



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.

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed Search **A**

Create RSS Create alert Advanced Help

Abstract **B**

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.

Nuttie X¹, 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**

Images from this publication. See all images (6) Free text **C**

Full text links

Save items

Similar articles

Cited by 7 PubMed Central articles

Related information

Articles frequently viewed together

Gene

Gene (GeneRIF)

Gene (nucleotide/PMC)

HomoloGene

Nucleotide

Nucleotide (RefSeq)

Nucleotide (Weighted)

Protein (RefSeq)

Protein (Weighted)

References for this PMC Article

Related Project

SRA **D**

Finding NGS Data Through PubMed’s SRA Links

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.

Display Settings: Summary, 20 per page

SRA Links for PubMed (Select 23892896)

Items: 1 to 20 of 1123 << First < Prev Page 1 of 57 Next > Last >>

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.

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.

The screenshot shows the SRA Advanced Search Builder interface. At the top, there is a search bar with the text "SRA" and a "Search" button (A). Below the search bar, there is a navigation menu with "Advanced" selected (B). The main content area is divided into three columns: "Getting Started" (with links like "Understanding and Using SRA"), "Tools and Software" (with links like "Download SRA Toolkit"), and "Related Resources" (with links like "dbGaP Home"). Below this is the "SRA Advanced Search Builder" section. It features a query box containing the text "platform illumina"[Properties] (E). To the left of the query box is an "Edit Builder" section with a dropdown menu set to "Properties" (C) and an "Edit" link (F). Below the query box is a list of indexed terms, with "platform illumina (973786)" highlighted (E). To the right of the list are links for "Hide index list", "Previous 200", "Next 200", and "Refresh index". Below the list is a "Show index list" link (D). At the bottom of the search builder, there is a "Search" button (G) and an "Add to history" link. Below the search builder is a "History" table with columns for "Search", "Add to builder", "Query", "Items found", and "Time". The table contains three entries: #14 (Search "platform illumina"[Properties]), #13 (SRA Links for PubMed (Select 23892896)), and #9 (Search NA12878). At the bottom of the page, there is a "Filters: Manage Filters" section with "Results by taxon" and "Top Bioprojects" sub-sections.

The screenshot shows the SRA search results page. At the top, there is a "Display Settings" dropdown set to "Summary, 20 per page" (H). Below this is the "Search results" section, which displays a list of items (1 to 20 of 478). The first item is "Illumina HiSeq 2000 paired end sequencing: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription." (I). The second item is "Illumina HiSeq 2000 sequencing: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription." The third item is "Illumina Genome Analyzer Iix sequencing: Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription." To the left of the search results is a "Filters" section (J) with a "Show additional filters" link. The filters include "Access" (Controlled (11), Public (467)), "Source" (DNA (468), RNA (9)), "Type" (exome (47), genome (343)), and "Other" (aligned data (167)). A "Clear all" link is also present. Below the filters is a "Show additional filters" link (J). Below the search results is a "Filters activated" section (K) showing "exome (47)" selected. Below this is a "Search results" section (1 to 20 of 47) with a "Filters activated: exome. Clear all to show 478 items." message. The first item is "NX WXS of NA12878, standardized to 100x" (1 ILLUMINA (Illumina HiSeq 2500) run: 57.5M spots, 16.4G bases, 7.6Gb downloads, Accession: SRX1100296). The second item is "SSCR WXS of NA12878, standardized to 100x" (1 ILLUMINA (Illumina HiSeq 2500) run: 44.4M spots, 8.9G bases, 3.7Gb downloads, Accession: SRX1100295). To the right of the search results is a "Send to:" dropdown and a "Filters: Manage Filters" link. Below the search results is a "Results by taxon" section with "Top Organisms" (Homo sapiens (477), unidentified (1)) and "Top Bioprojects" (Production ENCODE epigenomic... (6), Production ENCODE functional... (5), Production ENCODE transcript... (2)). Below the results by taxon is a "Find related data" section with a "Database:" dropdown set to "Select".

The Metadata Display

Click the title of an experiment 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 in the right-hand column (E).

