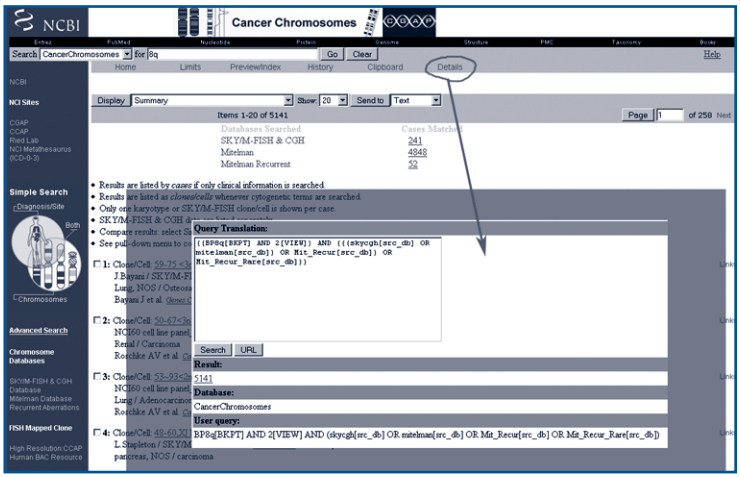

Figure 1. Results of an Entrez Cancer Chromosomes search for records using the query "8q". The number of cases from each database is given at the top of the result summary. Clicking on the Details link shows how the query was interpreted by Entrez.

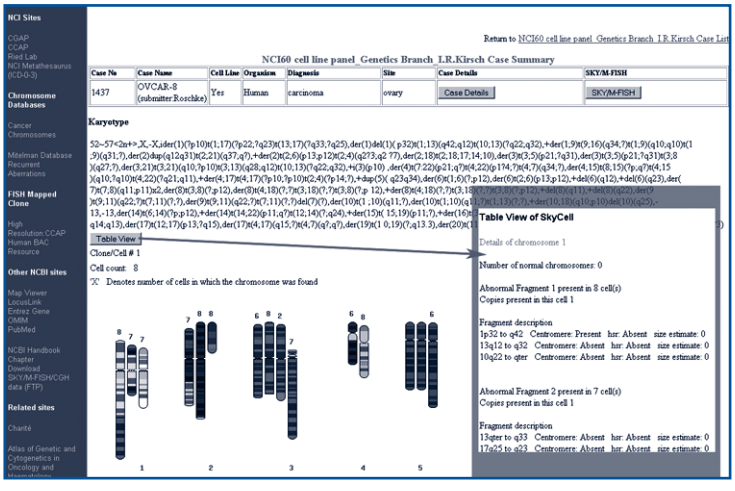

Figure 2. Detailed display of Case report #1437, with written karyotype, case summary, graphic display of colored ideogram, and chromosome abnormalities in tabular format.

The number of cases or cells/clones found for each search is displayed at the top of the results page, broken down into three totals: the total matches found in the SKY/M-FISH & CGH Database, the total matches found in the Mitelman database, and the total matches from the Mitelman Recurrent Database.

From the results list, users can access the pull-down menu and display a variety of features, including the corresponding literature from PubMed, the results as a list of UI (unique identifier) numbers, or view related reports based on common cytogenetic or diagnostic features. Users can also view Similarity reports, which show terms common to a group or records within several term categories such as diagnosis/site and cytogenetic abnormalities (including CGH) among the selected cases or clones/cells. Term co-occurences are listed at several levels: common to all cases, common to 50%-90% of cases, and common to less than 50% of cases. The common term or abnormality is shown in the left column and the number of affected cases is shown in the right column. The cytogenetic abnormalities are shown at all levels of resolution. Select 'Similarity Report (High Resolution Only)' to see similarities at a high level of resolution such as chromosome band.

The results of a sample query for records dealing with breakpoints of the chromosomal band 8q is shown in Figure 1.

Links in the search results summary lead to full reports such as the Case report #1437 shown in Figure 2. Display buttons provide access to additional views of the data such as chromosomal diagrams or a table view. Users can also access the case details or link to related resources from the original search summary by way of the Entrez Links menu.

—VP

BLAST Link (BLink) to Protein Alignments and Structures

Pre-computed sequence alignments, generated from routine all-against-all BLAST comparisons performed at NCBI, are available for each protein record in Entrez. The best 200 of these alignments can be displayed by clicking on the "BLink" link in the upper right-hand corner of Entrez protein reports. The BLink report for human MLH1 is shown in Figure 1. The report begins with a description of the query sequence, the sequence IDs of other entries in Entrez with identical sequences, and a set of controls, described below, used to customize the display. The alignments presented in the lower section of the report are depicted graphically and color-coded on the basis of the taxonomic origin of the aligned sequence. Each alignment is followed by its BLAST score, linked to a detailed alignment view, the accession number of the aligned sequence, linked to Entrez, and the GI number of the aligned sequence, linked to its own BLink report.

Customizing the Display

BLink reports can be customized using a combination of format buttons, taxonomic restriction controls, and a source database pulldown menu. Taxonomic and source database restrictions take effect when the "Select" button is pressed. Alignments may be sorted on the basis of sequence similarity to the query, the default, or by the taxonomic proximity of the source organisms using a link in the formatting area.

Six format buttons are used to select the display mode of the BLink report. The 'Best Hits' button displays a single line for each organism represented in the results, showing the alignment of the best hit in the organism group and a link, in the "N" column, to a Blink report limited to the group. A portion of a "Best Hits" display is shown in Figure 2.

