Differential Expression and Visualization in R

Taylor Reiter N. Tessa Pierce

Learning objectives:

- Create a gene-level count matrix of Salmon quantification using tximport
- Perform differential expression of a single factor experiment in DESeq2
- Perform quality control and exploratory visualization of RNA-seq data in R

Tximeta

We first need to read our data into R. To do that, we will use a package called tximeta. We use tximeta for two main reasons: first, it facilitates summarizing transcript-level counts from Salmon to the gene-level for differential expression analysis. Second, tximeta enhances incorporates metadata (e.g. transcript locations, transcript and genome source and version, appropriate chromosome lengths, etc) for each transcriptome. This ensures computational reproducibility by attaching critical annotation information to the data object, such that exact quantifications can be reproduced from raw data (all software versions are also attached to the data object).

For more information, see the tximeta vignette

Since we're working in binder, all of the software is already installed. If you'd like to do a similar analysis on your own system, you can use the following installation commands:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
 install.packages("BiocManager")
BiocManager::install("tximeta", version = "3.10")
BiocManager::install("DESeq2")
install.packages("ggplot2")
install.packages("dplyr")
install.packages("readr")
install.packages("pheatmap")
```

```
library(SummarizedExperiment)
library(tximeta)
library(DESeq2)
library(dplyr)
library(ggplot2)
library(readr)
library(pheatmap)
```

This lesson is executed in binder, so the relevant paths should work if we're running this from within the 2020-ggg-201b-rnaseq folder.

First let's read in our `rnaseq_samples.csv` file. This file designates the names of our samples, their conditions, and the path to their quantification files.

```
samples <- read_csv("rnaseq_samples.csv")
```

```
## Parsed with column specification:
## cols(
##   names = col_character(),
##   files = col_character(),
##   condition = col_character()
## )
```

names	files	condition
ERR458493	rnaseq/quant/ERR458493_quant/quant.sf	wt
ERR458494	rnaseq/quant/ERR458494_quant/quant.sf	wt
ERR458495	rnaseq/quant/ERR458495_quant/quant.sf	wt
ERR458500	rnaseq/quant/ERR458500_quant/quant.sf	snf2
ERR458501	rnaseq/quant/ERR458501_quant/quant.sf	snf2
ERR458502	rnaseq/quant/ERR458502_quant/quant.sf	snf2

Next we need to define our reference files. We will use ensembl files, as this simplifies the way we interact with this information.

```
indexDir <- file.path("rnaseq", "quant", "sc_ensembl_index")
fastaFTP <- c("ftp://ftp.ensembl.org/pub/release-99/fasta/saccharomyces_cerevisiae/cdna/Saccharomyces_c
gtfPath <- "ftp://ftp.ensembl.org/pub/release-99/gtf/saccharomyces_cerevisiae/Saccharomyces_cerevisiae.
gtfLocal <- "yeast_ref/Saccharomyces_cerevisiae.R64-1-1.99.gtf.gz"
fastaLocal <- "yeast_ref/Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa.gz"
```

With this information set, we can now build an object that saves this information.

```
makeLinkedTxome(indexDir = indexDir,
                source   = "Ensembl",
                organism = "Saccharomyces cerevisiae",
                release  = "99",
                fasta    = fastaLocal,
                gtf      = gtfLocal,
                genome   = "GCA_000146045.2",
                write    = FALSE)
```

```
## saving linkedTxome in bfc (first time)
```

To create an `se` object for using our `tximeta` information:

```
se <- tximeta(samples)
```

```
## importing quantifications
```

```
## reading in files with read_tsv
```

```
## 1 2 3 4 5 6
## found matching linked transcriptome:
## [ Ensembl - Saccharomyces cerevisiae - release 99 ]
## building EnsDb with 'ensembldb' package
## Importing GTF file ... OK
## Processing metadata ... OK
## Processing genes ...
##  Attribute availability:
##   o gene_id ... OK
##   o gene_name ... OK
##   o entrezid ... Nope
##   o gene_biotype ... OK
```

