Differential Gene Expression with edgeR

Note: First we need to install one more `R` *package. To do so follow these instructions:*

Once you are logged in to hydra:

```
$ module load tools/R/3.2.1
$ R
```

Now that you are in `R`, you need to load the biocLite function again:

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

You will be prompted to update your other libraries. Respond with `n`.

Create sample text file

You will need to create a tab delimited text file containing information for your different samples. In the first column, you will have the name of the condition and, in the second column, you will enter the name of the sample. Start a new text file, samples.txt, with nano.

```
$ nano samples.txt
```

Enter the following (keep in mind that the condition and names should be separated by tabs!).

```
GSNO    GSNO_SRR1582648
```

```
GSNO    GSNO_SRR1582647
GSNO    GSNO_SRR1582646
WT  wt_SRR1582651
WT  wt_SRR1582649
WT  wt_SRR1582650
```

Since software can be very picky about whether you specified config files correctly, it is sometimes good to check that you did, indeed, enter the correct characters. You can view special characters with:

```
$ cat -te samples.txt
```

If your file was specified correctly, it should look like this:

```
GSNO^IGSNO_SRR1582648$
GSNO^IGSNO_SRR1582647$
GSNO^IGSNO_SRR1582646$
WT^Iwt_SRR1582651$
WT^Iwt_SRR1582649$
WT^Iwt_SRR1582650$
```

`^I` characters are tabs and `$` characters are newlines. Make sure that your text file looks like the example above when using `cat -te`. If it doesn't, you'll need to edit it until it does.

Detect differentially expressed transcripts in `edgeR`

Now we are going to use the `run_DE_analysis.pl` script that is included with the `Trinity` package to detect differentially expressed transcripts in edgeR. Note that this will only work if the R packages from `Environment setup.md` were installed properly.

Create a new job file, and select the short queue and 2GB of RAM. Load the Trinity module. The command will look like this:

```
run_DE_analysis.pl \
      --matrix Trinity_trans.counts.matrix \
      --samples_file samples.txt \
      --method edgeR \
      --output edgeR_trans
```

Save the job file to your `/pool/genomics/<username>/RNAseq_workshop` directory and submit it to the cluster. Once it is finished, there will be a new directory called `edgeR_trans`. Take a look at its contents:

```
$ ls -lh edgeR_trans
```

There should be three files in the directory:

```
-rw-rw-r-- 1 frandsenp frandsenp  27K Jun  7 07:53 Trinity_t
rans.counts.matrix.GSNO_vs_WT.edgeR.DE_results
-rw-rw-r-- 1 frandsenp frandsenp  12K Jun  7 07:53 Trinity_t
rans.counts.matrix.GSNO_vs_WT.edgeR.DE_results.MA_n_Volcano.
pdf
-rw-rw-r-- 1 frandsenp frandsenp 1020 Jun  7 07:53 Trinity_t
rans.counts.matrix.GSNO_vs_WT.GSNO.vs.WT.EdgeR.Rscript
```

The file `Trinity_trans.counts.matrix.GSNO_vs_WT.edgeR.DE_results` contains the results from comparing the GSNO condition to the wt condition. Take a look:

```
$ head edgeR_trans/Trinity_trans.counts.matrix.GSNO_vs_WT.ed
geR.DE_results | column -t
```

```
logFC                   logCPM             PValue
 FDR
```

```
TRINITY_DN283_c0_g1_i1  8.86394339657974   14.1351668438638
 4.32887819405242e-47  1.29433458002167e-44
TRINITY_DN587_c0_g1_i1  5.74036873153572   14.5017490646562
 5.65375941609977e-40  8.45237032706915e-38
TRINITY_DN545_c0_g1_i1  2.48516846245412   15.0533519786141
 1.32755252392501e-29  1.32312734884526e-27
TRINITY_DN41_c0_g1_i1   5.0406890184565    13.8558652639267
 2.40875725742197e-26  1.80054604992293e-24
TRINITY_DN568_c0_g1_i1  -5.11456747445237  13.7656734840196
 1.0513434360315e-25   6.28703374746839e-24
TRINITY_DN8_c0_g1_i1    6.05737236243611   13.2471666144585
 2.0676204325293e-24   1.03036418221044e-22
TRINITY_DN300_c0_g1_i1  2.45950024584126   15.1423904751683
 3.48392323833275e-24  1.48813292608785e-22
TRINITY_DN442_c0_g1_i1  -3.19677994792633  14.0243706003162
 1.91541341804279e-21  7.15885764993493e-20
TRINITY_DN181_c0_g1_i1  5.20988783333878   13.2037202161663
 3.01350479084321e-21  1.00115325829124e-19
```

As you can see, `edgeR` calculates log fold change (`logFC`), the log counts per million (`logCPM`), the p-value from the exact test (`PValue`), and the false discovery rate (`FDR`).