The "Common Tree" button allows for the selective display of alignments to sequences arising from specific branches of the taxonomic tree. The related "Taxonomy Report" button lists the BLink results as a BLAST Taxonomy Report.

To limit the view to those sequences derived from structure records, press the "3-D Structures" button. In the "3-D Structures" display, shown in Figure 3, the dots are links to Conserved Domain and Cn3D structure displays.

The "CDD Search" button does not format the BLink report, but instead links to a precomputed conserved domain display for the query sequence.

The "GI" button links to the Entrez display of the protein sequences whose alignments are shown in the BLink report.

Taxonomic and Database Restrictions

Sequences derived from a taxonomic group may be selectively removed from the display by clicking on any of the color-coded taxon links; an "X" across the sequence count for the group indicates that the group will be removed from the display when the "Select" button is pressed. The BLink report can also be customized using the "Keep only" pulldown menu to limit the display to entries included in databases such as RefSeq, Protein Data Bank, SwissProt, COGs, and NCBI Complete Genomes.

Use BLink for Quick Insights into Protein Function

Because BLink reports are pre-computed it is possible to rapidly view a BLAST alignment without having to generate it. The graphical display of the aligned sequences provides a clear view of the distribution of conserved sequence blocks across taxonomic groups as an aid to understanding evolutionary and functional relationships. Added insight into protein function is provided by the CDD display of multiple sequence alignments for functional domains allowing one to evaluate position specific sequence conservation in the context of the biological function of the query protein. The "3D Structure" display provides a quick way to determine the availability of 3D structures to serve as modeling templates to use for further study.

—TT

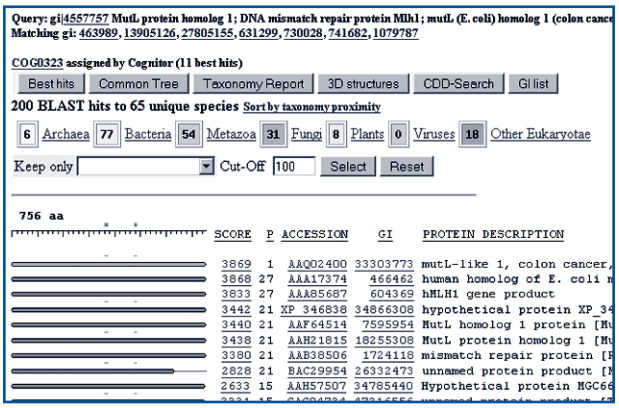

Figure 1. BLink report for the human MLH1protein. Format controls are located at the top of the report with alignments, colored by the taxonomic origin of the sequence match, given in the lower section.

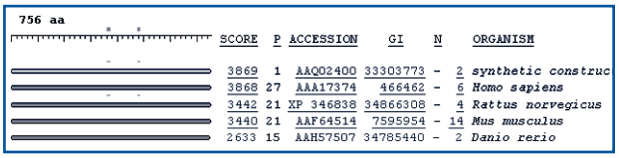

Figure 2. "Best Hits" Blink report format indicating two alignments to a sequence from a synthetic construct and six alignments to sequences from Homo sapiens. A graphical depiction of the best alignment in each organism group is shown on the left.

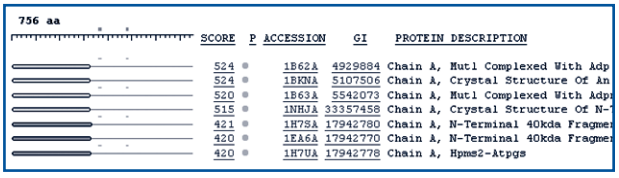

Figure 3. BLink report limited to sequences derived from structures using the "3-D Structures" button.

Debut of the HCT Database and Anthropology/Allele Frequencies in dbMHC

The International Histocompatibility Working Group (IHWG) in Hematopoietic Cell Transplantation (HCT) is a worldwide scientific collaborative effort to support the use of information on the properties of the HLA barrier to allogeneic transplantation to improve the safety, efficacy and availability of HCT. The IHWG HCT studies are designed to determine whether complete allele matching for HLA-A, B, C DRB1, DQB1 and DPB1 is necessary for successful transplantation. Data generated from the study are anticipated to offer new approaches to the selection of suitable donors for HCT.

The IHWG and the NCBI at the National Institutes of Health, have collaborated to create a public database, dbMHC, to store genotype and clinical data, including up-to-date information on matching and transplant outcomes, and to provide online tools for data analysis. The new database contains anonymous data for selected unrelated donor transplants performed worldwide for the treatment of both malignant and non-malignant blood disorders. More information and a link to a list of contributors to dbMHC is found on the home page at:

www.ncbi.nih.gov/mhc/ihwg.fcgi?cmd=page&page=HCTintro

A powerful way to view the data in dbMHC is to compute Kaplan-Meier survival plots using the HCT online plotting tool. Many default parameters for the plots can be adjusted in order to customize the data displayed. Help and FAQ documents are available for most plots. As an example, consider the Kaplan-Meier survival plot shown in Figure 1. The plot uses tick marks for censored data and shading to indicate the confidence intervals. To produce the plot, the Advanced Query Form is initialized with default parameters and one curve is plotted which allows no mismatches at HLA A, B, C, DRB1 and DQB1 loci and ignores HLA DPB1 locus. The parameters which affect the entire plot can be specified and a summary is displayed for each curve. A user may add a curve, modify a curve, delete a curve, view the plot, view plot data, view individual data, save the curve parameters, or restore saved parameters. When adding or modifying a curve, another form is displayed for entering the parameters for the curve and upon completion the user will return to the original page to view the updated selections.

