Week 2, Lecture 4

István Albert

Bioinformatics Consulting Center
Penn State, 2014
SRA – Sequence Read Archive
naming conventions

NCBI BioProject: **PRJN...** (aka SRA study **SRP...**)  
- the overall description of a single research initiative; a project will typically relate to multiple samples and datasets

NCBI BioSample: **SAMN...** (aka SRA Sample **SRS...**)  
- a description of biological source material; each physically unique specimen should be registered as a single BioSample with a unique set of attributes

SRA Experiment: **SRX...**  
- a unique sequencing library for a specific sample

SRA Run: **SRR...**  
- a manifest of data file(s) linked to a given sequencing library (experiment)

This contains the data
That’s not ALL – when it comes to biological data distribution confusion is the rule.

- The **Gene Expression Omnibus also** stores results from functional genomic experiments → but the raw data links back to SRA.

- GEO was originally designed for microarray data, later augmented for high throughput sequencing

- These organizations appear to be monolithic and it is not clear what entity is responsible for them, who makes what decisions and why.

- This is why groups of scientists want to form their own independently run information repositories.
GEO nomenclature

Words that start with G usually refer to GEO:

- **GPL**... will be a platform
- **GSM**... indicates a sample
- **GSE**... indicates a series

Then it all links back to SRA (hopefully)
Getting data from SRA

- You will need to install a software package called **sra-toolkit**

- This package can fetch and unpack data from SRA

- One of the skills that you need to develop is that of installing tools just by their name. Find the **sra toolkit**, download and install it.
Short diversion: making programs run

- One way to enable a program to run on UNIX is to add the path to the program to the `PATH` variable of your system:

  ```
  export PATH=$PATH:/src/somefolder/bin
  ```

To have this automatically applied every time you start a new terminal, append this line to the startup configuration file:

- `.profile` on a Mac
- `.bashrc` on Linux

(files that start with `.` are so called “hidden” files that don’t show up by default, this is to avoid accidentally changing them)

To show all files use: `ls -a`
You’ll need a “decent” text editor

Absolutely essential features:

• Needs to be able to show white-space (allow you to distinguish between tabs and spaces)

• Needs to be able to show line numbers
I use PyCharm for work – but it is a more complex and complicated/powerful tool.

During lectures we will use Komodo Edit.

You may use any other editor as long as you understand how they work.
For our own sanity (too keep it) it is essential that we are able to distinguish between spaces and tabs in a file.

Some editors will happily convert between the two without telling us so.

Extremely subtle or hard to trace errors may occur from tab/space mixups.
Set up your system

• Use your editor to add the Entrez Direct and the sra-toolkit to your system path

• How to know you were successful?

• Open a new terminal. The programs `esearch`, `prefetch`, `fastq-dump` should run as typed.
Let’s download some data

• Find the bioproject PRJNA257197, choose an experiment and a run (SRR) file.

The rest of the commands are in the code repository.
Homework 4

- Download and unpack at least five SRR runs.
  (a single FASTQ data unit (FASTQ sequence record) consist of four lines that start with the @SRR pattern)

- How many sequences are in each run? Check the number for at least one run via SRA.

- What does the following command do:
  
  ```bash
  fastq-dump -X 10 -Z SRR1553610
  ```