# Computing on TSCC

- Make a folder for the class and move into it
  - **mkdir –p /oasis/tscc/scratch/username/biom262_harismendy**
  - **cd /oasis/tscc/scratch/username/biom262_harismendy**

- Make symbolic link to data folder
  - **ln -s /projects/ps-yeolab/biom262_2016/data/ harismendy_data**

- Adjust your environment to have access to tools
  - **Source exportPATH.txt**

- Use screen
  - **screen –S name** to create
  - ^**A** ^**D** to detach
  - **Screen –x [PID].name** to attach
  - **exit** to kill

# UNIX commands

- Count number of lines
  - `wc –l TSG.bed`
- Select lines with “TP53”
  - `grep ‘TP53’ TSG.bed`
- Sort according to column 4
  - `sort –k 4 TSG.bed`
- Select genes on chromosome 1
  - `awk ‘$1==“chr1”’ TSG.bed`
- Calculate gene length
  - `awk ‘$5=$3-$2’ TSG.bed`

# Working with Intervals

Look at the file

```
    more harismendy_data/class/CGC.exons.bed
```

Count number of lines

```
    wc –l harismendy_data/class/CGC.exons.bed
```

Count the number of genes

```
    cut –f 4 harismendy_data/class/CGC.exons.bed | sort | uniq
| wc –l
```

Count number of intervals per gene

```
    cut –f 4 harismendy_data/class/CGC.exons.bed | sort | uniq
–c | sort –r | more
```

Calculate the size of each interval and sort by size

```
    awk ‘{size=$3-$2; print $0,size}’
harismendy_data/class/CGC.exons.bed | sort –n –r –k 5
```

Calculate the average interval size

```
    awk ‘{sum=sum+$3-$2} END {print sum/NR}’
harismendy_data/class/CGC.exons.bed
```

# fastqc

Inspect the fastq file

```
    zcat harismendy_data/class/file.fastq.gz | more
```

Read the help menu

```
    fastqc –h
```

Create a output directory

```
    mkdir fastqc
```

Run fastqc

```
    fastqc –o fastqc
harismendy_data/class/CGC.exons.bed/file.fastq.gz
```

Transfer the results to your desktop

Open results with web-browser

# BWA alignment and BAM files

Read the doc

```
   bwa mem or https://github.com/lh3/bwa
```

Start a Screen

```
   screen –S username
```

Alignment + convert to sorted bam

```
   bwa mem harismendy_data/resources/hg19_lite.fa –
t 1 R1.fq.gz R2.fq.gz | samtools view –buSh - >
sample.bam
```

Sort and index the bam file

```
   samtools sort –m 2G sample.bam sample.sorted
   samtools index sample.sorted.bam
```

# Remove Duplicate reads

-
  ```
  java -jar
  /opt/biotools/picard/picard.jar
  MarkDuplicates
  INPUT=harismendy_data/class/P21.sorted.bam OUTPUT=P21.sorted.rmdup.bam
  METRICS_FILE=myrmdupMetrics.txt
  REMOVE_DUPLICATES=true
  ASSUME_SORTED=true
  ```

# Stats and Slices

Calculate flag statistics

```
    samtools flagstat
harismendy_data/class/P21.sorted.bam > P21.flagstat.txt
```

How many “gapped reads” ?

```
    samtools view harismendy_data/class/P21.sorted.bam |
awk ‘$6~/[ID]/’ | wc -l
```

Subset the reads from an interval

```
    samtools view –bh –L harismendy_data/class/TSG.bed
harismendy_data/P21.sorted.bam > P21.sorted.TSG.bam
```

or

```
    samtools view -bh harismendy_data/P21.sorted.bam
chr1:120454175-120612317 > P21.NOTCH2.bam
```

# Visualize Alignments in IGV

- Transfer NOTCH2 BAM to your laptop (WinSCP, Fugu, cyberDuck, scp)
- Start IGV
- Use human hg19 genome reference

# BEDTOOLS coverage

Read the doc

**http://bedtools.readthedocs.org/en/latest/**

Read examples

https://github.com/arq5x/bedtools-protocols/blob/master/bedtools.md

Calculate coverage depth over CGC genes (chr1 only)

```
    bedtools coverage -hist -abam
harismendy_data/class/AA2253B_groupRealigned.chr1.bam -b
harismendy_data/class/CGC.exons.chr1.bed >
AA2253B.CGC.hist.cov.txt
```

Fraction of CGC bp covered at >30x ?

```
    grep ’^all' AA2253B.CGC.hist.cov.txt| awk '$2>30' | awk
'{sum=sum+$5} END {print sum}'
```

# Calculate HS Metrics

- 
  ```
  java -jar /opt/biotools/picard/picard.jar
  CalculateHsMetrics
  BAIT_INTERVALS=harismendy_data/resources/humanV4-baits_hg19_lite.interval_list
  TARGET_INTERVALS=harismendy_data/resources/resource/humanV4-targets_hg19_lite.interval_list
  INPUT=harismendy_data/class/P21.sorted.bam
  OUTPUT=P21.HsMetrics.txt
  ```