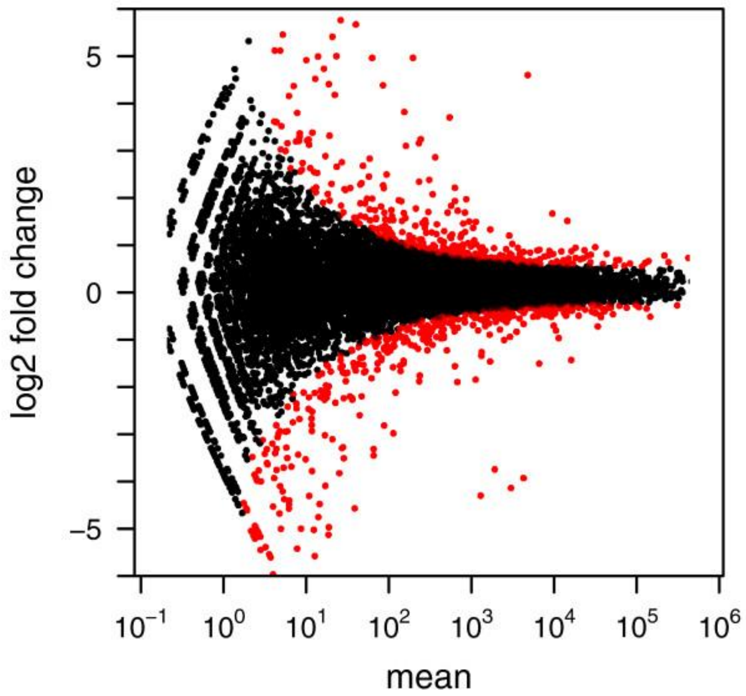
Differential Expression with DESeq2

Erin Osborne Nishimura

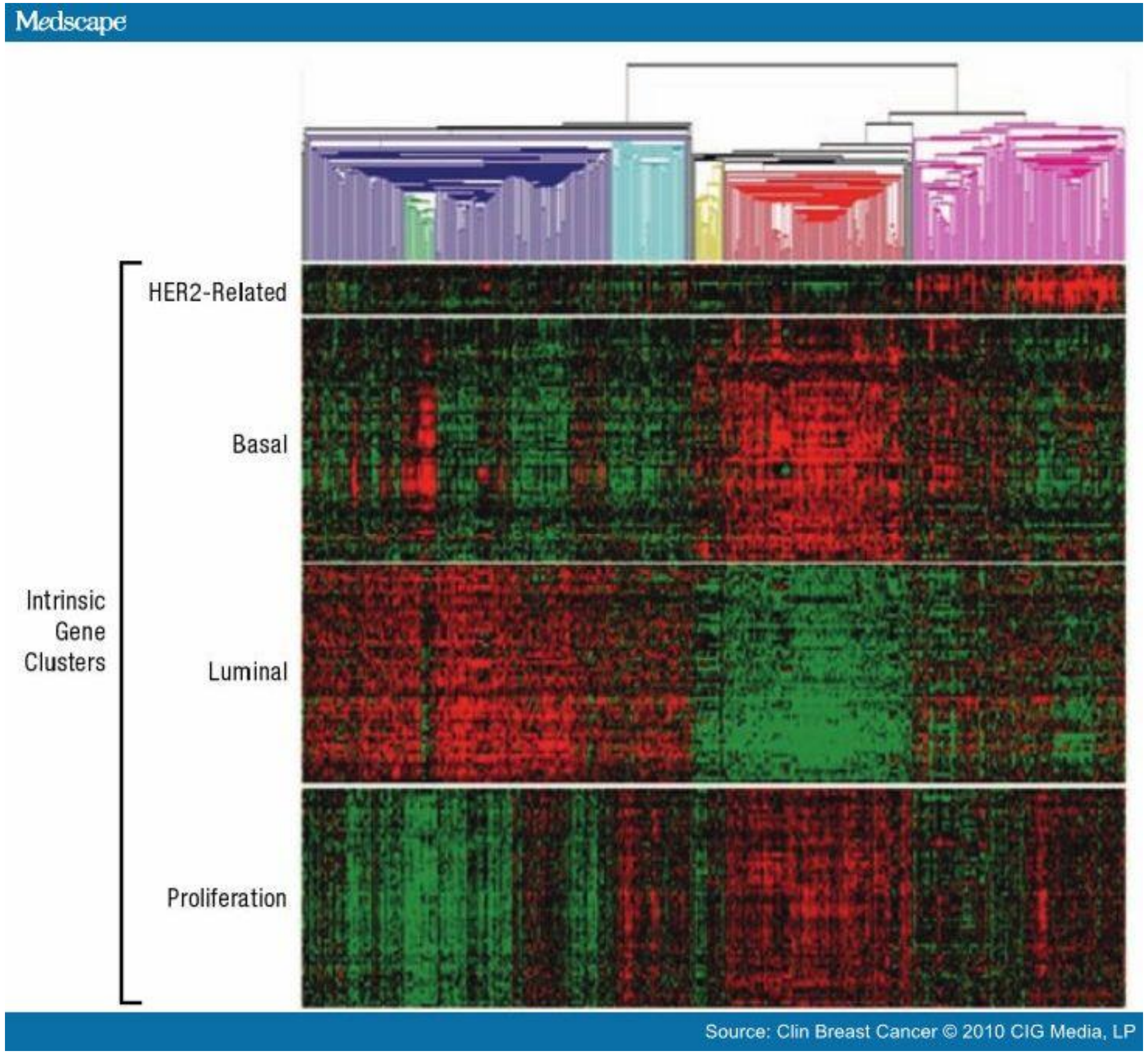
December 3, 2024

Assessing pairwise differential abundance, relatively simple



Anders, Simon

Identifying genes with shared patterns across multiple samples, complex

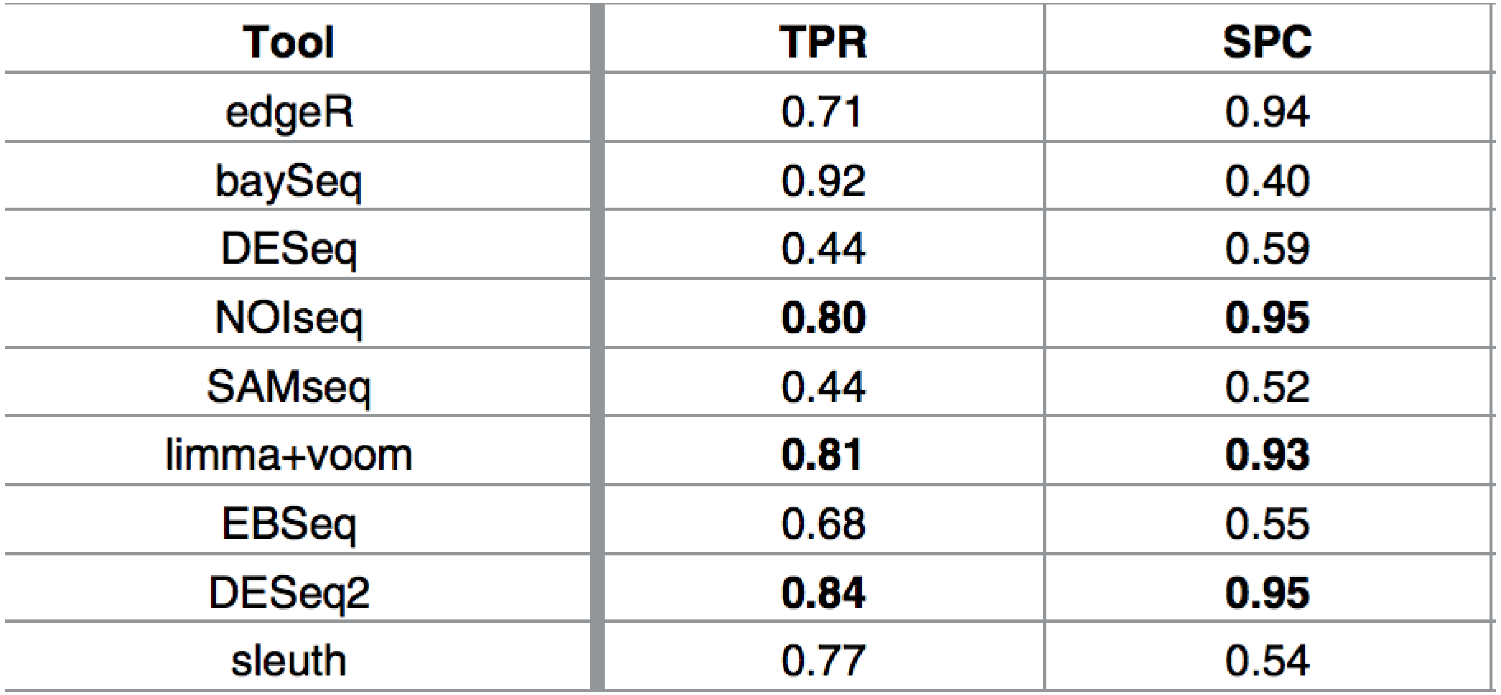


Source: Clin Breast Cancer © 2010 CIG Media, LP

Chuck Perou & Lab

Many algorithms exist to identify differentially expressed genes

True Positive Rate Specificity



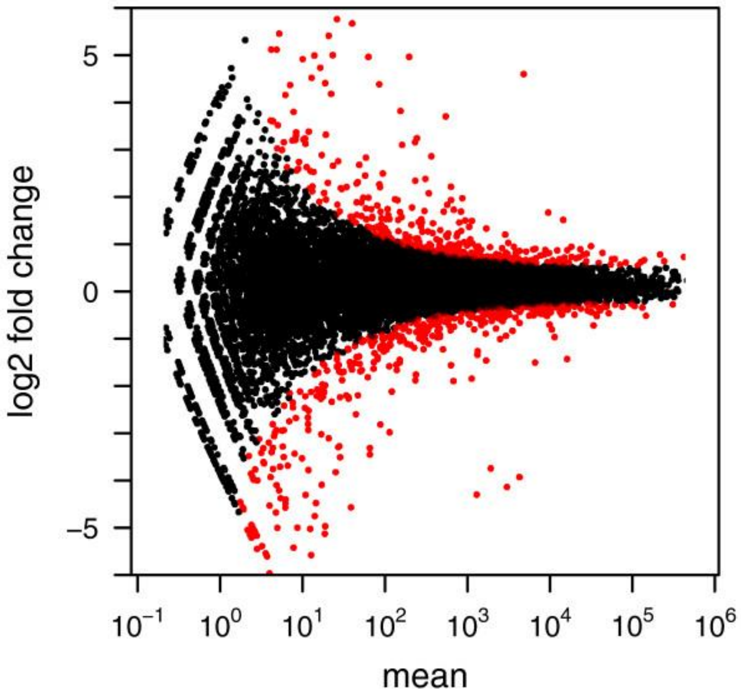
Tool	TPR	SPC
edgeR	0.71	0.94
baySeq	0.92	0.40
DESeq	0.44	0.59
NOIseq	**0.80**	**0.95**
SAMseq	0.44	0.52
limma+voom	**0.81**	**0.93**
EBSeq	0.68	0.55
DESeq2	**0.84**	**0.95**
sleuth	0.77	0.54

Costa-Silva et al., 2017

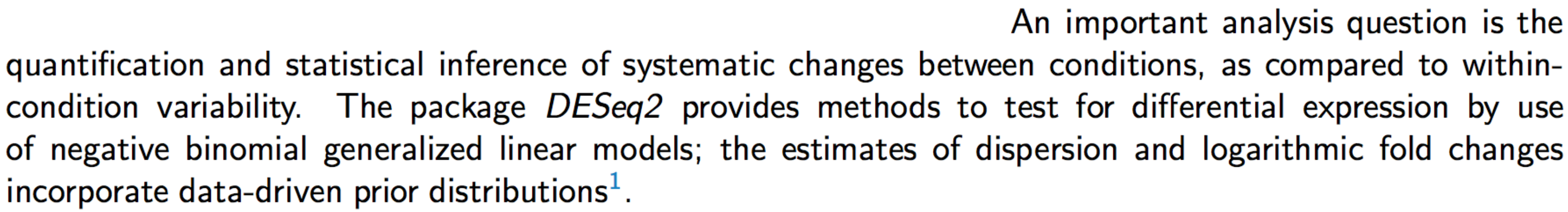
Why is this hard?

Why is this different from other types of data?

- Your question
- The data
  - Discreteness
  - Small numbers of replicates
  - Large dynamic range
  - Outliers
  - Data is over-dispersed
    - Variance does not scale linearly with mean
    - Breaks the assumptions of some inference tests



Why DESeq?



An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package *DESeq2* provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions[1].

- Original paper
  http://www.genomebiology.com/content/11/10/R106
- DESeq2 paper
  - http://www.genomebiology.com/2014/15/12/550
- Bioconductor
