“Proving” a hypothesis?

Given what we have just learned, is it ever possible to definitively prove a hypothesis using hypothesis testing?

NO, the best we can do is to assign a probability to the null hypothesis being true.

Adjusted p-value

- **If Mel performed RNA sequencing on the irises…**
  - –Some iris genomes have over 100,000 genes!
  - –Each gene is a variable
  - –RNA sequencing would give ~100,000 variables (genes) to compare between the iris species

- **For data with thousands of variables, it is important to mathematically adjust p-values for multiple comparison testing to avoid false positives.**
  - – Multiple comparison testing = comparing multiple variables between conditions

Why is it important to adjust p-values for multiple comparisons?

- **If Mel looked at the difference in expression of a random set of 20 genes:**
  - –*By chance* we would expect that on average 1 out of 20 genes would appear significant using the 0.05 (5%) p-value cutoff.
  - –This is a false positive - a result that appears significant but in reality is due to random chance.
  - –Mathematical p-value adjustment takes into account the probability of false positives given the number of variables

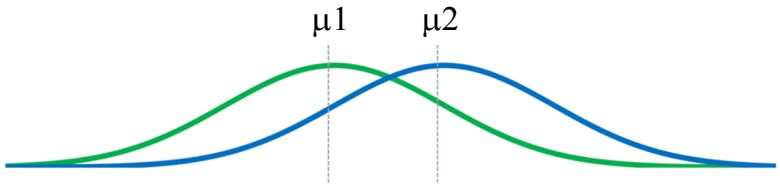
- **Many bioinformatics statistical tools will generate, for each variable:**
  - –a p-value (does not take false positives into account)
  - –an adjusted p-value (does take false positives into account)

Adjusted p-value questions

1. **What should be used when determining the significance of an experiment where ~100,000 genes (aka variables) are considered, p-value or adjusted p-value, and why?**
   – Adjusted p-value, because it takes into account the likelihood of false positives.
2. **When is it appropriate to use a p-value rather than an adjusted p-value to determine significance, given a p-value cutoff of 0.05 (1/20)?**
   – In an experiment where there are few enough variables that false positives are not reasonably expected. If you are using a p-value cutoff of 0.05 (i.e., 1 out of 20), you would want far fewer variables than 20.
3. **Which will generate more significantly differentially expressed genes in Mel's experiment, the p-value or the adjusted p-value?**
   – The p-value, since it is a less stringent metric.

How do we actually calculate a p-value?

- **Use a statistical hypothesis testing method such as a T-Test**
  - –Compares the means of all measured variables between two groups



Expression of Gene X in *setosa*

Expression of Gene X in *virginica*

$$tScore = \frac{signal}{noise} = \frac{difference\ in\ group\ means}{variability\ of\ groups} = \frac{\mu_1 - \mu_2}{\sigma_1 - \sigma_2}$$

μ_1 = mean expression of gene X in *setosa*
μ_2 = mean expression of gene X in *virginica*

σ_1 = standard deviation of gene X in *setosa*
σ_2 = standard deviation of gene X in *virginica*

Standard deviation comes in handy again!

What are the results returned from a T-test?

- **A *t-score* / *t-value* for each variable**
  - –The greater the t-value, the more difference there is within that variable between the two groups

- **A p-value associated with each t-value**
  - –Describes the likelihood of the null hypothesis being true for that variable

Another hypothesis testing method: **Wald test**

- **Creates a mathematical model for the null hypothesis and the testing hypothesis, and identifies which model best maximizes the likelihood of the data**
  - Mathematical model = mathematical description of a situation, specifying rules about how data should behave within that situation.

- **Put another way: the Wald test identifies a mathematical model for which the data we have is *the most likely result***

- **Returns a p-value for each variable**

Questions about T-tests and Wald Tests

1. **Suppose Mel's final gene expression data has measurements from 100,000 genes. If Mel performs a T-test to compare the two groups, how many t-values (and p-values) are expected?**
   – We would expect 100,000 t-values and 100,000 p-values: one for each gene, because in gene expression data, each gene is a variable.

2. **If Mel performs a Wald test to compare these two groups, how many p-values are expected?**
   – We would expect 100,000 p-values: one for each gene.

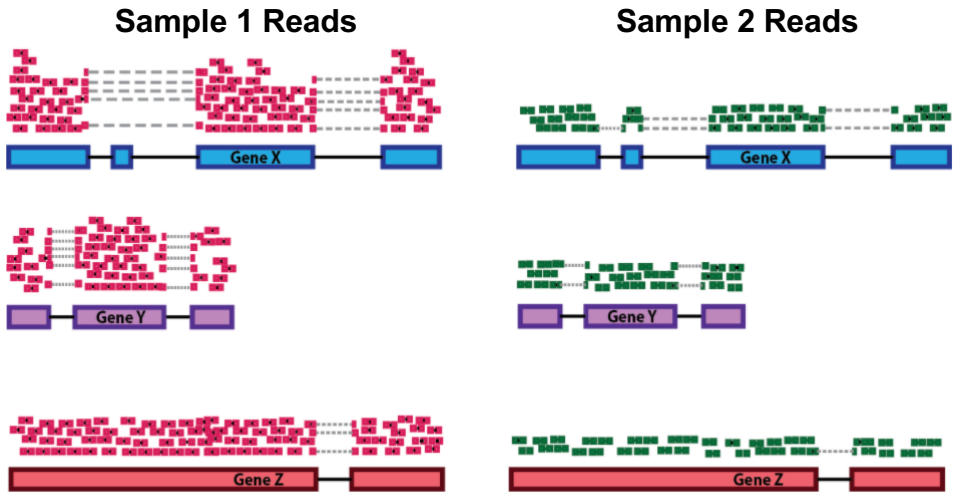
Normalizing gene expression data for differences in **read depth**

