#2.1: Scripting and data preparation

SRA-toolkit

Scripting

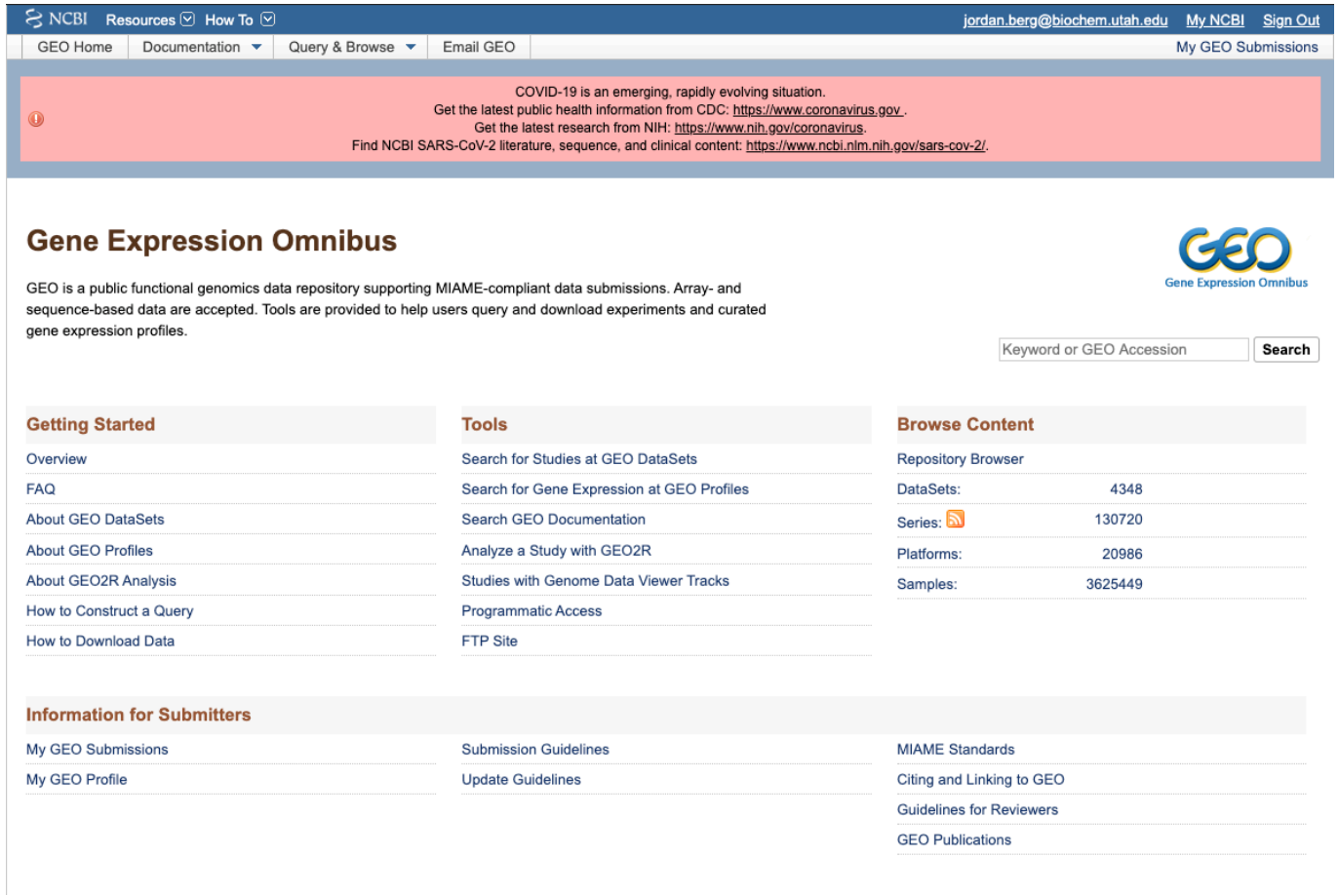
Downloading source files

Access the slides and files here:

https://github.com/j-berg/bioinformatics_bootcamp

Downloading publicly available data

Any published NIH-funding sequencing datasets *should* appear here



Downloading publicly available data

- SRA-toolkit
  - Download and unpackage

```
$ curl -OL http://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-centos_linux64.tar.gz
$ tar -zxvf sratoolkit.current-centos_linux64.tar.gz
```

