

Analysis of

RNA-seq Data

Bernard Pereira



UNIVERSITY OF
CAMBRIDGE



CANCER
RESEARCH
UK

The many faces of RNA-seq

illumina®

AREAS OF INTEREST ▾ TECHNIQUES ▾ SYSTEMS ▾

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

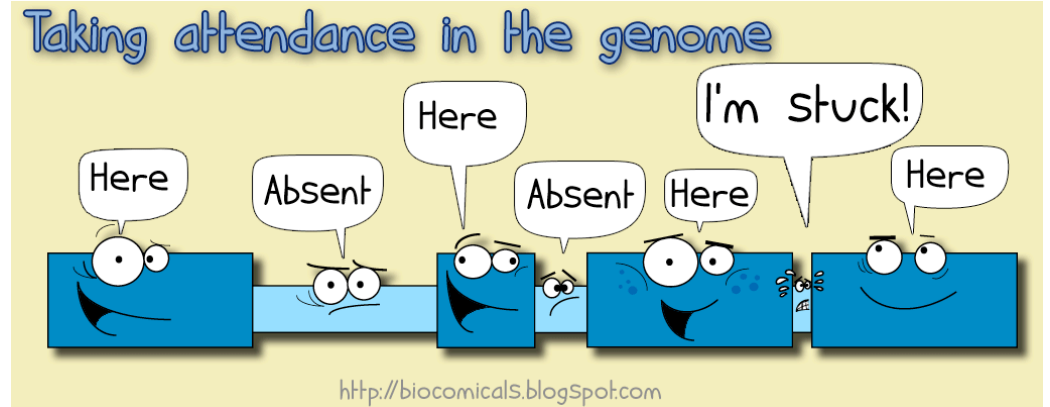
Ribosome Profiling

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

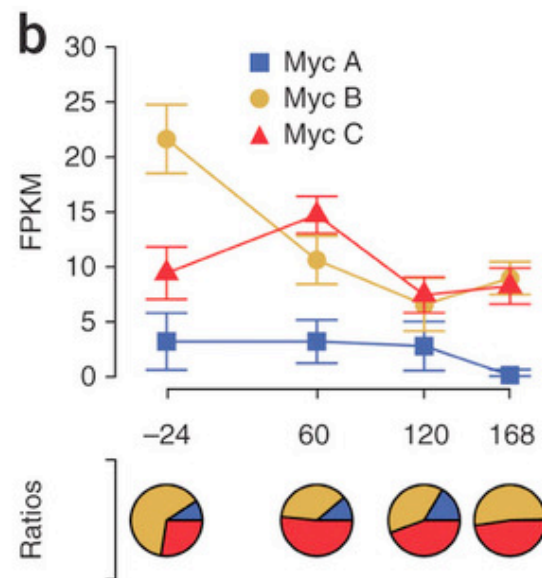
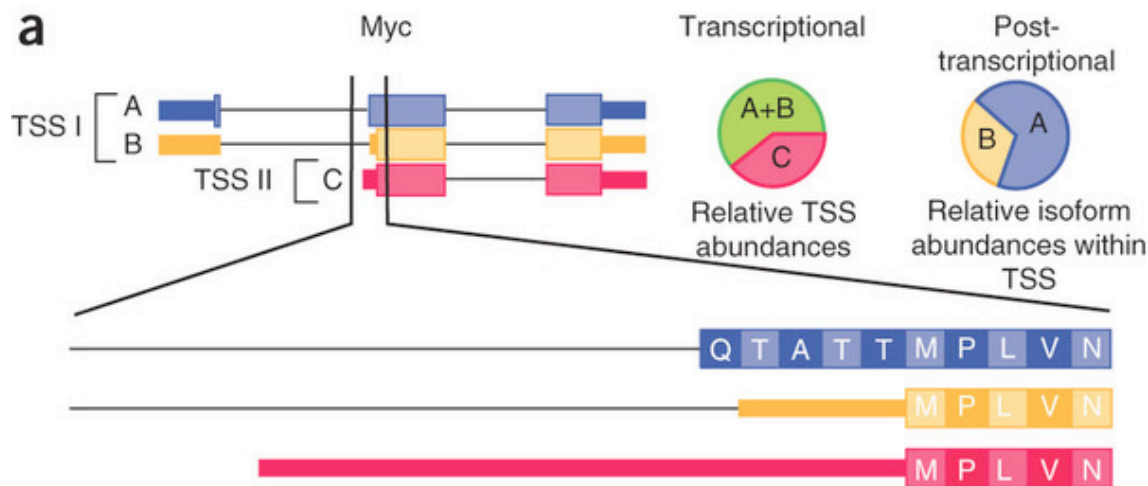
Applications

nature
biotechnology

LETTERS

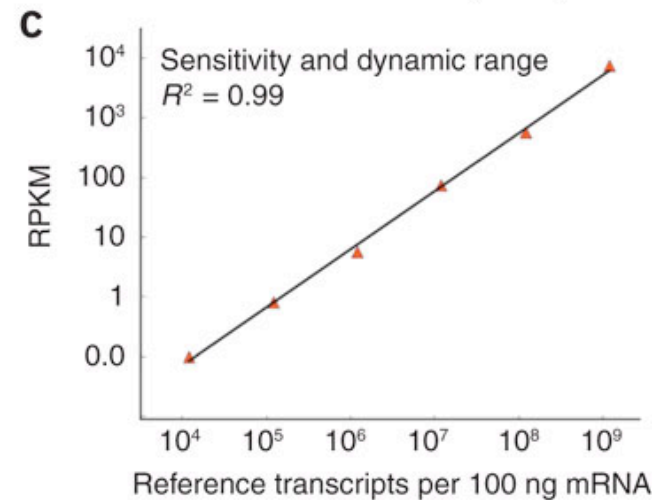
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell¹⁻³, 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}

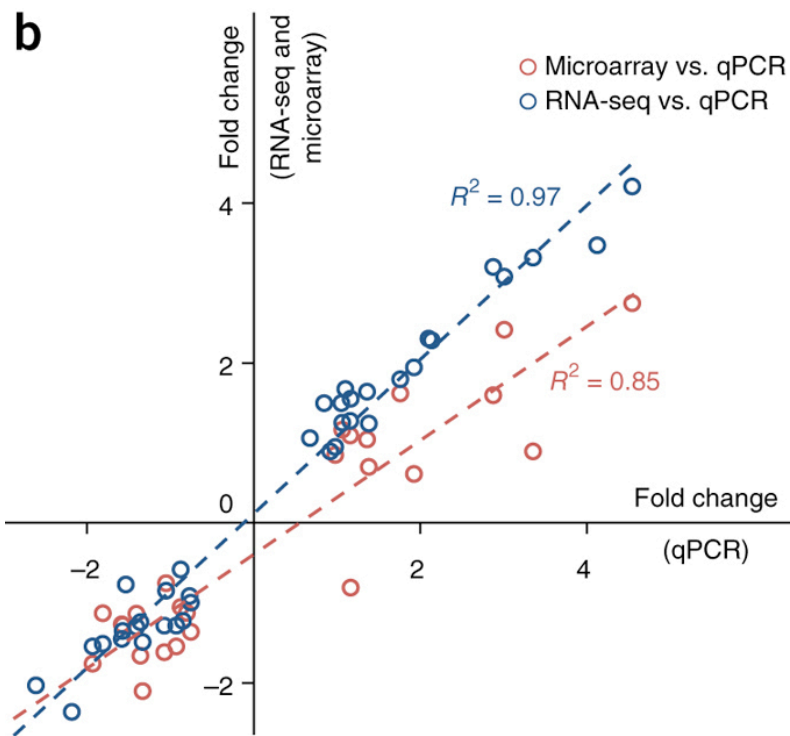
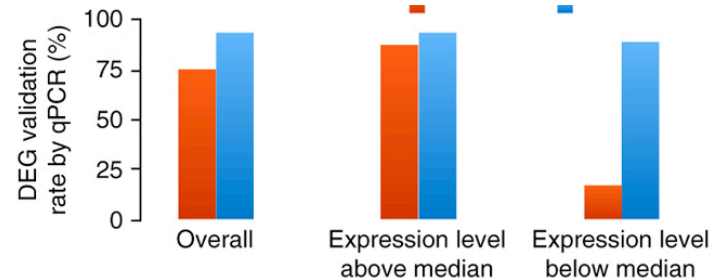
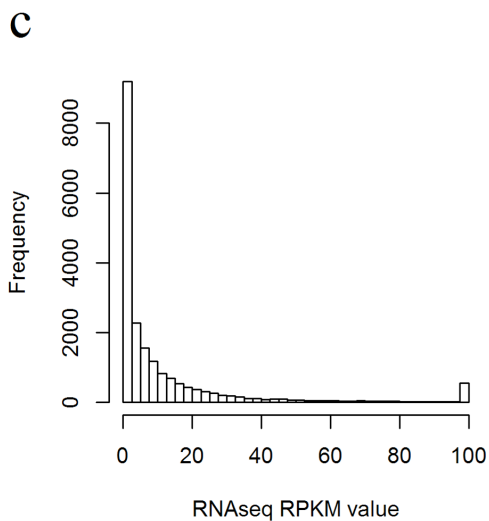
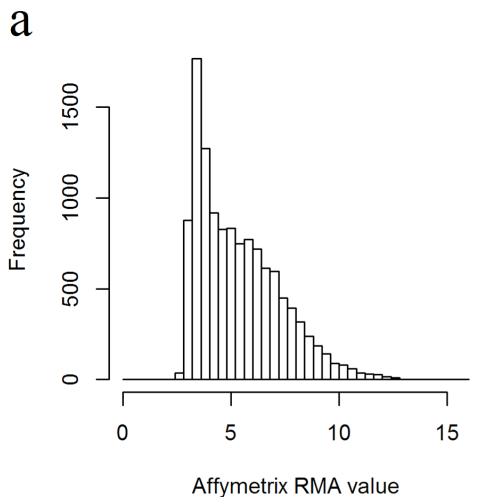


Differential Expression

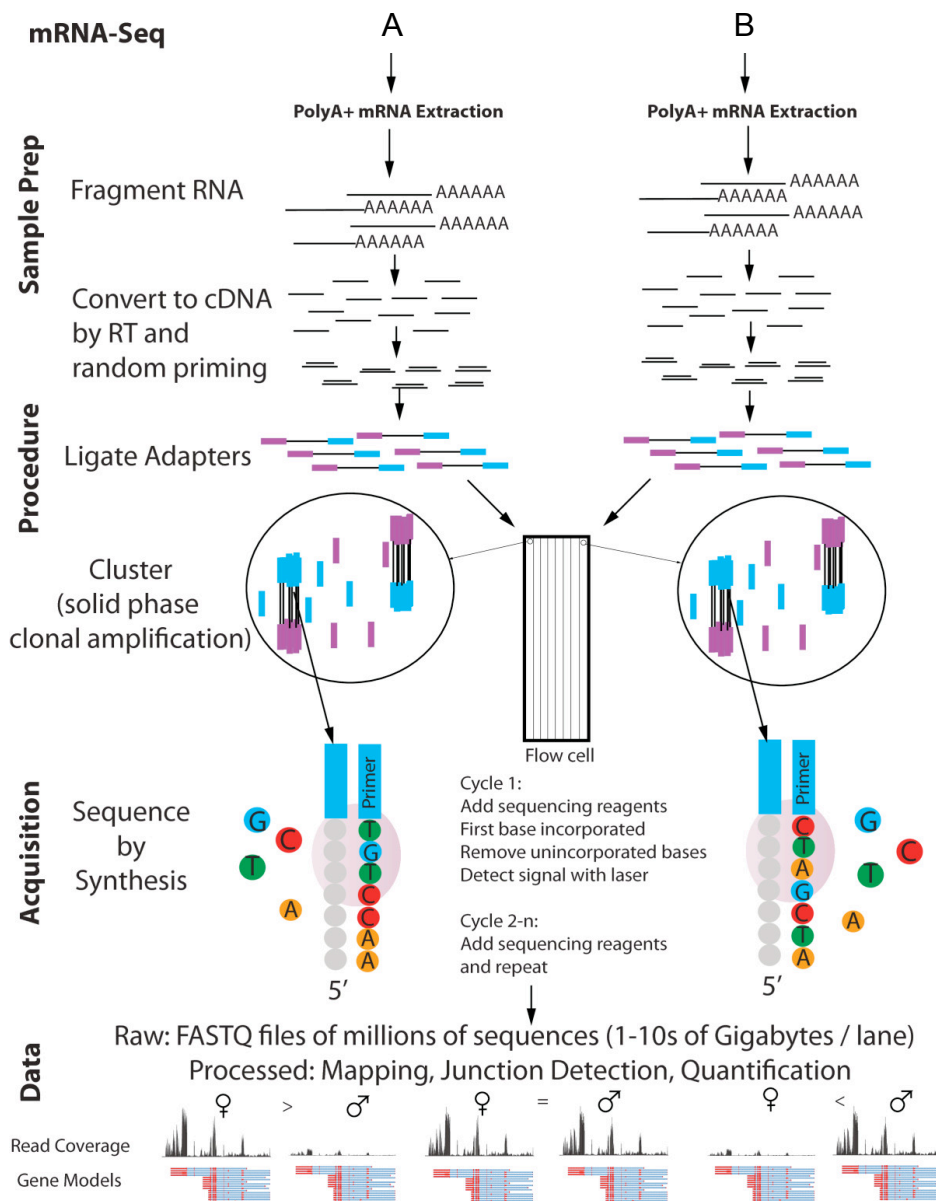
- Comparing feature abundance under different conditions
- Assumes linearity of signal over a range of expression levels
- When *feature=gene*, well-established pre- and post-analysis strategies exist



