

Analysis of RNA-seq Data

Bernard Pereira



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

The many faces of RNA-seq



AREAS OF INTEREST ▾ TECHNIQUES ▾ SYSTEM

RNA Sequencing

Overview >

[Targeted RNA Sequencing](#)

[mRNA-Seq](#)

[Total RNA-Seq](#)

[Small RNA-Seq](#)

[Low-Quality/FFPE RNA-Seq](#)

[Ultra-Low-Input & Single-Cell RNA-Seq](#)

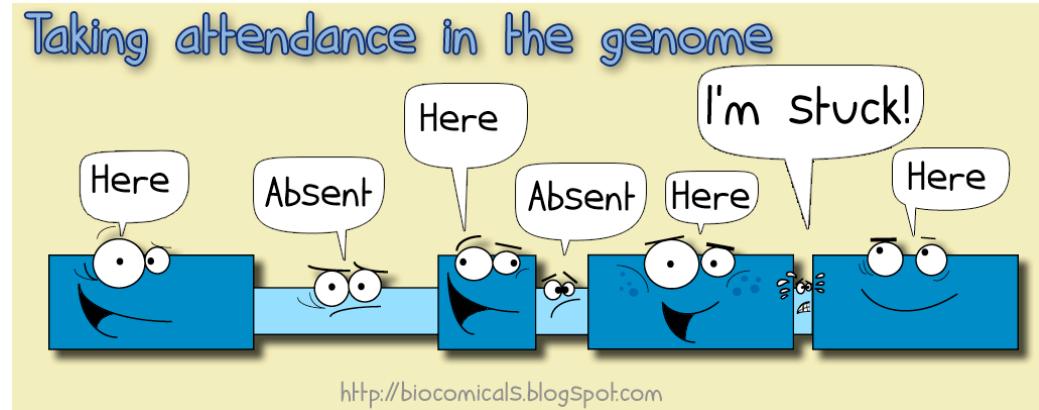
[Ribosome Profiling](#)

[RNA-Seq Data Analysis](#)

Applications

Discovery

- Find new transcripts
- Find transcript boundaries
- Find splice junctions



Comparison

Given samples from different experimental conditions, find effects of the treatment on

- Gene expression strengths
- Isoform abundance ratios, splice patterns, transcript boundaries

Applications

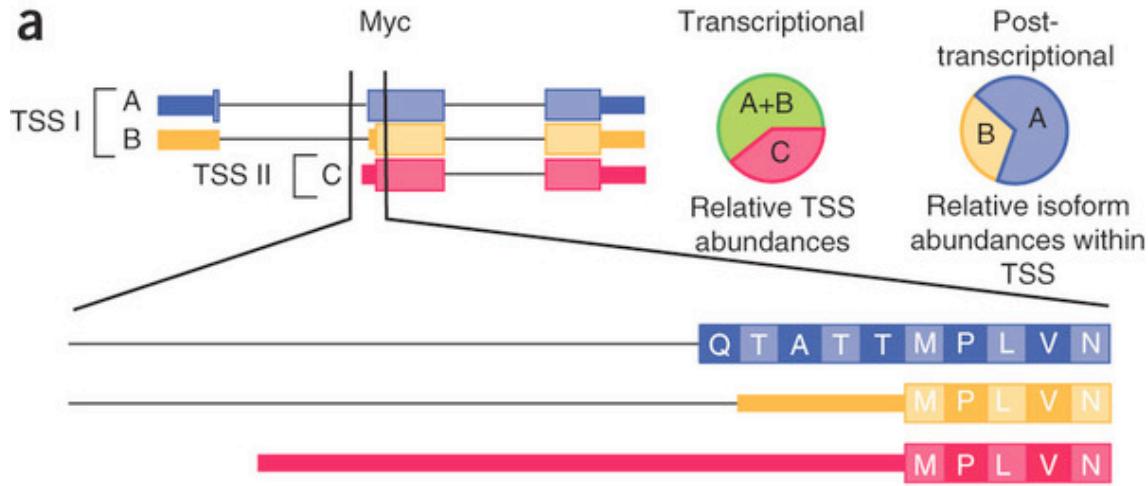
LETTERS

nature
biotechnology

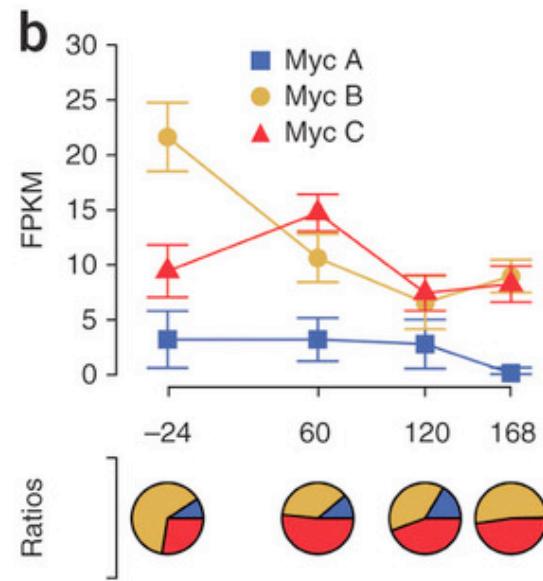
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell^{1–3}, Brian A Williams⁴, Geo Pertea², Ali Mortazavi⁴, Gordon Kwan⁴, Marijke J van Baren⁵, Steven L Salzberg^{1,2}, Barbara J Wold⁴ & Lior Pachter^{3,6,7}

a

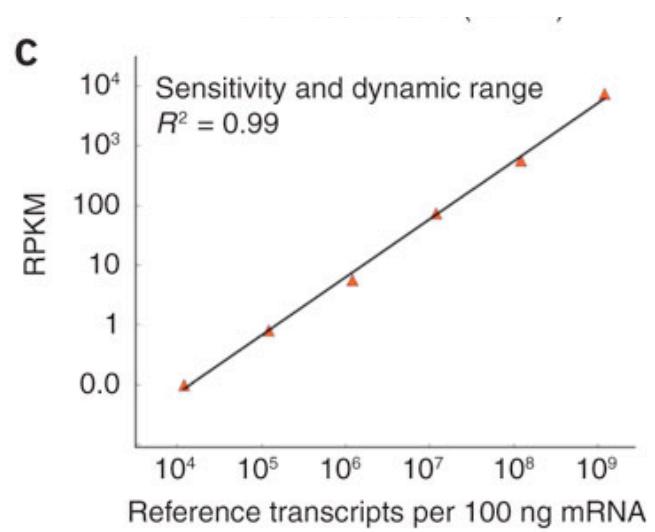


b

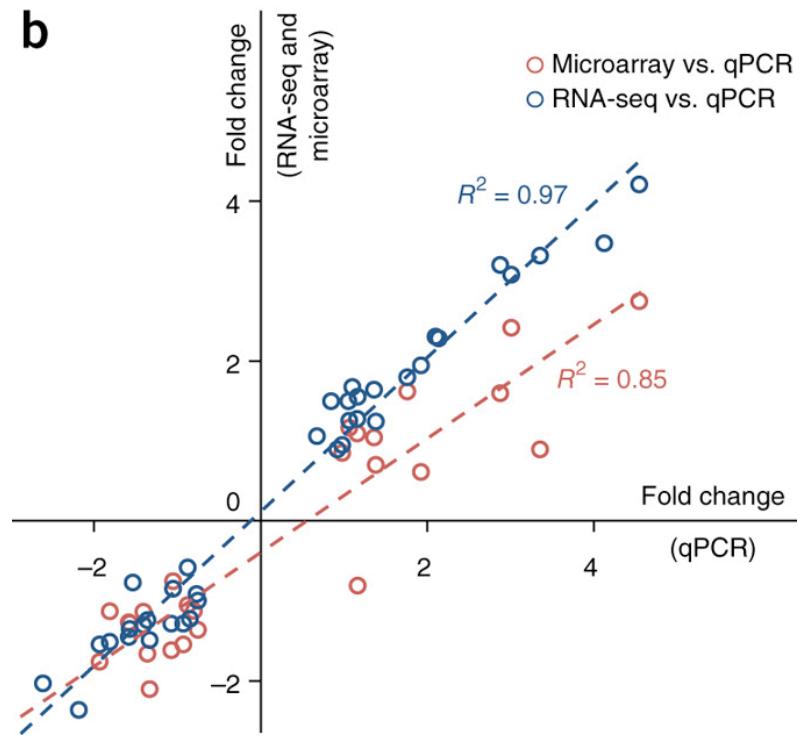
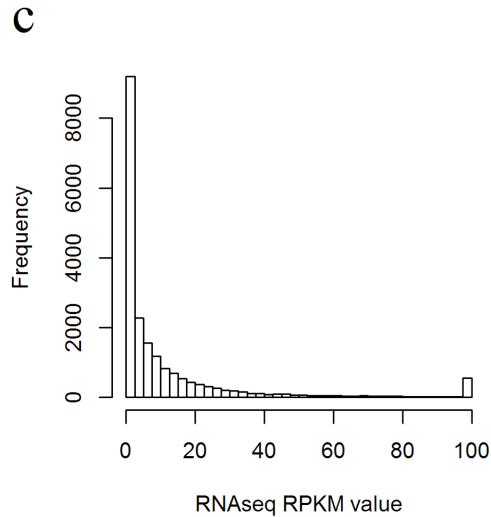
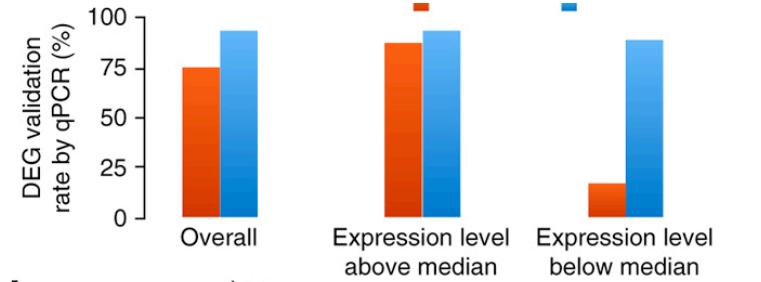
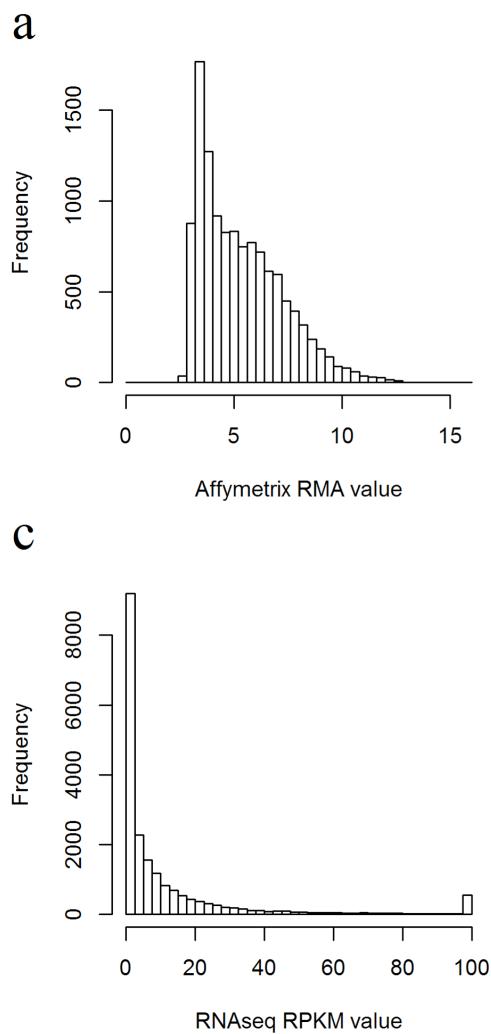


Differential Expression

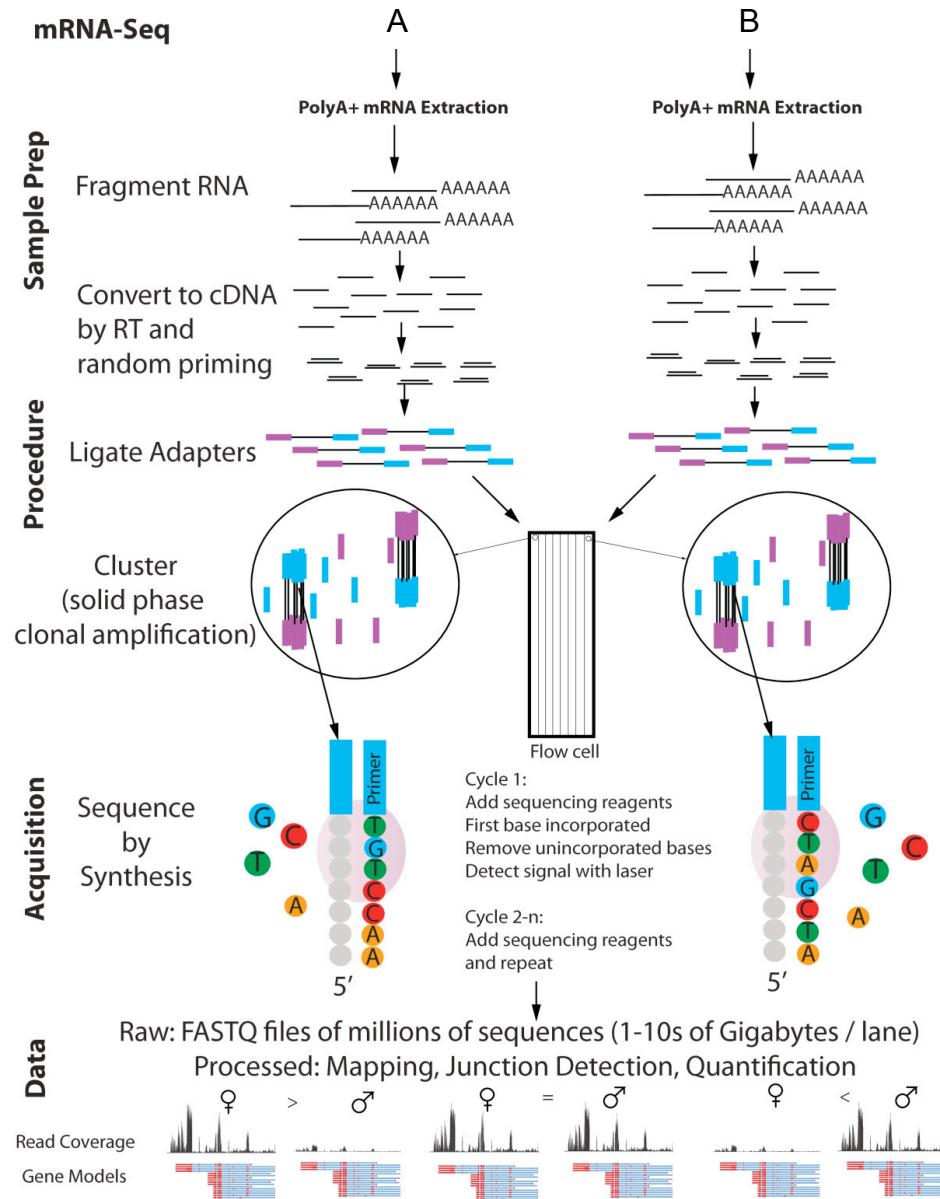
- Comparing feature abundance under different conditions
- Assumes linearity of signal over a range of expression levels
- When *feature=gene*, well-established pre- and post-analysis strategies exist