Another important resource stemming from the efforts of the IHWG is the Antrhopology/Allele Frequencies databank, created in an effort to determine HLA class I and class II allele and haplotype frequencies in various human populations. Studies of allelic diversity in different populations can shed light on the evolution of HLA polymorphism as well as on the evolution and migration of human populations. In a clinical context, knowledge of the allele frequency distributions in various populations is critical to the strategy of establishing and searching bone marrow donor registries as well as for studies of HLA-associated disease susceptibility.

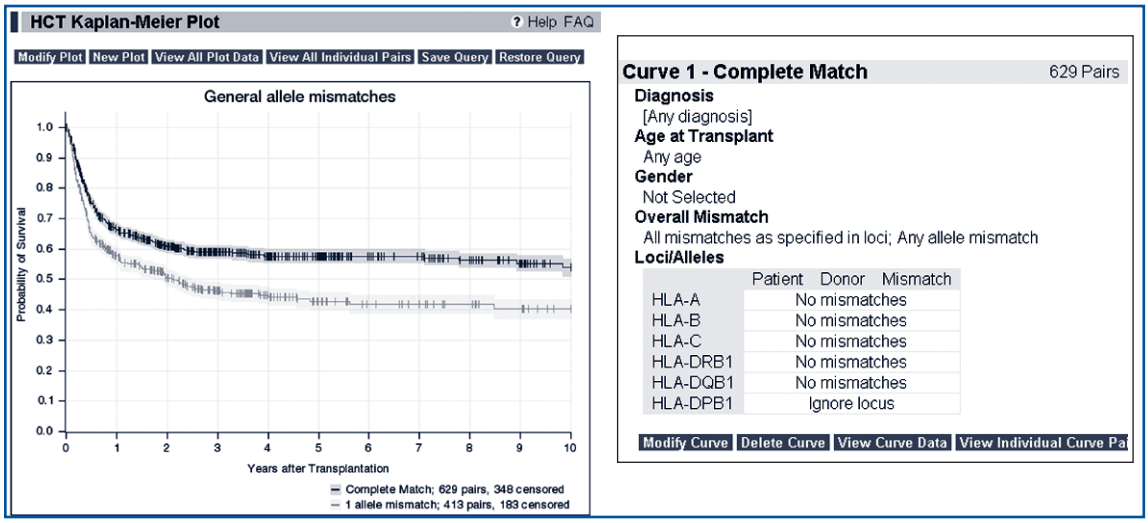

Figure 1. Kaplan-Meier survival plot generated on the web using analysis tools available from dbMHC. A summary of the parameters used to create the plot is displayed to the right. The plot reveals a significant effect on transplant survival times of a one allele mismatch.

Users of the resource can choose to view allelic frequencies found in individuals from certain regions of the world, or view data submitted by a particular group. Users can also specify the loci to be displayed in the output table. Additional information about the project from the links to project overview and data contributors is found at:

www.ncbi.nlm.nih.gov/mhc/ihwg.fcgi?ID=9&cmd=PRJOV

The HCT and Anthropology/Allele Frequencies resources are available from the dbMHC home page via links from the blue sidebar menu.

Questions and comments can be addressed to the NCBI Service Desk at:

info@ncbi.nlm.nih.gov

—VP

LocusLink to Entrez Gene

continued from page 1

(RefSeq) and intron, exon, and coding region information is available with genomic coordinates. Each accession given in this section is a link to a menu allowing the display of the sequence in several formats. Protein accessions provide menu options to navigate to BLink, CDD, or COG displays. This section is equivalent to the RNA-Genomic alignment available from the graphic at the top of a LocusLink entry. In the case of the Gene record for BMP7, NC_000020 is the accession number of the genomic contig that contains the gene. Clicking on the "NC_000020" link brings up a menu used to select one of several displays of the contig within the genomic range of the gene BMP7.

The Entrez Gene record's "General Gene Information" section summarizes information contained in LocusLink's "Function", "Relationships" and "Map Information" sections. This section includes several categories of information, such as Gene Ontology (GO), Homology, Phenotypes, Markers, Pathways and Relationships.

The remaining sections of an Entrez Gene record-"NCBI Reference Sequences", "Related Sequences", and "Additional Links"-are equivalent to the corresponding entries in the LocusLink report. The first section lists gene-specific NCBI RefSeqs, provides links to the appropriate Entrez sequence database, and gives descriptions of each transcript variant, the accession numbers of sequences used to support the RefSeqs, and a listing of conserved domains found in the encoded proteins. The "Related Sequences" section lists the nucleotide and protein accessions of sequences that are related to the gene, and provides links to the sequence records in Entrez. The "Additional Links" section provides a printable view of a subset of links to information both within and external to NCBI. Some of these links overlap those included in the Links menu. The intent of this section is to provide a printable report of, for example, MIM numbers, UniGene cluster numbers, and family-specific Web sites.

