SRA: Sequence Read Archive

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

http://www.ncbi.nlm.nih.gov/sra/ and http://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

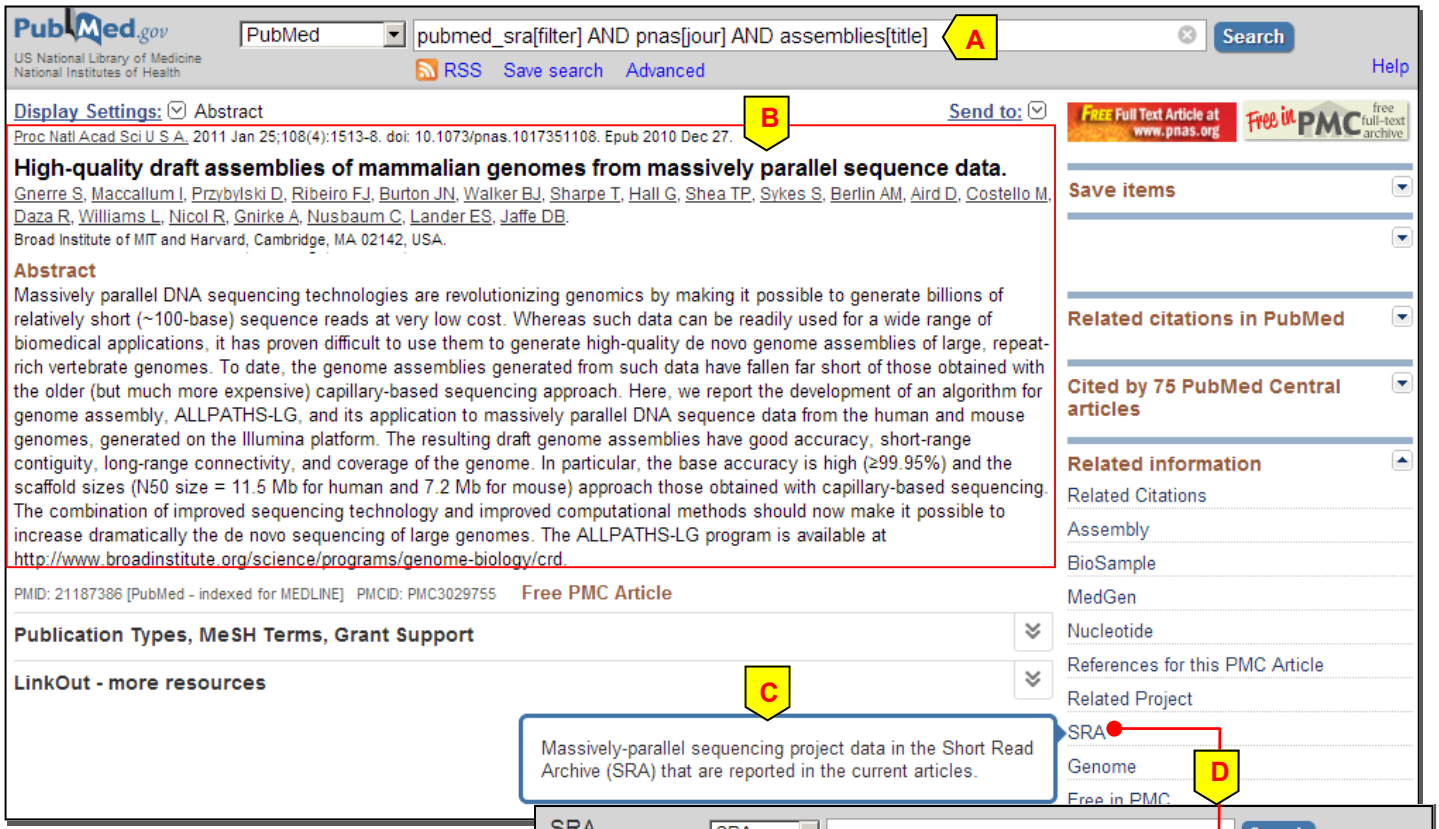
Sequence Read Archive (SRA) is the NCBI database which stores sequence data obtained from next generation sequence (NGS) technology. Through this database, the metadata for those sequences can be queried to locate the sequence dataset for subsequent download and further analysis. Specifically, SRA:

