limma – Differential Expression Analysis and Beyond

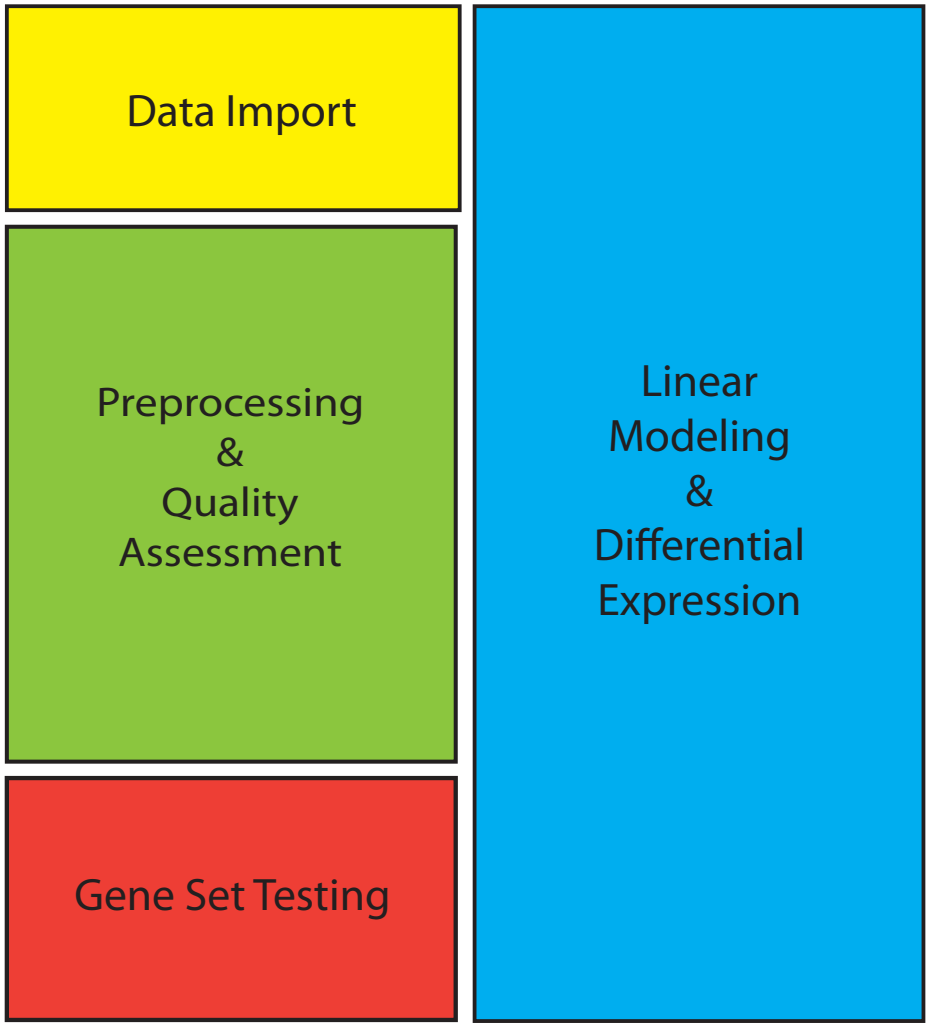
COMBINE RNA-seq Workshop, 23rd September 2016