LocusLink	Gene	Comments
Table of Contents	Not retained	
Alphabetic lists	Not retained	
Gene diagram	Transcripts and Products	Gene adds the function of Genomic context, to allow a quick view of nearby genes and links to their report pages.
Link to Evidence Viewer from Gene diagram	Evidence Viewer link in Links menu	The option to first see only the diagram of the alignment is not retained.
Button Links	Links menu	On the Gene Graphic/Default display, the number of links may be greater than in LocusLink.
Title bar with links to nomenclature source.	Initial text, with link to nomenclature source via LocusTag.	Links from LocusTag values may connect to an external database where official nomenclature has not been assigned.
	Overview Section	
RefSeq Summary	Summary	not changed
Locus type	Combination of Gene type and evidence type (under development)	The text values are not equivalent. LocusLink's Locus type values are being subdivided into Gene type and Evidence type categories.
Protein names	General protein information	not changed
Alternate symbols	Gene aliases	not changed
	Relationships Section	
Homology data	Links menu; HomoloGene	What is printed in LocusLink is still printed in Gene.
Related models	Related	not changed; limited to genomes being annotated by NCBI's pipeline.
	Function Section	
GeneRIFS	GeneRIFs	Not in a function section; indented under Bibliography
GO annotation	General gene information: GeneOntology	organization changed, but not content
Phenotype	General gene information: Phenotypes	organization changed, but not content
	Map Section	
Chromosome	Genomic context	not changed
Associated markers	General gene information: Sequence Tagged Site (Markers)	Entrez Gene added display of alternate marker names
	Sequence Section	
	RefSeq Subsection	
Category	RefSeq status	not changed
GenBank® source	Source sequence	not changed
Domain matches	Domains	CDD link also attached to the protein accession in **Transcripts and Products** and in Links menu.
BL (BLink)	BLink link attached to the protein accession in the Transcripts and Products section	The function was not changed, but the placement and visibility are different.
Variant name	After the protein accession in the Transcripts and Products section	content not changed
	Annotation Subsection	
Genomic contig	Transcripts and Products	function retained
gb: Link to gene-specific subsequence	GENBANK view from source NC, NT, or NW accession	function retained as GENBANK option from the genomic accession-based menu.
sv: Link to graphic display of gene-specific subsequence	GRAPHICS view from source NC, NT, or NW accession	function retained as GRAPHIC option from the genomic accession-based menu.
mv: Map Viewer	Map Viewer in Links menu	function retained
ev: Evidence Viewer	Evidence Viewer in Links menu	function retained
mm: Model Maker	Model Maker in Links menu	function retained
strain or haplotype	not retained	
	Related Sequences Subsection	
Accessions, type, and strain data	Related Sequences	content not changed
BL (BLink) from protein accessions	not retained	
Additional Link	Additional Links	

Table 1. LocusLink to Gene feature transition chart. The help documentation covers the conversion of the master LocusLink FTP file, "LL_tmpl", to the Entrezgene.asn format. The Entrezgene.asn data will be available on the Gene FTP site in the near future.

Entrez Gene can be considered as the successor to LocusLink, but Gene improves on LocusLink by providing coverage of more NCBI reference genomes, by providing additional display formats, and by its integration with other databases within NCBI's Entrez system. Users can query Gene via the powerful query features of Entrez, using Boolean operators, filters, and field limiters, such as accession number, gene name, protein name, disease/phenotype, and map location. Users can search for records in Entrez Gene using any of the search strategies in the shaded box.

(human [organism] OR mouse [organism] OR rat [organism]) AND bmp7

human [organism] AND (bmp7 OR bmp3)

human [orgn] AND (bmp7 [title] OR bmp3 [title])

Like other Entrez databases, Gene offers a number of display formats beyond the default "Graphical" format. Additional formats include an XML format and a "Gene Table" view, providing access to the sequences of each of the gene's exons and introns.

Entrez Gene is also accessible using the Entrez Programming Utilities (E-utilities), that provide access to Entrez from application programs and scripts.

Users interested in subscribing to email announcements of new Entrez Gene features are welcome to join the Gene-announce mailing list at:

www.ncbi.nlm.nih.gov/mailman/listinfo/gene-announce

—VP

350 kb Sequence Length Limit Removed by Sequence Database Collaboration

In 1995, the International Nucleotide Sequence Database Collaborators (GenBank, DDBJ, and EMBL) agreed to a 350 KB limit on the size of database sequence records in order to maintain compatibility with existing molecular biology software that was not able to work with large sequences.

At this time, a new GenBank division was created called "CON" for contig. The records in the CON division contain the instructions for the assembly of full-length contigs from the sequence data of multiple GenBank records. Although CON division records contain no sequence data, the assembly information they provide makes it possible for NCBI's Entrez search and retrieval system to show complete genomic sequences by dynamically assembling the data for display. Using the information in CON division records, FTP files are also regularly created by NCBI for download that contain megabase-scale genomic sequences as single FASTA files.

By 1998, GenBank, DDBJ, and EMBL were routinely accepting submissions from large scale sequencing projects of draft sequences, such as phase 1 and phase 2 high-throughput genomic sequences (HTGS), that were longer than 350 KB. To avoid breaking a huge amount of draft sequence into 350 KB chunks, the database collaborators agreed to relax the 350 KB limit in these cases. The 350 KB limit was also relaxed for assemblies of Whole Genome Shotgun (WGS) project data and for large eukaryotic genes.

Removal of 350 kb Limit

In 2003, the Database Collaborators agreed to remove the 350 KB limit for all sequences as of June 2004, since the increased ability of molecular biology software to analyze long sequences quickly has rendered the limit on sequence length unnecessary. To help software developers prepare for the change, some sample records with large sequences have been made available for testing:

ftp.ncbi.nih.gov/genbank/LargeSeqs

An example of the effect of the removal of the 350 KB limit on GenBank records may be seen in the case of accession U00096, the *Escherichia coli* K-12 MG1655 complete genome sequence. Under the 350 KB limit, this accession number refered to a contig record giving a list of short sequences that can be assembled to create the complete genome. With the removal of the 350 KB limit the accession now refers to the complete contiguous sequence for *Escherichia coli* K-12. The accessions for all 400 parts will appear as secondary accessions. The CON division will remain as a GenBank division to represent sequences which by their nature are assembles; Ex. genome scaffold records.