  - http://bioconductor.org/packages/release/bioc/html/DESeq2.html
- Analyzing RNA-seq with DESeq2
  - http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

What does DESeq2 do?

1) Requires raw read counts per gene feature.
    A. Ambiguous alignments discarded
2) Normalizes for library size
3) Estimates within-group variability
    A. Estimates variance (negative binomial distribution)
    B. Performs variance shrinkage
4) Estimates fold-change between conditions
5) Tests for significant differences in signals between groups.
    A. Fits a generalized linear model of the negative binomial family
    B. Performs a Wald test to calculate a p-value
    C. Performs multiple testing correction

1) DESeq2 requires raw read counts per gene feature

- Ambiguous reads are discarded (featureCounts)
- Why raw read counts versus FPKM (Fragments Per Kilobase Gene Length Per Million Mapped Reads)?
  - We will try to avoid comparisons between GeneA and GeneB. Only compare GeneA in Condition1 v. GeneA in Condition2.
  - Avoid comparisons between experiments
  - Raw read counts are discrete variables. By preserving the discrete nature of the raw data, we can take advantage of statistical approaches that work on discrete variables

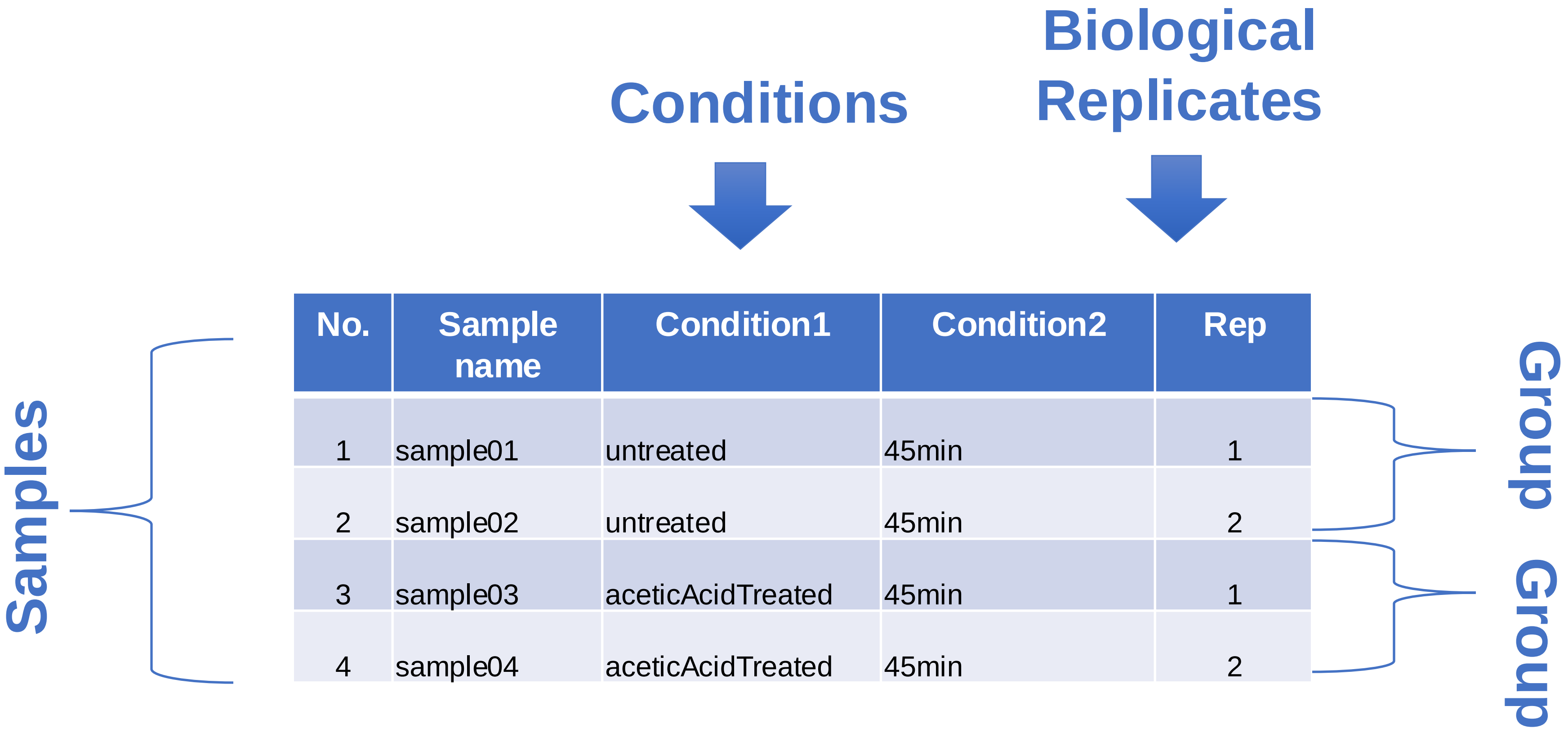
2) DESeq2 normalizes for library size

- Control for read depth

- Scaling factor is used

- Weights median expressed genes more heavily than outliers.