Range of detection



Library Prep i



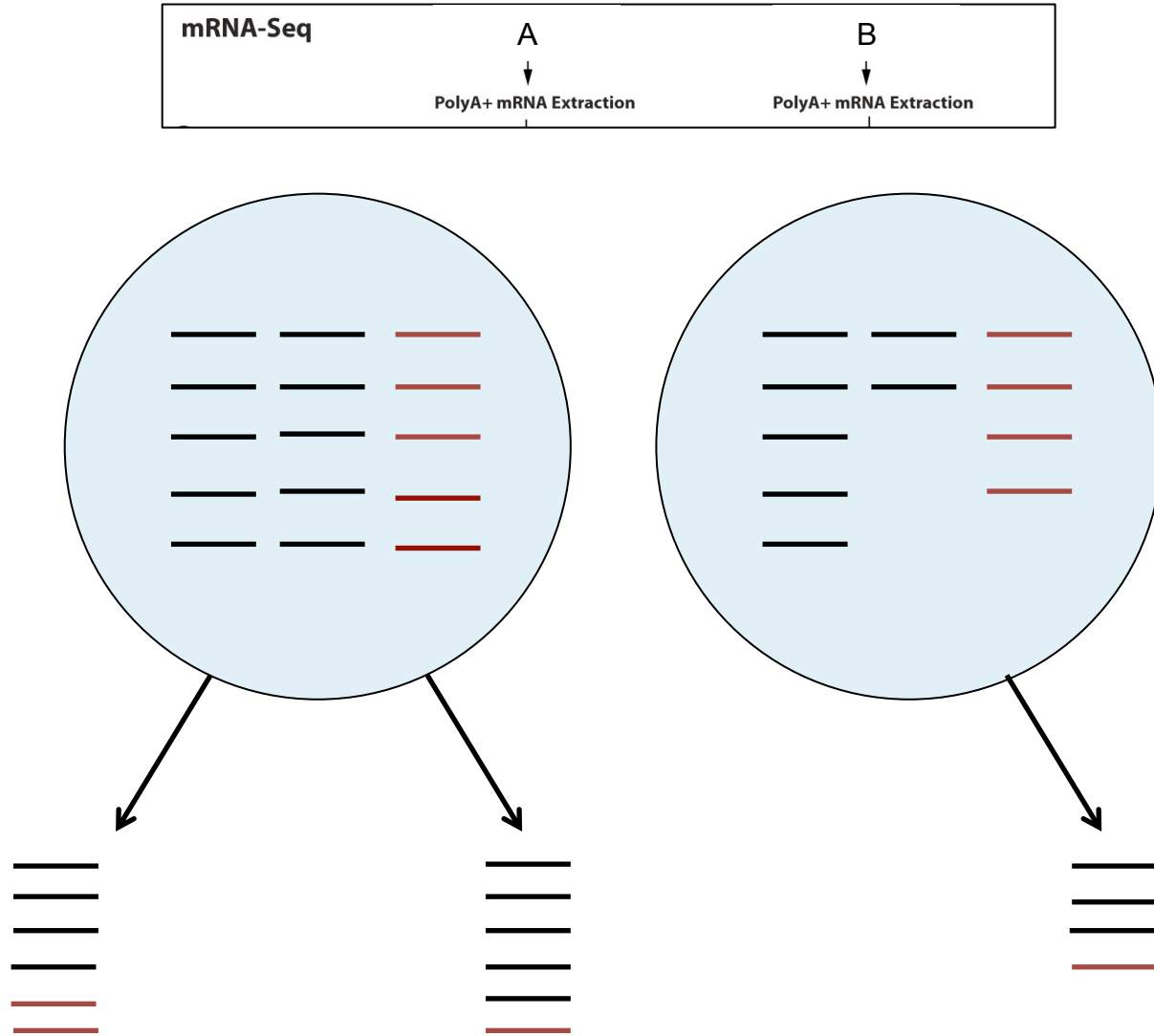
Library Prep ii



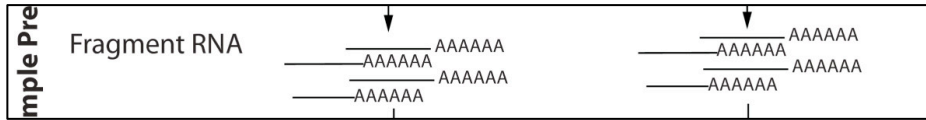
Biological

Technical

Library Prep iii



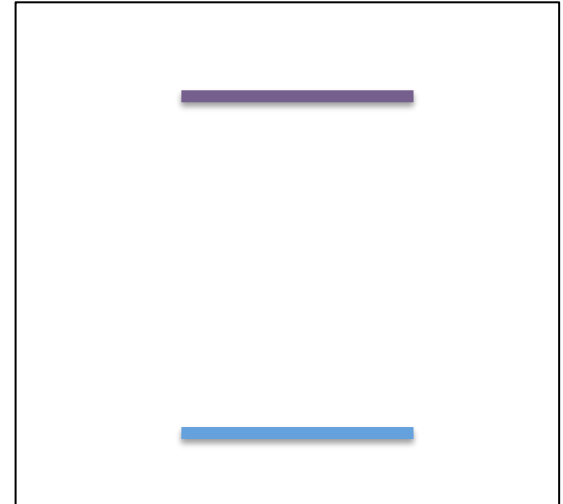
Library Prep iii



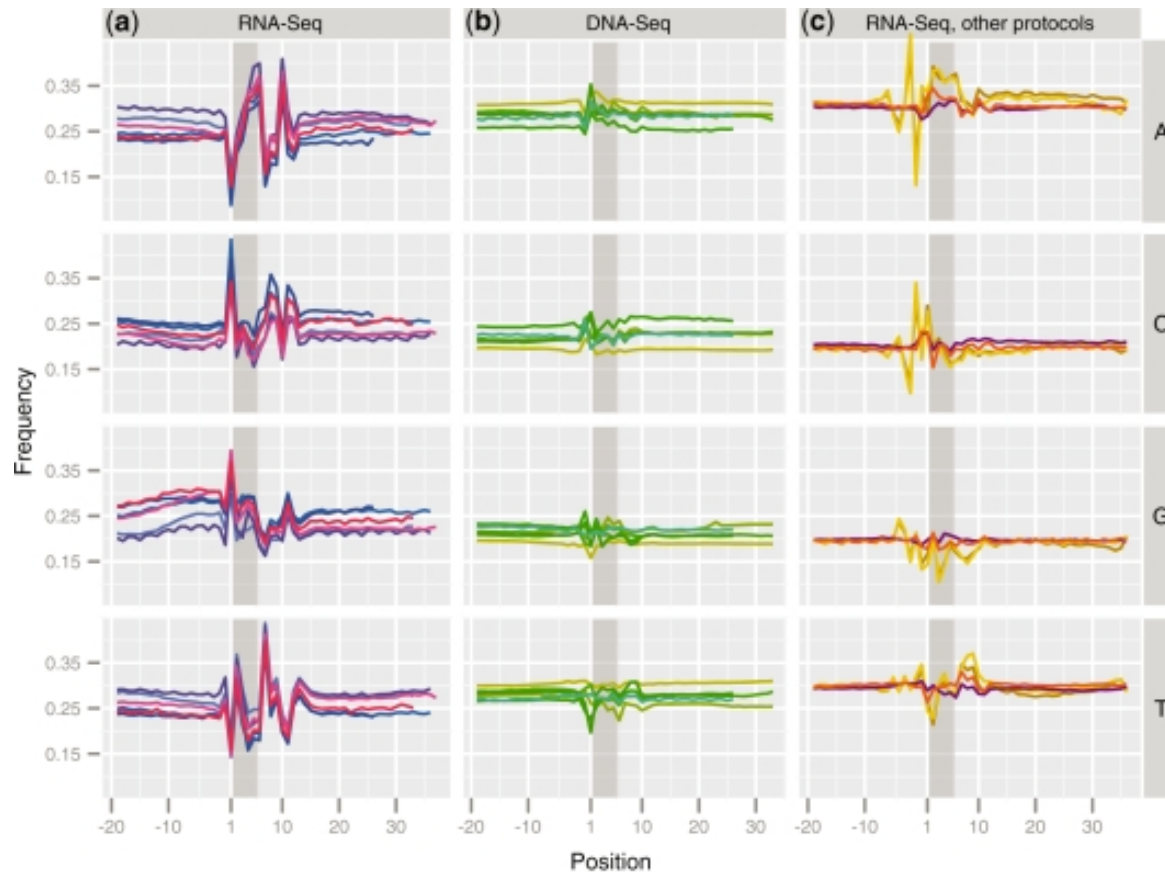
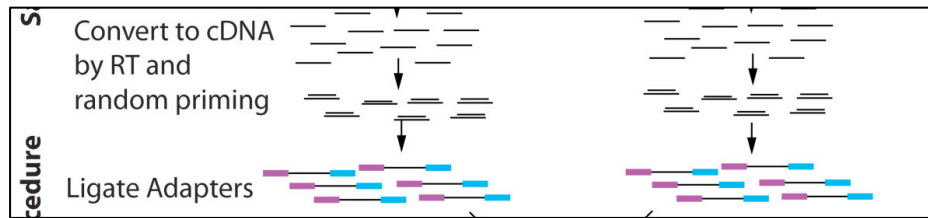
A



