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

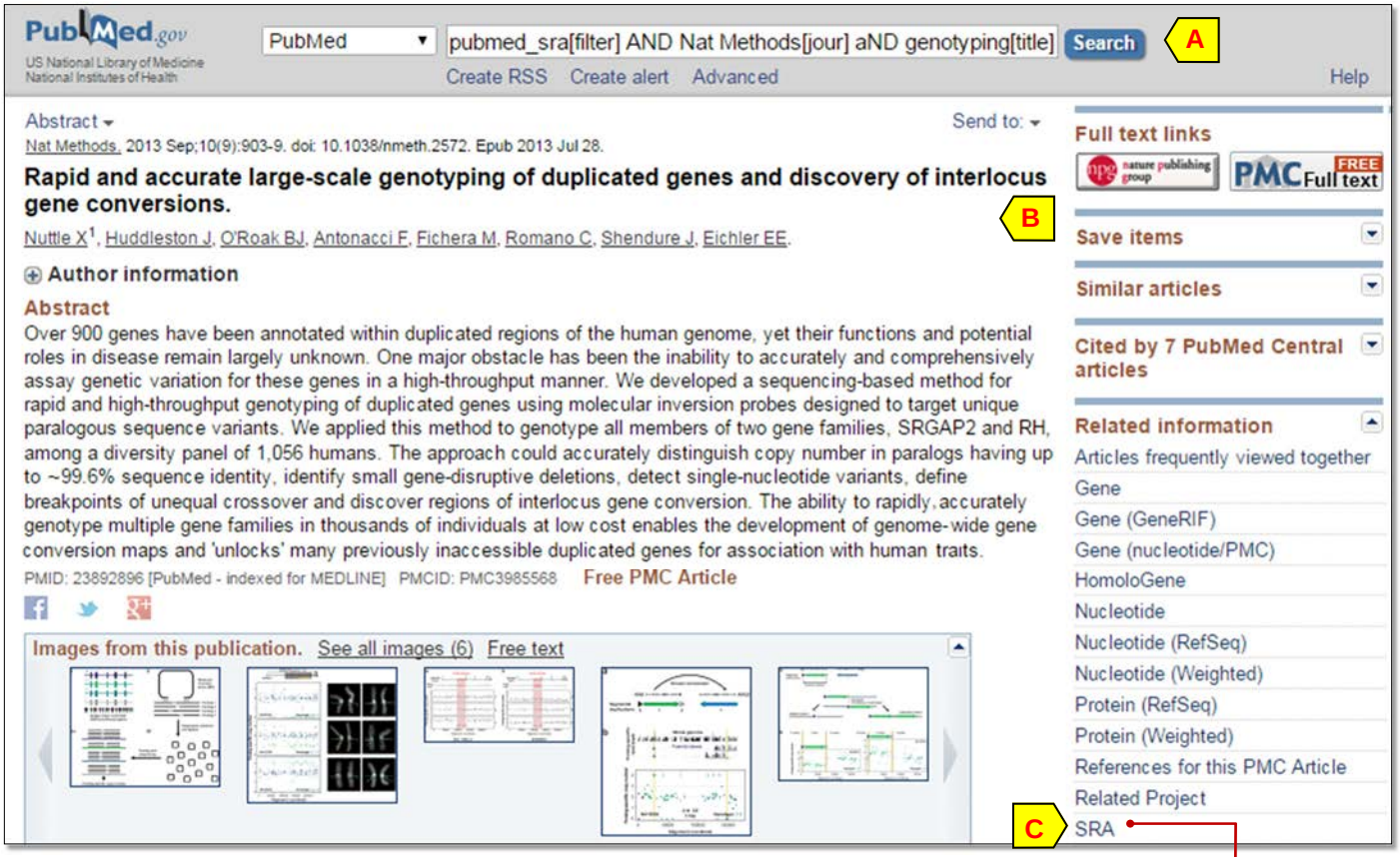
Scope and access

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



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.