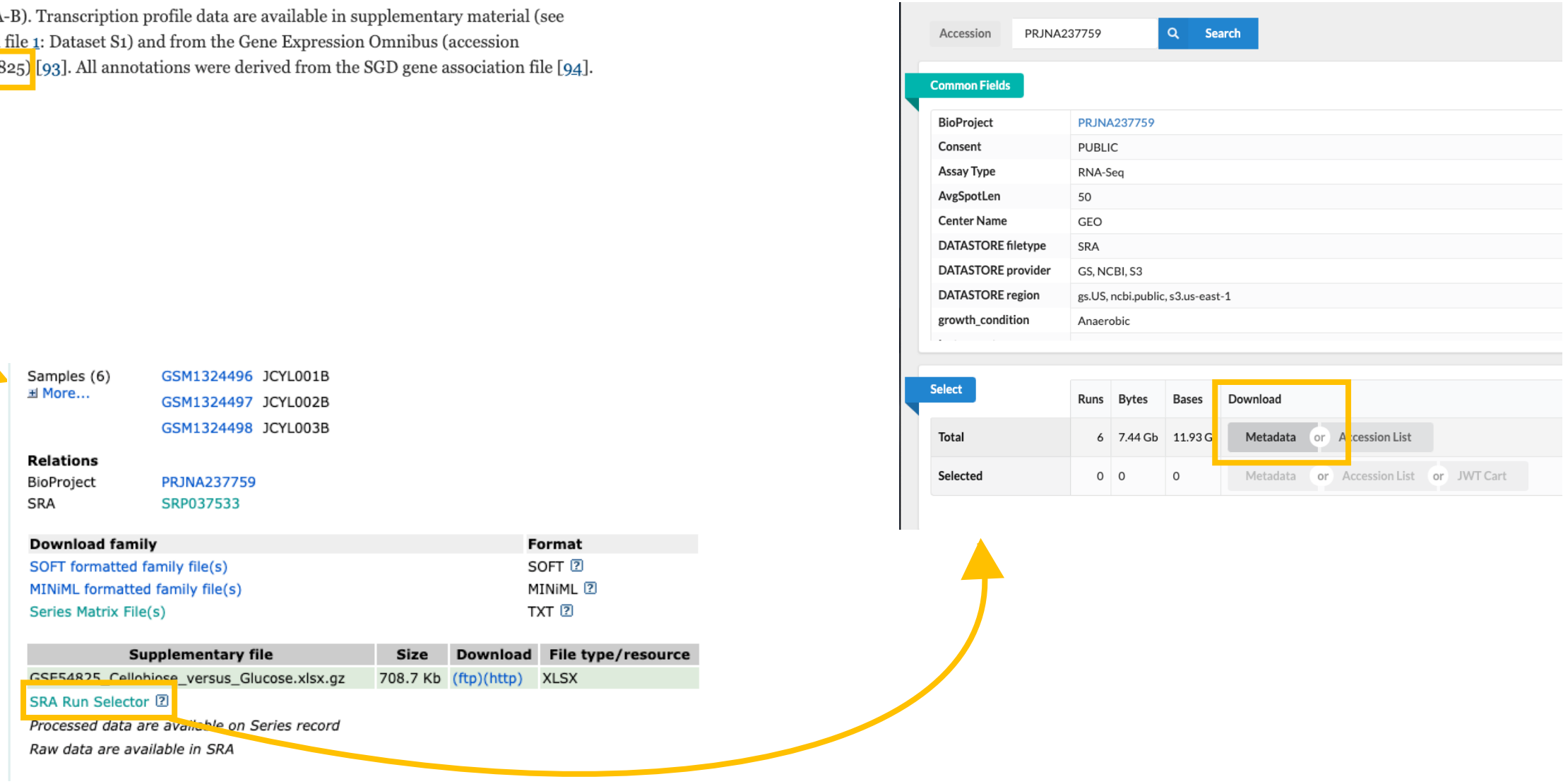
  - Rename

```
$ mv sratoolkit.2.10.7-centos_linux64 sratoolkit
```

  - Make SRA-toolkit findable to you anywhere in supercomputer

```
$ echo 'export PATH="/uufs/chpc.utah.edu/common/home/u0690617/sratoolkit/bin:$PATH"' >> ~/.bashrc
```

Figure S7A-B). Transcription profile data are available in supplementary material (see Additional file 1: Dataset S1) and from the Gene Expression Omnibus (accession no. GSE54825) [93]. All annotations were derived from the SGD gene association file [94].