- **Gene expression data often have different *read depth* levels per sample**
  - –Read depth = the total number of reads that come off the sequencer for each sample

- **The number of reads in a sequencing run is split up between the number of samples**
  - –Thus, the read depth for each sample depends on the total number of samples

Sample 1 has twice as many reads as Sample 2

- **If Mel directly compares Sample 1 and Sample 2 without normalization, it looks like all genes are more highly expressed in Sample 1**



- **Thus, Mel first must normalize the samples by their read depth, to allow for direct comparison of gene expression values between samples.**

Normalization with Median of Ratios Method (DESeq2 Step 1)

➢**DESeq2 normalizes gene expression data *by read depth* using the Median of Ratios method.**

1. For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

$$GM = \sqrt[n]{x_1 * x_2 * x_3 \ldots * x_n}$$

gene	Sample1	Sample2	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

Normalization with Median of Ratios Method

2. For every gene in a sample, the ratios (sample/pseudo-reference) are calculated. This is performed for each sample in the dataset. Since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample.

gene	Sample1	Sample2	pseudo-reference sample	ratio of Sample1/ref	ratio of Sample2/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

Normalization with Median of Ratios Method

3. The median (middle) value of all ratios for a given sample is taken as the normalization factor (size factor) for that sample.

gene	Sample1	Sample2	pseudo-reference sample	ratio of Sample1/ref	ratio of Sample2/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

```
normalization_factor_Sample1 <- median(c(1.28, 1.3, 1.29, 1.35, 0.59))
normalization_factor_Sample2 <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

Normalization with Median of Ratios Method

4. Each raw count value in each sample is divided by that sample's normalization factor to generate normalized count values.

Raw Counts

gene	Sample1	Sample2
EF2A	1489	906
ABCD1	22	13
...	...	...

Normalized Counts

gene	Sample1	Sample2
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

Normalization with Median of Ratios Method

1. **If we did not perform read depth normalization, what result would you expect if we compared gene expression values between Sample 1 (read depth = 5 million reads) and Sample 2 (read depth = 20 million reads)?**

   Most, if not all of the differentially expressed genes would display overexpression in Sample 2 because more raw counts are assigned to each gene in Sample 2 due to deeper sequencing depth.

