STAT718/BIOL703: Genomic Data Science
Gene-level DE Analysis with DESeq2 II

Dr. Yen-Yi Ho (hoyen@stat.sc.edu)

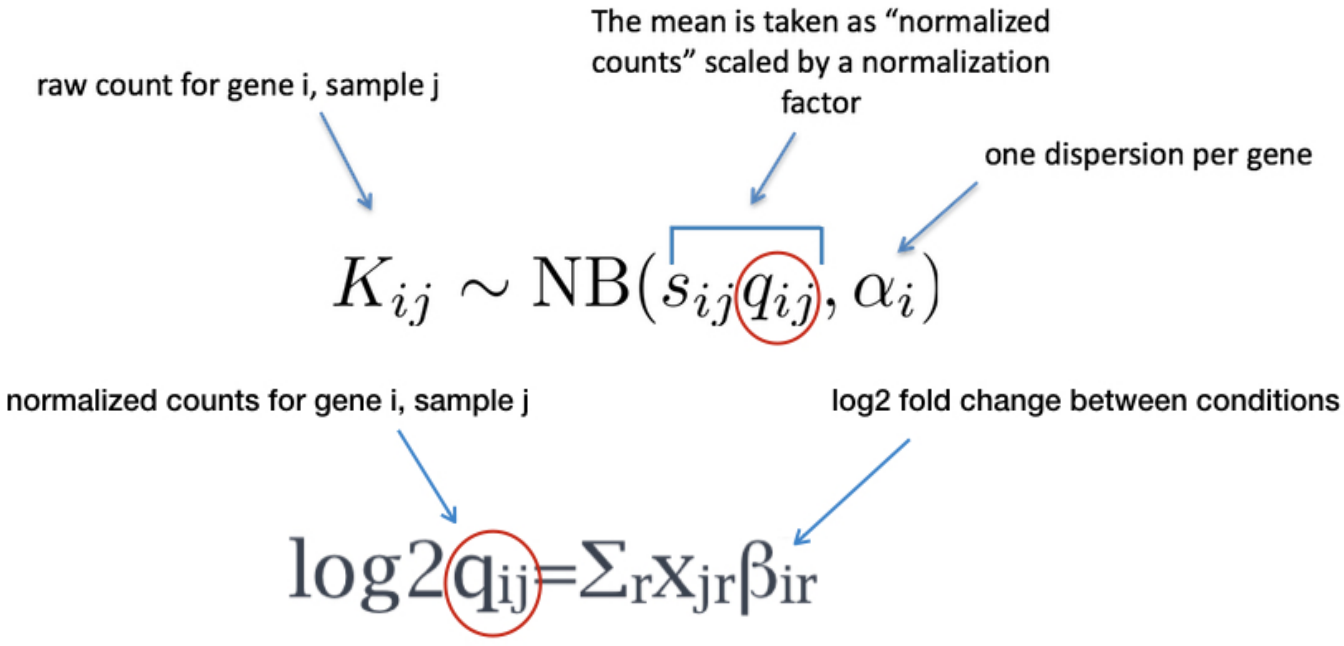
Learning Objectives

- Understanding the different steps in a differential expression analysis in the context of DESeq2
- Building results tables for comparison of different sample classes
- Summarizing significant differentially expressed genes for each comparison

Differential expression analysis with DESeq2: model fitting and hypothesis testing

