SRA: Sequence Read Archive

Collection of sequence data from next-generation sequencing technology for different organisms

https://www.ncbi.nlm.nih.gov/sra/ & https://www.ncbi.nlm.nih.gov/Traces/sra/

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

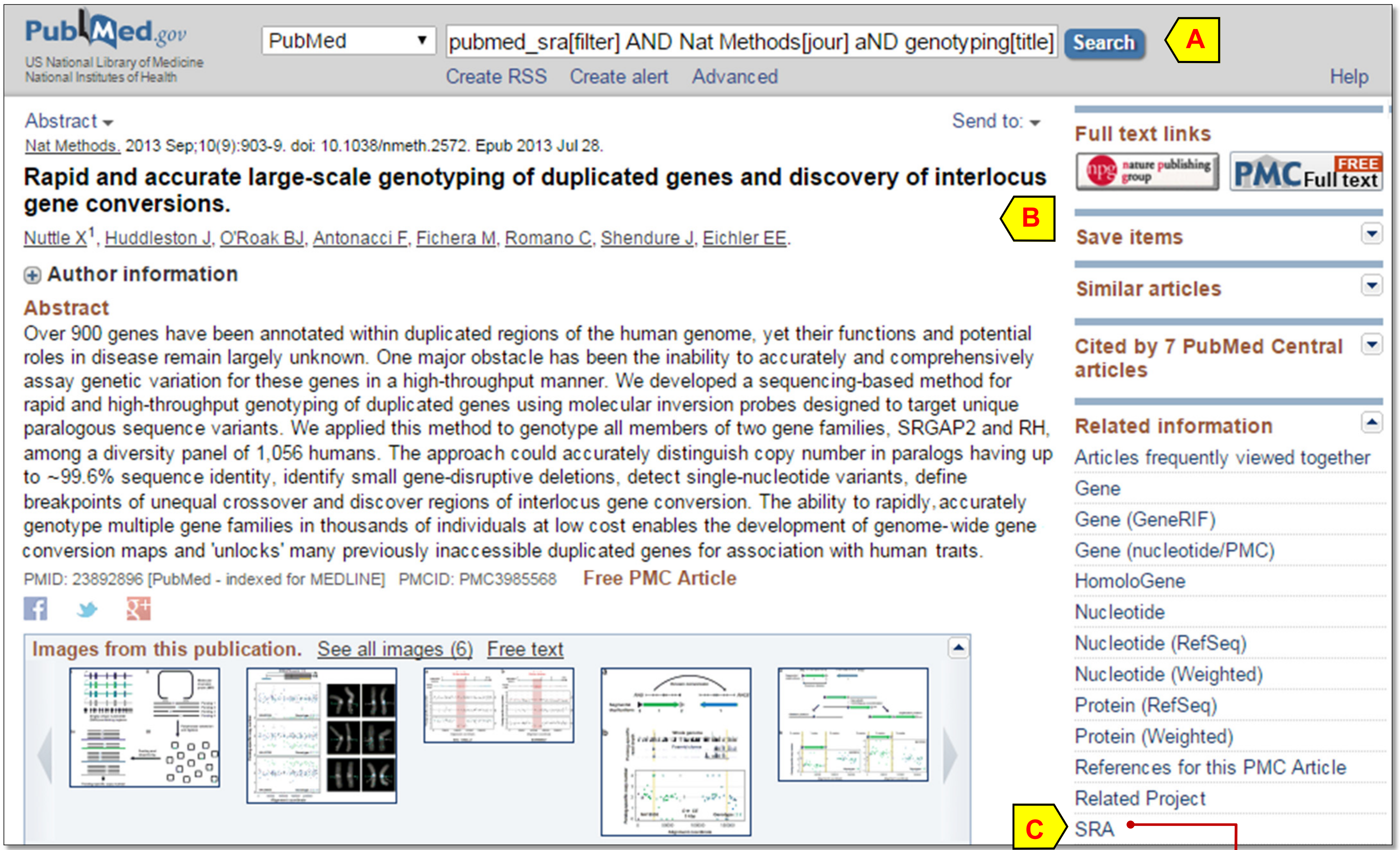
Scope and access

Sequence Read Archive (SRA) is the NCBI database which stores sequence data obtained from next generation sequence (NGS) technology. Through this database, you can search metadata for those sequences to locate the sequence reads for download and further downstream analyses. Specifically, SRA:

- Archives raw oversampling NGS data for various organisms from several platforms
- Shares submitted NGS data with EMBL and DDBJ
- Serves as a starting point for "secondary analyses"
- Provides access to data from human clinical samples to authorized users who agree to the datasets' privacy and usage mandates

You can query metadata from SRA through Entrez SRA page (www.ncbi.nlm.nih.gov/sra/), or browse the SRA project list and sequence data, or search and download them from its homepage (www.ncbi.nlm.nih.gov/Traces/sra/), respectively. You can also do sequence-based search using The "Search SRA by experiment" link under the "Specialized BLAST" section of the BLAST homepage (blast.ncbi.nlm.nih.gov/) to search against certain subsets of SRA reads. The NCBI sratoolkit, version 2.4.1 and newer, provides two command line tools to allow local BLAST searches against specific sra files directly. The downloading link is in the Entrez SRA page.

Finding NGS data through PubMed's SRA links



pubmed_sra[filter] AND Nat Methods[jour] aND genotyping[title]

Nat Methods. 2013 Sep;10(9):903-9. doi: 10.1038/nmeth.2572. Epub 2013 Jul 28.

Rapid and accurate large-scale genotyping of duplicated genes and discovery of interlocus gene conversions.

Nuttle X[1], Huddleston J, O'Roak BJ, Antonacci F, Fichera M, Romano C, Shendure J, Eichler EE.

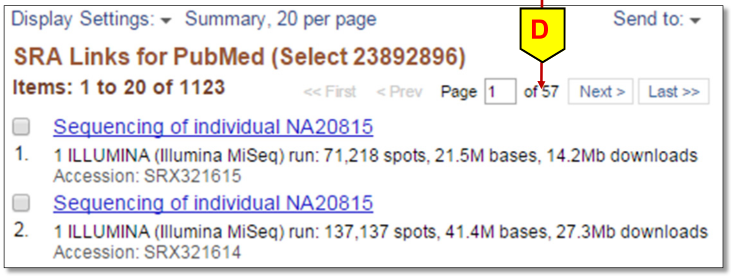
⊕ Author information

Abstract

Over 900 genes have been annotated within duplicated regions of the human genome, yet their functions and potential roles in disease remain largely unknown. One major obstacle has been the inability to accurately and comprehensively assay genetic variation for these genes in a high-throughput manner. We developed a sequencing-based method for rapid and high-throughput genotyping of duplicated genes using molecular inversion probes designed to target unique paralogous sequence variants. We applied this method to genotype all members of two gene families, SRGAP2 and RH, among a diversity panel of 1,056 humans. The approach could accurately distinguish copy number in paralogs having up to ~99.6% sequence identity, identify small gene-disruptive deletions, detect single-nucleotide variants, define breakpoints of unequal crossover and discover regions of interlocus gene conversion. The ability to rapidly, accurately genotype multiple gene families in thousands of individuals at low cost enables the development of genome-wide gene conversion maps and 'unlocks' many previously inaccessible duplicated genes for association with human traits.

PMID: 23892896 [PubMed - indexed for MEDLINE] PMCID: PMC3985568 Free PMC Article

Interests in a specific set of SRA data are often prompted by a publication. PubMed indexes abstracts with associated SRA data set through a field-limited term "**pubmed_sra [filter]**". Combining this with additional terms (**A**) retrieves a selective set of PubMed records with links to SRA data, such as the one in display (**B**). Click the SRA link (**C**) in the "Related Information" section to retrieves all the relevant datasets from SRA in the summary format (**D**), which lists the title of the experiment, the adopted platform, number of spots, number of bases, size of the download file, as well as accessions of the experiment.



SRA Links for PubMed (Select 23892896)

Items: 1 to 20 of 1123

□ Sequencing of individual NA20815
1. 1 ILLUMINA (Illumina MiSeq) run: 71,218 spots, 21.5M bases, 14.2Mb downloads
Accession: SRX321615