3) Estimates within-group variability
4) Estimates fold-change between conditions



Conditions

Biological Replicates

Samples

Group

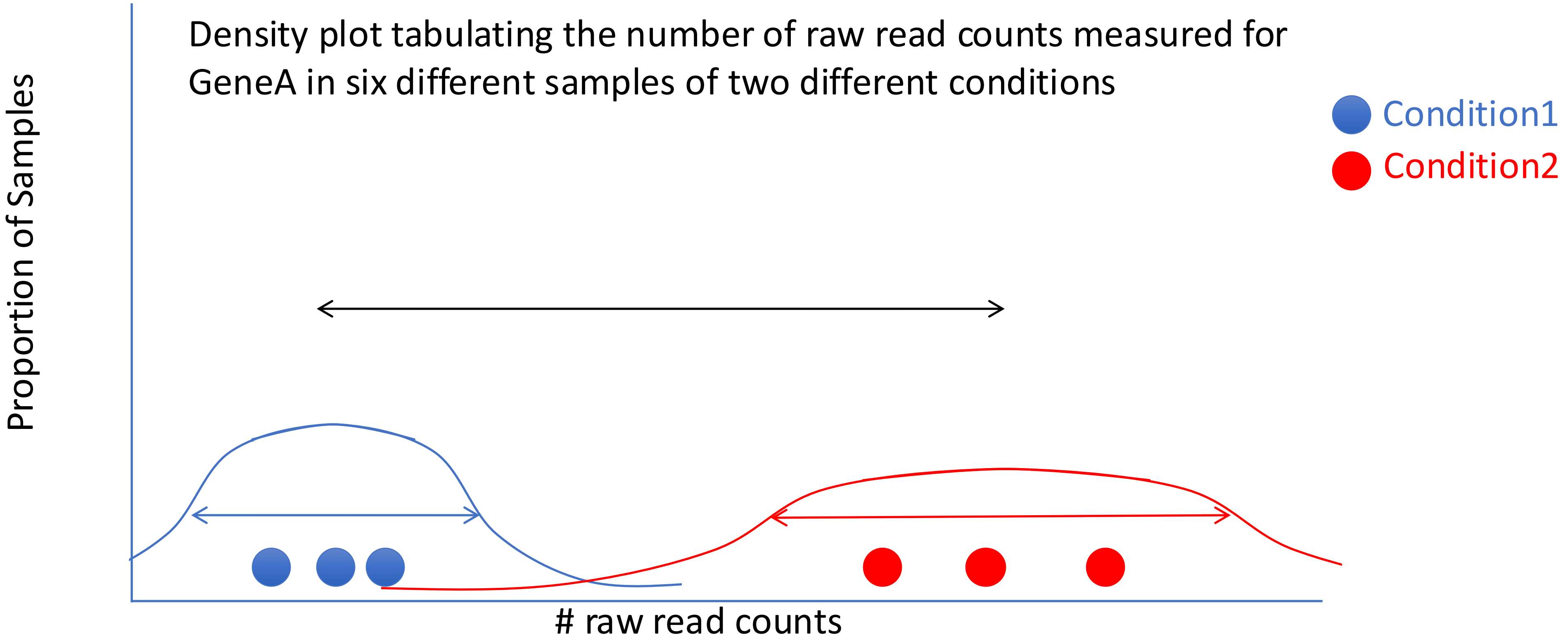
Group

No.	Sample name	Condition1	Condition2	Rep
1	sample01	untreated	45min	1
2	sample02	untreated	45min	2
3	sample03	aceticAcidTreated	45min	1
4	sample04	aceticAcidTreated	45min	2

What does DESeq2 do?

1) Requires raw read counts per gene feature.
   A. Ambiguous alignments discarded
2) Normalizes for library size
3) **Estimates within-group variability**
   A. Estimates variance (negative binomial distribution)
   B. Performs variance shrinkage.
4) **Estimates fold-change between conditions**
5) Tests for significant differences in signals between groups.
   A. Fits a generalized linear model of the negative binomial family
   B. Performs a Wald test to calculate a p-value
   C. Performs multiple testing correction

3) Estimates within-group variability
4) Estimates fold-change between conditions

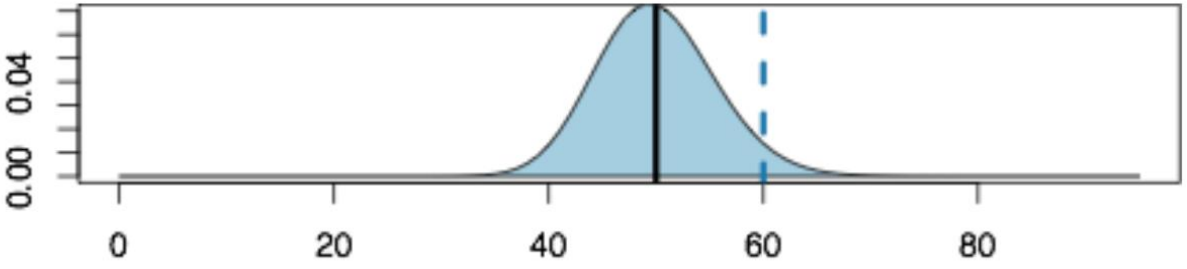

Density plot tabulating the number of raw read counts measured for GeneA in six different samples of two different conditions
Proportion of Samples
# raw read counts
● Condition1
● Condition2

- Analyzing RNA-seq with DESeq2
  - http://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

SELF STUDY

3) Estimates within-group variability

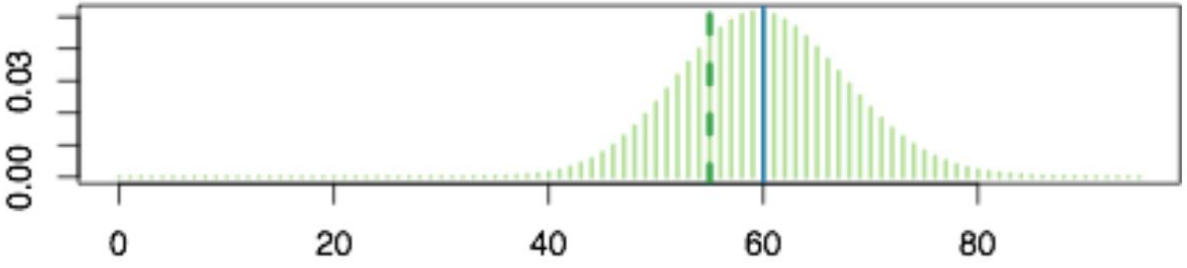
The negative binomial distribution



Normal distribution:

Mean = μ

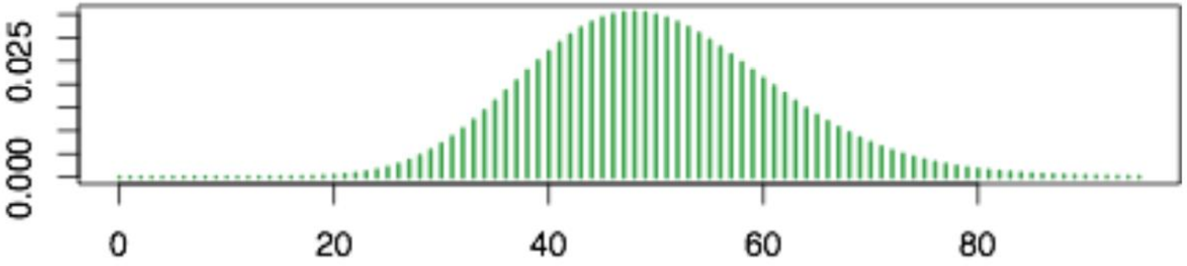
Variance = v



Poisson distribution:

Mean = μ

Variance = μ



Negative Binomial Distribution:

Mean = μ

Variance = $\mu + \alpha\mu^2$

A major insight of the DESeq2 approach was to use a different distribution to model the spread of the data. The negative binomial distribution assumes there will be greater variance in the data than what is predicted by the Poisson distribution. This approach was widely adapted by other tools after DESeq2 made this insight.

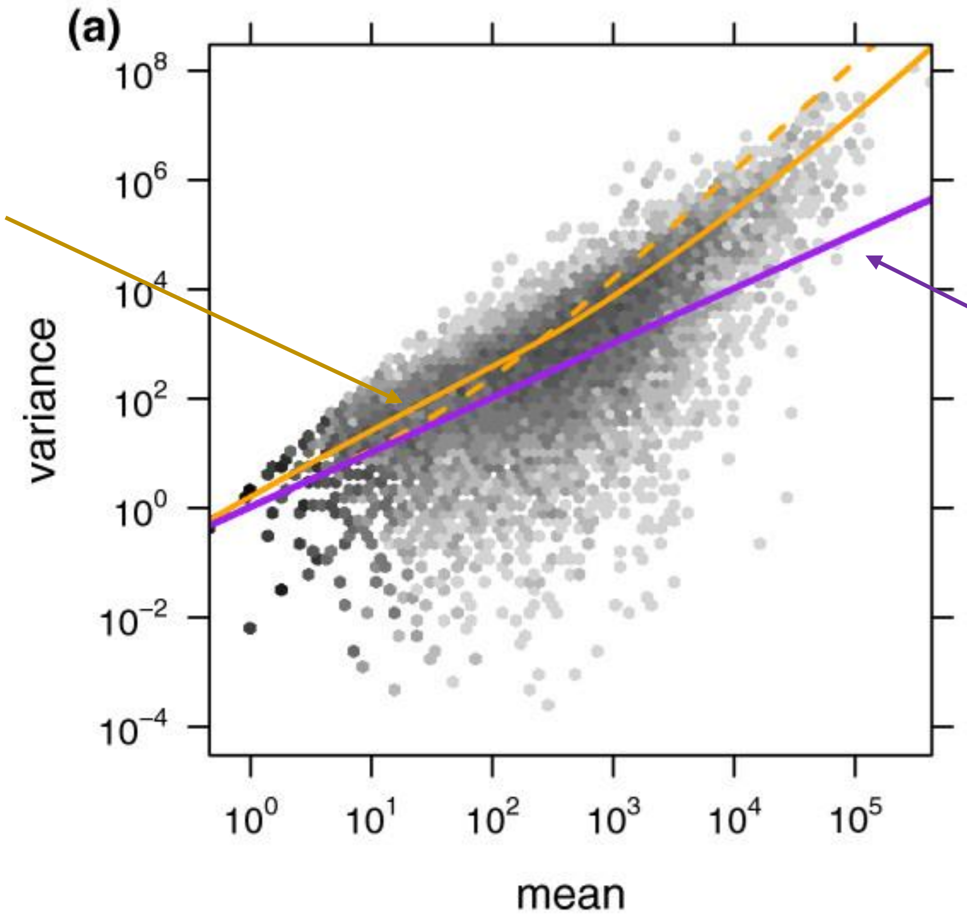
Simon Anders, https://www.bioconductor.org/help/course-materials/2014/CSAMA2014/2_Tuesday/lectures/DESeq2-Anders.pdf