limma package: Linear Models for Microarrays & RNA-seq

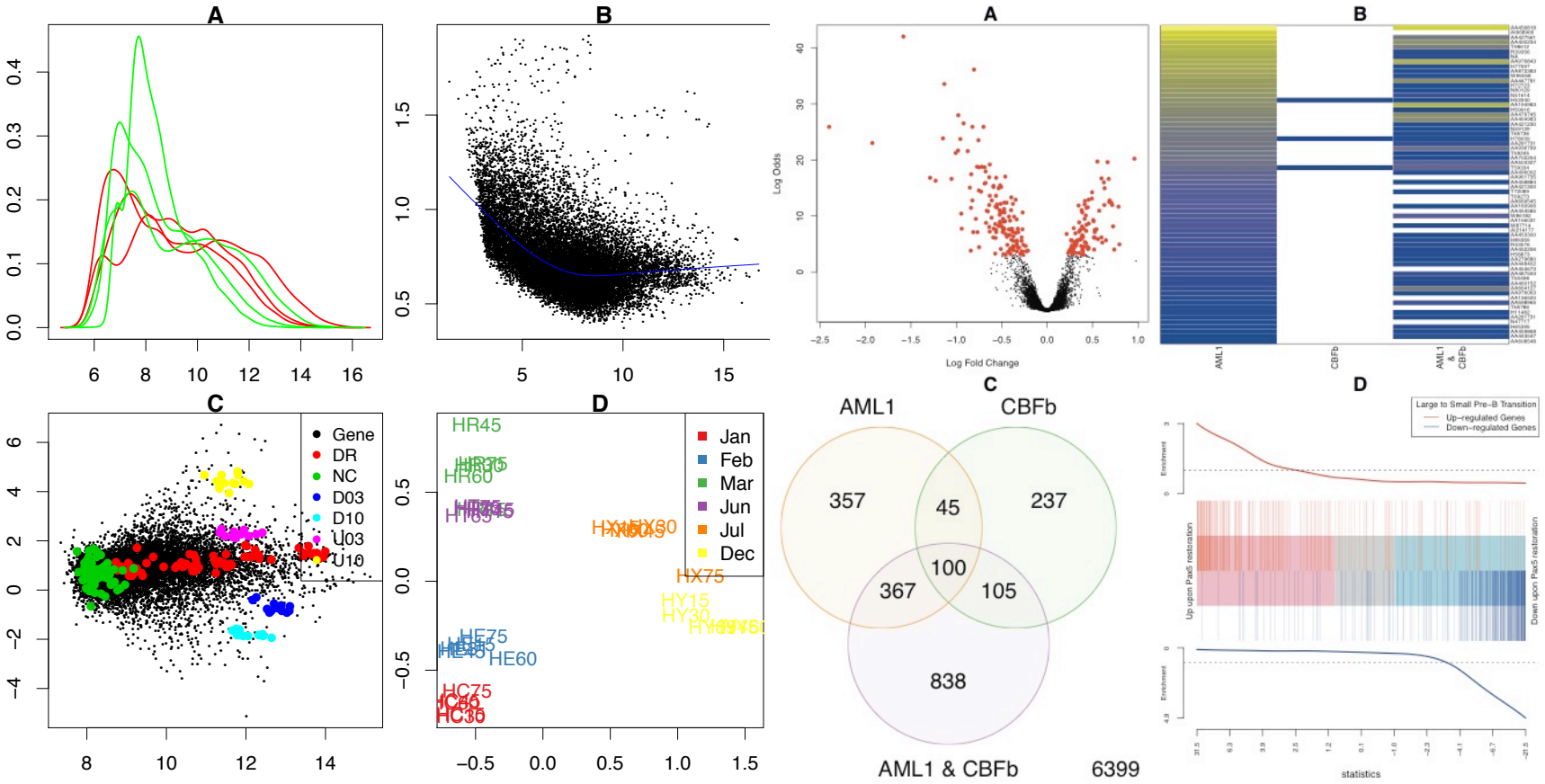
Professor Gordon Smyth

limma is celebrating its 13th birthday this year!

AUSTRALIAN MADE



Many plotting options available…



Linear models for differential expression

Matrix of expression values (from RNA-seq / microarray)

Gene ID	LSK_1	LSK_2	CMP_1	CMP_2
11303	478	619	4830	7165
11305	27	20	48	55
11306	132	200	560	408
11307	42	60	131	99
...	... tens of thousands more ...			

Gene-wise linear models

$$E(y_g) = X\beta_g$$

$$\text{var}(y_{gj}) = \sigma_g^2 / w_{gj}$$

Advanced statistical algorithms in *limma* that allow...

