Obtaining public datasets from NCBI

SRA Toolkit overview

SRA (**S**equence **R**ead **A**rchive) is an NCBI-defined interchange format for NGS data. The idea is that before submitting your data to NCBI, you convert whatever format it is in (fastq, bam, etc.) to SRA format using one of the "load" tools. Then, the data can be downloaded from NCBI by anyone and extracted in one of a number of different formats as desired (ABI ***csfasta/qual***, ***fastq***).

While this sounds like a great idea (someone else taking care of format interchange issues for you!), the tookit is somwhat obscure and quirky, so in practice it is used mostly to download ***fastq*** files from NCBI. However there is a lot of interesting data out there that's only available as SRAs so it is worthwhile knowing how to use it.

For example, you have aligned a ChIP-seq dataset to **hg19** and have a **.bam** file. You want to upload the data to NCBI. You use the **bam-load** tool:

```
bam-load -o mySRA.sra myAlignment.bam
```

The raw reads can be then be extracted to ***fastq*** using **fastq-dump**:

```
fastq-dump mySRA.sra
```

Looks deceptively simple but you can run into problems. For one thing, SRA toolkit versions change often and are not always compatible. So if you get any weird errors, check for a newer (or sometimes older) toolkit version. The SRA Toolkit documentation, such that it is, is located at the NCBI website.

Finding data

Submissions for a publication generally have the form **SRPnnnn**, with all data under an **a**ccession **SRAnnnn** (the n's have no relation to one another). Data is organized by e**x**periment (**SRXnnnn**) and sequencing **r**un (**SRRnnnn**).

The SRA search home page is where to start looking.

Exercise 1

Find and download RNAseq data from run **SRR390925**, of experiment **SRX112044**, publication **SRP009873**. Copy the file to your home directory on **Lonestar5** at TACC then extract the data in ***fastq*** format.

SRA search home page http://www.ncbi.nlm.nih.gov/sra

wget

module load biocontainers
module spider sratoolkit

fastq-dump

A solution