Applied Genomics 2018
Homework 02 - Alignment
Due: February 23rd, 11:59pm

The Experiment

The dataset that we are going to align is an RNA-seq dataset. There are 12 paired-end stranded RNA-seq samples (24 files) from an experiment in Arabidopsis root. It is a factorial design with two factors:

- oxidative stress (caused by the chemical "paraquat")
- knockdown of the RNA methylation gene "trm4b" and control samples are wild type (WT).

There are three biological replicates for each combination of factors.

For full details see this page: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80054

The experimental design is described in the table below. Each sample has two files (one for each end of the pair). Files are located in :

/scratch/courses/AppliedGenomics2018/Homework02/

SRA id	GEOid	
SRR3347441	GSM2111753	WT rep1_root_RNA-seq
SRR3347442	GSM2111754	WT rep2_root_RNA-seq
SRR3347443	GSM2111755	WT rep3_root_RNA-seq
SRR3347444	GSM2111756	trm4b rep1_root_RNA-seq
SRR3347445	GSM2111757	trm4b rep2_root_RNA-seq
SRR3347446	GSM2111758	trm4b rep3_root_RNA-seq
SRR3347447	GSM2111759	WT paraquat rep1_root_RNA-seq
SRR3347448	GSM2111760	WT paraquat rep2_root_RNA-seq
SRR3347449	GSM2111761	WT paraquat rep3_root_RNA-seq
SRR3347450	GSM2111762	trm4b paraquat rep1_root_RNA-seq
SRR3347451	GSM2111763	trm4b paraquat rep2_root_RNA-seq
SRR3347452	GSM2111764	trm4b paraquat rep3_root_RNA-seq

1) Align using Hisat2

Align all samples to the Arabidopsis reference genome using HISAT2. Use the Ensembl reference so we can use the GTF file like we did in class to identify where the reads mapped. The sequences and indexes related to this version is here :

```
/genomics/genomes/Public/Plant/Arabidopsis_thaliana/Ensembl/TAIR10
```

You **MUST** submit an SBATCH script (call it **netid_run_hisat.sbatch** where netid is **your** netid) that will execute this job for all sequences.

2) Use Samtools

Write another sbatch script (**netid_sort_bam.sbatch**) which will use Samtools to convert the file to BAM and sort it.

3) Use Bowtie2

HISAT is an alignment designed for aligning RNA-seq data, but how much information would we lose if we actually did not care that it is RNA-seq data? Create an sbatch script (**netid_run_bowtie.sbatch**) which will align the RNA-seq data to the Arabidopsis reference genome using Bowtie2.

I realize we did not review Bowtie2 in class, however you should be able to look up the documentation and forums that will help you determine how to use the different options. Also note that the index files that you will use for HISAT are different from the ones you need for Bowtie2. You will need to refer to the Bowtie2 database as:

/genomics/genomes/Public/Plant/Arabidopsis_thaliana/Ensembl/TAIR10/Arabidopsis_thaliana.TAIR10.dna.toplevel

4) Compare the HISAT2 and BOWTIE2 results

Compare the summary statistics of SRR3347441 aligned HISAT2 to BOWTIE2. Use the flagstat function in samtools to compare the results. Was it helpful to use HISAT?

5) Calculate Read Counts

Use the R packages (RSamtools, GenomicFeatures, and GenomicAlignments) to count the number of reads that are matching the different genes.

a) How many genes have no reads mapping across all samples ?
b) Which gene has the highest total number of reads?
c) What is the total number of mapped reads per sample?

Extra Credit:

How many Exon-Intron junctions were identified by HISAT2? How many by BOWTIE2?