- ***Information Borrowing***
- ***Quantitative Weighting***
- ***Variance Modelling***
- ***Data Pre-processing***

***limma* delivers powerful inference for differential expression analysis**

$$\hat{\beta}_g, s_g^2 *$$

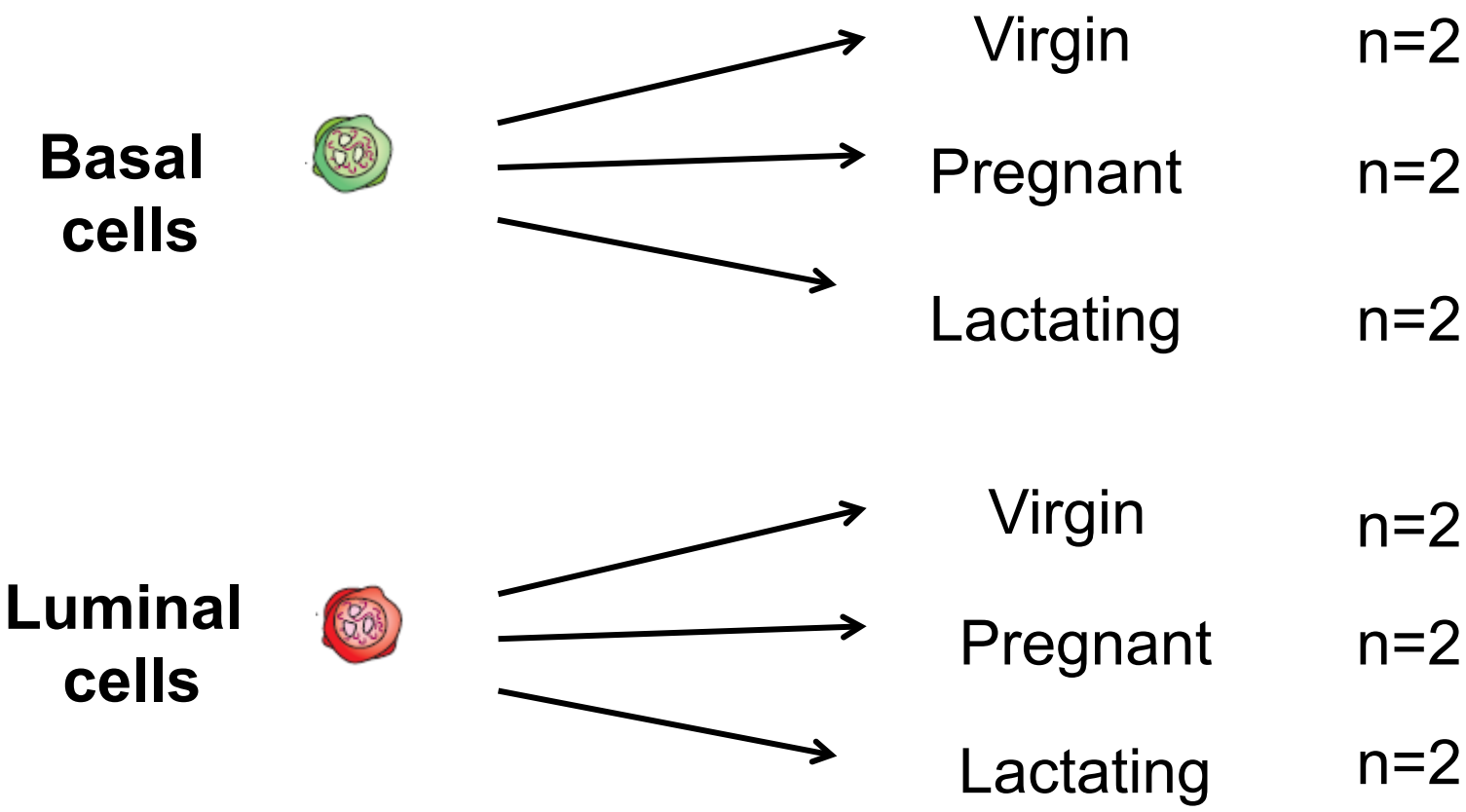
Estimated gene-specific parameters used for gene prioritization and gene set testing

limma package: Linear Models for Microarrays & RNA-seq

Analysis of differential expression studies

- arbitrarily complex experiments: linear models, contrasts
- empirical Bayes methods for differential expression: *t*-tests, *F*-tests, posterior odds
- analyse log-ratios, log-intensities, log-CPM values
- accommodate quality weights in analysis
- control of FDR across genes and contrasts
- many plotting functions to help visualize raw data and final results from statistical analysis
- gene set testing at various levels
- fast, numerically efficient methods

