



Statistical models for count data analysis

Mark D. Robinson, Institute of Molecular Life Sciences

- simple counting (and new alternatives ..)
- why the negative binomial distribution?
- dispersion estimation and information sharing
- normalization considerations
- how about transformations of count data → limma?



A screenshot of a GitHub repository page. The repository name is 'sta426hs2016 / answers'. The status is 'Private'. The 'Code' tab is selected. There are 0 issues, 0 pull requests, 0 projects, and a Wiki tab.



Announcements

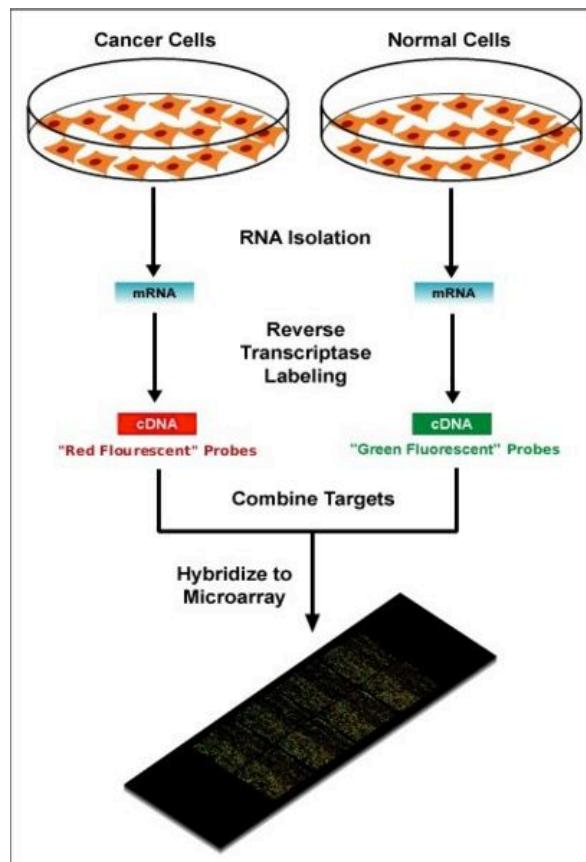
Plan for project: 28th November (send me a short description of your plan – 3-4 sentences)

Exercises: see comments as Github issues. If you haven't received feedback for week 6/7 exercises, let me know ASAP.

Exercises: solutions posted to 'answers' repo.

