# #1.2: Bioinformatics File Types and their manipulation

# Access the slides and files here:

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

# Package Managers

- Allow you to quickly download pre-compiled software

1. Download the installer

   $ curl -OL => download file as is, make sure to not just download a link but the actual file

2. Run the installer

   $ bash => run shell script (.sh)

3. Accept the license and let it set the PATH for you

~ => shortcut for home directory

```
(base) [u       @notchpeak2 ~]$ curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 81.1M  100 81.1M    0     0  87.3M      0 --:--:-- --:--:-- --:--:-- 87.2M
(base) [u       @notchpeak2 ~]$ bash ~/Miniconda3-latest-MacOSX-x86_64.sh
bash: /uufs/chpc.utah.edu/common/home/u       /Miniconda3-latest-MacOSX-x86_64.sh: No such file or directory
(base) [u       @notchpeak2 ~]$ bash ~/Miniconda3-latest-Linux-x86_64.sh

Welcome to Miniconda3 4.8.2

In order to continue the installation process, please review the license
agreement.
Please, press ENTER to continue
>>>
```

# Use conda to install a package

Make an environment to avoid package clashing:

$ conda create --name class

Activate the environment:

$ conda activate class

Will need to do this every time you log onto the supercomputer if you want to use the environment

```
(base) [u       @notchpeak2 ~]$ conda env create -n test

SpecNotFound: Invalid name, try the format: user/package

(base) [u       @notchpeak2 ~]$ conda create --name class
Collecting package metadata (current_repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
  current version: 4.8.2
  latest version: 4.8.3

Please update conda by running

    $ conda update -n base -c defaults conda



## Package Plan ##

  environment location: /uufs/chpc.utah.edu/common/home/u       /miniconda3/envs/class



Proceed ([y]/n)? y

Preparing transaction: done
Verifying transaction: done
Executing transaction: done
#
# To activate this environment, use
#
#     $ conda activate class
#
# To deactivate an active environment, use
#
#     $ conda deactivate
```

# Use conda to install a package

Install a package to the environment:

$ conda install -c bioconda samtools

-c tells conda the channel to search for the package in

```
(class) [u       @notchpeak2 ~]$ conda config --set channel_priority true
(class) [u       @notchpeak2 ~]$ conda config --add channels bioconda
(class) [u       @notchpeak2 ~]$ conda config --add channels conda-forge
(class) [u       @notchpeak2 ~]$ conda config --get channels
--add channels 'defaults'   # lowest priority
--add channels 'bioconda'
--add channels 'conda-forge'   # highest priority
```

```
(class) [u       @notchpeak2 ~]$ conda install -c bioconda samtools
Collecting package metadata (current_repodata.json): done
Solving environment: done


==> WARNING: A newer version of conda exists. <==
  current version: 4.8.2
  latest version: 4.8.3

Please update conda by running

    $ conda update -n base -c defaults conda



## Package Plan ##

  environment location: /uufs/chpc.utah.edu/common/home/u       /miniconda3/envs/class

  added / updated specs:
    - samtools
```

```
Proceed ([y]/n)? y


Downloading and Extracting Packages
libssh2-1.9.0        | 298 KB    | ############################################################################
libstdcxx-ng-9.2.0   | 4.5 MB    | ############################################################################
libedit-3.1.20170329 | 172 KB    | ############################################################################
libcurl-7.69.1       | 573 KB    | ############################################################################
bzip2-1.0.8          | 396 KB    | ############################################################################
htslib-1.9           | 1.2 MB    | ############################################################################
ca-certificates-2020 | 146 KB    | ############################################################################
krb5-1.17.1          | 1.5 MB    | ############################################################################
ncurses-6.1          | 1.3 MB    | ############################################################################
xz-5.2.5             | 430 KB    | ############################################################################
libdeflate-1.2       | 63 KB     | ############################################################################
curl-7.69.1          | 137 KB    | ############################################################################
zlib-1.2.11          | 105 KB    | ############################################################################
openssl-1.1.1g       | 2.1 MB    | ############################################################################
Preparing transaction: done
Verifying transaction: done
Executing transaction: done
```