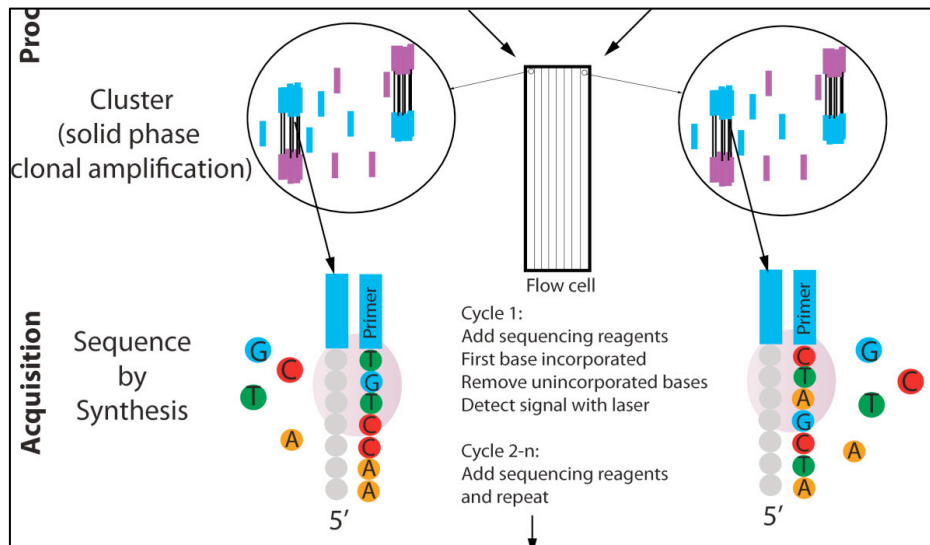
B



Library Prep iv

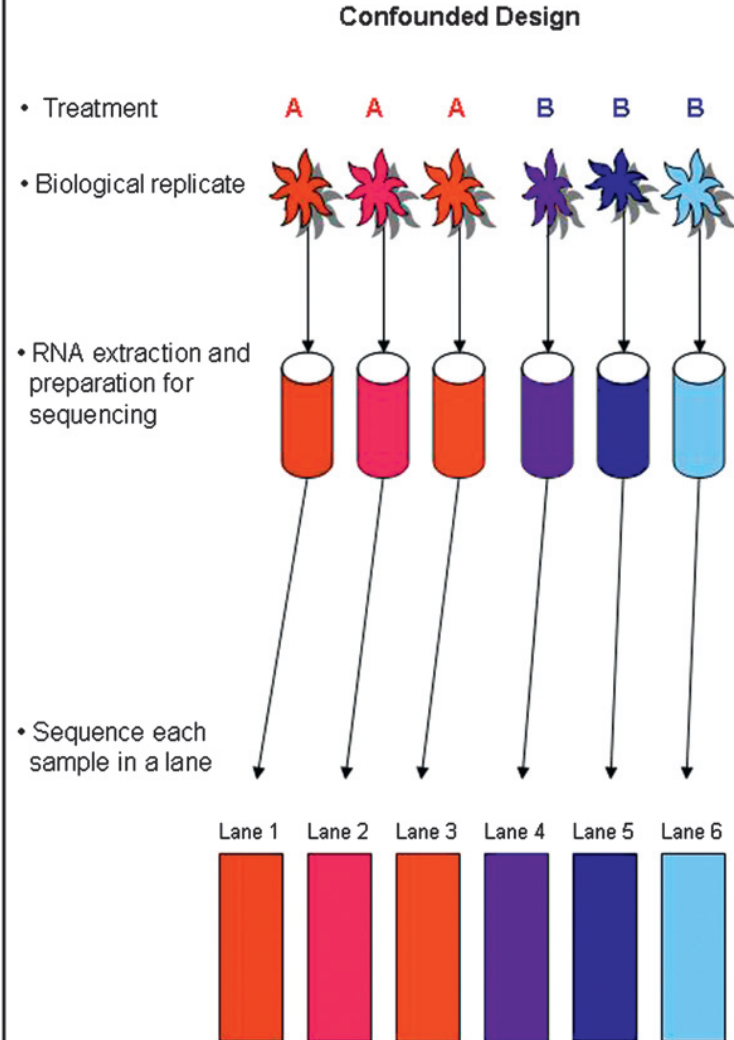
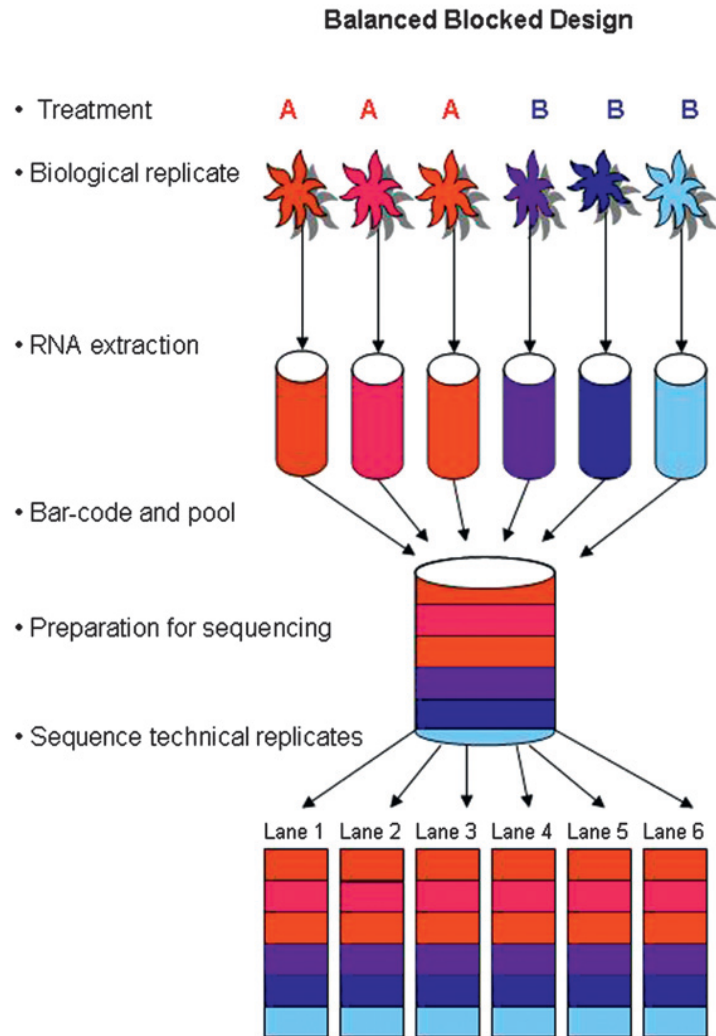


Library Prep v

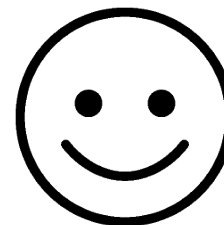


- Duplicates (optical & PCR)
- Sequence errors
- Indels
- Repetitive/problematic sequence

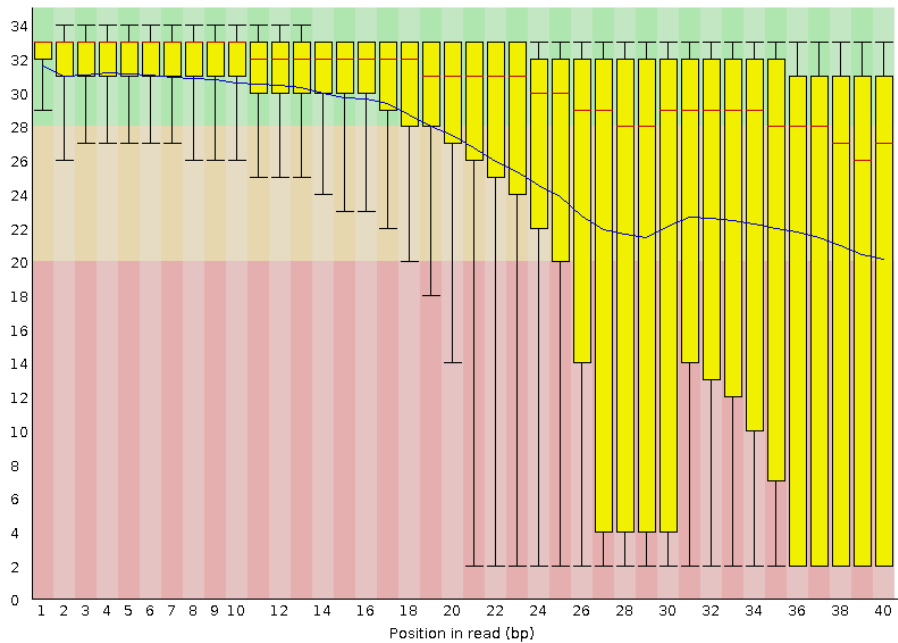
Hot off the sequencer...



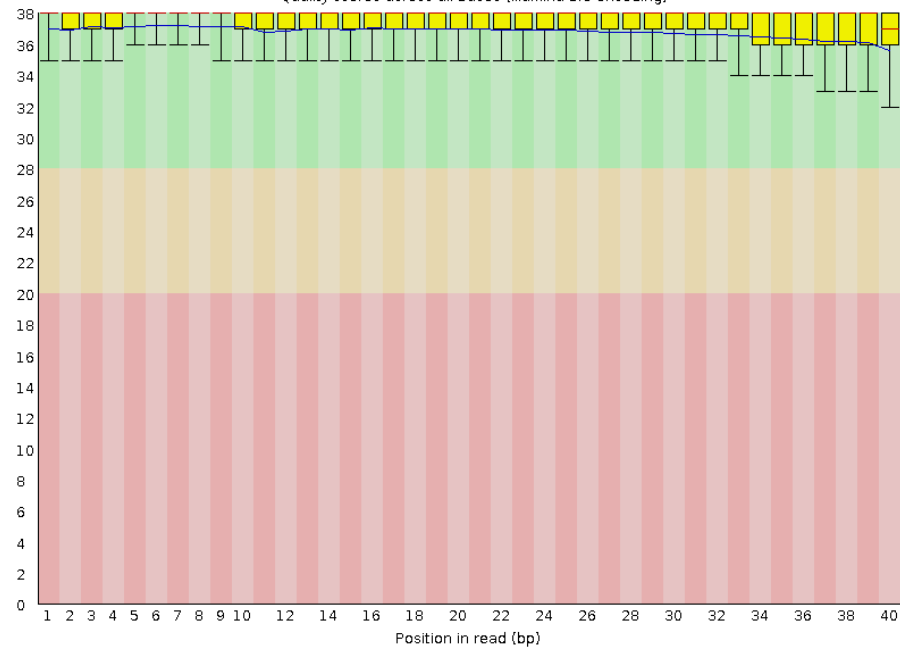
FASTQC



Quality scores across all bases (Illumina 1.5 encoding)

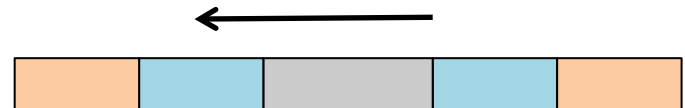


Quality scores across all bases (Illumina 1.5 encoding)

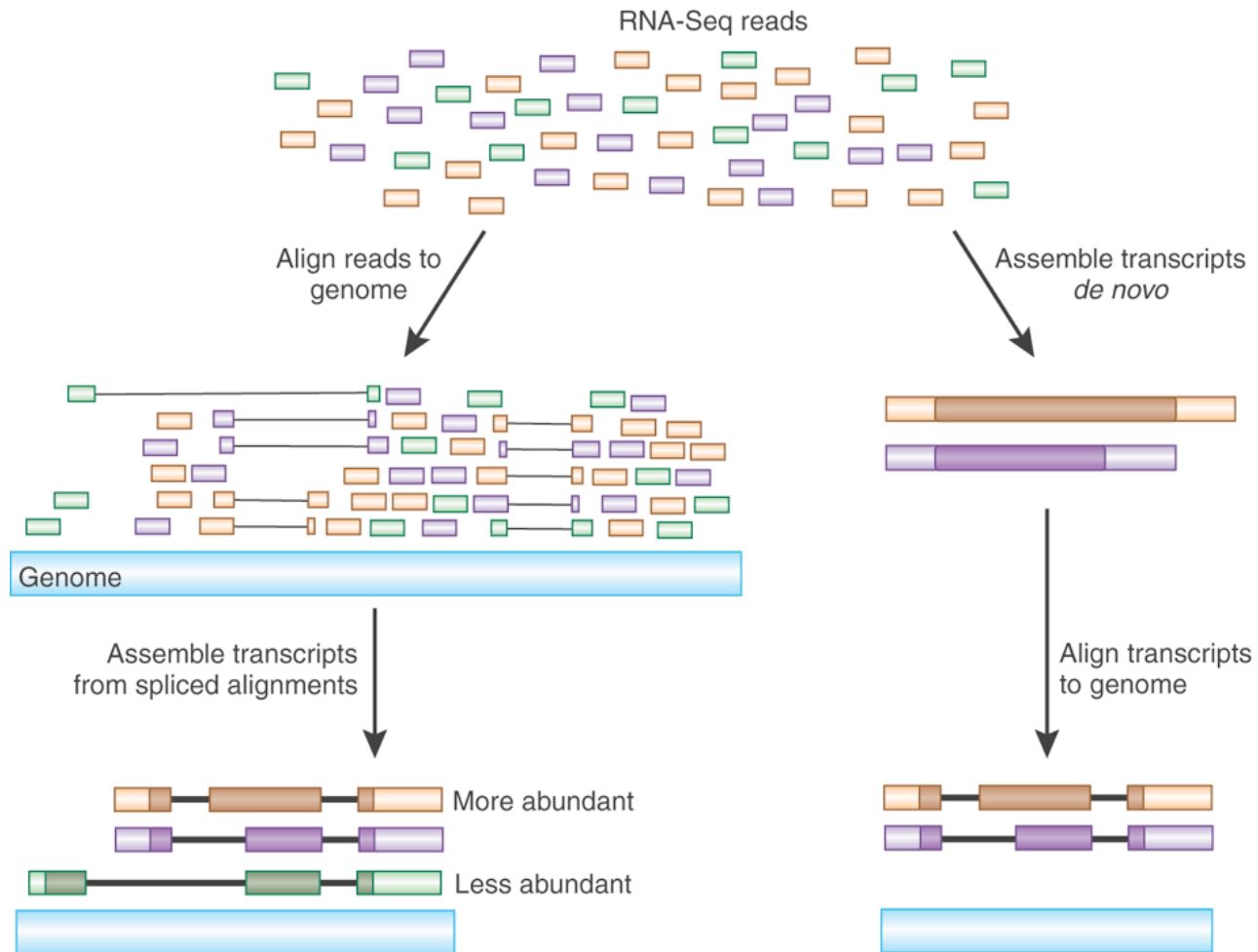


Trimming

- Quality-based trimming
- Adapter 'contamination'

