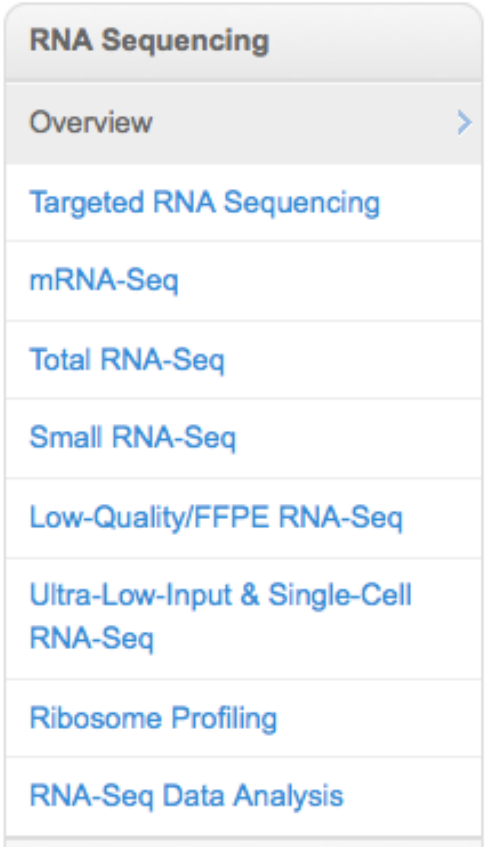
Analysis of RNA-seq Data

Bernard Pereira

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

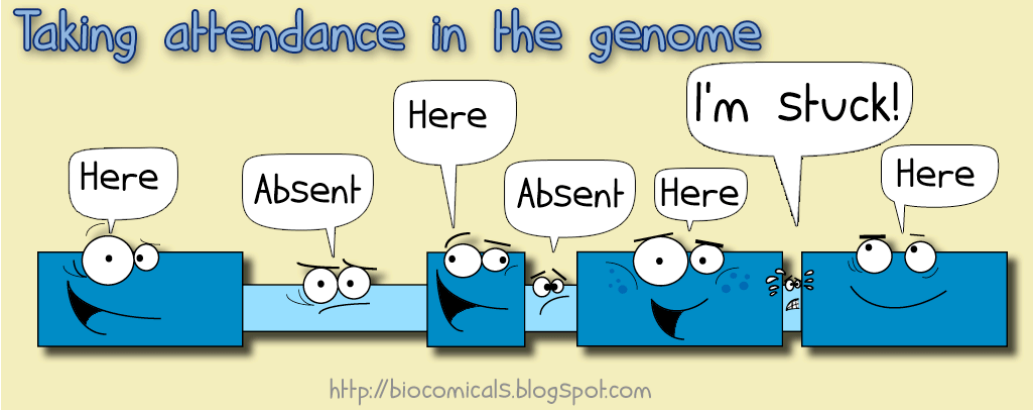
The many faces of RNA-seq



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

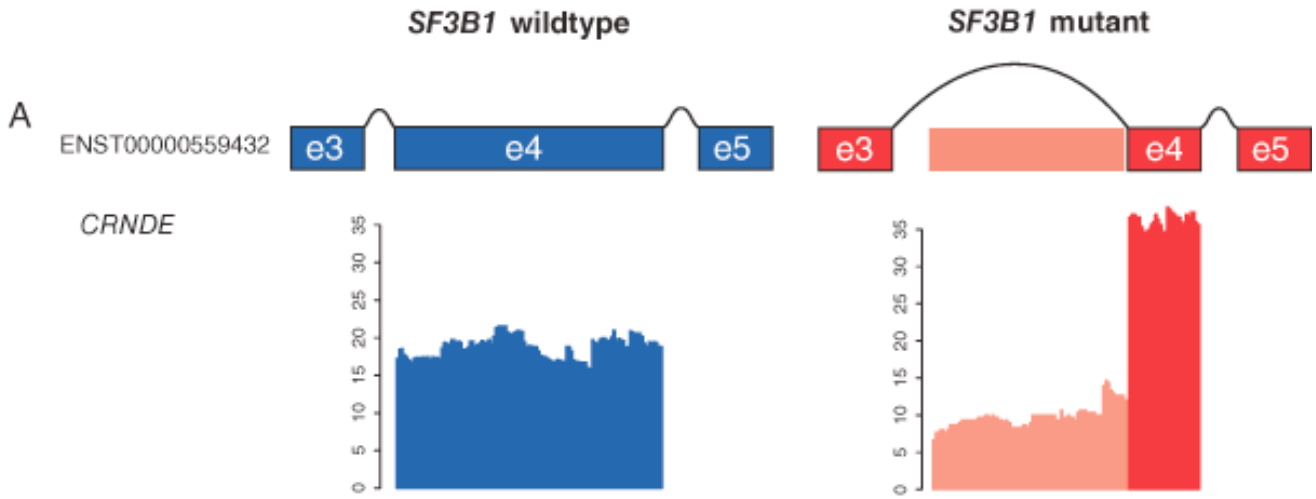
Applications

Journal of Pathology
J Pathol 2015; **235**: 571–580
Published online 22 December 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.4483

ORIGINAL PAPER

SF3B1 mutations constitute a novel therapeutic target in breast cancer

Sarah L Maguire,[1,†] Andri Leonidou,[1,2,†] Patty Wai,[1,2,†] Caterina Marchiò,[2,3] Charlotte KY Ng,[3,4] Anna Sapino,[2] Anne-Vincent Salomon,[5,6] Jorge S Reis-Filho,[3,4] Britta Weigelt[3,4] and Rachael C Natrajan[1,2,*]



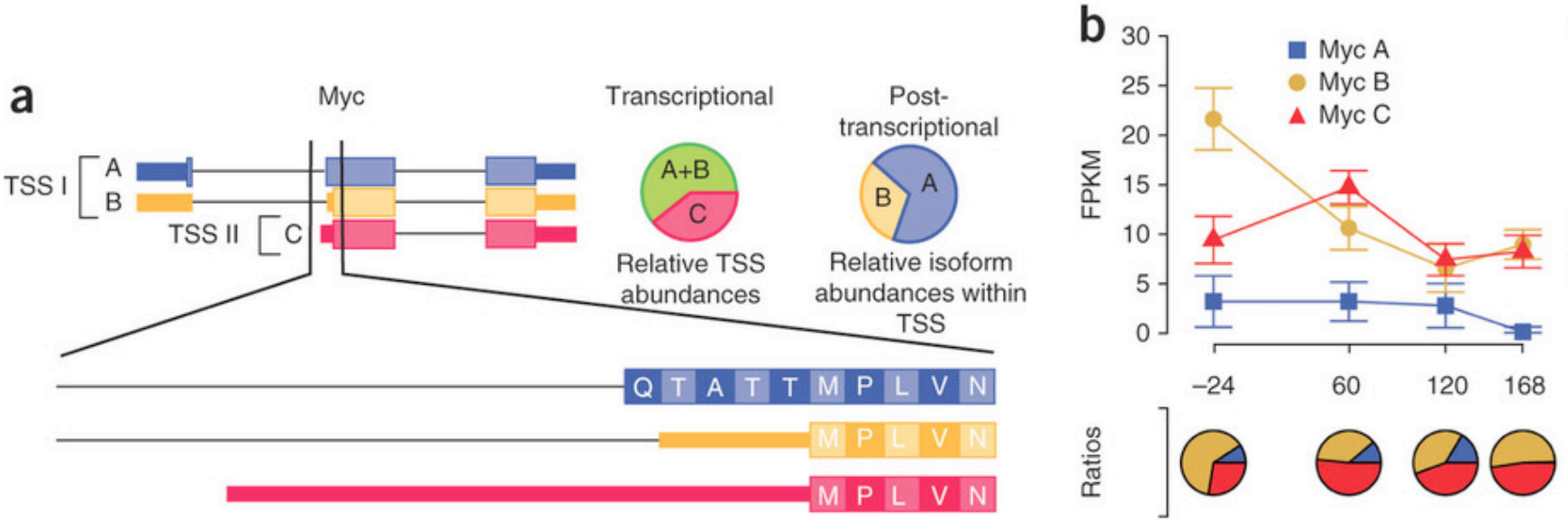
Applications

nature biotechnology

LETTERS

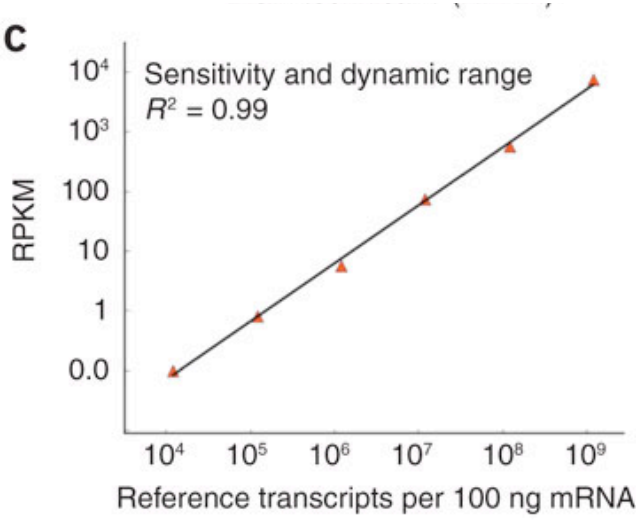
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Cole Trapnell[1–3], Brian A Williams[4], Geo Pertea[2], Ali Mortazavi[4], Gordon Kwan[4], Marijke J van Baren[5], Steven L Salzberg[1,2], Barbara J Wold[4] & Lior Pachter[3,6,7]



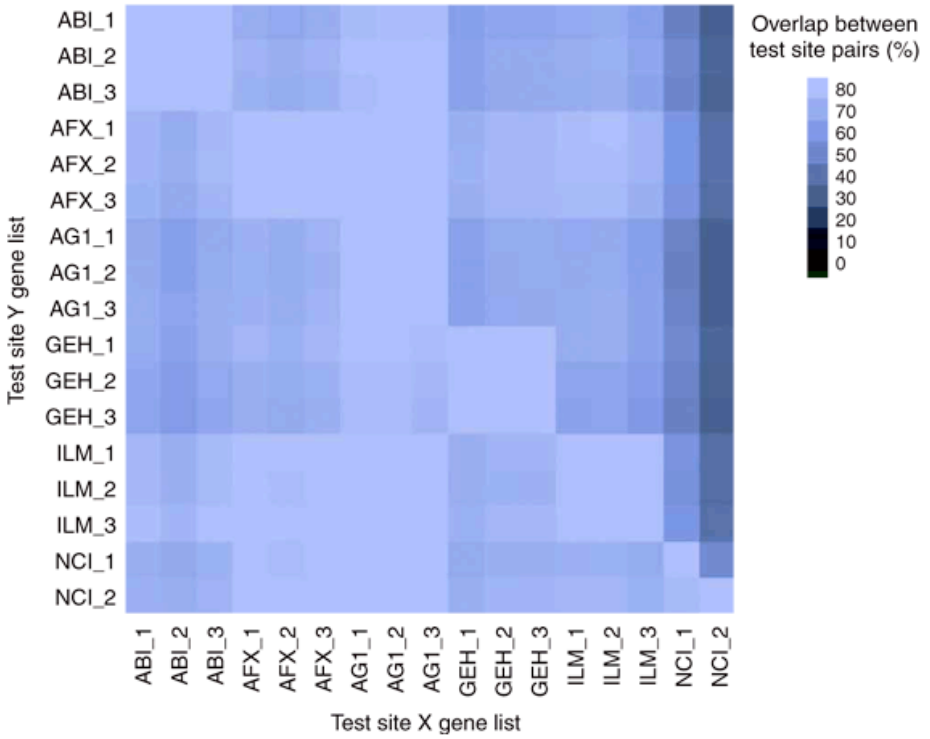
Differential Expression

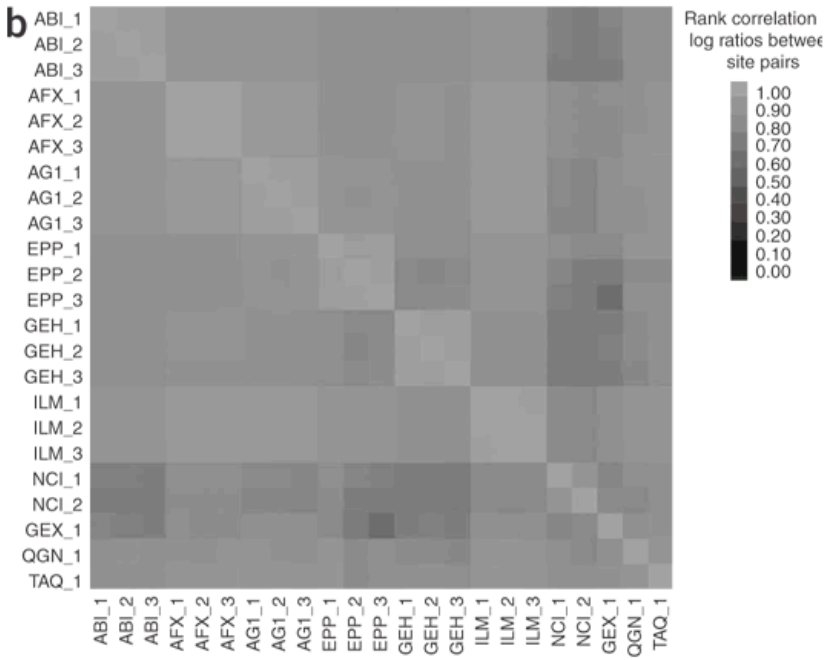
- Comparing feature abundance under different conditions
- Assumes linearity of signal
- When *feature=gene*, well-established pre- and post-analysis strategies exist