RNA-seq of Mouse mammary gland



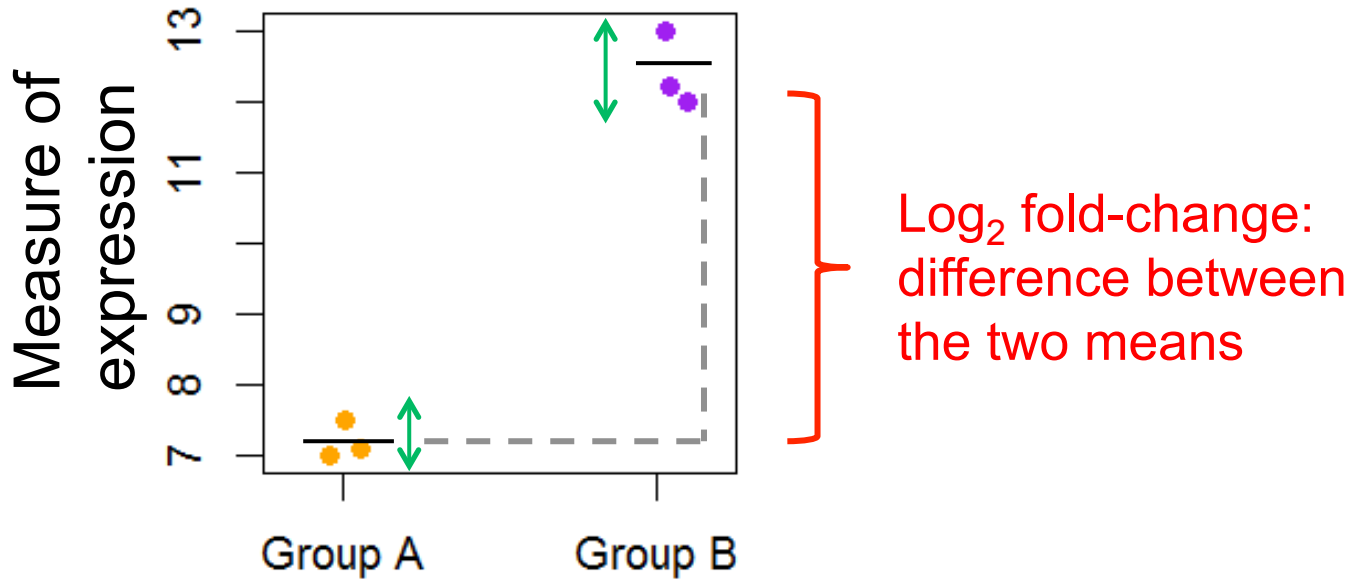
Fu *et al.* (2015) 'EGF-mediated induction of Mcl-1 at the switch to lactation is essential for alveolar cell survival' Nat Cell Biol

(some) questions we can ask

- Which genes are differentially expressed between basal and luminal cells?
- … between basal and luminal in virgin mice?
- … between pregnant and lactating mice?
- … between pregnant and lactating mice in basal cells?

What do we need to perform a statistical test?

- Measure of average expression
- Measure of variability



One of the most useful statistics: *t*-test

- We want to test the null hypothesis:

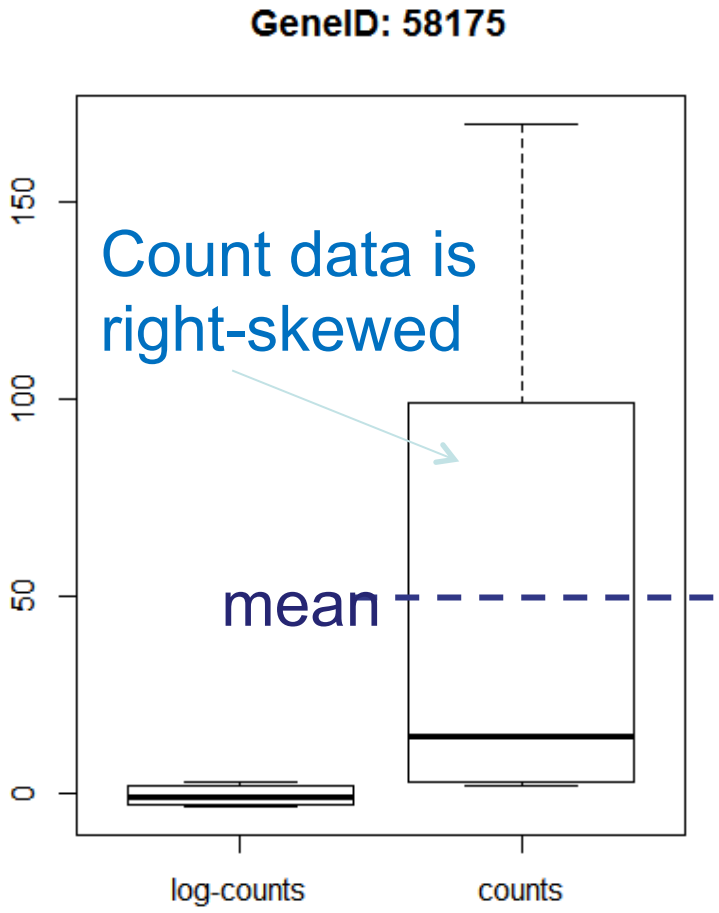
  H0: mean(GroupA) = mean(GroupB)

