NGS public data repositories

Estudios in silico en Biomedicina

(Máster en Bioinformática, Universidad de Valencia)

David Montaner

2014-12-16

Sequence Read Archive (SRA)

- originally called Short Read Archive
- provides a public repository for DNA sequencing data generated by High-throughput sequencing
- collaboration between the *NCBI*, the *EBI*, and the *DNA Data Bank of Japan*

- NCBI SRA: Sequence Read Archive.
- EBI SRA: Sequence Read Archive; part of the ENA database.
- DRA: DDBJ Sequence Read Archive.

A nice list of other possible repositories:
`http://omictools.com/primary-databases-c390-p1.html`

Goals

- Provide a central repository for next generation sequencing data.
- Provide links to other resources referencing or using this data.
- Provide users with retrieval based on ancillary information and sequence comparison.
- Track studies and experiments (project meta data).
- Allow flexible submission and retrieval of ancillary data.
- Improve database efficiency through normalization of data structures.
- Separate submission from content.
- Establish basis for user-interactive submission and retrieval.

Resources

- NCBI web page
- NCBI hand book
- SOFTWARE / TOOLS
- BLAST
- ...

Data organization

- Study or *project* (SRP)
- Experiment (SRX)
- Sample (SRS)
- Run, lane, slide, plate (SRR)
- Analyses (SRZ)

See details for all *tags* in the SRA Handbook.

Sequence Read Format (SRF)

- The Sequence Read Format (SRF) is generalized normalization of data structures.
- The SRA accepts any secondary analyses typically performed on the next generation data.
- These might include alignments, small-scale assemblies, oligo profiles.
- A **toolkit** is provided that supports conversion to several popular formats.

SRA Toolkit Tools

- prefetch: Allows command-line downloading of SRA, dbGaP, and ADSP data
- sra-stat: Generate statistics about SRA data (quality distribution, etc.)
- sra-pileup: Generate pileup statistics on aligned SRA data
- fastq-dump: Convert SRA data into fastq format
- sam-dump: Convert SRA data to sam format
- illumina-dump: Convert SRA data into Illumina native formats (qseq, etc.)
- abi-dump: Convert SRA data into ABI format (csfasta / qual)
- sff-dump: Convert SRA data to sff format

See more options at
www.ncbi.nlm.nih.gov/Traces/sra/?view=toolkit_doc

Downloads

Users are advised to switch from ftp to aspera for bulk downloads. Aspera provides:

- faster bandwidth
- higher level flow control
- user level encryption
- ability to download trees of components

Bioconductor package

A Bioconductor library is available to help you query information about the SRA datasets and samples.

- SRAdb : A compilation of metadata from NCBI SRA

References

- SRA docs
- SRA NCBI web page
- Wikipedia
- SRA Toolkit