Analysing Transcriptomics data

Estimating Expression Levels of Transcripts and Genes

Using the sratoolkit download the following samples from SRA and clean them, unless they are made available for you in some other way.

DRR016125	DRR016131	DRR016137
DRR016126	DRR016132	DRR016138
DRR016127	DRR016133	DRR016139
DRR016128	DRR016134	DRR016140
DRR016129	DRR016135	
DRR016130	DRR016136	

That data is form an experiment in Arabidopsis thaliana evaluating the effect of some abiotic stress on Arabidopsis plants (Wild type and some mutants). Check in the SRA for the details of the experiment.

From last week you shoudl have all the predicted cDNA in Arabidopsis. We will use that file to create an index file that will aid in the estimation of transcript levels. This estimation and the index creation will be carried out with Salmon, using an alignment-free approach.

First get Salmon: Go to https://github.com/COMBINE-lab/salmon/releases and get the latest binary release for your platform. Now from the terminal run the following:

```
cd
mkdir SalmonBinaries
cd SalmonBinaries
tar xvzf ~/Downloads/salmon_0.14.2_linux_x86_64.tar.gz
```

run:

```
~/SalmonBinaries/salmon-latest_linux_x86_64/bin/salmon
```

The Salmon help should be displayed in your terminal

Now let's create the index.

In your home create a folder with the name: `RNASeqPracticalTranscriptLevels and inseide that one create ArabidopsiscDNAINDEX`, go to that folder and put you cDNA file in there:

```
mkdir -p ~/RNASeqPracticalTranscriptLevels/ArabidopsiscDNAINDEX
cd ~/RNASeqPracticalTranscriptLevels/ArabidopsiscDNAINDEX
cp ~/Downloads/Araport11_genes.201606.cdna.fasta.gz .
gunzip  Araport11_genes.201606.cdna.fasta.gz
```

Now we will create an index file for Salmon using all the cDNA in the A. thaliana genome

```
~/SalmonBinaries/salmon-latest_linux_x86_64/bin/salmon index --transcripts Araport11_genes.201606.cdna.fasta --index Araport11_genes.201606.cdna --threads 3
```

Change folder one level up:

```
cd ~/RNASeqPracticalTranscriptLevels
```

And copy your cleand RNASeq data into that folder. There is copy in your instructor's server. Run the following to get it:

```
scp -r student20@192.168.105.106:/home/student20/RNASeqTranscriptLevelsCleanDATA .
```

Before computing the expression levels with Salmon we need a table, that stores the relationship between genes and their transcripts, i.e., which transcripts belong to which genes. How could you generate that table usign the information in the file `~/RNASeqPracticalTranscriptLevels/ArabidopsiscDNAINDEXraport11_genes.201606.cdna.fasta?` discuss with your instructor. Please name this file tx2gene.txt and put a copy of it in the same folder where you have the salmon index.

Now we can use Salmon and the index we created before to estimate expression levels for the genes and their transcripts. Please note that we have 16 samples, and we need to run Salmon for each of there. Please write a bash script using a for loop to do this. Note, that you must check the help from Salmon to decide on which parameters to use. Discuss with your instructor. Also, please make sure to store all the data in subfolder within /home/diriano/RNASeqPracticalTranscriptLevels/Salmon, these subfolder MUST have as name the name of the sample followed by a common string, e.g., _salmon, so that they look like:

DRR016125_salmon

Now have a look at the files quant.genes.sf and quant.sf, please identify all the information contained in them. Also check the salmon log files, identify the mapping rate and the type of library used? Was it an strand-specific library?

Now follow the instructions to continue working with these data from R