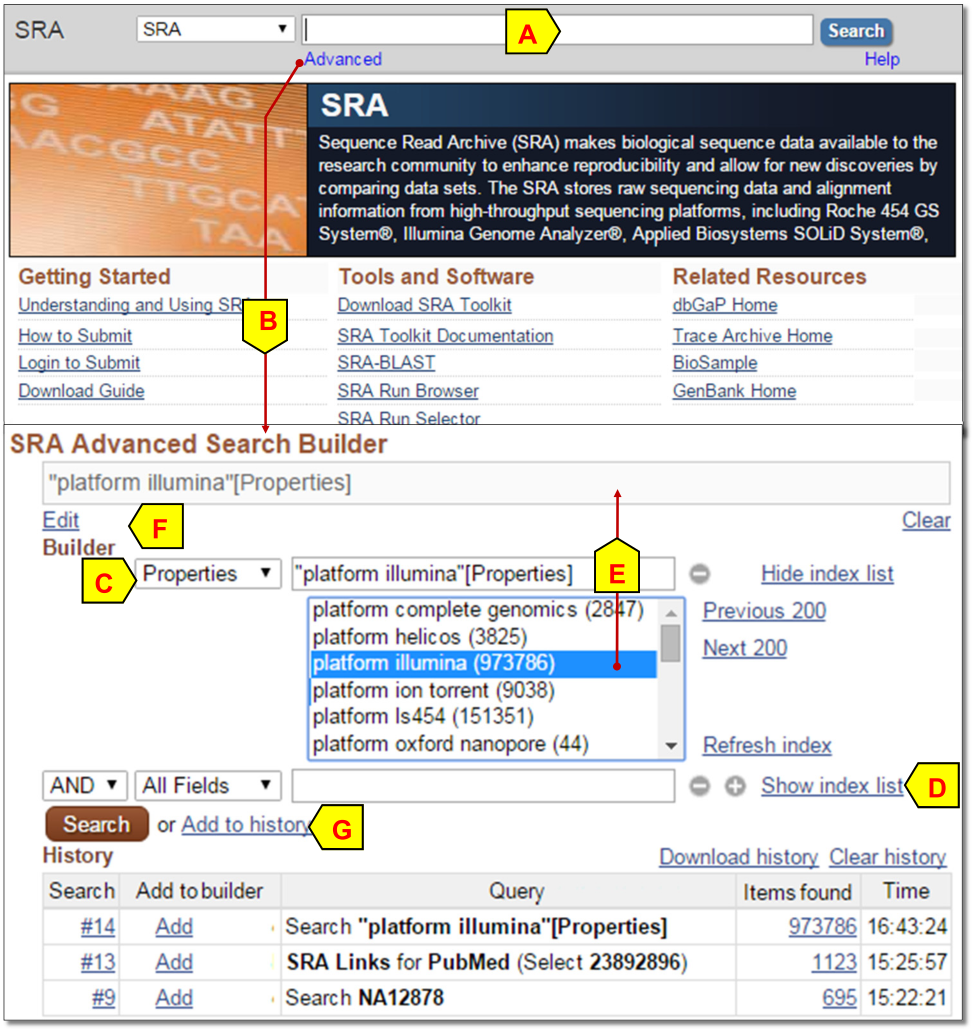
□ Sequencing of individual NA20815
2. 1 ILLUMINA (Illumina MiSeq) run: 137,137 spots, 41.4M bases, 27.3Mb downloads
Accession: SRX321614

Searching SRA metadata

You can search SRA metadata through the Entrez SRA page by entering desired terms and clicking the "Search" button (**A**). The Advanced (**B**) page provides access to indexing fields (**C**) and terms indexed under them through the "Show index list" link (**D**).

Highlight a term from the list to add it to the query box with the selected Boolean operator (**E**). Unlock the query box using the Edit link (**F**) to enter custom terms, such as history #, to construct complex queries. Click Add to history link (**G**) to preview the number of records retrieved by the terms in the query box, which also adds an entry to the History table (#4 and #5) at the bottom of the page.

The system displays initial search results in summary format (**H**), listing the title, platform and data file size, as well as the experiment accession. For details, click a title (**I**) to open that record in the "Full" display format.