Generalized Linear Model fit for each gene

Negative Binomial Model



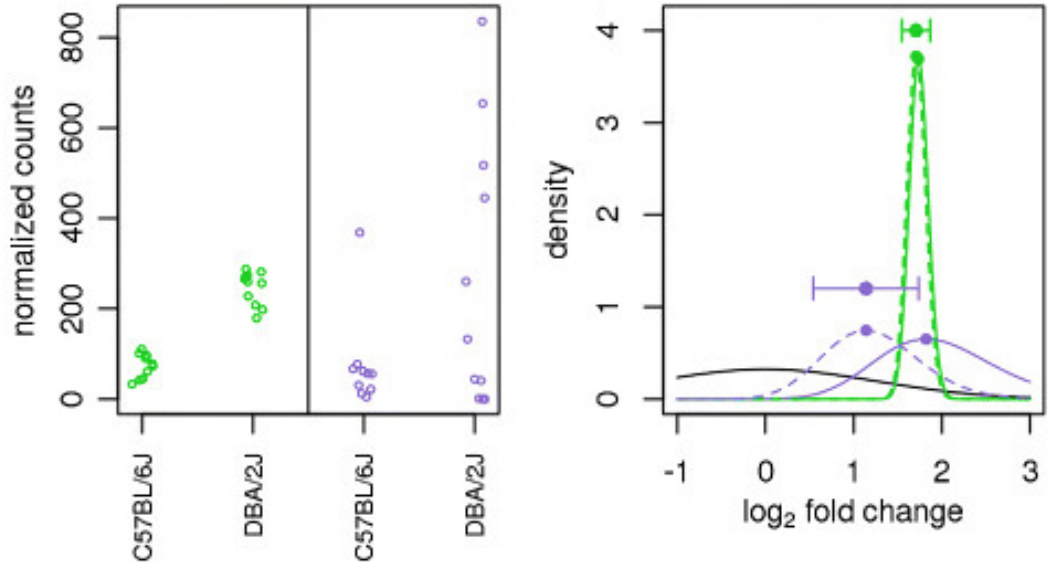
The coefficents are the estimates for the **log2 foldchanges** for each sample group.
log 2 (normalized_counts_group1 / normalized_counts_group2)

Shrunken log2 foldchanges (LFC)

To generate more accurate log2 foldchange estimates, DESeq2 allows for the **shrinkage of the LFC estimates toward zero** when the information for a gene is low, which could include:

- Low counts
- High dispersion values

As with the shrinkage of dispersion estimates, LFC shrinkage uses **information from all genes** to generate more accurate estimates.



Hypothesis testing using the Wald test

The first step in hypothesis testing is to set up a **null hypothesis** for each gene. In our case is, the null hypothesis is that there is **no differential expression across the two sample groups (LFC == 0)**.

A Wald test statistic is computed along with a probability that a test statistic at least as extreme as the observed value were selected at random. This probability is called the p-value of the test. If the p-value is small we reject the null hypothesis and state that there is evidence against the null (i.e. the gene is differentially expressed).

Contrast and Wald Tests

MOV10 DE analysis: contrasts and Wald tests

We have three sample classes so we can make three possible pairwise comparisons:

1. Control vs. primary colorectal cancer
2. Control vs. normal-looking surrounding colonic epithelium

Using the design formula we provided `~ sampletype`, indicating that this is our main factor of interest.

Contrast and Wald Tests

Building the results table

To build our results table we will use the `results()` function.

```
## Define contrasts, extract results table,
## and shrink the log2 fold changes

contrast_oe <- c("tissueype", "primary", "normal")

res_tableOE_unshrunken <- results(dds, contrast=contrast_oe, alpha = 0.05)

res_tableOE <- lfcShrink(dds, contrast=contrast_oe, res=res_tableOE_unshrunken)
```

The order of the names determines the direction of fold change that is reported. The name provided in the second element is the level that is used as baseline.

MA Plot

The MA plot shows the mean of the normalized counts versus the log2 foldchanges for all genes tested. The genes that are significantly DE are colored to be easily identified.