Range of detection

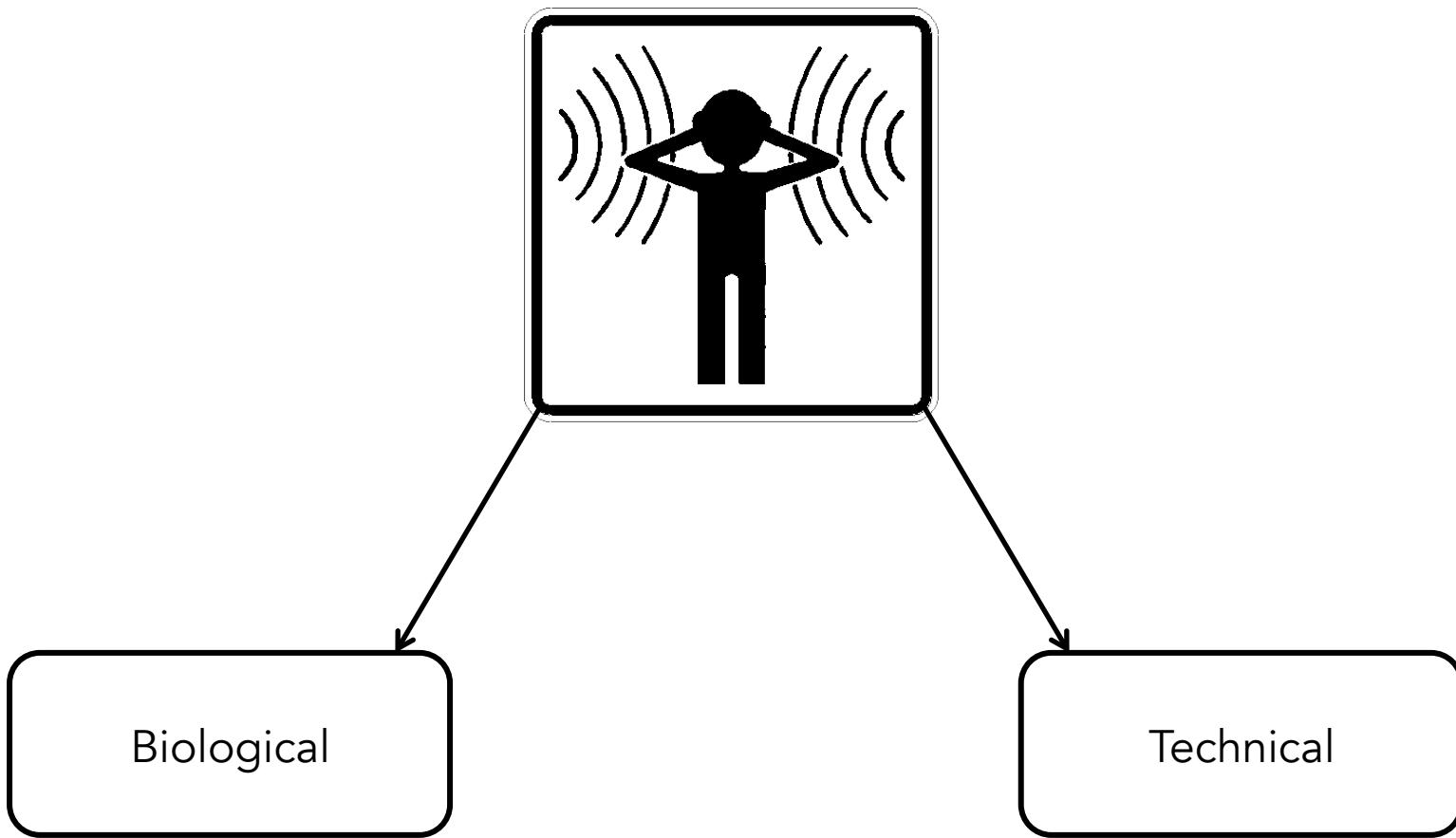


Library Prep i

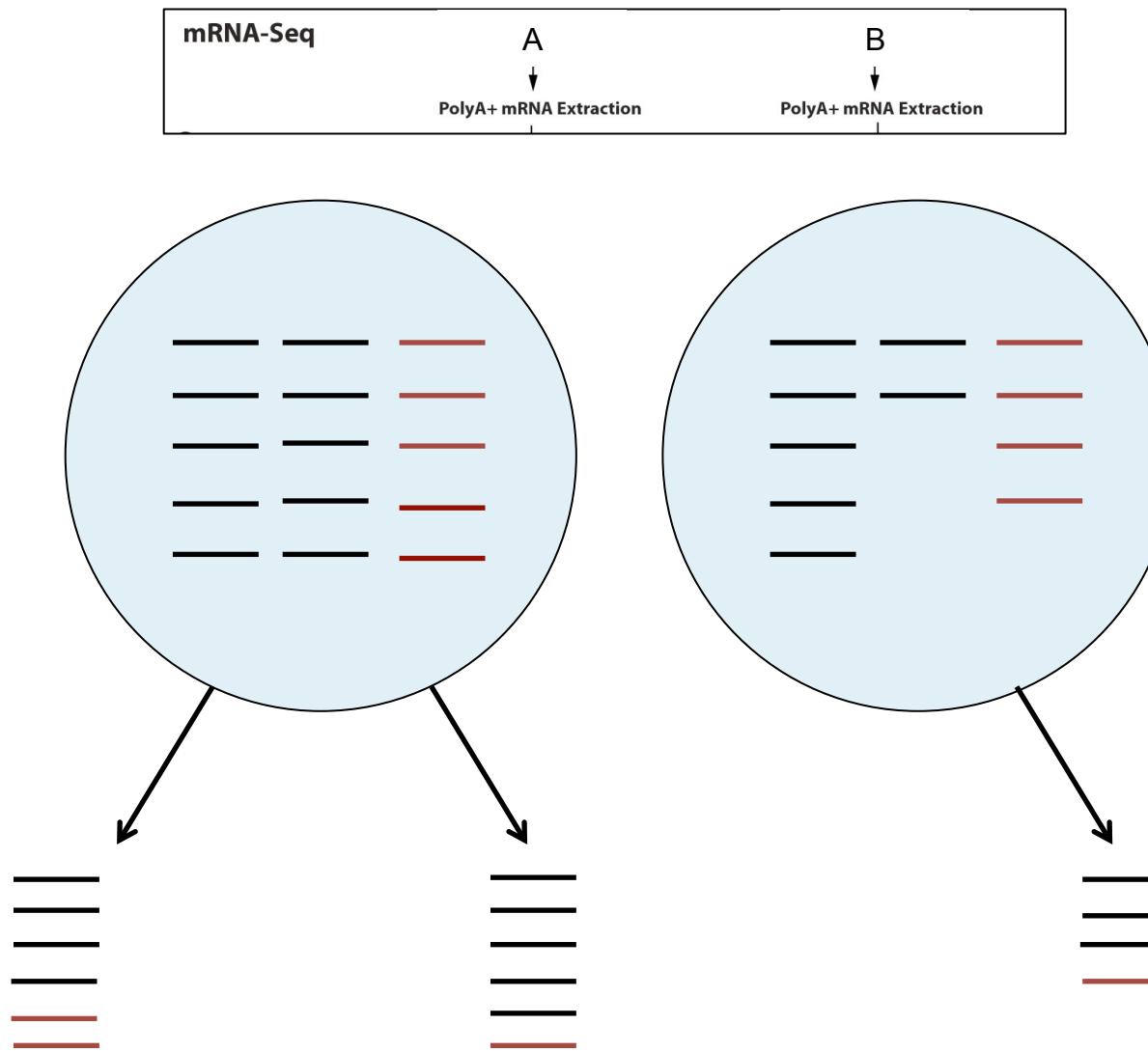


Malone, J.H. & Oliver, B.