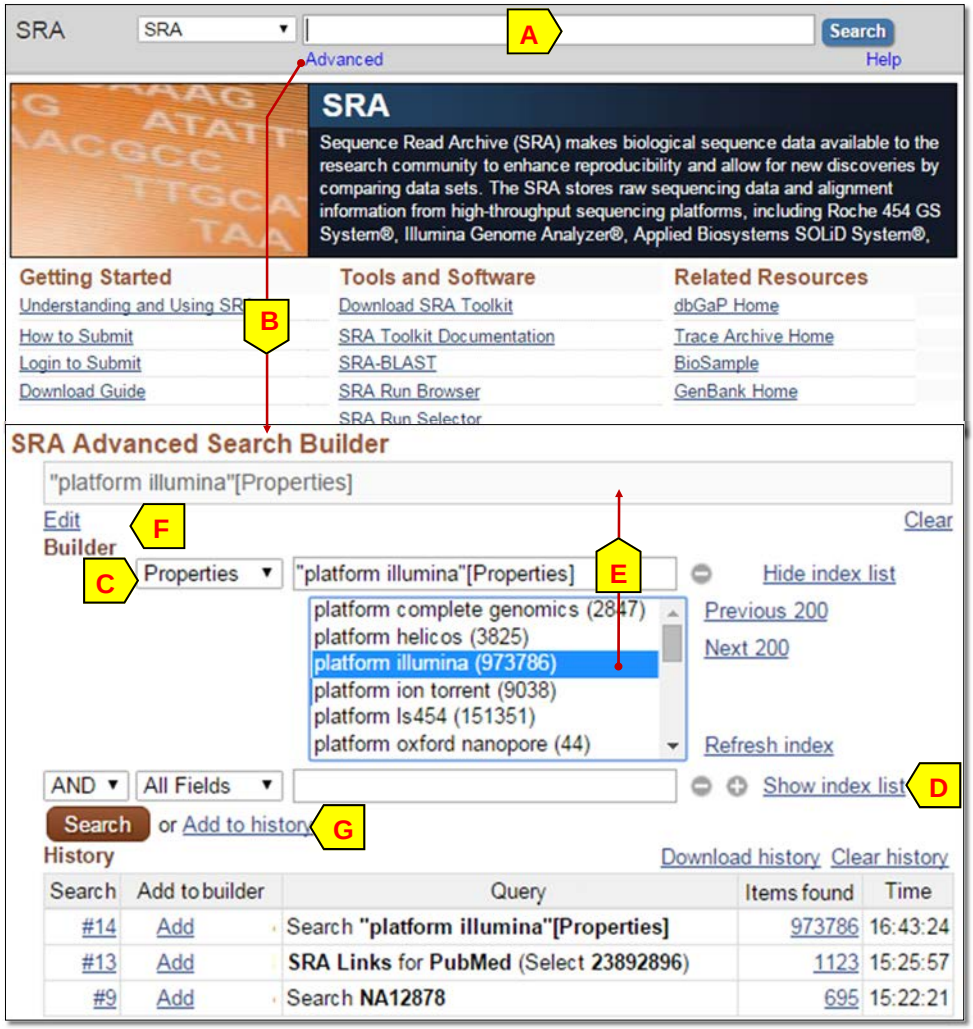


Searching SRA metadata

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**).