Mortazavi, A. et al (2008) *Nature Methods*

The good old days…

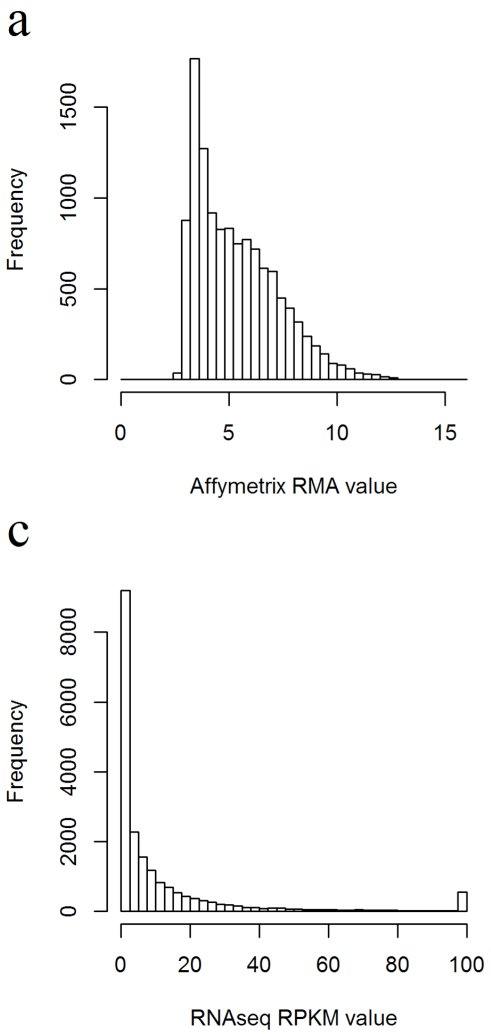




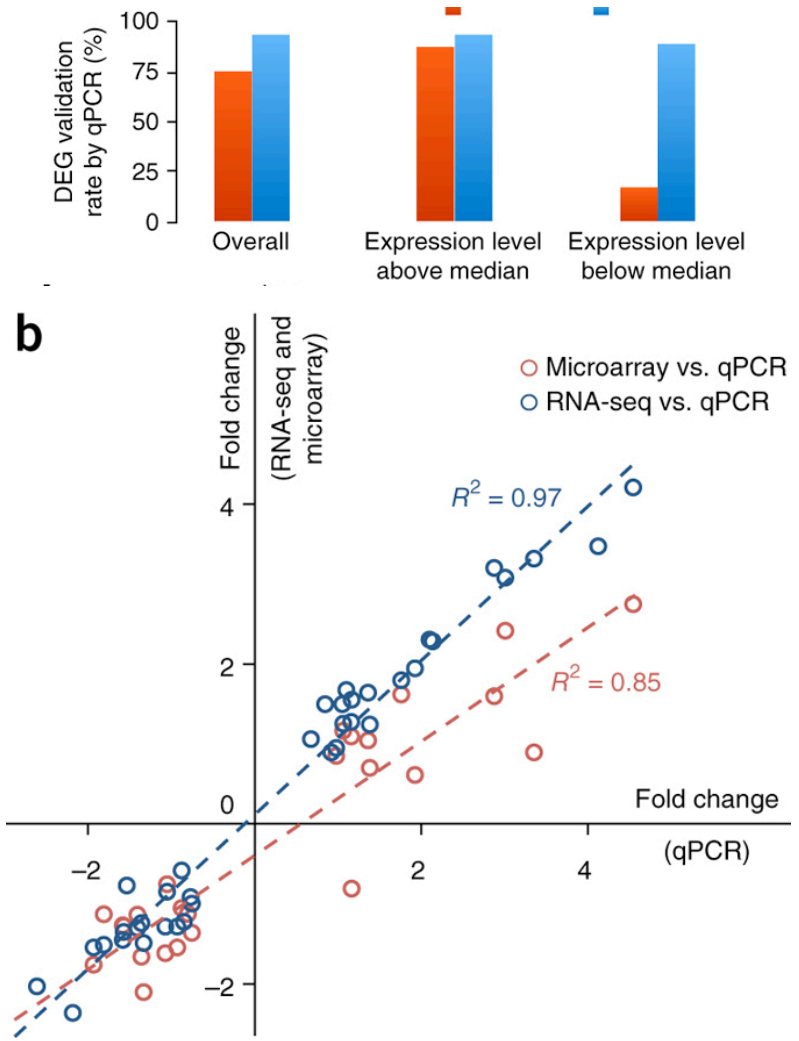
- The MicroArray Quality Control Project (led by the FDA)

MAQC Consortium (2006) *Nature Biotech.*

Range of detection

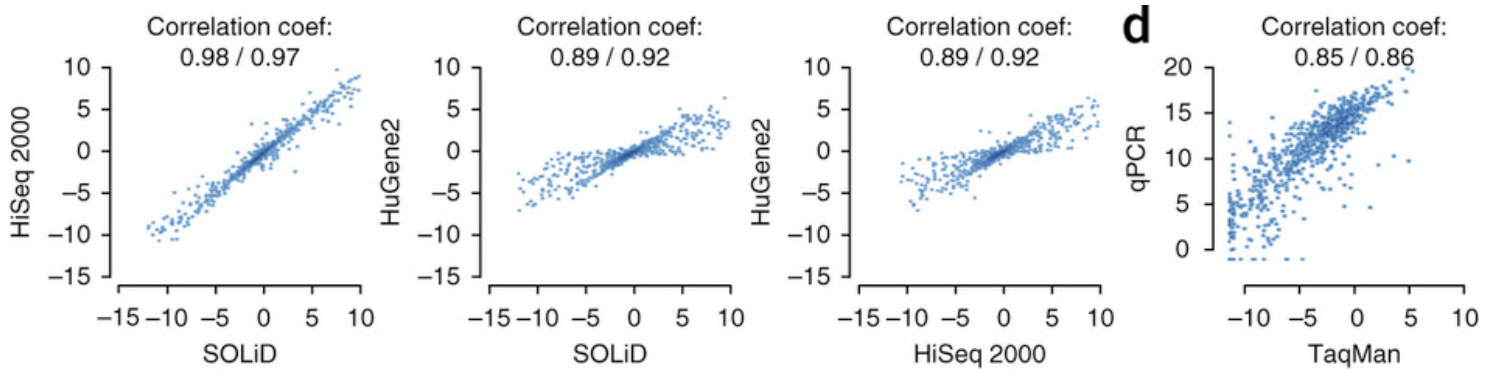


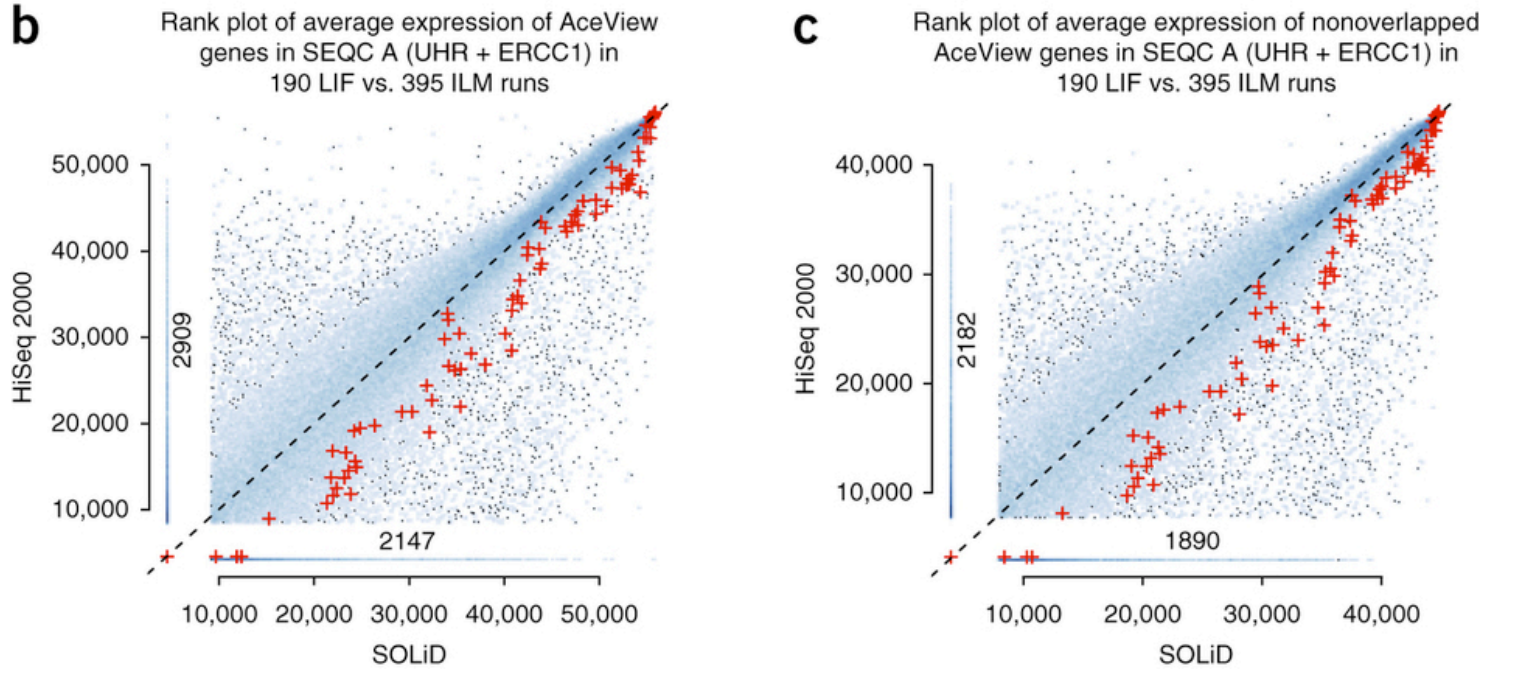
Guo et al. (2013) *Plos One*



Wang et al (2014) *Nature Biotech.*

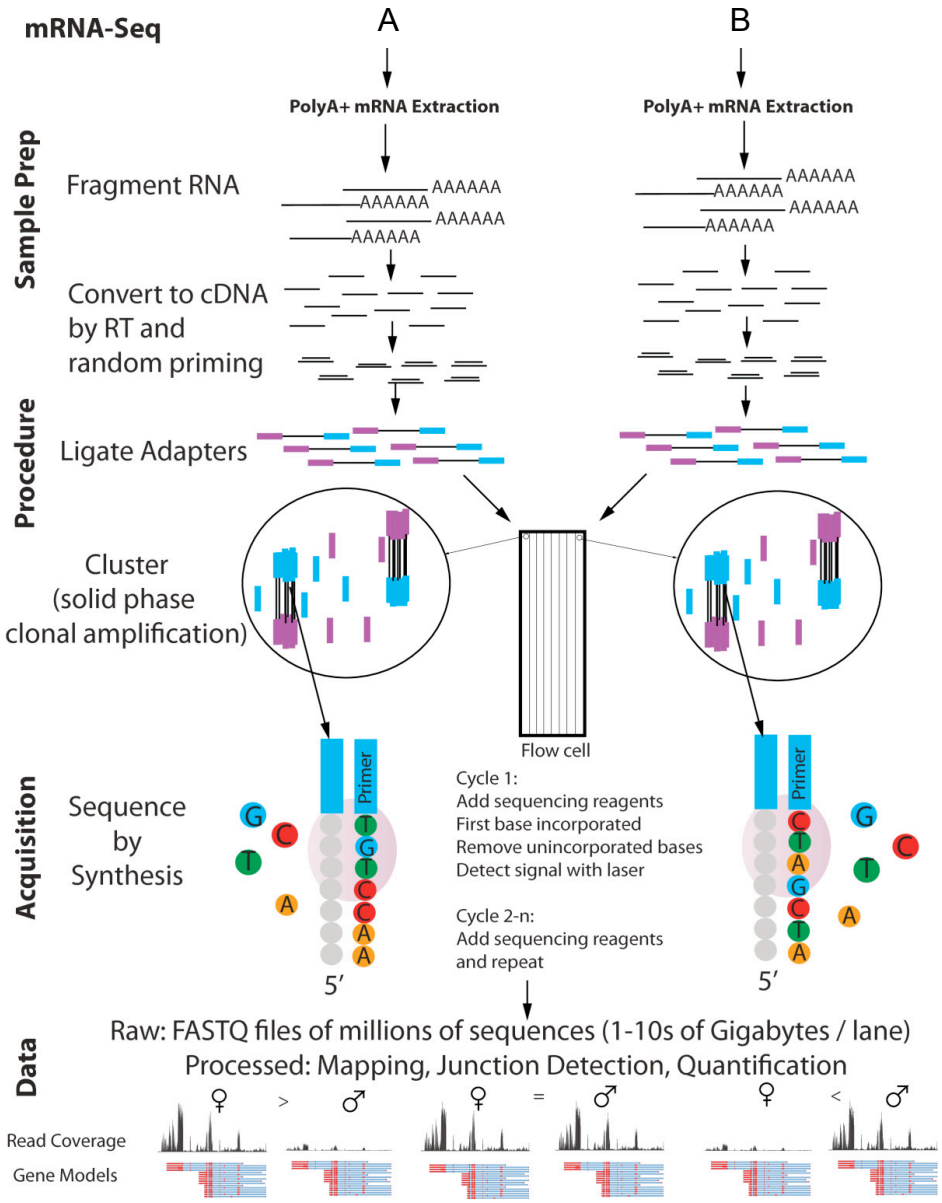
Into the future…





SEQC/MAQC-III Consortium (2014) *Nature Biotech.*

Library Prep i



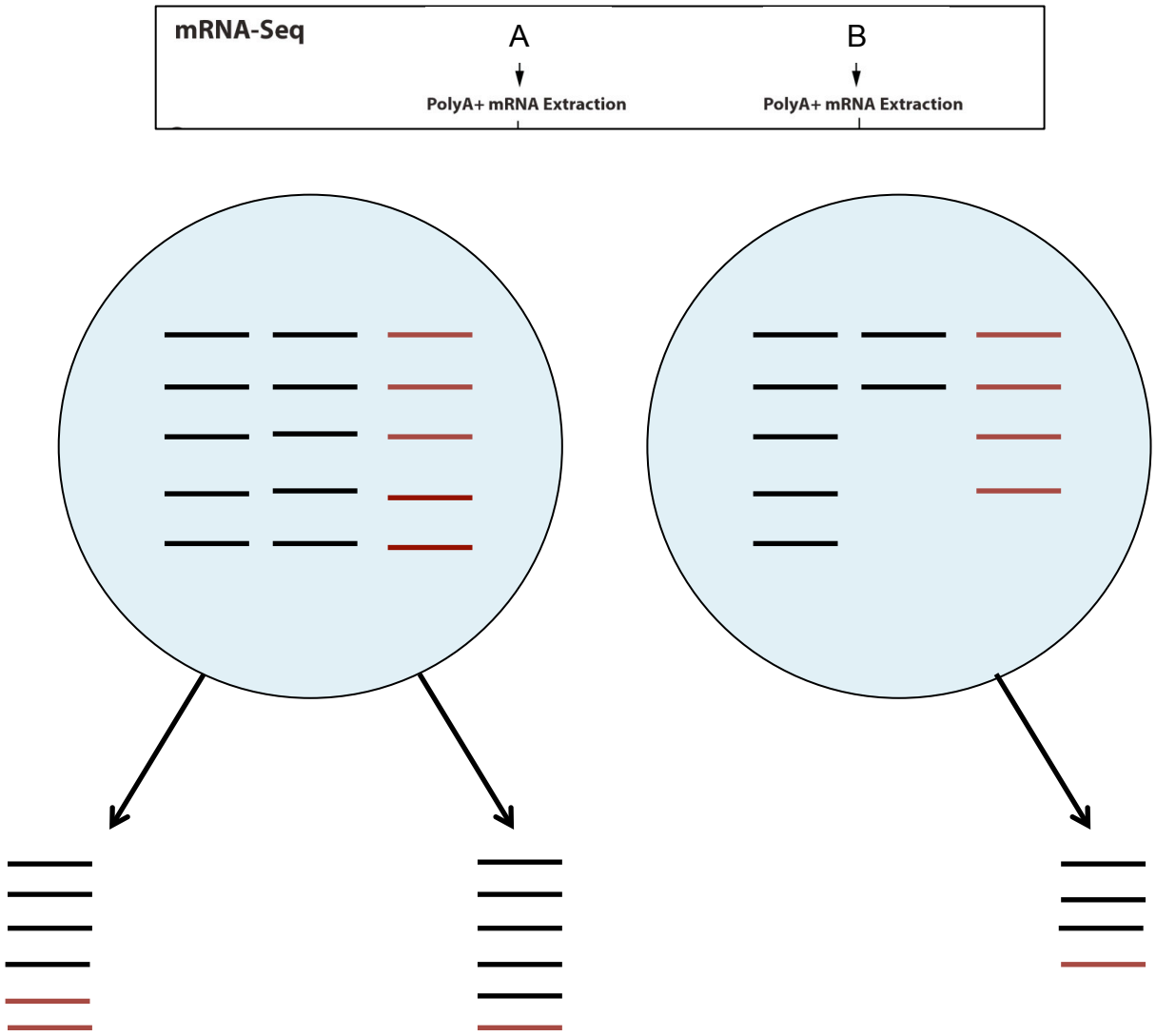
Malone, J.H. & Oliver, B.