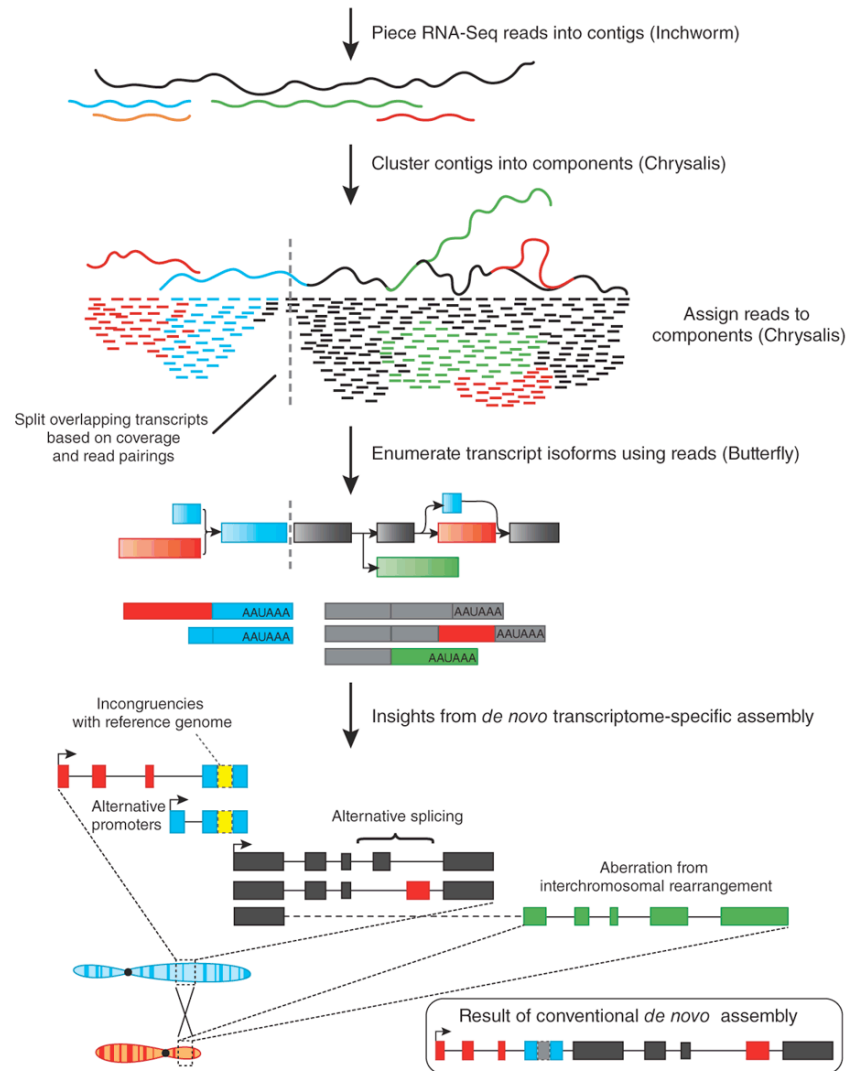


Sequence to sense

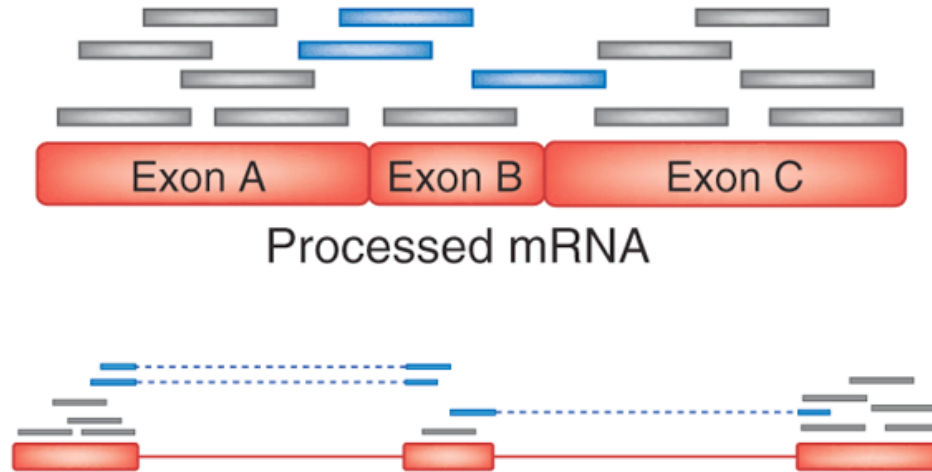


De novo assembly

- eg. Trinity



Reference-based assembly




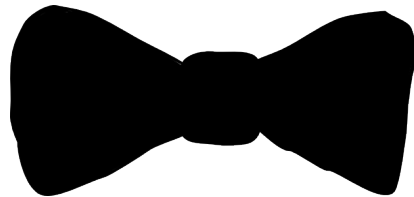
Genome mapping

- Can identify novel features
- Splice 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)



Bowtie
Extremely fast, general purpose short read aligner



TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites



Cufflinks package

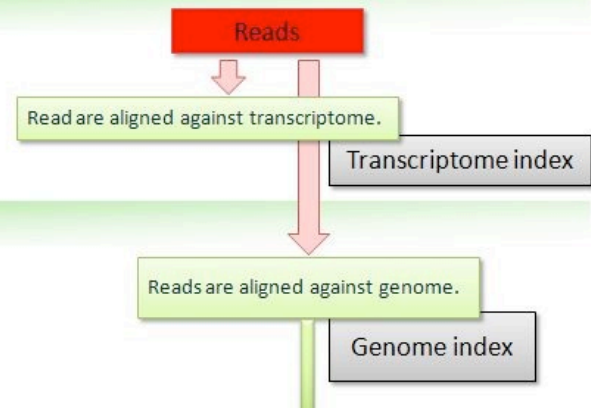
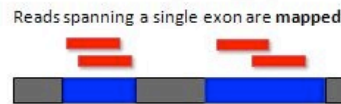
- Cufflinks**
Assembles transcripts
- Cuffcompare**
Compares transcript assemblies to annotation
- Cuffmerge**
Merges two or more transcript assemblies
- Cuffdiff**
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

Tophat/Bowtie

(1) Transcriptome alignment (optional)



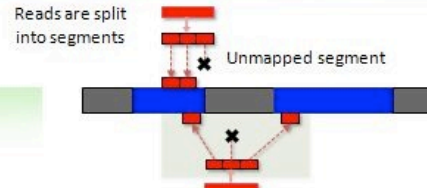
(2) Genome alignment



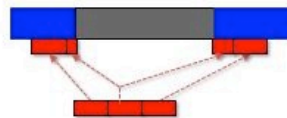
Tophat/Bowtie

(3) Spliced alignment

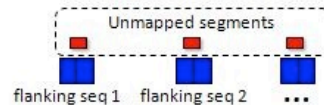
(3-1) Segment alignment to genome



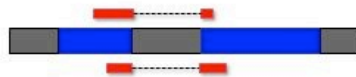
(3-2) Identification of splice sites (including indels and fusion break points)



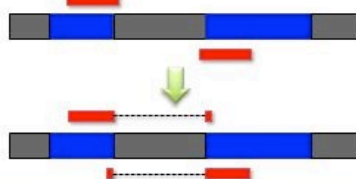
(3-3) Segments aligned to junction flanking sequences



(3-4) Segment alignments stitched together to form whole read alignments



(3-5) Re-alignment of reads minimally overlapping introns



Reads are split into smaller segments which are then aligned to the genome.

Genome index

Segment mappings are used to find potential splice sites usually when the distance between the mapped positions of the left and the right segments are longer than the length of the middle part of a read.

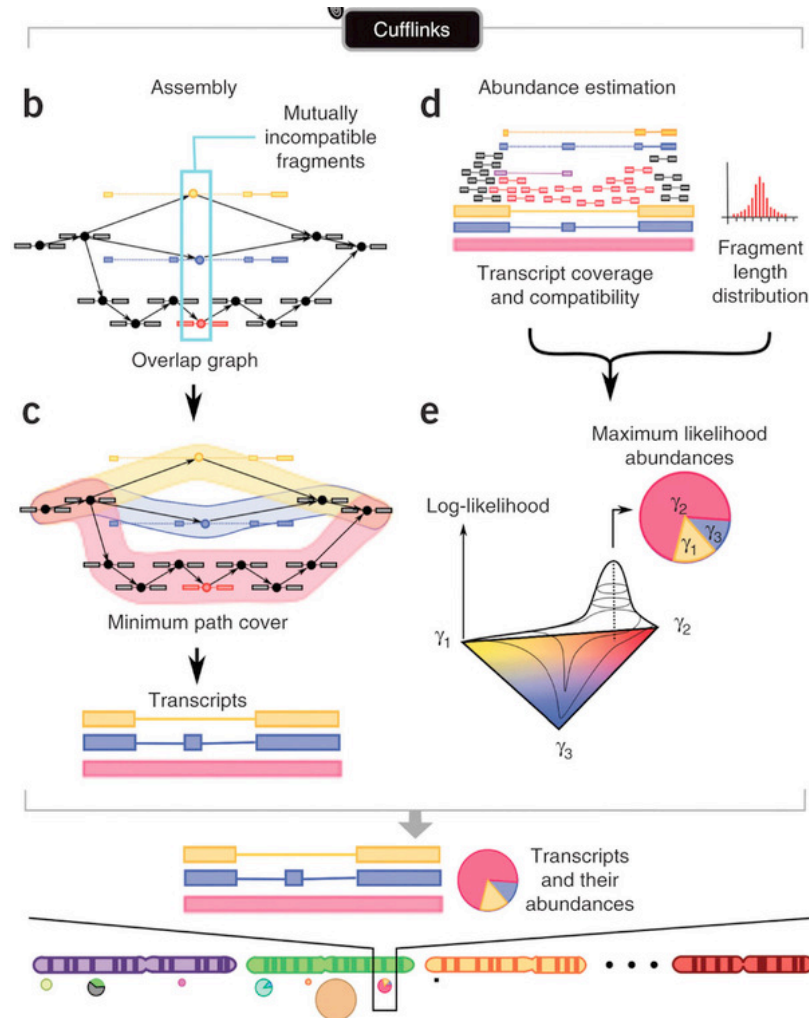
Sequences flanking a splice site are concatenated and segments are aligned to them.