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

Metadata from SRA can be queried from Entrez SRA page and the SRA project list and sequence data can be browsed, searched and downloaded from its homepage at www.ncbi.nlm.nih.gov/sra/ and www.ncbi.nlm.nih.gov/Traces/sra/, respectively. For sequence-based searches against certain subsets of SRA reads (long reads from the Roche 454 platform) using BLAST, a link is listed under the "Specialized BLAST" section on the BLAST homepage: blast.ncbi.nlm.nih.gov/ .

Finding NextGen data through PubMed's SRA links

The interest in a specific set of SRA data is often prompted from reading a publication. PubMed indexing makes SRA data retrieval from PubMed records a straightforward process. A field-limited term "**pubmed_sra[filter]**" (**A**) retrieves PubMed records with links to SRA data, such as the ChIP-seq paper displayed below in Abstract format (**B**). The relevant SRA experiments under this project are available through the SRA link (**C**) in the "Related Information" section.



Clicking the "SRA" link retrieves all the relevant datasets and displays them in summary format (**D**) listing the title of the experiment, the adopted platform, number of spots, number of bases, size of the download file, as well as the accession of the experiment.

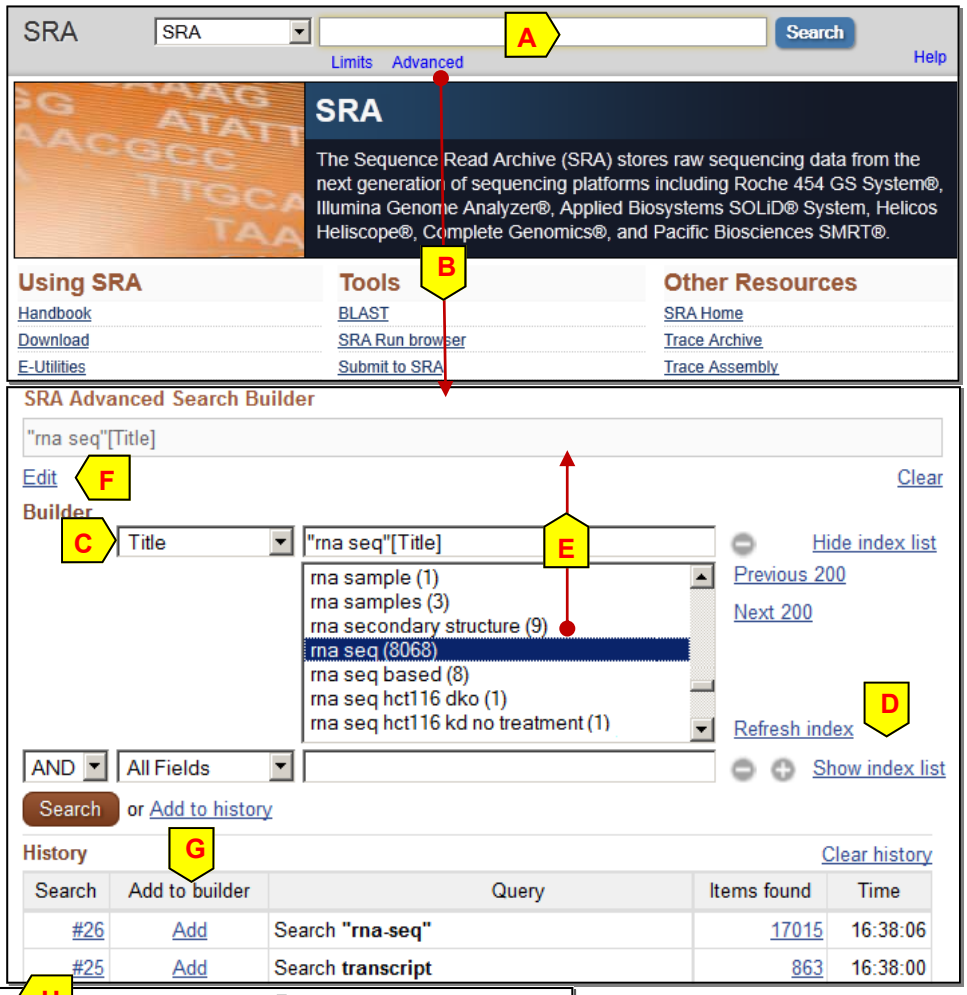