Download a file by SRR accessor

```
(base) [u       @notchpeak1 ~]$ prefetch SRR1166442

2020-06-07T13:26:44 prefetch.2.9.3: 1) Downloading 'SRR1166442'...
2020-06-07T13:26:44 prefetch.2.9.3:  Downloading via https...
2020-06-07T13:28:12 prefetch.2.9.3: 1) 'SRR1166442' was downloaded successfully
(base) [u       @notchpeak1 ~]$ fastq-dump --outdir ~/ --split-files /uufs/chpc.utah.edu/common/home/
   /ncbi/public/sra/SRR1166442.sra
Read 40585197 spots for /uufs/chpc.utah.edu/common/home/u       /ncbi/public/sra/SRR1166442.sra
Written 40585197 spots for /uufs/chpc.utah.edu/common/home/u       /ncbi/public/sra/SRR1166442.sra
```

Writing your first bash script

```
$ vim download_seqs.sh
```
<-- let's write a new bash script (.sh)

Commands:

- You will start in viewer mode
- Tap "d" twice = delete a line
- Arrow keys to move cursor
- Tap "i" to enter editor mode
- Write your code
- Press escape to exit editor mode
- Type ":wq" to save (write) and quit
- Type ":q" to quit without saving
- If you made changes and want to quit without saving, you will need to use ":q!"

Downloading sequencing files

```
$ vim download_seq.sh
```

Press “i”

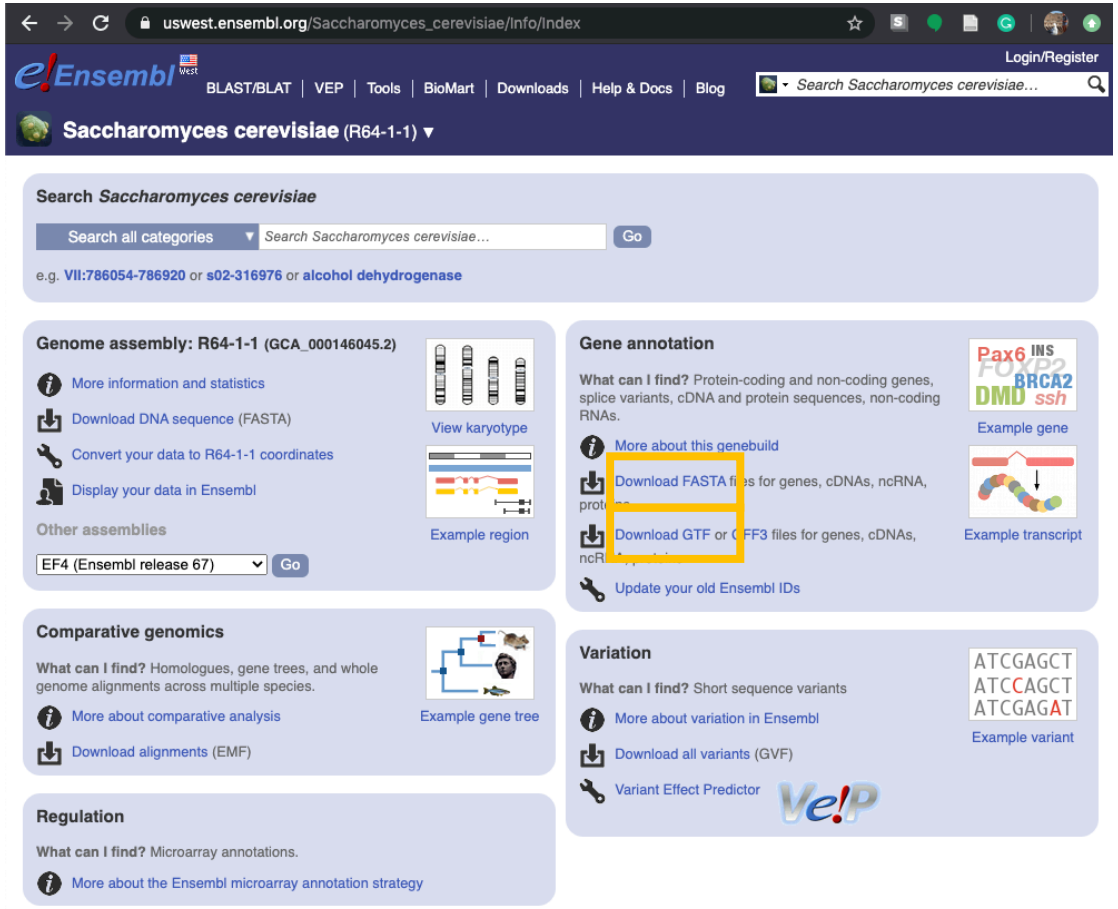
```
SRR_IDS=(SRR1166442 SRR1166443 SRR1166444 SRR1166445 SRR1166446 SRR1166447)
OUTPUT=~/seq_files
PRE_LOC=~/ncbi/public/sra

mkdir $OUTPUT
for X in ${SRR_IDS[@]}; do prefetch ${X}; done
for X in ${SRR_IDS[@]}; do fastq-dump --outdir $OUTPUT    --split-files $PRE_LOC/${X}.sra; done
```