2. **Why would this result be problematic?**

   The differentially expressed genes may or may not be representative of the true biology of the two groups, and it would be impossible to tell which genes' overexpression were due to true biology and which were due to the sequencing depth difference.

Differential expression analysis in DESeq2

- **After normalization, Mel is ready to perform differential expression analysis**

- **~100,000 hypothesis tests run at once (one for each gene)**

- **DESeq2 uses a Wald test for differential expression analysis hypothesis testing**

- **What is the null hypothesis we are testing for each gene?**

  "This gene does not change expression significantly between the tested groups"

Estimating gene-wise variance/dispersion before Wald test (DESeq2 Step 2)

- **For genes with low mean counts, variance/dispersion estimates are often unreliable**

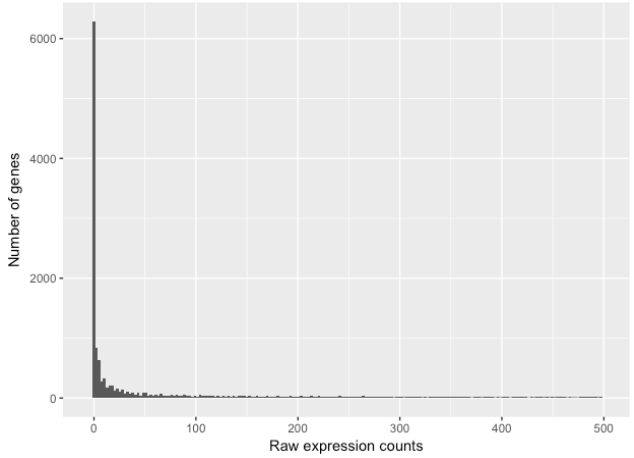
- **To solve this problem, DESeq2 estimates variance/dispersion by sharing information across genes**

  - With this approach, DESeq2 assumes that genes with similar expression levels have similar dispersion/variance

- **The dispersion for each gene is estimated using maximum likelihood estimation. This means that the most likely estimate of dispersion is calculated for each gene, given the gene expression values (aka gene counts) from all the replicates in each group.**