```
## OK
## Processing transcripts ...
##  Attribute availability:
##   o transcript_id ... OK
##   o gene_id ... OK
##   o transcript_biotype ... OK
## OK
## Processing exons ... OK
## Processing chromosomes ... Fetch seqlengths from ensembl ...
```

```
## Error in readLines(curl(my_url)) :
##   Server denied you to change to the given directory
## Error in readLines(curl(my_url)) :
##   Server denied you to change to the given directory
## Error in readLines(curl(my_url)) :
##   Server denied you to change to the given directory
## Error in readLines(curl(my_url)) :
##   Server denied you to change to the given directory
## Error in readLines(curl(my_url)) :
##   Server denied you to change to the given directory
```

```
## FAIL
## OK
## Generating index ... OK
##   -------------
## Verifying validity of the information in the database:
## Checking transcripts ... OK
## Checking exons ... OK
## generating transcript ranges
```

To look at our `se` object, we can use the following commands:

```
colData(se)
assayNames(se)
rowRanges(se)
seqinfo(se)
```

And to summarize to gene level, we can use `summarizeToGene()`.

```
gse <- summarizeToGene(se)
```

```
## loading existing EnsDb created: 2020-03-09 16:11:31
```

```
## obtaining transcript-to-gene mapping from TxDb
```

```
## generating gene ranges
```

```
## summarizing abundance
```

```
## summarizing counts
```

```
## summarizing length
```

And now let's take a look at the data

```
rowRanges(gse)
mcols(gse)
```

Why do we need to normalize and transform read counts

Given a uniform sampling of a diverse transcript pool, the number of sequenced reads mapped to a gene depends on:

- its own expression level,
- its length,
- the sequencing depth,
- the expression of all other genes within the sample.

In order to compare the gene expression between two conditions, we must therefore calculate the fraction of the reads assigned to each gene relative to the total number of reads and with respect to the entire RNA repertoire which may vary drastically from sample to sample. While the number of sequenced reads is known, the total RNA library and its complexity is unknown and variation between samples may be due to contamination as well as biological reasons. **The purpose of normalization is to eliminate systematic effects that are not associated with the biological differences of interest.**

Normalization aims at correcting systematic technical biases in the data, in order to make read counts comparable across samples. The normalization proposed by DESeq2 relies on the hypothesis that most features are not differentially expressed. It computes a scaling factor for each sample. Normalized read counts are obtained by dividing raw read counts by the scaling factor associated with the sample they belong to.

Differential Expression with `DESeq2`

Image credit: Paul Pavlidis, UBC

Differential expression analysis with DESeq2 involves multiple steps as displayed in the flowchart below. Briefly,

- DESeq2 will model the raw counts, using normalization factors (size factors) to account for differences in library depth.
- Then, it will estimate the gene-wise dispersions and shrink these estimates to generate more accurate estimates of dispersion to model the counts.
- Finally, DESeq2 will fit the negative binomial model and perform hypothesis testing using the Wald test or Likelihood Ratio Test.

We're now ready to use `DESeq2`, the package that will perform differential expression.

We'll start with a function called `DESeqDataSetFromTximport` which will transform our `txi` object into something that other functions in `DESeq2` can work on. This is where we also give information about our samples contain in the `samples` data.frame, and where we provide our experimental design. **A design formula tells the statistical software the known sources of variation to control for, as well as, the factor of interest to test for during differential expression testing. Here our experimental design has one factor with two levels.**

```
dds <- DESeqDataSet(se = gse, design = ~condition)
```

```
## using counts and average transcript lengths from tximeta
```

```
## Warning in DESeqDataSet(se = gse, design = ~condition): some variables in design
## formula are characters, converting to factors
```

Note: DESeq stores virtually all information associated with your experiment in one specific R object, called DESeqDataSet. This is, in fact, a specialized object of the class "SummarizedExperiment". This, in turn,is a container where rows (rowRanges()) represent features of interest (e.g. genes, transcripts, exons) and columns represent samples (colData()). The actual count data is stored in theassay()slot.

The first thing we notice is that both our counts and average transcript length were used to construct the `DESeq` object. We also see a warning message, where our condition was converted to a factor. Both of these messages are ok to see!

Now that we have a DESeq2 object, we can can perform differential expression.