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®,

Getting Started	Tools and Software	Related Resources
Understanding and Using SR	Download SRA Toolkit	dbGaP Home
How to Submit	SRA Toolkit Documentation	Trace Archive Home
Login to Submit	SRA-BLAST	BioSample
Download Guide	SRA Run Browser	GenBank Home
	SRA Run Selector	

SRA Advanced Search Builder

"platform illumina"[Properties]

Edit Clear

Builder

Properties ▾ "platform illumina"[Properties] Hide index list

platform complete genomics (2847)
platform helicos (3825)
platform illumina (973786)
platform ion torrent (9038)
platform ls454 (151351)
platform oxford nanopore (44)

Previous 200
Next 200
Refresh index

AND ▾ All Fields ▾ Show index list

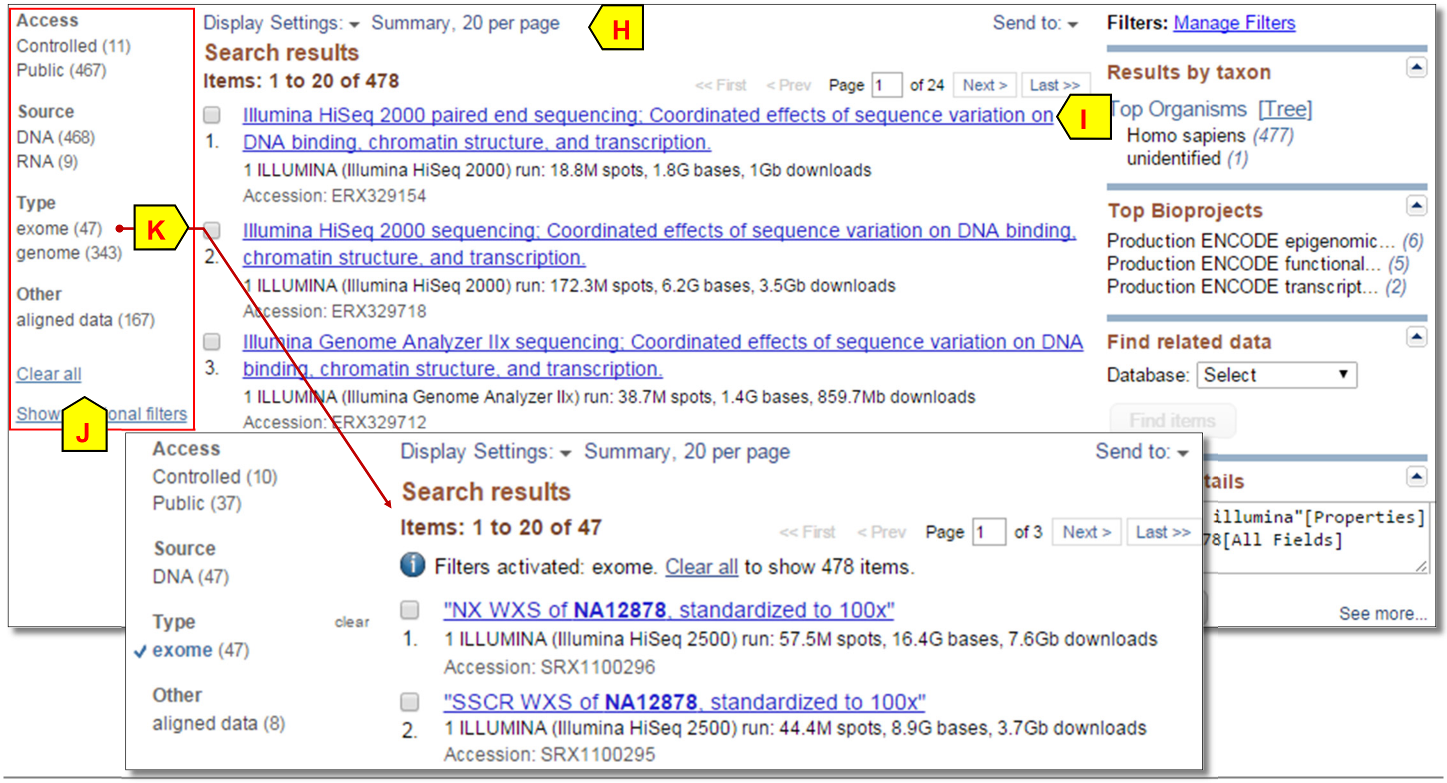
Search or Add to history

History Download history Clear history

Search	Add to builder	Query	Items found	Time
#14	Add	Search **"platform illumina"[Properties]**	973786	16:43:24
#13	Add	**SRA Links for PubMed** (Select 23892896)	1123	15:25:57
#9	Add	Search **NA12878**	695	15:22:21

Using pre-set filters

A search could retrieve a large number of experiments, which is hard to examine manually. You can use the preset filters listed in the left-hand column (**J**) to get experiments with more desirable characteristics. For example, you can click the "type: exome (47)" filter (**K**) to reduce the initial search set to those with exome (RNA-seq) data.