Contact: info@ncbi.nlm.nih.gov

Searching SRA metadata

Searching SRA metadata can be performed through the Entrez SRA page by entering desired terms and clicking the "Search" button (**A**). Complex queries can be constructed using functions on the "Advanced" (**B**) page, where indexing fields (**C**) and terms indexed under them can be examined using the "Show index list" link (**D**). Highlighting a term from the list automatically adds it to the query box with the selected Boolean operator (**E**).

The query box can be unlocked using the "Edit" link (**F**) so that custom terms, such as history #, can be entered. Clicking the "Add to history" link (**G**) previews the number of records retrieved by the terms in the query box and adds an entry to the History table (#26 and #25) at the bottom of the page.

Initial search results are displayed in summary format (**H**), listing the title, platform and data file size, as well as the experiment accession. Clicking the title (**I**) opens the "Full" display format, providing more details.





Using pre-set filters

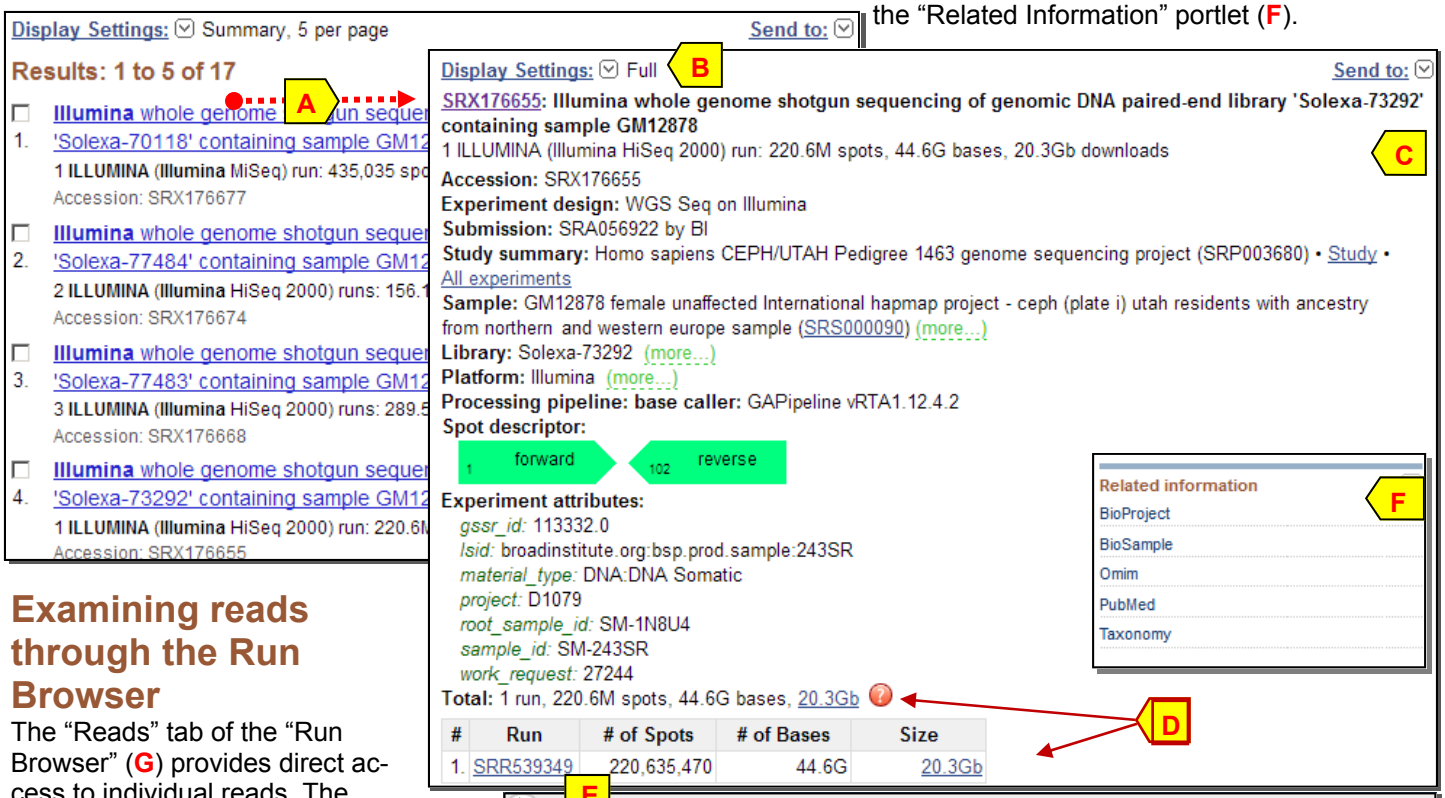
The number of experiments retrieved by a search can be large and time-consuming to exam. Additional links (**J**) in the right-hand column further filter the retrieved experiments into different categories according to their characteristics.