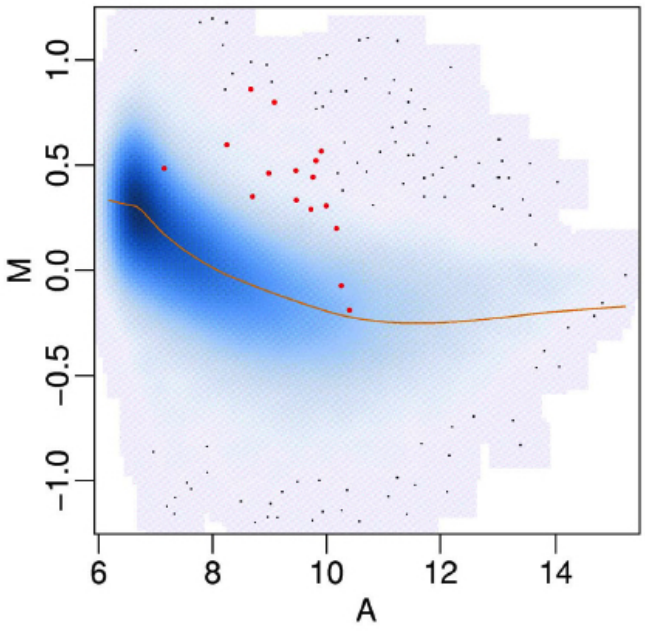
Let's start with the unshrunken results:

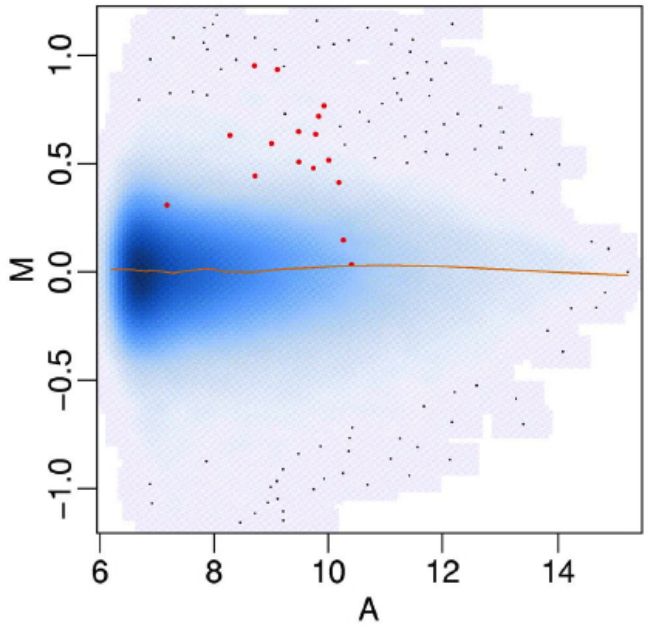
```
plotMA(res_tableOE_unshrunken, ylim=c(-2,2))
```