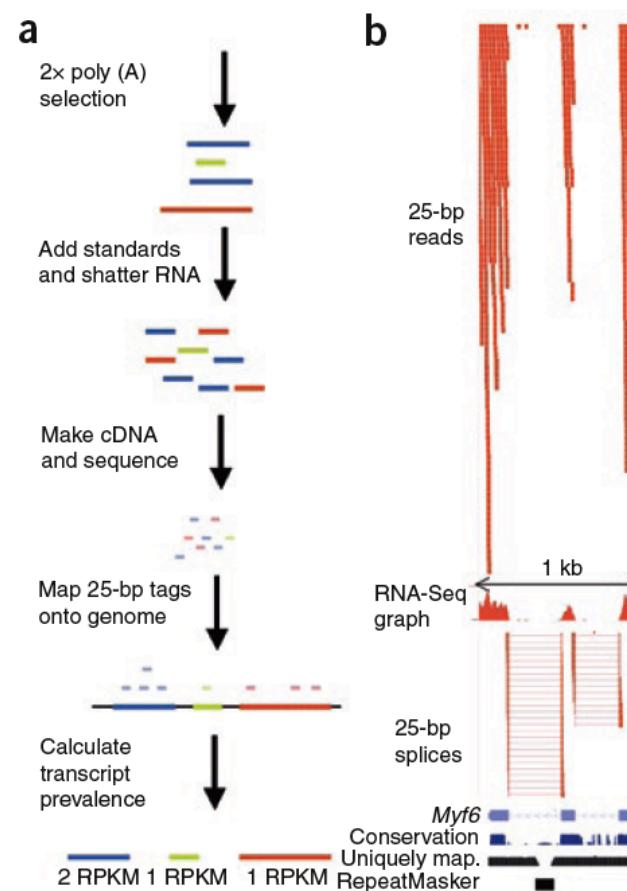


Abundance by Fluorescence Intensity



http://en.wikipedia.org/wiki/DNA_microarray

Abundance by Counting

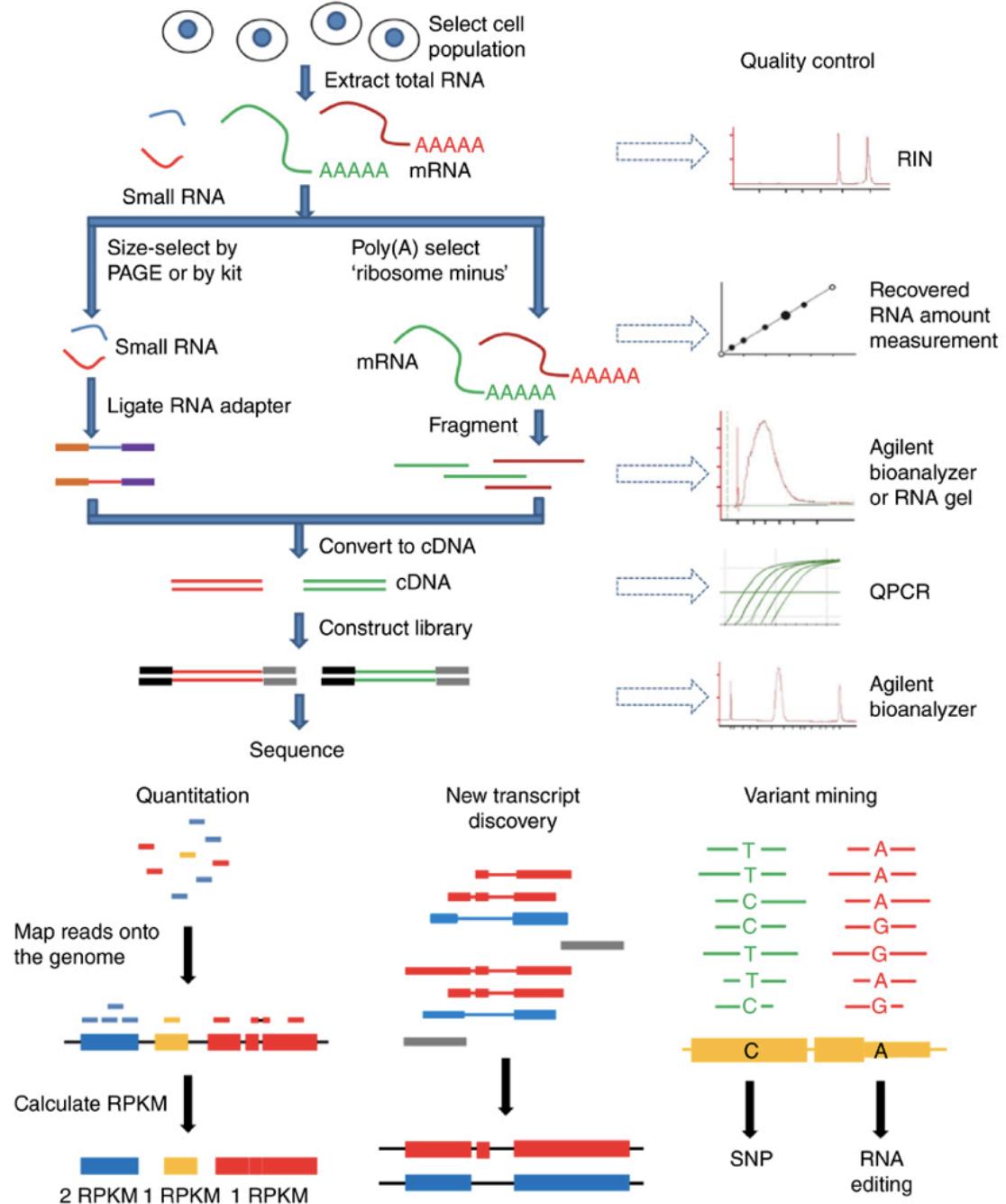


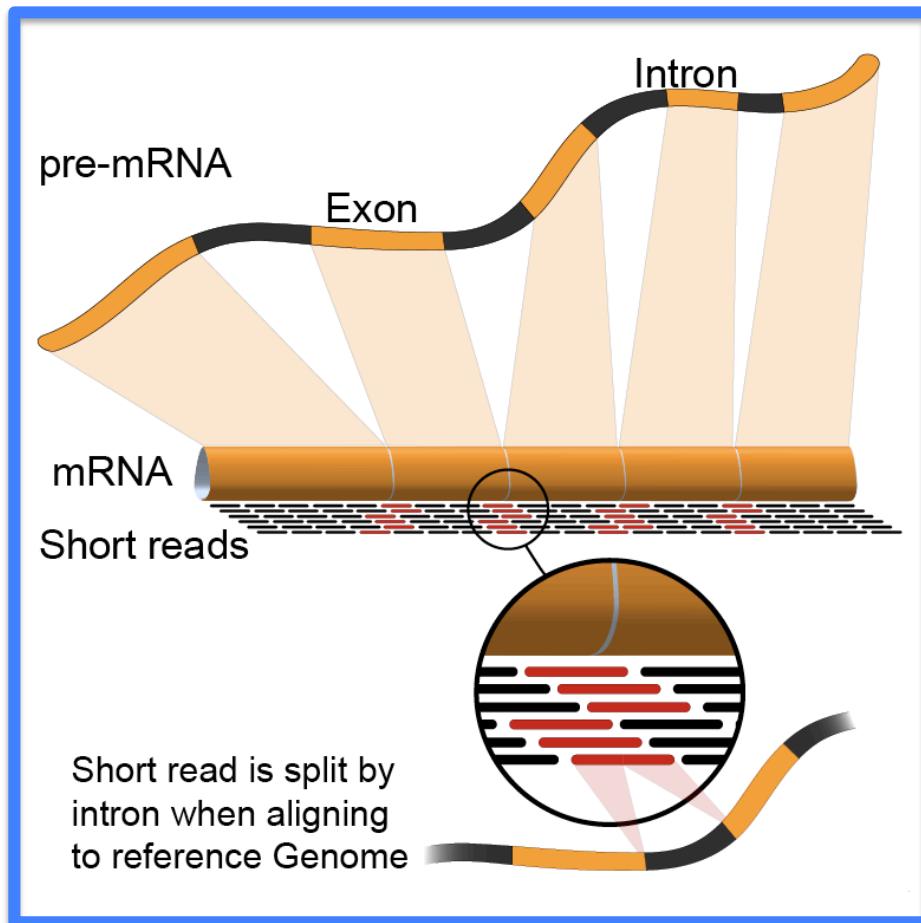
Mortazavi et al., Nature Methods, 2008



RNA-seq differential expression analyses

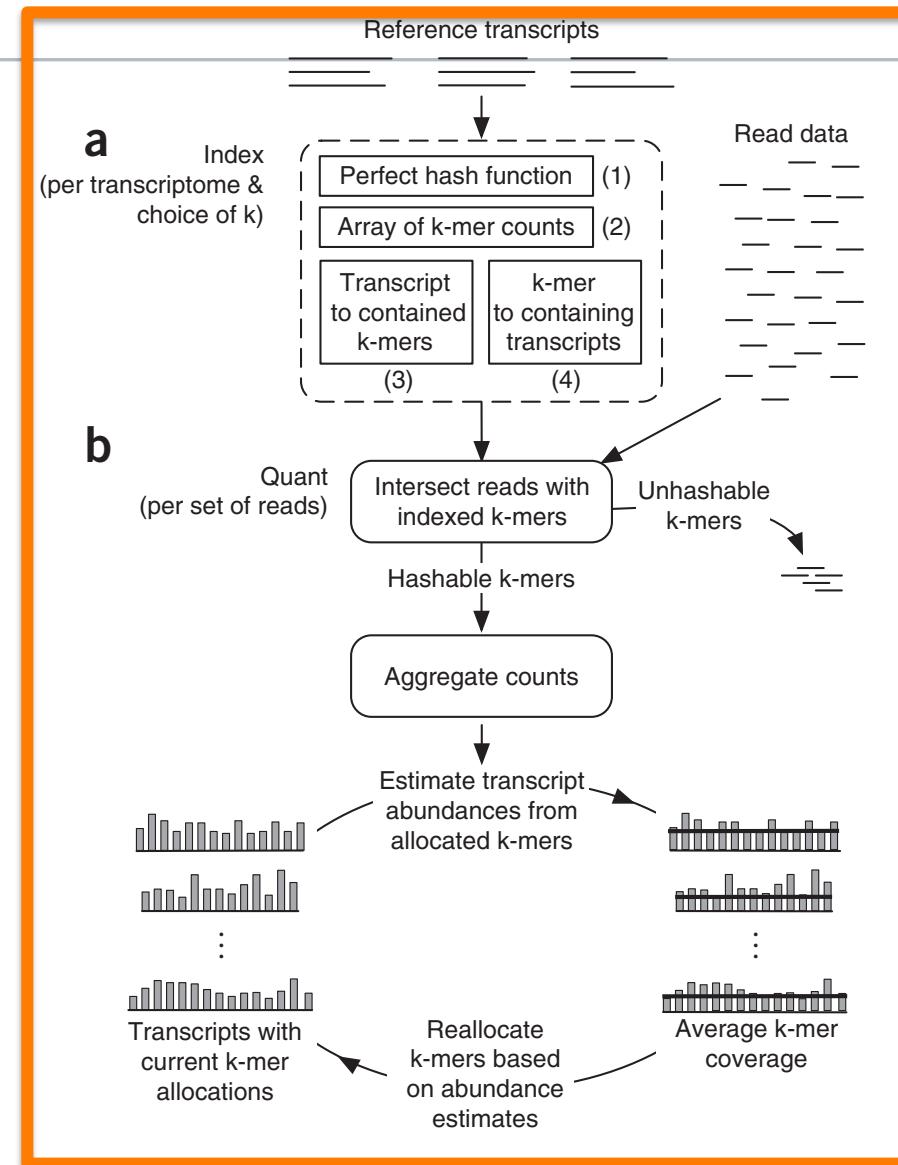
1. Map the reads to reference sequences
2. “Count” reads that map to genes (quantify)
3. Compute DE Statistics





<https://en.wikipedia.org/wiki/RNA-Seq>

Alignment versus quasi-alignment

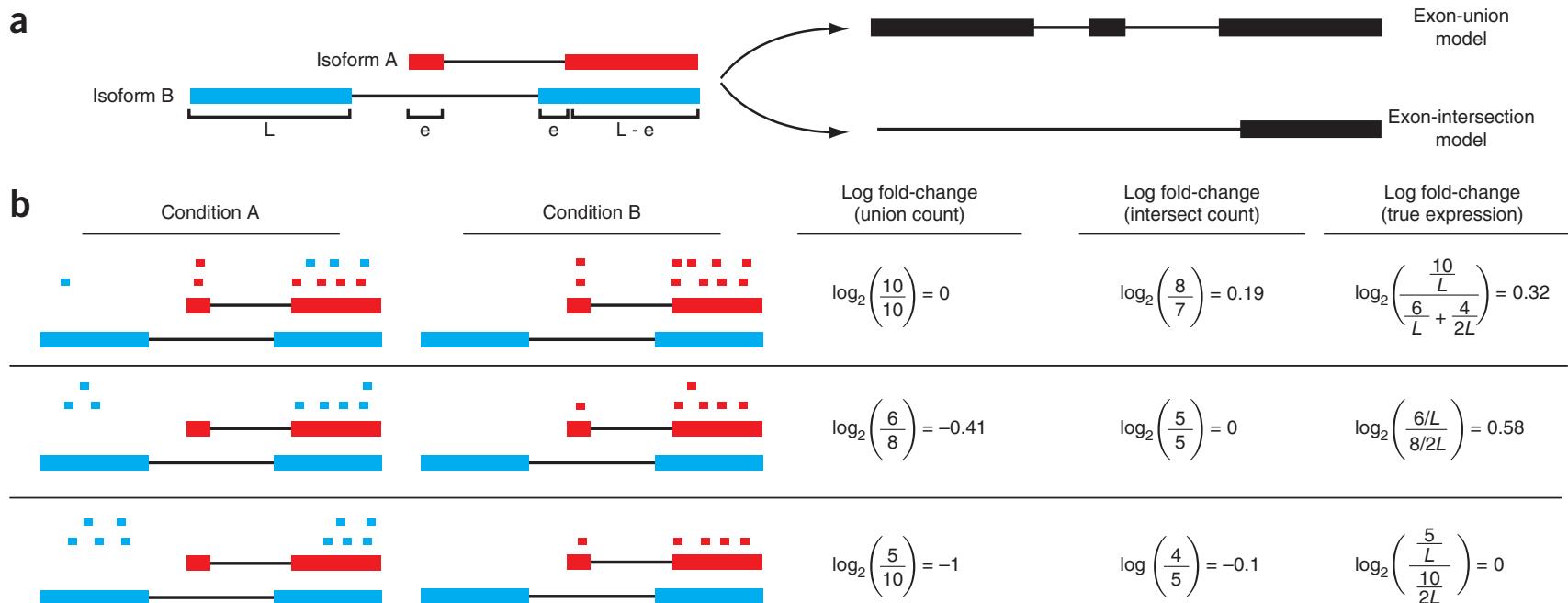


sailfish (Patro et al. 2014)



Caveat: simple gene-level counting not perfect, but good first approximation

Trapnell et al. 2013 Nat Biotech



Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene

Mar González-Porta¹, Adam Frankish², Johan Rung¹, Jennifer Harrow² and Alvis Brazma^{1*}