Junction flanking index

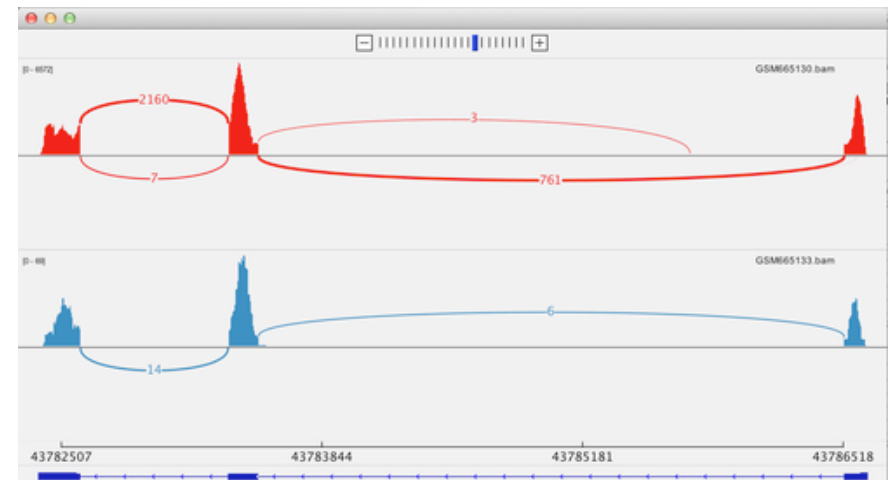
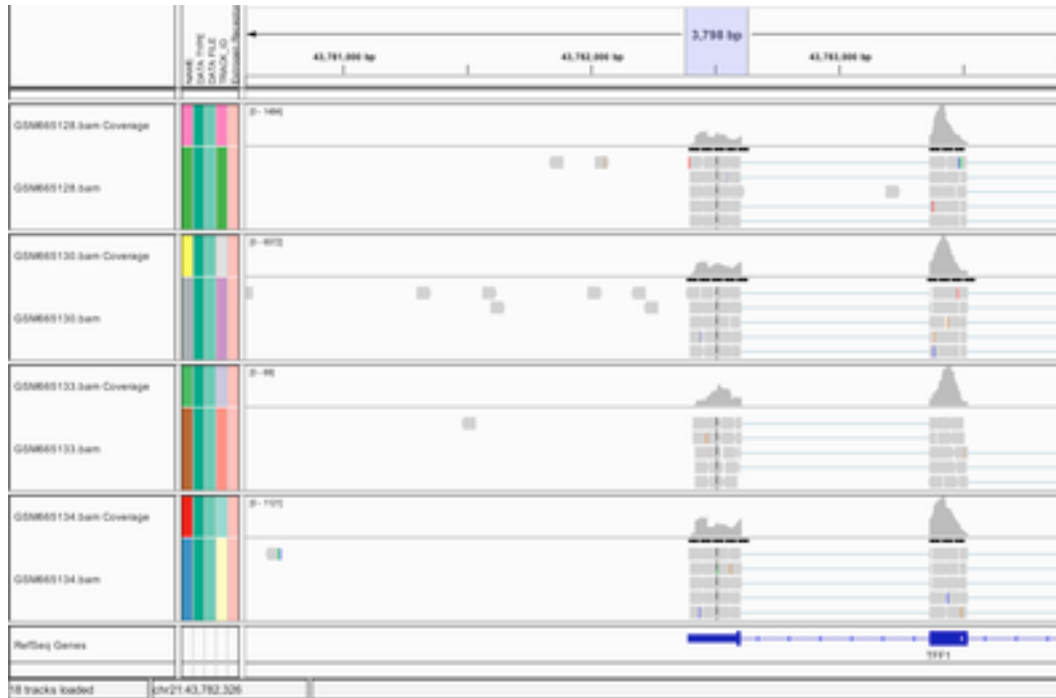
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

Genome mapped reads with alignments extending a few bases into introns are re-aligned to exons instead.

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

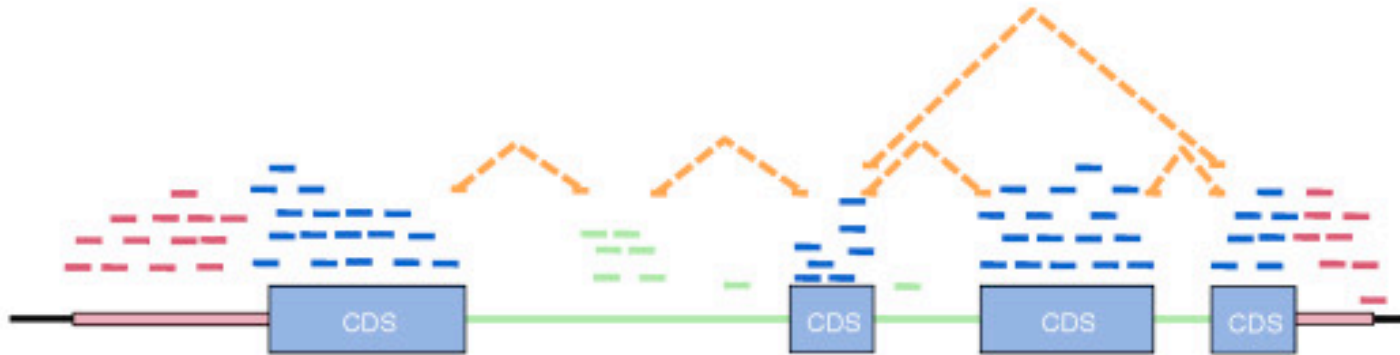


Platform/pipeline

Single-end vs paired-end

Counting

(b)



Genome-based features

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

Transcript-based features

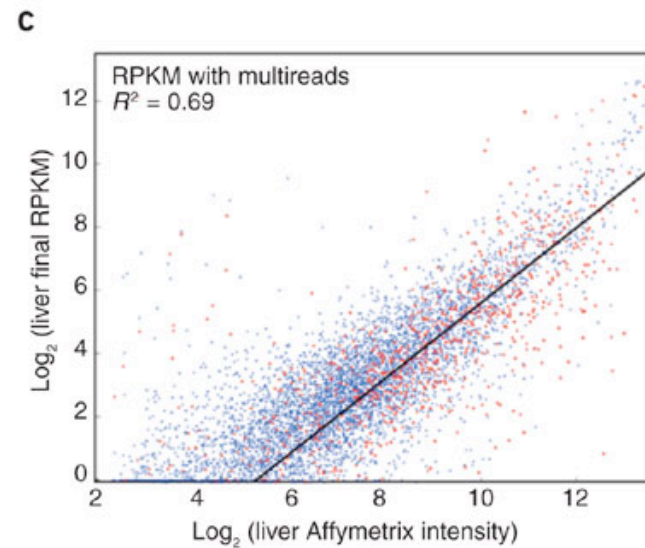
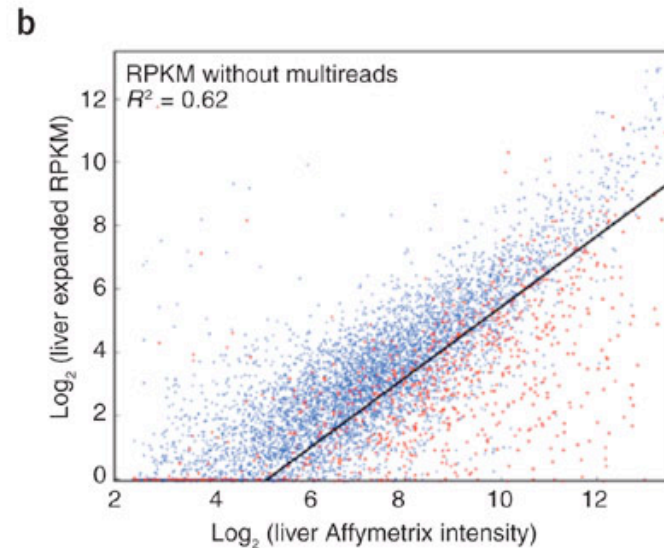
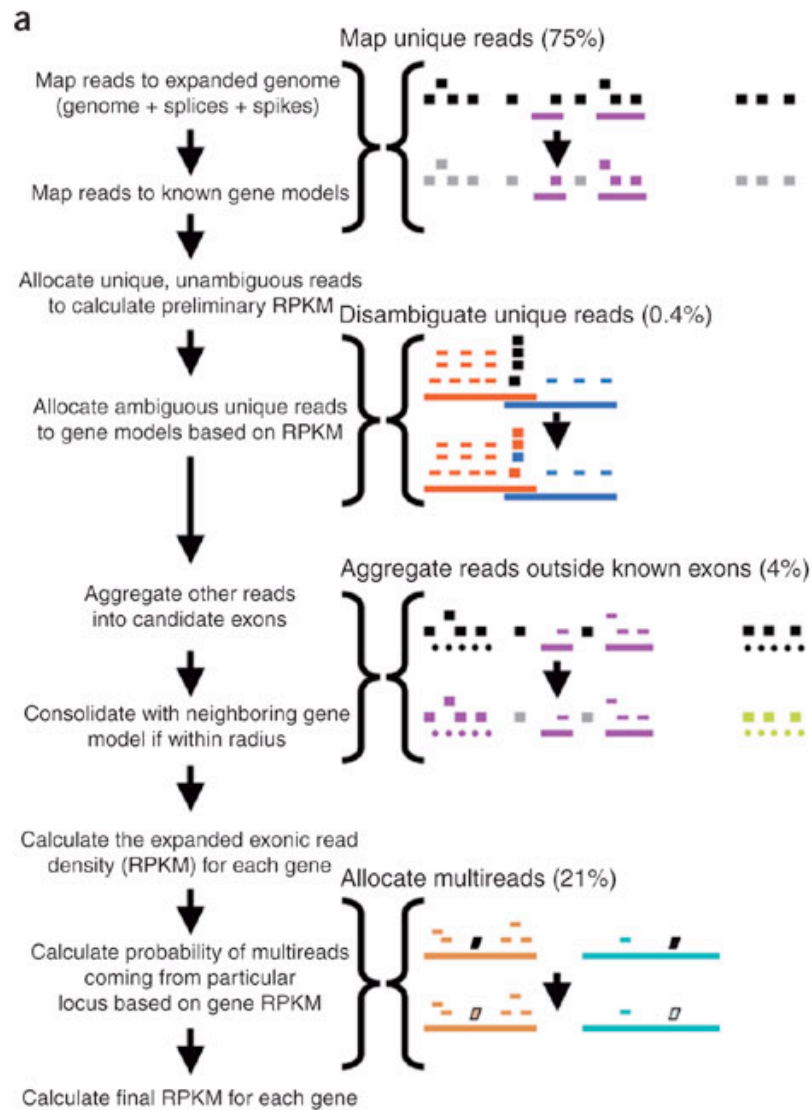
- Transcript assembly?
- Novel structures?
- Isoform multireads?

Counting

- eg. HTseq

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	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