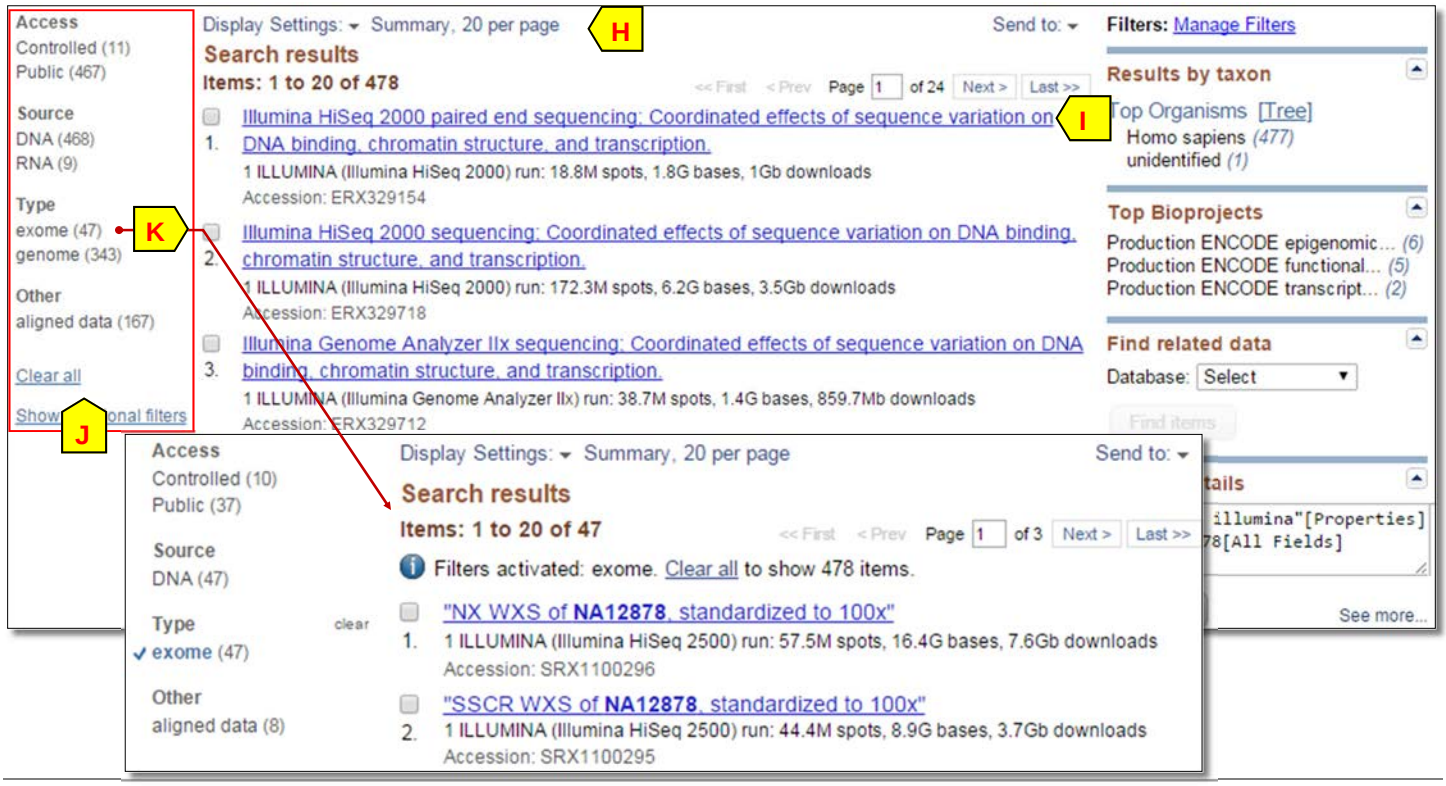
Highlight a term from the list to add it to the query box with the selected Boolean operator (**E**). Unlock the query box using the Edit link (**F**) to enter custom terms, such as history #, to construct complex queries. Click Add to history link (**G**) to preview the number of records retrieved by 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.