And now the shrunken results:

```
plotMA(res_tableOE, ylim=c(-2,2))
```

MA plots



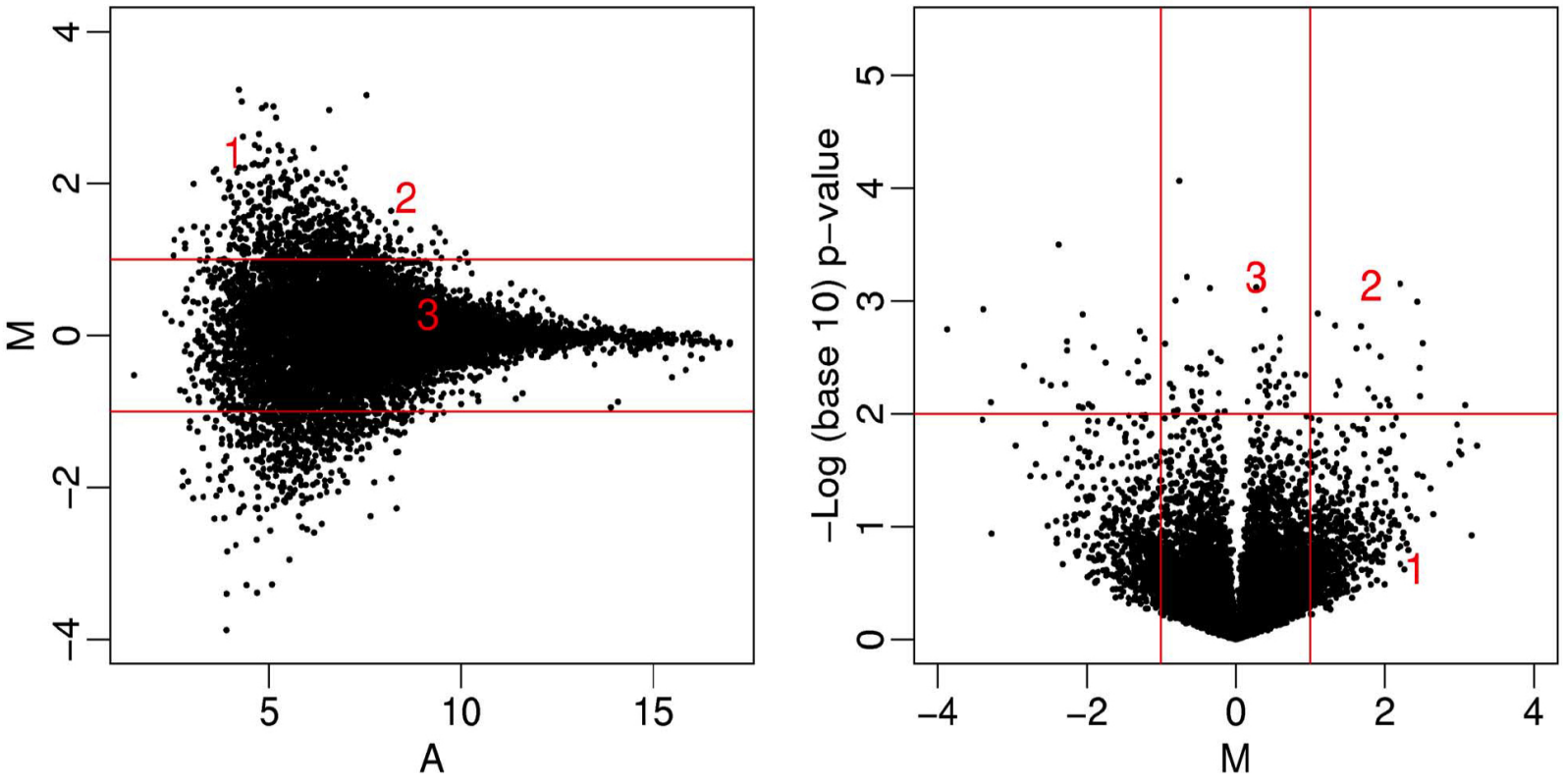


$$M = \log R_1 - \log R_2$$
$$A = (\log R_1 + \log R_2)/2$$

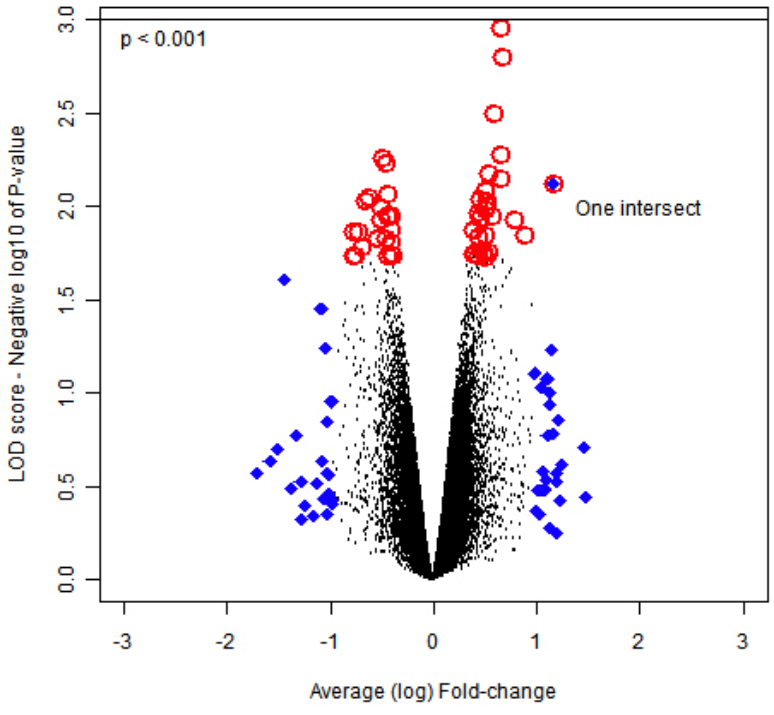
Volcano Plot