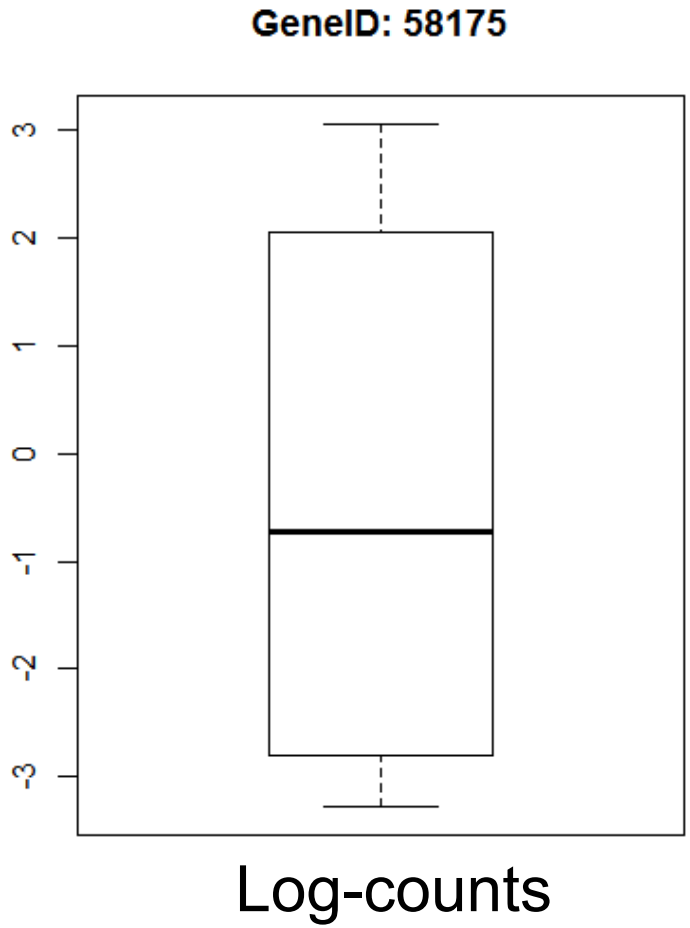
  against the alternative hypothesis:

  H1: mean(GroupA) ≠ mean(GroupB)

- An important assumption of the *t*-test is that the data is roughly normally distributed
- A statistician's best trick is to transform data that isn't normally distributed into something that looks more normally distributed

Log-counts vs counts for one gene





*A quick check to see how normal your data is: compare the mean and the median

We can perform *t*-tests on log-counts

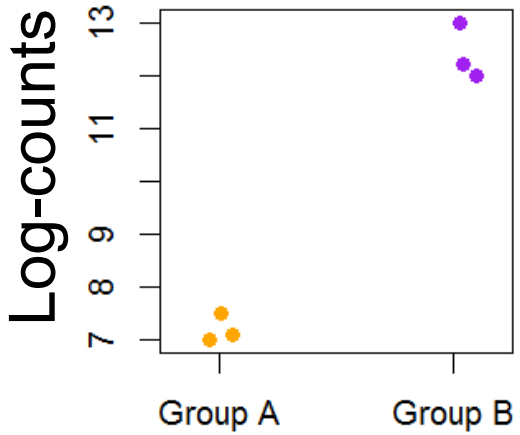
- Take into account different sequencing depths
- Take into account normalisation factors
- Take into account we can't log a zero
- The cpm(y,log=TRUE) function does this for you