# Accessing a tools documentation

- Standard command line tools:

$ man wc

- Other tools:

$ samtools --help

$ samtools view --help

```
(class) [u       @notchpeak2 ~]$ samtools --help

Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)

Usage:   samtools <command> [options]

Commands:
  -- Indexing
     dict           create a sequence dictionary file
     faidx          index/extract FASTA
     fqidx          index/extract FASTQ
     index          index alignment

  -- Editing
```

```
(class) [u       @notchpeak2 ~]$ samtools view --help
samtools view: unrecognised option '--help'

Usage: samtools view [options] <in.bam>|<in.sam>|<in.cram> [region ...]

Options:
  -b       output BAM
  -C       output CRAM (requires -T)
  -1       use fast BAM compression (implies -b)
  -u       uncompressed BAM output (implies -b)
  -h       include header in SAM output
  -H       print SAM header only (no alignments)
  -c       print only the count of matching records
```

# Connecting tasks

- The Pipe
- $ samtools view filename | wc -l | head
- Do this, then this, then this…

# Moving and renaming files

$ mv

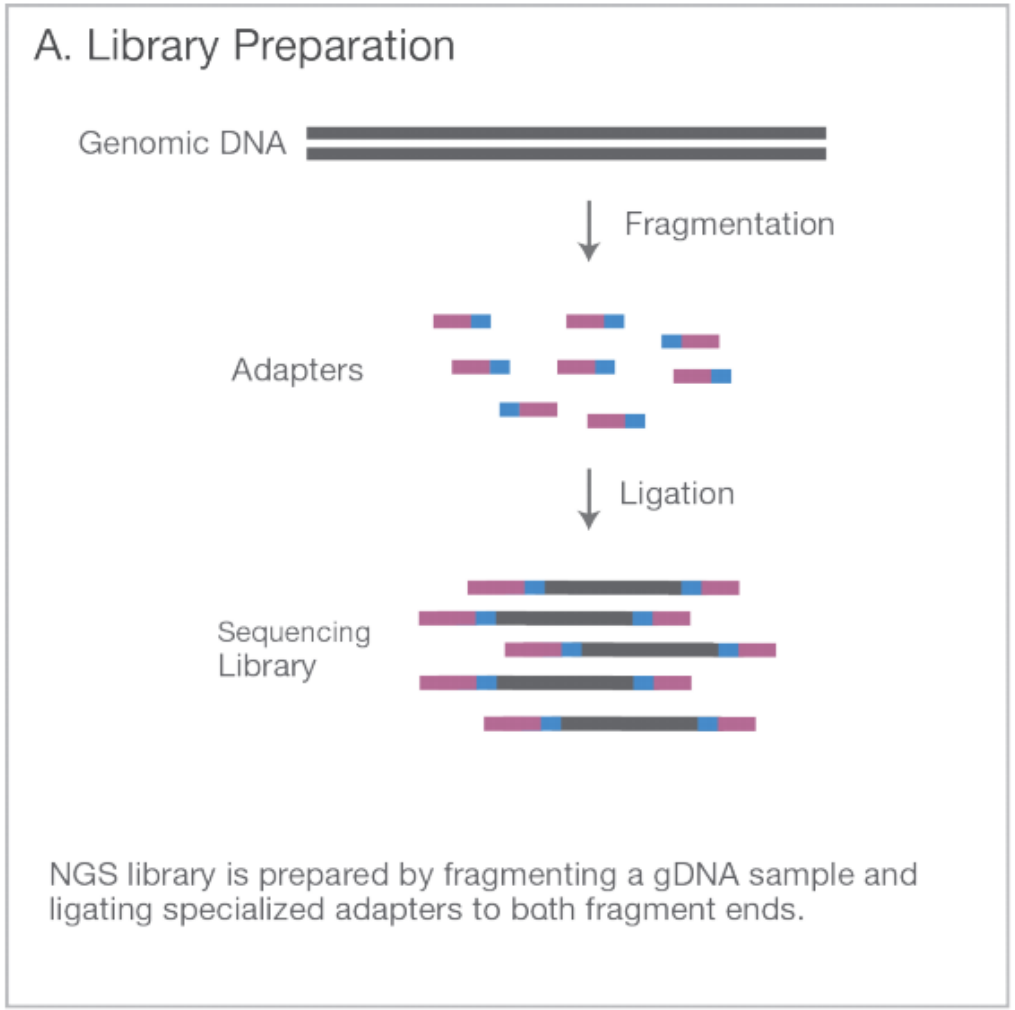
```
(class) [u       @notchpeak2 test]$ mv Saccharomyces_cerevisiae.R64-1-1.100.gtf home/
(class) [u       @notchpeak2 test]$ cd home/
(class) [u       @notchpeak2 home]$ ls
total 9.2M
drwxr-xr-x 3 u         rutter   48 May 30 10:56 han
drwxr-xr-x 2 u         rutter   10 May 30 10:31 leia
drwxr-xr-x 3 u         rutter   45 May 30 10:31 luke
-rw-r--r-- 1 u         rutter 9.2M May 30 12:11 Saccharomyces_cerevisiae.R64-1-1.100.gtf
(class) [u       @notchpeak2 home]$ mv Saccharomyces_cerevisiae.R64-1-1.100.gtf yeast_transcripts.gtf
(class) [u       @notchpeak2 home]$ ls
total 9.2M
drwxr-xr-x 3 u         rutter   48 May 30 10:56 han
drwxr-xr-x 2 u         rutter   10 May 30 10:31 leia
drwxr-xr-x 3 u         rutter   45 May 30 10:31 luke
-rw-r--r-- 1 u         rutter 9.2M May 30 12:11 yeast_transcripts.gtf
```