Note: Since there is no header for gene name, the headers are shifted one column to the right, i.e. `logFC` should be over the first column of floating point numbers.

`edgeR` also generated MA and Volcano plots for these data. We will now download them to our computer. If you are using Mac or Linux, we will do this with the `scp` command. Open a new terminal window and `cd` to the directory that you wish to download the files to. On Mac, I often download to my `Downloads` directory. You can go there with:

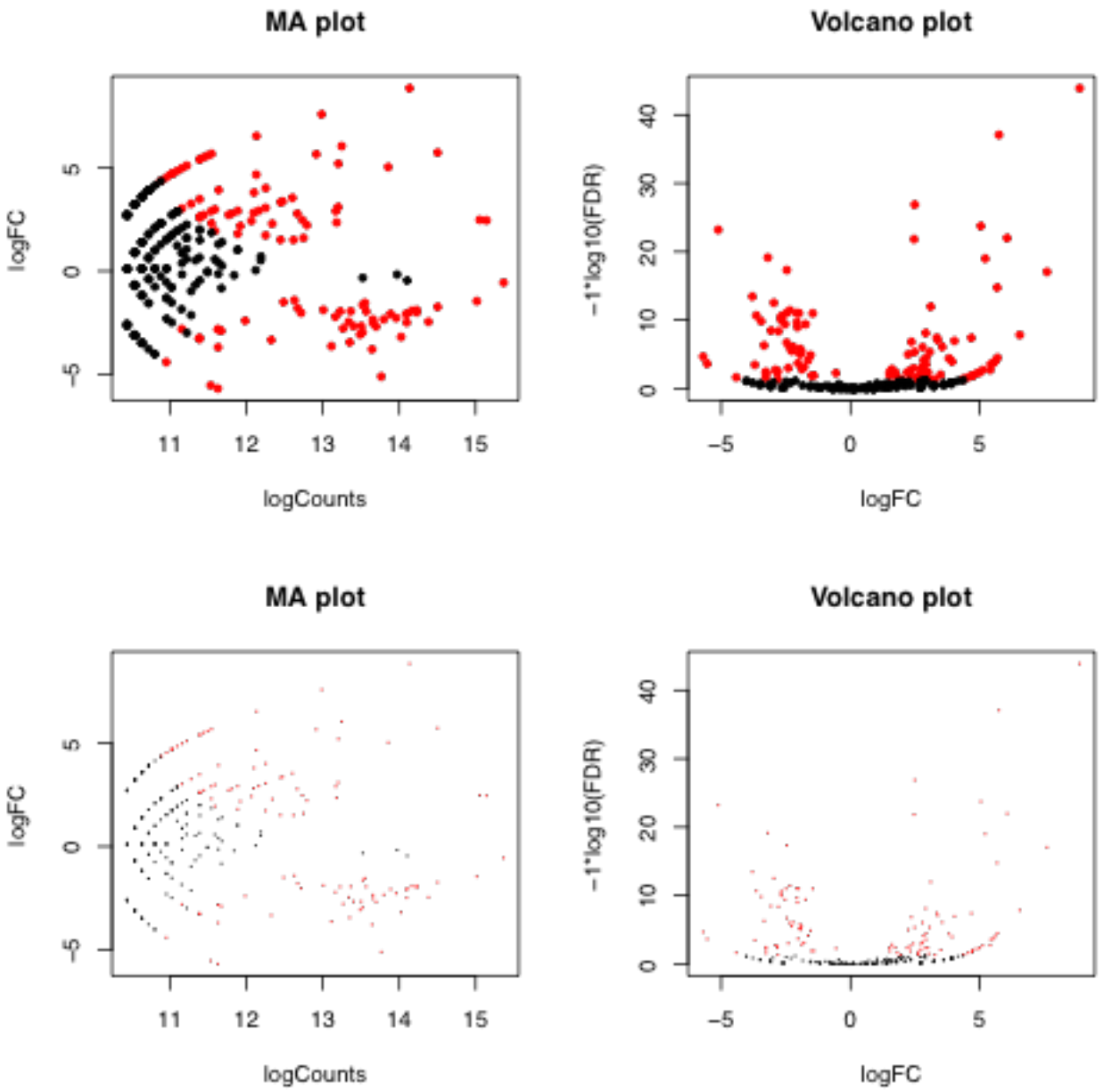
```
$ cd ~/Downloads
```

Now download the plot:

Now download the plot:

```
$ scp <username>@hydra-login01.si.edu:/pool/genomics/<username>/RNAseq_workshop/edgeR_trans/*.pdf .
```

Go ahead and open it to examine its contents.