Press Escape + “wq”, then Enter

Downloading sequencing files

```
(base) [u0690617@notchpeak1 ~]$ bash download_seq.sh

2020-06-07T14:33:11 prefetch.2.9.3: 1) Downloading 'SRR1166442'...
2020-06-07T14:33:11 prefetch.2.9.3:  Downloading via https...
2020-06-07T14:34:33 prefetch.2.9.3: 1) 'SRR1166442' was downloaded successfully

2020-06-07T14:34:35 prefetch.2.9.3: 1) Downloading 'SRR1166443'...
2020-06-07T14:34:35 prefetch.2.9.3:  Downloading via https...
2020-06-07T14:36:06 prefetch.2.9.3: 1) 'SRR1166443' was downloaded successfully

2020-06-07T14:36:10 prefetch.2.9.3: 1) Downloading 'SRR1166444'...
2020-06-07T14:36:10 prefetch.2.9.3:  Downloading via https...
2020-06-07T14:37:16 prefetch.2.9.3: 1) 'SRR1166444' was downloaded successfully

2020-06-07T14:37:18 prefetch.2.9.3: 1) Downloading 'SRR1166445'...
2020-06-07T14:37:18 prefetch.2.9.3:  Downloading via https...
```

Downloading reference files



Downloading reference files

$ vim download_refs.sh

Press “i”

```
# Set variables
OUTPUT=~/reference_yeast
GTF_URL=ftp://ftp.ensembl.org/pub/release-100/gtf/saccharomyces_cerevisiae/Saccharomyces_cerevisiae.R64-1-1.100.gtf.gz
FASTA_PATH_URL=ftp://ftp.ensembl.org/pub/release-100/fasta/saccharomyces_cerevisiae/dna
CHR=(I II III IV V VI VII VIII IX X XI XII XIII XIV XV XVI)

# Make output and navigate into folder
mkdir $OUTPUT
cd $OUTPUT

# Download GTF
curl -OL $GTF_URL

# Download all Chromosome FASTA files
for X in ${CHR[@]}; do curl -OL $FASTA_PATH_URL/Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.${X}.fa.gz; done

# Decompress all reference files and return to home directory
gzip -d *.gz
cd ~
```