Library Prep ii

Biological

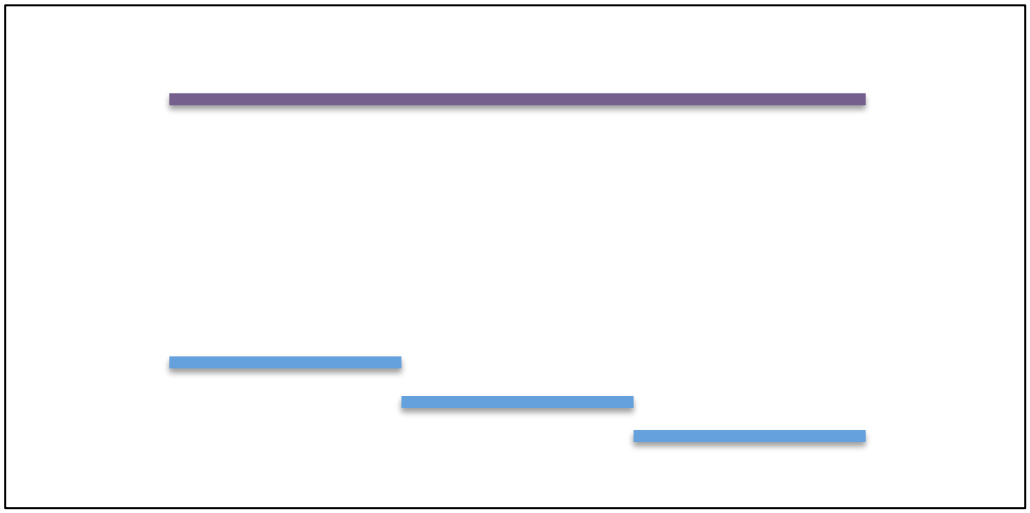
Technical

Library Prep iii

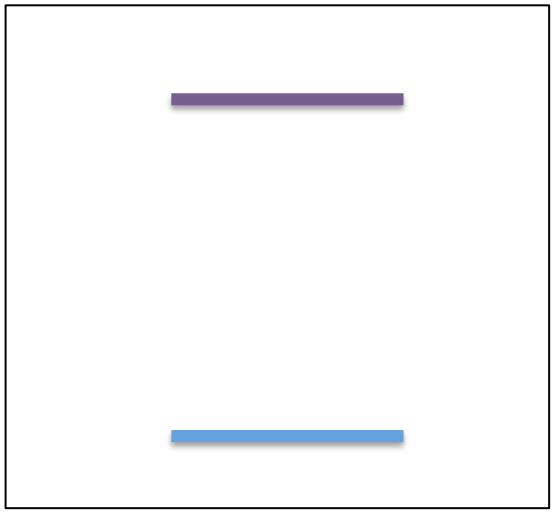


Library Prep iii

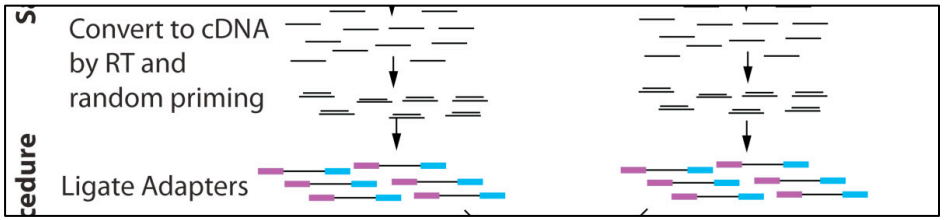
A

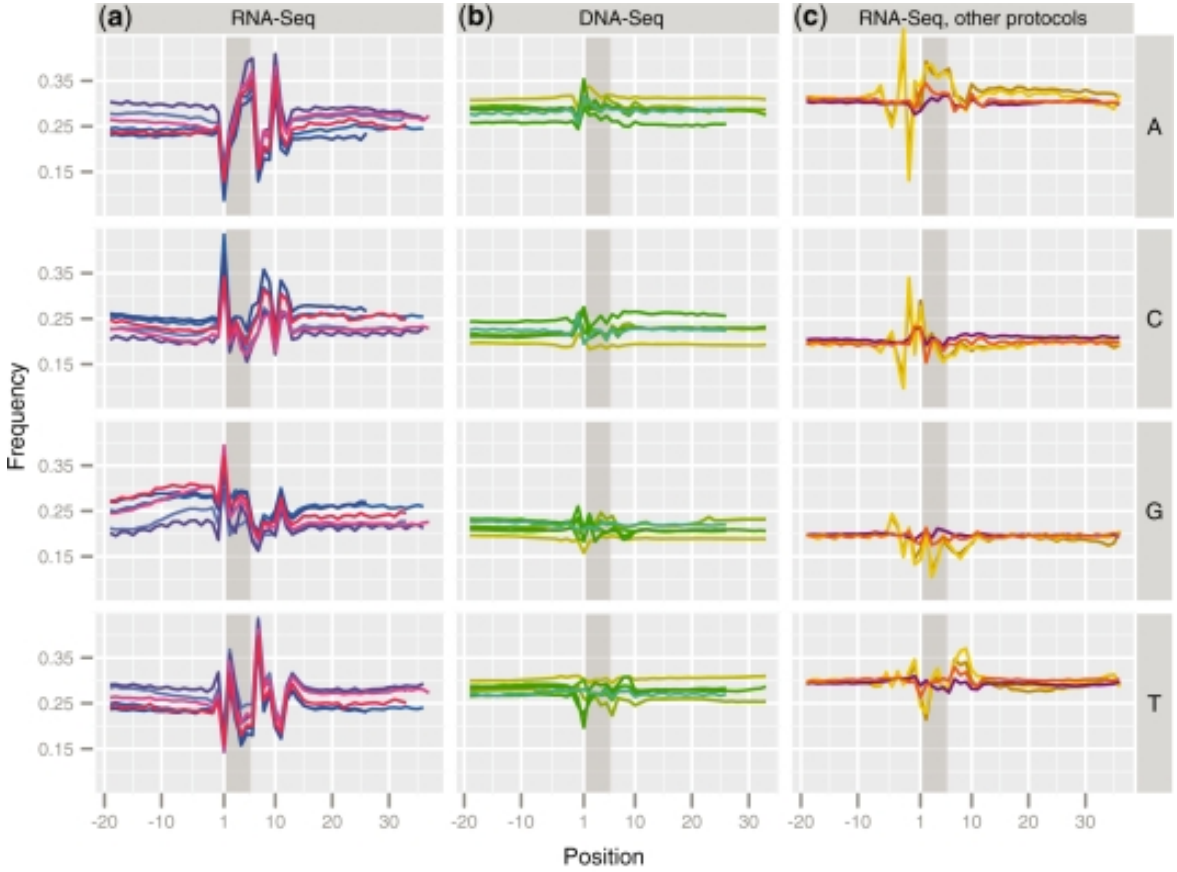


B



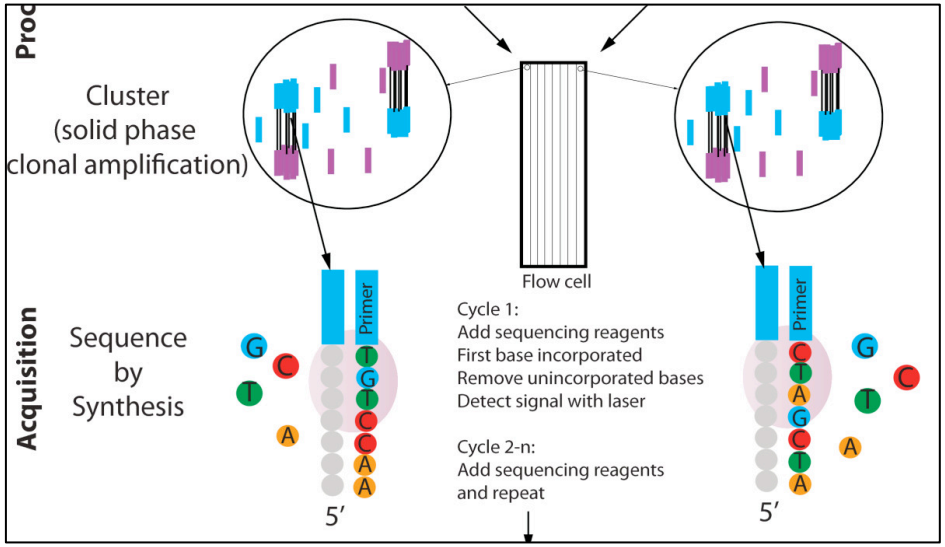
Library Prep iv





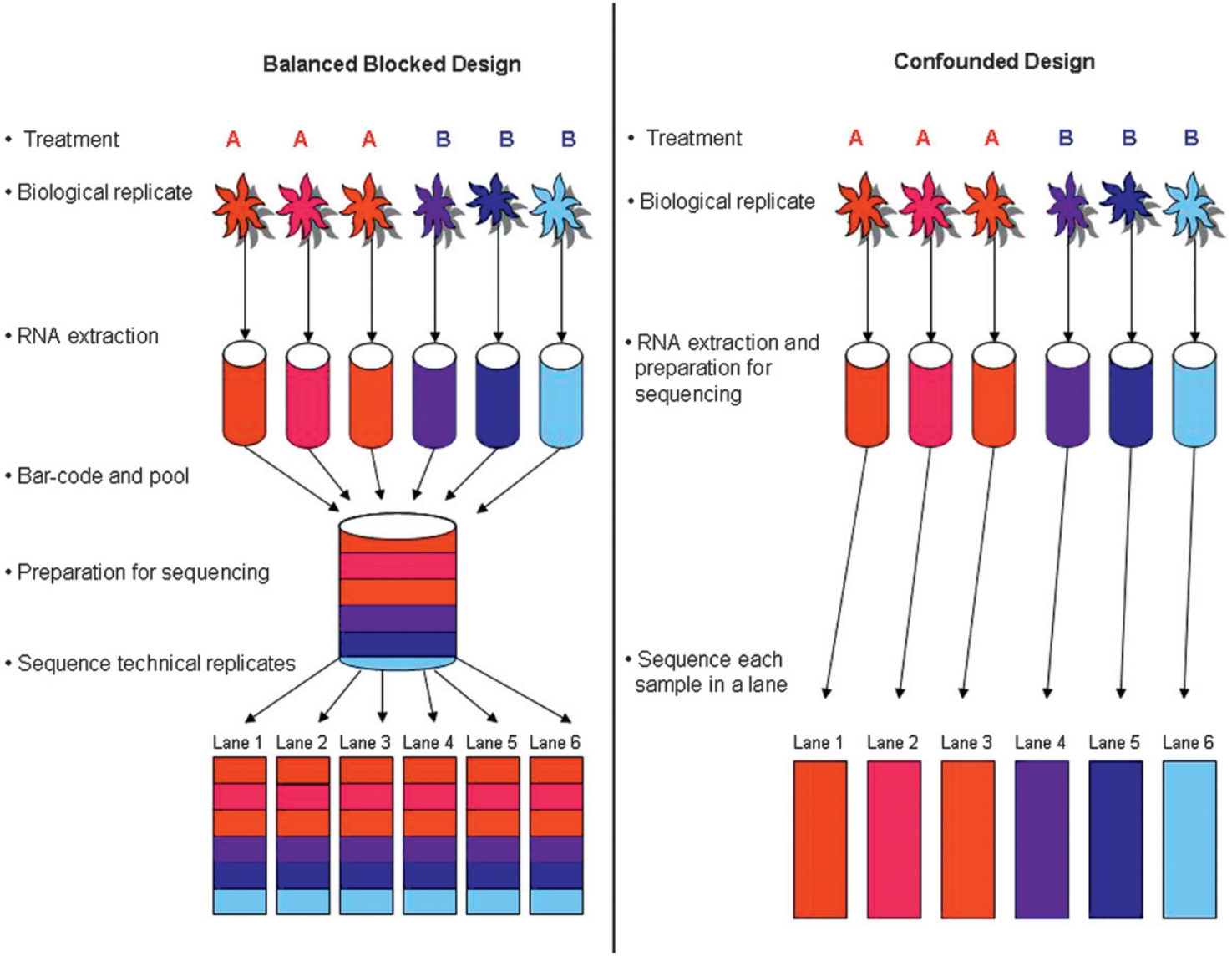
Hansen, K.D. et al. (2010) *Nuc. Acids Res.*

Library Prep v



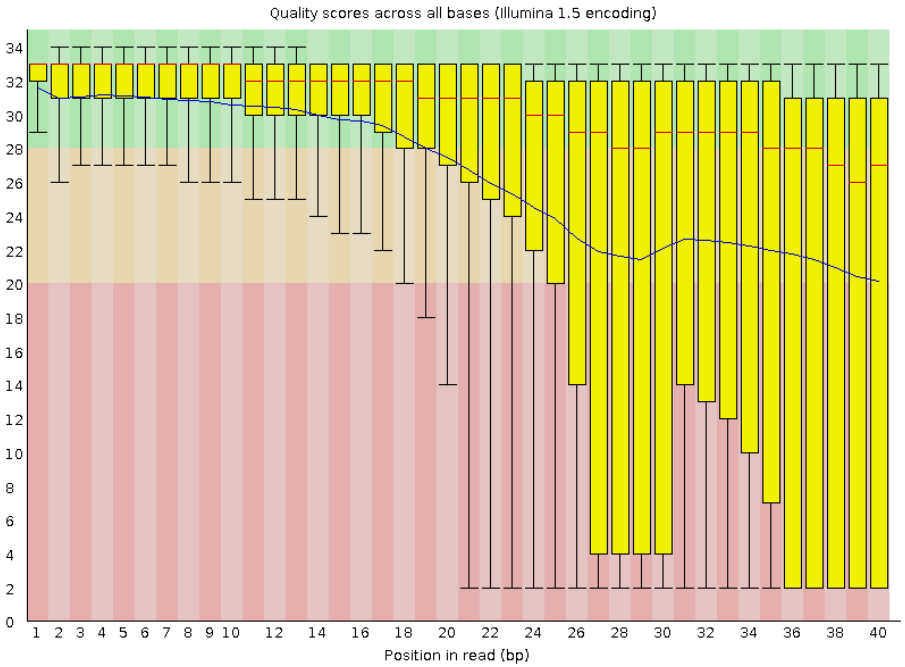
- Duplicates (optical & PCR)
- Sequence errors
- Indels
- Repetitive/problematic sequence

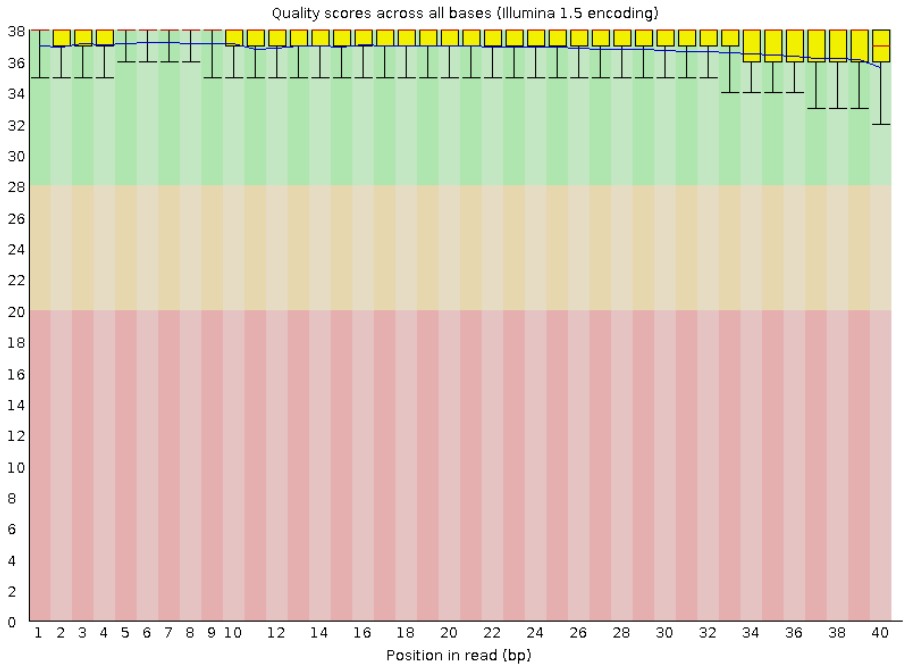
Hot off the sequencer…



Auer and Doerg (2010) *Genetics*

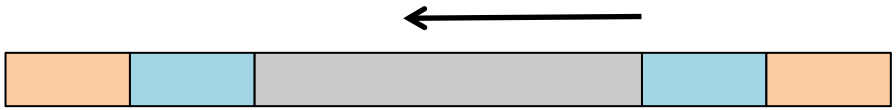
FASTQC

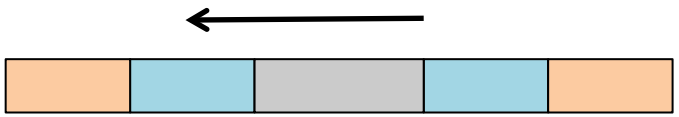

Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)