The points that are in red are determined to be significant with an `FDR` <= 0.05. To read more about these tests, you can follow the citations on the edgeR bioconductor page.

edgeR's bioconductor page.

You might wonder what you can do with these data. Luckily, Trinity also includes scripts to extract differentially expressed transcripts and to create heatmaps.

Change directories into your `edgeR_trans` directory:

```
$ cd edgeR_trans
```

Now we will extract any transcript that is 4-fold differentially expressed between the two conditions at a significance of `<= 0.001` .

Make another job file and choose the short queue and reserve the default RAM (1GB). Load the `bioinformatics/trinity/2.1.1` module. Your command will be:

```
analyze_diff_expr.pl \
      --matrix ../Trinity_trans.TMM.EXPR.matrix \
      --samples ../samples.txt \
      -P 1e-3 -C 2
```

This command will filter transcripts based on pvalue of less than Several files will be written as a part of this job. One is called `diffExpr.P1e-3_C2.matrix` . You can count the number of differentially expressed genes at this threshold by counting the number of lines:

```
$ wc -l diffExpr.P1e-3_C2.matrix
```

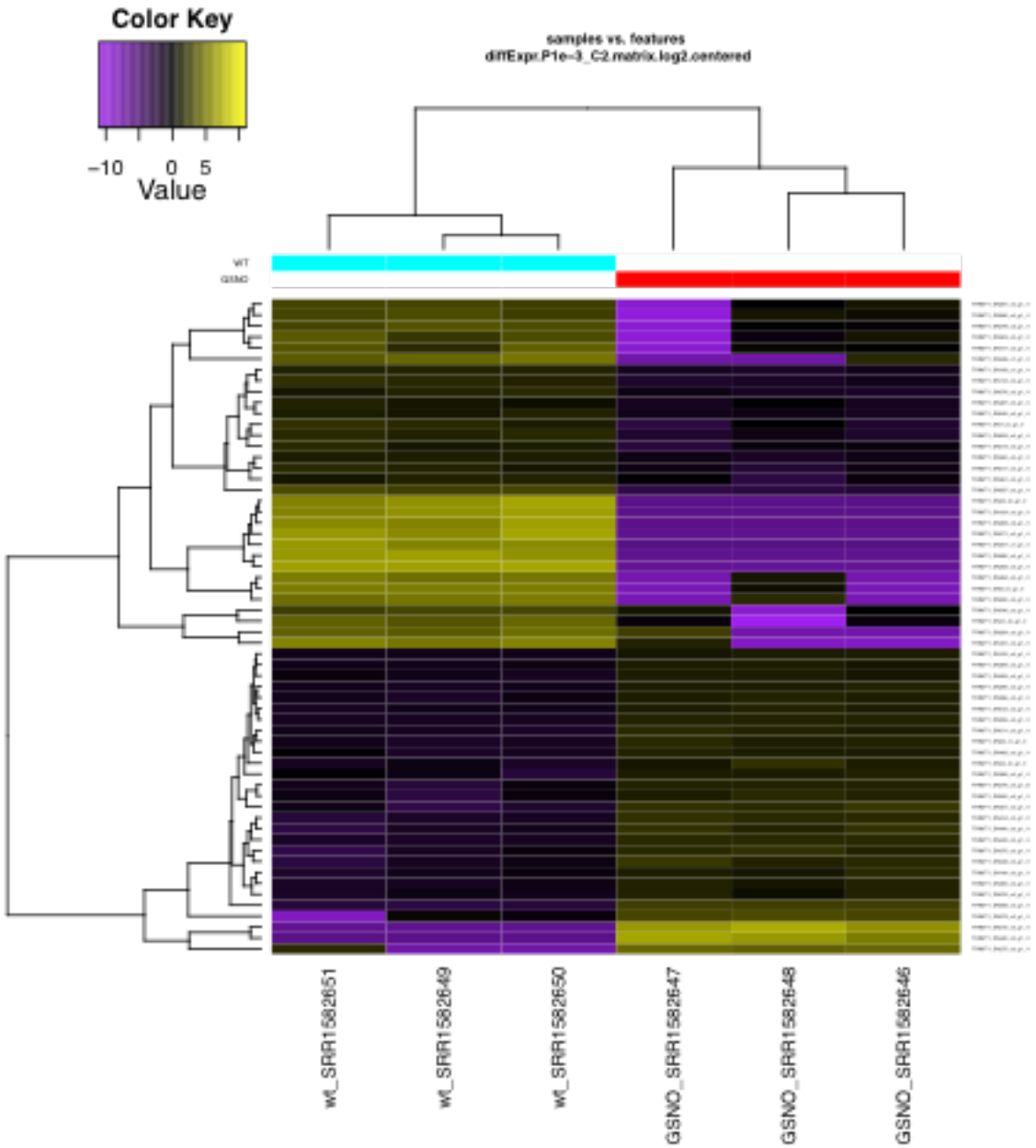
You should subtract 1 from the number since there is a header line.