Access
Controlled (11)
Public (467)

Source
DNA (468)
RNA (9)

Type
exome (47)
genome (343)

Other
aligned data (167)

Clear all
Show additional filters

Display Settings: Summary, 20 per page

Search results

Items: 1 to 20 of 478

<< First < Prev Page 1 of 24 Next > Last >>

1. Illumina HiSeq 2000 paired end sequencing; Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.
1 ILLUMINA (Illumina HiSeq 2000) run: 18.8M spots, 1.8G bases, 1Gb downloads
Accession: ERX329154

2. Illumina HiSeq 2000 sequencing; Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.
1 ILLUMINA (Illumina HiSeq 2000) run: 172.3M spots, 6.2G bases, 3.5Gb downloads
Accession: ERX329718

3. Illumina Genome Analyzer IIx sequencing; Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription.
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 38.7M spots, 1.4G bases, 859.7Mb downloads
Accession: ERX329712

Results by taxon
Top Organisms [Tree]
Homo sapiens (477)
unidentified (1)

Top Bioprojects
Production ENCODE epigenomic... (6)
Production ENCODE functional... (5)
Production ENCODE transcript... (2)

Find related data
Database: Select ▾

...tails
illumina"[Properties]
78[All Fields]

Access
Controlled (10)
Public (37)

Source
DNA (47)

Type clear
✓ exome (47)

Other
aligned data (8)

Display Settings: Summary, 20 per page

Send to: ▾

Search results

Items: 1 to 20 of 47

<< First < Prev Page 1 of 3 Next > Last >>

Filters activated: exome. Clear all to show 478 items.

1. "NX WXS of NA12878, standardized to 100x"
1 ILLUMINA (Illumina HiSeq 2500) run: 57.5M spots, 16.4G bases, 7.6Gb downloads
Accession: SRX1100296

2. "SSCR WXS of NA12878, standardized to 100x"
1 ILLUMINA (Illumina HiSeq 2500) run: 44.4M spots, 8.9G bases, 3.7Gb downloads
Accession: SRX1100295