```
dds <- DESeq(dds)
```

```
## estimating size factors
## using 'avgTxLength' from assays(dds), correcting for library size
## estimating dispersions
## gene-wise dispersion estimates
## Warning in seq.default(from = minLogAlpha, to = maxLogAlpha, length = 20):
## partial argument match of 'length' to 'length.out'
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing
```

Everything from normalization to linear modeling was carried out by the use of a single function! This function prints out a message for the various steps it performs.

And look at the results! The results() function lets you extract the base means across samples, log2-fold changes, standard errors, test statistics etc. for every gene.

```
res <- results(dds)
```

```
head(res)
```

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Q0010	0.0000000	NA	NA	NA	NA	NA
Q0017	0.0000000	NA	NA	NA	NA	NA
Q0032	0.0000000	NA	NA	NA	NA	NA
Q0045	0.1128050	0.155906	4.080473	0.0382078	0.9695220	NA
Q0050	0.3403302	-1.327244	2.729003	-0.4863477	0.6267207	NA
Q0055	0.1128743	0.155906	4.080473	0.0382078	0.9695220	NA

The first thing that prints to the screen is information about the "contrasts" in the differential expression experiment. By default, DESeq2 selects the alphabetically first factor to the be "reference" factor. Here that doesn't make that much of a difference. However, it does change how we interpret the log2foldchange values. We can read these results as, "Compared to *SNF2* mutant, WT had a decrease of -0.2124 in log2fold change of gene expression.

Speaking of log2fold change, what do all of these columns mean?

baseMean	giving means across all samples
log2FoldChange	log2fold changes of gene expression from one condition to another. Reflects how different the expression of a gene in one condition is from the expression of the same gene in another condition.
lfcSE	standard errors (used to calculate p value)
stat	test statistics used to calculate p value)
pvalue	p-values for the log fold change
padj	adjusted p-values

We see that the default differential expression output is sorted the same way as our input counts. Instead, it can be helpful to sort and filter by adjusted p value or log2 Fold Change:

```
res_sig <- subset(res, padj<.05)
res_lfc <- subset(res_sig, abs(log2FoldChange) > 1)
```