The effect of the changes on the NCBI GenBank FTP files and the BLAST database files available for download is expected to be minimal. As sequences become secondary to primary records, the overall size of the databases should not change drastically. However, the number of megabased-sized records will increase, therefore NCBI recommends that software be tested with the example large sequence records, mentioned above.

—SM

New Eukaryotic Genomes at NCBI

NCBI has created new Map Viewer displays for several organisms, including the honey bee, cat, and the fungi *Eremothecium gossypii* and *Encephalitozoon cuniculi.* In addition, new genome guides are available for the honey bee, cat, chicken, frog and sea urchin.

The Map Viewer display for *Apis mellifera*, the honey bee, includes a variety of sequence-based maps, such as maps for contigs, UniGene clusters, genes, and transcripts, as well as the Solignac microsatellite-based linkage map. The current honey bee genome build, Amel_1.1, is a composite of whole genome shotgun sequence and BAC sequence from clones isolated by a clone-array pooled strategy.

The *Apis mellifera* whole genome shotgun (WGS) project has been assigned the project accession AADG00000000 and the Amel_1.1 assembly displayed and annotated in Map Viewer, is comprised of accessions AADG02000001-AADG020-30074.

The NCBI Map Viewer for *Felis catus* presents an RH map of the cat genome. The map contains 1,126 markers, including microsatellite-based markers and coding loci. An integrated genetic map is also included.[1,2]

The sequencing, annotation, and analysis of the *Eremothecium gossypii* genome is described in Dietrich et al.[3] Its chromosomes have been assigned RefSeq accessions NC_005782 to NC_005789. The sequenced-based maps- contig, gene, and transcript maps- are provided in the Map Viewer for this filamentous fungus.

The sequencing, annotation, and analysis of the *Encephalitozoon cuniculi* genome is described in Katinka et al.[4] Its chromosomes have been assigned RefSeq accessions NC_003242, and NC_003229 to NC_003238. The contig and gene maps are available in the Map Viewer for this microsporidium.

New Genome Guide pages, created by NCBI in cooperation with the genomic research communities to provide links to an array of genome-

continued on page 11

Environmental Samples Make Big Splash

The technology of Whole Genome Shotgun (WGS) sequencing is now being applied to quickly assemble large sets of genomic sequences taken from organisms inhabiting a particular ecological niche. Sequence data collected in this manner provides a snapshot of the genetic diversity existing at a particular locale and is especially important in providing data for organisms which are difficult or impossible to culture in the laboratory. Recently, The Institute for Biological Energy Alternatives sampled water from the Sargasso Sea, one of the most well-characterized regions of the world's oceans.[1] The larger of two sets of samples collected produced over 1.3 gigabases of sequence in the form 1.66 million WGS reads. These reads were assembled into contigs containing about 1 gigabase of non-redundant sequence. In addition, over 1 million protein sequences were derived from the annotation of open reading frames on the genomic sequences. Contigs constructed from the WGS reads and the remaining single reads have been deposited in the WGS division of GenBank, under the project accession number AACY01000000. Scaffolds assembled from these contigs are available within the accession ranges CH004436-CH004736, and CH004737-CH236877. The raw sequencing data is available in the Trace Archive.

The Sargasso Sea dataset along with other environmental sample datasets, such as sequences from an acid mine drainage biofilm submitted by the DOE Joint Genome Institute,[2] can be queried using the new "Environmental Samples" BLAST page at:

www.ncbi.nlm.nih.gov/BLAST/Genome/EnvirSamplesBlast.html

continued on page 12

New HIV Protein-Interaction Database

Documenting the interaction of human immunodeficiency virus type 1 (HIV-1) proteins with those of the host cell is crucial to our understanding of the processes of HIV-1 replication and pathogenesis. To meet this need, the Division of Acquired Immunodeficiency Syndrome (DAIDS) of the National Institute of Allergy and Infectious Diseases (NIAID), in collaboration with the Southern Research Institute and NCBI, has begun compiling a comprehensive "HIV Protein-Interaction Database" to provide a concise summary of documented interactions between HIV-1 proteins and host cell proteins, other HIV-1 proteins, or proteins from disease organisms associated with HIV or AIDS. For each documented protein-protein interaction the following information is collected, if available:

- Protein Reference Sequence accession numbers.
- Entrez Gene ID numbers.
- The Amino acids from each protein that are known to be involved in the interaction.
- A Brief description of the protein-protein interaction.
- Keywords to support searching for interactions.
- PubMed identification numbers for all journal articles describing the interaction.

The HIV Protein-Interaction Database may be searched through an online interface at:

www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/index.html

Clicking on a link for an HIV-1 protein in the "Reports and downloads" section of the page displays its interaction report.