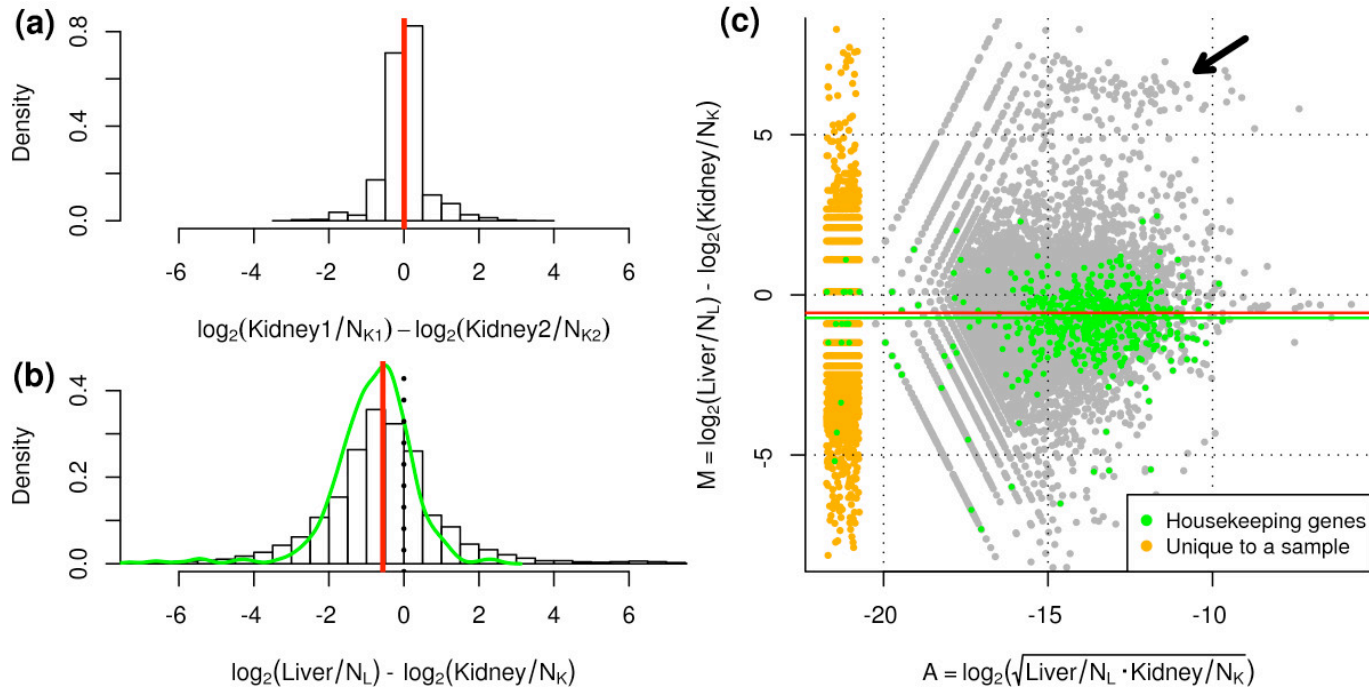
Library composition

- 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

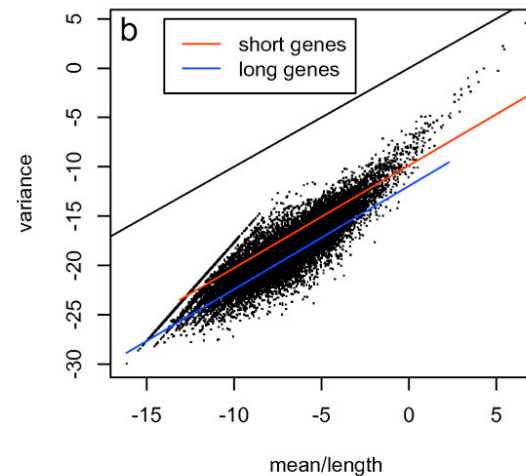
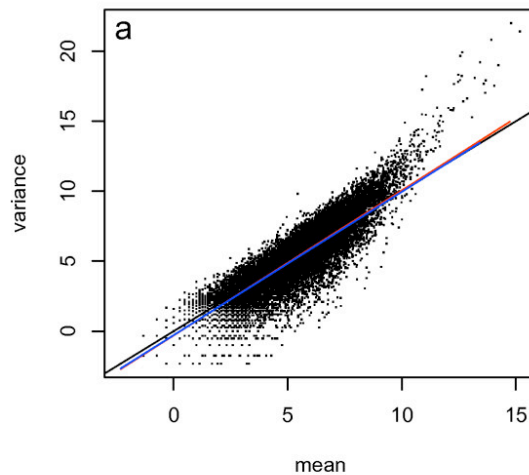


Normalisation ii

RPKM

- Reads per kilobase per million =

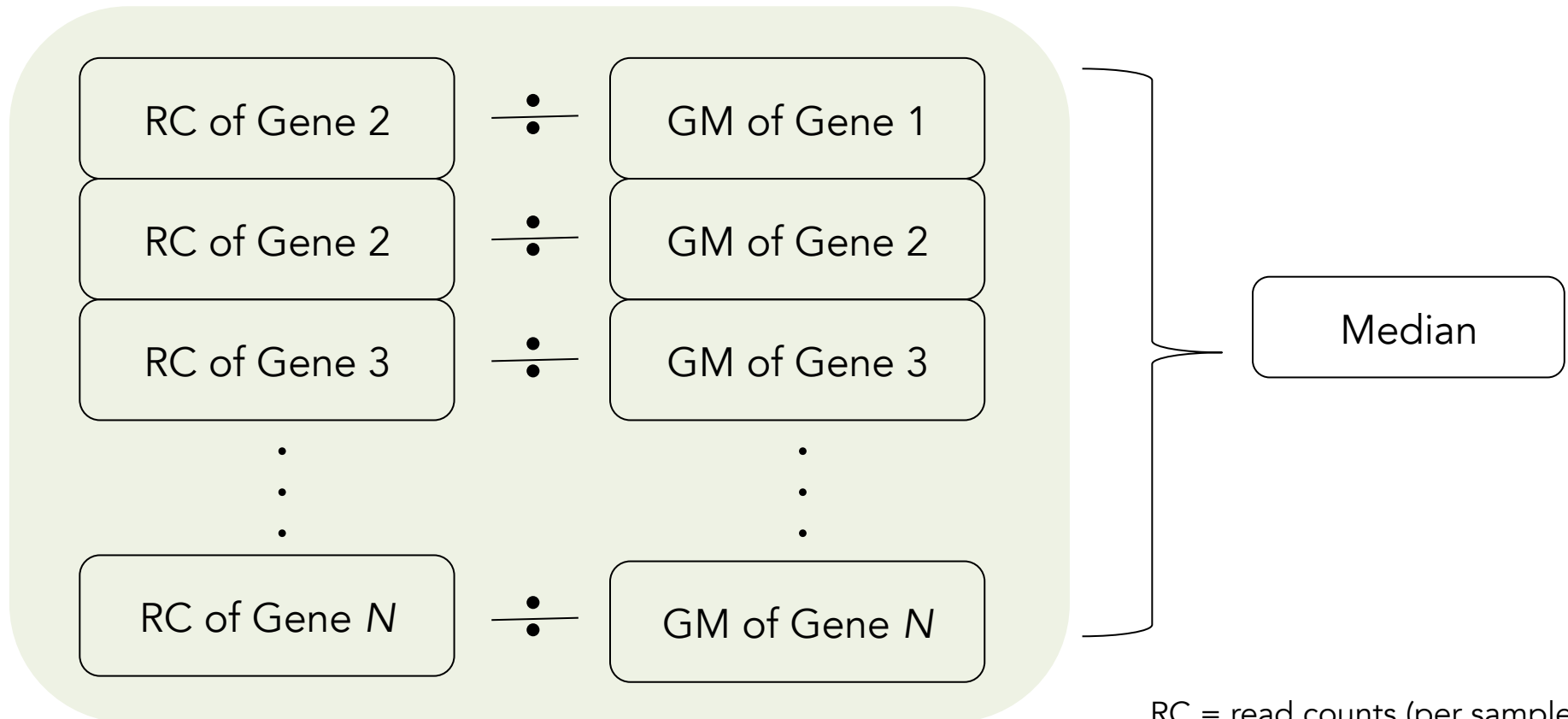
$$\frac{\text{reads for gene A}}{\text{length of gene A} \times \text{Total number of reads}}$$



Normalisation iii

Geometric scaling factor

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

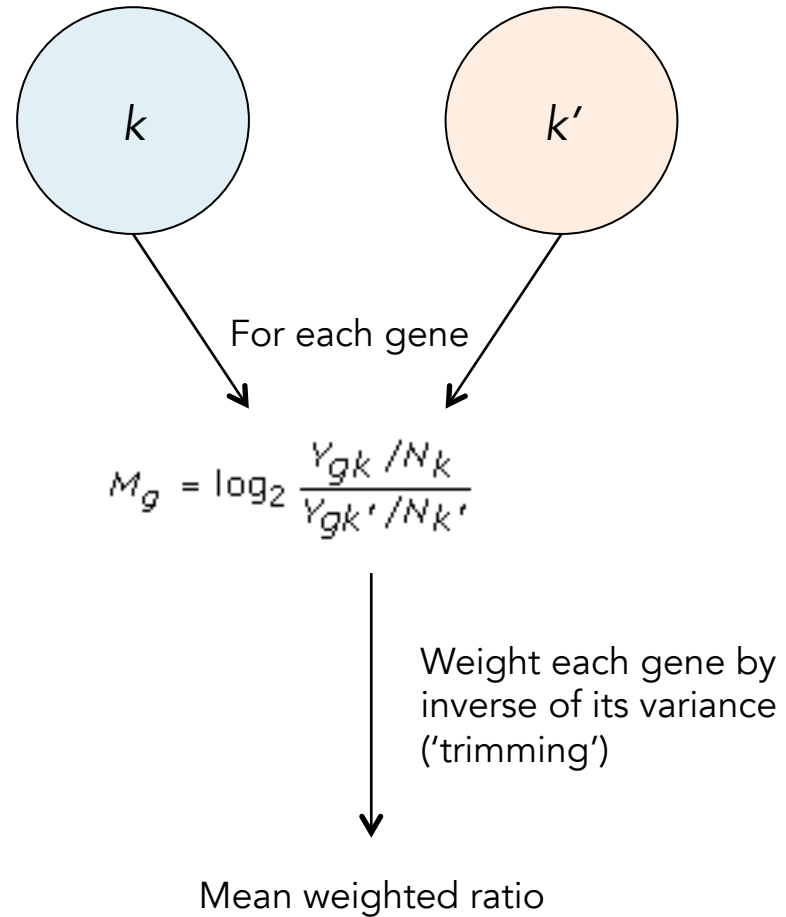


RC = read counts (per sample)
GM = geometric mean (all samples)

Normalisation iv

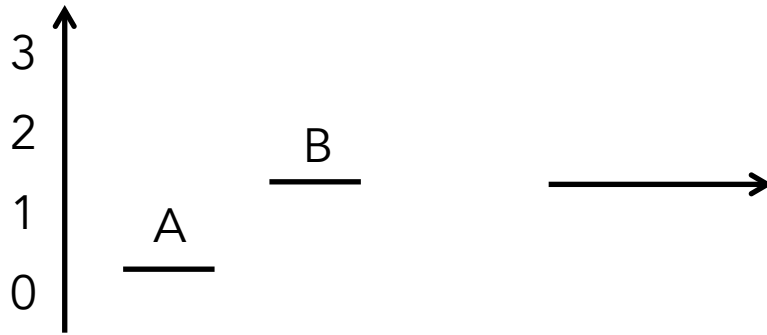
Trimmed mean of M

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



Differential expression

- Simple



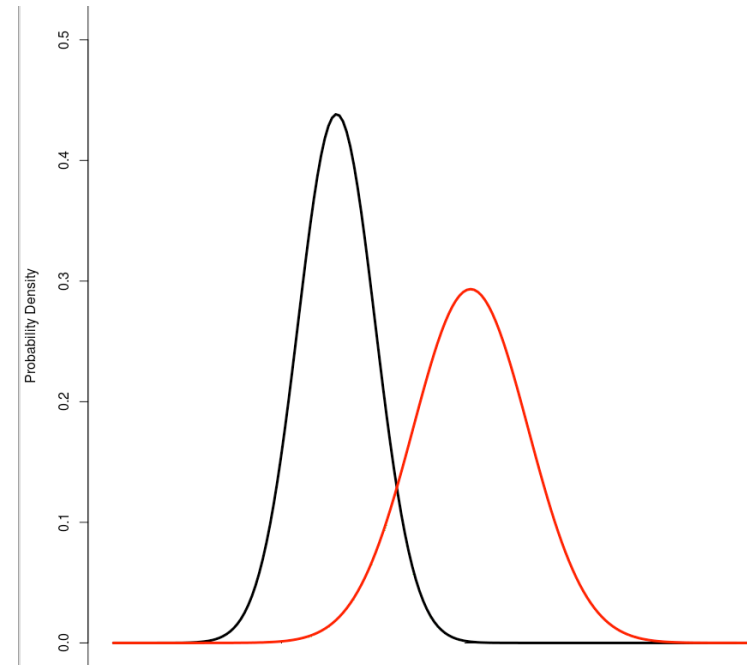
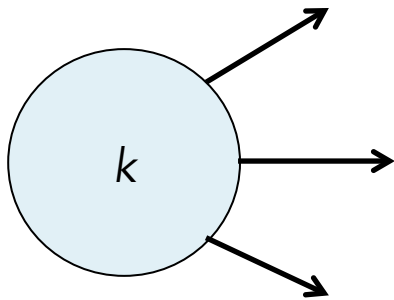
	Cond A	Cond B
Gene X		
Other		

