Homework 7

Chandrima Bhattacharya

25 February 2019

Question 1

First, I loaded the featureCounts which is present in subread package. Then I ran it for gene count.

```
spack load subread@1.6.2
featureCounts -a /home/frd2007/ANGSD_2019/RNA-seq/refGenome_S_cerevisiae/sacCer3.gtf -o featureCounts2_g
```

After this I ran it for exon count.

```
featureCounts -f -t exon -O -a /home/frd2007/ANGSD_2019/RNA-seq/refGenome_S_cerevisiae/sacCer3.gtf -o fe
```

The parameters I have used include "-f", "-t", "-O".

- The "-f" option performs read counting at feature level. This specifies the level of summarization.
- The "-t" option specifies feature type in GTF annotation. It is set to'exon' by default. Features used for read counting is extracted from annotation using the provided value. It is related with the annotations created.
- The "-O" option assigns reads to all their overlapping meta-features. If "-f" is specified, then the following assigns reads to the features instead. It is used to define overlap between reads and the features. As seen while implementing, when we have "-O", "-f" and "-t" specifying exon, Unassigned_ambiguity value comes to 0.

Question 2

I first copied all the data in my local file system.

```
scp chb4004@pascal.med.cornell.edu:angsd_hw/hw3/feature* Chandrima/Desktop/
```

I used R to plot values for gene count as below.

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.4
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.4
```

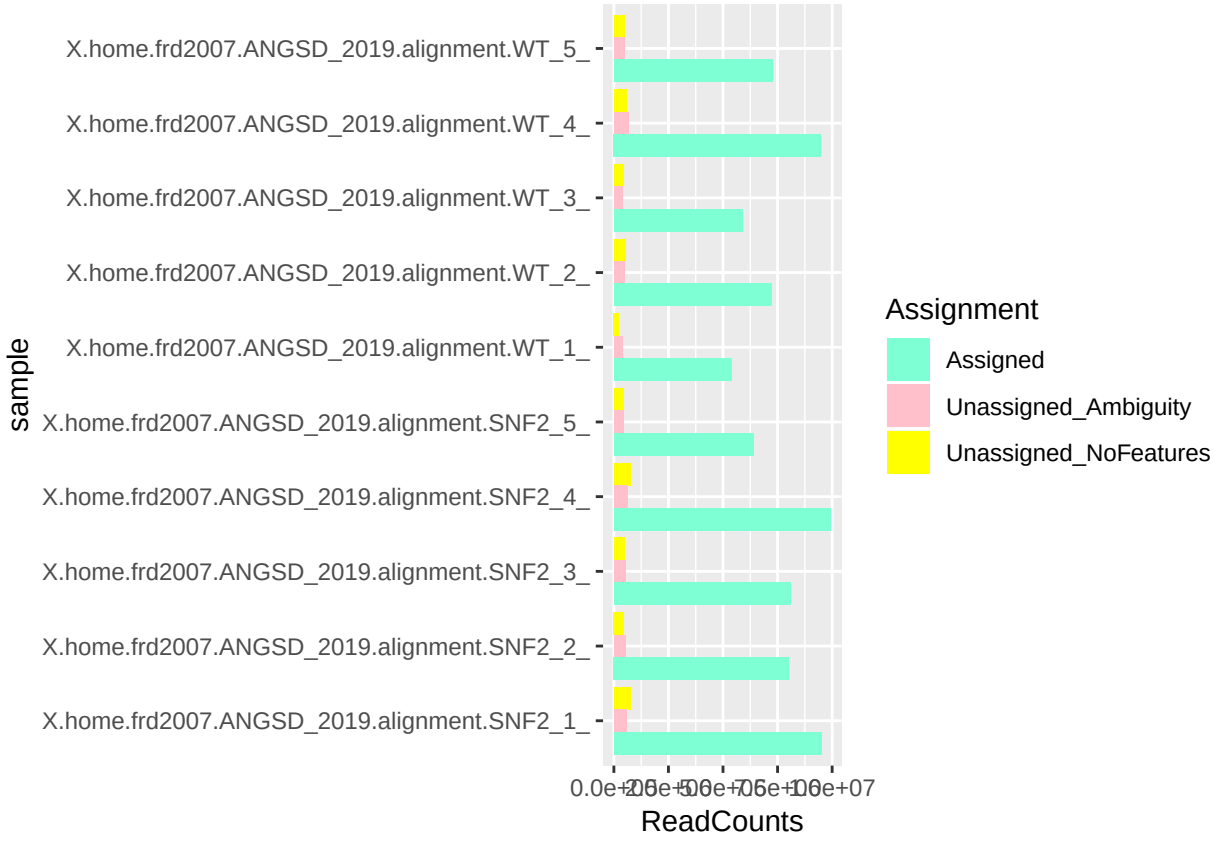
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
d <- read.delim("featureCounts2_gene.txt.summary", row.names=1) %>%
  # transpose and turn it from a matrix to a data frame
  t %>%
  as.data.frame %>%
  # make a "sample" column from the rownames of the object.
  # remove the ".star.Aligned.out.sam" from the pattern
  mutate(sample=gsub("Aligned.sortedByCoord.out.bam", "", rownames(.))) %>%
  gather(Assignment, ReadCounts, -sample)
```