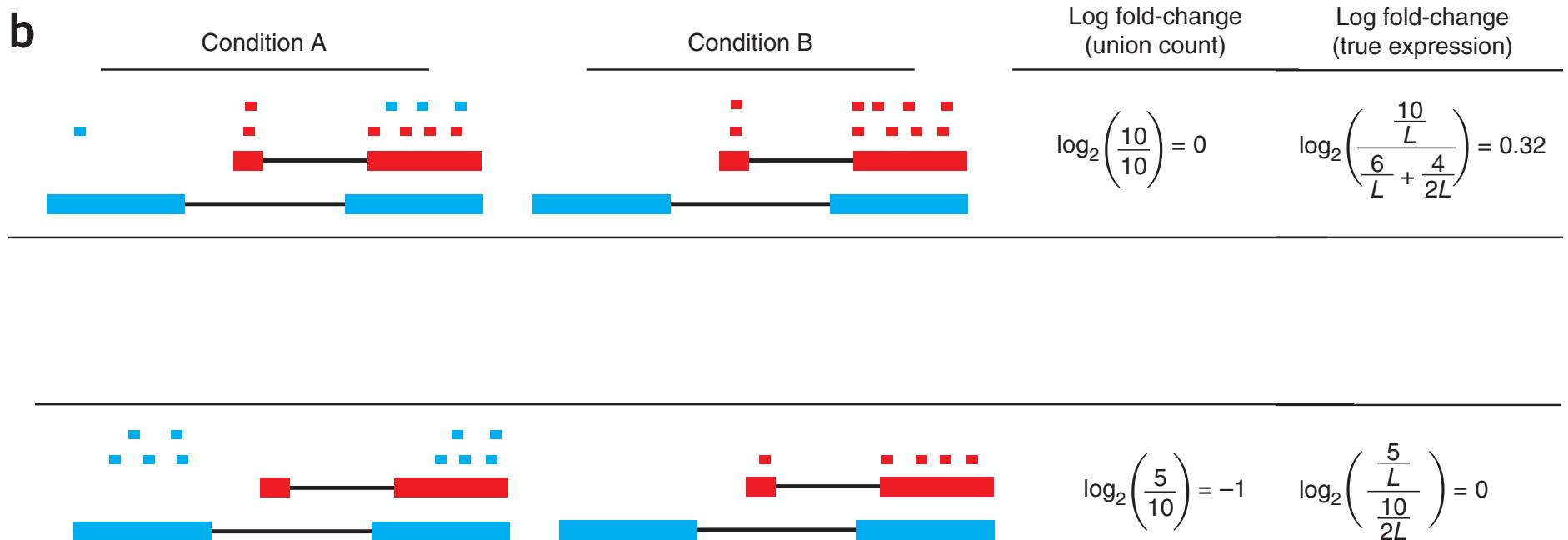
Mark D. Robinson, IMLS, UZH

Page 6



Counting/Quantification

- union counters → simple sum of all reads
transcript counters → sum of length-normalized reads
(often unknown which reads map to which transcript → portioning)



adapted from Trapnell et al. 2013 Nat Biotech



Quick plug: open science, PPR

F1000Research

F1000Research 2016, 4:1521 Last updated: 05 APR 2016



METHOD ARTICLE

REVISED Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]

Charlotte Soneson^{1,2}, Michael I. Love^{3,4}, Mark D. Robinson^{1,2}

¹Institute for Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland

²SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, 8057, Switzerland

³Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA, 02210, USA

⁴Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA, 02115, USA

v2 First published: 30 Dec 2015, 4:1521 (doi: [10.12688/f1000research.7563.1](https://doi.org/10.12688/f1000research.7563.1))

Latest published: 29 Feb 2016, 4:1521 (doi: [10.12688/f1000research.7563.2](https://doi.org/10.12688/f1000research.7563.2))



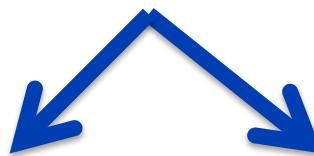


Differential expression: why not use methods developed for microarrays?

Count data is discrete, not continuous.

Methods designed for microarrays are not directly applicable and suboptimal (**more on this later**)

Two options:

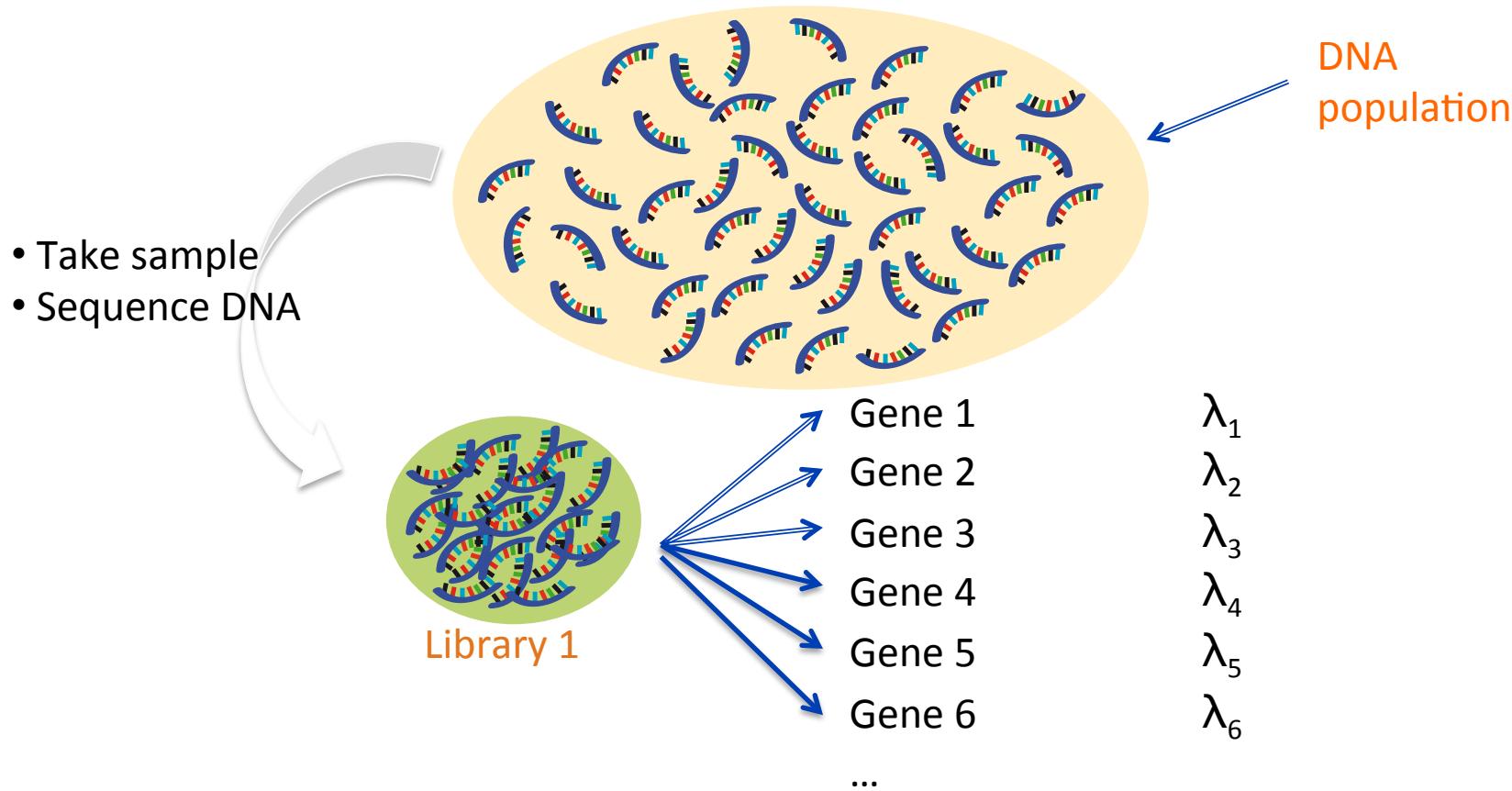


Transform count data
and apply standard
methodology

Analyze using
models for count
data

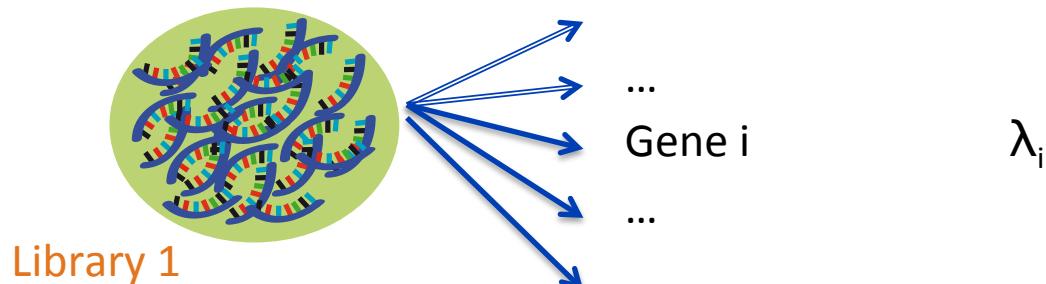


Sampling reads from population of DNA fragment is multinomial





For a single gene, it's a coin toss, i.e. Binomial



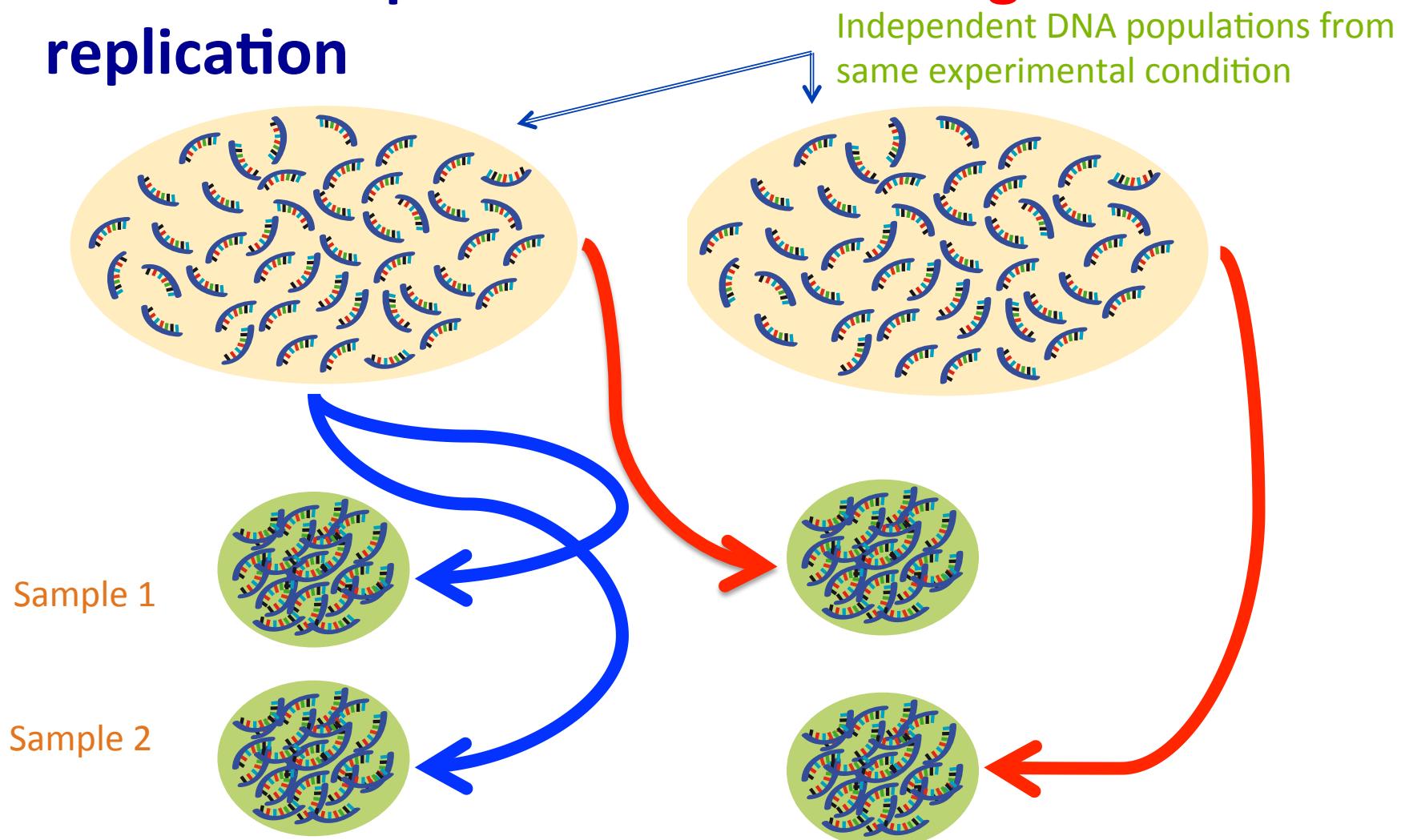
$$Y_i \sim \text{Binomial}(M, \lambda_i)$$

- Y_i - observed number of reads for gene i
- M - total number of sequences
- λ_i - proportion

Large M , small $\lambda_i \rightarrow$ approximated well by Poisson($\mu_i = M \cdot \lambda_i$)

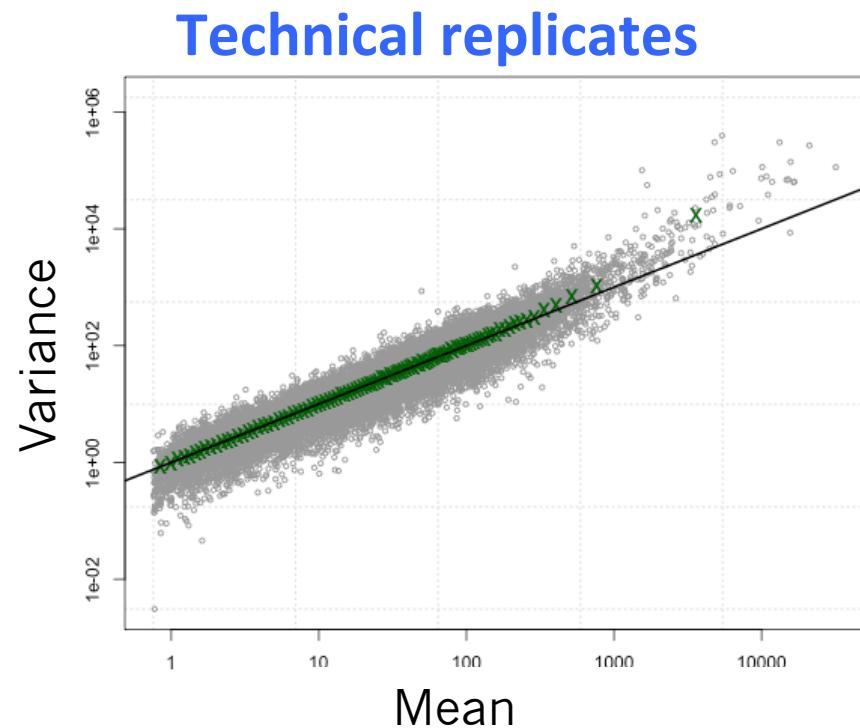


Technical replication versus biological replication

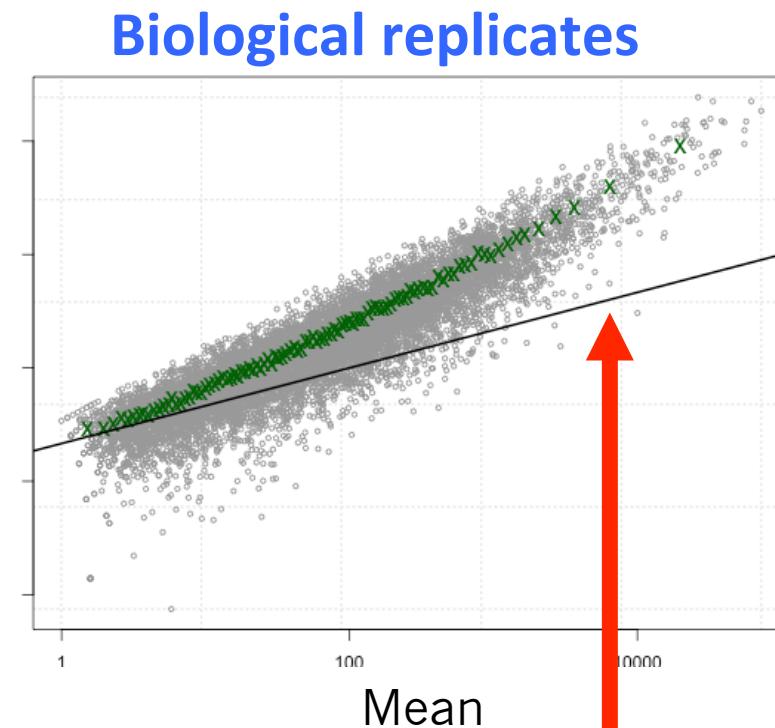




Mean-Variance plots: What we see in real data



Data from Marioni et al. *Genome Research* 2008



Data from Parikh et al.
Genome Biology 2010

mean=variance
(Poisson assumption)



Count data modeling assumptions

Poisson adequately describes technical variation

$$Y_i \sim \text{Pois}(M * \lambda_i)$$

$$\text{mean}(Y_i) = \text{variance}(Y_i) = M * \lambda_i$$

Negative binomial (gamma-Poisson) model is a natural extension that allows biological variability:

$$Y_i \sim \text{NB}(\mu_i = M * \lambda_i, \phi_i)$$

Same mean, variance is quadratic in the mean:

$$\text{variance}(Y_i) = \mu_i (1 + \mu_i \phi_i)$$

M = library size

λ_i = relative contribution of gene i



Similar interpretation

$$Y_i \sim NB(\mu_i = M * \lambda_i, \phi_i)$$

$$E(y_{gi}) = \mu_{gi} = N_i \pi_{gi}.$$

Coefficient of variation = standard deviation/mean

$$\text{var}(y_{gi}) = E_\pi[\text{var}(y|\pi)] + \text{var}_\pi[E(y|\pi)] = \mu_{gi} + \phi_g \mu_{gi}^2.$$

