Lab 4: RNA-Seq Mapping and Differential Expression Analysis

In the Google Drive, download the salmon.zip file and put in your udrive. Unzip it.

We are going to be determining the differential expression of mRNAs from a 4-condition experiment in *Arabidopsis thaliana*. The conditions are listed below.

Sample	Genotype	Growth condition	Ros1&3 present	Dml2 &3 present	Osmotic stress condition
DRR016125	wt	Control	+	+	-
DRR016126	wt	ABA	+	+	-
DRR016127	wt	Saline	+	+	+
DRR016128	wt	Dry	+	+	+
DRR016129	Ros1-3	Control	-	+	-
DRR016130	Ros1-3	ABA	-	+	-
DRR016131	Ros1-3	Saline	-	+	+
DRR016132	Ros1-3	Dry	-	+	+
DRR016133	Dml2;3	Control	+	-	-
DRR016134	Dml2;3	ABA	+	-	-
DRR016135	Dml2;3	Saline	+	-	+
DRR016136	Dml2;3	Dry	+	-	+
DRR016137	Ros1,dml2;3	Control	-	-	-
DRR016138	Ros1,dml2;3	ABA	-	-	-
DRR016139	Ros1,dml2;3	Saline	-	-	+
DRR016140	Ros1,dml2;3	Dry	-	-	+

These data are from the European Nucleotide Archive. They are from study PRJDB2508. For your write-up you will need to obtain some data from this website.
You will also find more information in the SRARunTable_DRR0161XX.csv in the folder.

We are going to start by mapping reads to the transcriptome of the *Arabidopsis*. To do this, open a terminal window and navigate to the folder you just unzipped and into the salmon folder.

We're going to be using the tool Salmon[1] (https://github.com/COMBINE-lab/salmon). As discussed in class, this program maps reads to the transcriptome (not the genome). It uses quasi-mapping, so it is fast. Though quasi-mapping might lead you to think it's less accurate, the reverse is true. The algorithm is less prone to being confused by SNP errors, so can lead to more accurate mappings and quantifications than alignment based algorithms, like the Tuxedo Suite.

Today we will be essentially following the tutorial for how to use Salmon. Next week, you will have an opportunity to run this procedure on data of your choosing.

Start by downloading the Mac file from https://github.com/COMBINE-lab/salmon/releases.
Unzip it and move the unzipped folder into the Applications folder.

Though one can generate a transcriptome from the RNA-Seq reads, we are going to use the annotated transcriptome from ENSEMBL for *Arabidopsis thaliana*. In the folder you will find a file called `athal.fa.gz`. This file was downloaded using the following command:

```
$ curl ftp://ftp.ensemblgenomes.org/pub/plants/release-28/fasta/arabidopsis_thaliana
```

In the salmon directory, in the terminal window you are going to build the index of the transcriptome that salmon will use for matching.

```
$ DYD_FALLBACK_LIBRARY_PATH=/Applications/Salmon-0.8.2_macOX.10.12/lib
/Applications/Salmon-0.8.2_macOX.10.12/bin/salmon index —t ahtal.fa.gz
—i athal_index
```

While this is running, you can watch to see what it's doing, and you can look at our data. You will find a few things in this folder. In the folder data you will find 2 folders DRR016125-126. We are only going to map the first 2 datasets here. The data are LARGE, and take a long time to download (as you saw). Also, though salmon is fast, it still takes a while to align, so we aren't going to wait that long. We are going to use the Julia Child technique for this class. We'll start the mapping and then pull mapped data from the oven.

I have made a file for you that will go through each of the data sets in the folders and generate read quants. This file is called quant_tut_samples. You should open it up in TextEdit and look at it. We'll talk about what Salmon is doing.

To run it:

```
$ bash quant_tut_samples
```

In the folder quants, you will find the mapped data for the other 14 sets of reads. Open one and together we'll look at the data and see what they tell us.

Once we're done with the mapping and quantitation we want to move to expression analysis – looking to see which genes are changing in expression.

In class, we will be using Wasabi (https://github.com/COMBINE-lab/wasabi) and Sleuth[2] (http://pachterlab.github.io/sleuth/) and R to analyze these data.