For example, the "type: exome (6)" filter selectively displays only the subset of experiments with exome (RNA-seq) data, while the "aligned data (456)" filter (**K**) selectively displays only the subset that have alignment to master genome assemblies.

The filtering effect is indicated by the highlighted filter and the number of Results shown above the list (**L**).

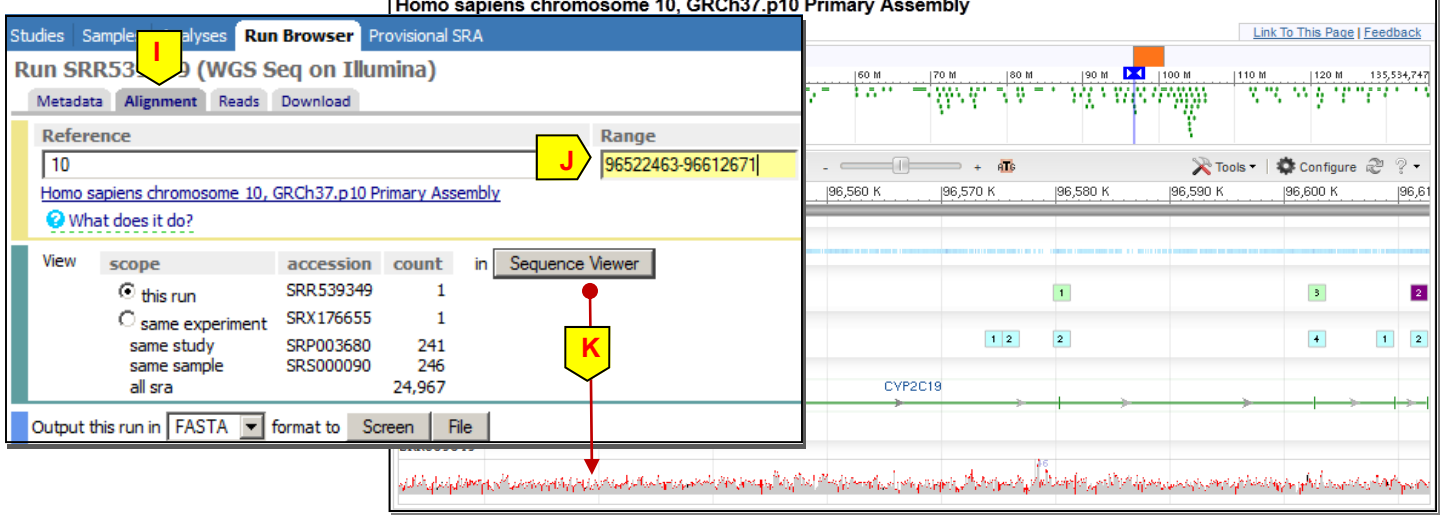
The metadata display

Clicking the title of an experiment (**A**) opens the record in "Full" display format (**B**) to provide more details about the experiment. In this display, the summary of the experiment is shown at the top (**C**), which is followed by data download links (**D**) and links to individual run data in the SRA browsers by clicking on SRR accessions (**E**). Entries in other databases related to this experiment, such as PubMed listings of the publications describing the experiment, are shown in the "Related Information" portlet (**F**).