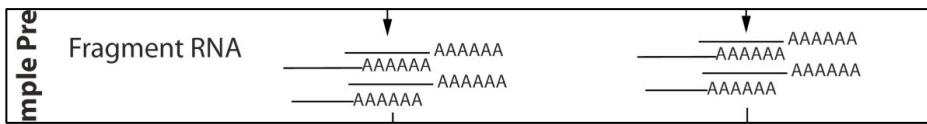
Library Prep ii



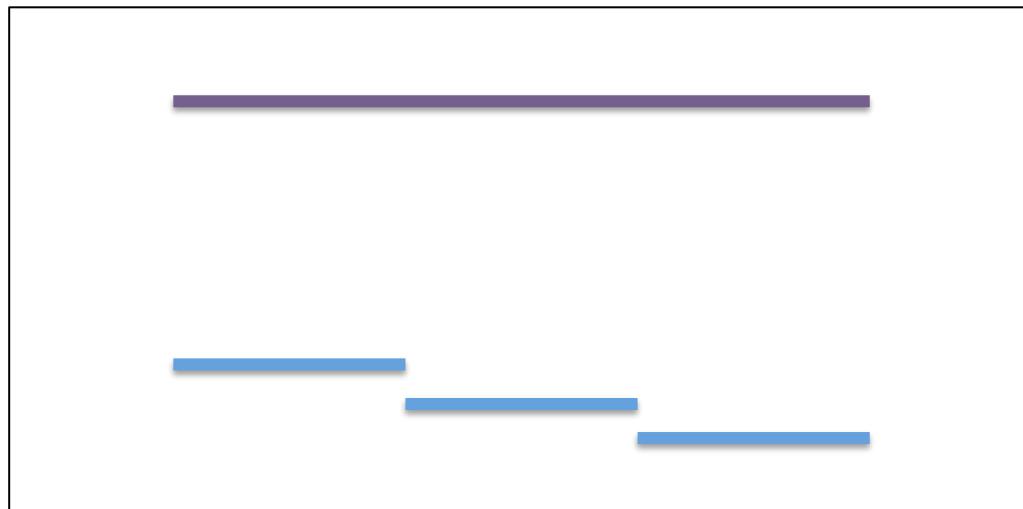
Library Prep iii



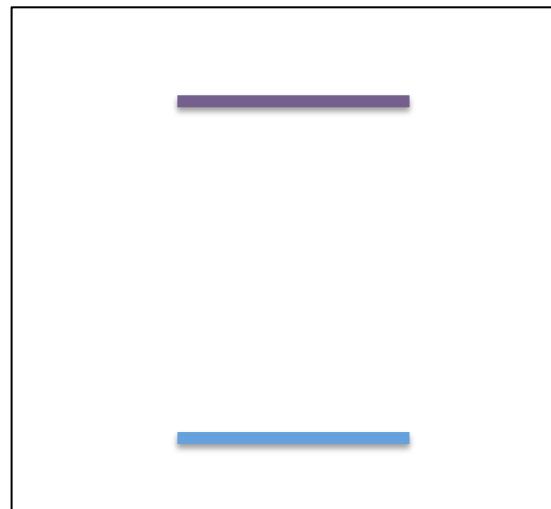
Library Prep iii



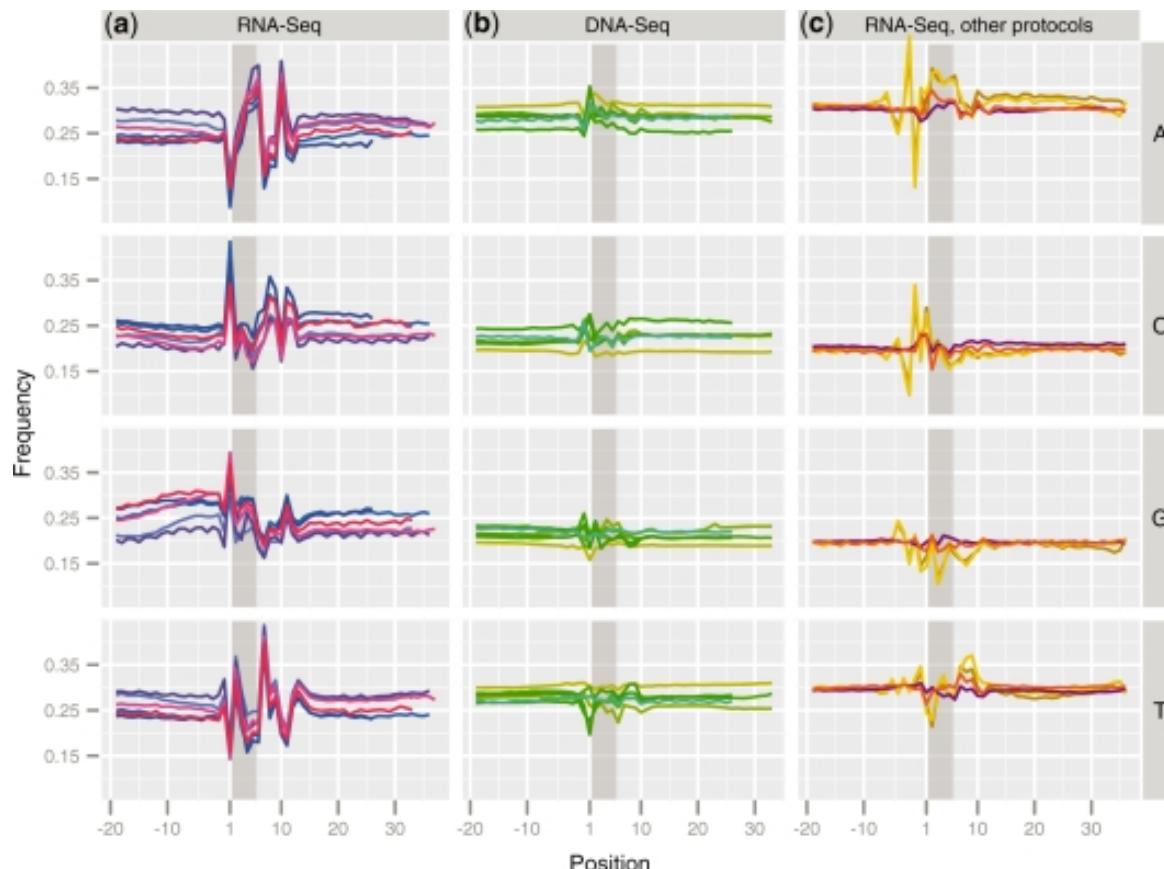
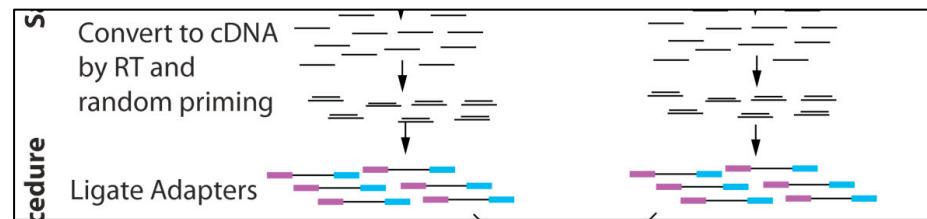
A



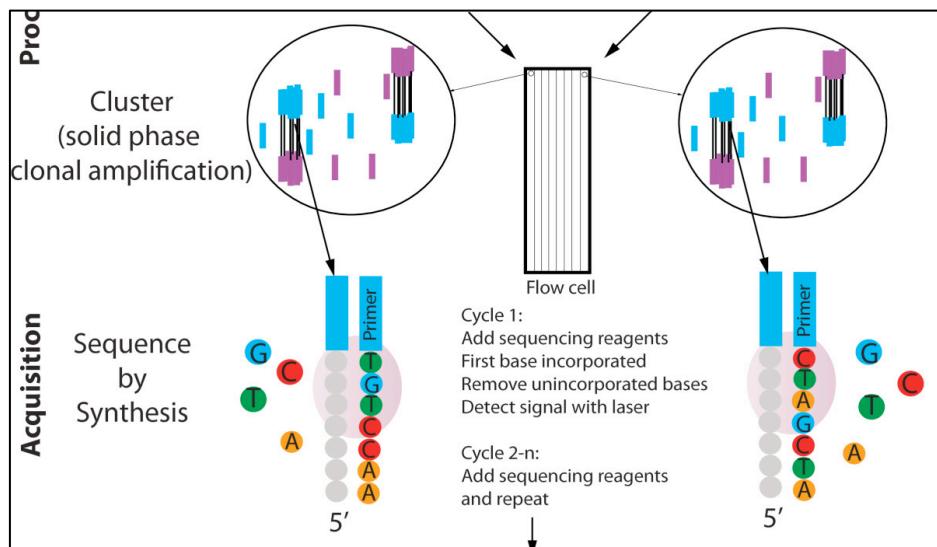
B



Library Prep iv

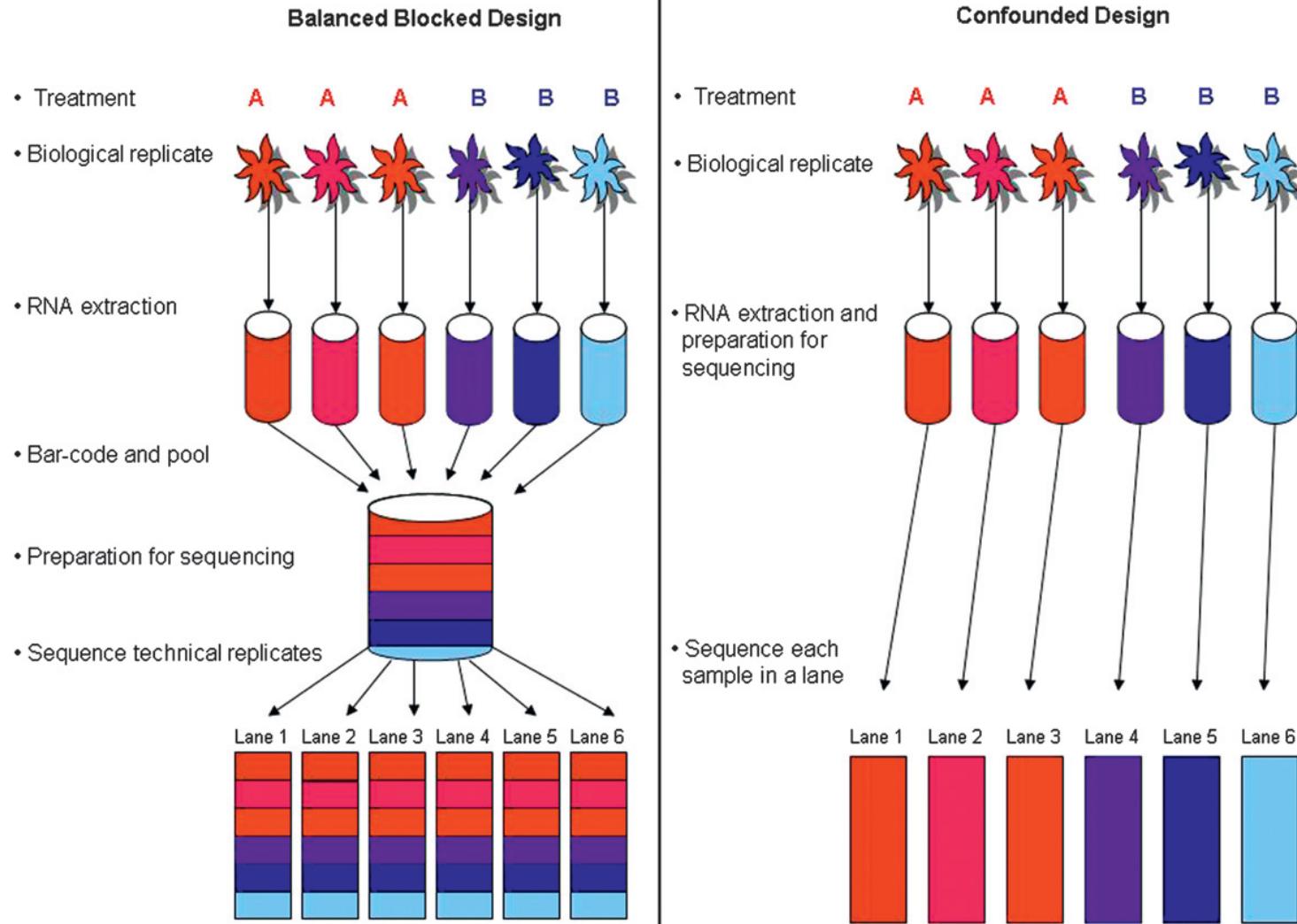


Library Prep v

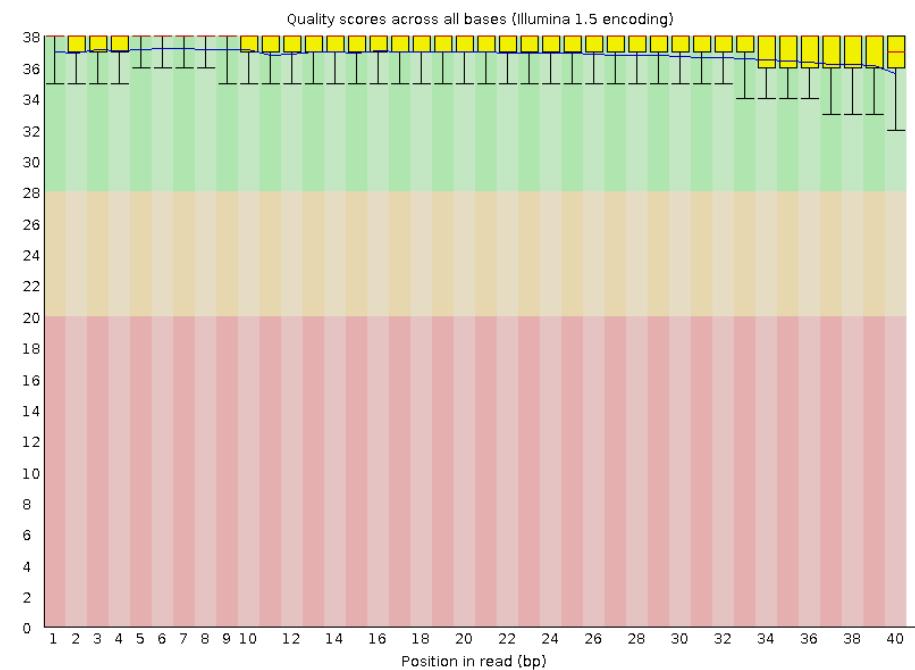
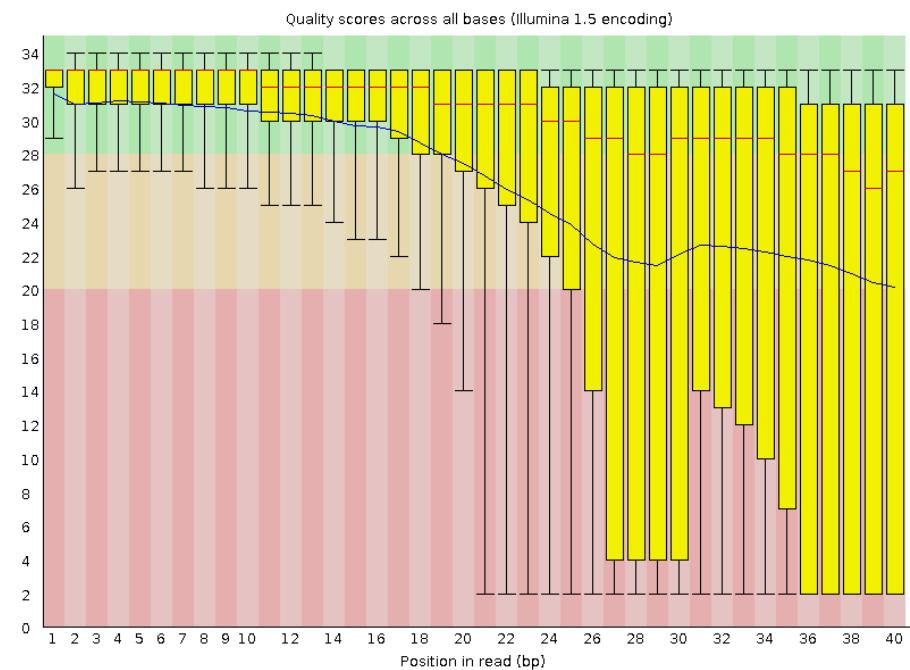


- Duplicates (optical & PCR)
- Sequence errors
- Indels
- Repetitive/problematic sequence

Hot off the sequencer...