Press Escape + “wq”, then Enter

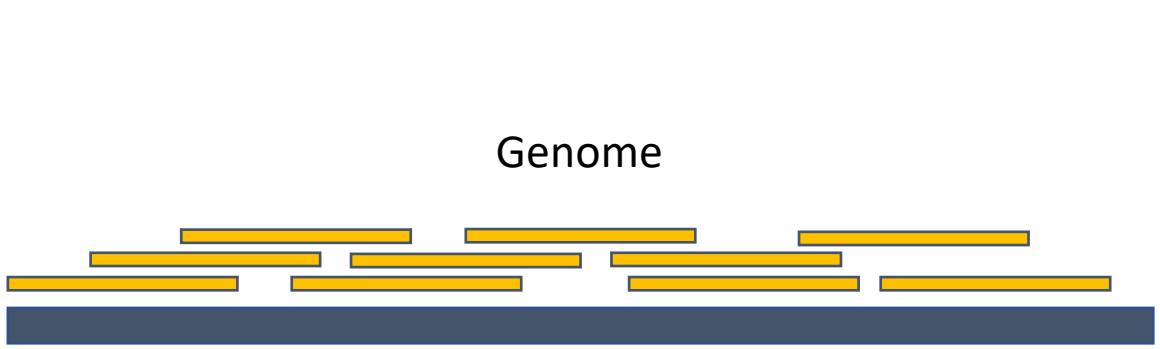
Downloading reference files

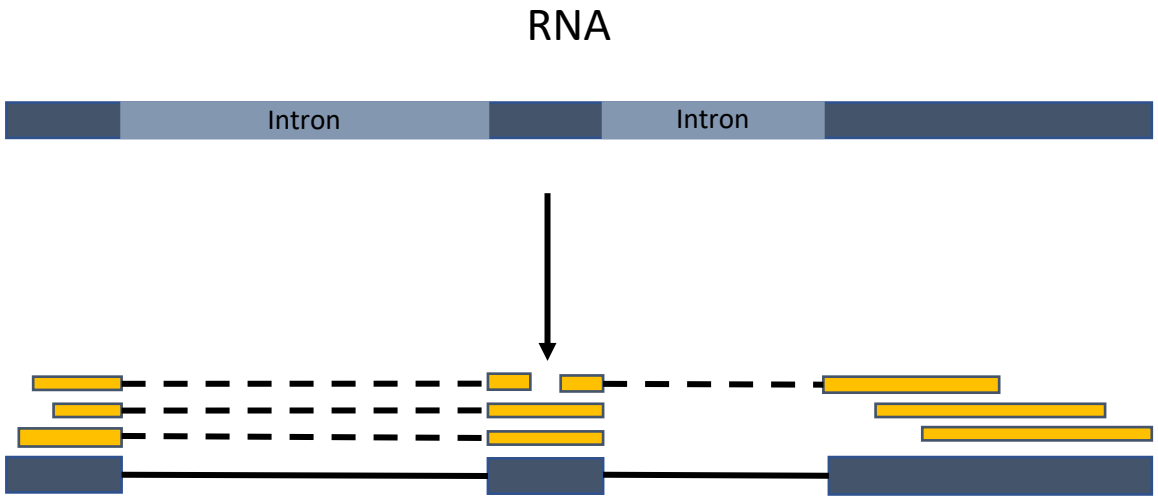
```
(base) [u0690617@notchpeak1 ~]$ bash download_refs.sh
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  539k  100  539k    0     0   143k      0  0:00:03  0:00:03 --:--:--  143k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 71530  100 71530    0     0  32662      0  0:00:02  0:00:02 --:--:-- 32677
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  248k  100  248k    0     0    99k      0  0:00:02  0:00:02 --:--:--   99k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 99014  100 99014    0     0  45108      0  0:00:02  0:00:02 --:--:-- 45108
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  464k  100  464k    0     0   169k      0  0:00:02  0:00:02 --:--:--  169k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  176k  100  176k    0     0  76418      0  0:00:02  0:00:02 --:--:-- 76451
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 84644  100 84644    0     0  38509      0  0:00:02  0:00:02 --:--:-- 38492
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  332k  100  332k    0     0   128k      0  0:00:02  0:00:02 --:--:--  128k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  171k  100  171k    0     0  74883      0  0:00:02  0:00:02 --:--:-- 74883
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  134k  100  134k    0     0  58845      0  0:00:02  0:00:02 --:--:-- 58845
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  226k  100  226k    0     0  93451      0  0:00:02  0:00:02 --:--:-- 93451
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  203k  100  203k    0     0  89049      0  0:00:02  0:00:02 --:--:-- 89087
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  323k  100  323k    0     0   132k      0  0:00:02  0:00:02 --:--:--  132k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  281k  100  281k    0     0   115k      0  0:00:02  0:00:02 --:--:--  115k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  239k  100  239k    0     0  99193      0  0:00:02  0:00:02 --:--:-- 99234
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  333k  100  333k    0     0   128k      0  0:00:02  0:00:02 --:--:--  128k
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  289k  100  289k    0     0  116k      0  0:00:02  0:00:02 --:--:--  116k
```



```
(base) [u0690617@notchpeak1 ~]$ cd reference_yeast/
(base) [u0690617@notchpeak1 reference_yeast]$ ls
Saccharomyces_cerevisiae.R64-1-1.100.gtf
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.I.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.II.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.III.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.IV.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.IX.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.V.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VI.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VII.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.VIII.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.X.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XI.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XII.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XIII.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XIV.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XV.fa
Saccharomyces_cerevisiae.R64-1-1.dna.chromosome.XVI.fa
```