Dividing both sides by μ_{gi}^2 gives

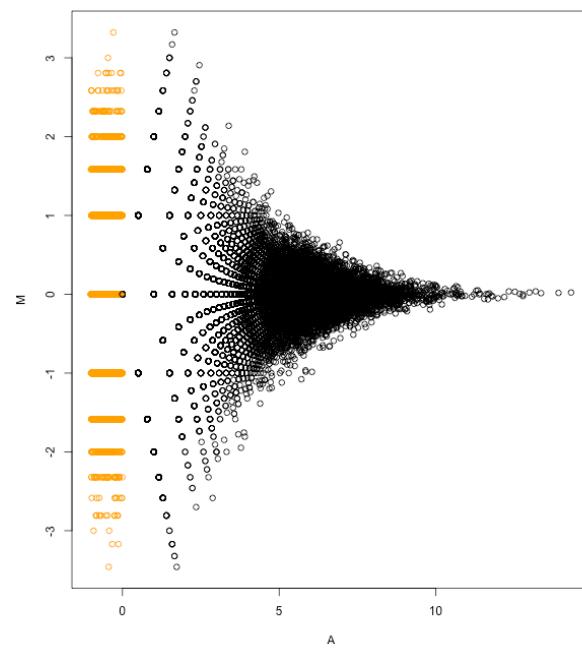
$$CV^2(y_{gi}) = 1/\mu_{gi} + \phi_g.$$



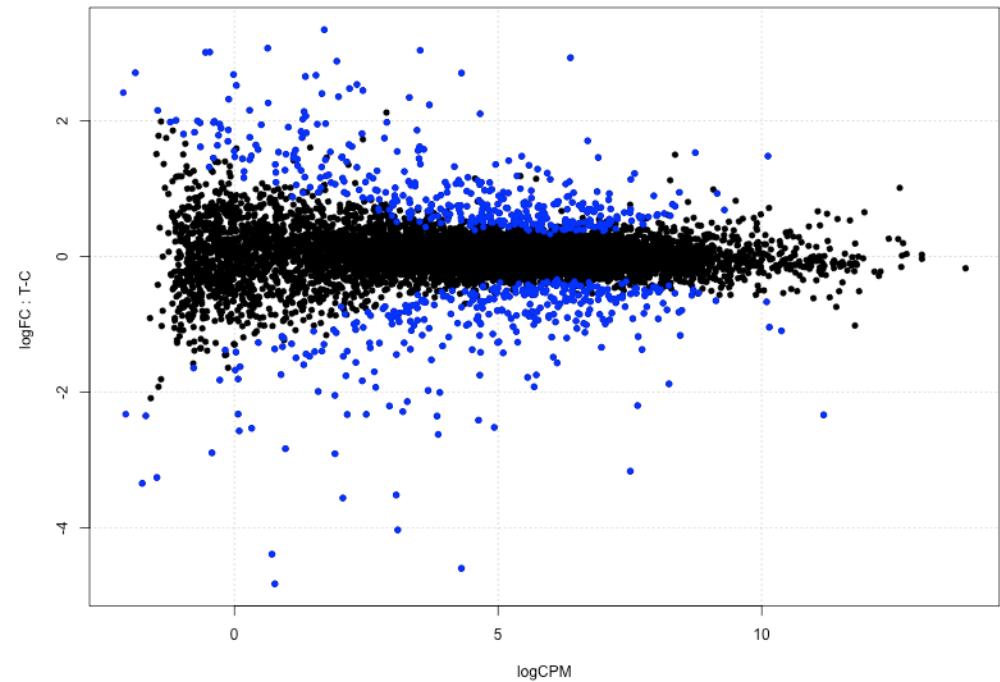
$$CV^2(y_{gi}) = 1/\mu_{gi} + \phi_g.$$

A confirmation of what the theory states

Technical replicates
(~Poisson)



Biological replicates





What was successful with microarray data: classical/moderated/shrunken t-tests

$$t_g = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{s_g c}$$

Feature-specific

$$\tilde{t}_g = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{\tilde{s}_g u}$$

Moderated

$$t_{g,\text{pooled}} = \frac{\bar{y}_{\text{mu}} - \bar{y}_{\text{wt}}}{s_0 c}$$

Common



Let's try the same strategy with counts

At one extreme, assume all genes have same dispersion (too strong)

At other extreme, estimate dispersion separately/independently for each gene (poor estimates)

Shrink individual estimates toward common/trend (how?)

No hierarchical model (e.g. limma) to do this:
approximations, weighted likelihood

No t-distribution theory to formulate statistical tests.



Count data modeling assumptions

Poisson adequately describes technical variation

$$Y_i \sim \text{Pois}(M * \lambda_i)$$

$$\text{mean}(Y_i) = \text{variance}(Y_i) = M * \lambda_i$$

Negative binomial (gamma-Poisson) model is a natural extension that allows biological variability:

$$Y_i \sim \text{NB}(\mu_i = M * \lambda_i, \phi_i)$$

Same mean, variance is quadratic in the mean:

$$\text{variance}(Y_i) = \mu_i (1 + \mu_i \phi_i)$$

M = library size

λ_i = relative contribution of gene i



First challenge: getting good estimates of dispersion in small samples

Several choices here:

- Maximum Likelihood (MLE)

$$Y_{gij} \sim \text{NegBin}(\mu_{gi} = M_j \lambda_{gi}, \phi)$$

$$(\hat{\lambda}_{MLE}, \hat{\phi}_{MLE}) = \arg \max_{\lambda, \phi} l(\lambda, \phi)$$

- Pseudo-Likelihood (PL)

$$X^2 = \sum_{gij} \frac{(y_{gij} - \hat{\mu}_{gi})^2}{\hat{\mu}_{gi}(1 + \hat{\phi}_{PL}\hat{\mu}_{gi})} = G(n_1 + n_2 - 2)$$

- Quasi-Likelihood (QL)

$$D = 2 \sum_{gij} \left\{ y_{gij} \log \left[\frac{y_{gij}}{\mu_{gi}} \right] - (y_{gij} + \phi_{QL}^{-1}) \log \left[\frac{y_{gij} + \phi_{QL}^{-1}}{\mu_{gi} + \phi_{QL}^{-1}} \right] \right\}$$

- Conditional Maximum Likelihood (CML)

- Approximate Conditional Inference (Cox-Reid)

- *quantile-adjusted Maximum Likelihood (qCML)*



Conditional likelihood

Likelihood for single **negative binomial** observation:

$$f(y; \mu, \phi) = P(Y = y) = \frac{\Gamma(y + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y + 1)} \left(\frac{1}{1 + \mu\phi}\right)^{\phi^{-1}} \left(\frac{\mu}{\phi^{-1} + \mu}\right)^y$$

If all libraries are the same size (i.e. $m_i \equiv m$),
the sum $Z = Y_1 + \dots + Y_n \sim \text{NB}(nm\lambda, \phi n^{-1})$.

Can form conditional likelihood:

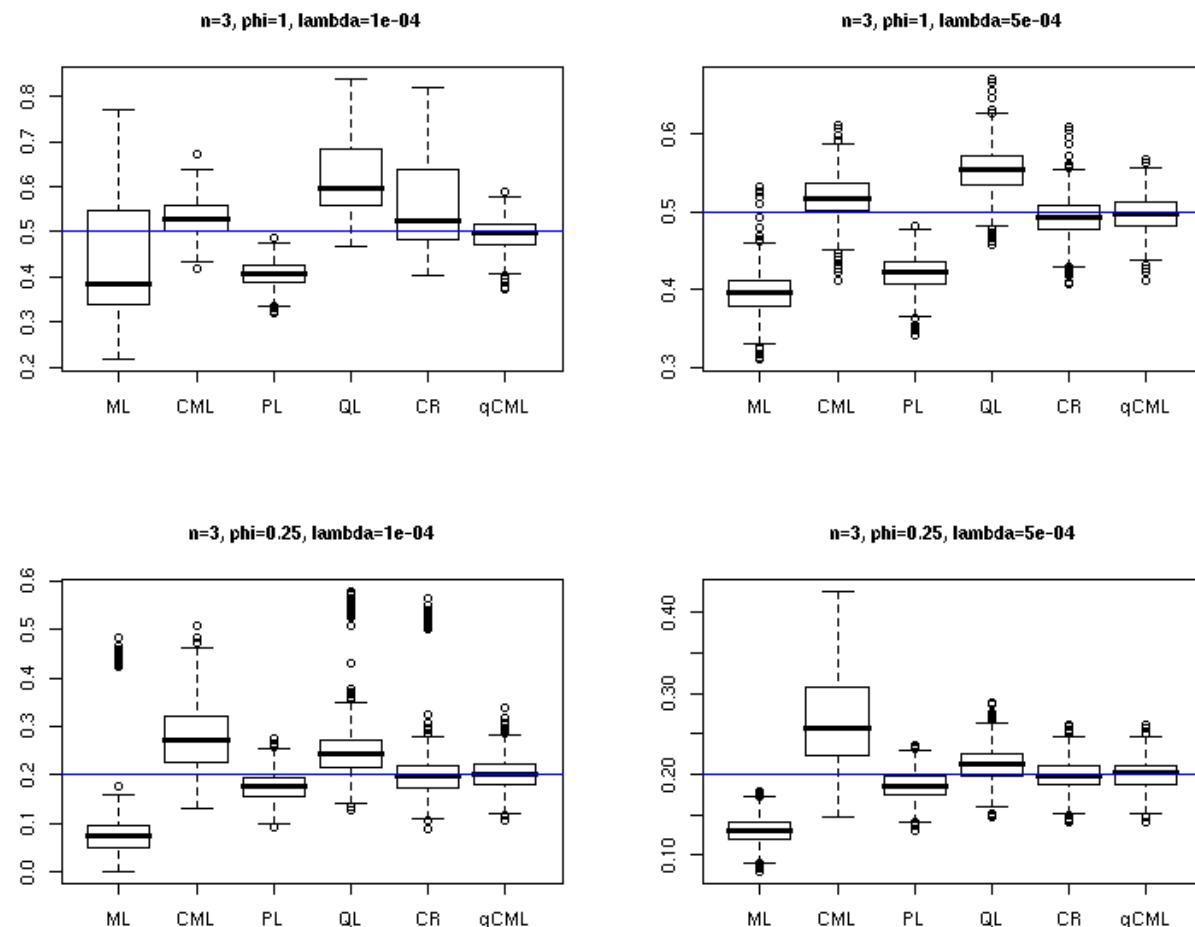
$$l_{Y|Z=z}(\phi) = \left[\sum_{i=1}^n \log \Gamma(y_i + \phi^{-1}) \right] + \log \Gamma(n\phi^{-1}) - \log \Gamma(z + n\phi^{-1}) - n \log \Gamma(\phi^{-1}).$$



Comparison of Estimators (Common Dispersion)

Horizontal blue line is
TRUE value.

qCML performs best
under a wide range
of conditions.