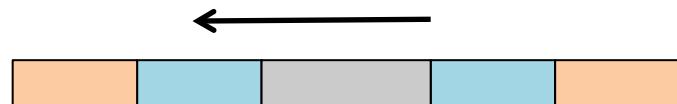
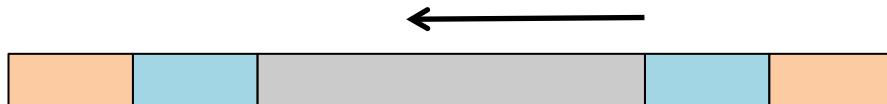


FASTQC

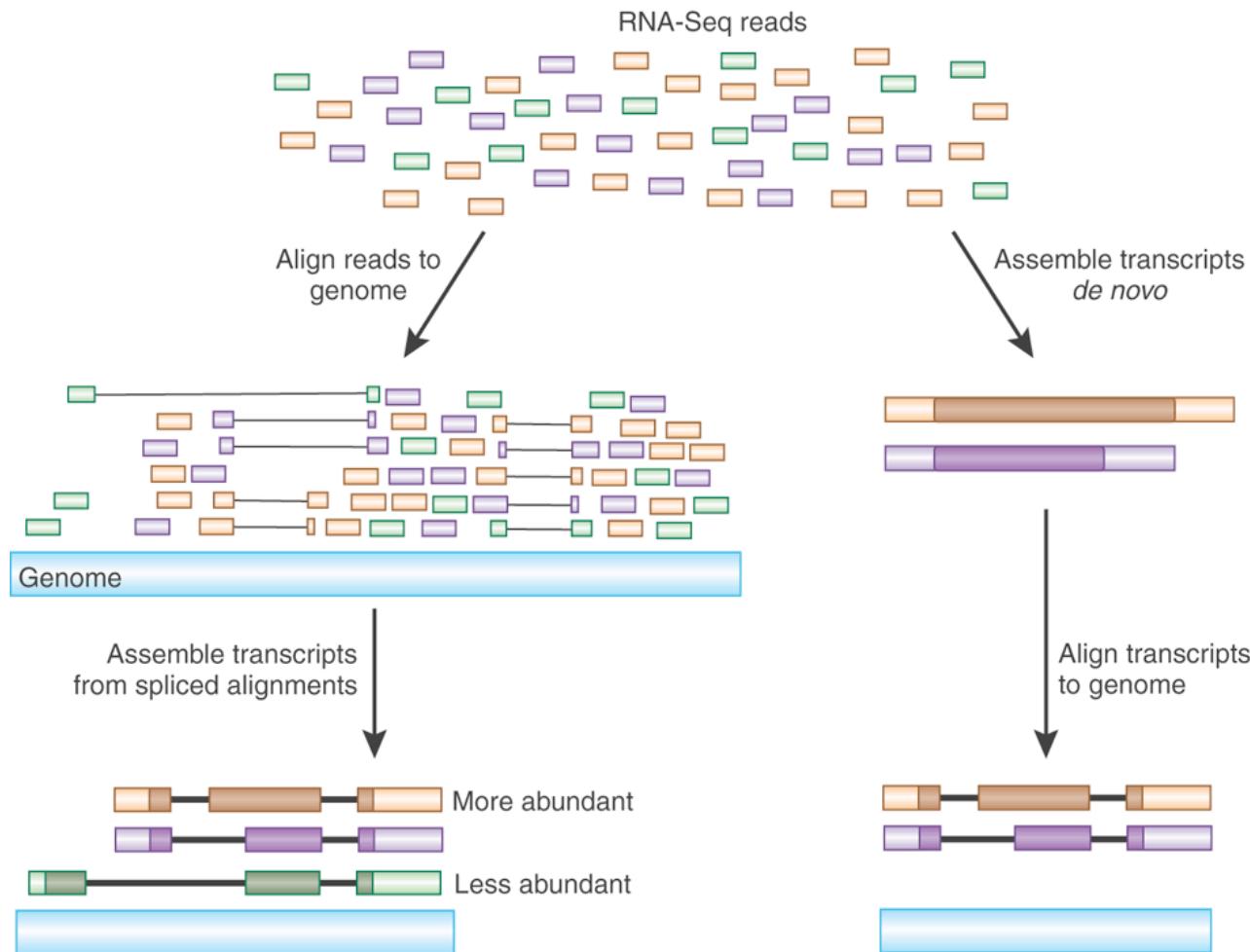


Trimming

- Quality-based trimming
- Adapter 'contamination'

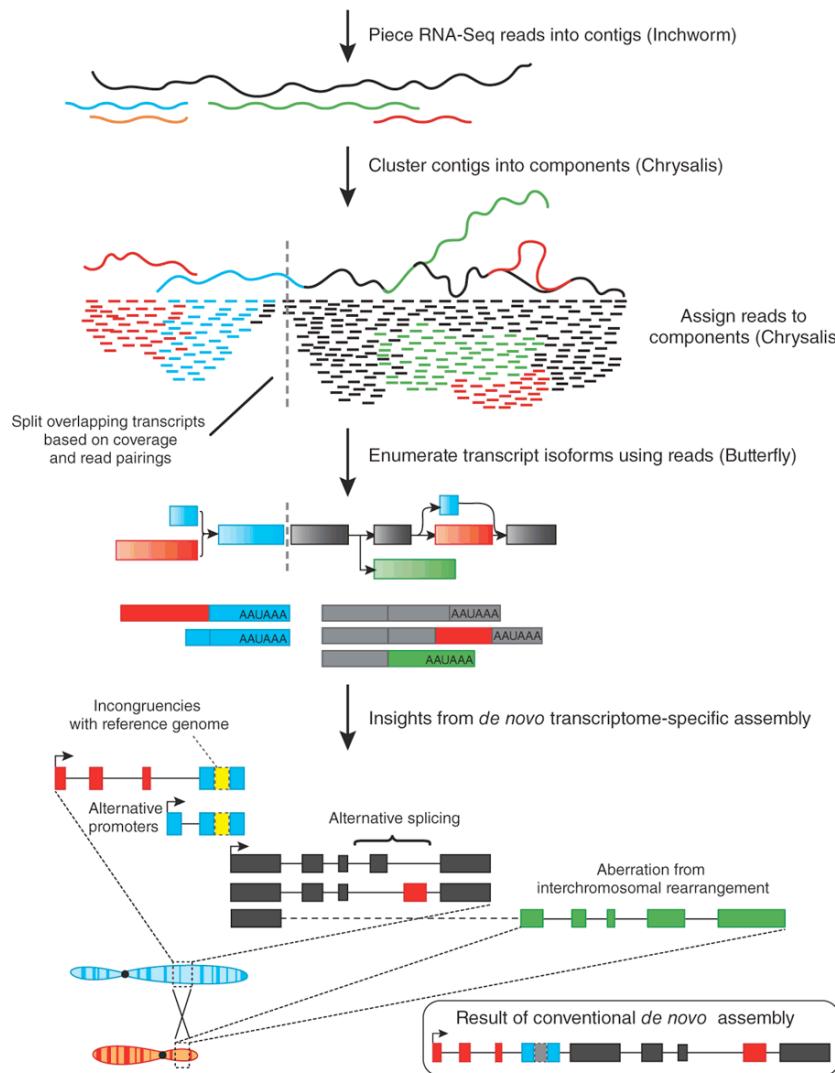


Sequence to sense

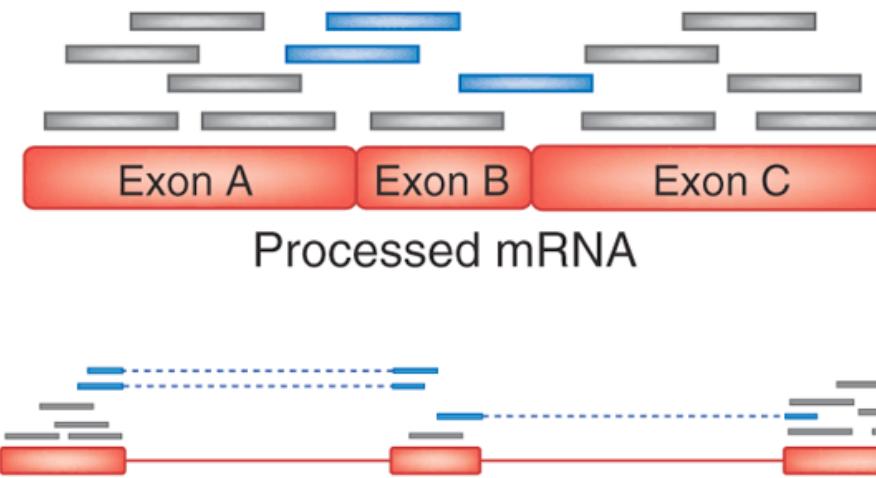


De novo assembly

- eg. Trinity



Reference-based assembly



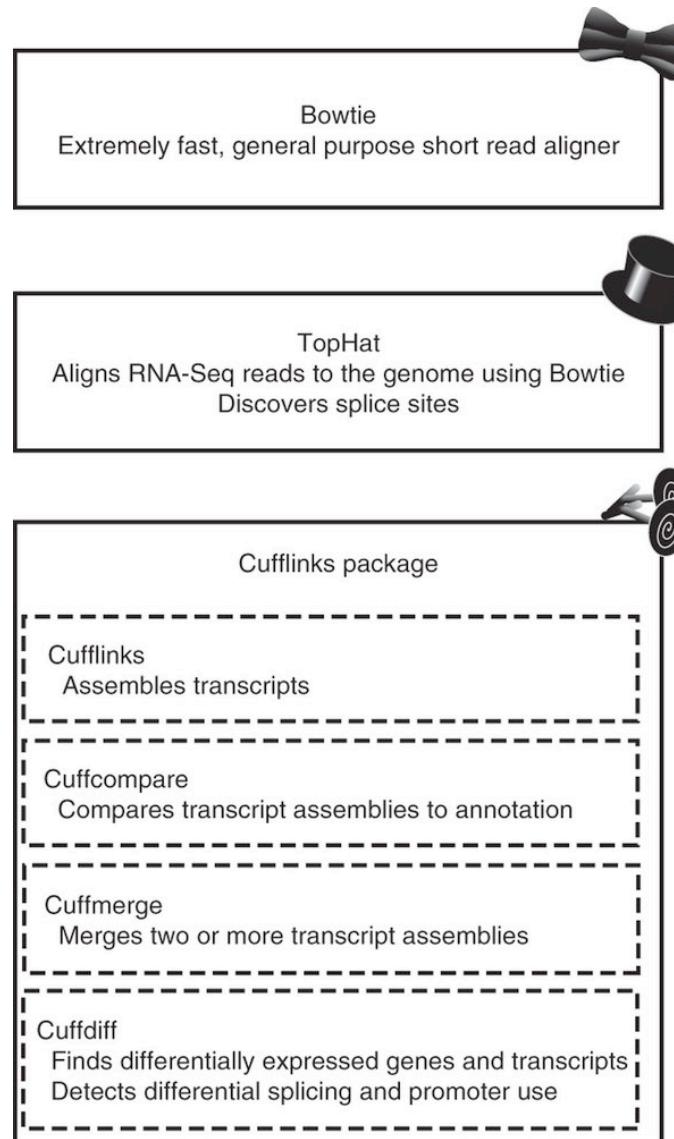
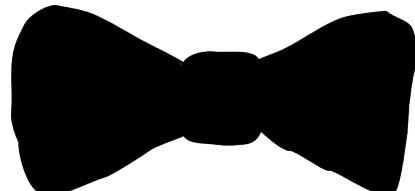
Genome mapping

- Can identify novel features
- Spice aware?
- Can be difficult to reconstruct isoform and gene structures

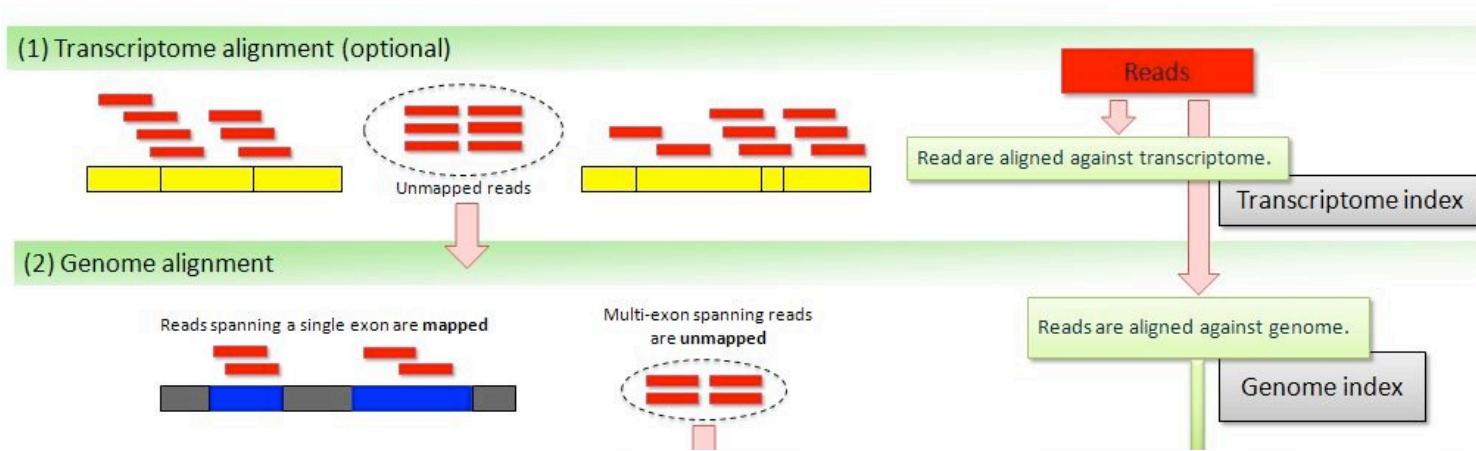
Transcriptome mapping

- No repetitive reference
- Overcomes issues of complex structures
- Novel features?
- How reliable is the transcriptome?