Now we have log-counts

- Calculate means and variances on the log-counts

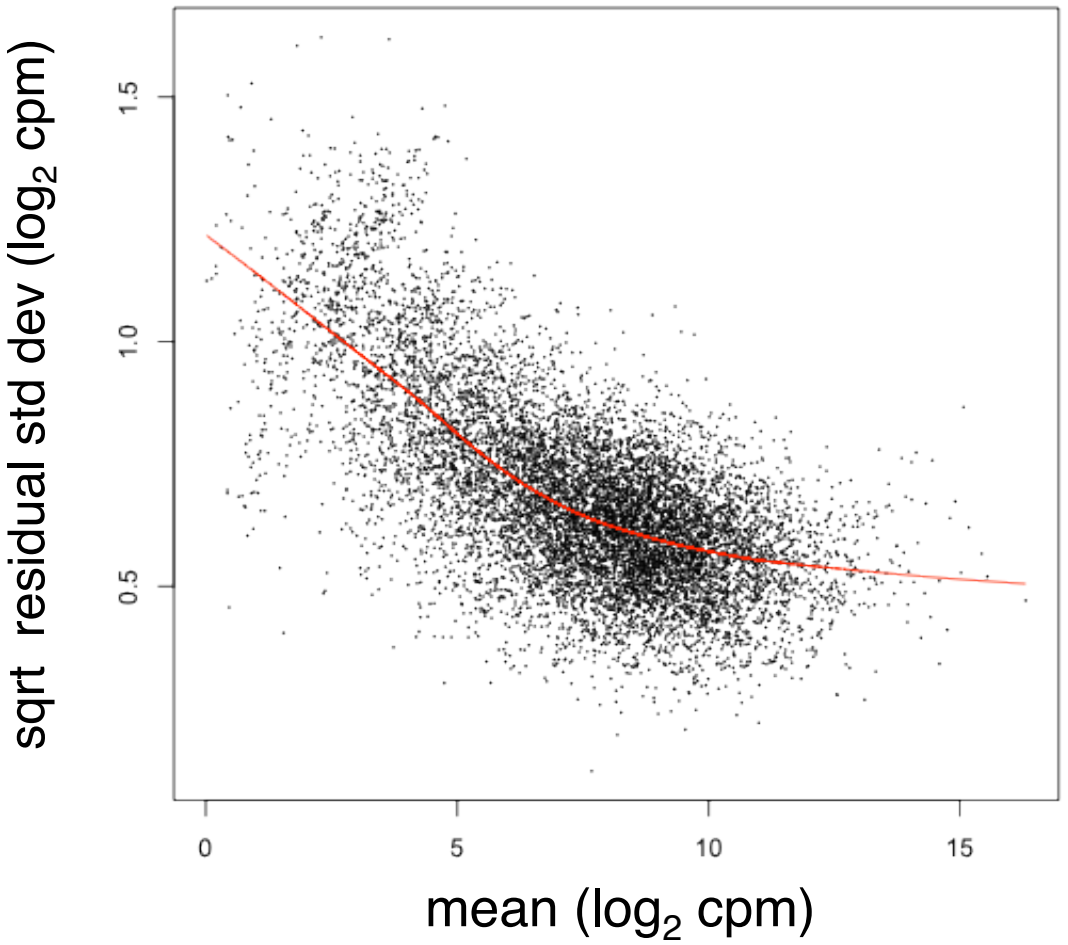
$$T = logFC / StdDev / \sqrt{n}$$

- logFC is the difference in means between the two groups for the log-counts
- The t-statistic is t-distributed on n-1 degrees of freedom
- P-values!