Likelihood ... Weighted likelihood

Maximum Likelihood Estimation (MLE)

Likelihood: $L(X; \theta) = \prod_i^n f(x_i; \theta)$

log-likelihood:

$$l(X; \theta) = \log(L(X; \theta)) = \sum_i^n \log(f(x_i; \theta))$$

MLE: $\hat{\theta} = \arg \max_{\theta} l(X; \theta)$



Likelihood ... Weighted likelihood

Weighted likelihood

$WL(X; \theta) = \prod_i^n f(x_i; \theta)^{w_i}$, where w_i is weight.

$$wl(X; \theta) = \log(WL(X; \theta)) = \sum_i^n w_i \log(f(x_i; \theta))$$

$$\hat{\theta} = \arg \max_{\theta} wl(X; \theta)$$



Second challenge: Moderate dispersion estimate

Weighted likelihood -- individual log-likelihood plus a weighted version of the **common** log-likelihood:

$$WL(\phi_g) = l_g(\phi_g) + \alpha l_C(\phi_g)$$

↑
 $(1-\alpha)$

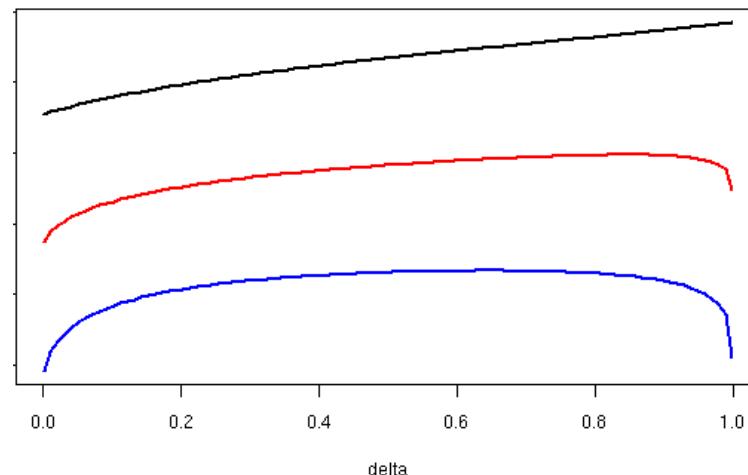
L_g - quantile-adjusted conditional likelihood

Black: single tag

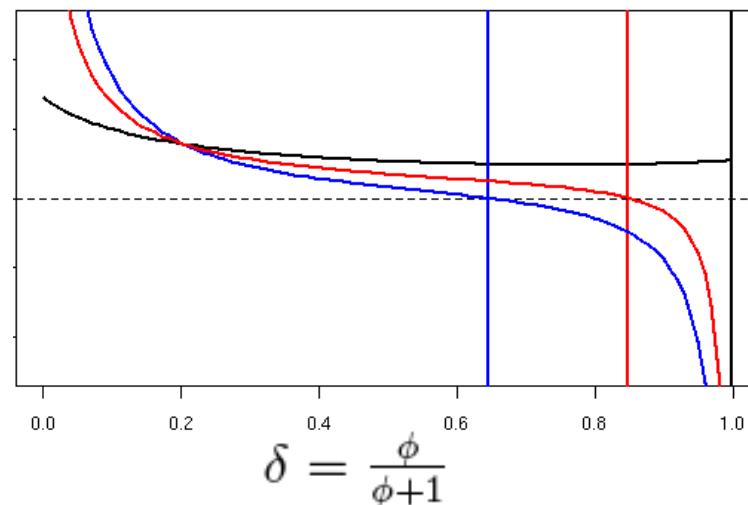
Blue: common dispersion

Red: Linear combination of the two

Log-Likelihood

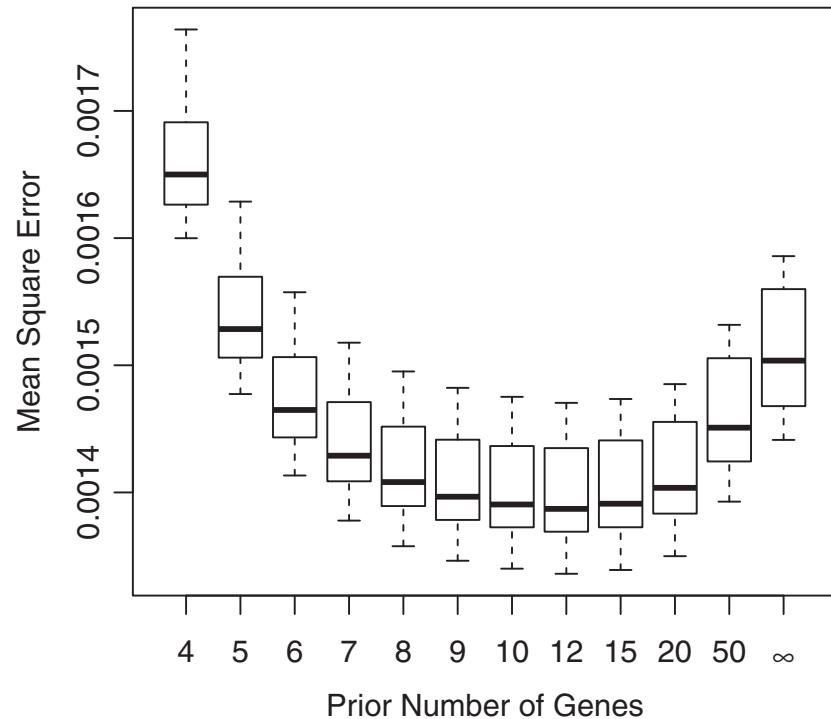


Score (1st derivative of LL)





How much to shrink?



Simulations suggest there is an optimal amount to shrink.

Challenge: choosing/estimating how much

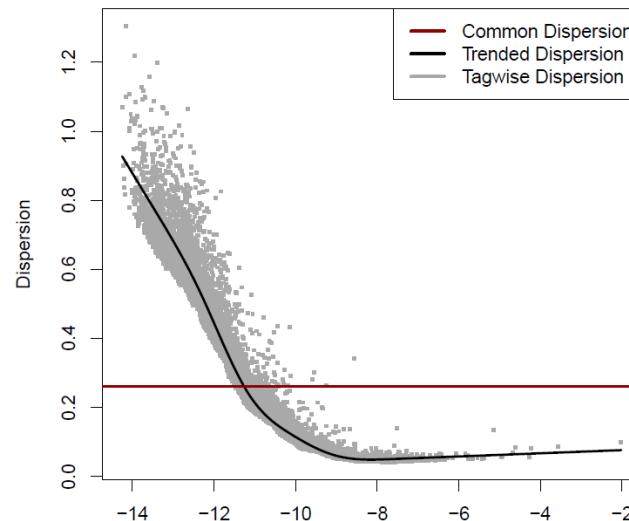
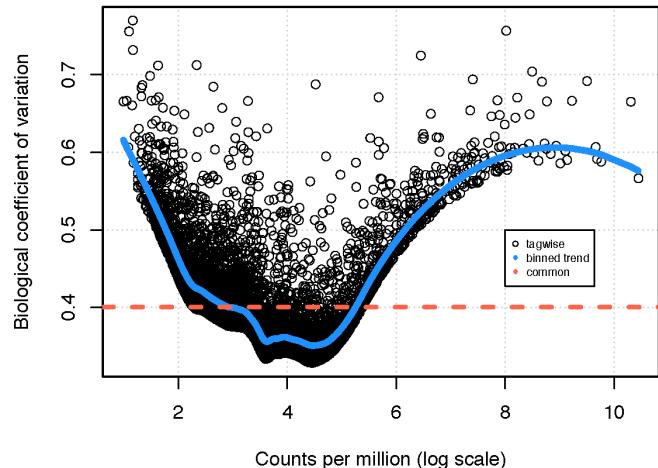
Figure 4. Mean-square error with which empirical Bayes genewise dispersions estimate the true dispersion (BCV^2), when true dispersions are randomly generated. In this case, the optimal prior weight is 10–12 prior genes, equivalent to 20–24 prior degrees of freedom. The common BCV estimator is equivalent to using infinite weight for the prior. Boxplots show results for 10 simulations.