SELF STUDY

3) Estimates within-group variability
The over-dispersion problem

The DESeq2 developers had the insight that the Poisson distribution was not the proper way to model the data because when they plotted real RNA-seq data, they always found that the variance was higher than the Poisson distribution predicted, especially for highly expressed genes.

… but here's some real RNA-seq data, and the variance is way higher than the mean.



The Poisson distribution estimates that the mean and variance should be equal

*** Dashed line is the variance estimation from edgeR, a previous program. That looks closer.**

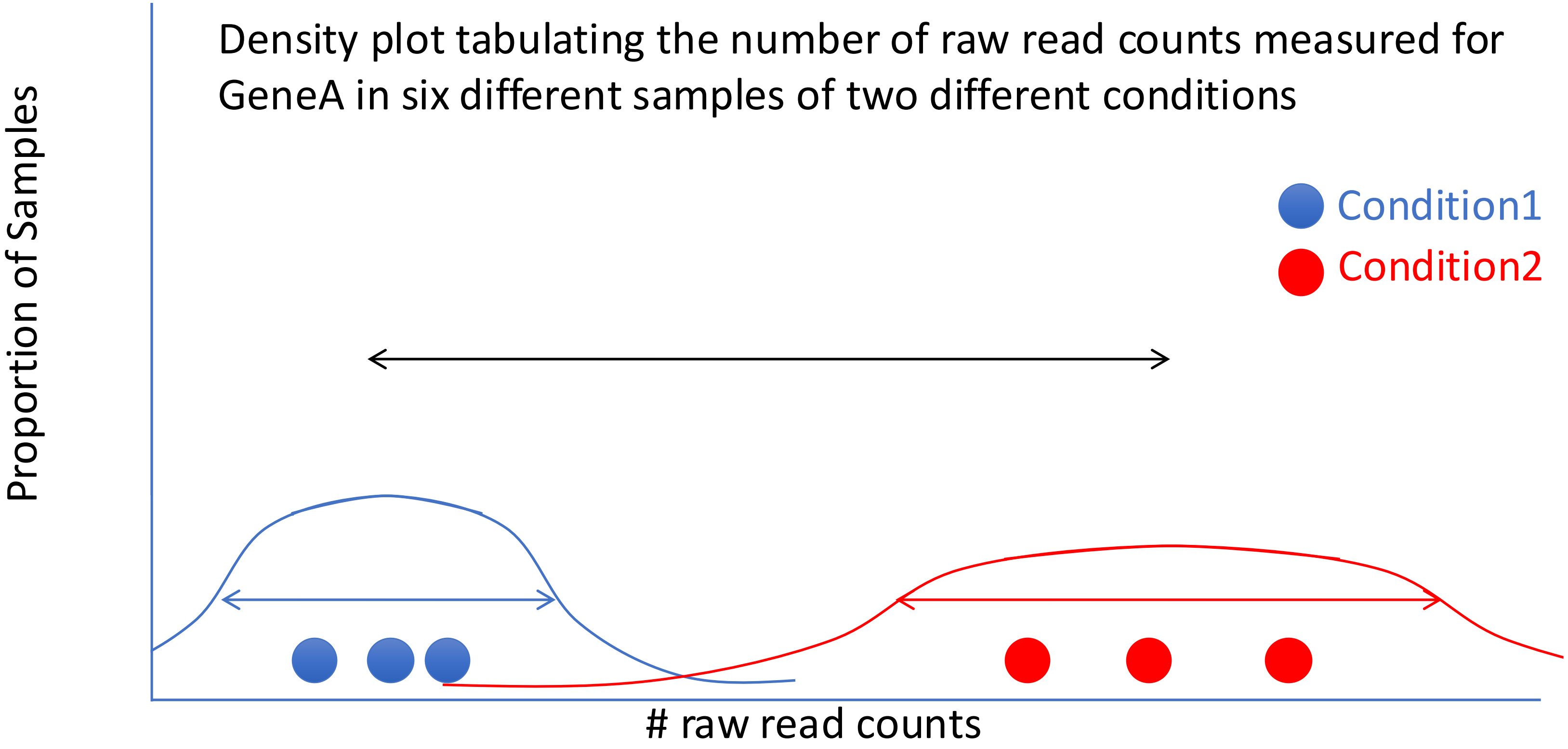
Anders et al., 2010

SELF STUDY

3) Estimates within-group variability

Estimates variance (negative binomial distribution)

OPTIONAL - Performs variance shrinkage.



Density plot tabulating the number of raw read counts measured for GeneA in six different samples of two different conditions

The next thing that the developers decided to do was to “borrow” insight across genes of similar expression levels. This seems totally illegal, but here is what they thought. The issue is that we just don’t have very many replicates. Maybe 3, sometimes 4, often 2. It’s not very many samples. So this is a problem.

Anders and co. thought, well, one gene expressed at 1000 copies or so tends to have a similar variance as a second gene expressed at 1000 copies or so. So, they used this to “borrow” read counts across similarly expressed genes as a “stand in” for more replicates. By “pretending that all the data from gene1 and gene2 came from gene1 (or gene2), they can double the replicates. This narrows the variance.