```
head(res_lfc)
```

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
YAL005C	1819.26712	1.419577	0.0443426	32.013882	0.0000000	0.0000000
YAL025C	62.81774	1.059898	0.2297107	4.614057	0.0000039	0.0000387
YAL026C-A	15.59573	-1.452192	0.5120298	-2.836147	0.0045661	0.0224178
YAL038W	8285.92305	1.607959	0.0219170	73.365700	0.0000000	0.0000000
YAL044C	200.55284	1.731466	0.1377790	12.566978	0.0000000	0.0000000
YAL054C	62.63066	-1.179617	0.2465120	-4.785232	0.0000017	0.0000175

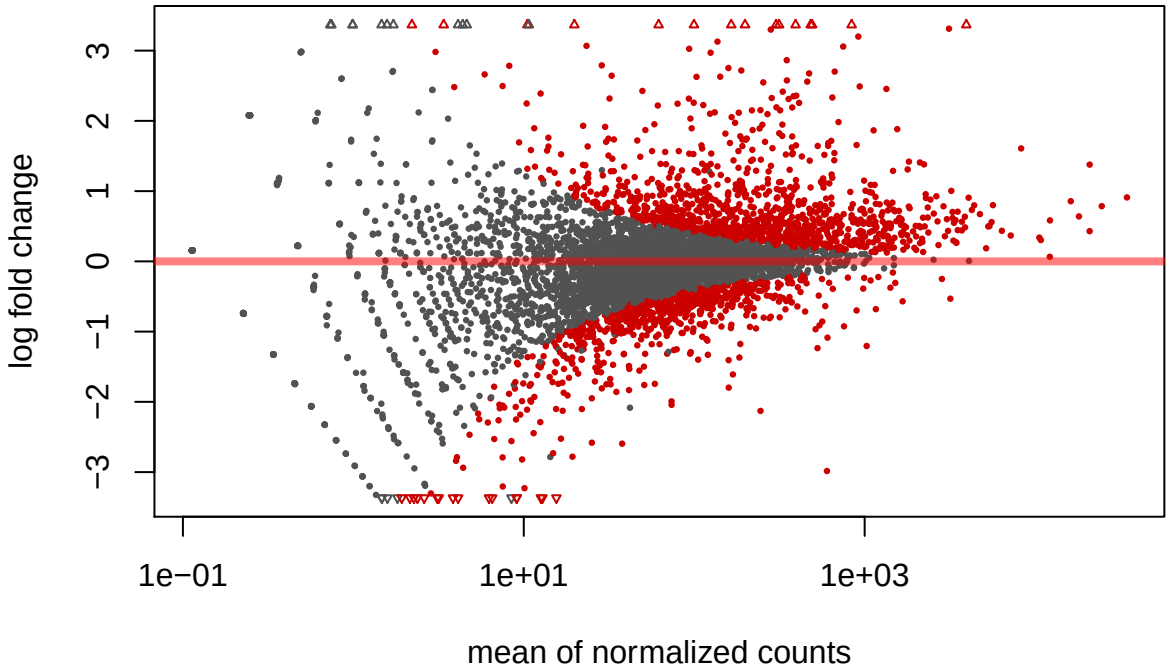
Visualization of RNA-seq and Differential Expression Results

Looking at our results is great, but visualizing them is even better!

MA Plot