This script also generates a heatmap that compares the differentially expressed transcripts. The file is called, `diffExpr.P1e-3_C2.matrix.log2.centered.genes_vs_samples_heatmap.pdf` .

Download that file and examine it on your computer.

Hint: you can use `scp` *as above. Or you can use a GUI interface like Filezilla/Cyberduck.*

Now examine the heatmap



You can use the heatmap to compare the two conditions. The left columns with the turquoise line on top are those under wt and the right columns under the red line are under GSNO. Upregulated expression is in yellow and

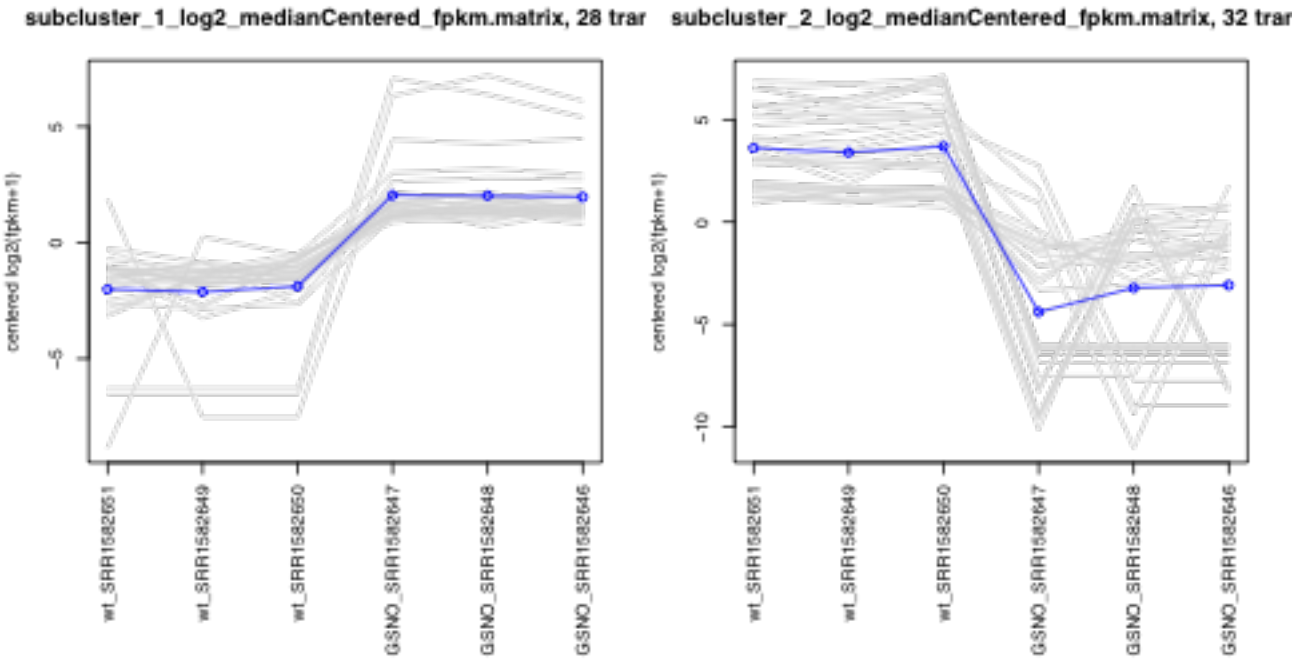
red line are under GSNO. Upregulated expression is in yellow and downregulated expression is in purple. This is a nice visual way to compare expression across conditions.

View transcript clusters

You can also cut the dendrogram to view transcript clusters that share similar expression profiles. To do this, run the following command into a job file. Be sure to load the `bioinformatics/trinity/2.1.1` module and choose a serial job with 1GB of RAM:

```
define_clusters_by_cutting_tree.pl --Ptree 60 -R diffExpr.P1e-3_C2.matrix.RData
```

You should have a new output that looks like the following graph, which shows transcripts with similar expression profiles:



Now run on genes

Now we will run differential expression analysis on the gene level. This will be