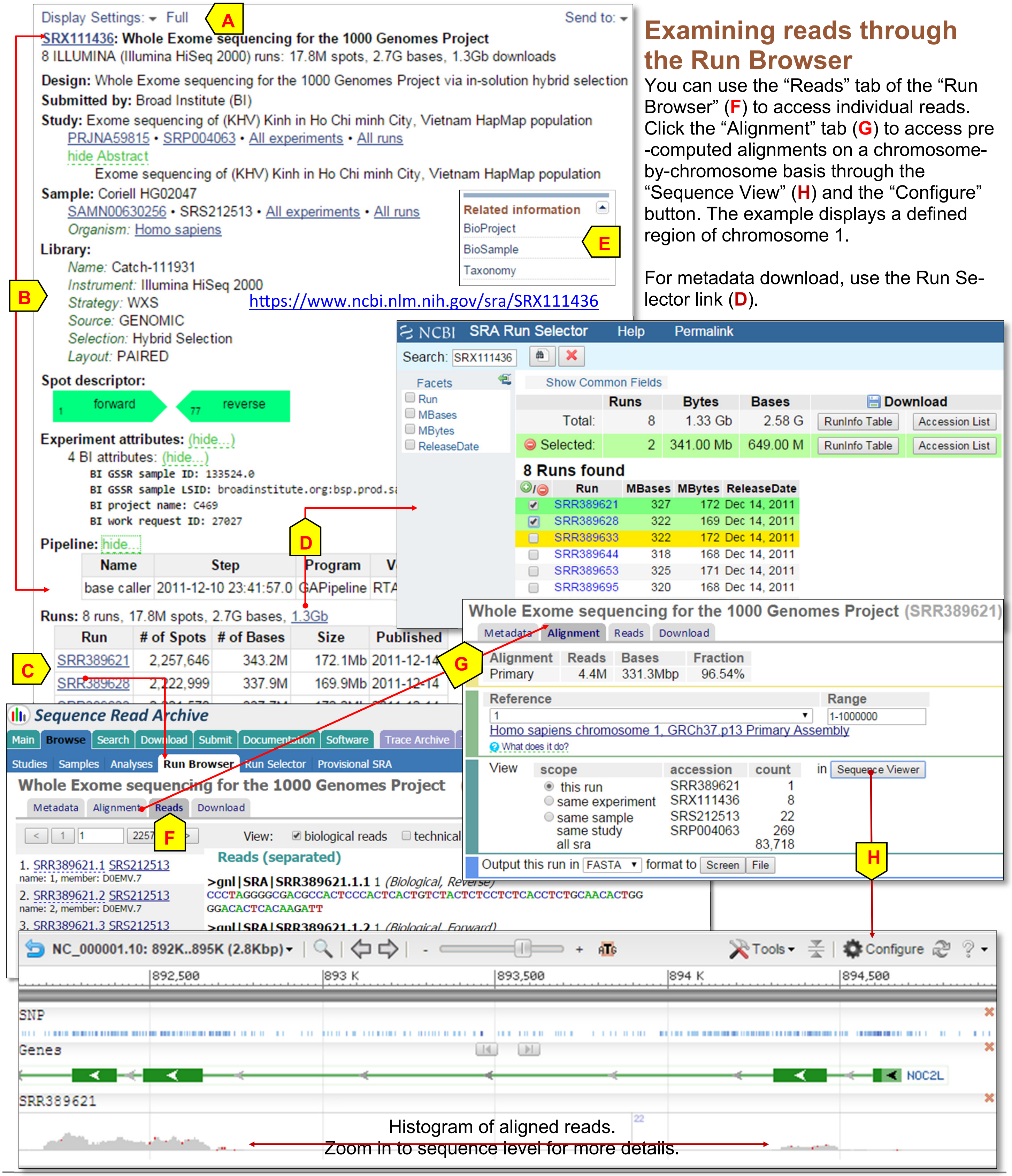
The metadata display

Click the title of an experiment retrieved from a search to open the record in "Full" display format (**A**) for more details about the experiment. In this display, the summary of the experiment is at the top (**B**), which is followed by links to individual run data in the SRA Run Browsers (**C**) and collection of runs in the Run Selector (**D**). Entries in other databases related to this experiment, such as BioSample, Taxonomy, and PubMed (if available), are shown in the "Related Information" portlet (**E**).

Examining reads through the Run Browser

You can use the "Reads" tab of the "Run Browser" (**F**) to access individual reads. Click the "Alignment" tab (**G**) to access pre-computed alignments on a chromosome-by-chromosome basis through the "Sequence View" (**H**) and the "Configure" button. The example displays a defined region of chromosome 1.

For metadata download, use the Run Selector link (**D**).



https://www.ncbi.nlm.nih.gov/sra/SRX111436

SRX111436: **Whole Exome sequencing for the 1000 Genomes Project**
8 ILLUMINA (Illumina HiSeq 2000) runs: 17.8M spots, 2.7G bases, 1.3Gb downloads

Design: Whole Exome sequencing for the 1000 Genomes Project via in-solution hybrid selection
Submitted by: Broad Institute (BI)
Study: Exome sequencing of (KHV) Kinh in Ho Chi minh City, Vietnam HapMap population
PRJNA59815 • SRP004063 • All experiments • All runs
hide Abstract
Exome sequencing of (KHV) Kinh in Ho Chi minh City, Vietnam HapMap population
Sample: Coriell HG02047
SAMN00630256 • SRS212513 • All experiments • All runs
Organism: Homo sapiens
Library:
Name: Catch-111931
Instrument: Illumina HiSeq 2000
Strategy: WXS
Source: GENOMIC
Selection: Hybrid Selection
Layout: PAIRED
Spot descriptor:
Experiment attributes: (hide...)
4 BI attributes: (hide...)

```
BI GSSR sample ID: 133524.0
BI GSSR sample LSID: broadinstitute.org:bsp.prod.s
BI project name: C469
BI work request ID: 27027
```

Pipeline: hide...

