Author: Cameron Ross MacPherson

Date: 2017-11-27
Revision: A

An overview Guide to Using The Sequence Read Archive

The Sequence Read Archive (SRA) is one of several data warehouses maintained by NCBI to ensure the storage, use, and dissemination of knowledge gained from high throughput sequencing experiments. The SRA functions as a raw data store serving unambiguous and unbiased data. This guide will inform the reader of the necessary steps to find, download, and analyze SRA data using Qlucore Omics Explorer (QOE). Several steps are required, and the guide suggests using the SRA toolkit to download data.

STEP 1 - START BY UNDERSTANDING THE SRA HIERARCHY

- All data is organized into **studies**.
- **Studies** contain **samples**.
- **Samples** have ex**periments** performed on them.
- **Experiments** have results recorded by **runs**.
- **Runs** describe results, how, and what made them.
- Altogether, there is a hierarchy of provenance.

STEP 2 - UNDERSTANDING IMPORTANT SRA IDS

- Individual data files are called runs. **Runs** have run IDs beginning with an "SRR" prefix. Runs are described by other IDs such as Biosample.
- Runs are contained in an SRA study.
- An SRA study simply describes the raw data. A similar entity might exist in GEO.
- The GEO study would typically describe analyses. Both datasets would be linked by a BioProject.
- A **BioProject ID** encapsulates an entire study. The screenshot details these IDs for one dataset.

Assay Type:	RNA-Seq
AvgSpotLen:	49
BioProject:	PRJNA216917
Center Name:	GEO
Consent:	public
InsertSize:	0
Instrument:	Illumina HiSeq 2000
LibraryLayout:	SINGLE
LibrarySelection:	cDNA
LibrarySource:	TRANSCRIPTOMIC
LoadDate:	2013-08-23
Organism:	Schizosaccharomyces pombe
Platform:	ILLUMINA
ReleaseDate:	2013-09-20
SRA Study:	SRP029197

STEP 3A - FINDING DATA BY TOPIC

- Be sure to use NCBI's search engine.
- See this SRA specific guide to finding data you need: ncbi.nlm.nih.gov/sra/docs/srasearch/
- Try this query: ((("mus musculus"[Organism]) AND BALB/c*) AND "lymph*") AND "rna seq"[Strategy]
- Use the "Send to" link to download a "File" using the "Accession List" format.
- Go to step 4, review steps 3b and 3c for interest.

STEP 3B - FINDING DATA LISTED IN AN ARTICLE

- Authors usually list accession IDs toward the end of the article just before the reference section.
- If the BioProject ID is listed, copy it and proceed.

Author: Cameron Ross MacPherson

Date: 2017-11-27
Revision: A

STEP 3C - OBTAIN THE SRR IDS YOU WANT

- Make use of the SRA Run Selector here: ncbi.nlm.nih.gov/Traces/study/?go=home
- The tool will accept BioProject, SRA study, and GEO study accessions.
- It will also accept sample, experiment, and run IDs.
- Remember the hierarchy and note that whichever ID you put in, the Run Selector will only report runs back to you that belong to that ID. This is why the parent BioProject ID is so important to record.
- Use the “Facets” section to subselect Runs you want
- If you click on the “Accession List” button to download a text file with all the accessions.

STEP 4 - INSTALL THE SRA TOOLKIT

- You will need the SRA toolkit installed on your computer to download the files.
- Use these instructions for Linux, Mac, or Windows: trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std

STEP 5 - DOWNLOAD AND EXTRACT FASTQ FILES

- The SRA toolkit makes working with SRA intuitive, but requires that you know the SRR IDs beforehand.
- Follow this guideline to become more familiar: ncbi.nlm.nih.gov/sra/docs/sradownload/

STEP 6 - MAP FASTQ FILES TO THE GENOME

- This important step can be done numerous ways. Choose the method you are used to.

STEP 7 - FINISH BY IMPORTING INTO QLUCORE OMICS EXPLORER

- Make use of QOE’s BAM file importer, or use the NGS Module to analyze an entire project with multiple samples and conditions.
- The BioProject ID will allow you to find all data files.
- It is unlikely the BioProject ID is listed. In this case, you will need to find it.
- Without the BioProject ID, you will have to find the SRR IDs for each file manually and risk overlooking some datasets.
- Remember, there is a hierarchy for each accession ID that will help you navigate to the BioProject ID.