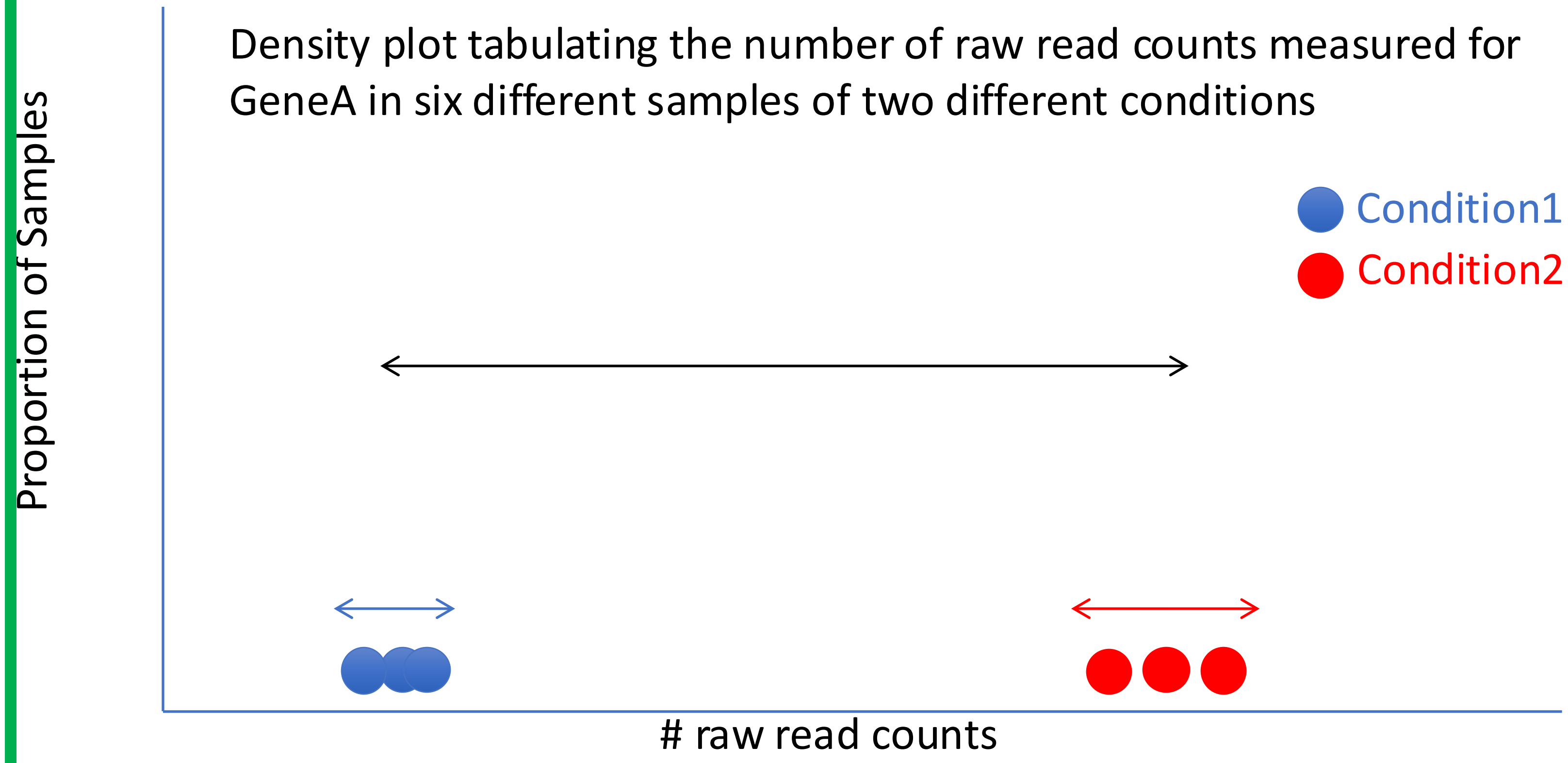
SELF STUDY

3) Estimates within-group variability

Estimates variance (negative binomial distribution)

OPTIONAL - Performs variance shrinkage.

And here we go… shrink!



Density plot tabulating the number of raw read counts measured for GeneA in six different samples of two different conditions
● Condition1
● Condition2
Proportion of Samples
# raw read counts

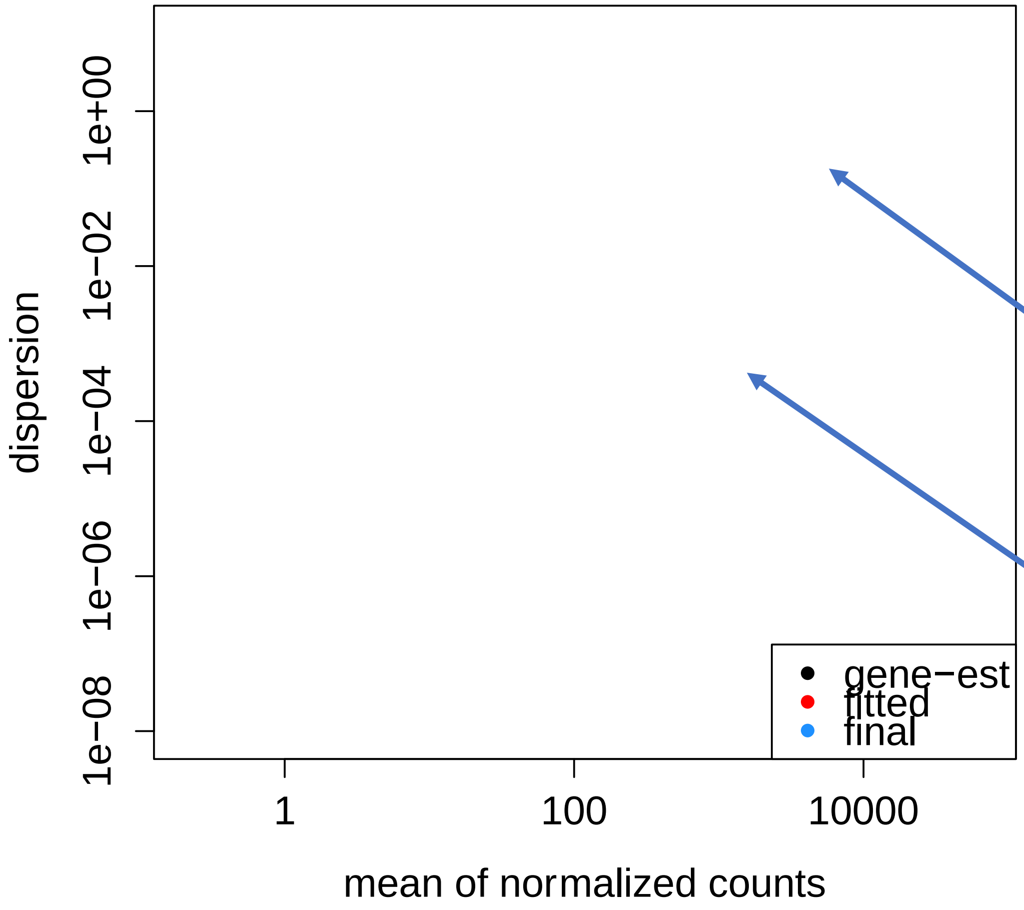
Now the variance is small and we're zooming in on, hopefully, the more 'accurate' measurement.

It seems totally wrong to do this. And I wouldn't try it at home, folks. But with some careful modeling, this is what they have done and this has become accepted over time.

Now, you can see that with the shrunken variance, the within-condition ranges are smaller, and this makes the between-conditions comparisons clearer.

SELF STUDY

3) DESeq2 estimates within group variance and performs variance shrinkage.

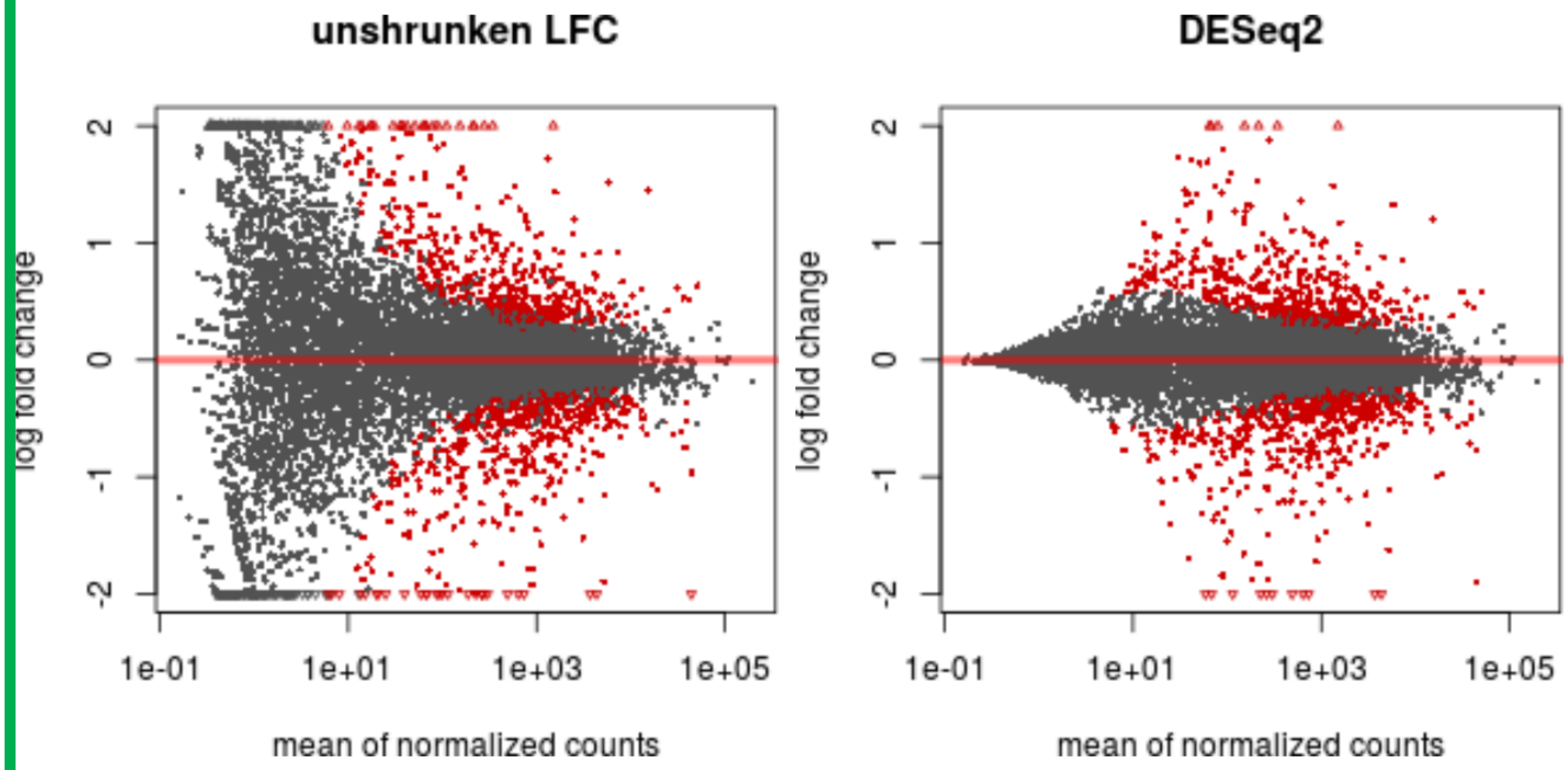


The key assumption:
Genes of the same intensity should have similar variance.