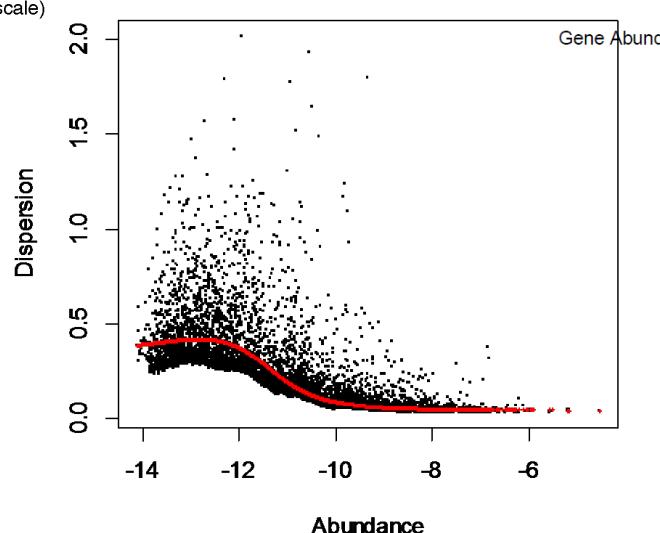
Dispersion varies with mean: moderate dispersion towards trend

Data:

Tuch et al.,
2008



Mouse
hemopoietic
stem cells



Mouse
lymphomas

Advantage: genes are allowed to have their own variance.



Nature Reviews Genetics | AOP, published online 18 November 2008; doi:10.1038/nrg2484

INNOVATION

RNA-Seq: a revolutionary tool for transcriptomics

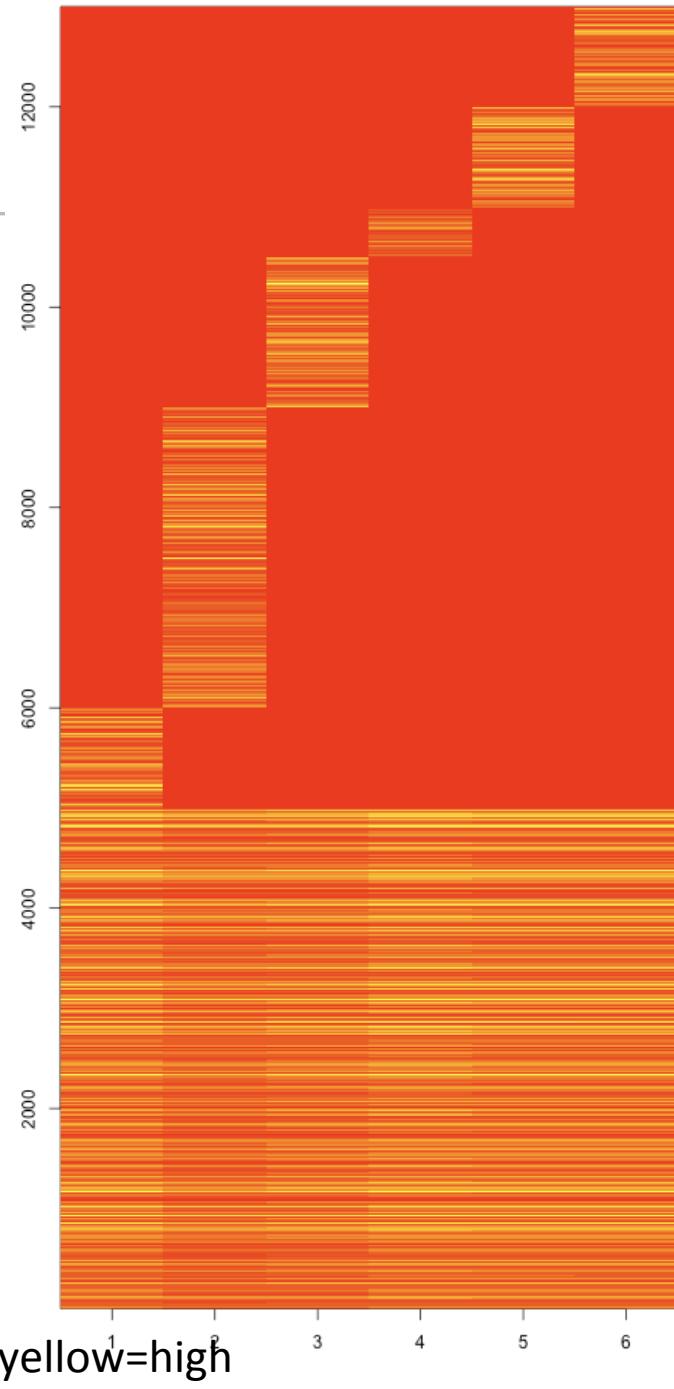
Zhong Wang, Mark Gerstein and Michael Snyder

One particularly powerful advantage of RNA-Seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets^{19,20,22}.



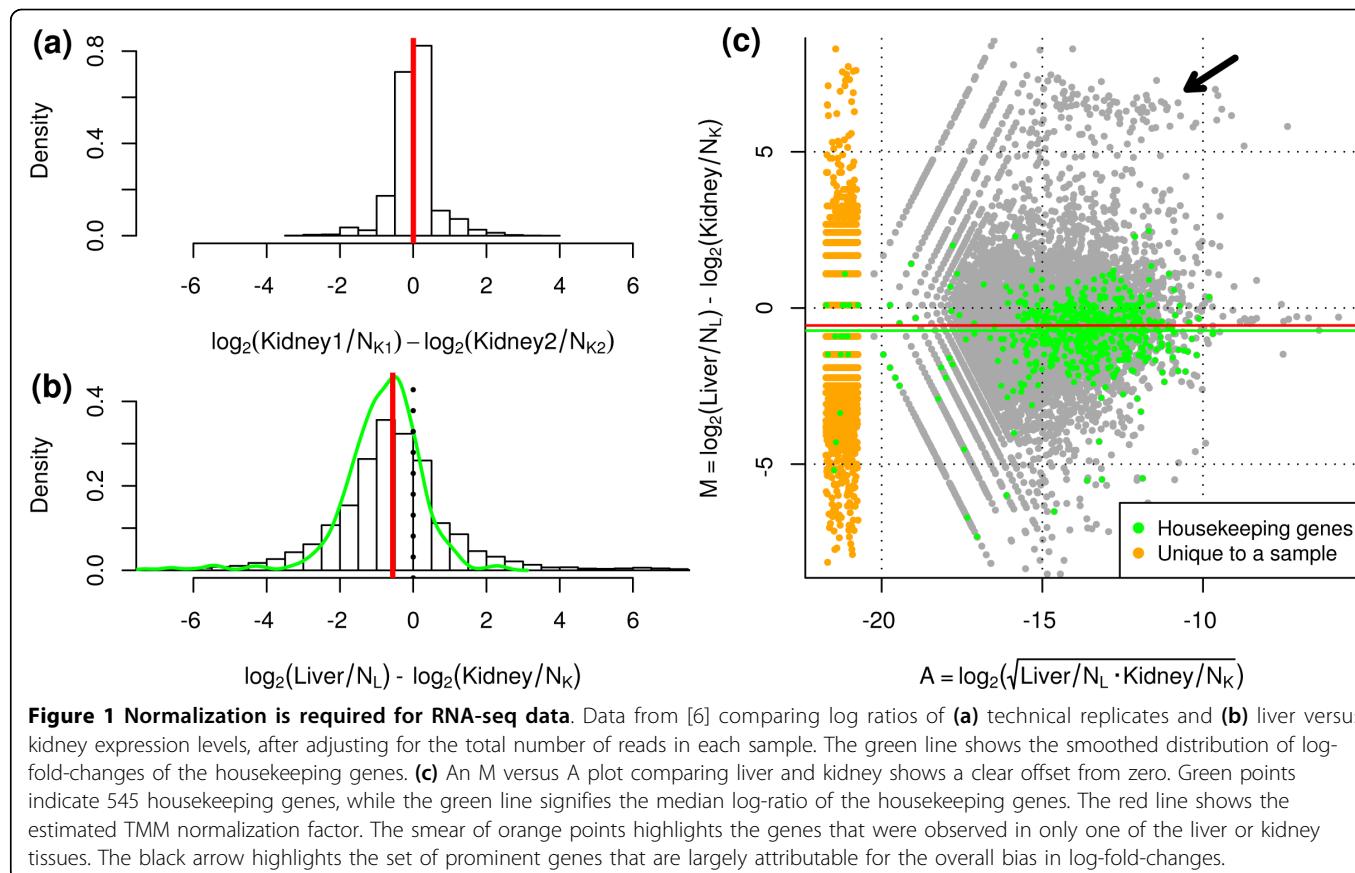
“Composition” or “Diversity” can affect read depth

- Hypothetical example: Sequence 6 libraries to the **same** depth, with varying levels of *unique-to-sample* counts
- Read depth is affected not only by expression (and length), but also expression levels of other genes
- Composition can induce (sometimes significant) differences in counts





Kidney and Liver RNA have very different composition





Use scaling factor (“offset”) in statistical model

Assumption: core set of genes/loci that do not change in expression.

Our Pick a reference sample, compute a weighted trimmed mean of M-values (TMM) to reference

Adjustment to statistical analysis:

- Use “effective” library size (edgeR)
- Use additional offset (GLM)

Note: count data is not modified



Differential expression: why not use methods developed for microarrays?