Name	Step	Program	V
base caller	2011-12-10 23:41:57.0	GAPipeline	RTA

Runs: 8 runs, 17.8M spots, 2.7G bases, 1.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR389621	2,257,646	343.2M	172.1Mb	2011-12-14
SRR389628	2,222,999	337.9M	169.9Mb	2011-12-14

Search: SRX111436

	Runs	Bytes	Bases	Download	
Total:	8	1.33 Gb	2.58 G	RunInfo Table	Accession List
Selected:	2	341.00 Mb	649.00 M	RunInfo Table	Accession List

8 Runs found

	Run	MBases	MBytes	ReleaseDate
☑	SRR389621	327	172	Dec 14, 2011
☑	SRR389628	322	169	Dec 14, 2011
□	SRR389633	322	172	Dec 14, 2011
□	SRR389644	318	168	Dec 14, 2011
□	SRR389653	325	171	Dec 14, 2011
□	SRR389695	320	168	Dec 14, 2011

Whole Exome sequencing for the 1000 Genomes Project (SRR389621)

Alignment	Reads	Bases	Fraction
Primary	4.4M	331.3Mbp	96.54%

Reference: 1 — Range: 1-1000000
Homo sapiens chromosome 1, GRCh37.p13 Primary Assembly
What does it do?

View	scope	accession	count
	● this run	SRR389621	1
	○ same experiment	SRX111436	8
	○ same sample	SRS212513	22
	same study	SRP004063	269
	all sra		83,718

in Sequence Viewer

Output this run in FASTA format to Screen File

Whole Exome sequencing for the 1000 Genomes Project

< 1 1 2257 > View: ☑ biological reads □ technical

1. SRR389621.1 SRS212513
name: 1, member: D0EMV.7
2. SRR389621.2 SRS212513
name: 2, member: D0EMV.7
3. SRR389621.3 SRS212513

Reads (separated)

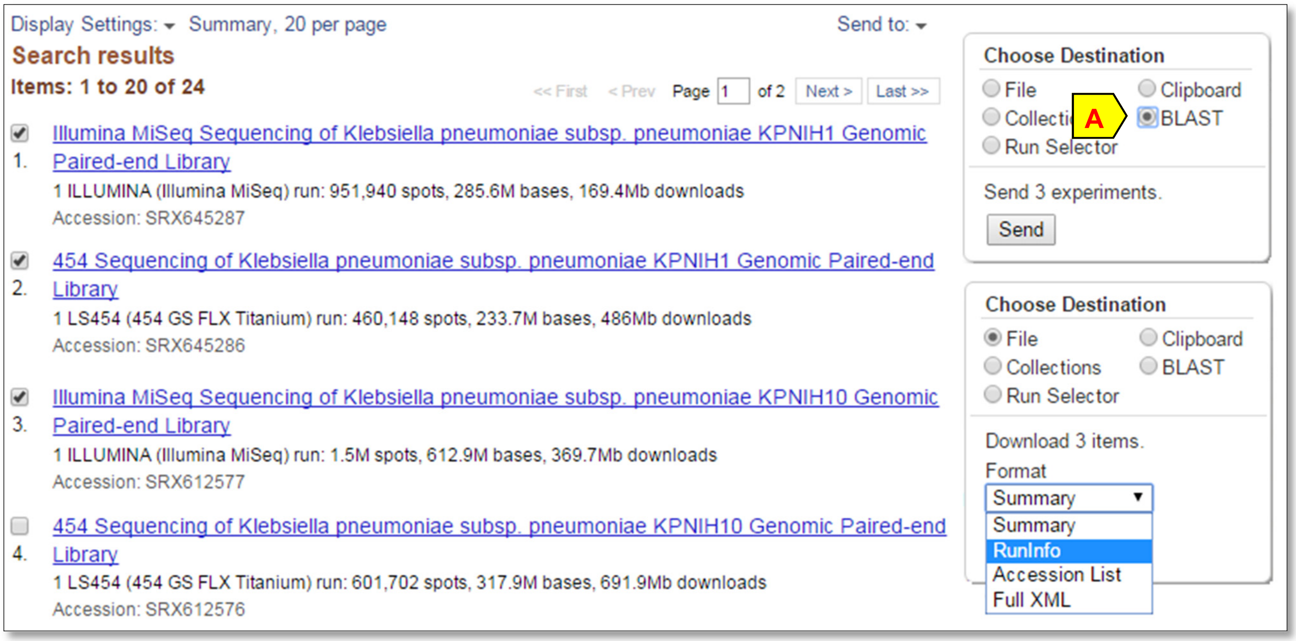
```
>gnl|SRA|SRR389621.1.1 1 (Biological, Reverse)
CCCTAGGGGCGACGCCACTCCCACTCACTGTCTACTCTCCTCTCACCTCTGCAACACTGG
GGACACTCACAAGATT
>gnl|SRA|SRR389621.1.2 1 (Biological, Forward)
```

NC_000001.10: 892K..895K (2.8Kbp)

Histogram of aligned reads.
Zoom in to sequence level for more details.