This is a plot you can generate easily in DESeq2. It is a graphical representation of the "dispersion" calculated from the data direction (black), the modeled dispersion they expect for genes of a given expression intensity (pink), and then when they apply the modeled dispersion to the measured dispersion, they get a new value for each gene (blue). You'll notice some values worsen the dispersion estimate. That's ok, because they're already low. Other values don't change much. Some of these are outliers or very wonky to begin with. Because those are difficult, they are often excluded from the variance shrinkage.

SELF STUDY

4) Estimates fold-change between conditions



Another cool trick DESeq2 does is to shrink the log fold change to remove that weird “flaring out” effect in the lowly expressed genes. This makes it more comparable to look at log-fold-change values across different expression levels.

This is typically called “r-stabilized” log fold change.

This makes the data “homoscedastic”, a feature valued among statisticians.

https://en.wikipedia.org/wiki/Homoscedasticity

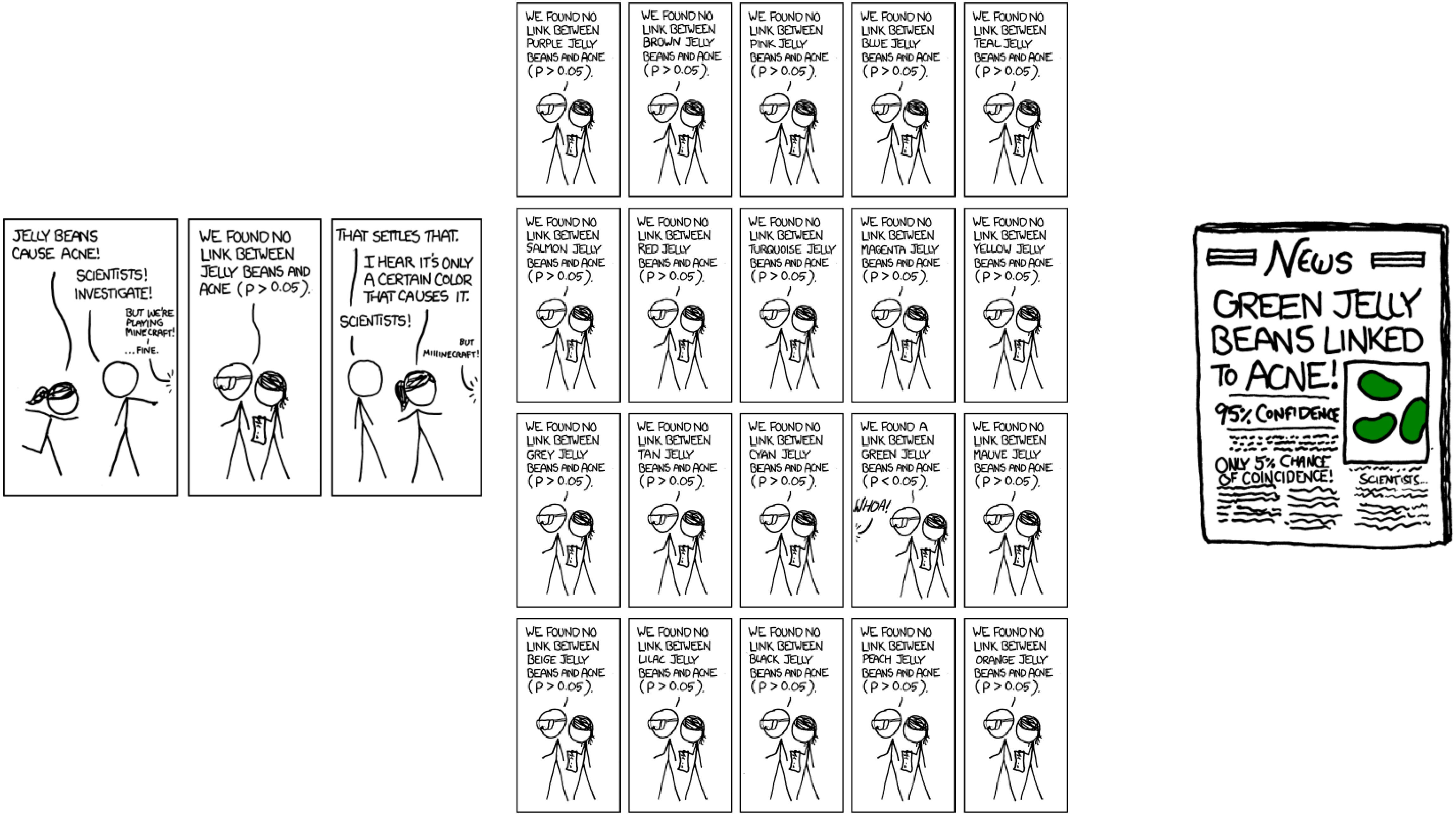
Love et al., 2014

What does DESeq2 do?

1) Requires raw read counts per gene feature
   A. Ambiguous alignments discarded
2) Normalizes for library size
3) Estimates within-group variability
   A. Estimates variance (negative binomial distribution)
   B. Performs variance shrinkage.
4) Estimates fold-change between conditions
5) **Tests for significant differences in signals between groups**
   A. Fits a generalized linear model of the negative binomial family
   B. Performs a Wald test to calculate a p-value
   C. Performs multiple testing correction

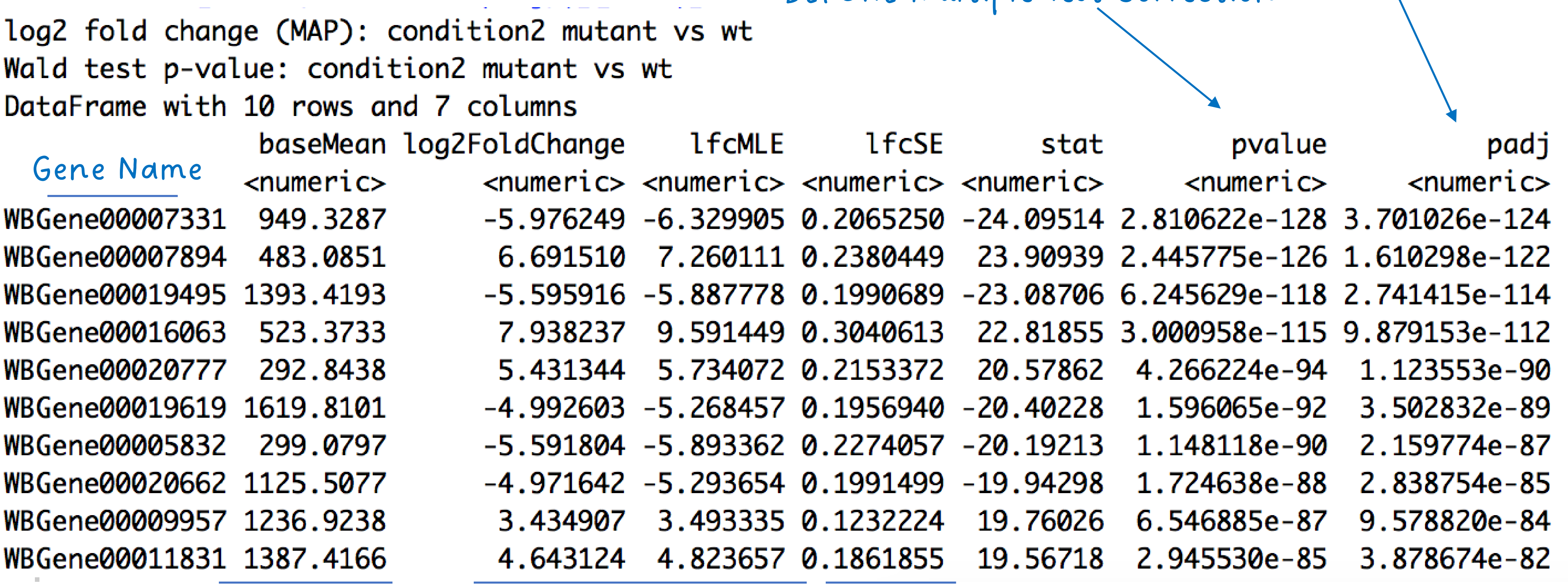
5) Tests for significant differences in signals between groups

- Wald test → p-value
- **Multiple testing correction** using **Benjamini–Hochberg Procedure**
  - P-value -> p-adj
  - Imagine you test 10,000 genes for significant differences between two identical samples at p-value < 0.05 (Correct 95 % of the time)
  - Say, 1,500 genes have a p-value < 0.05
  - How many could be false positives?
    - *A maximum of 500 of your 1,500 genes could be false positives.*
    - *Yikes – that's a lot*
    - *Caveat - is not likely to truly be that high (most p values will be well under 0.05)*
    - *But still, we should account for this*