Genome vs RNA-seq

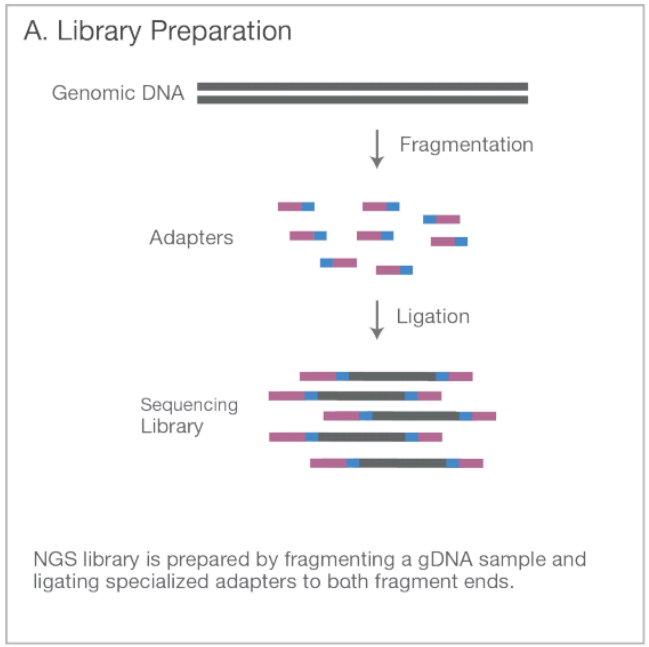


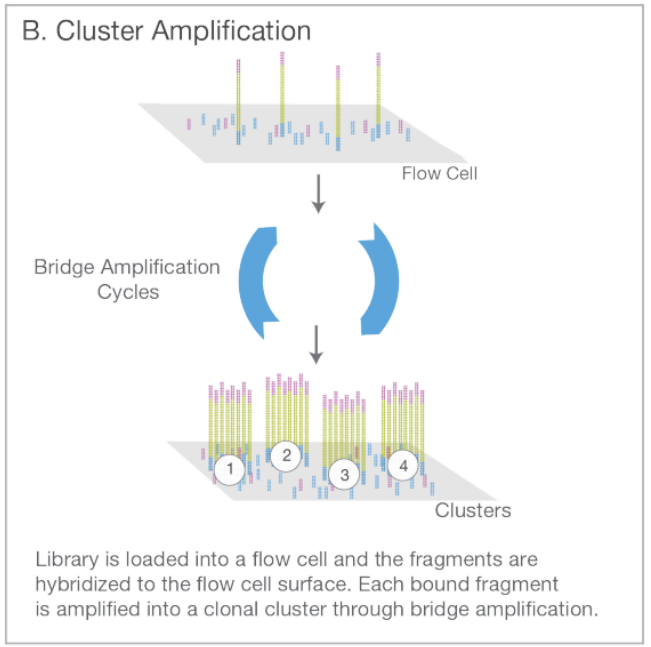


Different conditions sometimes call for different tools

Steps: Pre-processing

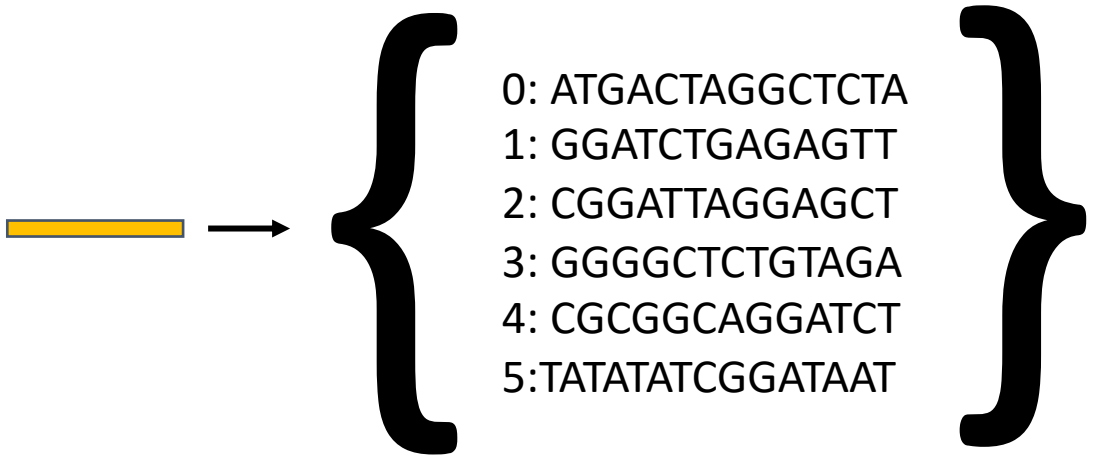
- Fastp
  - Remove adapters (will prevent alignment with synthetic sequence)
  - Remove low quality bases and short reads

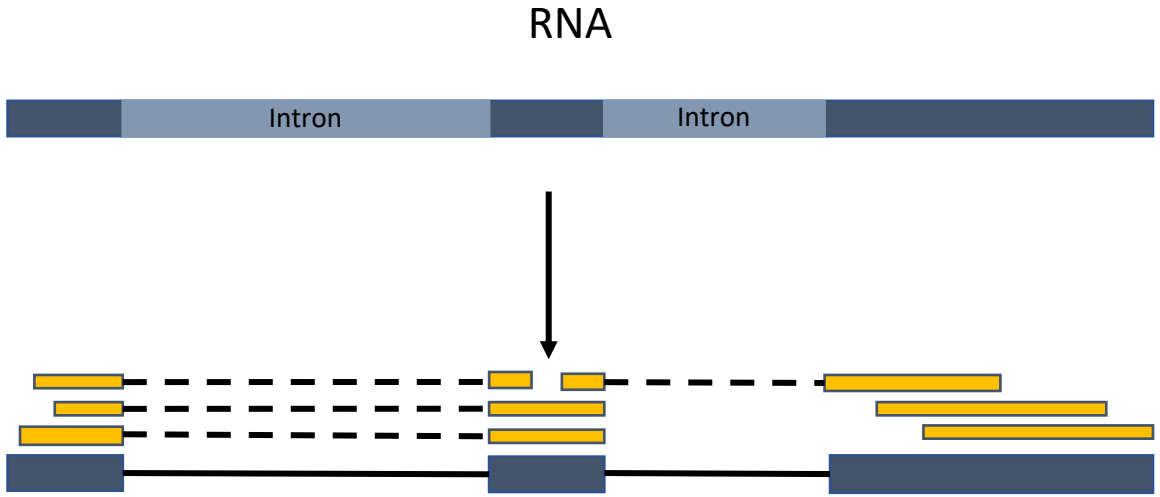




Steps: Splice-aware Alignment

- STAR
  - Generate a genome index
  - Perform splice-aware alignment of sequencing reads to genome index





Steps: Post-processing

- Samtools
  - Sort alignments and index

Steps: Quantification

- HTSeq-count
  - Generate read counts for each gene

Gene A:

6 reads = 6 counts

Steps: Quality Control

- FASTQC
- MultiQC
  - Make sure sequencing library is high-quality and reliable

Downloading required software for RNA-seq

- Make sure channel priority is set

```
$ conda activate class
$ conda install fastp STAR samtools htseq multiqc fastqc
```

Homework:

- Find an RNA-seq GEO dataset with 8 or fewer total samples
- Create a new folder for the dataset
- Copy the class sequence file download script and modify as needed to download each of the files from the GEO dataset
- Copy one of the files with a new name and decompress the file
- Use the head command to determine the average length of the first 10 reads of this copied file
- Delete the copied file (but do not remove the original downloads)

- To a new directory in your home folder on CHPC, download the GTF and genome FASTA files for the organism used in your downloaded RNA-seq dataset