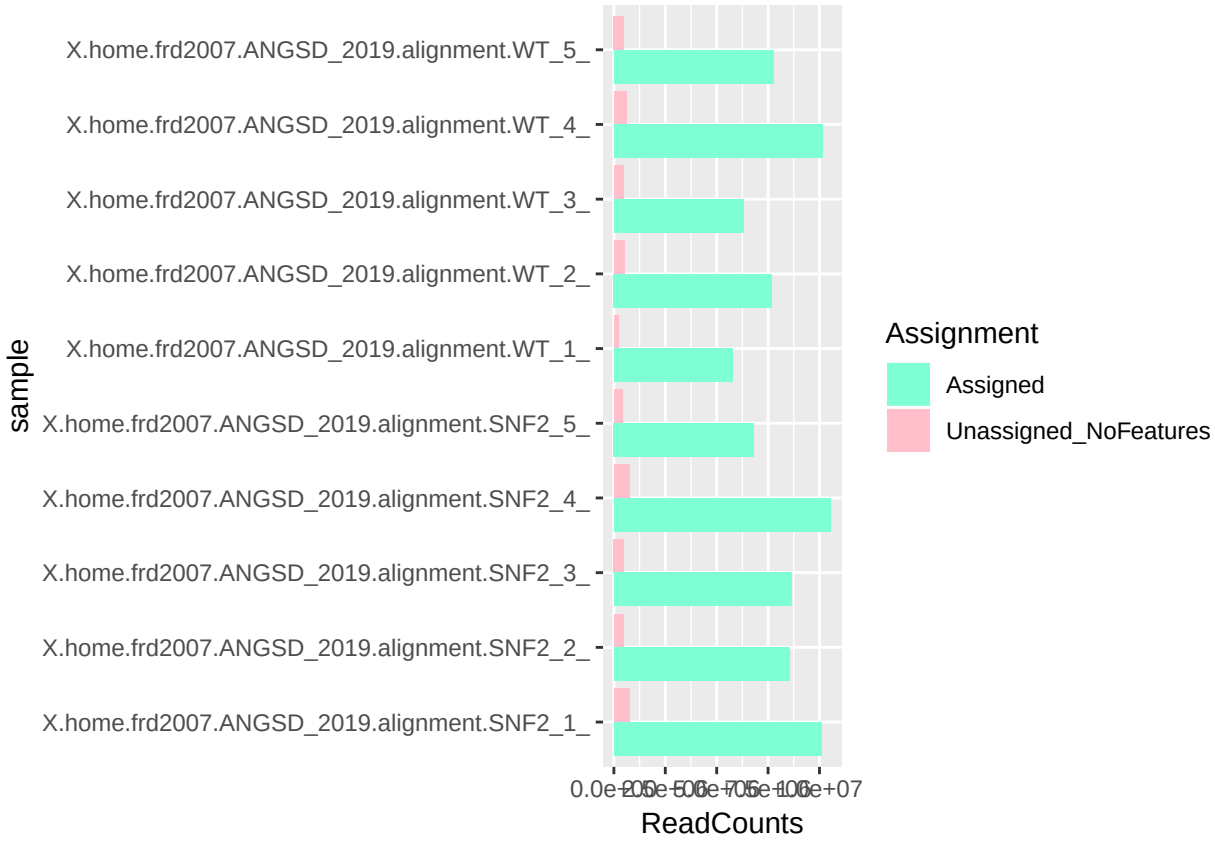
```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

```
d <- d[c(which(d$ReadCounts!=0)),]
d %>%
ggplot(aes(sample, ReadCounts)) + geom_bar(stat="identity", aes(fill=Assignment), position="dodge") + s
```



I used R to also plot values for exon count as below.

```r
library(dplyr)
library(tidyr)
library(ggplot2)
d <- read.delim("featureCounts2_star.txt.summary", row.names=1) %>%
  # transpose and turn it from a matrix to a data frame
  t %>%
  as.data.frame %>%
  # make a "sample" column from the rownames of the object.
  # remove the ".star.Aligned.out.sam" from the pattern
  mutate(sample=gsub("Aligned.sortedByCoord.out.bam", "", rownames(.))) %>%
  gather(Assignment, ReadCounts, -sample)
d <- d[c(which(d$ReadCounts!=0)),]
d %>% ggplot(aes(sample, ReadCounts)) + geom_bar(stat="identity", aes(fill=Assignment), position="dodge
```



Question 3

- There is no Unassigned_ambiguity in the run where for exon.
- If we did not have "-O", then there is Unassigned_ambiguity even for exon counts. Also, value of assigned decreases in this case.
- The levels of assigned reads between the exon and gene counts are exactly same.
- The value for assigned reads is more for exon than gene counts.

Reference

- https://rpubs.com/turnersd/processing-featurecounts