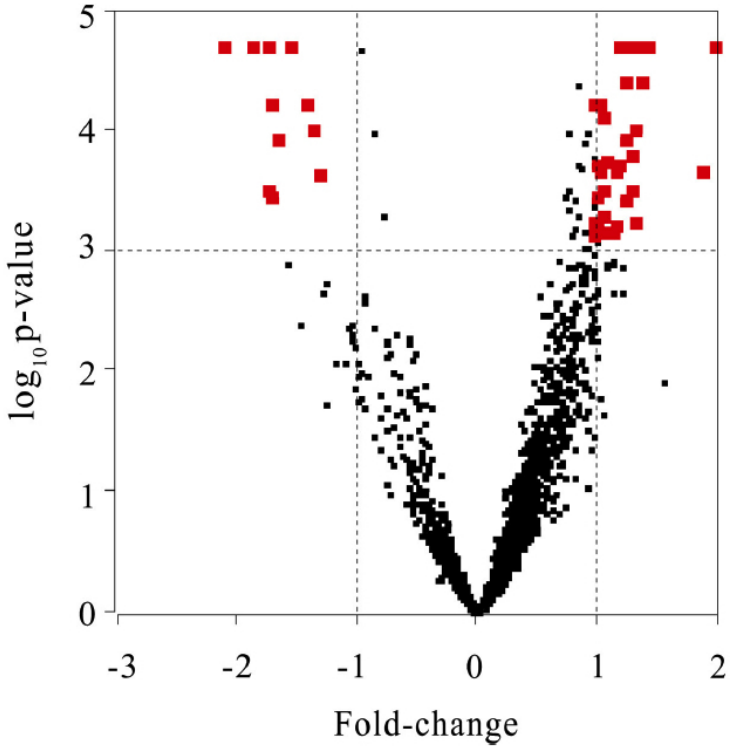
- A diagnostic plot to visualize the test results
- Scatter plot of statistical significance (log p values) versus biological significance (log fold-changes)
- Ideally the two should agree with each other