All we need

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

Modelling – old trends

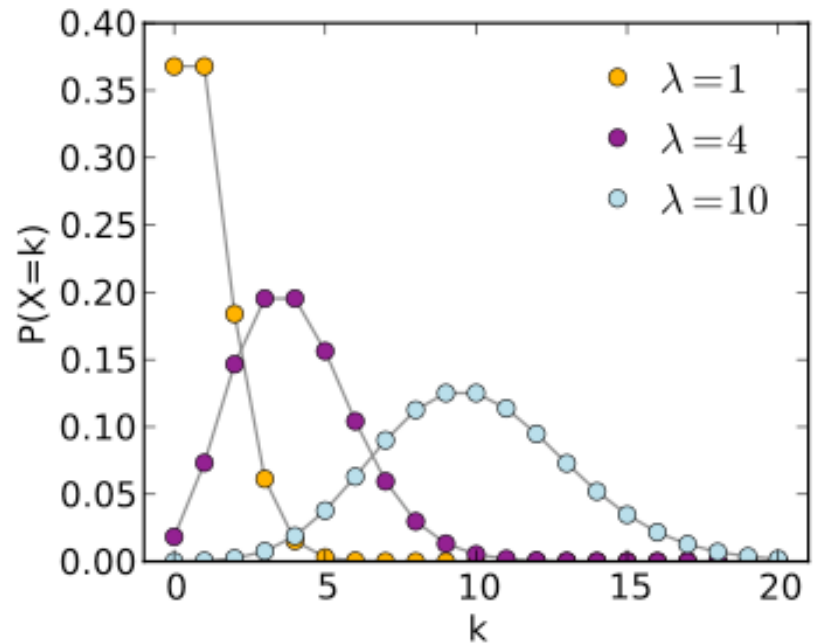
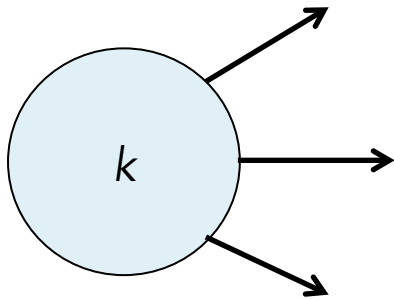
- Technical replicates introduce some variance



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

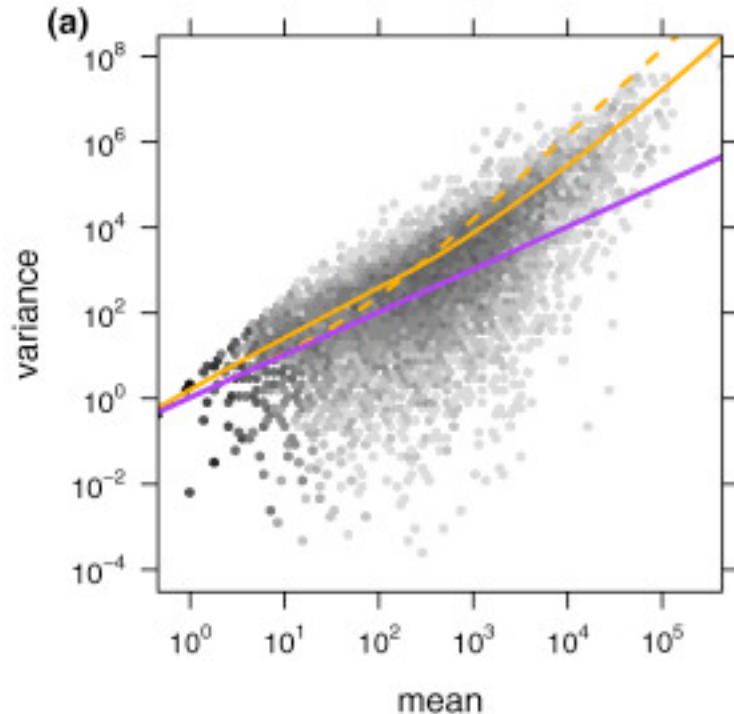
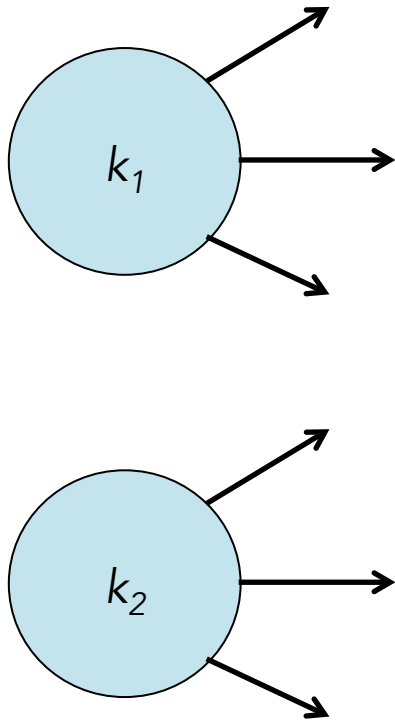
Modelling – in fashion

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



Modelling – in fashion

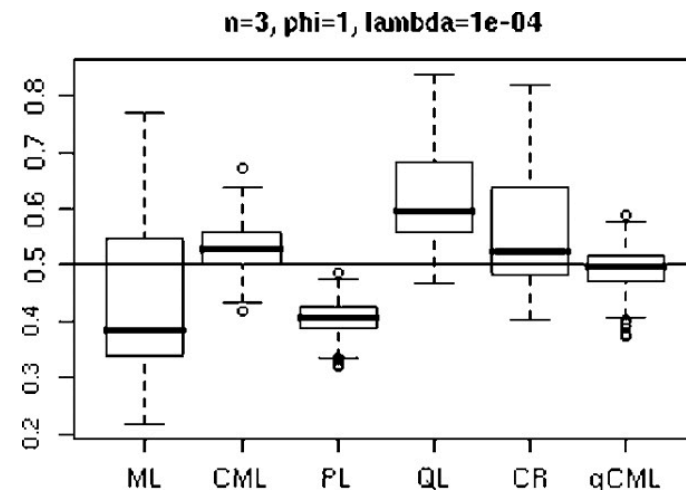
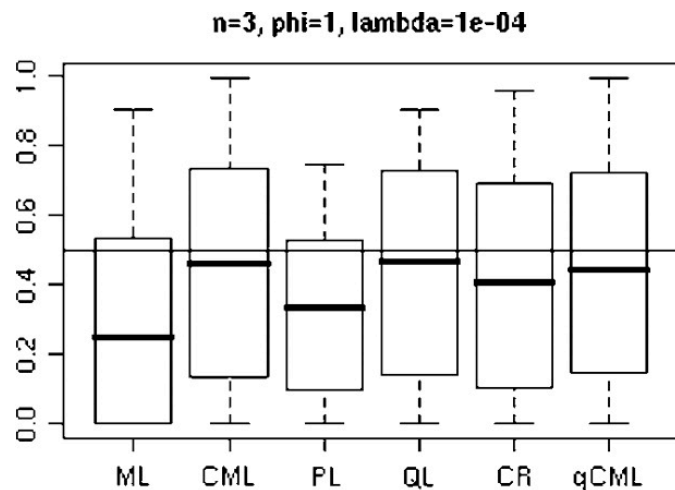
- Biology is never that simple...