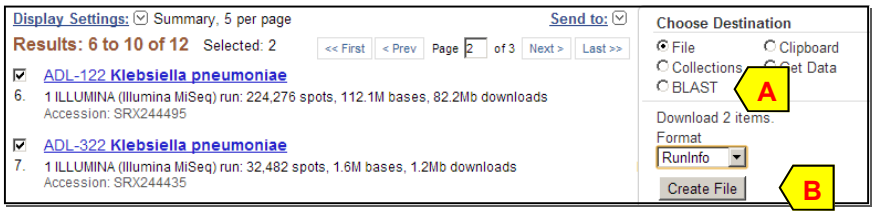
Examining reads through the Run Browser

The "Reads" tab of the "Run Browser" (**G**) provides direct access to individual reads. The "advanced options" link (**H**) pops up a window with options to allow for further customization of the reads display. The "Alignment" tab (**I**) provides access to the pre-computed alignment data on a chromosome-by-chromosome basis. In this example, a defined region of chromosome 10 (**J**) is selected. Clicking the "Sequence Viewer" button (**K**) opens the graphical display for the selected region showing the reads alignment as a separate track.



BLAST searching and downloading the sequence data

Selected SRA dataset can be searched with BLAST by using the "Send to" widget's BLAST option (**A**). The action generates a preconfigured BLAST page setting these dataset as the target database.



The ftp link to the dataset can also be saved through the "Send to" by selecting "File" and the "RunInfo" format (**B**). This "RunInfo" file provides URLs for direct download of the read data files from the NCBI's FTP site. Selected columns from an example RunInfo file are as follows.

Col 1	Col 5	Col 9
Run	**size_MB**	**download_path**
ERR145655	4385	ftp://ftp-private.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/ERR/ERR145/ERR145655/ERR145655.sra
SRR519754	8478	ftp://ftp-private.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR519/SRR519754/SRR519754.sra

For multiple files, simple scripts could be written to read URLs in the **download_path** column directly and call wget or another clients to perform the download. An example wget command line is shown:

```
wget ftp://ftp-private.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/ERR/ERR145/ERR145655/
ERR145655.sra
```

With Aspera installed, the command line tool **ascp** provides a faster alternative to regular FTP. However, since Aspera uses a different scheme to specify domain names, **ascp** cannot use these FTP URL listed in the download_path column directly. Instead, the top level domain name needs to be modified. The Aspera counterparts for the two FTP URLs shown above are now displayed below with the domain name portion in red:

Runs	Aspera and ascp recognizable URL
ERR145655	anonftp@ftp-private.ncbi.nlm.nih.gov:/sra/sra-instant/reads/ByRun/sra/ERR/ERR145/ERR145655/ERR145655.sra
SRR519754	anonftp@ftp-private.ncbi.nlm.nih.gov:/sra/sra-instant/reads/ByRun/sra/SRR/SRR519/SRR519754/SRR519754.sra

The text below shows the output from a sample **ascp** download session performed on a Windows PC. Note the download speed offered by **acsp** (**C**). At 292 Megabits/sec, it took only about 4 min to download an 8 GB file:

```
C:\ascp -QT -l 300M -m 64M -L aspera -i "C:\Program Files (x86)\Aspera\Aspera Conn

ect\etc\asperaweb_id_dsa.putty" -k 1 anonftp@ftp-private.ncbi.nlm.nih.gov:/sra/sra-instant/reads/
ByRun/sra/SRR/SRR519/SRR519754/SRR519754.sra .

SRR519754.sra                                100% 8478MB  292Mb/s    04:11

Completed: 8681786K bytes transferred in 252 seconds

 (282149K bits/sec), in 1 file.
```

B

NCBI provides an SRA toolkit so that data in a chosen format, such as FASTA or FASTQ, can be extracted from the downloaded .sra files. The SRA handbook provides additional technical details. The SRA Toolkit page, third link below, has a direct link to that section.

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 are available from this page:
www.ncbi.nlm.nih.gov/Traces/sra/?view=software
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