Large Data Analysis

- **HTS (High Throughput Sequencing) datasets:**
  - **RNA-Seq**
  - **ChIP-Seq**
  - **CLIP-Seq, ATAC-Seq, Microbiome-Seq etc.**
- **Other large datasets:**
  - **Proteome datasets**
  - **Literature-based and derived datasets**
  - **…**

Source of HTS data

- Your own (sequencing service).
- Public databases, such as NCBI/GEO.
- Major genomic /epigenomic projects, such as ENCODE (ENCylopedia Of DNA Elements); the Cancer Genome Project, etc.
- Other internet sources.

Retrieving HTS data

- Retrieving HTS data from the web using wget.

- Loading to and unloading data from UFHPC - check with HPC (HiPerGator) instructions.

Retrieval of information.

- Using web interface.
- Using FTP client

- Using command line tools.
  - Generic Linux file transfer tools - always available in Linux/MacOs.
  - Specialized tool – fastq-dump
  - Let the script do the job – when you need large amount of files or large file that will take hours to download.

Practice: log into UFHPC / Linux server.

Mac user, type in terminal:

$ ssh username@hpg2.rc.ufl.edu

Windows, Open in Putty:

hpg2.rc.ufl.edu

once you are in, move to your working dir:

>cd /blue/gms6014/share/<firstname>

HTS – Download dataset

- Commend line (with a few samples):
  - $ module load sra
  - $ fastq-dump --gzip SRRxxxx SRRyyyy
- With the .sbatch job file (for large data set)
  - $sbatch myjob.sbatch
  - Use “$ squeue –u <yourUserName>” to monitor progress.
  - Use “$ls –l” to make sure files size are correct. (Use checksum to verify)

HTS – Download dataset

```
#!/bin/sh
#SBATCH --job-name=GetSRA
#SBATCH --mail-type=ALL
#SBATCH --mail-user=xxxxx@ufl.edu
#SBATCH --output=GetSRA_%j.log
#SBATCH -t 12:00:00
#SBATCH --cpus-per-task=1
#SBATCH --mem-per-cpu=3gb

pwd; date

module load sra/2.10.3

fastq-dump --gzip SRR1618640 SRR1618641 SRR1618642 SRR1618643
```

Transfer the file to your folder in HiPerGator and submit the job ($sbatch *filename*)

Large Data Set Analysis.

Hardware considerations:

1.) Data storage.

- FASTA record of a protein (1,000 aa) ~ 1 **KB**.
- Human proteome, or Chromosome 21 ~ 50 **MB**
- Human genome ~ 1.5 **GB**
- **HTS transcriptome analysis (4 samples @ 40 million reads each) original and derived data sets ~ 200 GB**

Large Data Set Analysis.

Hardware considerations:

2.) Processors and RAM.

- Comparison: tbalstn of 5 protein sequences against 1.2GB genome, ~15 sec CPU time. Map a single 10 M reads illumina run to human genome ~15,000 CPU sec (> 4 hours).

- RAM < data size will greatly slow down the process.

Large Data Set Analysis.

Hardware considerations:

3.) Operating system determines the availability of tools.

- Linux is the default development system for most bioinformatics groups. It is also the OS of the UFHPC.
- Easy control and automation.
- Portable to Mac OSX, but some requires recompiling the source code.

Preparation for HTS project

- Make a folder in your local computer for
  - The GSE web page.
  - Files associated the project (RunInfo, Summary, Accession list).
  - Job files, results, etc.
- Make a folder (can be of the same name) in your working directory on HPC.
- **Use script/job files for analysis – keep it as a record**.