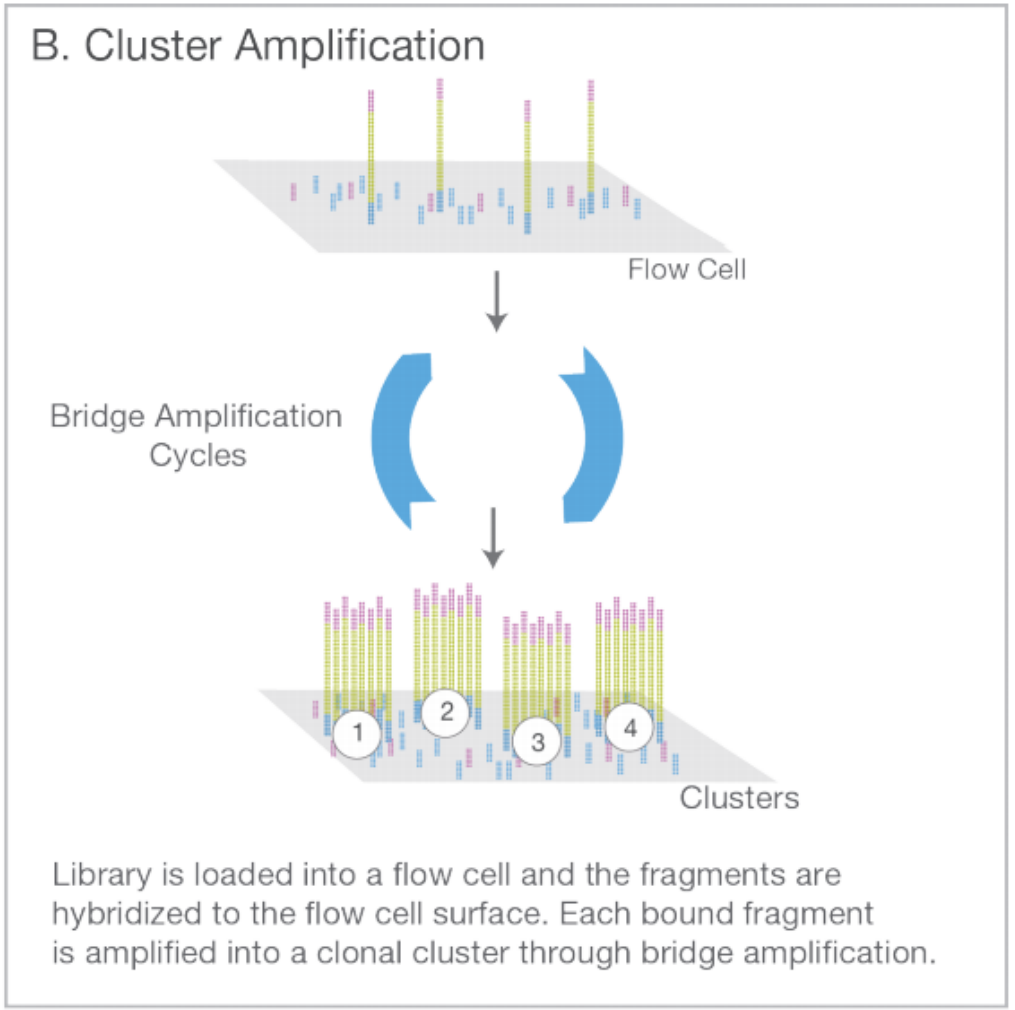
# Copying Files

$ cp

```
(class) [u       @notchpeak2 home]$ cp yeast_transcripts.gtf copy.gtf
(class) [u       @notchpeak2 home]$ ls
total 20M
-rw-r--r-- 1 u         rutter 9.2M May 30 12:25 copy.gtf
drwxr-xr-x 3 u         rutter   48 May 30 10:56 han
drwxr-xr-x 2 u         rutter   10 May 30 10:31 leia
drwxr-xr-x 3 u         rutter   45 May 30 10:31 luke
-rw-r--r-- 1 u         rutter 9.2M May 30 12:11 yeast_transcripts.gtf
(class) [u       @notchpeak2 home]$ cp copy.gtf luke/copy2.gtf
(class) [u       @notchpeak2 home]$ cd l
leia/ luke/
(class) [u       @notchpeak2 home]$ cd leia/
(class) [u       @notchpeak2 leia]$ cd ../luke/
(class) [u       @notchpeak2 luke]$ ls
total 10M
-rw-r--r-- 1 u         rutter 9.2M May 30 12:25 copy2.gtf
-rw-r--r-- 1 u         rutter    0 May 30 10:31 hi.txt
drwxr-xr-x 2 u         rutter   52 May 30 10:31 proj1
```