Interaction reports for HIV-1 proteins are displayed in a 4-column format beginning from the left with the HIV-1 protein name, linked to the LocusLink report for the gene, and continuing with phrases taken from the literature indicating, in the second column, a type of interaction, and, in the third column, a description of the interaction partner. The fourth column displays links to LocusLink and Entrez Gene reports for the interaction partner. Interaction reports can be filtered in a number of ways by phrases appearing in the reports using pull down phrase lists. Full or filtered reports may be downloaded as text in a tab-delimited format that includes fields for the name of the subject HIV-1 protein, its RefSeq accession number, the interaction phrase, the description of the interaction partner, the partner's RefSeq accession number, the title of the publication reporting the interaction, and its PubMed identifier. For example, the first line in the tab-delimited file produced by filtering the interaction report for the "tat" protein by the interaction "upregulates" is shown in Figure 1.

Interaction reports are currently available for 7 of the 9 proteins produced by HIV-1 including, gag, pol, rev, tat, vif, vpr, and vpu. Reports for the remaining proteins nef and env, will be completed soon. All protein-protein interactions documented in the HIV Protein-Interaction Database are listed in Entrez Gene reports in the "HIV-1 protein interactions" section.

—DW

Tat, p14 NP_057853 upregulates B-cell lymphoma 6 protein NP_001697 HIV-1
Tat upregulates the expression of BCL-6 in Kaposi's sarcoma cells 11994280

Figure 1. First line in the tab-delimited file produced by filtering the interaction report for the "tat" protein by the interaction "upregulates".

e-PCR and Reverse e-PCR: Greater Sensitivity, More Options

Electronic PCR (e-PCR) is used to identify sequence landmarks called Sequence Tagged Sites (STSs) within a nucleotide sequence. Electronic-PCR works by looking for matches to STS primer pairs with the orientation and spacing required to produce an amplicon of the expected size.

Two types of e-PCR can now be performed from the e-PCR home page: the original, Forward e-PCR, and a new application, Reverse e-PCR. Forward e-PCR is used to determine if a user-supplied nucleotide sequence contains any known STS. Queries are made against the markers in NCBIs UniSTS database, a public collection of PCR primer pairs used in mapping and other types of genome analysis for a wide range of organisms. UniSTS contains over 270,000 markers and includes data from the STS division of GenBank, The Radiation Hybrid Database, The Genome Database, Mouse Genome Informatics, the Rat Genome Database, Zebrafish Information Network and PubMed Central. If an STS is found using the online version of Forward e-PCR, the chromosomal location and a link to UniSTS are provided. Mapped markers can be displayed from UniSTS reports using the MapViewer and viewed in the context of the other available genomic data.

To improve the sensitivity of a Forward e-PCR search, there is now an option to search using discontiguous, or imperfect, matches between the query sequence and the STSs in UniSTS. To increase the probability of finding STSs that may have been missed with the contiguous word option, the size of the segment to be matched, called the word size, number of allowable mismatches, and number of permissible gaps can be adjusted. The size of the STS can also be adjusted in order to allow for deviations which may arise from the amplification of a region in the genome that shows length polymorphism.

Reverse e-PCR can be used to estimate the genomic binding site, amplicon size and specificity for user-supplied primer pairs. The Reverse e-PCR query page can accept up to 20 primer pairs or STS identifiers as input for searches against organism-specific databases. Primer pairs or STS identifiers can be searched against the genomic and transcript databases of *Anopheles gambiae, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Homo sapiens, Mus musculus* and *Rattus norvegicus*.

Interested in seeing how the forward and reverse e-PCR tools work? Try your hand at some of the e-PCR examples that are provided on the two search pages found at:

www.ncbi.nlm.nih.gov/sutils/e-pcr

For those who need to perform large batches of searches or who need to search a custom database, a standalone version of e-PCR has been developed for the Windows, Linux and Unix operating systems. These binaries, along with the source code for compiling e-PCR on other operating systems, are available via FTP at:

ftp.ncbi.nlm.nih.gov/pub/schuler/e-PCR

—SD

New Organisms in UniGene

UniGene now covers 45 animals and plants and can be searched using the Entrez search system where it is linked to nucleotide records. Recent additions to UniGene include:

Canis familiaris (dog) with 15,281 transcript sequences in 4,544 clusters, *Helianthus annuus* (sunflower), with 21,155 transcript sequences in 1,904 clusters, *Salmo salar* (Atlantic salmon), with 23,111 sequences in 1,050 clusters, *Bombyx mori* (domestic silkworm), with 60,481 sequences in 2,050 clusters, *Apis mellifera* (honey bee), with 14,386 transcripts in 5,190 clusters, *Lotus corniculatus* (Birdsfoot trefoil), with 59,678 transcripts in 8,158 clusters, *Physcomitrella patens* (physcomitrella moss) with 83,948 transcripts in 6,950 clusters, *Lactuca sativa* (garden lettuce) with 53,347 transcripts in 9,656 clusters, *Malus x domestica* (Apple) with 34,733 transcripts in 3,996 clusters, *Hydra magnipapillata* with 69,049 transcripts in 5,362 clusters, *Populus tremula x Populus tremuloides* (aspen) with 20,820 transcripts in 2,707 clusters, and *Ovis aries* (sheep) with 2,852 transcripts in 1,164 clusters.

RefSeq Accession Numbers Get Longer as Rat Gets Last 6-digit Accession

Rattus norvegicus has scurried to the end of the possibilities offered by the six digit RefSeq accession format by making off with NCBI protein RefSeq accession NP_999999, required for its olfactory receptor Olr1386. To compensate, RefSeq accessions have now been extended to a length of 9 digits, e.g., NP_123456789. Preexisting accessions will not be changed and accession numbers in both the old and new extended formats, such as NM_000100 and NM_01000000, will coexist.