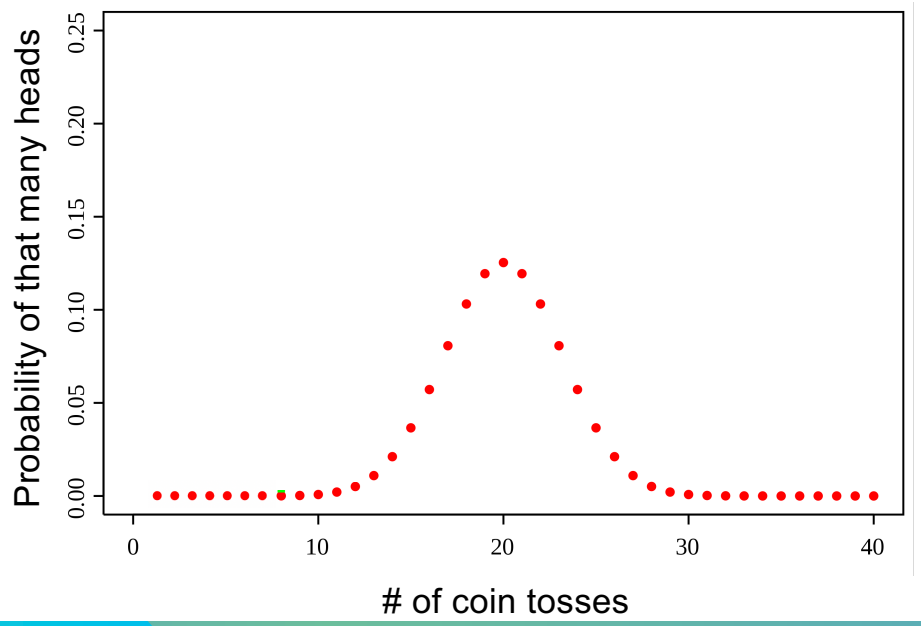
Picking a mathematical model for a Wald test on gene counts data (DESeq2 Step 3)

- **Recall that the Wald test creates a mathematical model for the null hypothesis and the testing hypothesis, and identifies which model best maximizes the likelihood of the data**
- **The type of model is dependent on the shape ("distribution") of the data**
  - A model = a distribution shape, and the numbers that specify that shape (e.g. mean and variance)
  - The Wald test has to be told which general type of distribution to use
- **RNAseq data has a specific shape: the majority of the genes have an expression value close to 0, and there is a long right tail because there is no upper limit for maximum expression value:**



Count data is sometimes modeled with the **binomial distribution**

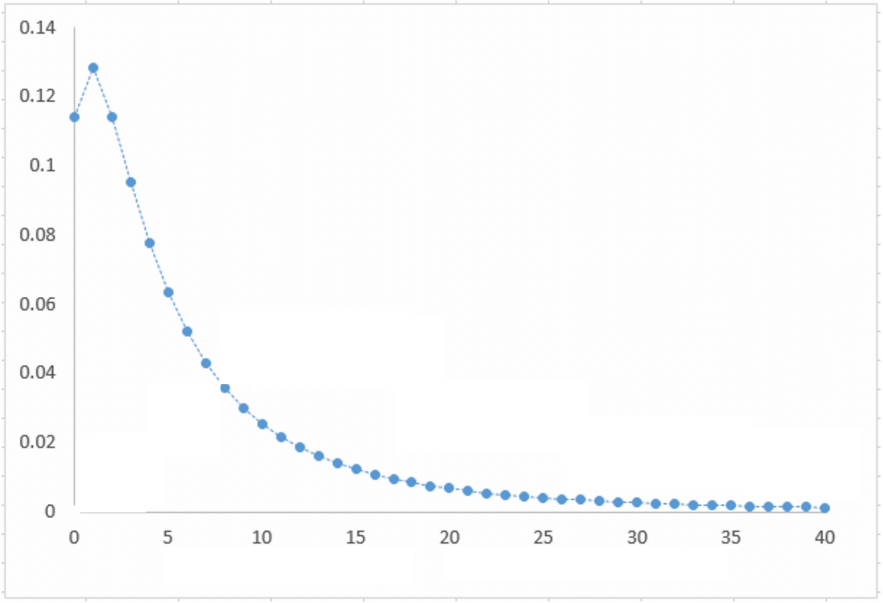
- **The binomial distribution is the “coin toss” distribution**
- **Only appropriate for relatively small number of data points**

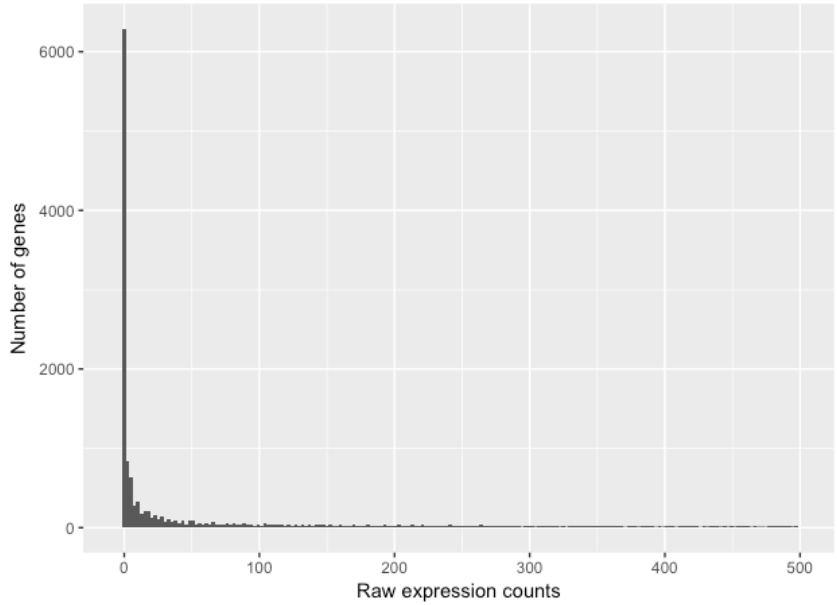


https://en.wikipedia.org/wiki/Binomial_distribution#/media/File:Binomial_distribution_pmf.svg

The negative binomial distribution is appropriate for RNA sequencing data