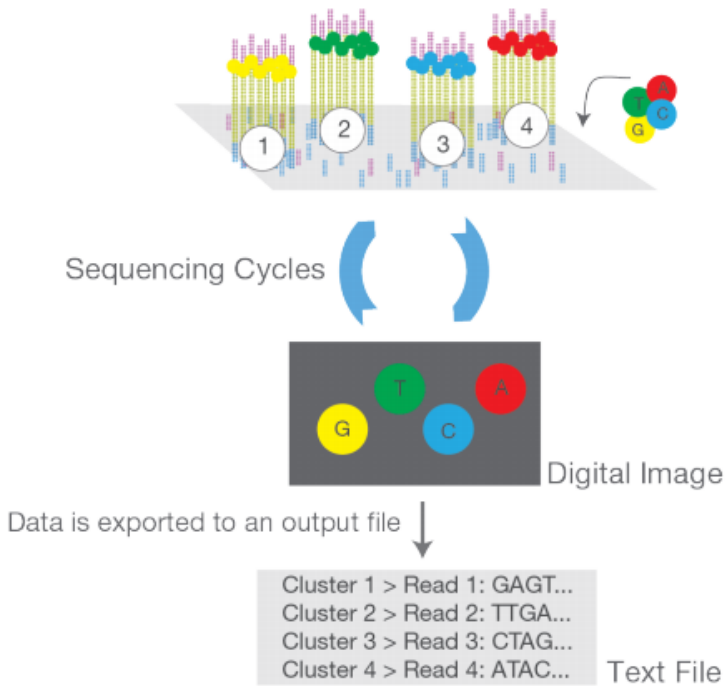
# Sequencing



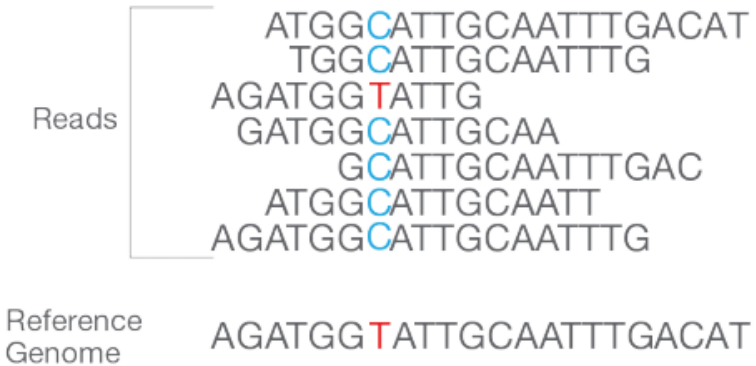

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

## C. Sequencing




Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated "n" times to create a read length of "n" bases.

## D. Alignment and Data Anaylsis




Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

# Bioinformatics File Types

# FASTA

- A nucleotide sequence file
- > indicates record ID/name
- Second line is the sequence
- One file can have this pattern repeat for multiple FASTA records

>record
ATATGTGTATACTCTATAGAGAGGATCTAGAGTATAGC
TCGCGTATAGAGATCTTCGCGATATAGAGAGTCTGCG
AAGGCTCTCGCGCGCAAAGAGAGAGATATTCGCGC

# FASTQ

- A sequencing file with paired nucleotide sequence and quality score for each read

@SRR2075930.2:UMI_NTGCG HS1:450:C5WTEACXX:7:1101:1276:2113 length=50
CTACGTGTGGAGGCTCANGCAGCGCTTCTGGCTGGAACGGGGAA
+
:@<??@?:>??<=??@>#189==?????>?<??<??>??55985

Line 1: read ID (will start with an @)
Line 2: read sequence
Line3: spacer (sometimes will have the the read ID repeated)
Line4: The corresponding PHRED score