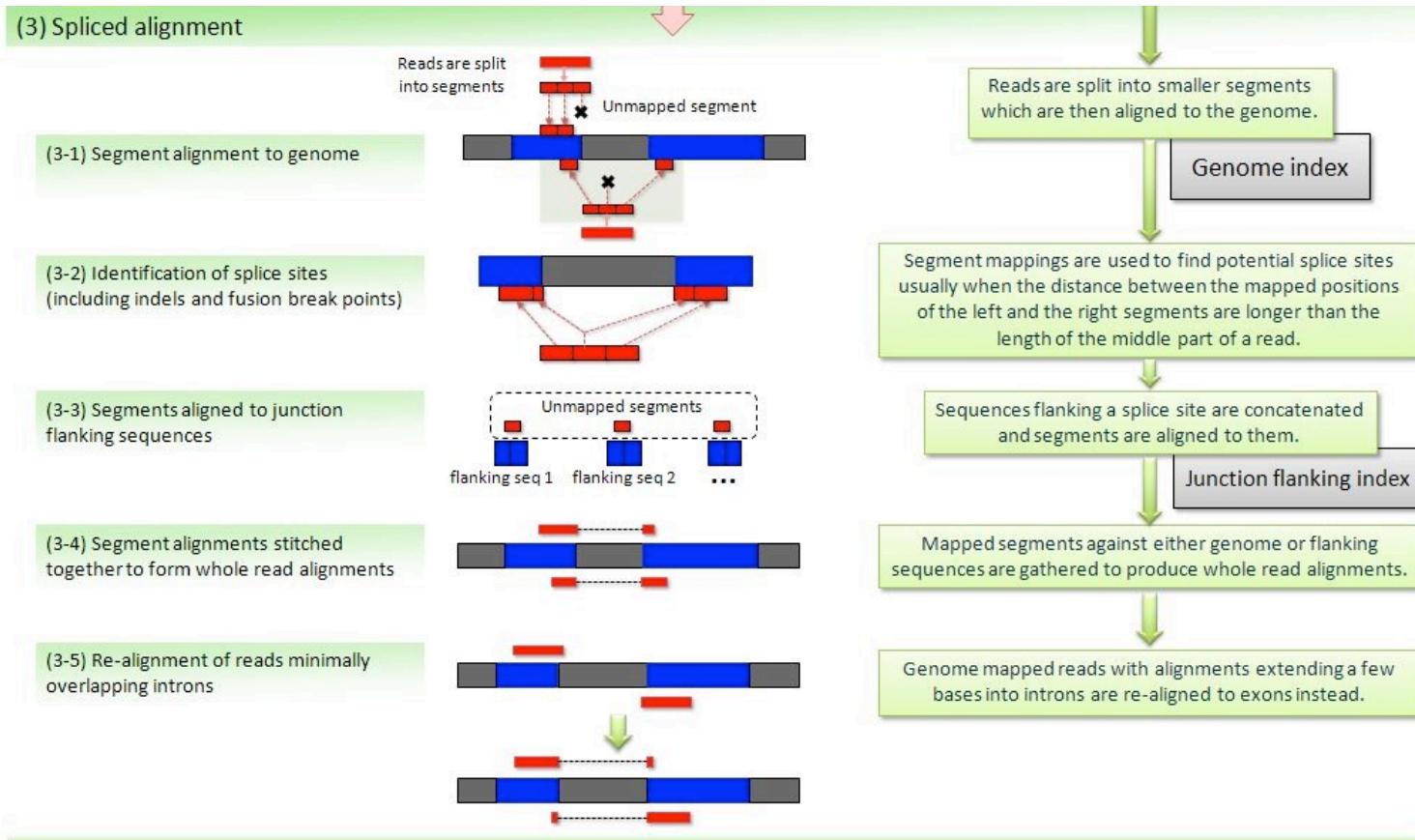
A smart suit(e)



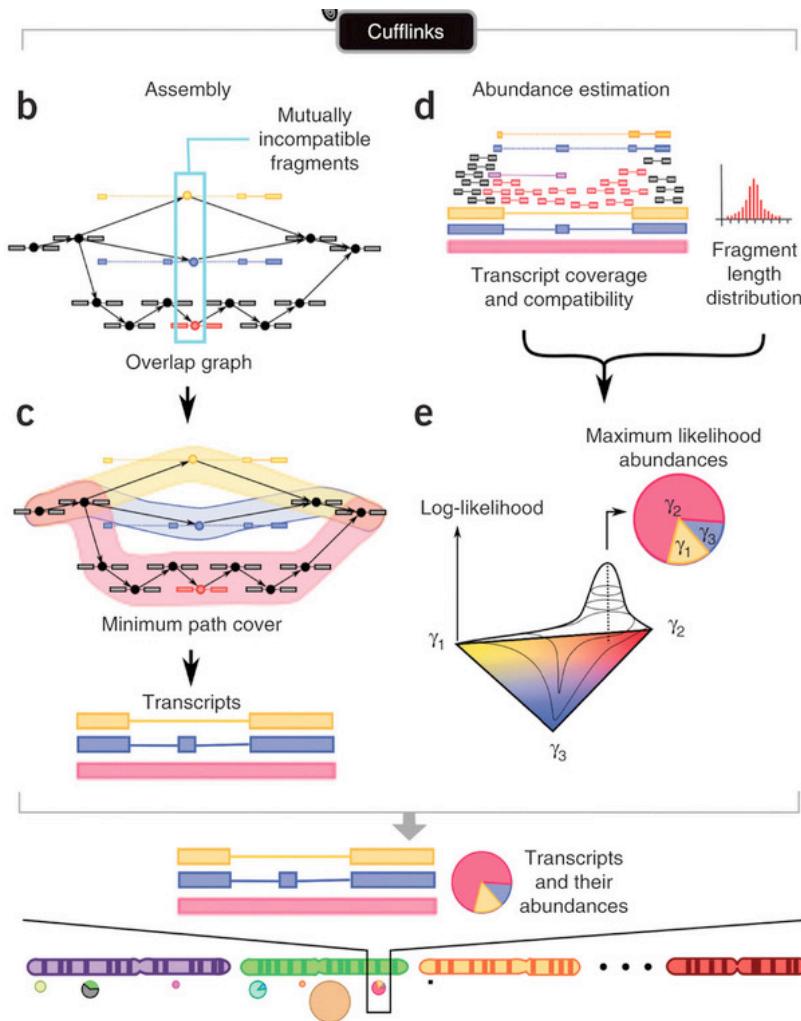
Tophat/Bowtie



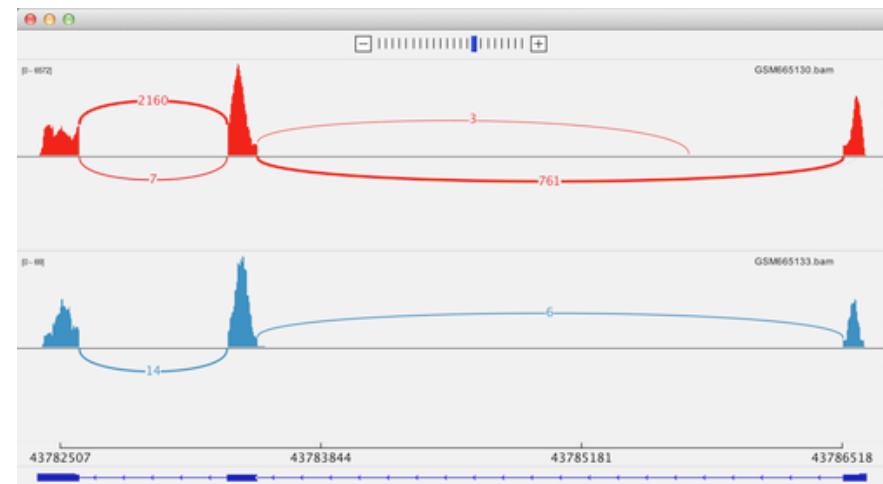
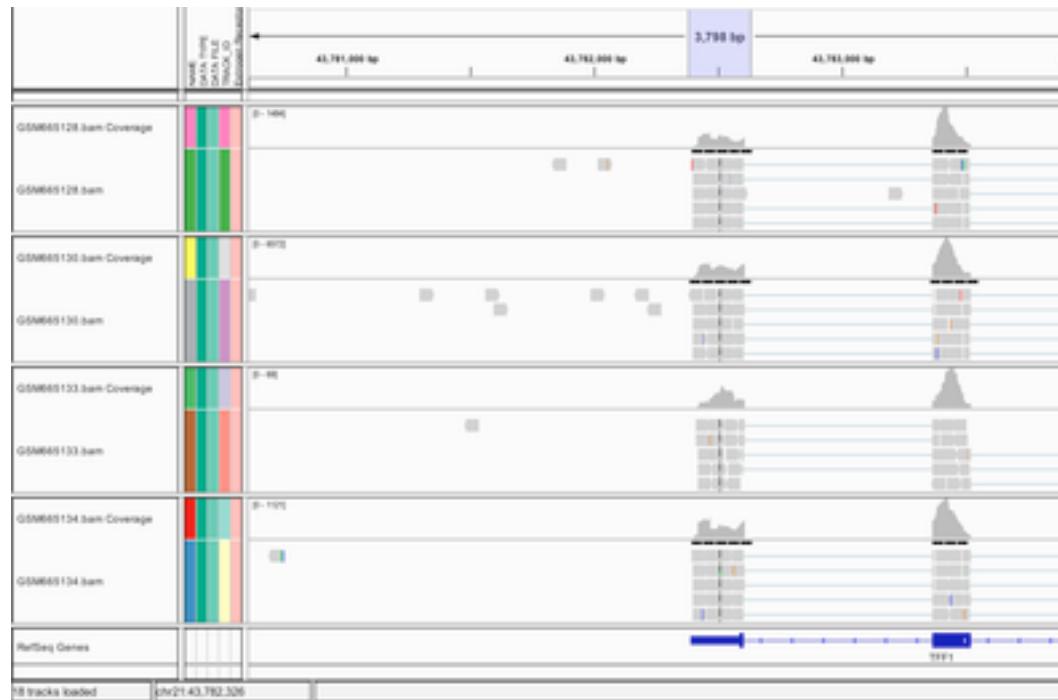
Tophat/Bowtie



Cufflinks



How do we look?



Duplicates & RNA-seq

Intrinsically lower complexity

Highly expressed genes

Model as part of counting process

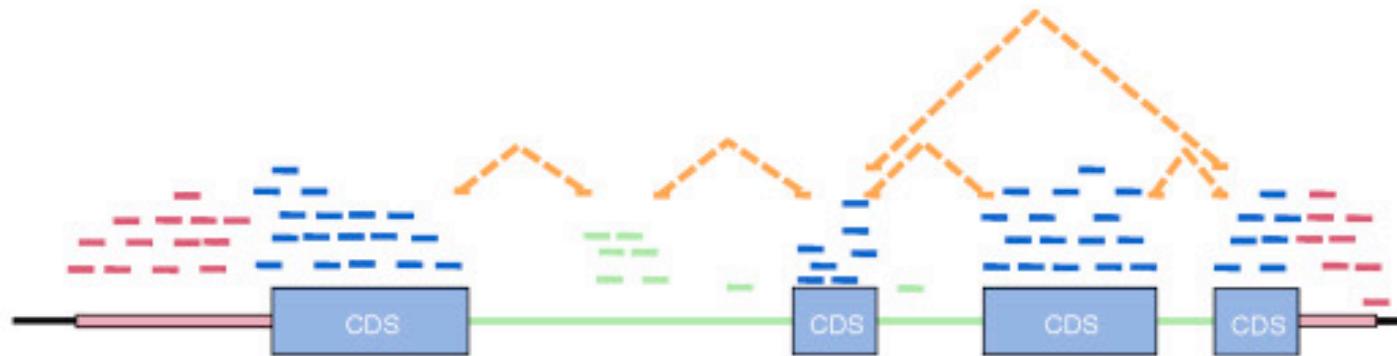
Variant calling vs DE analysis

Platform/pipeline

Single-end vs paired-end

Counting

(b)



Genome-based features

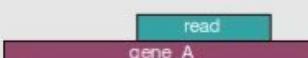
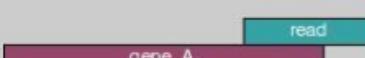
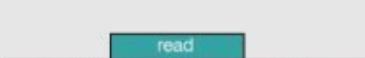
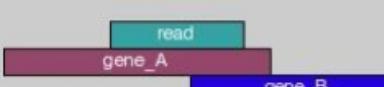
- Exon or gene boundaries?
- Isoform structures?
- Gene multireads?

Transcript-based features

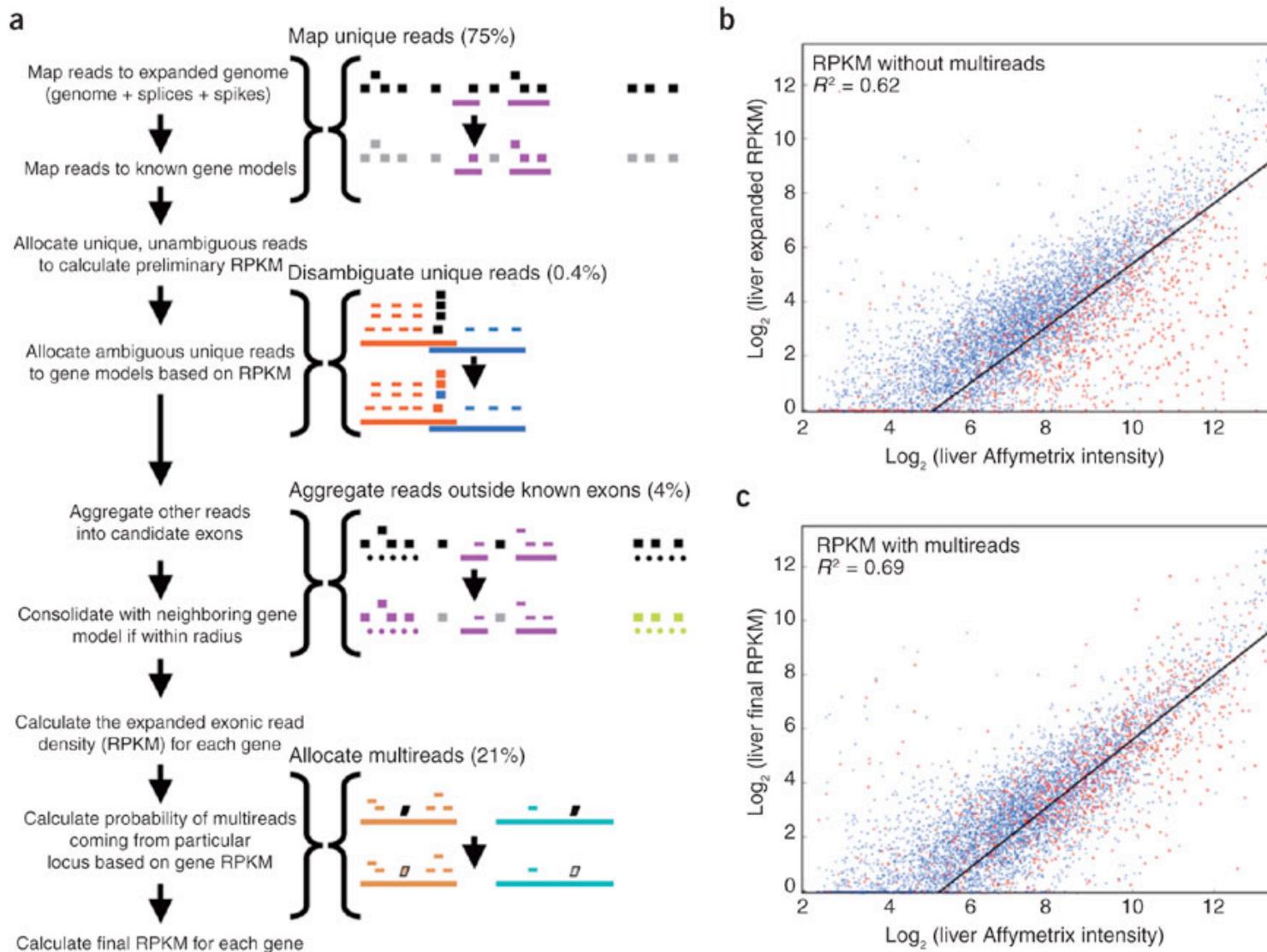
- Transcript assembly?
- Novel structures?
- Isoform multireads?