- **Therefore, DESeq2 uses the negative binomial distribution to create models in the Wald test for differential gene expression analysis**





https://www.vosesoftware.com/riskwiki/BetaNegBindistribution.php

Visualizing RNA sequencing data

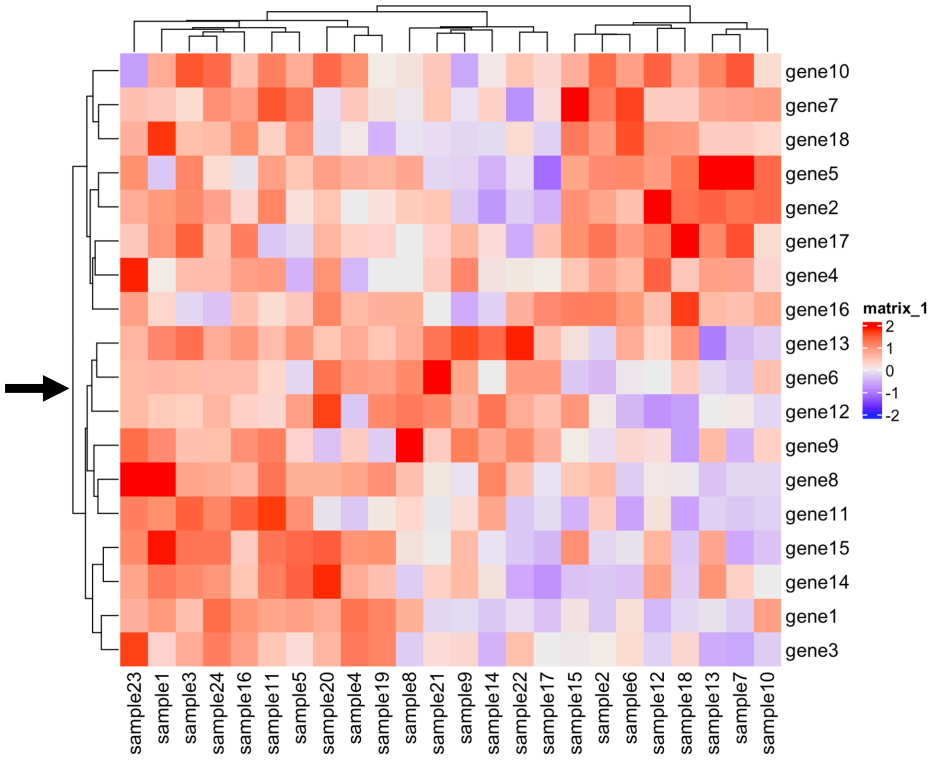
- **After differential expression analysis, we may have hundreds or thousands of differentially expressed genes**

- **For the human eye, it is easier to interpret a visualization or figure**

- **Two common visualizations for RNA sequencing are**
  - Heatmap
  - Volcano plot

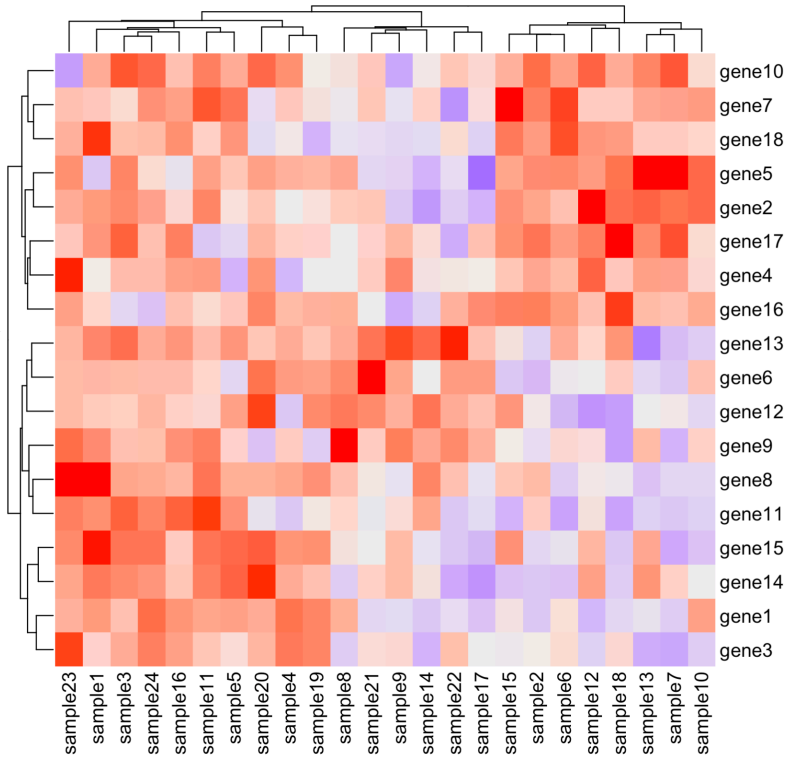
Heatmap

	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	sample 7	sample 8	sample 9	…
gene1	0.9	-0.35	0.5	1.27	0.83	0.16	-0.29	0.68	-0.16	…
gene2	0.91	0.79	1.07	0.01	0.14	0.47	1.28	0.39	-0.35	…
gene3	0.28	0.03	0.71	1.22	0.17	0.21	-0.64	-0.31	0.22	…
gene4	0.03	0.76	0.53	-0.5	-0.53	0.57	0.81	-0.01	1.08	…
gene5	-0.33	1.03	1.12	0.67	0.45	1.04	2.62	0.76	-0.23	…
gene6	0.58	-0.47	0.55	0.87	-0.2	-0.04	-0.32	1.06	0.78	…
gene7	0.4	1.15	0.21	0.41	1.26	1.63	0.81	-0.04	-0.1	…
gene8	2.2	0.55	0.77	0.79	0.68	-0.29	-0.22	0.5	-0.11	…
gene9	1.03	-0.13	0.46	0.37	0.28	0.23	-0.51	1.89	1.18	…
gene10	0.69	1.32	1.5	1.01	0.69	0.83	1.49	0.13	-0.63	…
gene11	1.01	0.35	1.41	-0.32	0.99	-0.66	-0.36	0.27	0.17	…
gene12	0.39	0.06	0.33	-0.33	0.84	-0.48	0.05	1.2	0.66	…