Are you wondering which organism got the first 9-digit RefSeq protein accession? That would be the energetic and inquisitive *Rattus norvegicus* again, for another of its very important olfactory receptors!

Slots available for FieldGuidePlus Training Course Onsite at NCBI

The popularity of the free NCBI training course for life scientists, A Field Guide to NCBI Resources, continues to grow, with over 40 courses presented throughout the United States in the first half of 2004. Portions of the Field Guide have also been presented as more detailed modules on specific tools and databases, including a 3D structures course, a gene expression course, and a BLAST course. In addition to the 2-day Field Guide, NCBI offers several 2-hour problem-centered Mini-Courses. This Spring, the NCBI conducted the first enhanced Field Guide, the FGPlus, at the National Library of Medicine, that combines the best of the standard Field Guide with the modular Field Guide courses and the Mini-Courses.

The FGPlus provides more detailed information, practical tips, and more extensive hands-on practice than the standard course. In addition, the FGPlus offers complete BLAST and structure modules, as well as a Mini-Course on disease genes. *Following the success of the Spring edition, NCBI will again offer the FGPlus course on August 24 and 25th, 2004 at the National Library of Medicine's Lister Hill auditorium.* For more details and registration information please visit the FGPlus page:

www.ncbi.nlm.nih.gov/Class/FieldGuide/FGPlus

or write to Peter Cooper (cooper@ncbi.nlm.nih.gov)

RefSeq Release 6 on FTP Site

RefSeq Release 6 is now available by anonymous FTP at:

ftp.ncbi.nih.gov/refseq/release

Release 6 includes genomic, transcript, and protein sequences available as of July 5, 2004 from 2,467 organisms. The number of RefSeq accessions in Release 6 and their combined lengths is given in the shaded box.

RefSeq releases are posted bimonthly and the next release is scheduled for September. Release notes documenting the scope and content of the release are provided at:

ftp.ncbi.nih.gov/refseq/release/release-notes

For more information, visit the NCBI RefSeq Web Site at:

www.ncbi.nih.gov/RefSeq

	# Accessions	# Basepairs/Residues
Genomic	68,592	8,263,102,565
RNA	247,639	433,269,151
Protein	1,050,975	365,446,682

Exponential Growth of GenBank Continues with Release 142

Over the past decade, the growth of GenBank has followed an exponential curve with a doubling time of between 12 and 15 months. As shown in Figure 1, the trend continues with release 142 for which close-of-data was June 16. In the eight week period between the close dates for GenBank releases 141.0 and 142.0, the non-WGS portion of GenBank grew by 1,335,978,783 base pairs and by 1,855,785 sequence records. The number of base pairs of sequence in release 142 for several organisms of interest is shown in Figure 2.

GenBank release 142 is available on the NCBI FTP site and at two mirror sites.

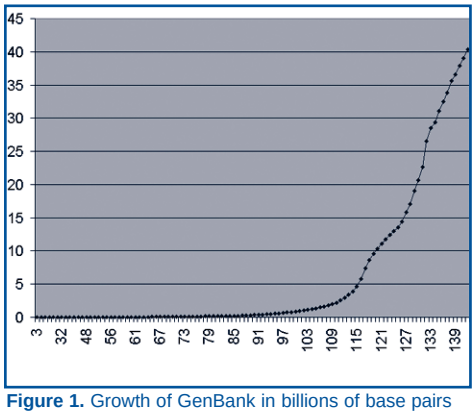

Figure 1. Growth of GenBank in billions of base pairs from release 3 in April of 1994 to the current release, 142.

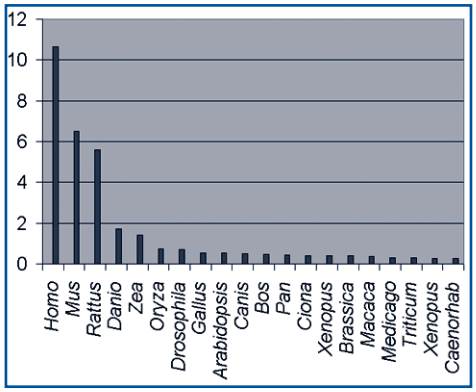

Figure 2. Billions of base pairs of sequence in GenBank release 142 for selected organisms.

Primary FTP site at NCBI:

ftp.ncbi.nih.gov/genbank

San Diego SuperComputer Center mirror:

genbank.sdsc.edu/pub

Indiana University mirror:

bio-mirror.net/biomirror/genbank

Uncompressed, the Release 142.0 flatfiles require approximately 136 gigabytes while the more compact ASN.1 version requires 119 gigabytes.

Entrez Tools is a 'Hot Spot'

Look for a new "Hot Spot" on the NCBI homepage called "Entrez Tools". Entrez Tools provides single-page access to a group of specialized Entrez resources. These resources include Batch Entrez, the Entrez Cubby, help with advanced Entrez searches, documentation for the Entrez utilities, and a guide to creating links to the interactive Entrez pages.

BLAST Lab

Using BLASTClust to Make Non-redundant Sequence Sets

BLASTClust is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. In the case of proteins, the blastp algorithm is used to compute the pairwise matches; in the case of nucleotide sequences, the Megablast algorithm is used.

In the simplest case, BLASTClust takes as input a file containing catenated FASTA-format sequences, each with a unique identifier at the start of the definition line. BLASTClust formats the input sequence to produce a temporary BLAST database, performs the clustering, and removes the database at completion. Hence, there is no need to run formatdb in advance to use BLASTClust. The output of BLASTClust consists of a file, one cluster to a line, of sequence identifiers separated by spaces. The clusters are sorted from the largest cluster to the smallest.