Quality scores across all bases (Illumina 1.5 encoding)
Position in read (bp)

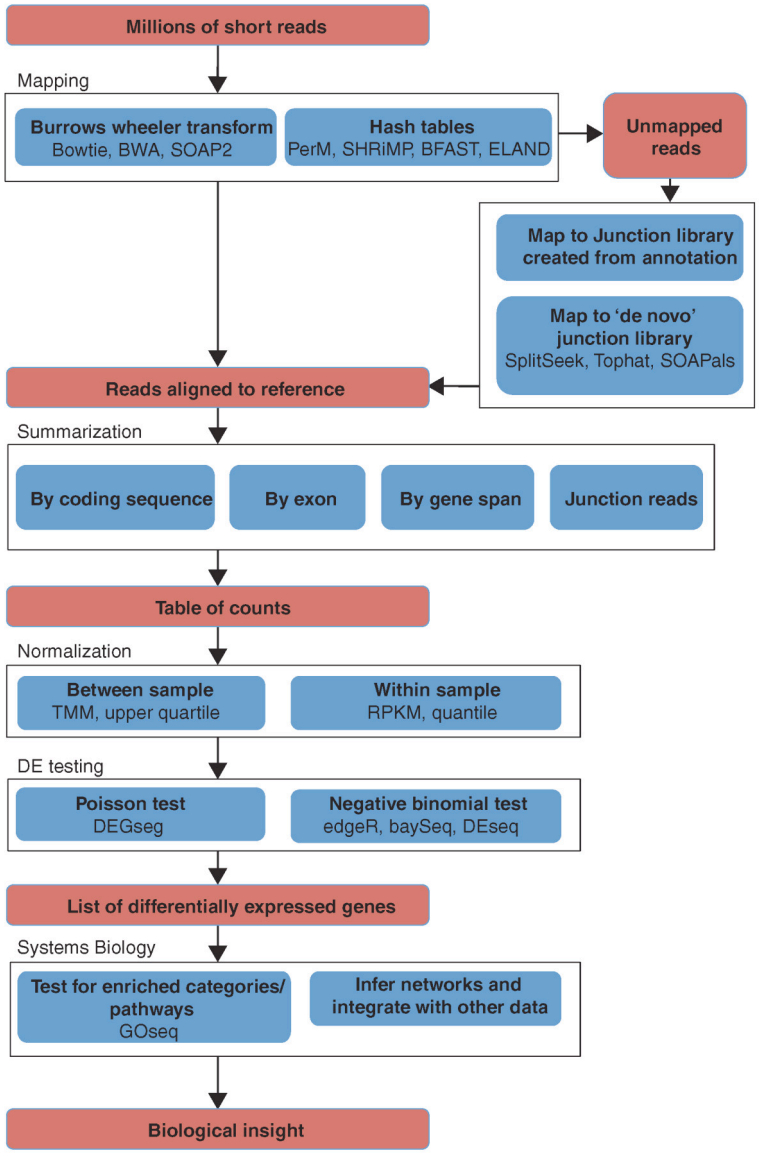
Trimming

- Quality-based trimming
- Adapter ‘contamination’



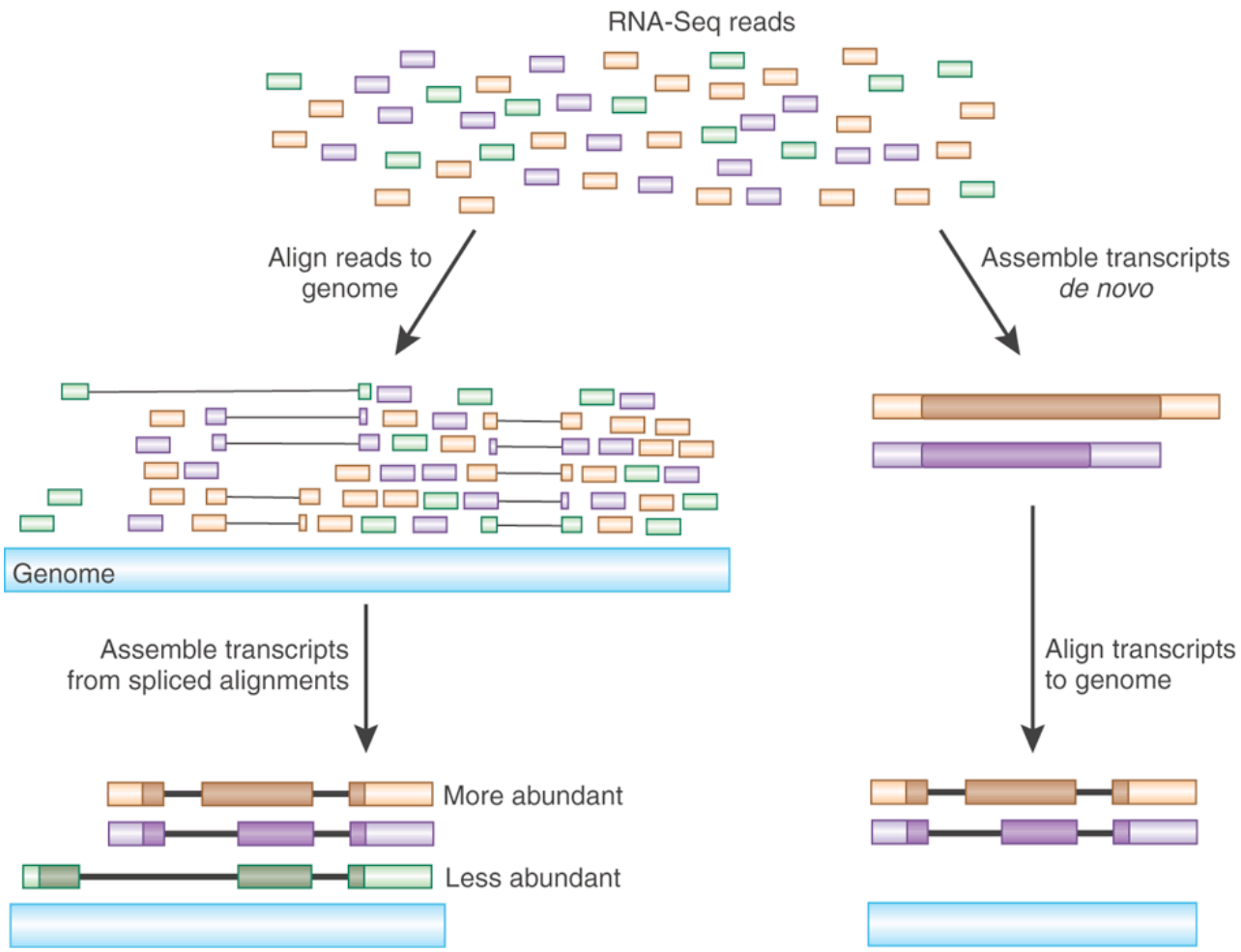


Analysis overview



Oshlack, A. et al (2010) *Genome Biology*

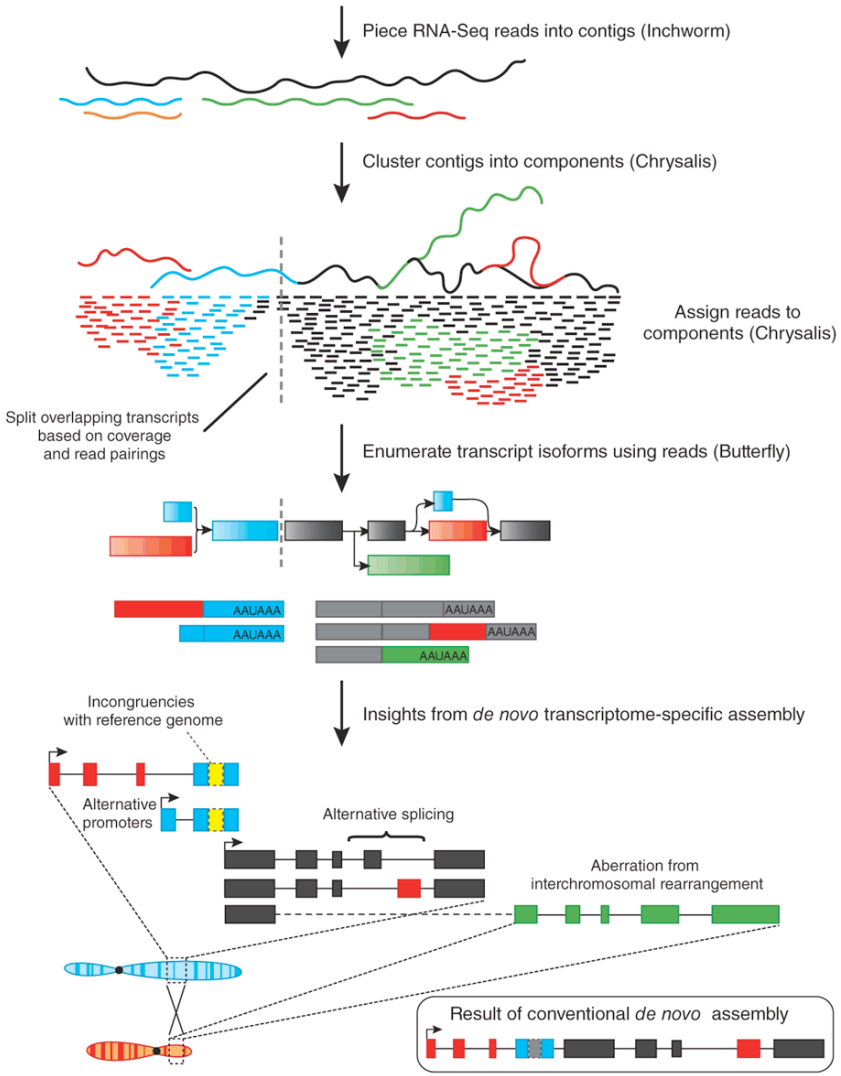
Sequence to sense



Haas, B.J. & Zody, M.C. (2010) *Nature Biotech.*

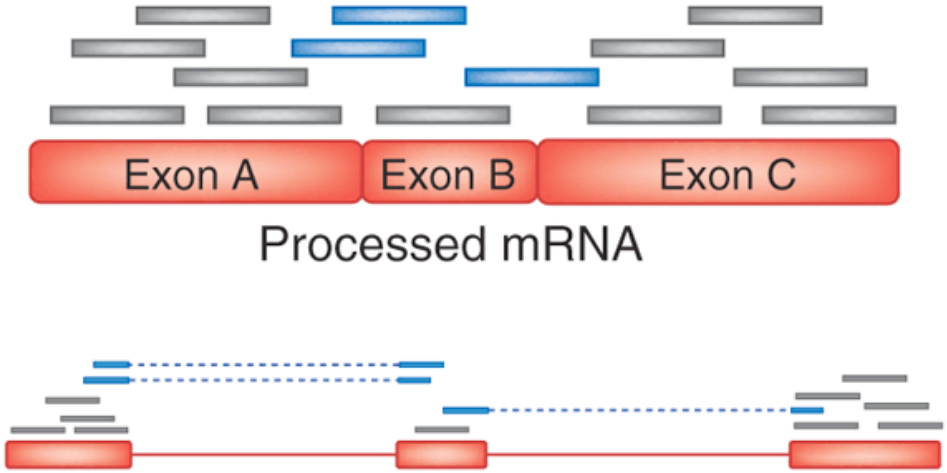
De novo assembly

- eg. Trinity



Haas, B.J.. et al (2013) *Nature Protocols*

Reference-based assembly



Genome mapping

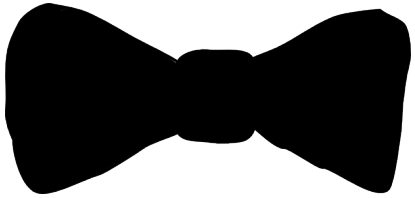
- Can identify novel features
- Spice aware?
- Can be difficult to reconstruct isoform and gene structures

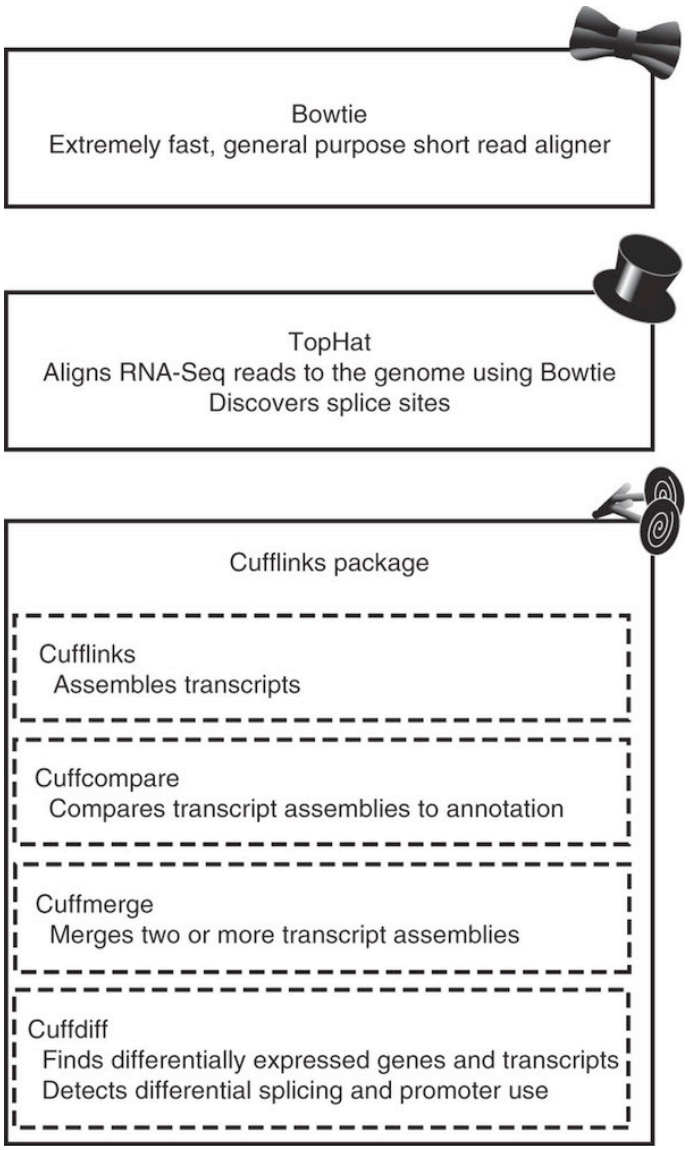
Transcriptome mapping

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

Trapnell & Salzberg (2009) *Nature Biotech*

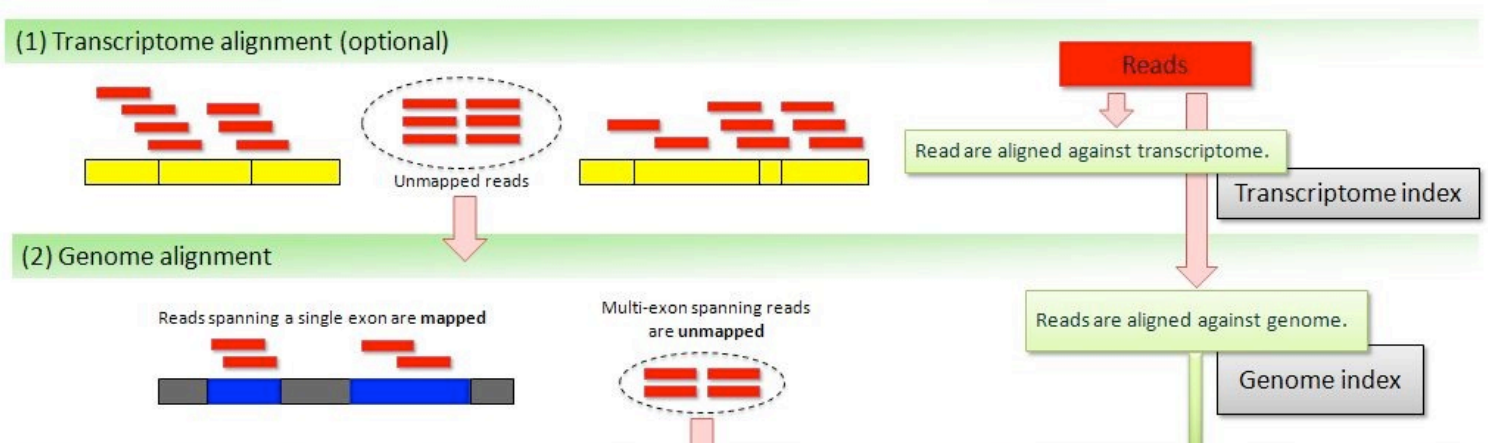
A smart suit(e)