The MA plot provides a global view of the relationship between the expression change between conditions (log ratios, M), the average expression strength of the genes (average mean, A) and the ability of the algorithm to detect differential gene expression: genes that pass the significance threshold are colored in red

```
plotMA(res)
```

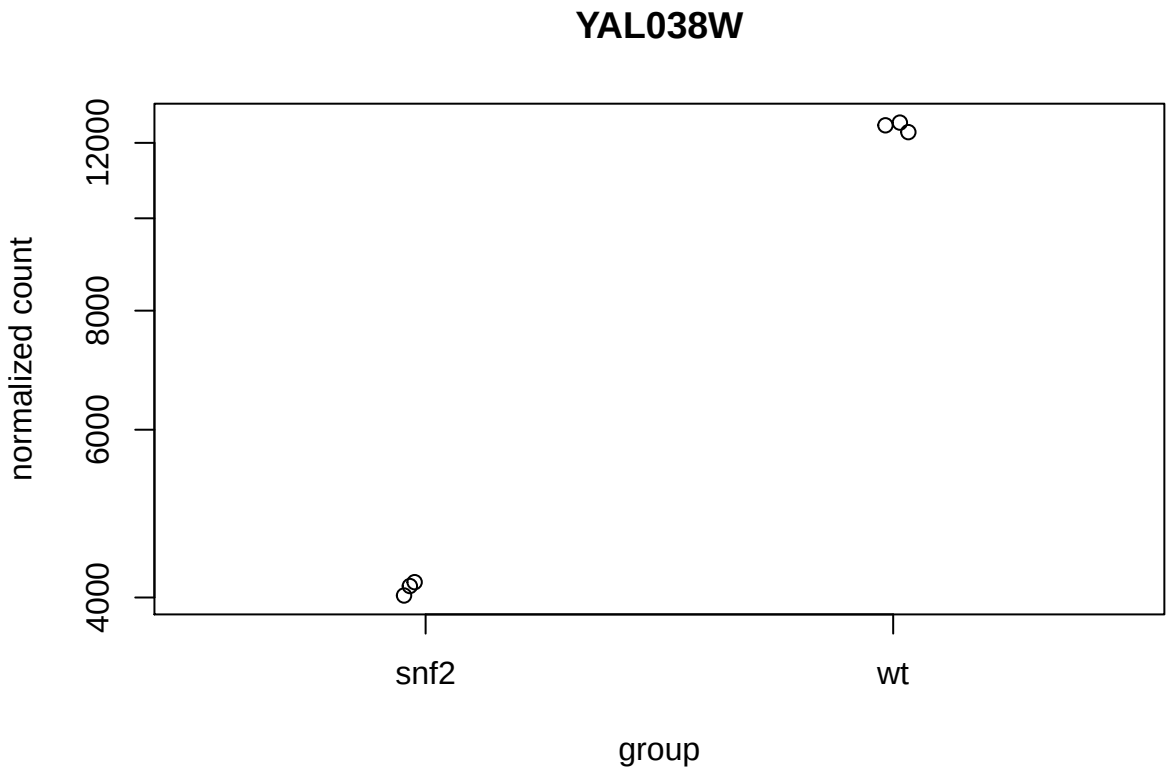


Question
Why are more genes grey on the left side of the axis than on the right side?

Plotting individual genes

Although it's helpful to plot many (or all) genes at once, sometimes we want to see how counts change for a specific gene. We can use the following code to produce such a plot:

```
plotCounts(dds, gene=which.min(res$padj), intgroup="condition")
```



Question
What gene is plotted here (i.e., what criteria did we use to select a single gene to plot)?

Normalization & Transformation

`DESeq2` automatically normalizes our count data when it runs differential expression. However, for certain plots, we need to normalize our raw count data. One way to do that is to use the `vst()` function. It perform variance stabilized transformation on the count data, while controlling for library size of samples.

```
vsd <- vst(dds)
```

```
## Warning in seq.default(from = 1, to = length(o), length = nsub): partial
## argument match of 'length' to 'length.out'
```

```
## Warning in seq.default(from = minLogAlpha, to = maxLogAlpha, length = 20):
## partial argument match of 'length' to 'length.out'
```

MDS Plot

An MDS (multi-dimensional scaling) plot gives us an idea of how our samples relate to each other. The closer two samples are on a plot, the more similar all of their counts are. To generate this plot in DESeq2, we need to calculate "sample distances" and then plot them.

```
# calculate sample distances
sample_dists <- assay(vsd) %>%
  t() %>%
  dist() %>%
  as.matrix()
```

```
head(sample_dists)
```

	ERR458493	ERR458494	ERR458495	ERR458500	ERR458501	ERR458502
ERR458493	0.00000	19.18457	19.22584	35.03510	35.16412	35.23432
ERR458494	19.18457	0.00000	18.67720	35.14104	35.08395	35.15756
ERR458495	19.22584	18.67720	0.00000	35.19157	35.34372	35.52140
ERR458500	35.03510	35.14104	35.19157	0.00000	12.71084	12.45395
ERR458501	35.16412	35.08395	35.34372	12.71084	0.00000	12.74785
ERR458502	35.23432	35.15756	35.52140	12.45395	12.74785	0.00000

Next, let's calculate the MDS values from the distance matrix.

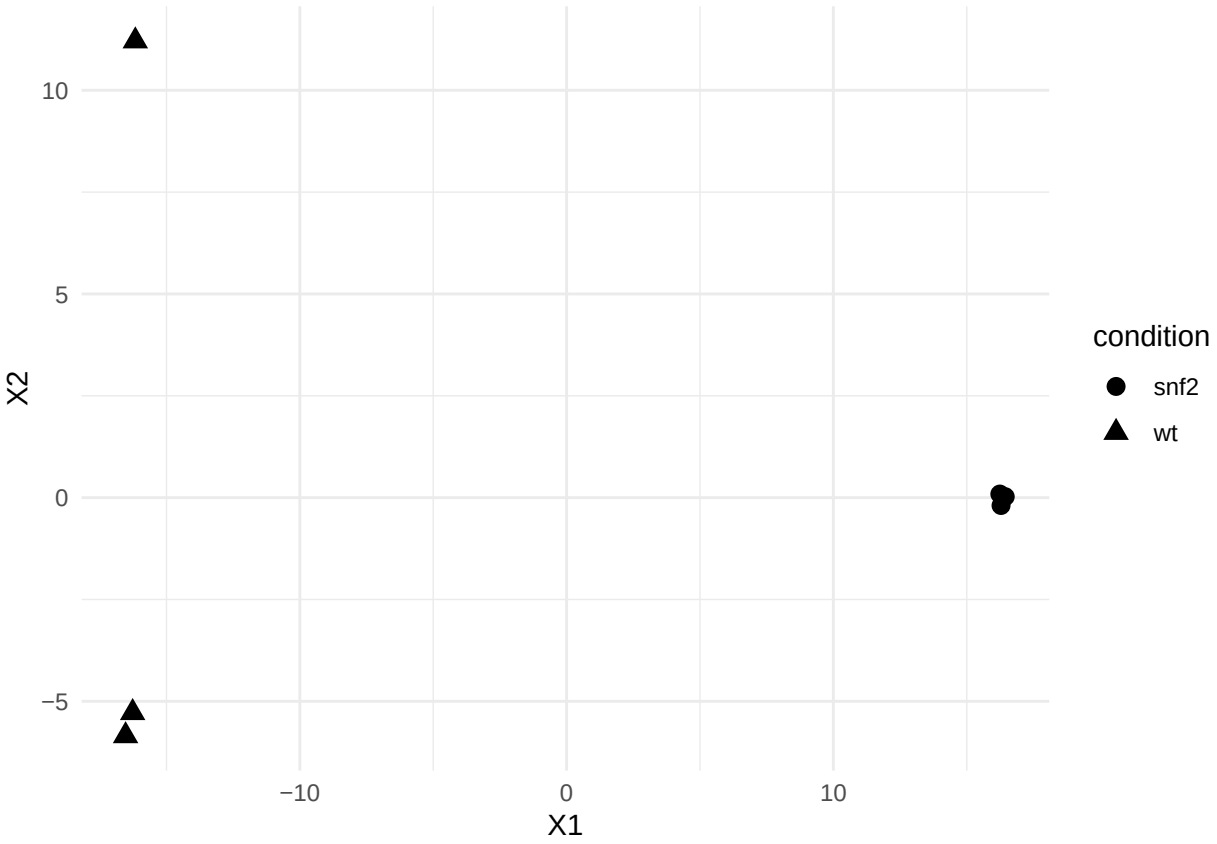
```
mdsData <- data.frame(cmdscale(sample_dists))
mds <- cbind(mdsData, as.data.frame(colData(vsd))) # combine with sample data
```

```
head(mds)
```

	X1	X2	names	condition
ERR458493	-16.16718	11.2092088	ERR458493	wt
ERR458494	-16.26755	-5.2798227	ERR458494	wt
ERR458495	-16.53186	-5.8473570	ERR458495	wt
ERR458500	16.23973	0.0878134	ERR458500	snf2
ERR458501	16.28481	-0.1931173	ERR458501	snf2
ERR458502	16.44205	0.0232748	ERR458502	snf2

And plot with `ggplot2`!

```
ggplot(mds, aes(X1, X2, shape = condition)) +
  geom_point(size = 3) +
  theme_minimal()
```



Question
How similar are the samples between conditions?

Heatmap

Heatmaps are a great way to look at gene counts. To do that, we can use a function in the `pheatmap` package. Here we demonstrate how to install a package from the CRAN repository and then load it into our environment.

Next, we can select a subset of genes to plot. Although we could plot all ~6000 yeast genes, let's choose the 20 genes with the largest positive log2fold change.

```
genes <- order(res_lfc$log2FoldChange, decreasing=TRUE)[1:20]
```

We can also make a data.frame that contains information about our samples that will appear in the heatmap. We will use our samples data.frame from before to do this.