Counting

- eg. HTseq

	union	intersection _strict	intersection _nonempty
 A single read overlaps with gene_A.	gene_A	gene_A	gene_A
 A single read overlaps with gene_A, but the read starts after the gene ends.	gene_A	no_feature	gene_A
 A single read overlaps with gene_A, but the read ends before the gene ends.	gene_A	no_feature	gene_A
 Two reads overlap with gene_A, but they are separated by a gap.	gene_A	gene_A	gene_A
 A read overlaps with both gene_A and gene_B.	gene_A	gene_A	gene_A
 A read overlaps with both gene_A and gene_B, but the read starts after gene_A ends.	ambiguous	gene_A	gene_A
 A read overlaps with both gene_A and gene_B, but the read ends before gene_B starts.	ambiguous	ambiguous	ambiguous

Counting



Counting & normalisation

- An estimate for the *relative* counts for each gene is obtained
- Assumed that this estimate is representative of the original population

Library size

- Sequencing depth varies between samples

Gene Properties

- GC content, length, sequence

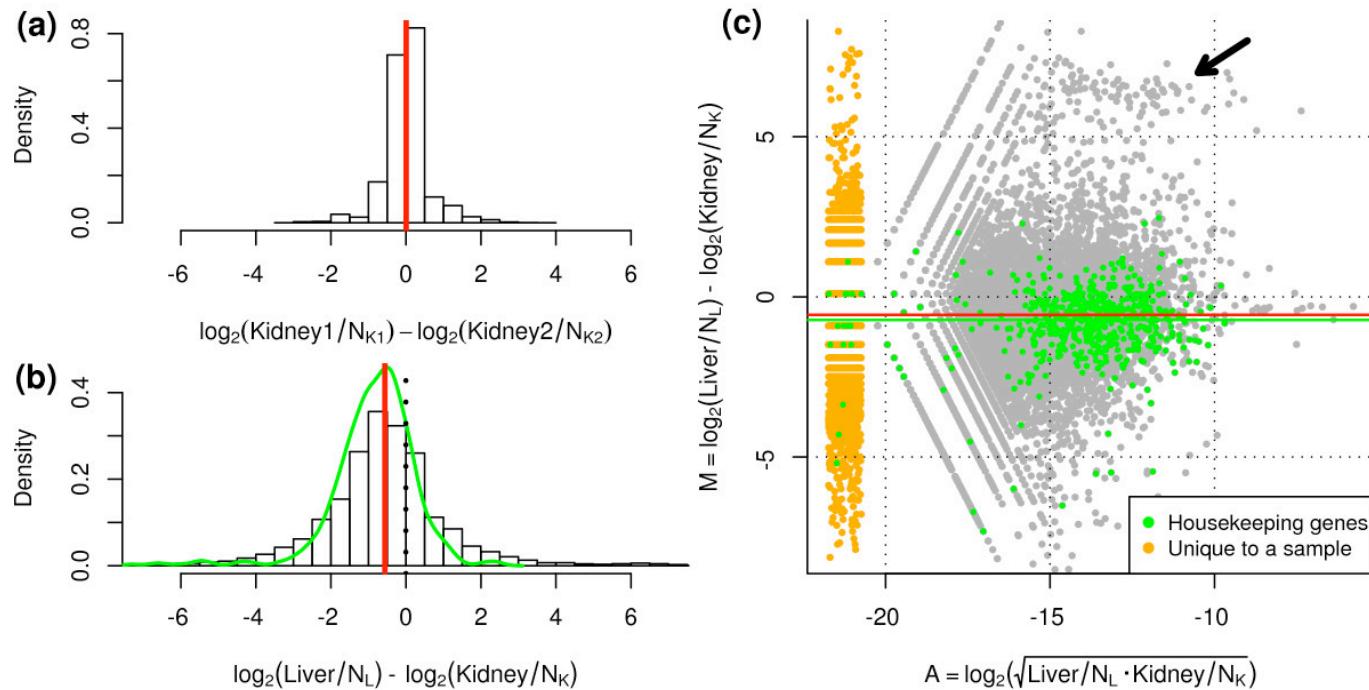
Library composition

- Highly expressed genes overrepresented at cost of lowly expressed genes

Normalisation i

Total Count

- Normalise each sample by total number of reads sequenced.
- Can also use another statistic similar to total count; eg. median, upper quartile

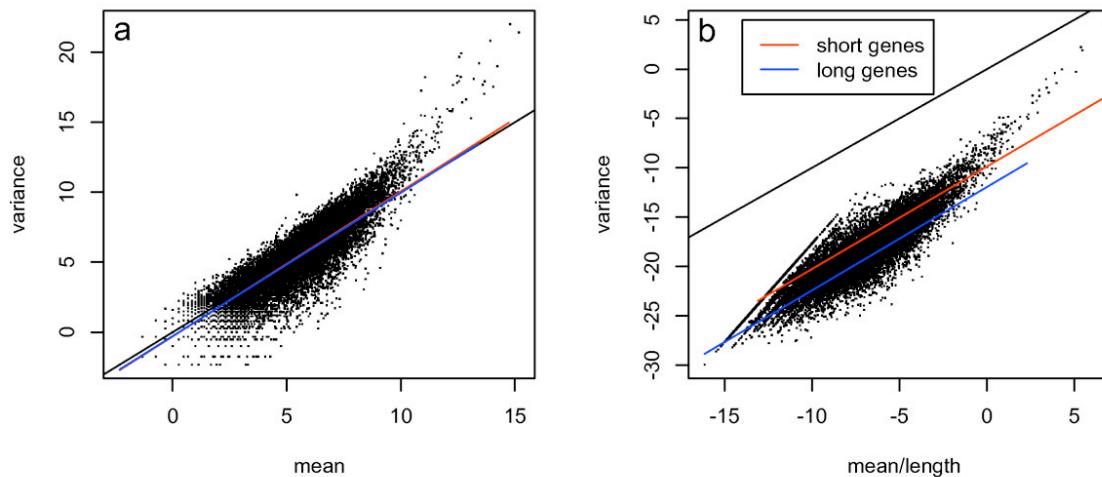


Normalisation ii

RPKM

- Reads per kilobase per million =

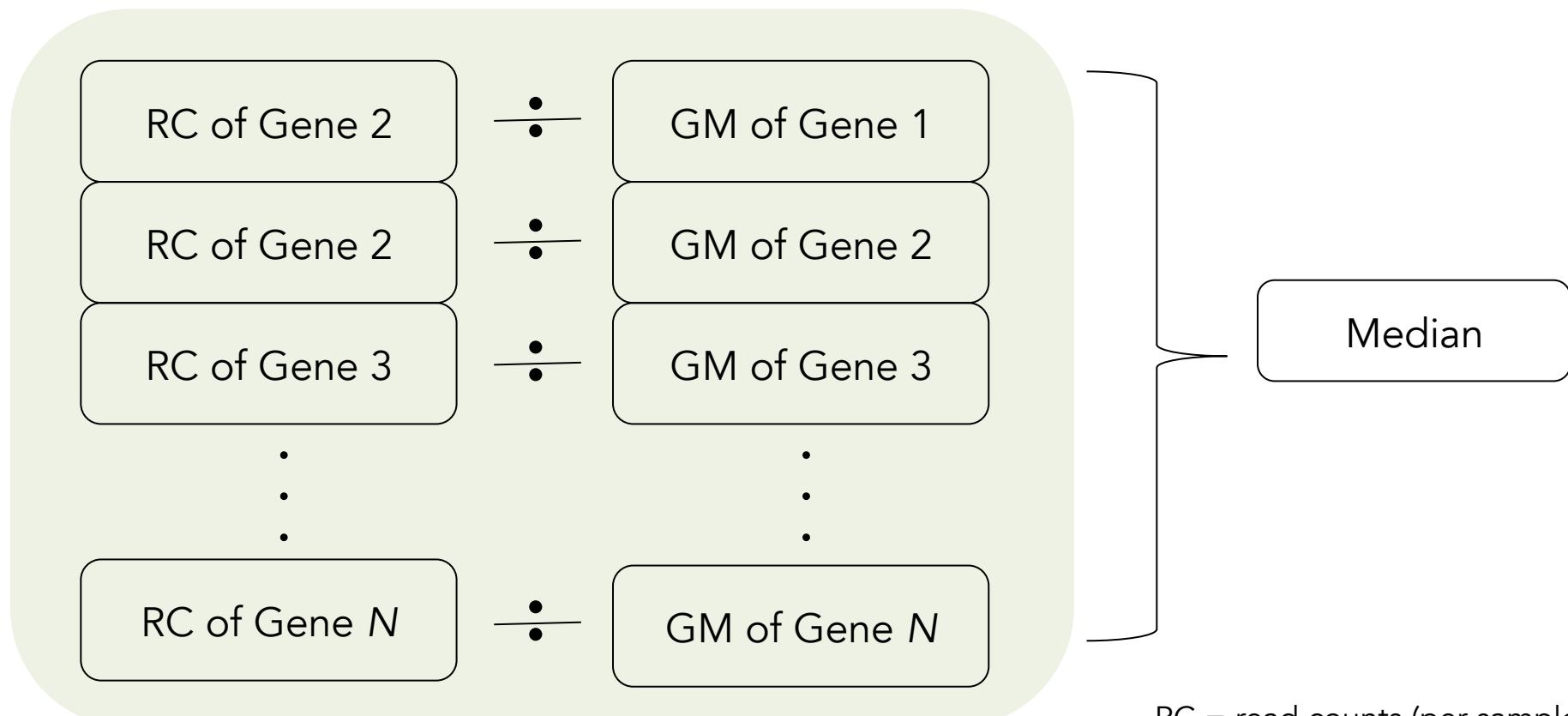
$$\frac{\text{reads for gene A}}{\text{length of gene A} \times \text{Total number of reads}}$$



Normalisation iii

Geometric scaling factor

- Implemented in DESeq
- Assumes that most genes are not differentially expressed



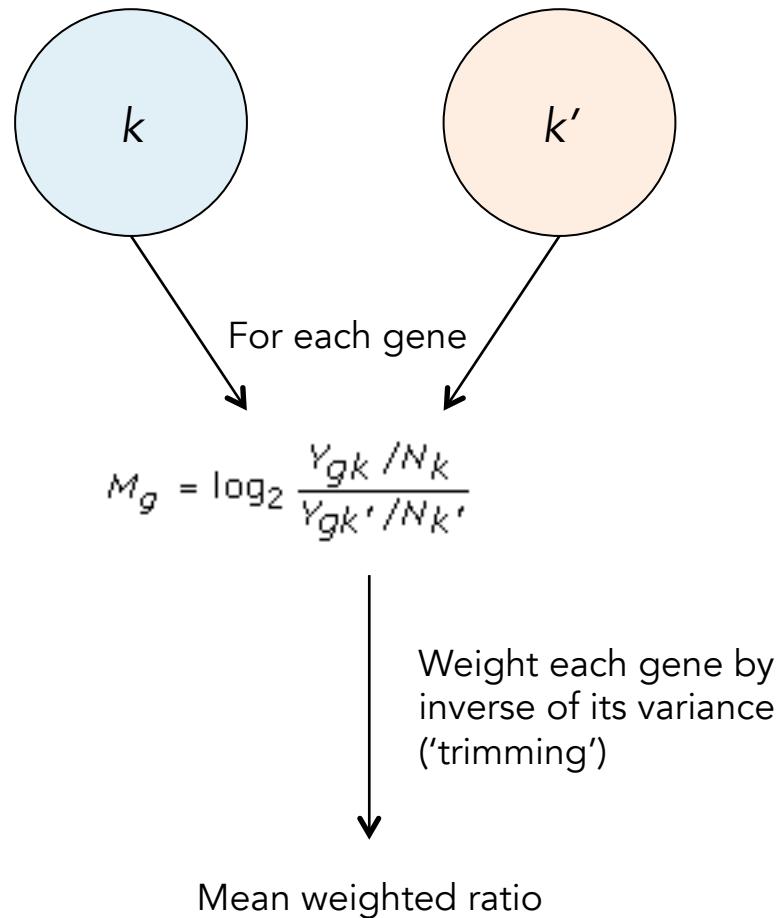
RC = read counts (per sample)

GM =geometric mean (all samples)