BLAST searching and downloading the sequence data

For selected SRA dataset, yon can use "Send to" >> "BLAST" (**A**) to generate a preconfigured BLAST page with the dataset set as the target database.



Display Settings: ▾ Summary, 20 per page Send to: ▾

Search results

Items: 1 to 20 of 24 << First < Prev Page 1 of 2 Next > Last >>

☑ 1. Illumina MiSeq Sequencing of Klebsiella pneumoniae subsp. pneumoniae KPNIH1 Genomic Paired-end Library
1 ILLUMINA (Illumina MiSeq) run: 951,940 spots, 285.6M bases, 169.4Mb downloads
Accession: SRX645287

☑ 2. 454 Sequencing of Klebsiella pneumoniae subsp. pneumoniae KPNIH1 Genomic Paired-end Library
1 LS454 (454 GS FLX Titanium) run: 460,148 spots, 233.7M bases, 486Mb downloads
Accession: SRX645286

☑ 3. Illumina MiSeq Sequencing of Klebsiella pneumoniae subsp. pneumoniae KPNIH10 Genomic Paired-end Library
1 ILLUMINA (Illumina MiSeq) run: 1.5M spots, 612.9M bases, 369.7Mb downloads
Accession: SRX612577

☐ 4. 454 Sequencing of Klebsiella pneumoniae subsp. pneumoniae KPNIH10 Genomic Paired-end Library
1 LS454 (454 GS FLX Titanium) run: 601,702 spots, 317.9M bases, 691.9Mb downloads
Accession: SRX612576

Command line tools from the NCBI SRA Toolkit (www.ncbi.nlm.nih.gov/Traces/sra/?view=software) can remotely prefetch data from the NCBI SRA site and process them locally, when fed a valid SRR accession as input. For local BLAST search against specific SRA datasets specified with SRR accessions, you can use the newly introduced blastn_vdb and tblastn_vdb command line tools. This prefetch function can take advantage of the faster download speed provided by through Aspera plugin, if you have already installed it on your computer. The example command line below uses tblastn_vdb to do a translated search with a drug resistance protein sequence from *Escherichia coli* (-query mdr_sequence.aa), against two *Klebsiella pneumoniae* datasets (-db "SRR1427233 SRR55906"), ask for tabular output (-outfmt 6), and save the results to a file (-out sra_tblastn.tab). The system automatically fetches the data from NCBI if you do not have the data files already downloaded locally.

```
tblastn_vdb -query mdr_sequence.aa -db "SRR1427233 SRR515906" -outfmt 6 -max_target_seqs 2500 -out
sra_tblastn.tab
```

Given an XRR (SRR/ERR/DRR) accession, you can use the following steps to reconstruct the FTP path for the .sra file:

- The base FTP path is **ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/**
- Append /XRR to get to the different source directory (with X being S, E, or D)
- Append /XRR### with the # being the first three digits of the XRR accession, for SRR1427233, use /SRR142
- Append XRR full accession, for SRR1427233, use /SRR1427233
- Append the full accession with .sra extension, for SRR1427233, use /SRR1427233.sra to arrive at:

ftp.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR142/SRR1427233/SRR1427233.sra
For ascp, replace the ftp.ncbi.nlm.nih.gov with **anonftp@ftp-private.ncbi.nlm.nih.gov:** to arrive at:
anonftp@ftp-private.ncbi.nlm.nih.gov:/sra/sra-instant/reads/ByRun/sra/SRR/SRR142/SRR1427233/SRR1427233.sra

References

SRA help documentation is available from the NCBI Bookshelf at:
www.ncbi.nlm.nih.gov/books/NBK47528/
The software package for processing downloaded SRA data (sratoolkit) are available from this page:
www.ncbi.nlm.nih.gov/Traces/sra/?view=software
Document on sratoolkit is available from this page:
www.ncbi.nlm.nih.gov/Traces/sra/?view=toolkit_doc
A handout for Sequence Viewer is at:
ftp.ncbi.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf
SRA-specific comments and submission-related questions can be addressed to
sra@ncbi.nlm.nih.gov