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

https://xkcd.com/882/

DESeq2 output



```
log2 fold change (MAP): condition2 mutant vs wt
Wald test p-value: condition2 mutant vs wt
DataFrame with 10 rows and 7 columns
                 baseMean log2FoldChange    lfcMLE     lfcSE      stat        pvalue          padj
                <numeric>      <numeric> <numeric> <numeric> <numeric>     <numeric>     <numeric>
WBGene00007331   949.3287      -5.976249 -6.329905 0.2065250 -24.09514 2.810622e-128 3.701026e-124
WBGene00007894   483.0851       6.691510  7.260111 0.2380449  23.90939 2.445775e-126 1.610298e-122
WBGene00019495  1393.4193      -5.595916 -5.887778 0.1990689 -23.08706 6.245629e-118 2.741415e-114
WBGene00016063   523.3733       7.938237  9.591449 0.3040613  22.81855 3.000958e-115 9.879153e-112
WBGene00020777   292.8438       5.431344  5.734072 0.2153372  20.57862  4.266224e-94  1.123553e-90
WBGene00019619  1619.8101      -4.992603 -5.268457 0.1956940 -20.40228  1.596065e-92  3.502832e-89
WBGene00005832   299.0797      -5.591804 -5.893362 0.2274057 -20.19213  1.148118e-90  2.159774e-87
WBGene00020662  1125.5077      -4.971642 -5.293654 0.1991499 -19.94298  1.724638e-88  2.838754e-85
WBGene00009957  1236.9238       3.434907  3.493335 0.1232224  19.76026  6.546885e-87  9.578820e-84
WBGene00011831  1387.4166       4.643124  4.823657 0.1861855  19.56718  2.945530e-85  3.878674e-82
```

Wald t-test p-value
AFTER Multiple Test Correction

Wald t-test p-value
BEFORE Multiple Test Correction

Gene Name

Mean # of
normalized
Reads across
ALL samples

One - Two Types of Log
Fold Change
Estimates

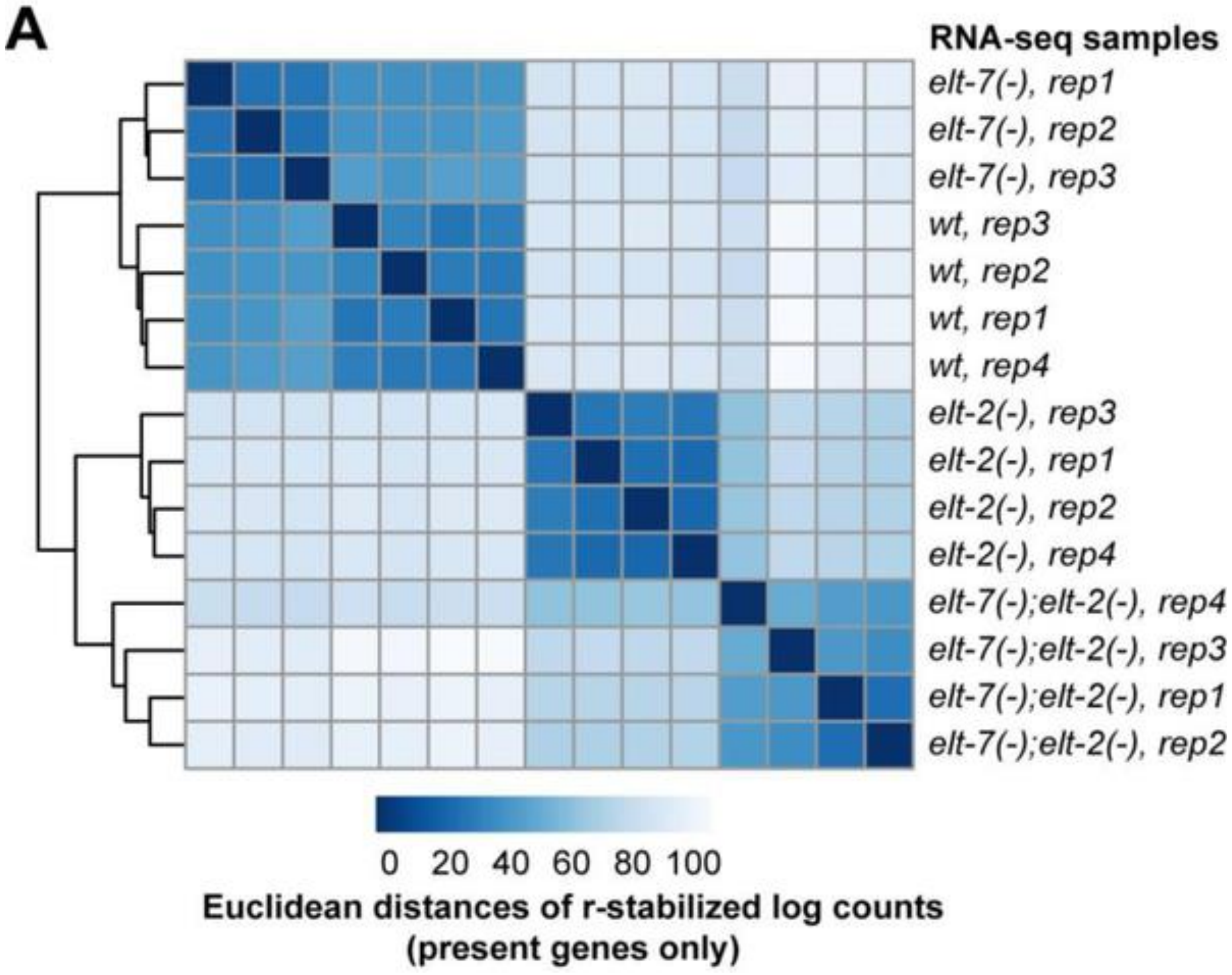
Standard
Error

Summary

- To assess whether fold changes between conditions are significant, the within-condition variance must first be estimated.
- RNA-seq data is over-dispersed and is better approximated by a negative-binomial distribution than a Poisson distribution.
- To improve variance estimates from datasets with small sample sizes, information can be shared across genes.
- Once variance is shrunken (r-stabilized), the log-transformed values are better for visualization, interpreting effect sizes, clustering, and ranking.

Meet the plots

Distance matrix: a bird's eye view

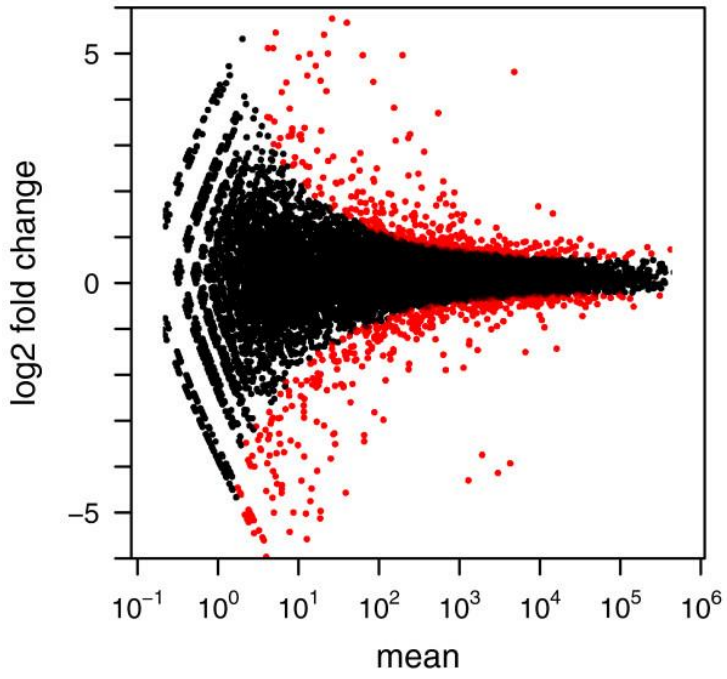


Dineen and Osborne Nishimura, et al. (2018) Dev Biol. pii: S0012-1606(17)30690-5.

- Each square is a comparison between two samples in the dataset
  - Order is the same on the Y and X axes
- The color represents how similar the two samples are to one another based on the values for ALL genes in the dataset
- Here, Euclidian distance is calculated… smaller value, more similar
- Correlation coefficients can also be used and those are called "Correlation Matrices". Larger value – more similar
- (matrix – singular; matrices - pl.)
- https://en.wikipedia.org/wiki/Distance_matrix