# PHRED

- Base call confidence score
- ASCII characters alongside the read sequences

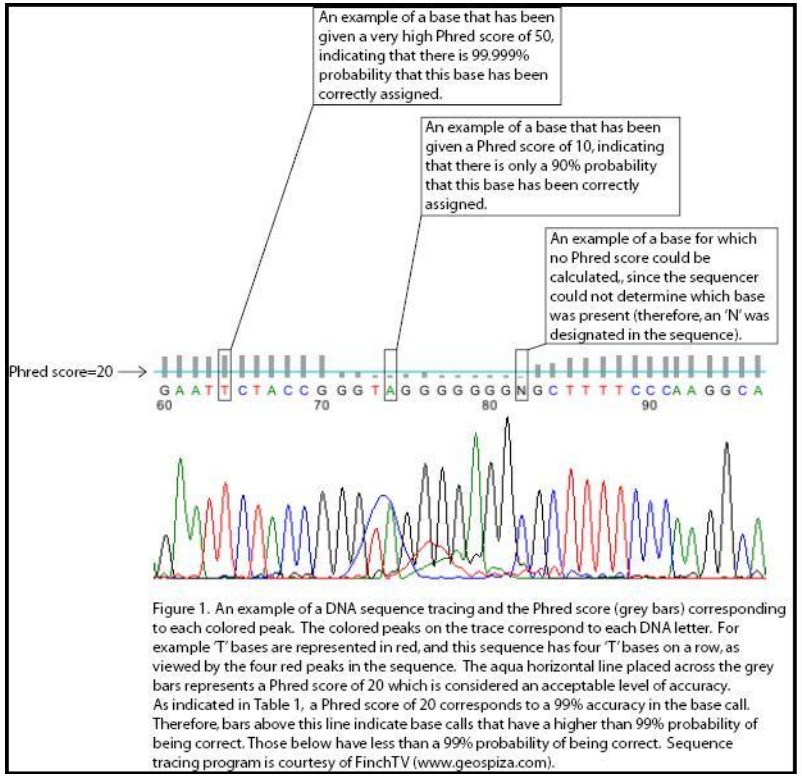



Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

https://en.wikipedia.org/wiki/Phred_quality_score
https://en.wikipedia.org/wiki/Fred_Flintstone#/media/File:Fred_Flintstone.png

# SAM/BAM/CRAM file

- Sequence Alignment Map
- Binary Alignment Map
  - YOU NEED SAMTOOLS TO READ BAM FILES
- CRAM: lossy compression BAM file
- Tab-delimited

- To view BAM files, we need samtools (downloaded previously)

```
$ samtools view sample.bam | column -t -s $'\t' | less -N -S
```

separate out data by tabs

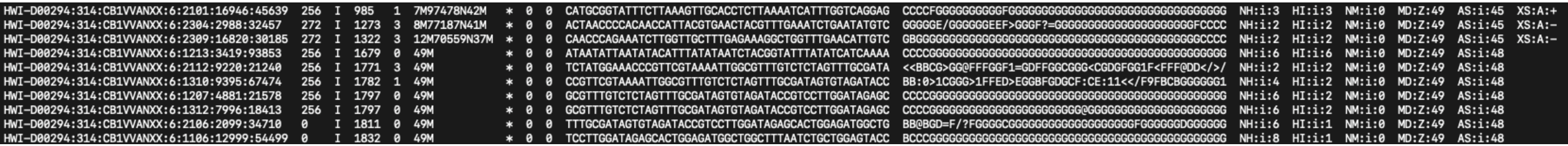
# SAM/BAM file

Flags indicating mapping quality

Leftmost mapping position

Mate read info (for paired-end)