MA and Volcano Plots



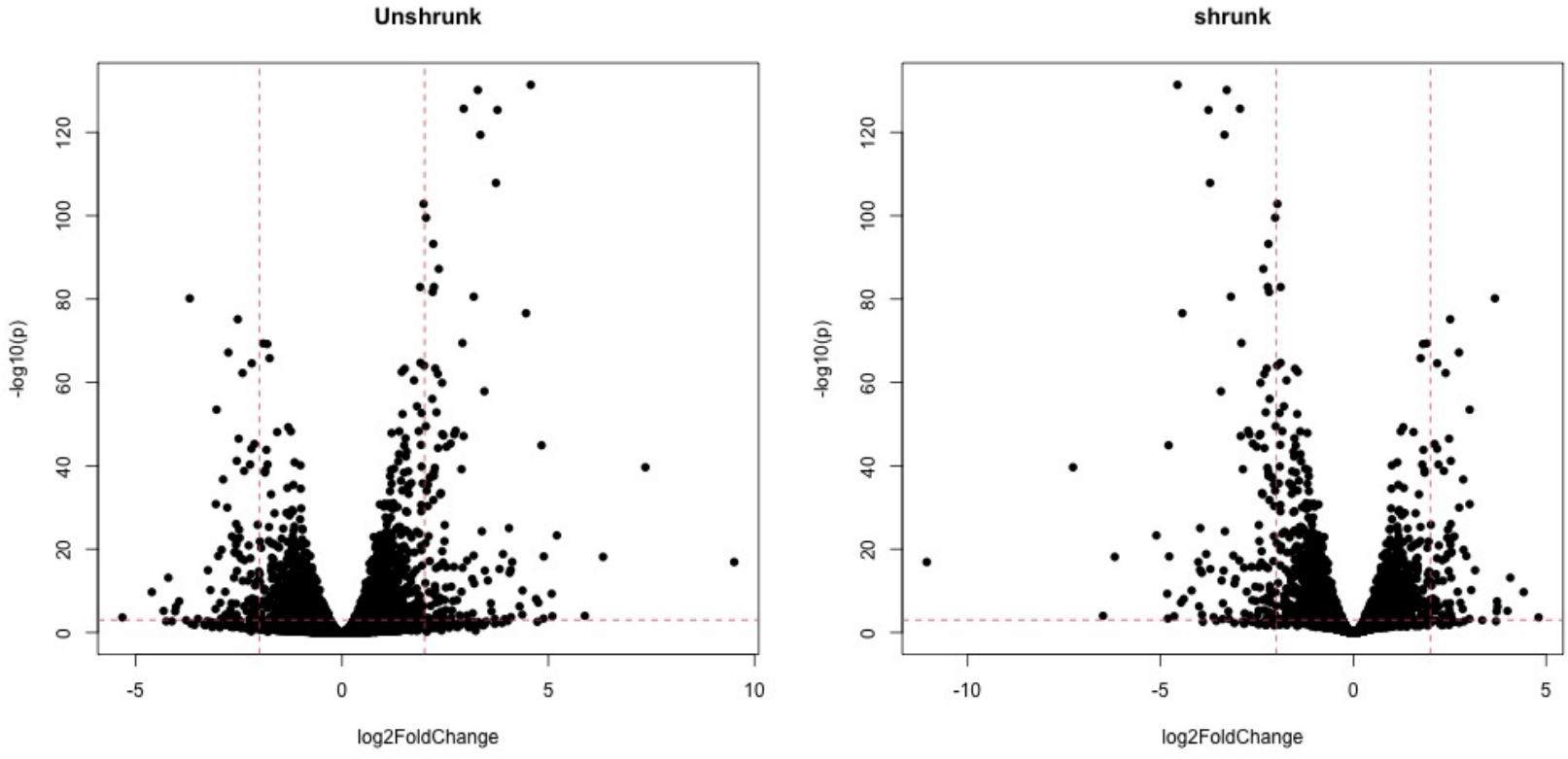
Volcano Plots: Bad Versus Good





When sample size is small, SD estimates in t-test are unstable.

Volcano Plots: unshrunken vs. shrunken



Multiple Comparisons

There are a few common approaches:

- **Bonferroni:** The adjusted p-value is calculated by: p-value * m (m = total number of tests). **This is a very conservative approach with a high probability of false negatives**, so is generally not recommended.
- **FDR/Benjamini-Hochberg:** Benjamini and Hochberg (1995) defined the concept of FDR and created an algorithm to control the expected FDR below a specified level given a list of independent p-values. **An interpretation of the BH method for controlling the FDR is implemented in DESeq2 in which we rank the genes by p-value, then multiply each ranked p-value by m/rank**.
- **Q-value / Storey method:** The minimum FDR that can be attained when calling that feature significant. For example, if gene X has a q-value of 0.013 it means that 1.3% of genes that show p-values at least as small as gene X are false positives

  So what does FDR $<$ 0.05 mean? *By setting the FDR cutoff to $<$ 0.05, we're saying that the proportion of false positives we expect amongst our differentially expressed genes is 5%. For example, if you call 500 genes as differentially expressed with an FDR cutoff of 0.05, you expect 25 of them to be false positives.*

More about multiple comparisons to come ...