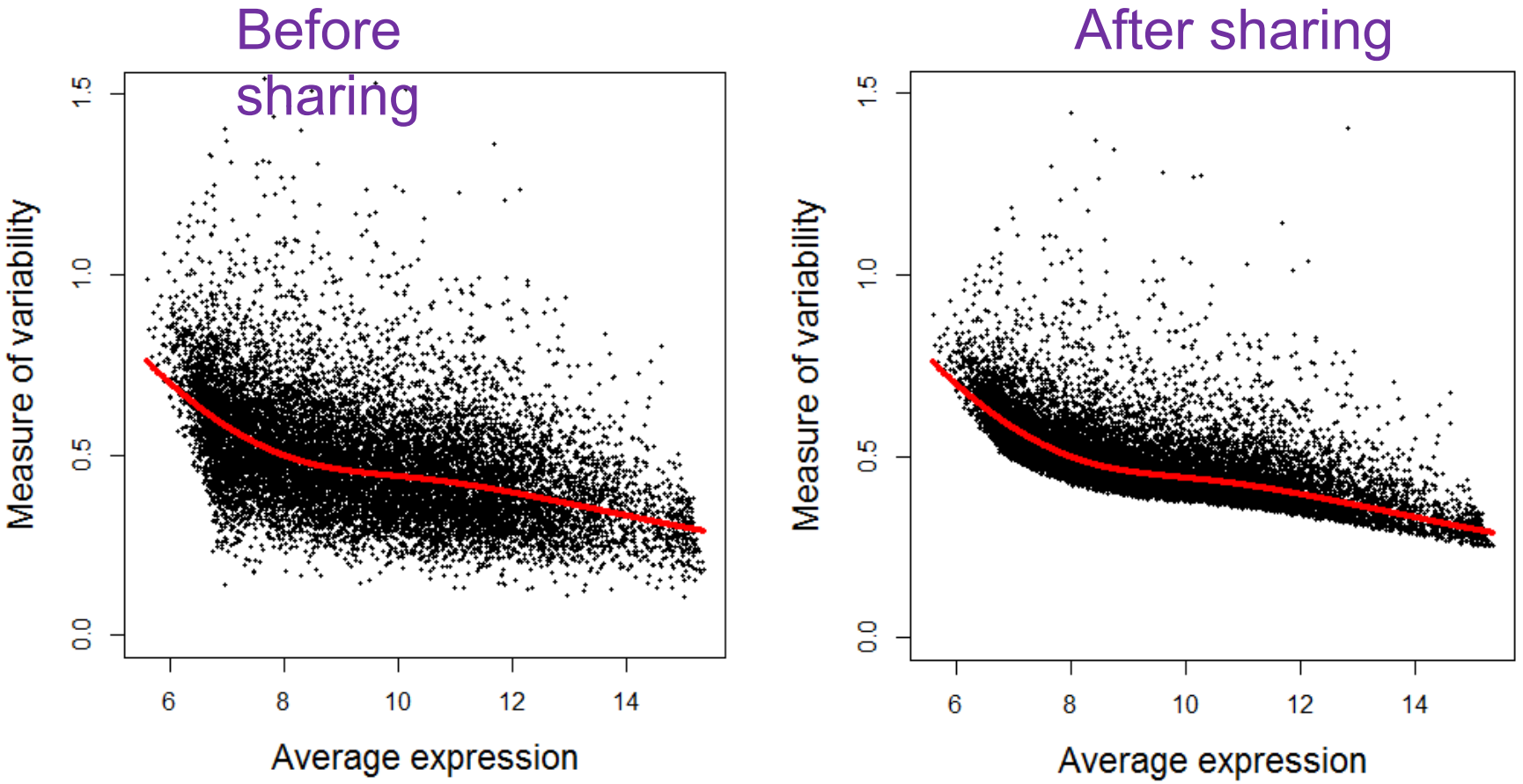
RNA-seq data is more complicated

- Mean-variance relationship. Use *voom*



lowess fit

Although we test one gene at a time, we can share information about all the genes to help with testing



Multiple testing burden

- Problem: We are performing tens of thousands of tests, which increases our chances of getting false discoveries
- Solution: Calculate false discovery rates ("adjusted *p*-values" in *limma*)
- Interpretation: If there are 100 genes significant at FDR< 5%, we are willing to accept that 5 will be false discoveries

Linear modelling analysis pipeline for RNA-seq data

- `model.matrix / makeContrasts`
- `voom`
- `lmFit`
- `contrasts.fit`
- `treat`
- `eBayes`
- `topTable / topTreat`
- `decideTests`