Normalisation iv

Trimmed mean of M

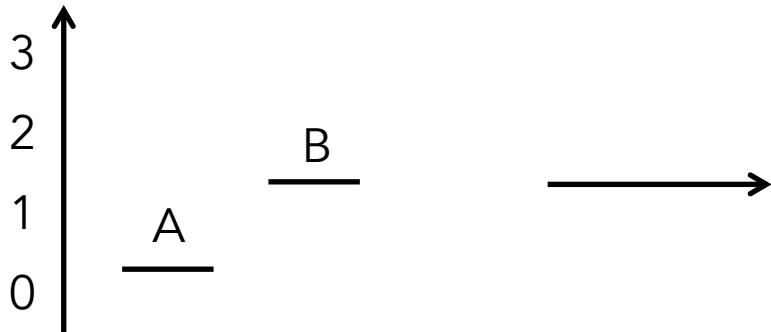
- Implemented in edgeR
- Assumes most genes are not differentially expressed



g = each gene

Differential expression

- Simple



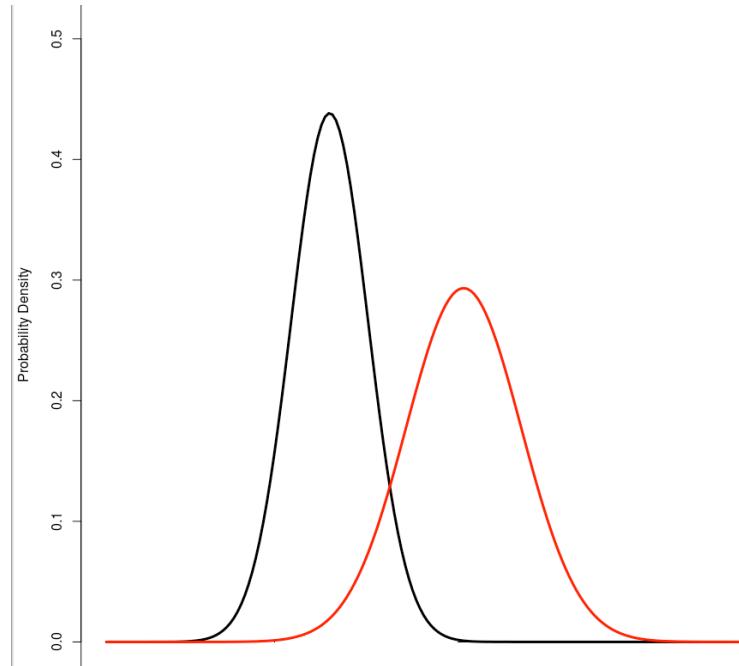
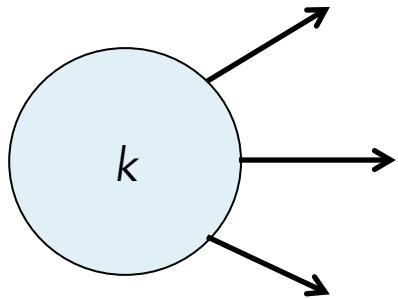
	Cond A	Cond B
Gene X		
Other		

All we need

- Know what the data looks like
- Some measure of difference

Modelling – old trends

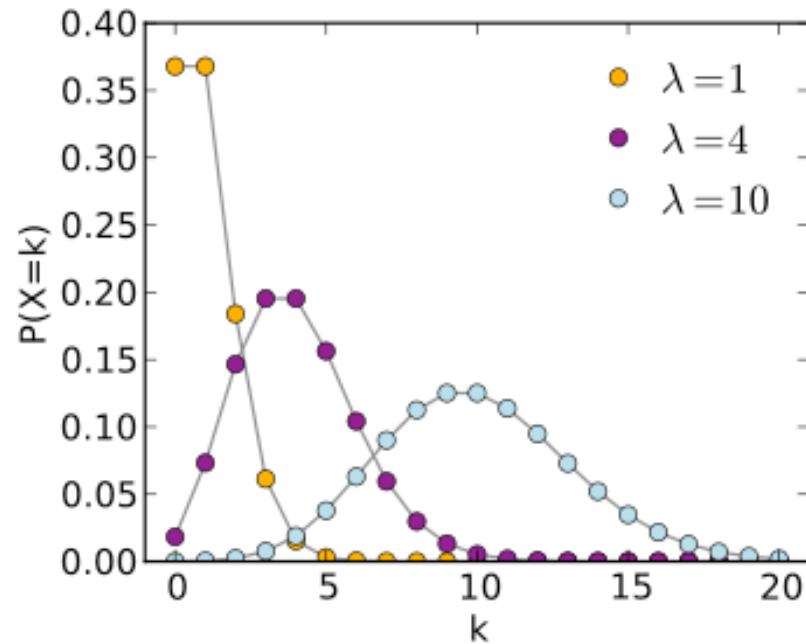
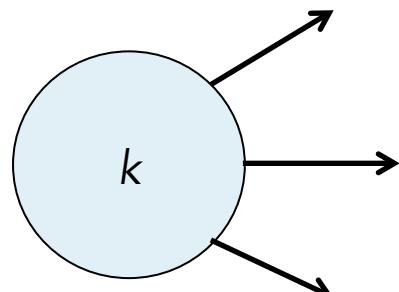
- Technical replicates introduce some variance



- What the data looks like: **normal distribution**
- Some measure of difference: **t-test etc**

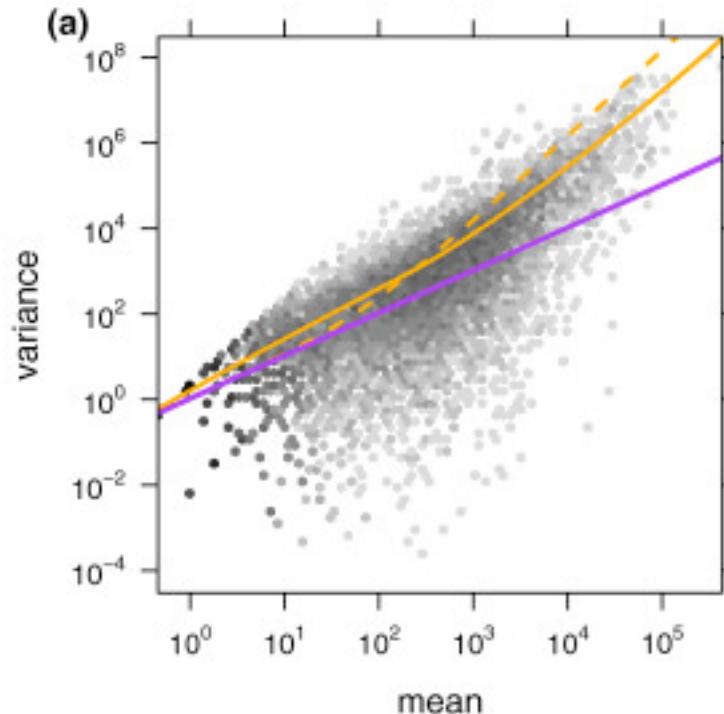
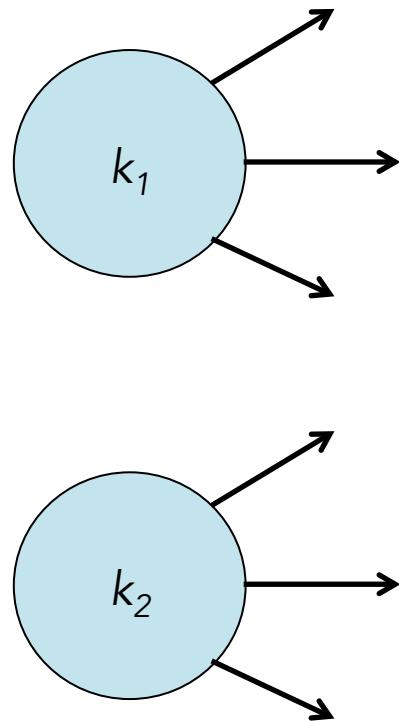
Modelling – in fashion

- Use the Poisson distribution for count data from technical replicates
- Just one parameter required – the mean



Modelling – in fashion

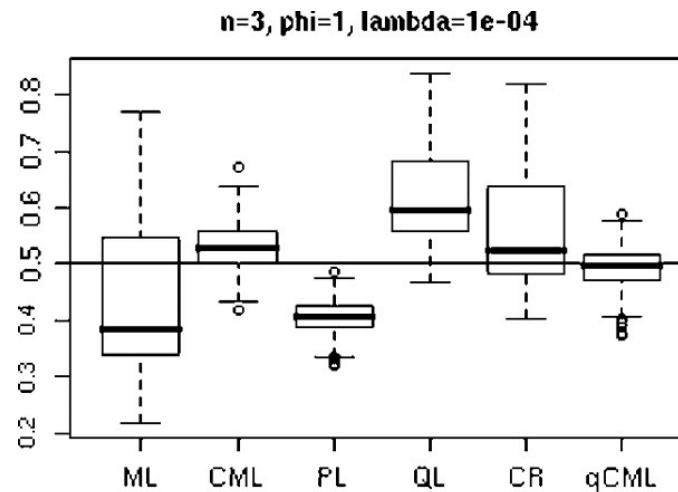
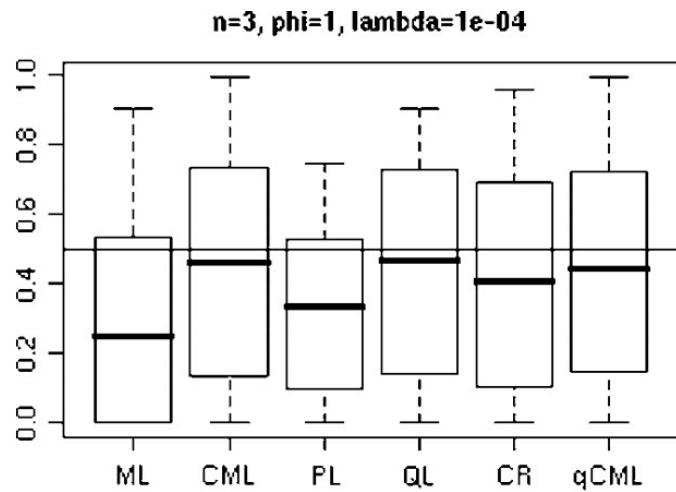
- Biology is never that simple...



- The negative binomial distribution represents an *overdispersed* Poisson distribution, and has parameters for both the mean and the overdispersion.

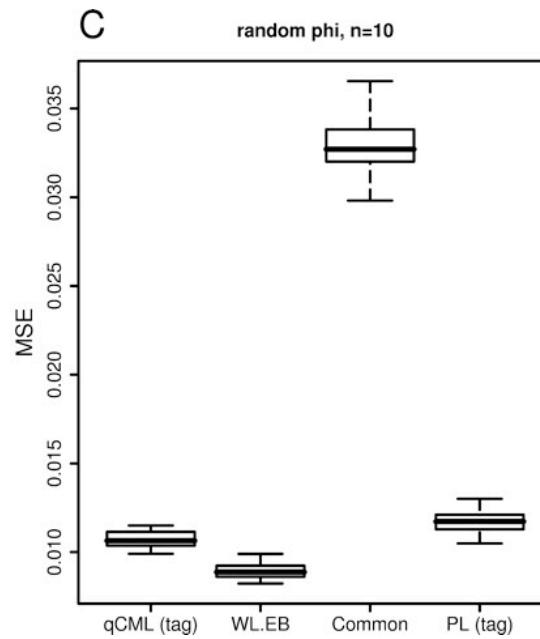
Modelling – in fashion

- Estimating the dispersion parameter can be difficult with a small number of samples
- 'Share' information from all genes to obtain global estimate



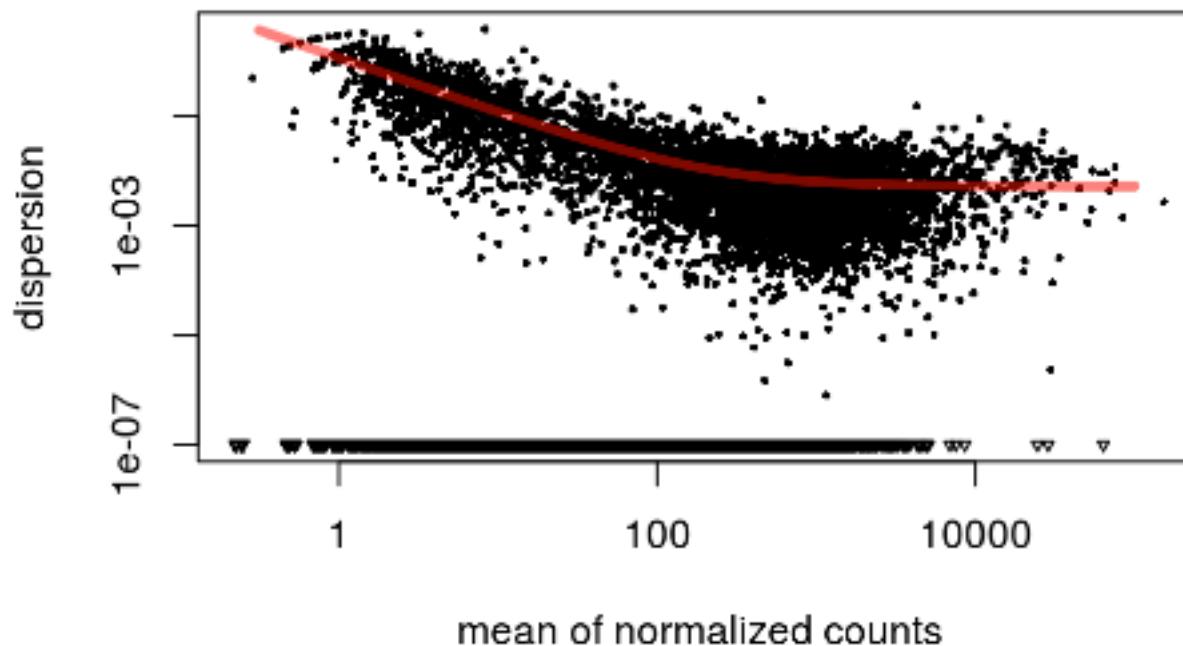
Shrinkage

- Genes do not share a common dispersion parameter
- 'Moderated' estimate – assign a per-gene weight to the combined estimate