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

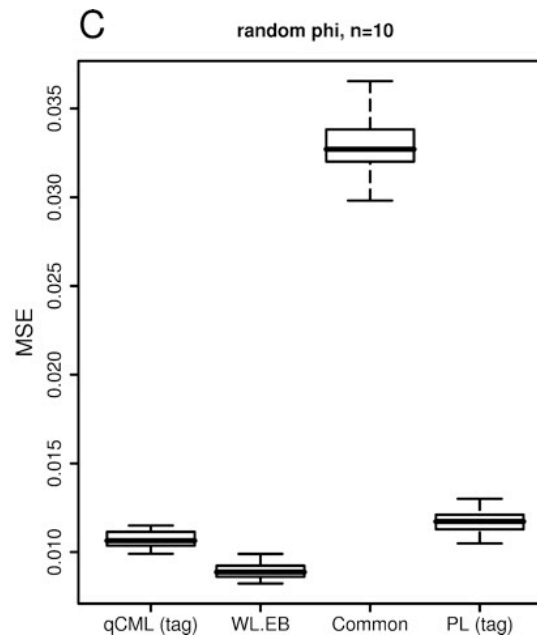
Modelling – in fashion

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



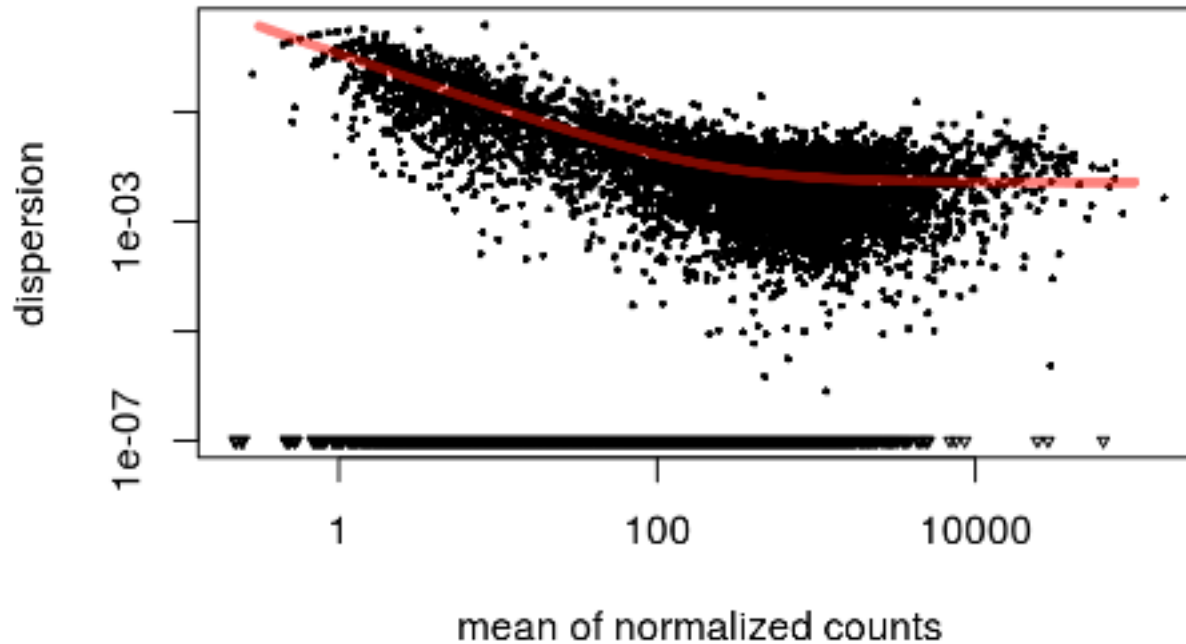
Shrinkage

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



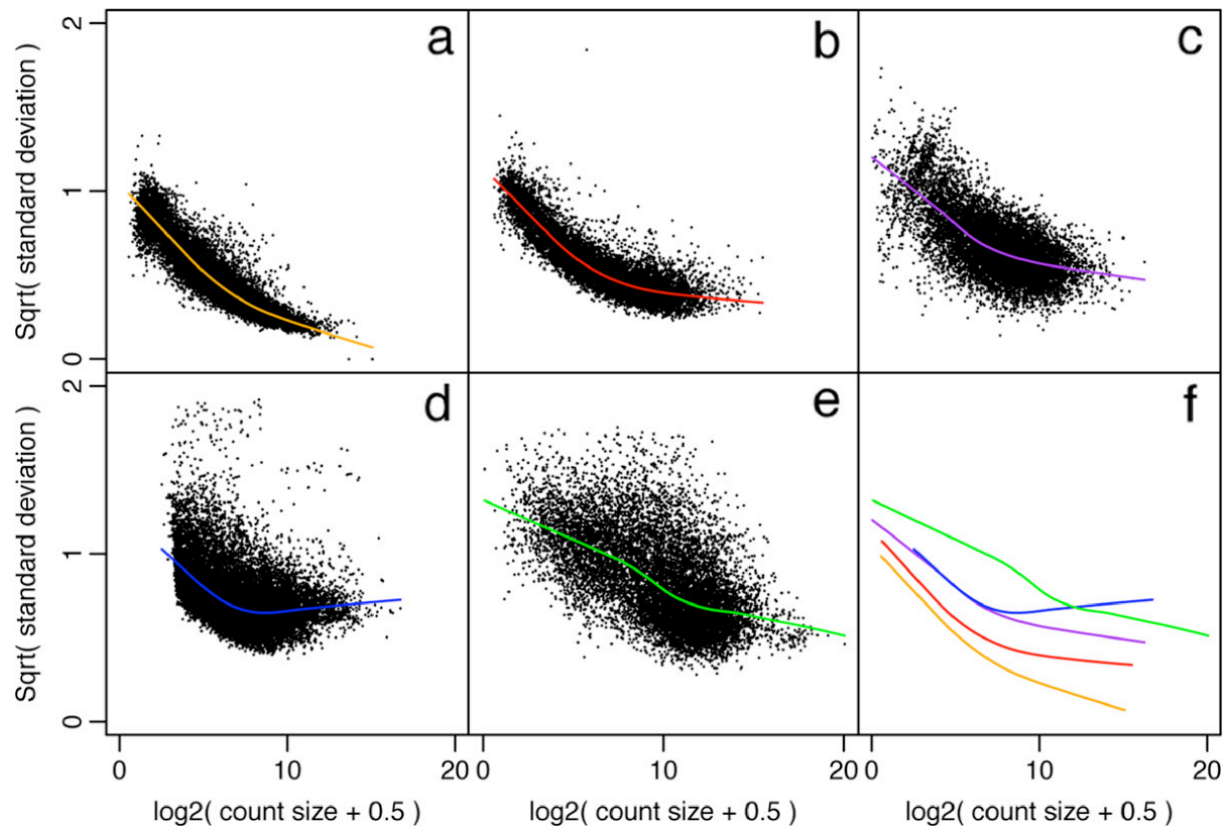
DESeq

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



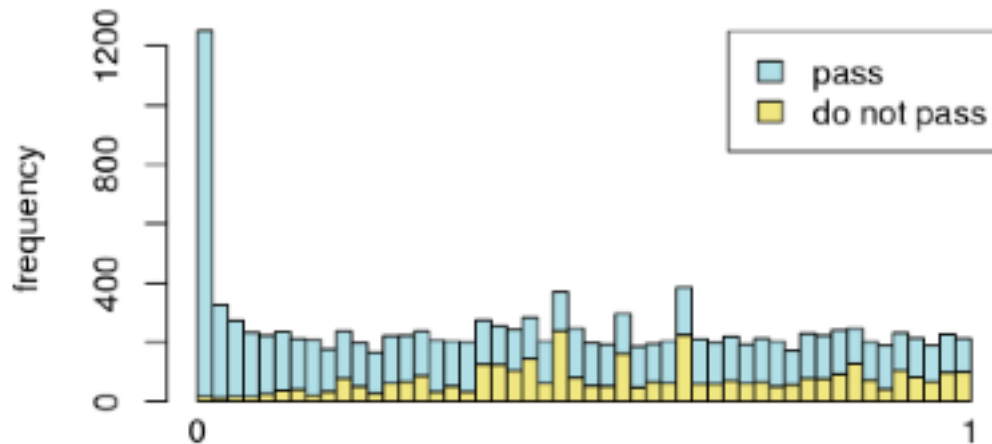
The mean-variance relationship

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

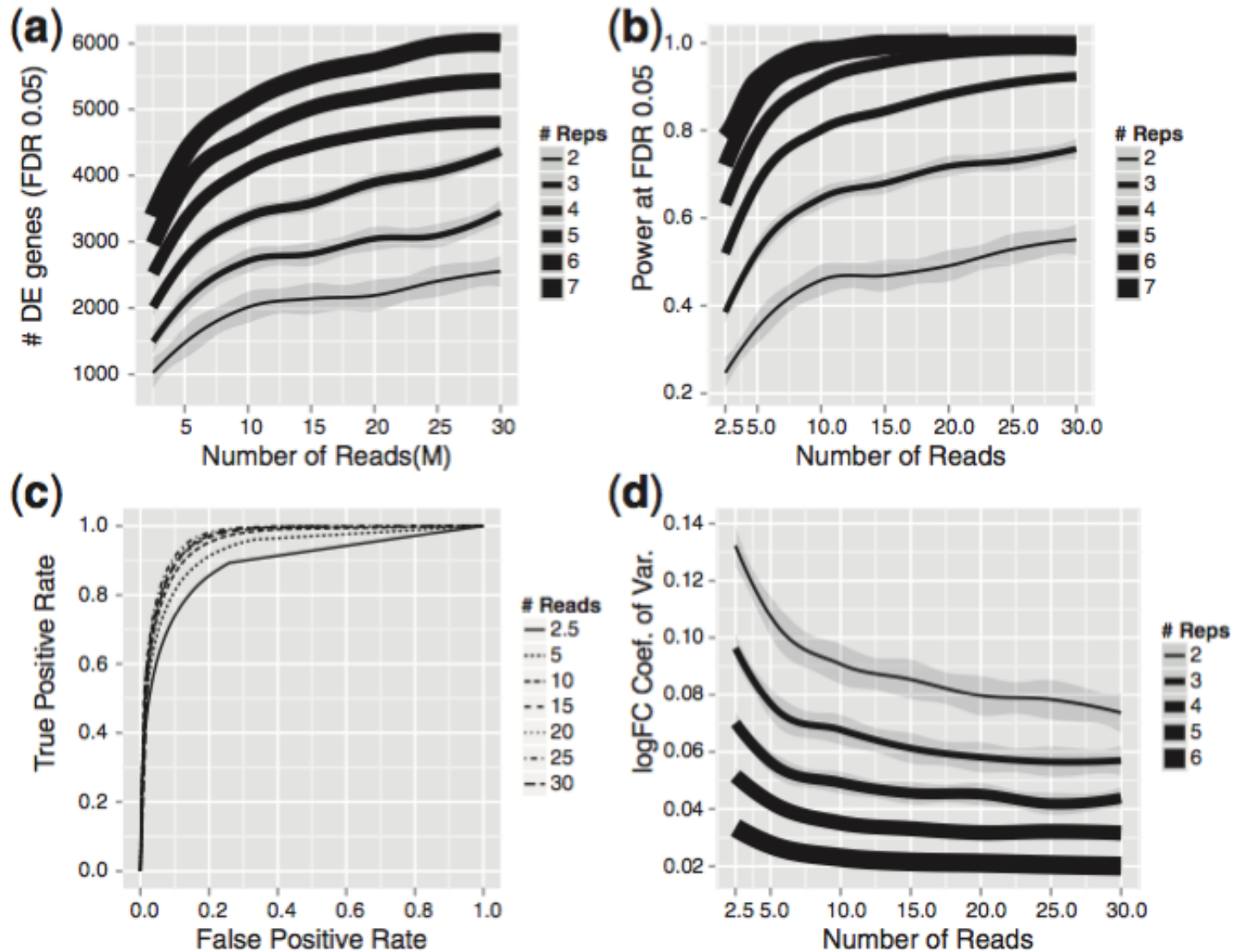


Filtering

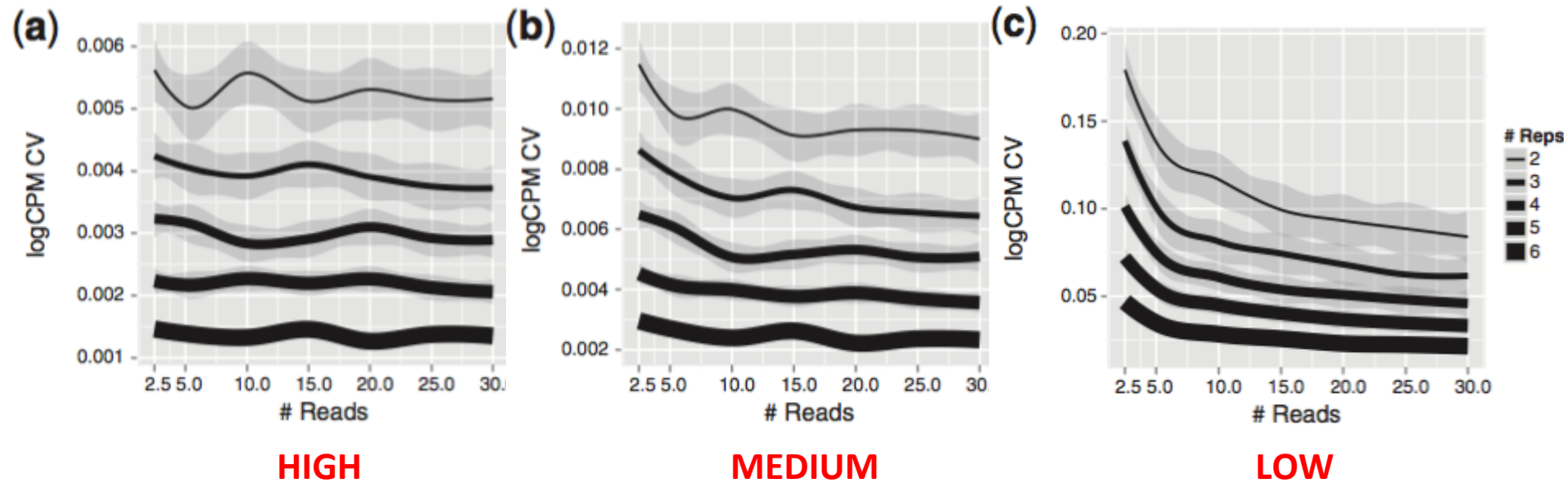
- *Independent filtering* = remove genes that have little chance of showing DE
- Can use eg. total count



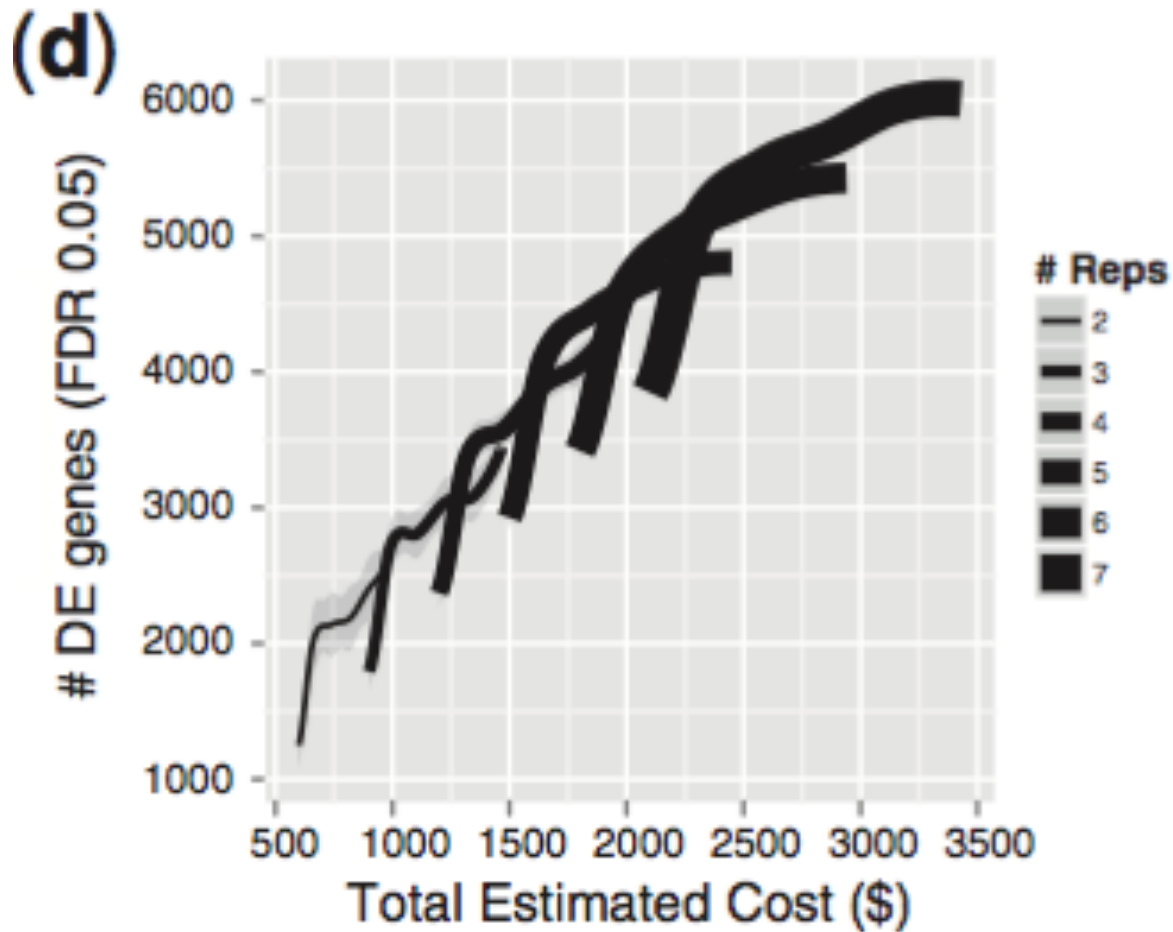
On replicates...



On replicates...



On replicates...



Summary