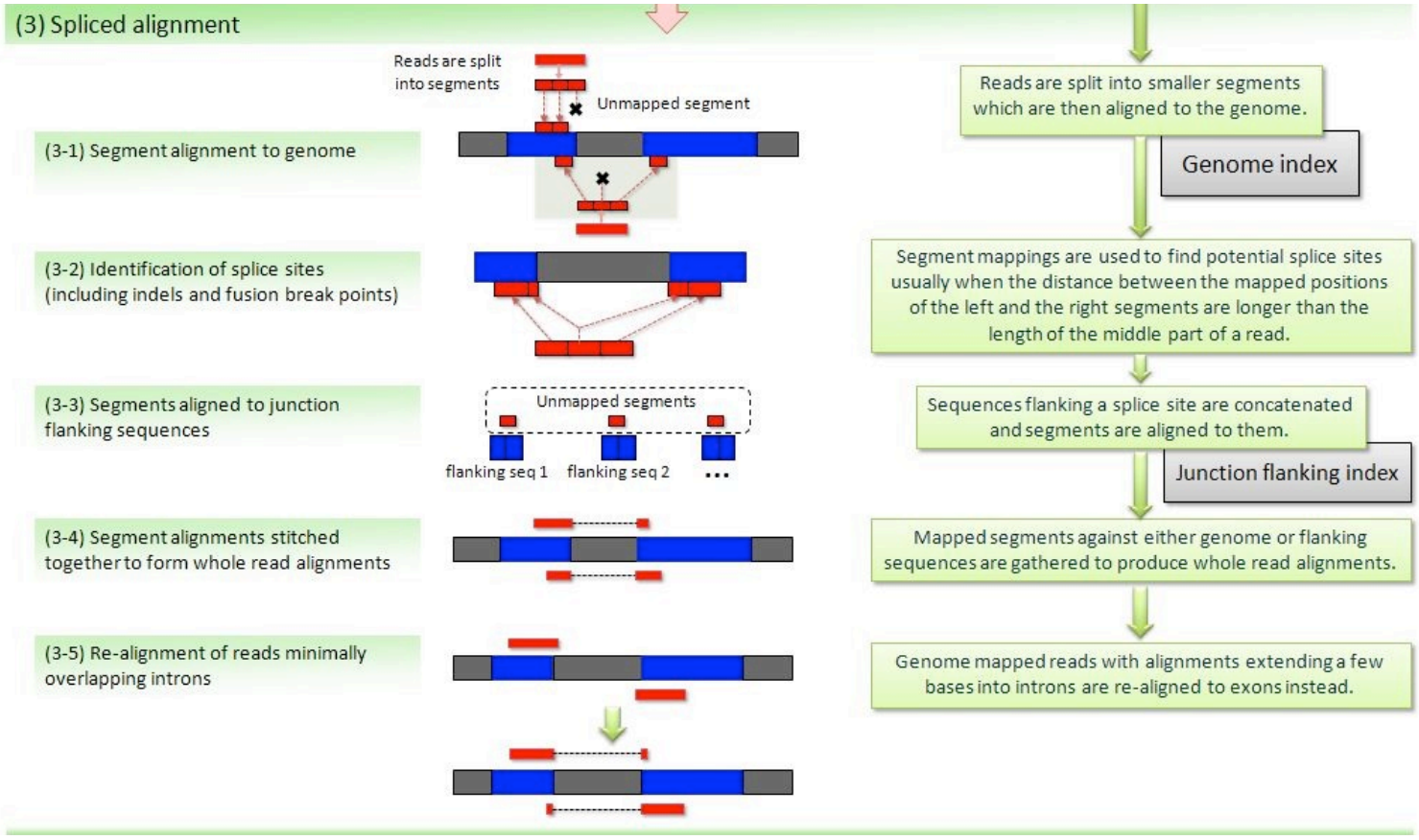
Trapnell, C. et al (2012) *Nature Protocols*

Tophat/Bowtie



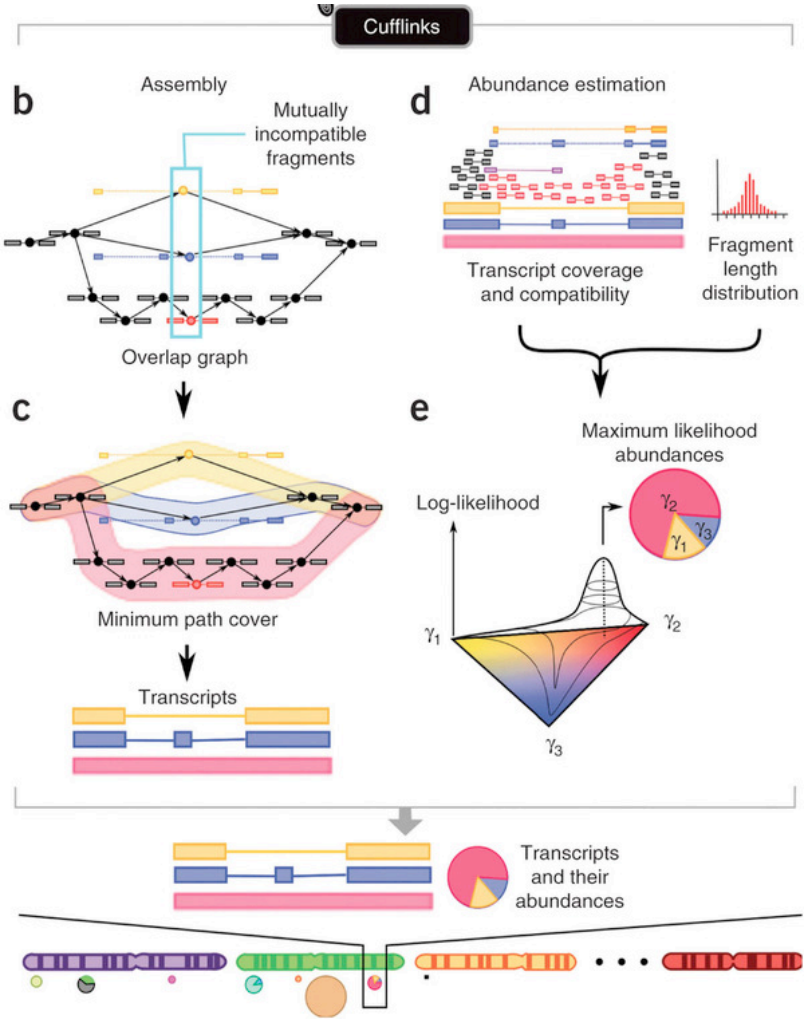
Kim, D. et al (2012) *Genome Biology*

Tophat/Bowtie



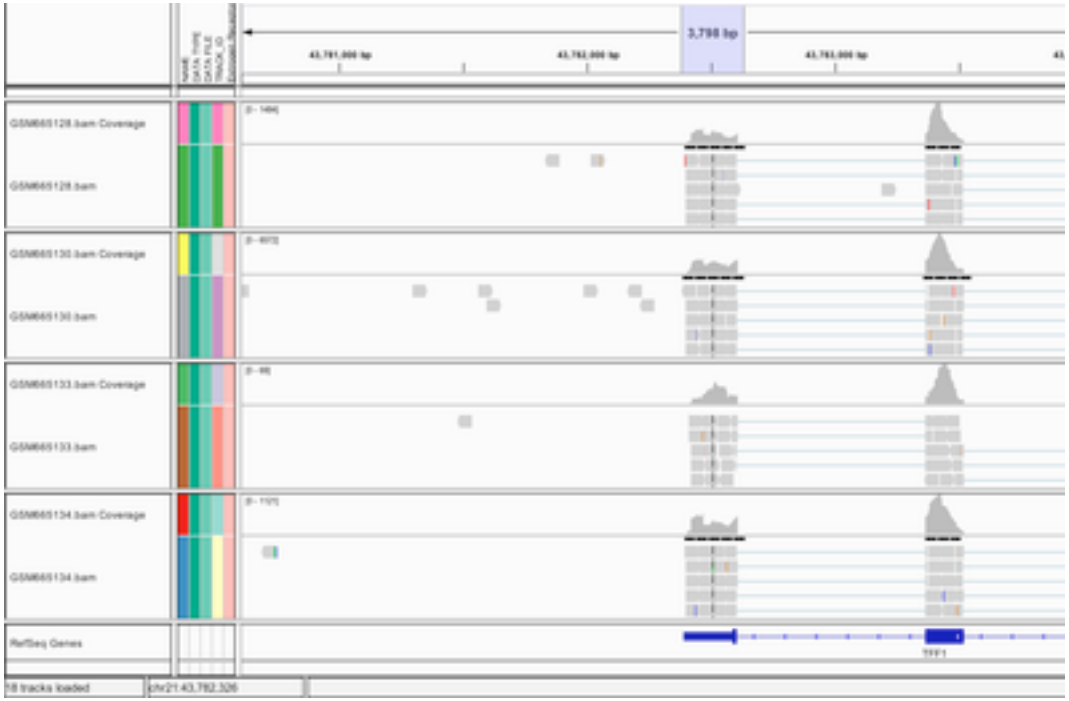
Kim, D. et al (2012) *Genome Biology*

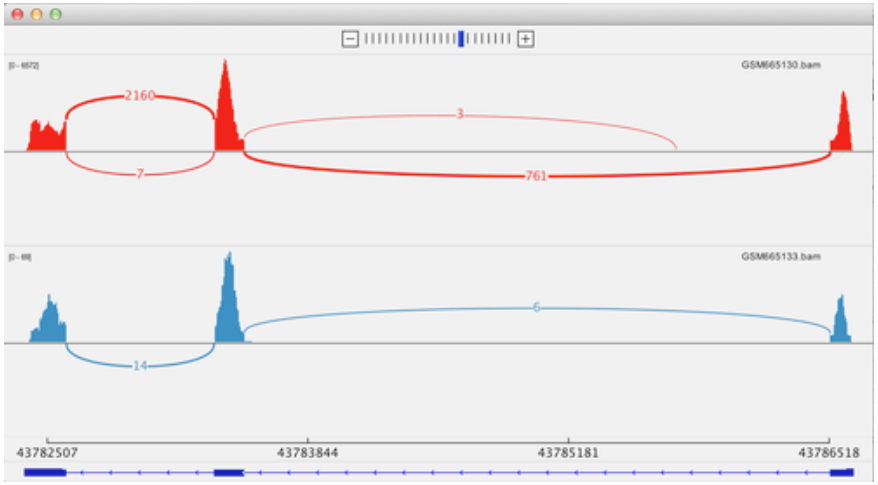
Cufflinks



Trapnell, C. et al. (2010) *Nature Biotech.*

How do we look?





Duplicates & RNA-seq

Intrinsically lower complexity

Highly expressed genes

Platform/pipeline

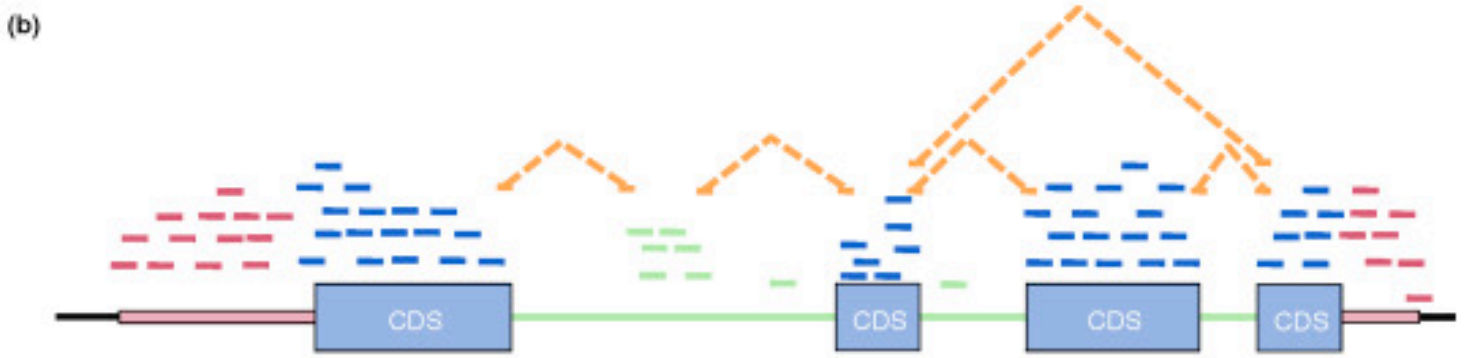
?