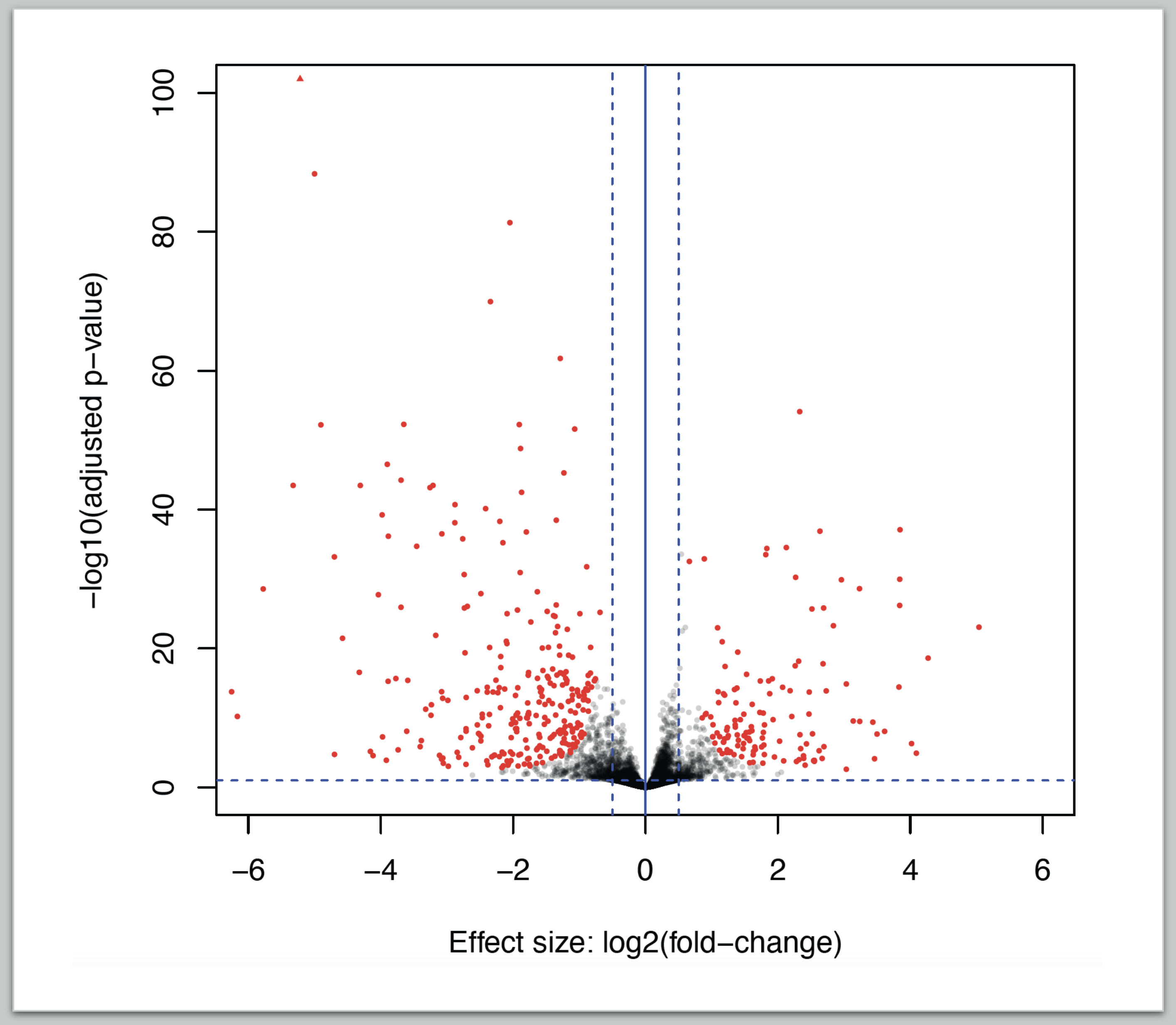
MA-plot

- Each point is data for a given gene
- X-axis: mean (M) expression intensity
  - Log10
- Y-axis: fold change between one condition and another (A)
  - Log2
- Red color – under a certain threshold p-value
- Why we use this plot:
  - To compare differences between conditions against the expression intensity for every gene.
  - Red genes are over-abundant or under-abundant
- Works on 2 conditions
- https://en.wikipedia.org/wiki/MA_plot



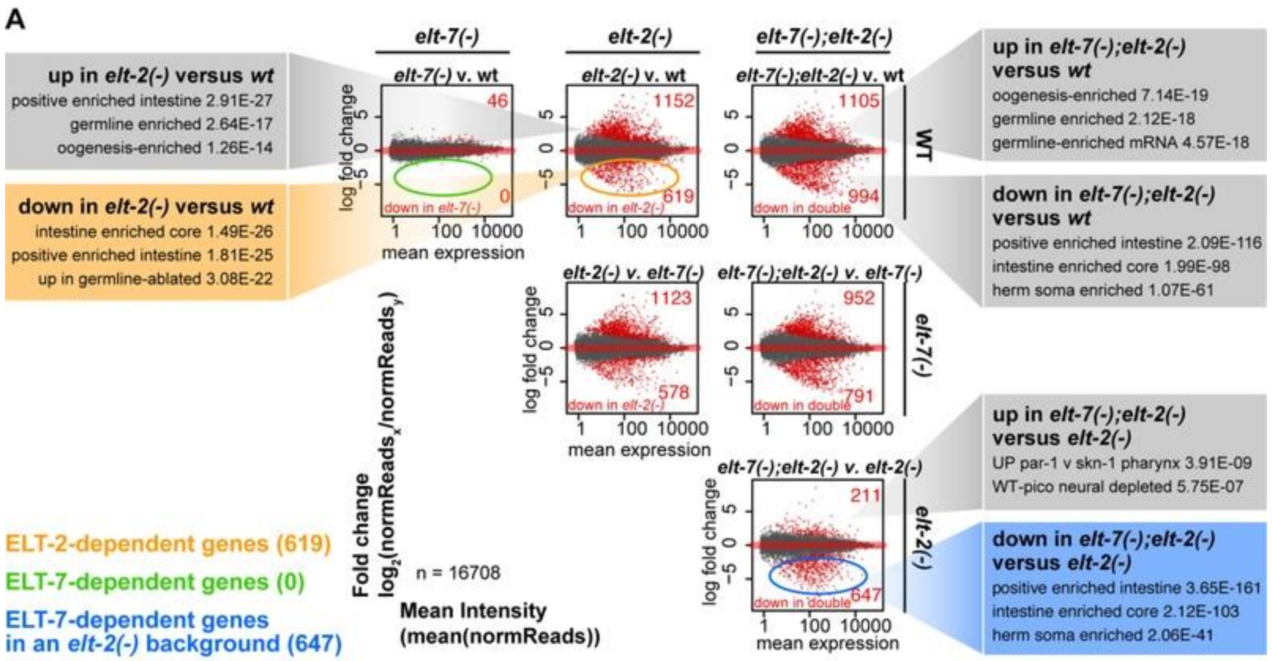
Volcano Plot

- Each point is data for a given gene
- X-axis: fold change between one condition and another (A)
  - Log2
- Y-axis: p-value
  - Log10
- Why we use this plot:
  - To find the genes that are most dramatically differentially expressed
- Works on 2 conditions
- https://en.wikipedia.org/wiki/Volcano_plot_(statistics)



What if you have more than two conditions?

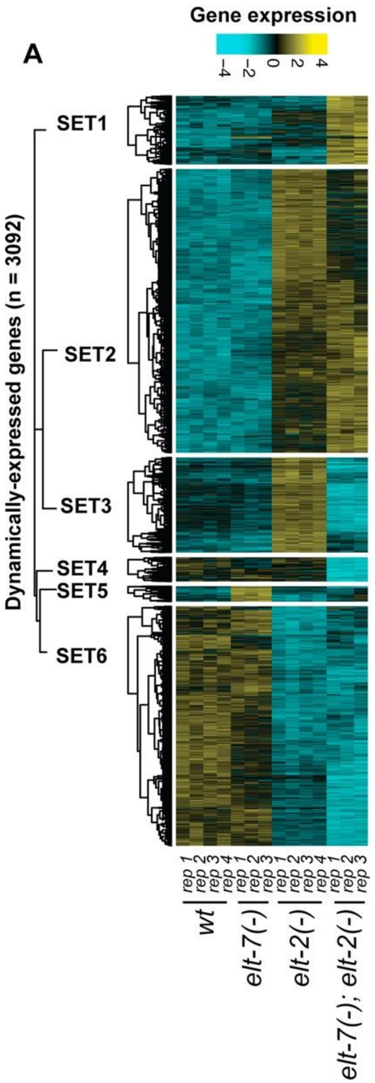
- Option 1 – array of MA plots



Dineen and Osborne Nishimura, et al. (2018) Dev Biol. pii: S0012-1606(17)30690-5.

What if you have more than two conditions?

- Heatmap
- Identify all differentially expressed genes between all pair-wise combinations
- Create a heatmap of the DE genes
- Cluster the heatmap based on expression
- Identify sets of genes with similar expression dynamics across the dataset
- A complete guide to heatmaps:
  - https://chartio.com/learn/charts/heatmap-complete-guide/



Dineen and Osborne Nishimura, et al. (2018) Dev Biol. pii: S0012-1606(17)30690-5.

Plots are fun!

- R Graph Gallery – a resource with code
    - https://r-graph-gallery.com/

Best Practices & Next Steps

- It has become very difficult to figure out what some of these DE algorithms are doing.
    - Try multiple approaches – do they concur? How do they differ?
    - Look at genome browser shots – do things make sense?
    - Consult with others in your field
    - Read methods sections of papers in your field – what is the standard? Why?
    - Follow up with experimental verification

Best Practices & Next Steps

- Gene Ontology (GO) – Figure out what “types” of genes are in your lists
    - Cell Category – where are these gene products located in the cell?
    - Molecular function – what do these gene products do?
    - Biological process – what overarching processes are these gene products involved in?
    - Relies on already-collected data that has been curated
    - Try DAVID , GOrilla, or clusterProfiler (works in R)
- KEGG Pathways – look for enriched metabolic pathways
- Use smFISH, Western blots, GFP imaging to verify differential expression of key genes/products
- How many genes should be in my gene list?
    - Consider lists of different sizes for different downstream analyses
        - GO Ontology typically works best with lists of 200 – 1000
        - smFISH, Western blot, GFP verification typically works best with the top 1 – 10 candidates
    - Publish your lists in Supplemental so others (and you) can more easily use them