Other metadata



```
HWI-D00294:314:CB1VVANXX:6:2101:16946:45639  256  I  985   1  7M97478N42M  *  0  0  CATGCGGTATTTCTTAAAGTTGCACCTCTTAAAATCATTTGGTCAGGAG  CCCCFGGGGGGGGGGFGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG  NH:i:3  HI:i:3  NM:i:0  MD:Z:49  AS:i:45  XS:A:+
HWI-D00294:314:CB1VVANXX:6:2304:2988:32457   272  I  1273  3  8M77187N41M  *  0  0  ACTAACCCCACAACCATTACGTGAACTACGTTTGAAATCTGAATATGTC  GGGGGE/GGGGGGEEF>GGGF?=GGGGGGGGGGGGGGGGGGGFCCCC  NH:i:2  HI:i:2  NM:i:0  MD:Z:49  AS:i:45  XS:A:-
HWI-D00294:314:CB1VVANXX:6:2309:16820:30185  272  I  1322  3  12M70559N37M  *  0  0  CAACCCAGAAATCTTGGTTGCTTTGAGAAAGGCTGGTTTGAACATTGTC  GBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGCCCC  NH:i:2  HI:i:2  NM:i:0  MD:Z:49  AS:i:45  XS:A:-
HWI-D00294:314:CB1VVANXX:6:1213:3419:93853   256  I  1679  0  49M          *  0  0  ATAATATTAATATACATTTATATAATCTACGGTATTTATATCATCAAAA  CCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG  NH:i:6  HI:i:6  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:2112:9220:21240   256  I  1771  3  49M          *  0  0  TCTATGGAAACCCGTTCGTAAAATTGGCGTTTGTCTCTAGTTTGCGATA  <<BBCG>GG@FFFGGF1=GDFFGGCGGG<CGDGFGG1F<FFF@DD</>/  NH:i:2  HI:i:2  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:1310:9395:67474   256  I  1782  1  49M          *  0  0  CCGTTCGTAAAATTGGCGTTTGTCTCTAGTTTGCGATAGTGTAGATACC  BB:0>1CGGG>1FFED>EGGBFGDGCF:CE:11<</F9FBCBGGGGGG1  NH:i:4  HI:i:2  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:1207:4881:21578   256  I  1797  0  49M          *  0  0  GCGTTTGTCTCTAGTTTGCGATAGTGTAGATACCGTCCTTGGATAGAGC  CCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG  NH:i:6  HI:i:2  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:1312:7996:18413   256  I  1797  0  49M          *  0  0  GCGTTTGTCTCTAGTTTGCGATAGTGTAGATACCGTCCTTGGATAGAGC  CCCCGGGGGGGGGGGGGGGGGGGGGGG@GGGGGGGGGGGGGGGGGGGG  NH:i:6  HI:i:2  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:2106:2099:34710   0    I  1811  0  49M          *  0  0  TTTGCGATAGTGTAGATACCGTCCTTGGATAGAGCACTGGAGATGGCTG  BB@BGD=F/?FGGGGCGGGGGGGGGGGGGGGGGGFGGGGGGDGGGGGG  NH:i:6  HI:i:1  NM:i:0  MD:Z:49  AS:i:48
HWI-D00294:314:CB1VVANXX:6:1106:12999:54499  0    I  1832  0  49M          *  0  0  TCCTTGGATAGAGCACTGGAGATGGCTGGCTTTAATCTGCTGGAGTACC  BCCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG  NH:i:8  HI:i:1  NM:i:0  MD:Z:49  AS:i:48
```

read ID

chr

mapping quality

mapping metadata

Nucleotide sequence

PHRED score

# GTF

- Transcriptome Annotation File (Gene Transfer File)

$ gzip -d  => Unzip a compressed file (.gz)

gzip: compression file format and tool for compression/decompression

- gzip filename -> compress
- gzip –d filename -> decompress

```
(class) [u       @notchpeak2 test]$ curl -OL ftp://ftp.ensembl.org/pub/release-100/gtf/saccharomyces_cerevisiae/Saccharomyces_cerevisiae.R64-1-1.100.gtf.gz
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  539k  100  539k    0     0   154k      0  0:00:03  0:00:03 --:--:--  154k
(class) [u       @notchpeak2 test]$ gzip -d Saccharomyces_cerevisiae.R64-1-1.100.gtf.gz
(class) [u       @notchpeak2 test]$ ls
total 10M
drwxr-xr-x 2 u        rutter   10 May 30 10:53 away
drwxr-xr-x 2 u        rutter   10 May 30 10:31 bin
drwxr-xr-x 5 u        rutter   57 May 30 10:52 home
-rw-r--r-- 1 u        rutter 9.2M May 30 12:11 Saccharomyces_cerevisiae.R64-1-1.100.gtf
drwxr-xr-x 2 u        rutter   10 May 30 10:31 vol
```