Full A

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

Design: Whole Exome sequencing for the 1000 Genomes Project
Submitted by: Broad Institute (BI)
Study: Exome sequencing of (KHV) Kinh in Ho Chi minh City, Vietnam
[PRJNA59815](#) • [SRP004063](#) • [All experiments](#) • [All runs](#)

Sample: Coriell HG02047
[SAMN00630256](#) • [SRS212513](#) • [All experiments](#) • [All runs](#)
Organism: [Homo sapiens](#)

Library: www.ncbi.nlm.nih.gov/sra/SRX111436
Name: Catch-111931
Instrument: Illumina HiSeq 2000
Strategy: WXS
Source: GENOMIC
Selection: Hybrid Selection
Layout: PAIRED

Spot descriptor:
1 forward 77 reverse

Experiment attributes: [\(hide...\)](#)
4 BI attributes: [\(hide...\)](#)
BI GSSR sample ID: 133524.0
BI GSSR sample LSID: broadinstitute.org:bsp.
BI project name: C469
BI work request ID: 27027

Pipeline: [\(hide...\)](#)

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

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

Run	# of Spots	# of Bases	Size
SRR389621	2,257,646	343.2M	172.1
SRR389628	2,222,999	337.9M	169.9
SRR389633	2,221,570	337.7M	172.2
SRR389644	2,194,040	333.5M	168.2

Related information E

- Assembly
- BioProject
- BioSample
- Taxonomy

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.

NCBI SRA Run Selector

Accession: SRP004063

Filters List

- Assay Type
- AvgSpotLen
- Bases
- BI_Run_Barcode
- BI_run_name
- BI_work_request_ID (Experiment)
- BI_work_request_ID (Run)
- Bytes
- Center Name
- flowcell_barcode
- gssr_id (Experiment)
- gssr_id (Run)
- Instrument
- Instrument_name
- lane
- LibrarySelection

Common Fields

BioProject: PRJNA59815
Consent: PUBLIC
DATASTORE filetype: FASTQ_SRA
DATASTORE provider: ENA, GS, NCBI, S3
DATASTORE region: ena, gs.us, ncbi.public, s3.us-east-1
LibraryLayout: PAIRED
LibrarySource: GENOMIC

Select

	Runs	Bytes	Bases	Download
Total	702	1.20 Tb	1.87 T	RunInfo Table or Accession List
Selected	0	0	0	RunInfo Table or Accession List

Found 702 Items

Run	BioSample	Assay Type	AvgSpotLen	Bases	Bytes	Center
<input type="checkbox"/> 1	ERR047698	SAMN00249811	WGS	180	12956817600	8927985040 BGI

Sequence Read Archive

Main Browse Search Download Submit Software Trace Archive

Studies Samples Analysis **Run Browser** Run Selector Provisional SRA

Whole Exome sequencing for the 1000 Genomes Project (SRR389633) [Change accession...](#)

Metadata Alignment Analysis Reads Data access

View: biological reads technical reads

Reads (separated)

- [SRR389633.1 SRS212513](#)
name: 1, member: D0EMV.1
- [SRR389633.2 SRS212513](#)
name: 2, member: D0EMV.1

Alignment G

Primary 4.3M 325.6Mbp 96.43%

Reference: [Homo sapiens chromosome 1, GRCh37, p13 Primary Assembly](#)
This run has 75 references. Only first 40 are shown.

View scope: this run (SRR389633, count: 1) same experiment (SRX111436, count: 8) same sample (SRS212513, count: 22) same study (SRP004063, count: 269) all sra (180,089)

Output this run in [FASTA] format to [Screen] [File]

NC_000001.10 Find: []

Genes, NCBI Homo sapiens Annotation Release 105.20190806

IL6R [46] IL6R-AS1 PSM8P1 NR_147855.1 exon

SRR389633

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

Pile-up, log 2 scaled

BLAST Searching and Downloading the Sequence Data

For selected SRA dataset, you can use the "Send to" menu (A) to further process the set. For example, "Send to" >> "BLAST" (B) generate a preconfigured BLAST page with the dataset set as the target database so you can align your query against the set, and "Send to" >> File (C) retrieves or downloads the set in specified format.

The screenshot shows the NCBI SRA search results page for the term 'kpni10'. The search results list four items, with two selected. The 'Send to' menu is open, showing options for 'File', 'Clipboard', 'Collections', 'BLAST', and 'Run Selector'. The 'BLAST' option is selected, and the 'Format' dropdown menu is open, showing options for 'RunInfo', 'Summary', 'RunInfo', 'Accession List', and 'Full XML'.

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 `tblastn_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 already have it installed and configured. 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 SRR515906"`), ask for 2500 hits if that many were found (`-max_target_seqs 2500`) in 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" -max_target_seqs 2500 -outfmt 6 -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/](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](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](ftp://ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR142/SRR1427233/SRR1427233.sra) 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>

NOTE: The above is only for your reference. We recommend that you use `sratoolkit`'s prefetch programs to download SRA data files, which can take advantage of Aspera when that setup exists. Other data conversion programs will be needed to check the integrity of the .sra file and dump fastq and bam formatted data from downloaded .sra files.

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.nlm.nih.gov/pub/factsheets/Factsheet_Graphical_SV.pdf

A handout for BLAST search sra reads locally is at:

ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_Local_SRA_BLAST.pdf

SRA-specific comments and submission-related questions can be addressed to

sra@ncbi.nlm.nih.gov