**DIFFERENTIAL EXPRESSION ANALYSIS**

We are now going to do a basic differential gene expression analysis of our NGS data. Note that this part of the analysis is similar regardless of whether you have a pre-assembled genome to work with or are working with a transcriptome you assembled yourself (as we did yesterday). We will keep working with our STG dataset for a bit.

The general workflow is as follows:

1. align trimmed reads to reference transcriptome (bowtie, bwa, shrimp, etc)

2. generate matrix summary of read statistics (RSEM, eXpress)

3. statistically compare read-contig values (edgeR, DEseq2, limma-voom)

4. generate figures (heatmaps, volcano plots, etc)

5. As our friend Harold Zakon says, now is when the real work begins: read and think about your data!

Before we begin we again need to load a couple of modules:

$ module load samtools/1.3

$ module load bowtie/1.1.1

$ module load gcc/4.7.1

$ module load trinityrnaseq/2.0.6

$ module load express/1.3.0

$ module load R/2.15.1

$ export PATH=/work/04148/efischer/programs/bowtie2-2.2.9:/work/04148/efischer/programs/RSEM-1.2.31:$PATH

**Align reads & estimate abundance**

Before we can run a differential gene expression we need to estimate the abundance of the reads in each sample. To do this we will have to map each sample back to the reference assembly (i.e. transcriptome or genome). We will continue with Trinity here, but there are other programs that can do this and Trinity is really just calling on some of those to do the mapping.

$ $TACC\_TRINITY\_DIR/util/align\_and\_estimate\_abundance.pl --transcripts /work/04148/efischer/STG/data/references/Cborealis\_ref.fa --est\_method RSEM --aln\_method bowtie --prep\_reference

The first step in this process is to prepare the reference for mapping.

Now we can map. Note: we will do this for each pair of R1 and R2 samples individually using only the paired reads!

$ $TACC\_TRINITY\_DIR/util/align\_and\_estimate\_abundance.pl --transcripts /work/04148/efischer/STG/data/references/Cborealis\_ref.fa --seqType fq --left PD\_01\_R1\_trim\_pair.fq.gz --right PD\_01\_R2\_trim\_pair.fq.gz --est\_method RSEM --aln\_method bowtie --SS\_lib\_type RF --thread\_count 16 --output\_prefix PD\_01

Modify the above code to map your second sample pair.

Now that we have mapped all the samples we can combine them into a single matrix of counts that we will use for our downstream analyses:

$ $TACC\_TRINITY\_DIR/util/abundance\_estimates\_to\_matrix.pl --est\_method RSEM --out\_prefix genes PD\_01.genes.results PD\_02.genes.results GM\_01.genes.results GM\_02.genes.results

**Differential expression analysis**

As with all the steps along the way there are many different programs you can use for differential expression analysis. All these programs claim to be the best one, so it’s up to you to make an informed decision based on your need. I personally like to use a couple of different approaches to understand whether I’m getting the same ‘answer’ across the board, and also because different programs have different strengths and weaknesses. In that vein, we’re going to show you a couple of options.

*EdgeR*

The first thing we will do is take advantage of the Trinity pipeline to run EdgeR. This is a program that runs in R, so you could also do this on your own that way, but it’s quite a bit easier to just use what Trinity already has set up (at least after you get everything installed properly …). Let’s get everything installed first.

We’ve already loaded R, but now we need to open it to install some libraries. Later on we will use R directly, but note that you may want to run R this way if you are doing something computationally intensive enough that your personal computer cannot handle it.

$ R

(lots of stuff will happen here because R will open – once this happens continue with the code below)

> source("http://bioconductor.org/biocLite.R")

> biocLite('edgeR')

> biocLite('limma')

> biocLite('DESeq2')

> biocLite('ctc')

> biocLite('Biobase')

> install.packages('gplots')

> install.packages('ape')

Once the above is completed without throwing errors you can type q() to exit R.

The other thing you need for differential expression analysis is a tab delimited file that tells the program which sampled are which. This should have the format:

group1 sample1

group1 sample2

group2 sample3

group2 sample4

I’ve created this file for you, so you should be good to go but NOTE: many programs that you’re likely to use to make these kinds of files (e.g. Microsoft excel) use a different format of line breaks and you need to make sure your file has unix line breaks otherwise things will not work because the file will not be read properly. From the directory that contains your counts matrix and the samples file run the following code:

$ $TACC\_TRINITY\_DIR/Analysis/DifferentialExpression/run\_DE\_analysis.pl --matrix genes.counts.matrix --method edgeR --samples\_file samples.txt --output edgeR\_DE

Once the above is completed we can cd into the edgeR\_DE directory it generated and run the following code to make a heatmap of our data. The -P option specifies the p-value cut off; -C specifies the min abs(log2(a/b) fold change where 2 means 2^2=4-fold.

$ $TACC\_TRINITY\_DIR/Analysis/DifferentialExpression/analyze\_diff\_expr.pl --matrix ../genes.TMM.fpkm.matrix --samples ../samples.txt -P 1e-2 -C 2

What happened when you did this?

**Working locally in R**

Now we’re going to transition and use R directly to run some additional visualizations as well as differential expression analysis using a second program called DESeq2. We’re also going to use a different dataset.

The dataset we will look at is from my PhD work and has fish from two populations of guppies, a high-predation and a low-predation population. High- and low-predation fish differ in all sorts of characteristics (morphology, life-history, behavior) and so we were interested in looking at differences in whole-brain gene expression in these fish.

*EdgeR visualizations*

We just ran EdgeR via Trinity, but really this is just an R program that you can also run yourself on your own computer (there are pros and cons to being able to do things on your laptop vs needing the HPC). To give you an idea of what is happening behind the scenes when we run this program in Trinity we’re now going to run it in R directly, mostly with the goal of doing some data visualization.

Open the “edgeRcode” R script in the folder of files I gave you. This should automatically open in R and then we can go from there.

*DESeq2*

Another program (which I prefer) for differential expression analysis is DESeq2. There are statistical differences between the programs, but the other thing I like about DESeq2 is that there documentation is much more helpful / easier to understand. I’ve provided you the pdf for this in the course folder.

Open the “DESeq2code” R script in the folder of files I gave you. This should automatically open in R and then we can go from there.