$ cat Saccharomyces_cerevisiae.R64-1-1.100.gtf | column -t -s $'\t' | less -N -S


```
1  #!genome-build R64-1-1
2  #!genome-version R64-1-1
3  #!genome-date 2011-09
4  #!genome-build-accession GCA_000146045.2
5  #!genebuild-last-updated 2018-10
6  IV  sgd  gene          1802  2953  .  +  .  gene_id "YDL248W"; gene_name "COS7"; gene_source "sgd"; gene_biotype "protein_coding";
7  IV  sgd  transcript    1802  2953  .  +  .  gene_id "YDL248W"; transcript_id "YDL248W_mRNA"; gene_name "COS7"; gene_source "sgd"; gene_
8  IV  sgd  exon          1802  2953  .  +  .  gene_id "YDL248W"; transcript_id "YDL248W_mRNA"; exon_number "1"; gene_name "COS7"; gene_so
9  IV  sgd  CDS           1802  2950  .  +  0  gene_id "YDL248W"; transcript_id "YDL248W_mRNA"; exon_number "1"; gene_name "COS7"; gene_so
10 IV  sgd  start_codon    1802  1804  .  +  0  gene_id "YDL248W"; transcript_id "YDL248W_mRNA"; exon_number "1"; gene_name "COS7"; gene_so
11 IV  sgd  stop_codon     2951  2953  .  +  0  gene_id "YDL248W"; transcript_id "YDL248W_mRNA"; exon_number "1"; gene_name "COS7"; gene_so
12 IV  sgd  gene          3762  3836  .  +  .  gene_id "YDL247W-A"; gene_source "sgd"; gene_biotype "protein_coding";
13 IV  sgd  transcript    3762  3836  .  +  .  gene_id "YDL247W-A"; transcript_id "YDL247W-A_mRNA"; gene_source "sgd"; gene_biotype "prote
14 IV  sgd  exon          3762  3836  .  +  .  gene_id "YDL247W-A"; transcript_id "YDL247W-A_mRNA"; exon_number "1"; gene_source "sgd"; ge
15 IV  sgd  CDS           3762  3833  .  +  0  gene_id "YDL247W-A"; transcript_id "YDL247W-A_mRNA"; exon_number "1"; gene_source "sgd"; ge
16 IV  sgd  start_codon    3762  3764  .  +  0  gene_id "YDL247W-A"; transcript_id "YDL247W-A_mRNA"; exon_number "1"; gene_source "sgd"; ge
17 IV  sgd  stop_codon     3834  3836  .  +  0  gene_id "YDL247W-A"; transcript_id "YDL247W-A_mRNA"; exon_number "1"; gene_source "sgd"; ge
18 IV  sgd  gene          5985  7814  .  +  .  gene_id "YDL247W"; gene_name "MPH2"; gene_source "sgd"; gene_biotype "protein_coding";
19 IV  sgd  transcript    5985  7814  .  +  .  gene_id "YDL247W"; transcript_id "YDL247W_mRNA"; gene_name "MPH2"; gene_source "sgd"; gene_
```

# Delimited file

- .csv  => comma separated file
- .tsv  => tab separated file
- .txt

- Used for data tables

# Samtools

View alignment records

$ samtools view *filename.bam*| less –S

$ samtools view *filename.bam*| head –n 40 | less -S

# Determining sex of individual

- samtools view NA06984.454.MOSAIK.SRP000033.2009_11.bam | grep "Y" | column -t -s $'\t' | wc -l

  1098

- samtools view NA06984.454.MOSAIK.SRP000033.2009_11.bam | grep ”X" | column -t -s $'\t' | wc -l

  32012

Why are there such fewer reads for the Y chromosome in this individual?

# Homework

Run and copy these commands and outputs to a file

1. Log into the supercomputer
2. Download a 3 BAM files from the Thousand Genomes Project https://www.internationalgenome.org/data/
3. Determine the sex of the individual
4. Remove the BAM files to free up the space from your home directory

The size of a user's home directory space is enforced with a quota. There is a soft quota of 50GB and a hard quota of 75GB. Once a user's directory exceeds the soft quota, they have seven days to clean up and return to below the soft quota amount. After 7 days, they will no longer be able to write in their home directory until they clean up so that they are under the soft quota. If an user exceeds the hard quota, they immediately will no longer be able to write to their home directory until they clean up so they are not longer over this quota.