Variant calling vs DE analysis

Platform/pipeline

Single-end vs paired-end

Counting



Genome-based features

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

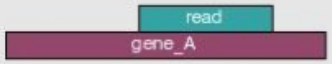
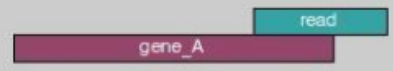
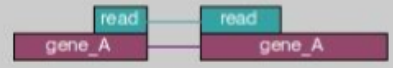
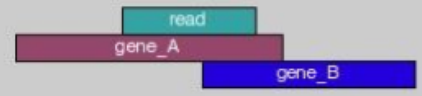
Transcript-based features

- Transcript assembly
- Novel structures
- Isoform multireads

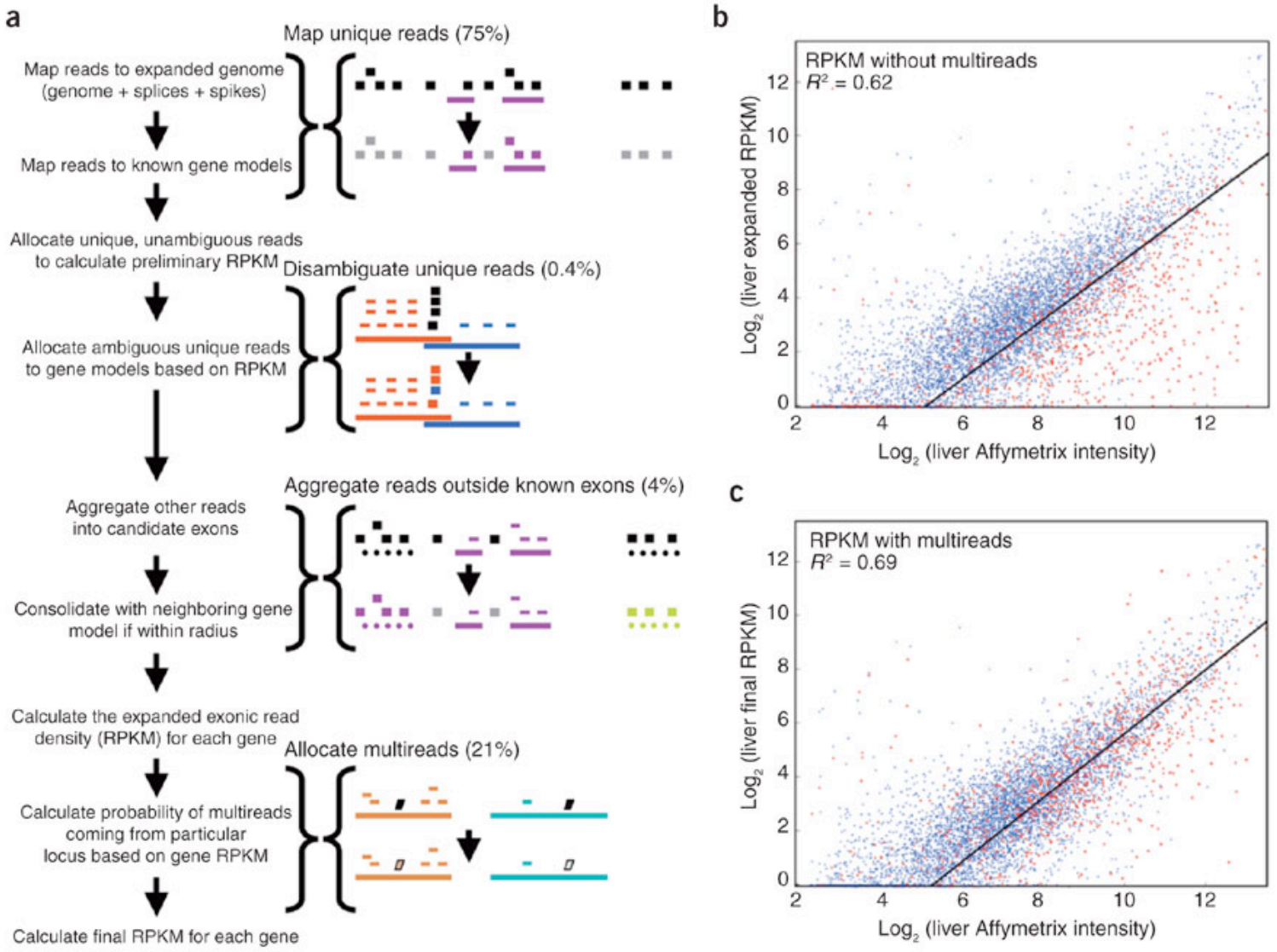
Oshlack, A. et al. (2010) *Genome Biology*

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



Mortazavi, A. et al (2008) *Nature Methods*

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

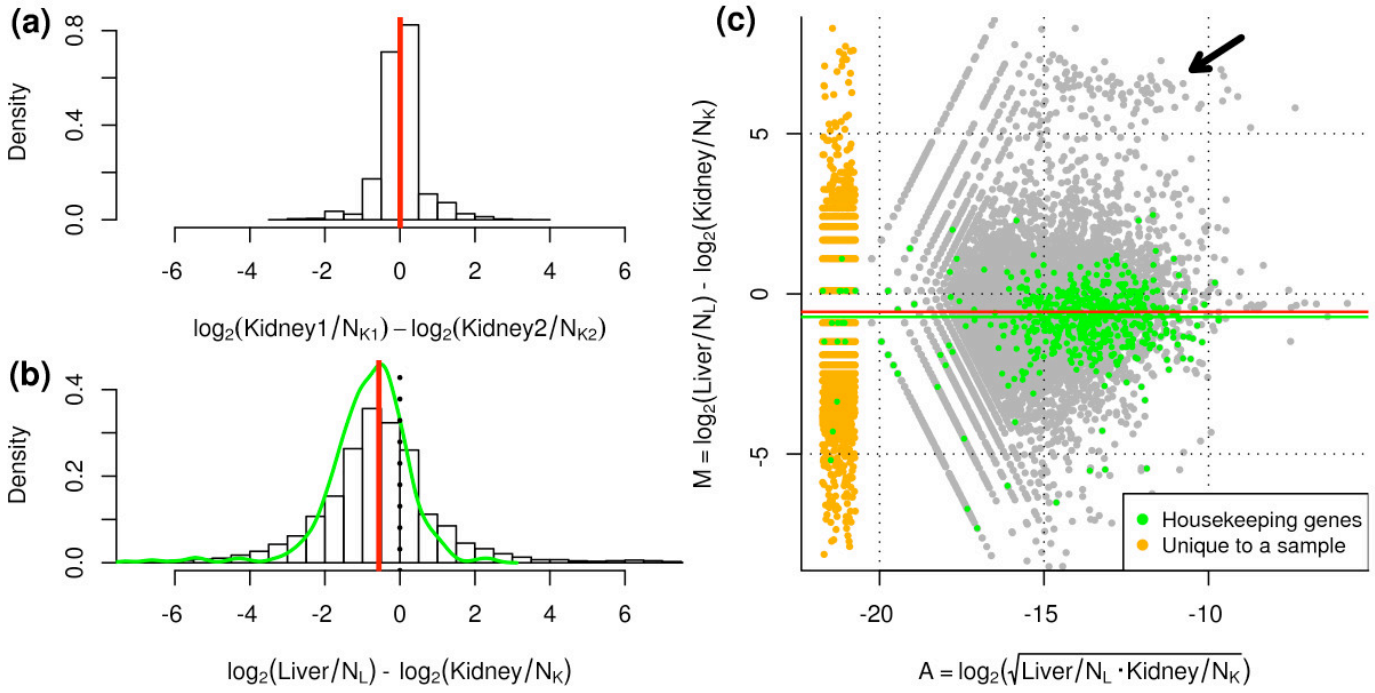
Library composition

- 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 tototal count; eg. median, upper quartile



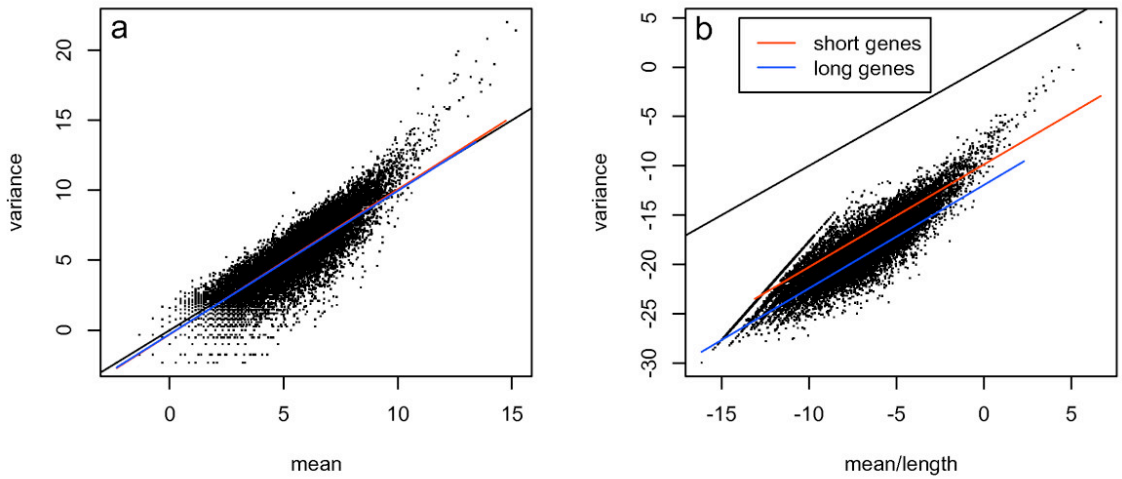
Robinson, M.D. & Oshlack, A. (2010) *Genome Biology*

Normalisation ii

RPKM

- Reads per kilobase per million =

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

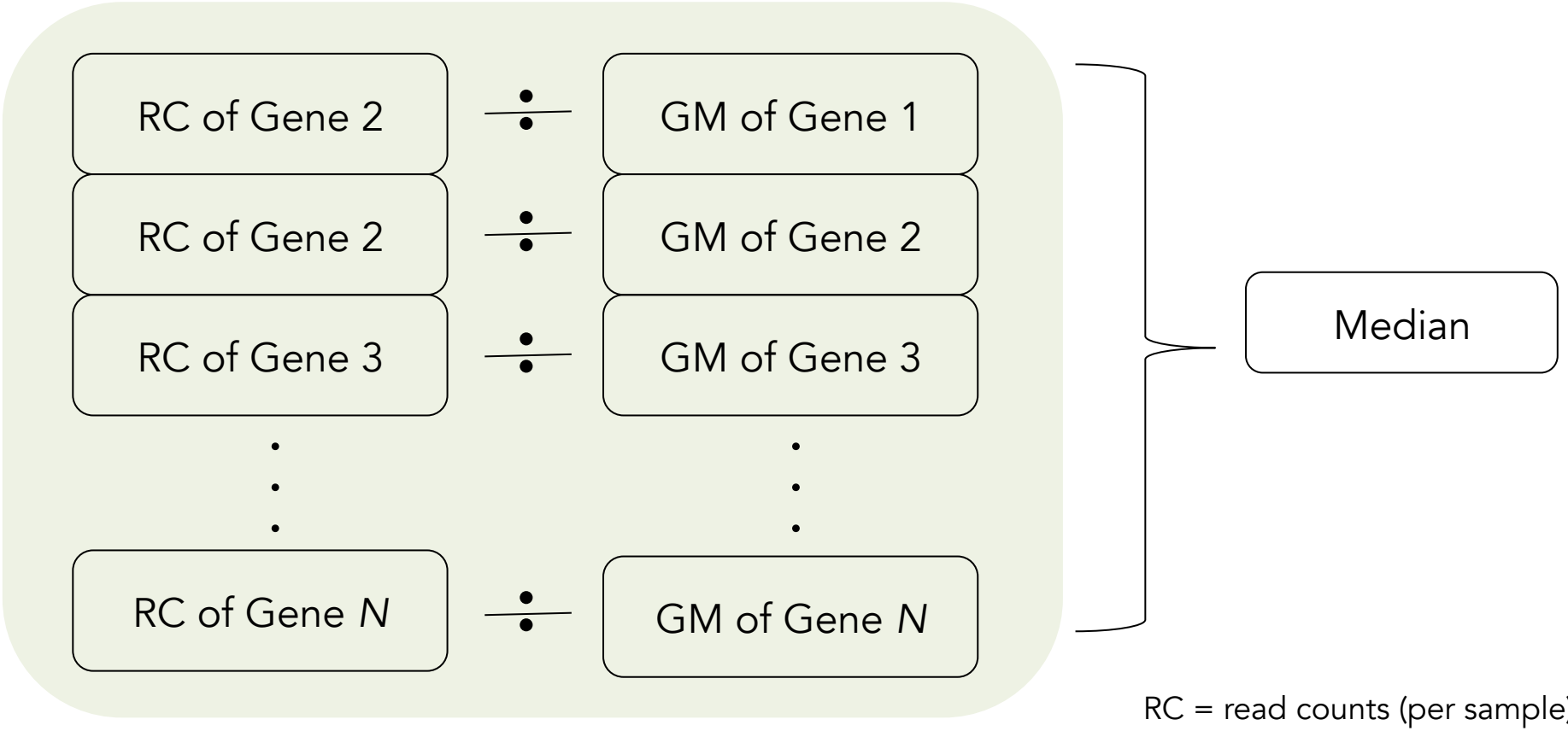


Oshlack, A. & Wakefield, M.J. (2009) *Biology Direct*

Normalisation iii

Geometric scaling factor

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

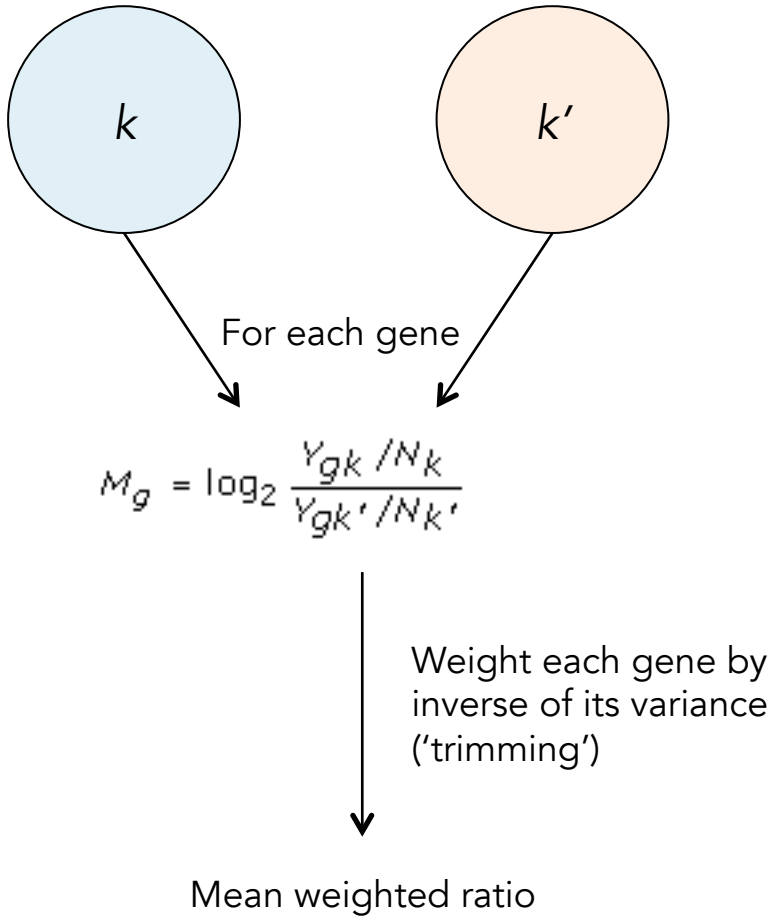


RC = read counts (per sample)
GM =geometric mean (all samples)