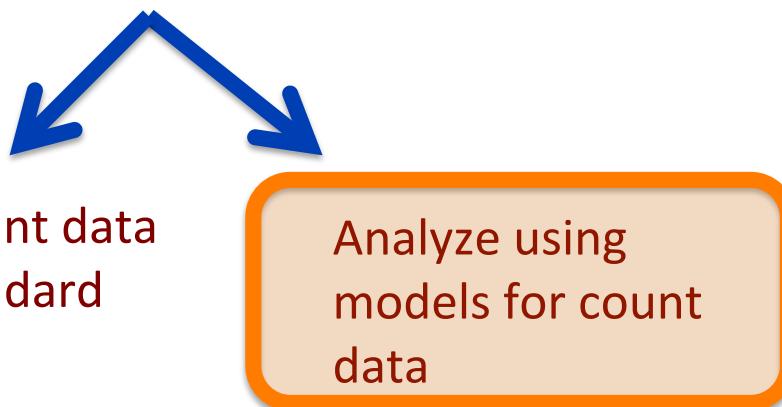
Count data is discrete, not continuous.

Methods designed for microarrays are not directly applicable and suboptimal

Transforming count data with logs, with some special treatment, can give very good results

Two options:

Transform count data
and apply standard
methodology





What does transformation do to M-V relationship?

For Poisson data, square-root should stabilize

Logarithm is too strong – variance decreases to asymptote (Neg Bin) or 0 (Poisson)

How to pick? Doesn't matter ... voom

voom: mean-variance modeling at the observational level

voom

package:limma

R Documentation

Transform RNA-Seq Data Ready for Linear Modelling

Description:

Transform count data to log2-counts per million, estimate the mean-variance relationship and use this to compute appropriate observational-level weights. The data are then ready for linear modeling.



Model log counts per million

log counts per million:

$$z_{gi} = \log_2 \left(1e6 \frac{\text{count}_{gi} + 0.5}{\text{libsize}_{gi} + 1.0} \right) = \log_2 \left(1e6 \frac{y_{gi} + 0.5}{M_{gi} + 1.0} \right)$$

normalize libsize in advance or normalize z_{gi} as for microarrays.

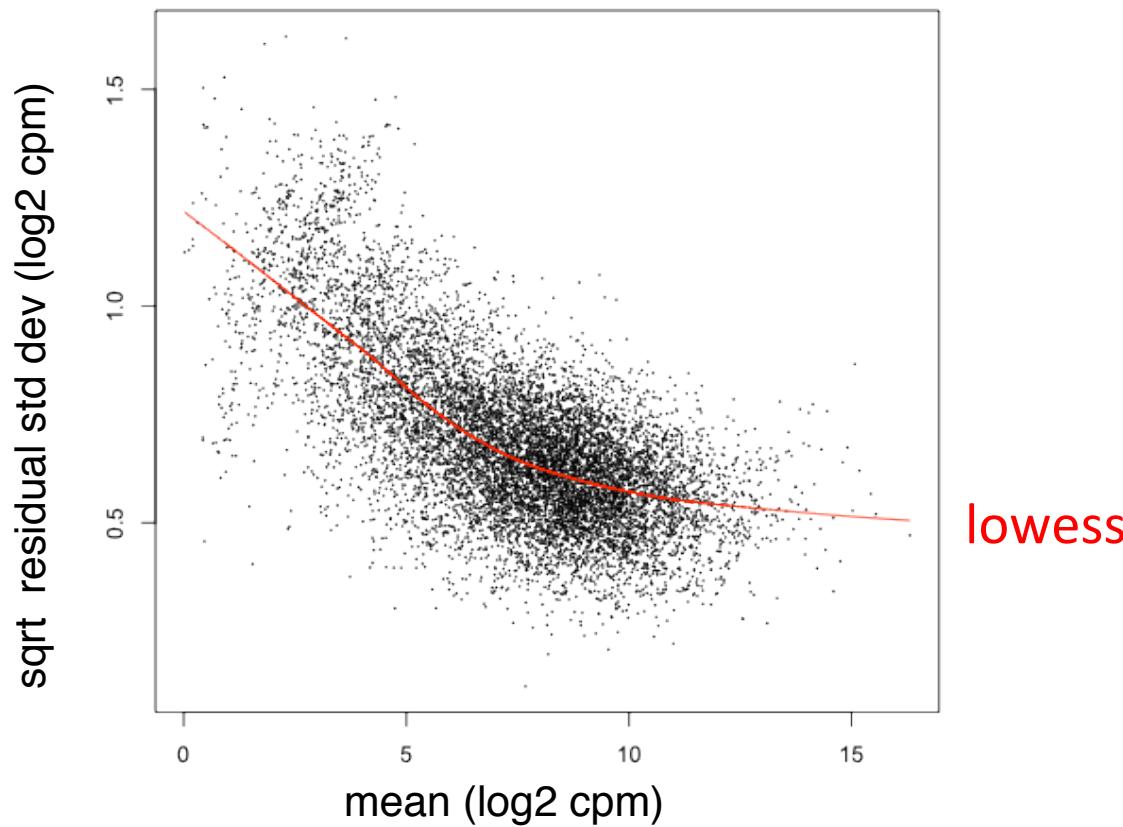
Linear modelling:

$$E(z_{gi}) = \mu_{gi} = x_i^T \beta_g$$

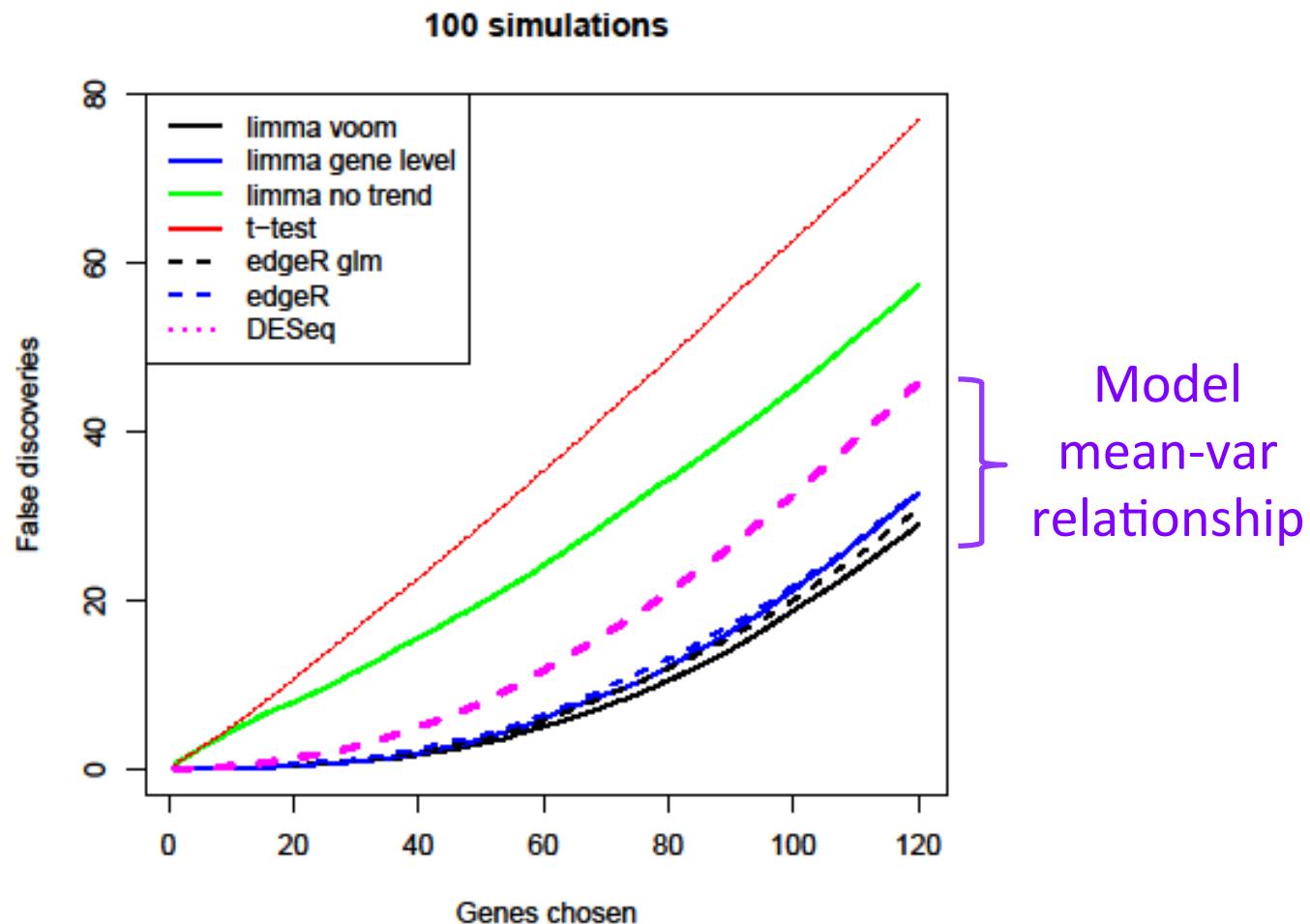
$$\text{var}(z_{gi}) = s(\mu_{gi}) \sigma_g^2$$



***voom* fits a lowess trend to the mean-variance relationship ...**



→ Use weights ($1/\text{var}$) in limma analysis



Law et al. 2014 Genome Biology – also a good candidate for journal club.