Open R-studio.
In the command line section of R-Studio install BioConductor, Sleuth, and Wasabi using the following commands:

1. Install/activate bioconductor as we've done before.
2. `> biocLite('rhdf5')`
3. `> biocLite('COMBINE-lab/wasabi')`

```
4. > library(sleuth)
5. > library(wasabi)
```

Using the bottom right box, set the working directory to salmon
Then using the command line:

```
1. > sf_dirs <- file.path ( 'DRR_quants', c('DRR016125_quant',
   'DRR016126_quant', 'DRR016127_quant', 'DRR016128_quant',
   'DRR016129_quant', 'DRR016130_quant', 'DRR016131_quant',
   'DRR016132_quant', 'DRR016133_quant', 'DRR016134_quant',
   'DRR016135_quant', 'DRR016136_quant', 'DRR016137_quant',
   'DRR016138_quant', 'DRR016139_quant', 'DRR016140_quant'))

2. > prepare_fish_for_sleuth(sf_dirs)
```

We have 16 conditions: wild type, each gene knocked out individually, and both knocked out together.

3. Open and Source Genotype.R This will label the genotypes.

We need to describe the experiment in terms of genotype and growth condition. We also summarize the gene knockouts and osmotic stress status for each sample.

4. Open and source Experimental Conditions.R

You should see the data frame s2c upper in the upper right frame.

Next we want to model the experimental condition of interest, osmoic stress.

5. Open and source Model.R

In the command line:

```
6. > sleuth_live(so)
```

This will start the application Shiny. In there we can explore the differential expressions.

Go to the Maps tab and select PCA.
Find a setting that allows you to find separable populations. Is there a gene knowckout that makes a difference under one of the stress conditions?

This is the principal component analysis. This compacts a multi-dimensional variable set down into a combination of the variables that allows for the greatest separation between samples.

If you compare PCA 1 and 2, how many groups do you see? What if you compare 2&3? Can you identify what characteristics determine these separate groups?

If you go to the analyses tab and select test table, you will find a list of the genes that are most different between the control and experimental conditions of the model. Find the most changed transcript. Go to the analysis tab and select transcript view. Put the most differently expressed transcript in the box and select osmotic stress for the color. What do you see? Which of the osmotic stress conditions causes the largest change in expression?

Look at the volcano plot on the analyses tab. What does this show you?

You can also run Model2.R and see how that changes your volcano plot.

References

1. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Meth* **14,** 417–419 (2017).
2. Pimentel, H. J., Bray, N., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-Seq incorporating quantification uncertainty. *bioRxiv* 58164 (2016). doi:10.1101/058164

Lab Report Guidelines: Total 100 pts. Due May 19, 11:59 PM

Section 1, Understanding the data (15 pts):
Who sequenced these plants? What technology was used? Are these single reads or paired reads? What are the growth conditions? What are the genes that are knocked out? What are their functions? How many reads are there in on of the RNA-Seq datasets?

Section 3, Mapping the reads (15 pts):
We used the ENSEMBL transcriptome for this analysis, why? What are the benefits of using an annotated transcriptome versus a *de novo* transcriptome? When would it be better to use a *de novo* transcriptome? Looking at the mappings, are there transcripts with no reads assigned to them? Is this surprising? Given how long it took to map these reads, how long would it take to map all 16 sets of RNA-Seq data? What would make this faster?

Section2, Analyzing the data (35 pts):
In comparing PCA 1 and 2, how many groups do you see? What happens you compare 2 and 3? What characteristics determine these separate groups? What does this tell you about the differences in these experimental conditions? What causes the biggest difference in gene expression?

What are the five genes with the largest changes of expression under osmotic stress conditions? How large is the change? Do all these genes change in the same direction, ie, all upregulated during osmotic stress? Are these changes significant? Create figures that illustrate the magnitude of the changes in these genes in context with other genes. You may use the graphs created by Sleuth in Shiny, but you should annotate them as needed and provide a descriptive caption.

Section 3, Interpreting the results (25 pts):
Identify the names of genes in your list of top 5 most differentially expressed genes. What do these genes do? How are they be influenced by osmotic stress? How might the mutations seen in this experiment influence these genes? Is there support for your hypothesis in the data collected? What is it? How would you confirm your hypothesis? Provide any figures and tables you think necessary for a reader to understand your interpretation.

Overall (10 pts):
Is this report written in paragraph form with the questions answered in context, not at bullet points? Is scientific vocabulary used appropriately? Is it grammatically correct with few typos? Does it read well? Have you cited the tools and sources you used?