STAT718/BIOL703 Homework 5

DUE: Friday Oct 25, 2024 before calss

Total Points: 40

Download Data from GitHub and Set Up.

In this homework, we will use the dataset from Howard et al (2013). The FASTA files were downloded from GEO (Accession SRP010938). The dataset contains 18 paired-end (PE) read sets from *Arabidposis thaliana* These FASTA files data have been aligned to the reference genome and the corresponding BAM files can be downloaded from Yen-YiHo's GitHub Repo STAT718_Homework5 (https://github.com/Yen-YiHo/STAT718_Homework5/tree/main).

Part 1 Unstranded and strand-specific read counting

To work with this data set,

- (1) Clone the GitHub repository described above
- (2) Set your RStudio sesssion to this directory
- (3) Next, load all libraries, the annotation and the BAM files requires for the read counting and downstream analysis steps

```
library(systemPipeR)
library(GenomicAlignments)
library(GenomicFeatures)
library(BiocParallel)
library(ggplot2)
setwd("/Users/hoyen/Desktop/STAT718/Homework/RNAseqHW/rnaseq")
txdb <- loadDb("./data/tair10.sqlite")
eByg <- exonsBy(txdb, by = c("gene"))
outpaths <- list.files('./results/hisat2_mapping/', pattern='sorted.bam$',
↪  full.names=TRUE)
bfl <- BamFileList(outpaths, yieldSize = 50000, index = character())
```

In the read quantification step with **summarizeOverlaps** generate count tables for exons by genes (**eByg**) of the following three strand modes:

- Unstranded
- Strand-specific for positive (sense) strand
- Strand-specific for negative (antisense) strand

The codes below can be used to generate the unstranded read counts,

```
unstranded <- summarizeOverlaps(eByg, bfl, mode="Union",
↪  ignore.strand=TRUE,inter.feature=FALSE, singleEnd=TRUE)
unstranded <- assays(unstranded)$counts
unstranded[1:4,]
```

```
##           A12A.sorted.bam A12B.sorted.bam A1A.sorted.bam A1B.sorted.bam
## AT1G01010            1080             887            748            435
## AT1G01020             263             229            287            329
## AT1G01030              67             114            347            152
## AT1G01040            2719            1468           2110           1618
##           A6A.sorted.bam A6B.sorted.bam M12A.sorted.bam M12B.sorted.bam
## AT1G01010           1065            653             427             521
## AT1G01020            235            215             135             335
## AT1G01030             53            155             102             118
## AT1G01040           1805           1335            1362            2249
##           M1A.sorted.bam M1B.sorted.bam M6A.sorted.bam M6B.sorted.bam
## AT1G01010            586            626            524            287
## AT1G01020            219            338            312            280
## AT1G01030            256            267            154             58
## AT1G01040           1866           1798           1790           1272
##           V12A.sorted.bam V12B.sorted.bam V1A.sorted.bam V1B.sorted.bam
## AT1G01010            1331            1408           1181            608
## AT1G01020             160             319            351            308
## AT1G01030             367             967            309            402
## AT1G01040            1380            2084           2089           1845
##           V6A.sorted.bam V6B.sorted.bam
## AT1G01010           1561           1148
## AT1G01020            423            442
## AT1G01030            242            464
## AT1G01040           3349           3248
```

Read the vignette Counting reads with `summarizeOverlaps` and the help file for `?summarizeOverlaps`. Then perfrom the following tasks:

- Generate strand-specific for positive (sense) strand and Strand-specific for negative (antisense) strand. (20 points)
- Sum the two strand-specific read count table and compare to the unstranded count table. Are they similar? (10 points)
- Explain the experimental conditions when the different strand counting modes would result in different read counts. (10 points)

Reference:

1. Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, He L, Winter-Sederoff H, Gassmann W, Veronese P, Heber S. High-throughput RNA sequencing of pseudomonas-infected Arabidopsis reveals hidden transcriptome complexity and novel splice variants. PLoS One. 2013 Oct 1;8(10):e74183. doi: 10.1371/journal.pone.0074183. PMID: 24098335; PMCID: PMC3788074.