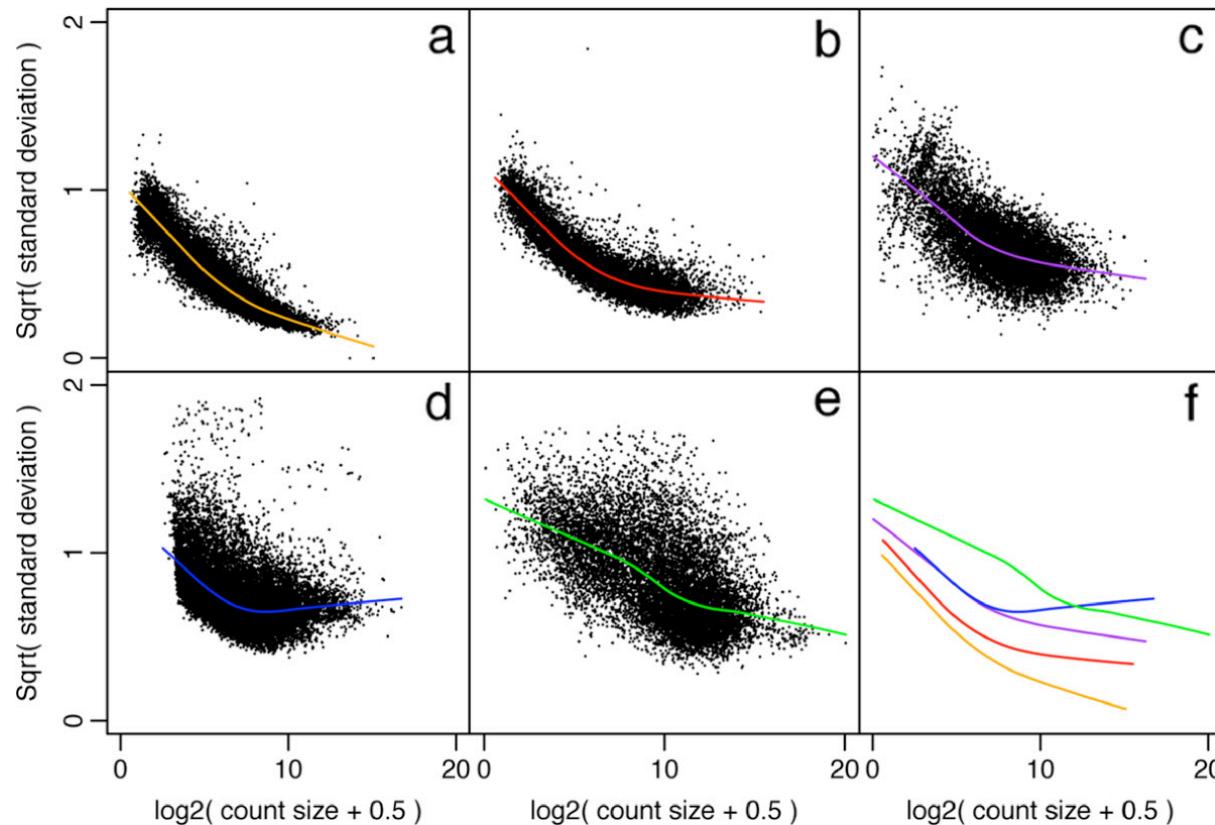
DESeq

- DESeq fits a mean/dispersion relationship model
- Shifts individual estimates to regression line



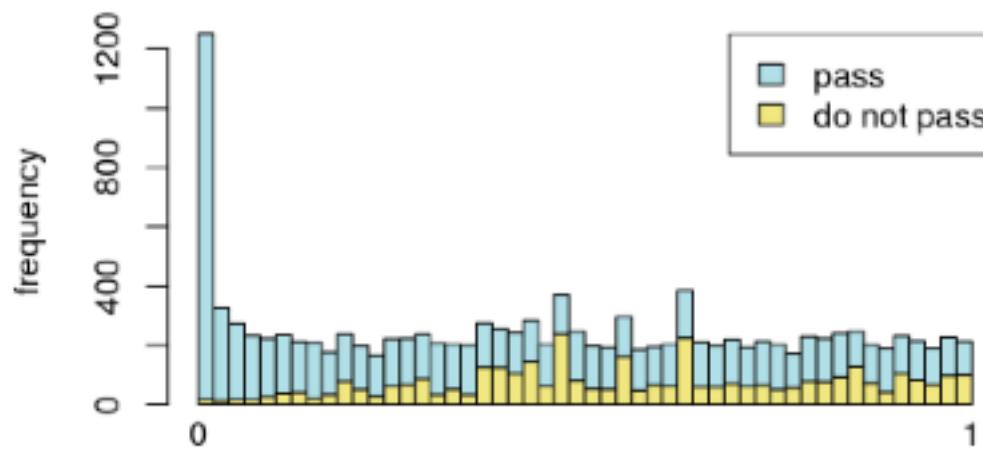
The mean-variance relationship

- Variance = Technical (variable) + Biological (constant)
- A=technical replicates ---> E =(very) biologically different replicates

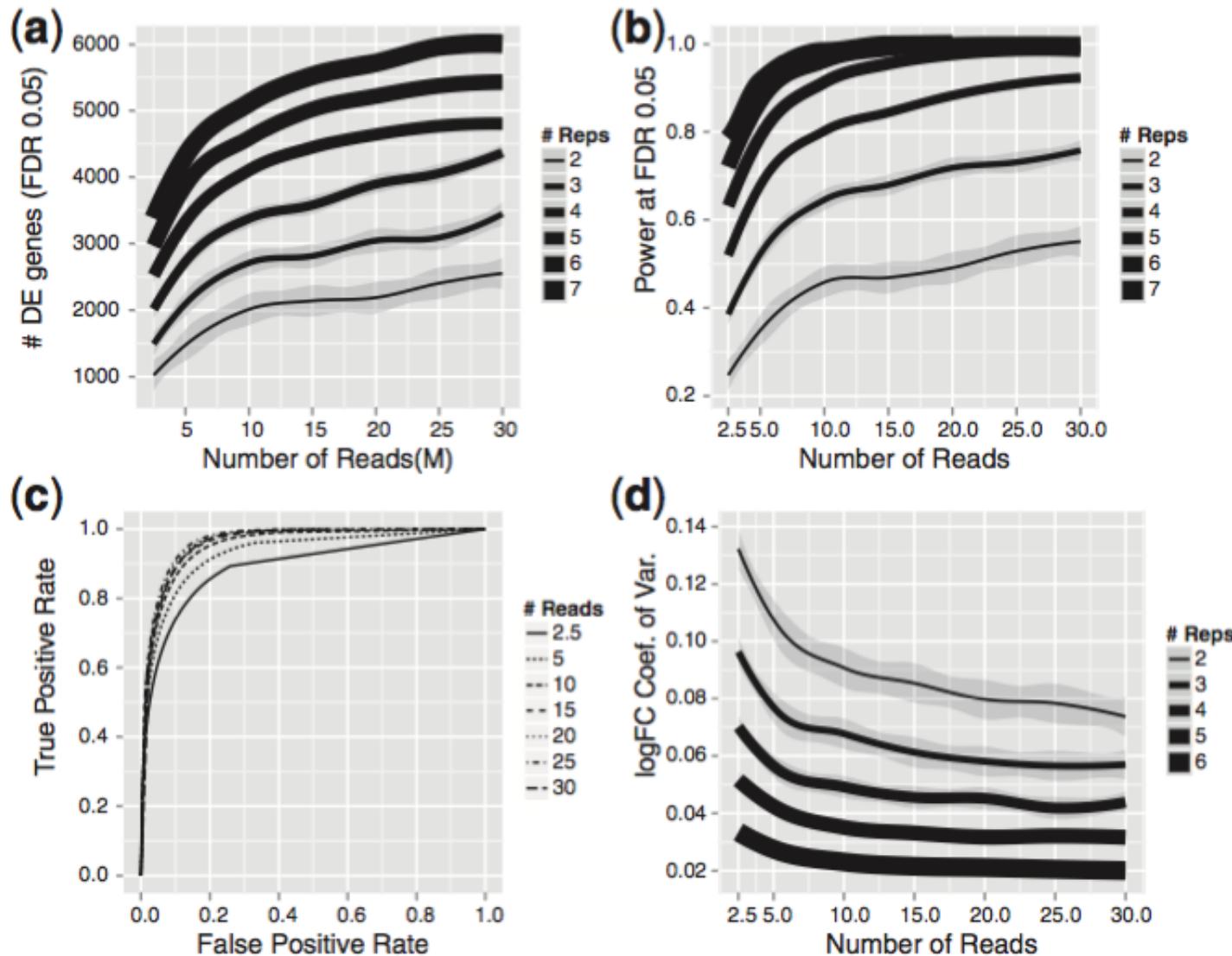


Filtering

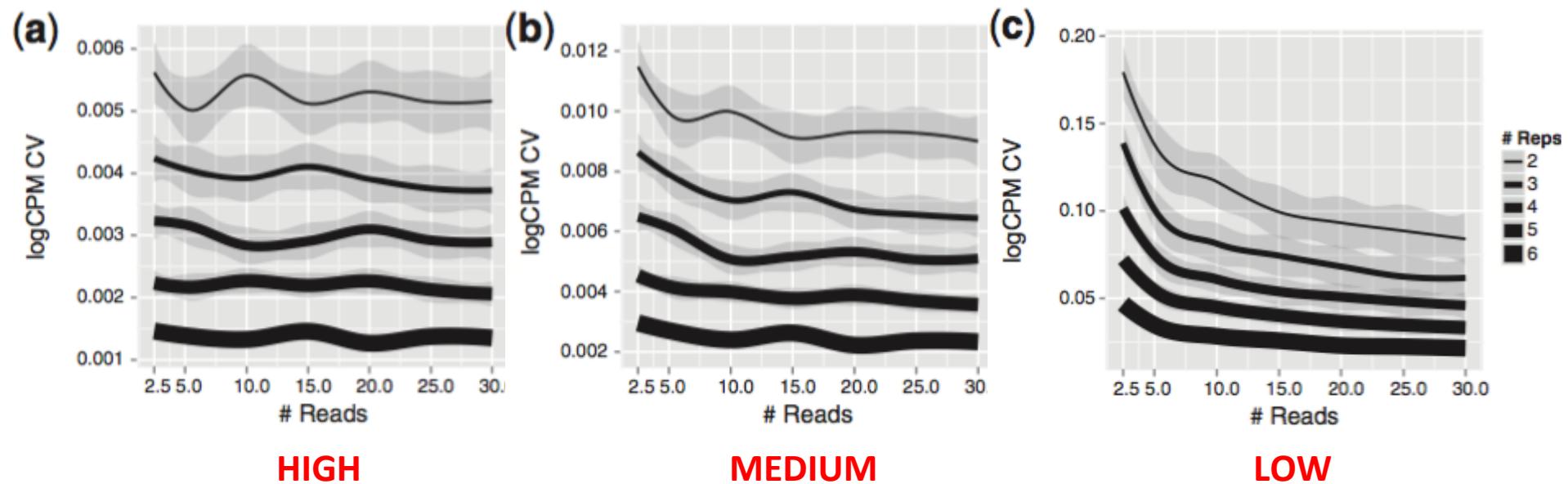
- *Independent filtering* = remove genes that have little chance of showing DE
- Can use eg. total count



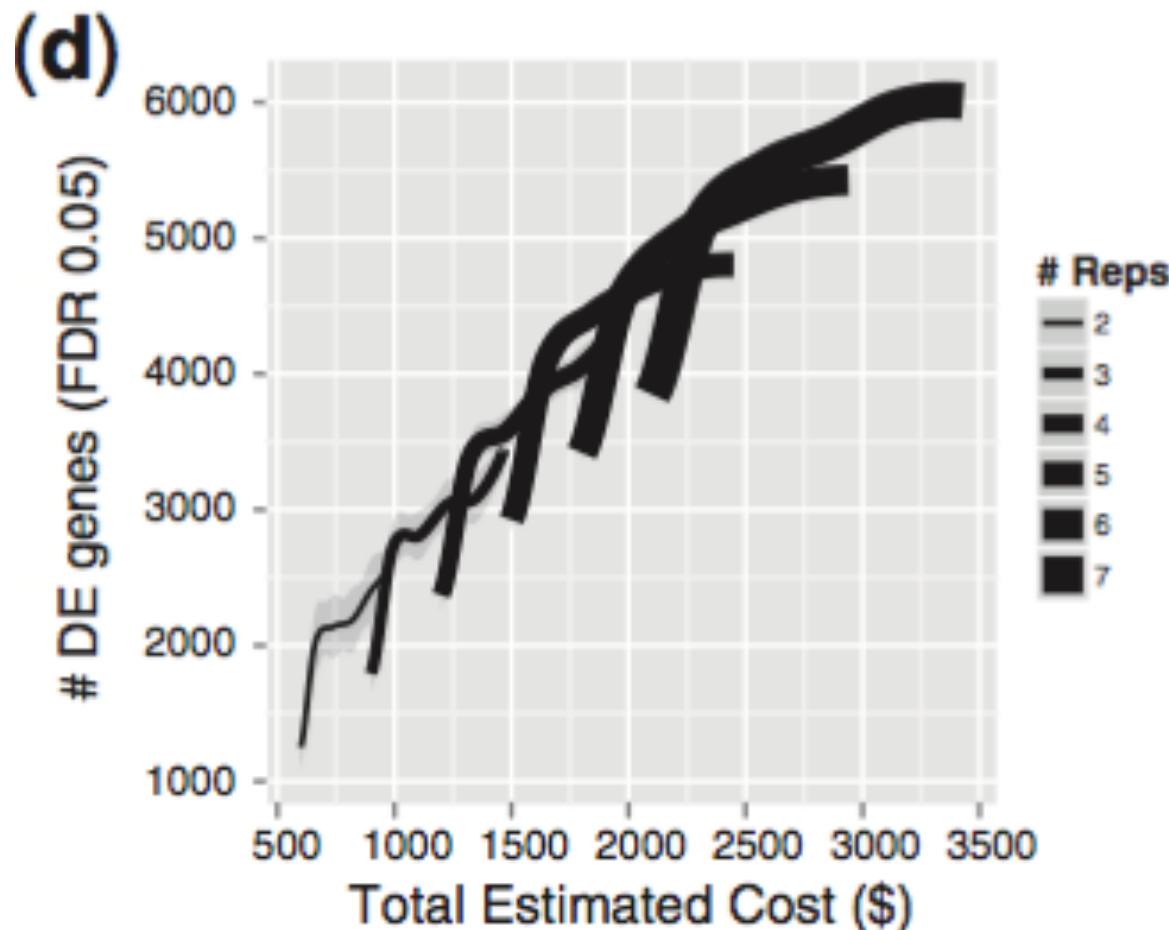
On replicates...



On replicates...



On replicates...



Summary