gene13	1.08	-0.25	1.3	0.69	0.92	0.72	-0.42	0.72	1.61	…
gene14	1.19	-0.36	1.07	0.72	1.38	-0.39	0.31	-0.3	0.57	…
gene15	1.82	-0.18	1.26	0.95	1.34	-0.09	-0.62	0.12	0.55	…
gene16	0.27	1.18	-0.18	0.56	0.4	0.9	0.48	0.68	-0.56	…
gene17	0.93	1.26	1.43	0.29	-0.18	0.88	1.53	-0.02	0.6	…
gene18	1.71	0.89	0.51	0.05	0.93	1.55	0.37	-0.11	-0.2	…



https://jokergoo.github.io/ComplexHeatmap-reference/book/a-single-heatmap.html

Heatmap



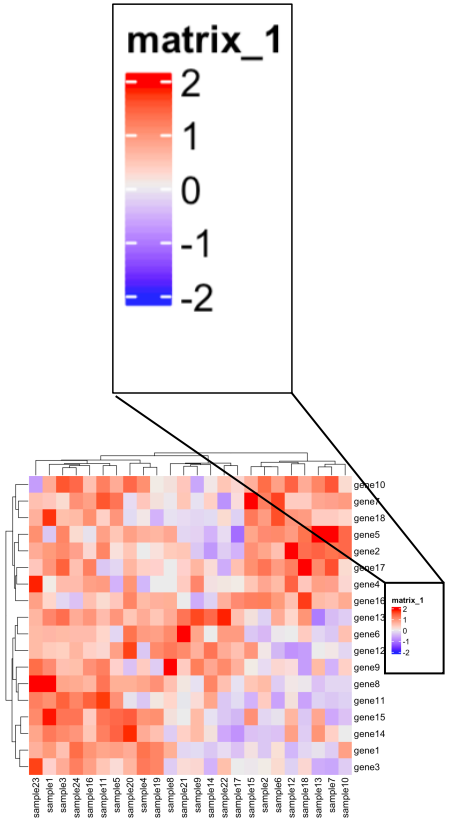
Columns clustered based on similarity of values

Legend/color bar

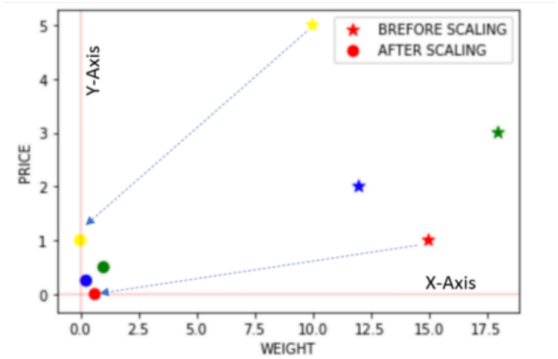
Rows clustered based on similarity of values

Colors mapped to the range of values in the matrix

Scaling data



- Do you think that these data values naturally fall between -2 and 2?



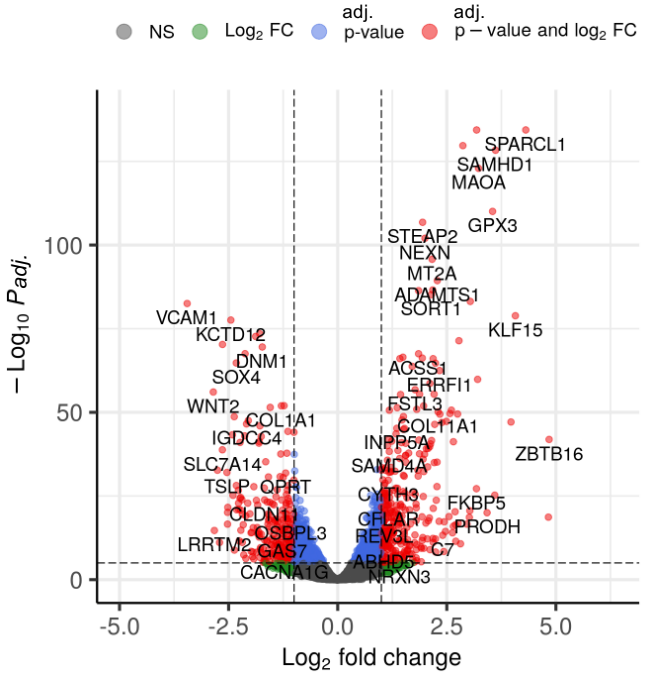
- These data have been SCALED
- Pre-scaling, gene count values can range from ~30,000 to 1. The larger values can skew the data non-representatively
- Scaling brings all the values into similar proportions so they can be reasonably compared

Volcano Plot

Results of a differential expression analysis:

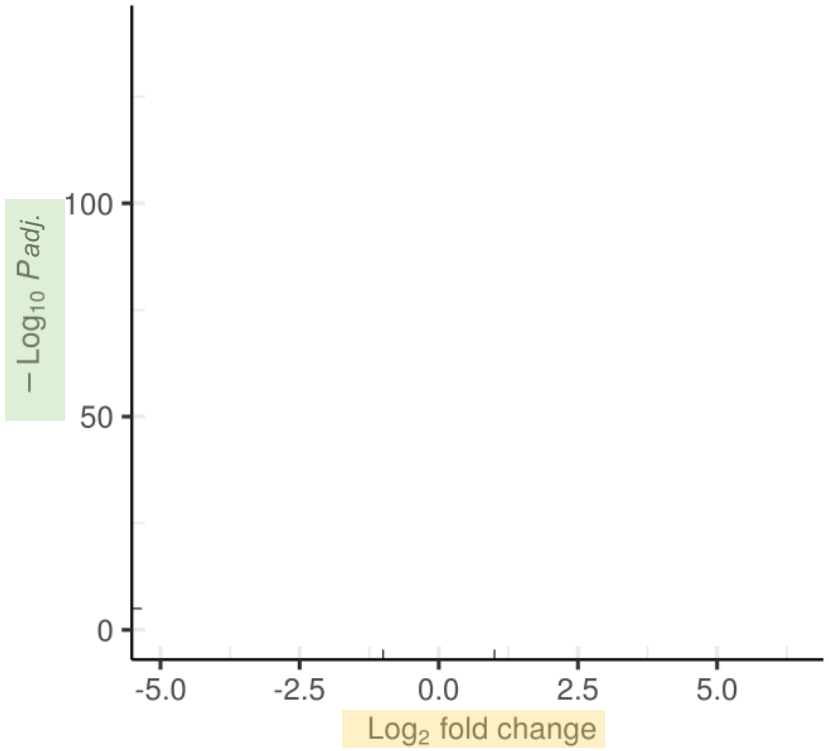
	log2FoldChange	pvalue	padj
TSPAN6	-0.38	0	0
TNMD	NA	NA	NA
DPM1	0.2	0.07	0.19
SCYL3	0.03	0.81	0.91
C1orf112	-0.09	0.75	0.88
FGR	-0.08	0.69	NA
CFH	0.42	0	0
FUCA2	-0.24	0.01	0.03
GCLC	-0.05	0.76	0.88