BLASTClust accepts a number of parameters that can be used to control the stringency of clustering including thresholds for score density, percent identity, and alignment length. The BLASTClust program has a number of applications, the simplest of which is to create a non-redundant set of sequences from a source database. As an example, one might have a library of a few thousand short nucleotide sequence reads and wish to replace these with a non-redundant set. To produce the non-redundant set, one might use:

```
blastclust -i infile -o outfile -p F -L .9 -b T -S 95
```

The sequences in "infile" will be clustered and the results will be written to "outfile". The input sequences are identified as nucleotide (-p F); "-p T", or protein, is the default. To register a pairwise match two sequences will need to be 95% identical (-S 95) over an area covering 90% of the length (-L .9) of each sequence (-b T) . Using "-b F" instead of "-b T" would enforce the alignment length threshold on only one member of a sequence pair. The parameter "S", used here to specify the percent identity, can also be used to specify, instead, a "score density." The latter is equivalent to the BLAST score divided by the alignment length. If "S" is given as a number between 0 and 3, it is interpreted as a score density threshold; otherwise it is interpreted as a percent identity threshold.

To create a stringent non-redundant protein sequence set, use the following command line:

```
blastclust -i infile -o outfile -p T -L 1 -b T -S 100
```

In this case, only sequences which are identical will be clustered together. The "blastclust.txt" file in the standalone BLAST package details the full range of BLASTClust parameters.

—DW

New Eurkaryotic Genomes
continued from page 7

specific resources, are available for honey bee, cat and chicken. These pages can be found at:

www.ncbi.nlm.nih.gov/genome/guide/bee

www.ncbi.nlm.nih.gov/genome/guide/cat

www.ncbi.nlm.nih.gov/genome/guide/chicken

A Map Viewer display for the chicken genome, *Gallus gallus*, will be available soon, but the sequences deposited into GenBank under the whole genome shotgun (WGS) sequencing project accession AADN00000000 (accessions AADN01000001-AADN01111864), are available now in Entrez and on the GenBank FTP site within the "WGS" directory.

—VP

[1]Menotti-Raymond et al, 2003a, PMID 14970716
[2]Menotti-Raymond et al, 2003b, PMID 12692169
[3]Dietrich et al. Science 304: 304-307, 2004 PMID 15001715
[4]Katinka et al. Nature 414:450-453, 2001 PMID 11719806

New Microbial Genomes in GenBank

Organism	GenBank | RefSeq Accession Numbers
Mycobacterium avium subsp. paratuberculosis str. k10	AE016958 | NC_002944
Bdellovibrio bacteriovorus	BX842601 | NC_005363
Treponema denticola ATCC 35405	AE017226 | NC_002967
Lactobacillus johnsonii NCC 533	AE017098 | NC_005362
Wolbachia endosymbiont of Drosophila melanogaster	AE017196 | NC_002978
Wolbachia endosymbiont of Drosophila melanogaster	pending | pending
Mycoplasma mobile 163K	AE017308 | NC_006908
Bacillus anthracis strain Ames 0581	
Picrophilus torridus	AE017261 | NC_005877

For more detailed information, see the online version of the Spring 2004 NCBI News, or use the GenBank or RefSeq Accession Number to query the Entrez "Genome" database using the query box on the NCBI Home Page.

Entrez Quiz

What is the total number of records in each of the 23 Entrez databases? If you've ever asked yourself this question, try the following query in the Entrez global search:

All[filter]

Each of the Entrez databases supports the "all" filter term which returns the total number of records indexed in a database. You can also collect this information using an E-utilities URL, e.g.:

eutils.ncbi.nlm.nih.gov/entrez/eutils/egquery.fcgi?term=all[filter]

The results of such a query as of mid-June are plotted in the Figure 1.

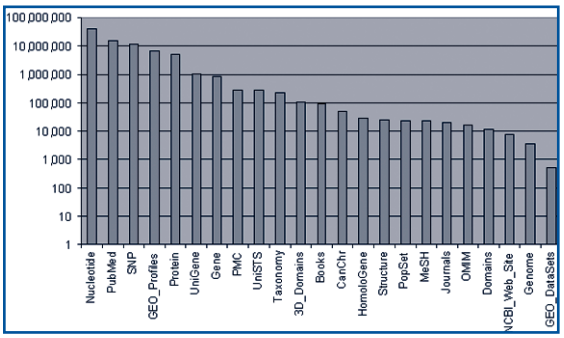

Figure 1. Number of records in the 23 Entrez databases as of mid June 2004. Numbers are plotted on a logarithmic scale.

Environmental Samples
continued from page 8

Environmental sample data can also be searched using two newly-created standard BLAST databases, "env_nt" or "env_nr" for nucleotide and protein sequences respectively. The environmental sample data contained within these two new databases is no longer contained within the "nt" or "nr" BLAST databases.

—VP

[1] Venter, J.C., et.al., Environmental Genome Shotgun Sequencing of the Sargasso Sea, *Science*, 2004 Apr 2;304(5667):66-74.

[2] Tyson, G.W., et.al., Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature*, 2004 Mar 4;428(6978):37-43.

Department of Health and Human Services
Public Health Service, National Institutes of Health
National Library of Medicine
National Center for Biotechnology Information
Bldg. 38A, Room 3S308
8600 Rockville Pike
Bethesda, Maryland 20894