```
annot_col <- samples %>%
        tibble::column_to_rownames('names') %>%
        select(condition) %>%
        as.data.frame()
```

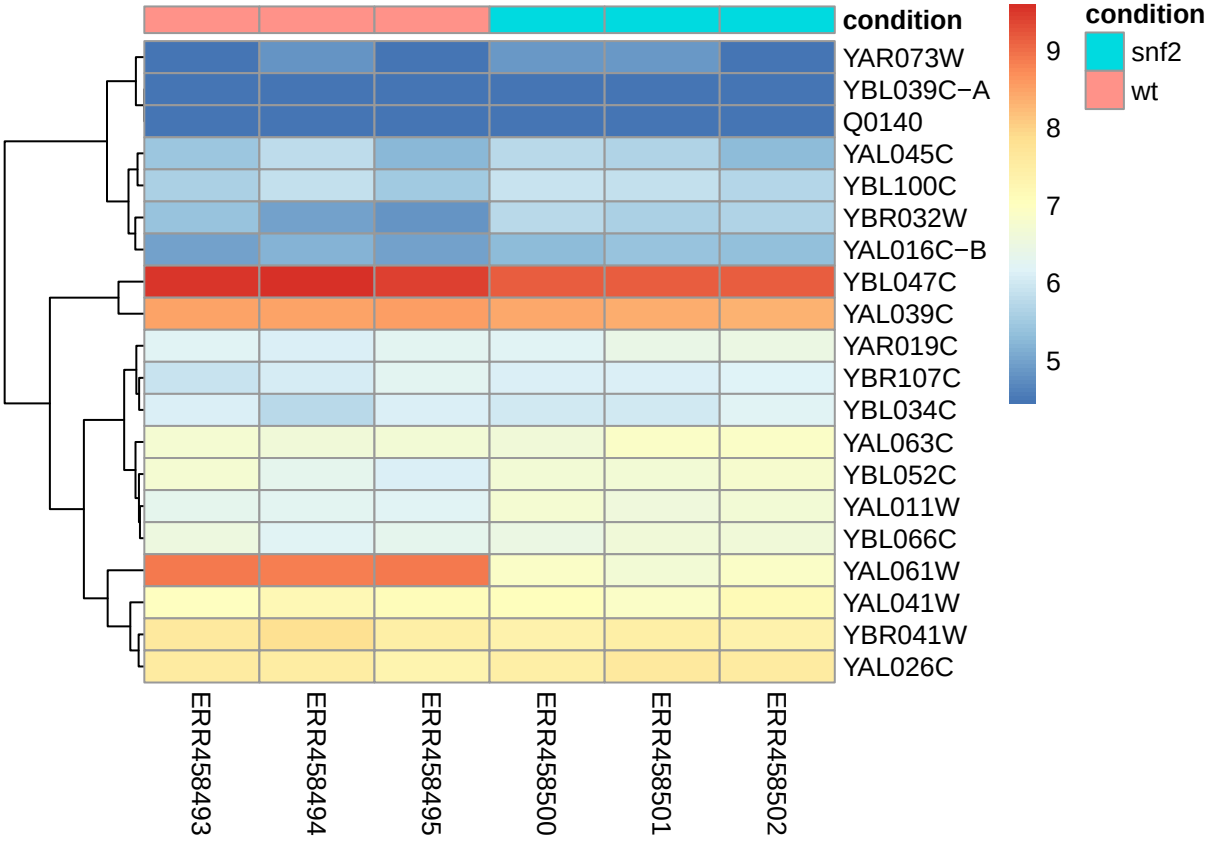
```
head(annot_col)
```

	condition
ERR458493	wt
ERR458494	wt
ERR458495	wt
ERR458500	snf2
ERR458501	snf2
ERR458502	snf2

And now plot the heatmap!

```
pheatmap(assay(vsd)[genes, ], cluster_rows=TRUE, show_rownames=TRUE,
         cluster_cols=FALSE, annotation_col=annot_col)
```

```
## Warning: partial match of 'just' to 'justification'
```



We see that our samples do cluster by condition, but that by looking at just the counts, the patterns aren't very strong. How does this compare to our MDS plot?

Question
When are heatmaps useful?
What other types of heatmaps have you seen in the wild?

Further Notes

Here are some helpful notes or resources for anyone performing differential expression.

- Introduction to differential gene expression analysis using RNA-seq (Written by Friederike Dündar, Luce Skrabanek, Paul Zumbo). Click here
- Introduction to DGE - click here