NFYA	-0.49	0	0
STPG1	-0.12	0.48	0.69
...	...	...	...

genes



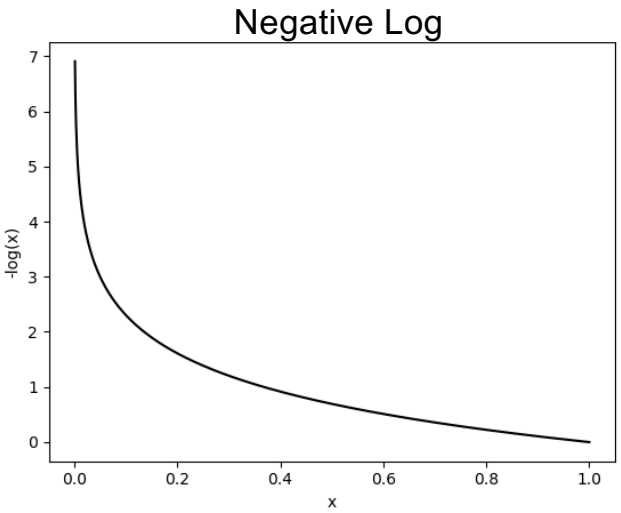
Volcano Plot

	log2FoldChange	pvalue	padj
TSPAN6	-0.38	0	0
TNMD	NA	NA	NA
DPM1	0.2	0.07	0.19
SCYL3	0.03	0.81	0.91
C1orf112	-0.09	0.75	0.88
FGR	-0.08	0.69	NA
CFH	0.42	0	0
FUCA2	-0.24	0.01	0.03
GCLC	-0.05	0.76	0.88
NFYA	-0.49	0	0
STPG1	-0.12	0.48	0.69
…	…	…	…

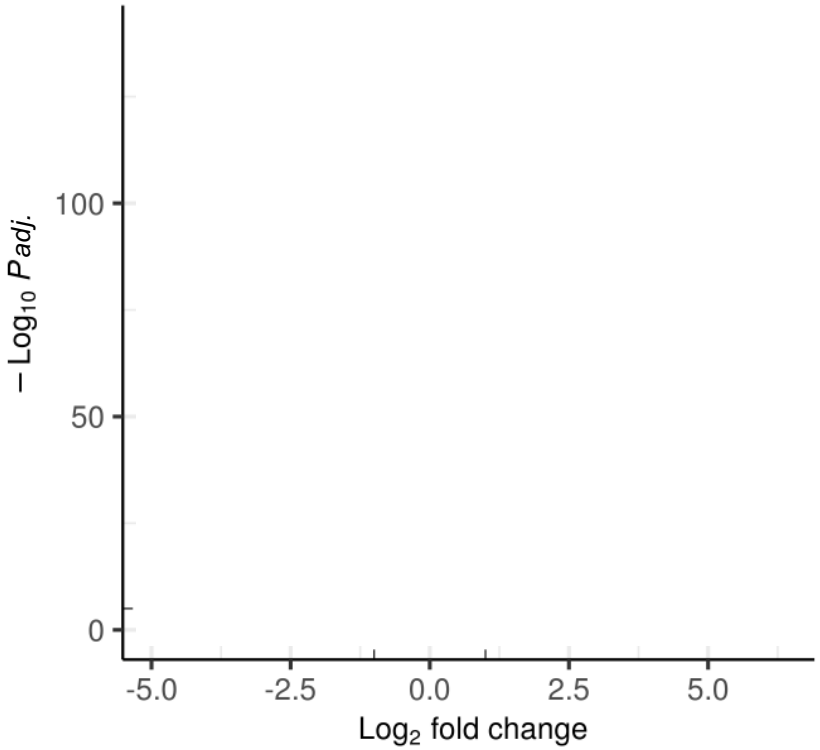


https://livmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-loglikelihood/

Volcano Plot



- As the adjusted pvalue gets smaller, the negative log adjusted pvalue gets bigger!



https://livmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/

Volcano Plot

<u>Significance cutoffs:</u>

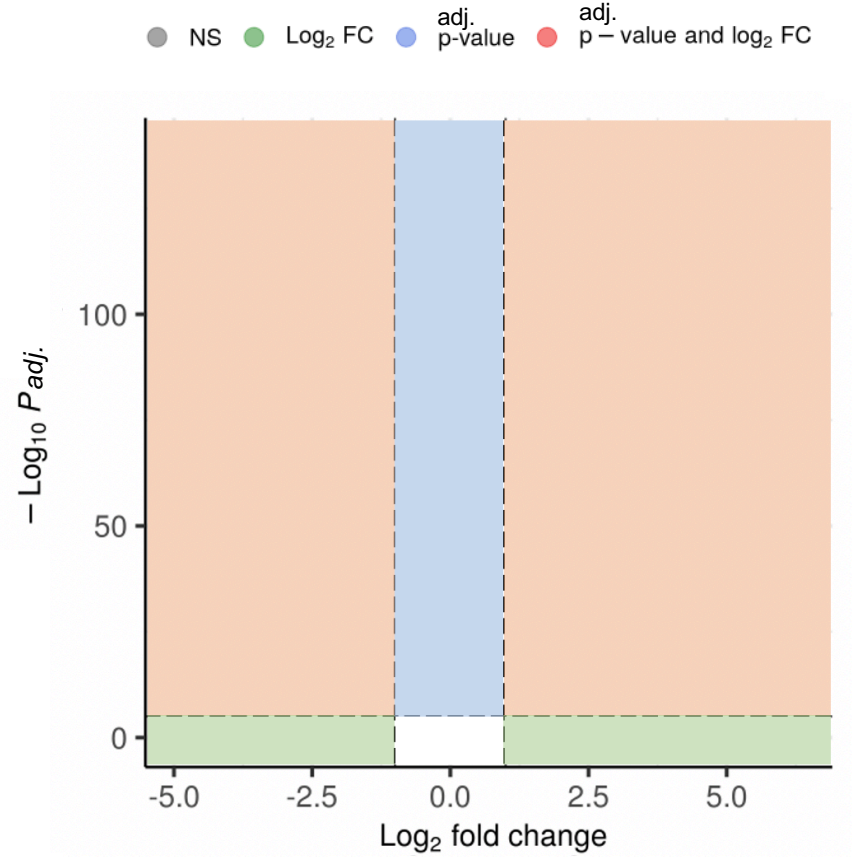
NS Log_2 FC p-value p – value and log_2 FC

On the plot, these cutoffs are shown with dotted lines

A gene is colored green if:
- Its LFC between the compared conditions is >|2|

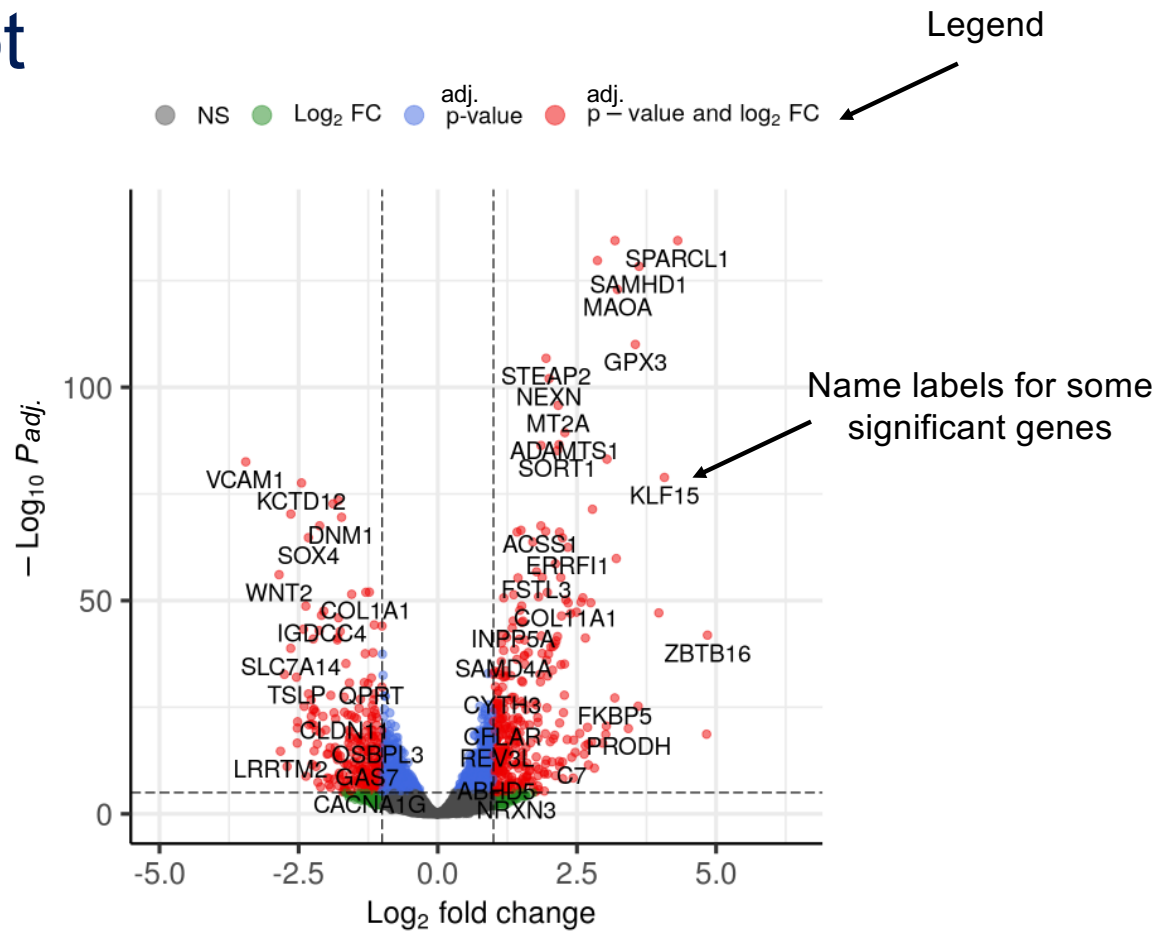
A gene is colored blue if:
- Its adj pvalue is <10e-6

A gene is colored red if:
- Its LFC is >|2| AND its adj p value is <10e-6



https://livmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/

Volcano Plot



Now you are ready to run DESeq2 on your gene counts data to find differentially expressed genes between spaceflight and ground control conditions!