Normalisation iv

Trimmed mean of *M*

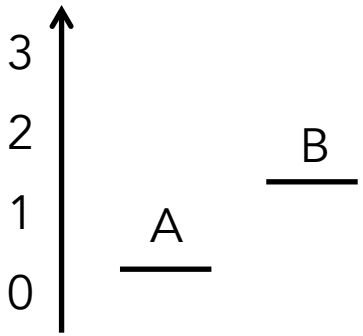
- Implemented in edgeR
- Assumes most genes are not differentially expressed



g = each gene

Differential expression

- Simple

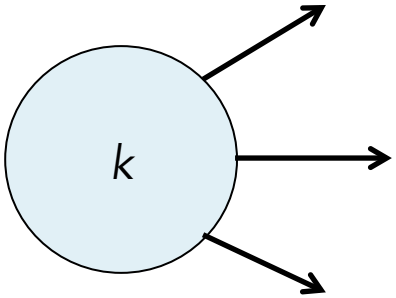


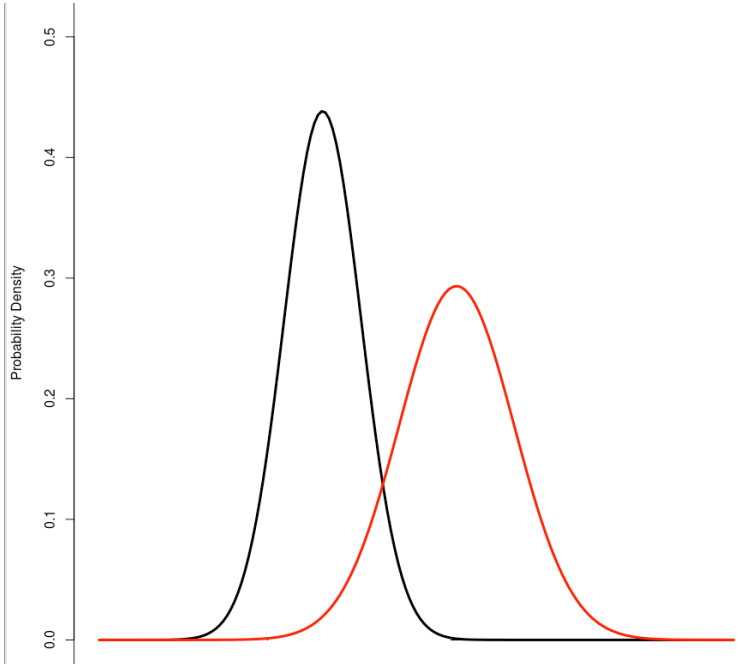
All we need

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

Modelling – old trends

- Technical replicates introduce some variance

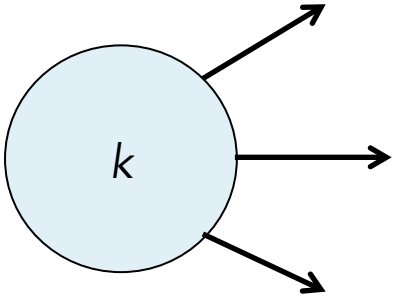


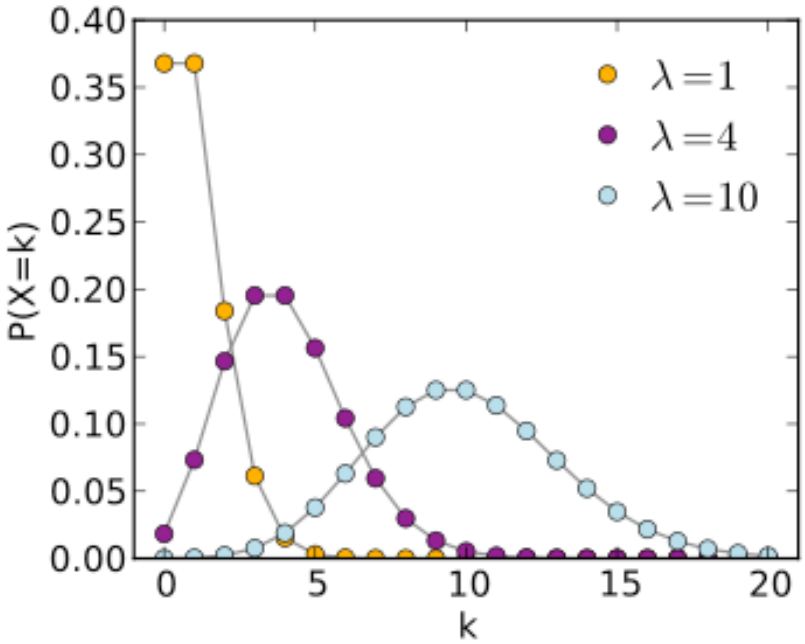


- What the data looks like: **normal distribution**
- Some measure of difference: ***t-test***

Modelling – in fashion

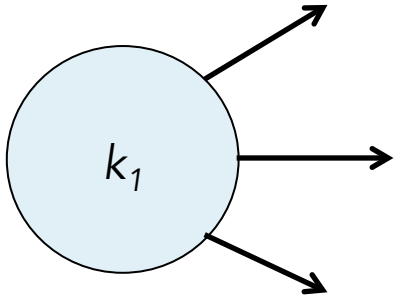
- Use the Poisson distribution for count data from technical replicates
- Just one parameter required – the mean

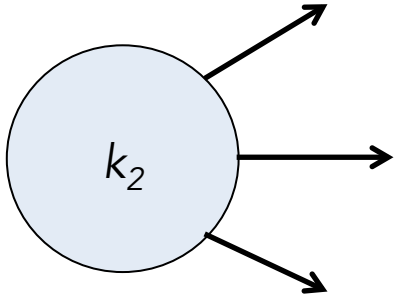


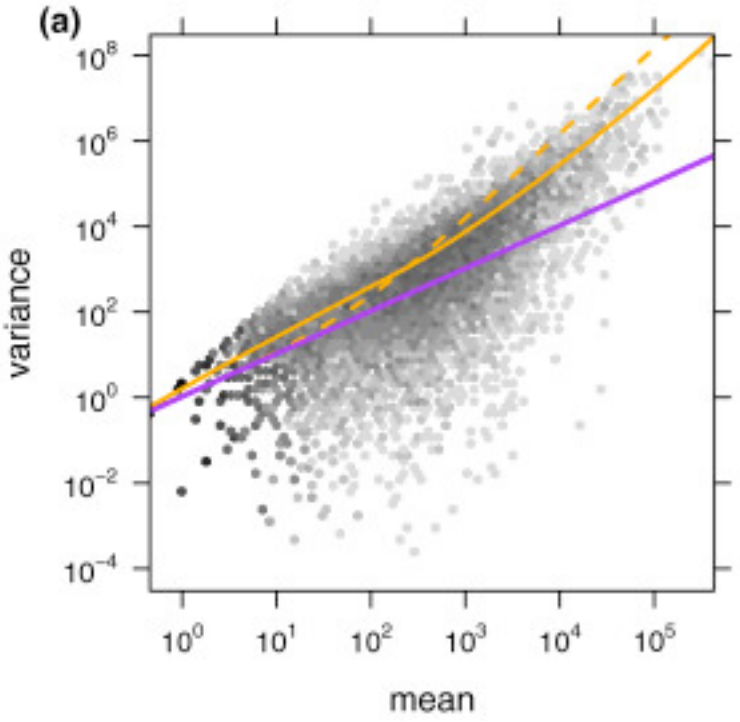


Modelling – in fashion

- Biology is never that simple…





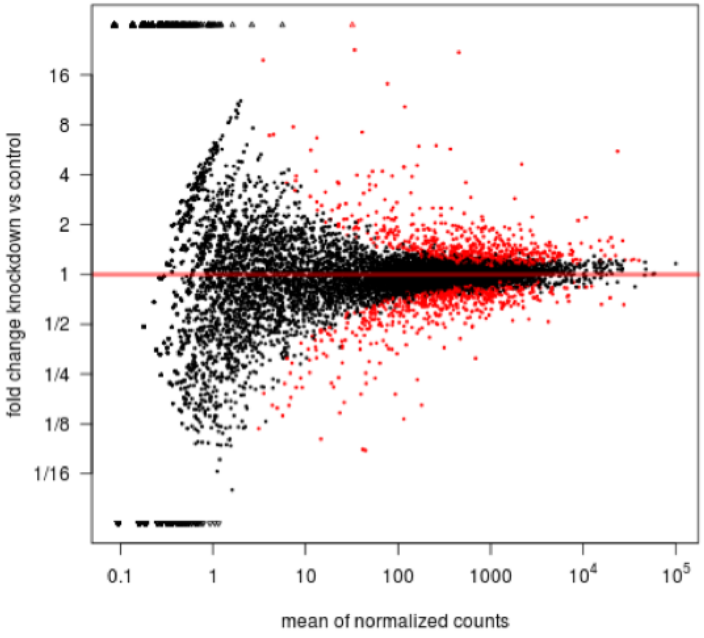


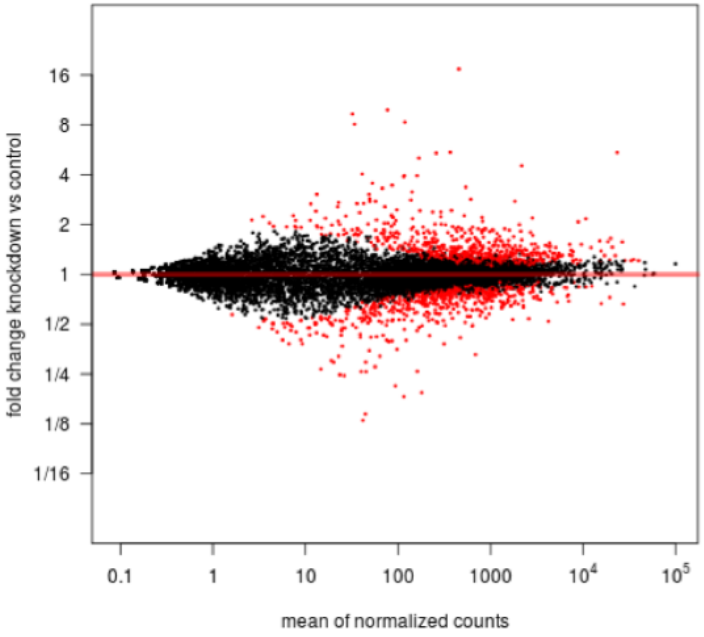
- The negative binomial distribution represents an *overdispersed* Poisson distribution, and has parameters for both the mean and the overdispersion.

Anders, S. & Huber, W. (2010) *Genome Biology*

Modelling – in fashion

- Estimating the dispersion parameter can be difficult with a small number of samples
- edgeR: models the variance as the sum of technical and biological variance
- 'Share' information from all genes to obtain global estimate - *shrinkage*