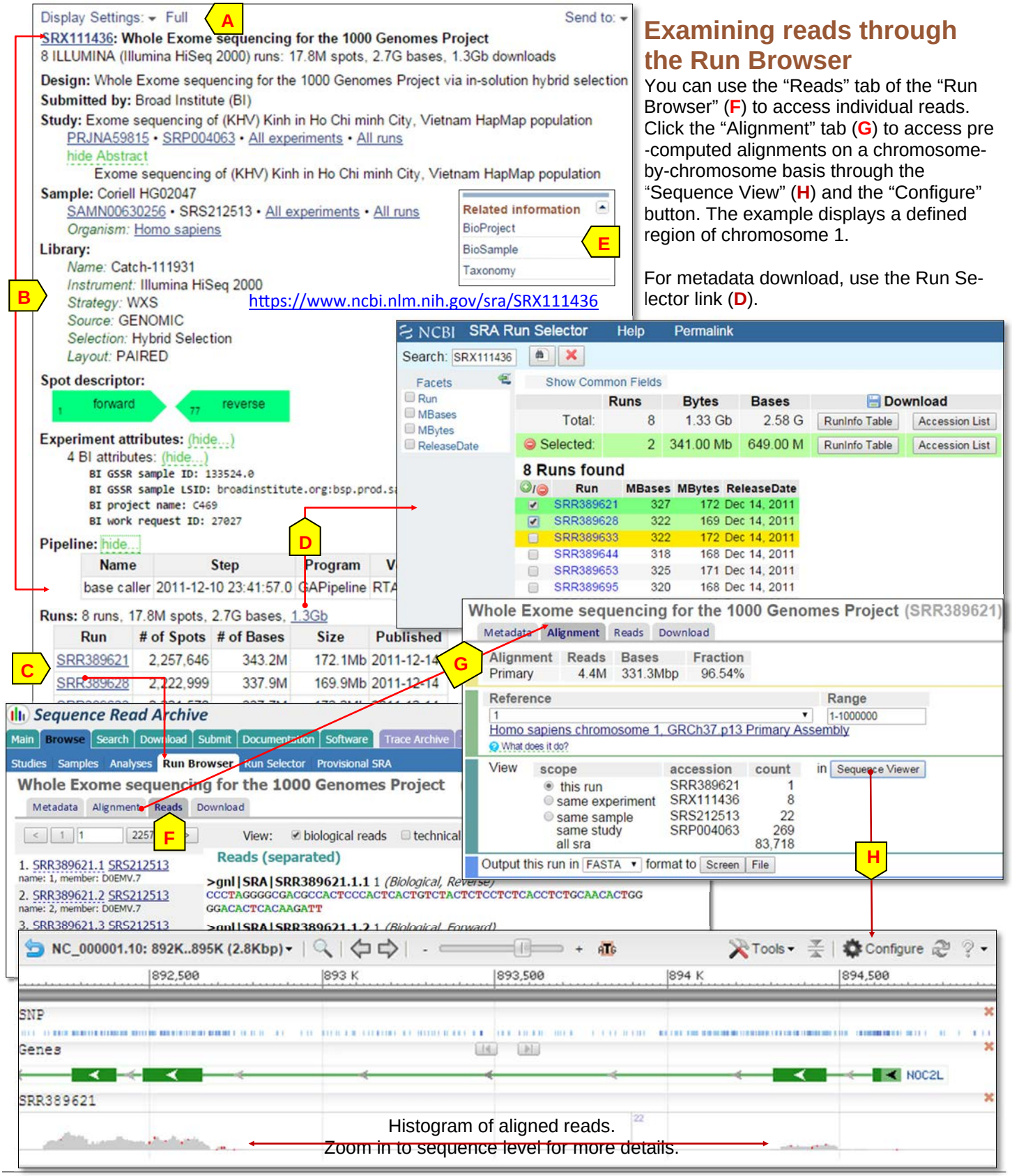
The metadata display

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

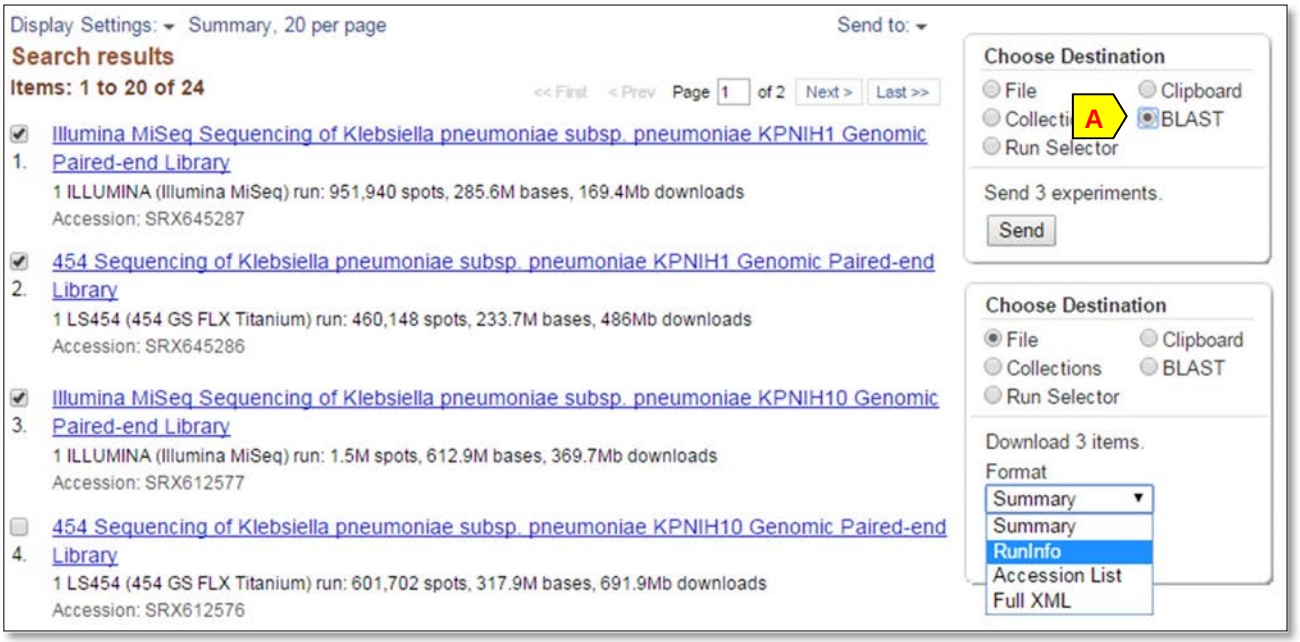
You can use the "Reads" tab of the "Run Browser" (**F**) to access individual reads. Click the "Alignment" tab (**G**) to access pre-computed alignments on a chromosome-by-chromosome basis through the "Sequence View" (**H**) and the "Configure" button. The example displays a defined region of chromosome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

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



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.



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