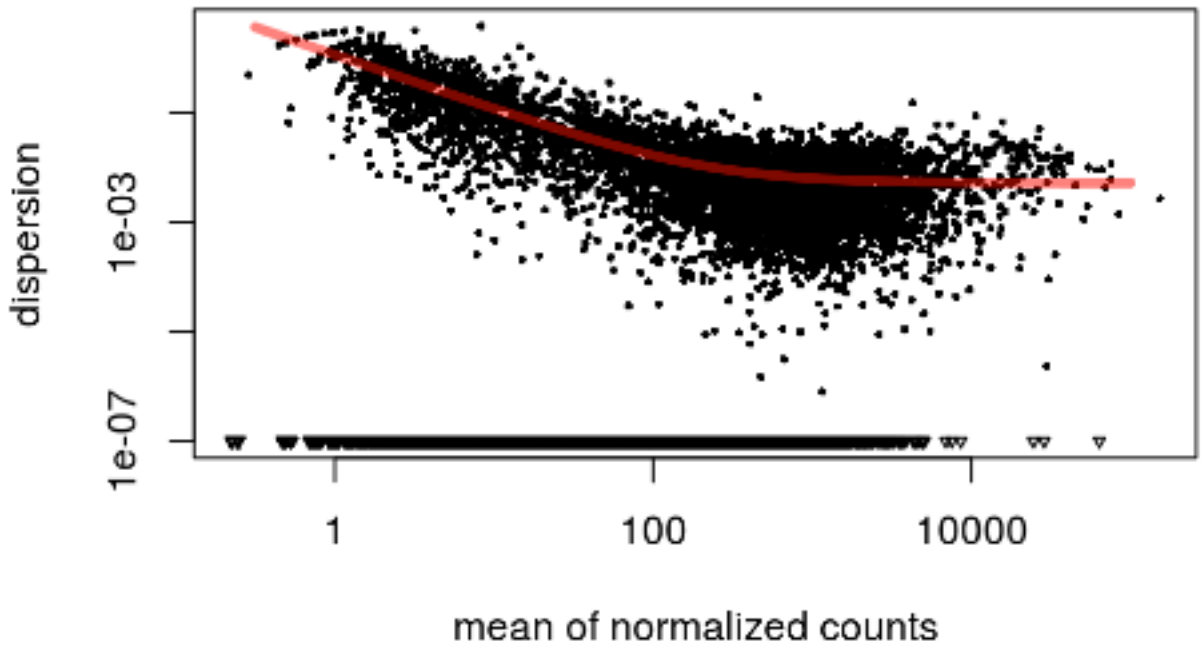


Simon Anders

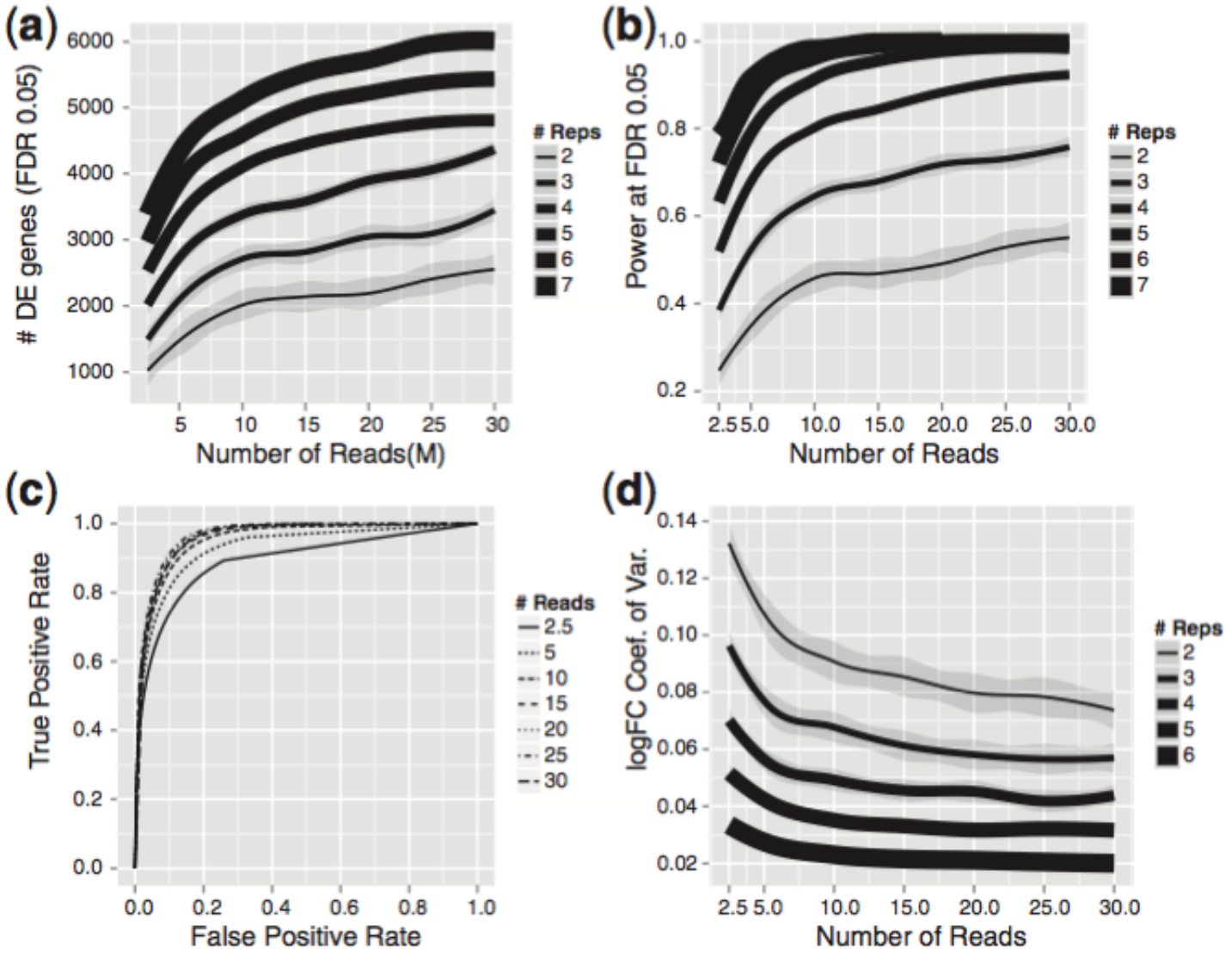
Modelling – in fashion

- DESeq uses a similar formulation of the variance term

$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}} .$$

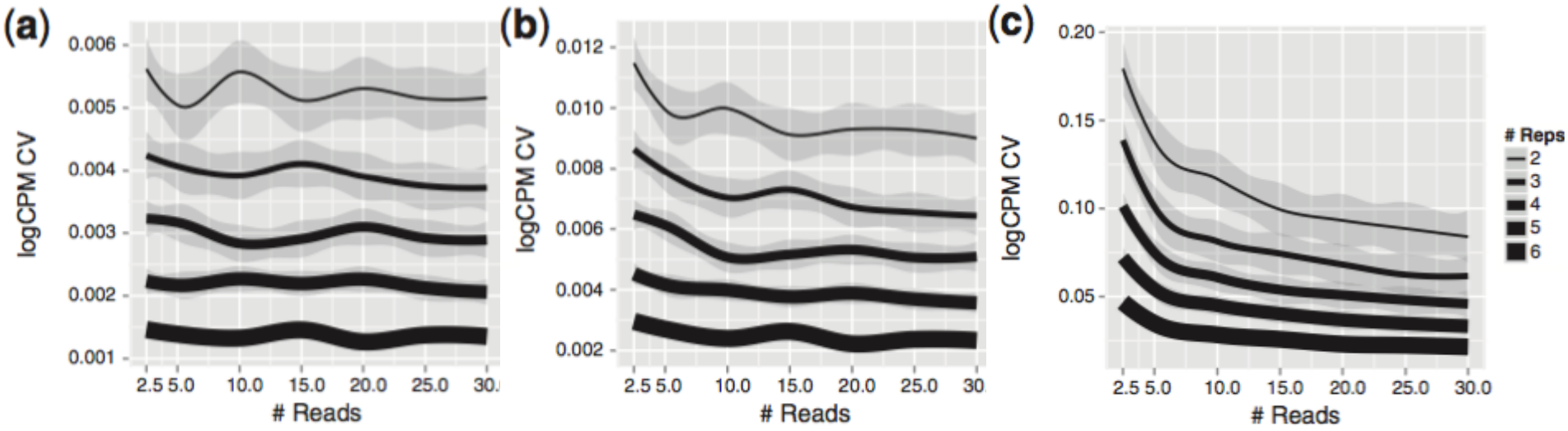


On replicates…



Liu et al. (2014) *Bioinformatics*

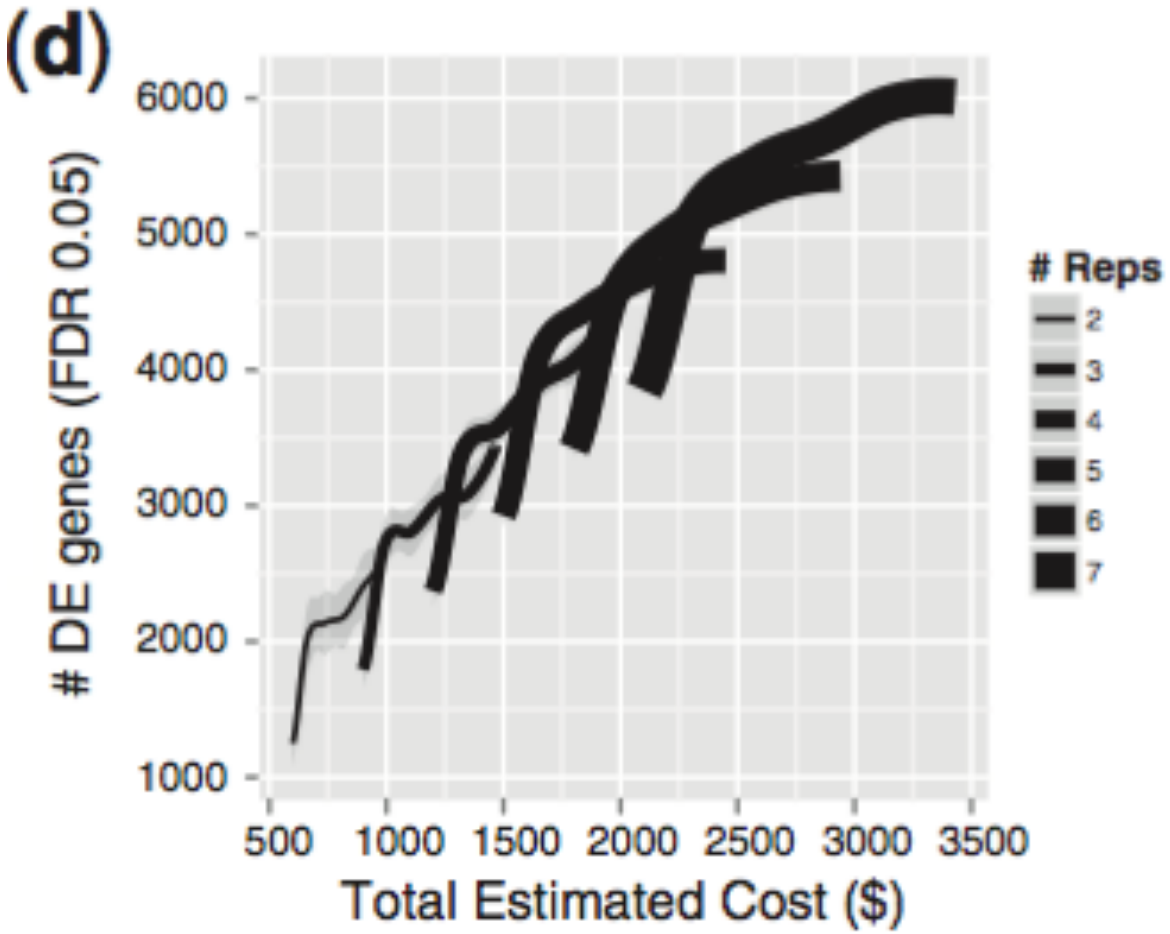
On replicates…



HIGH **MEDIUM** **LOW**

Liu et al. (2014) *Bioinformatics*

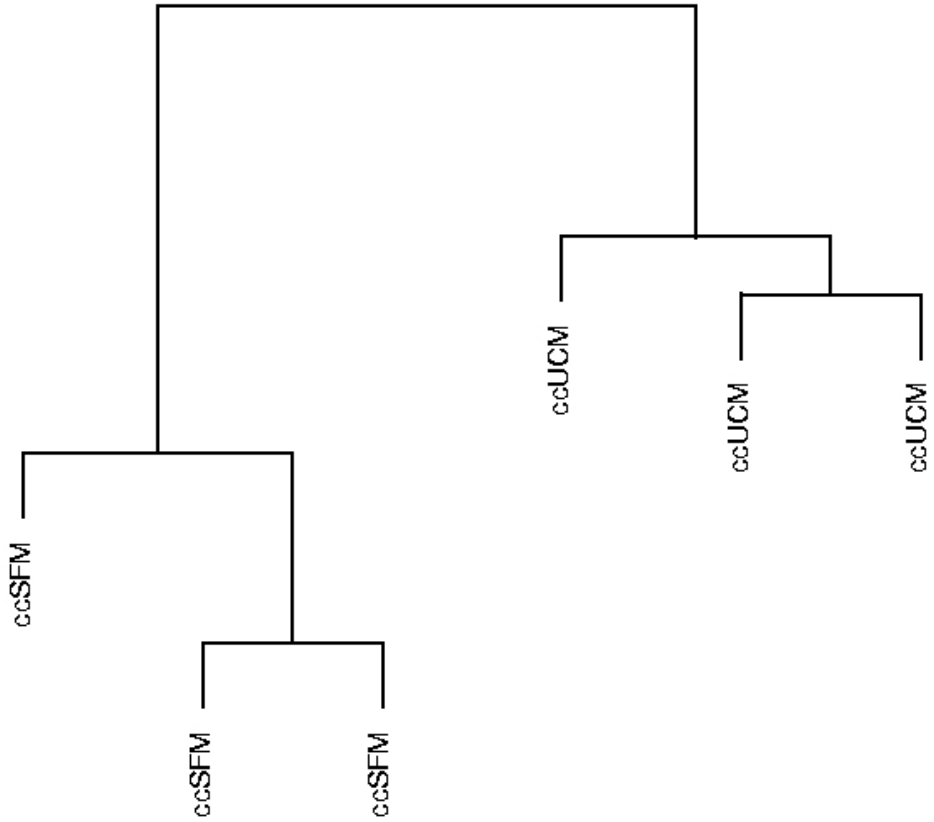
On replicates…

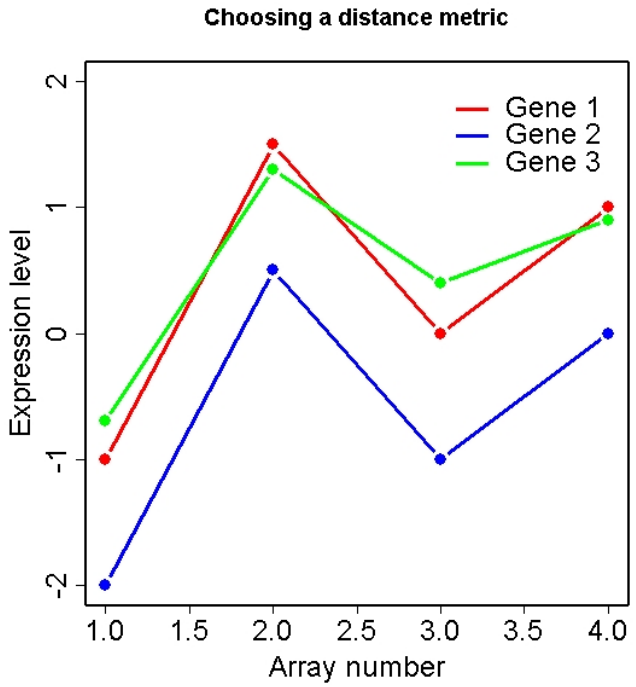


Liu et al. (2014) *Bioinformatics*

What next?

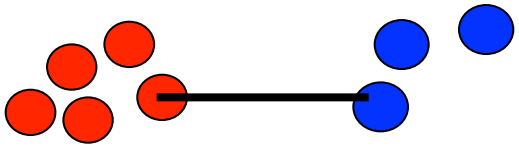
- Hierarchical clustering = define metric & look for similarities



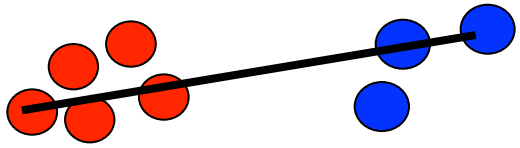


What next?

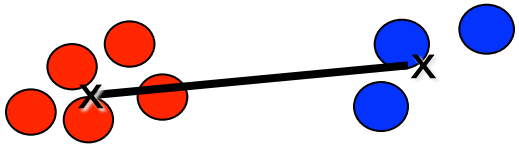
- Merging clusters according to a metric



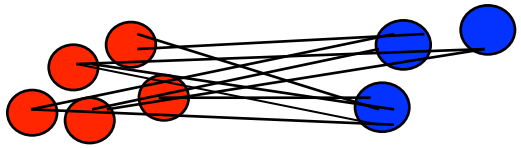
Single
(min. of pairwise distances)



Complete
(max. of pairwise distances)

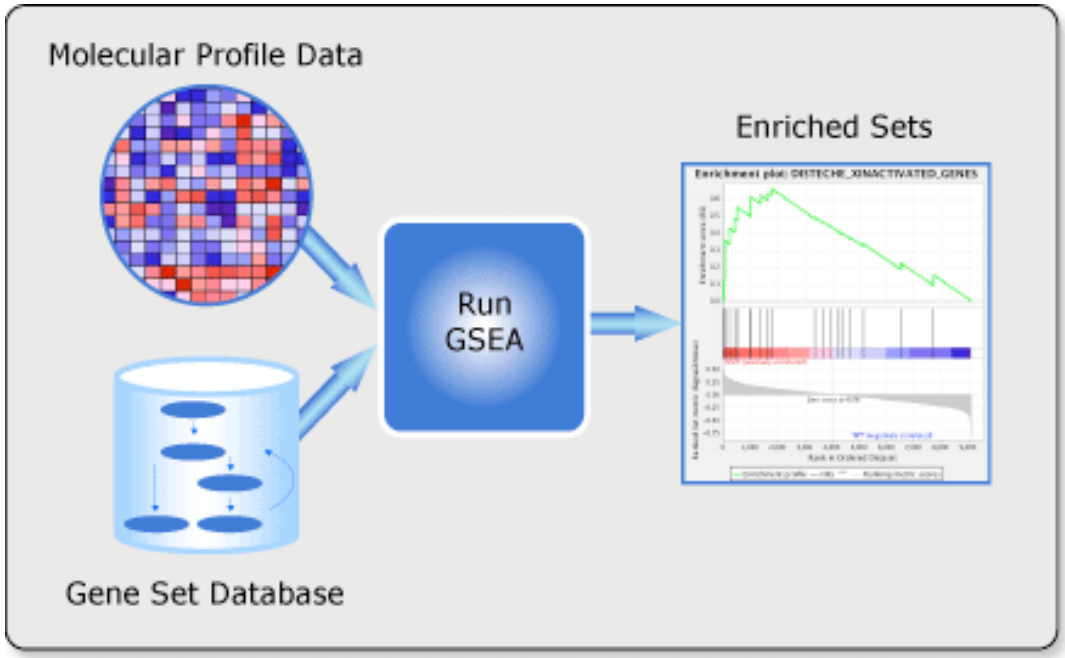


Distance between centroids



Average linkage
(mean of all pairwise distances)

What next?



- **H** (hallmark gene sets, 50 gene sets)
- **C1** (positional gene sets, 326 gene sets)
  - by chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
- **C2** (curated gene sets, 4725 gene sets)
  - **CGP** (chemical and genetic perturbations, 3395 gene sets)
  - **CP** (Canonical pathways, 1330 gene sets)
  - **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets)
  - **CP:KEGG** (KEGG gene sets, 186 gene sets)
  - **CP:REACTOME** (Reactome gene sets, 674 gene sets)
- **C3** (motif gene sets, 836 gene sets)
  - **MIR** (microRNA targets, 221 gene sets)
  - **TFT** (transcription factor targets, 615 gene sets)
- **C4** (computational gene sets, 858 gene sets)
  - **CGN** (cancer gene neighborhoods, 427 gene sets)
  - **CM** (cancer modules, 431 gene sets)
- **C5** (GO gene sets, 1454 gene sets)
  - **BP** (GO biological process, 825 gene sets)
  - **CC** (GO cellular component, 233 gene sets)
  - **MF** (GO molecular function, 396 gene sets)
- **C6** (oncogenic signatures, 189 gene sets)
- **C7** (immunologic signatures, 1910 gene sets)