(A brief introduction to the downstream)

Analysis of RNA-seq Data

RNA-seq analysis course - May 2018 - Day 2

Stephane Ballereau
based on slides by Bernard Pereira
and Ashley Sawle

UNIVERSITY OF CAMBRIDGE

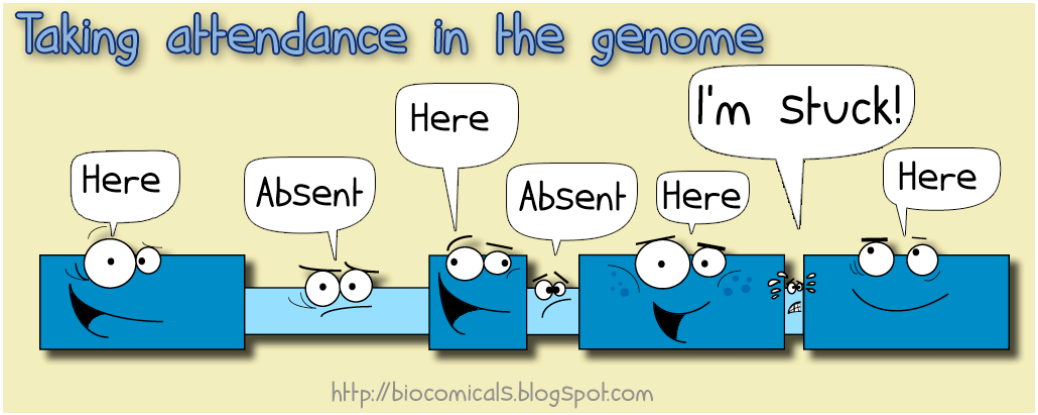
CANCER RESEARCH UK

Recap of Part 1

The many faces of RNA-seq – Applications

Discovery

- Transcripts
- Isoforms
- Splice junctions
- Fusion genes

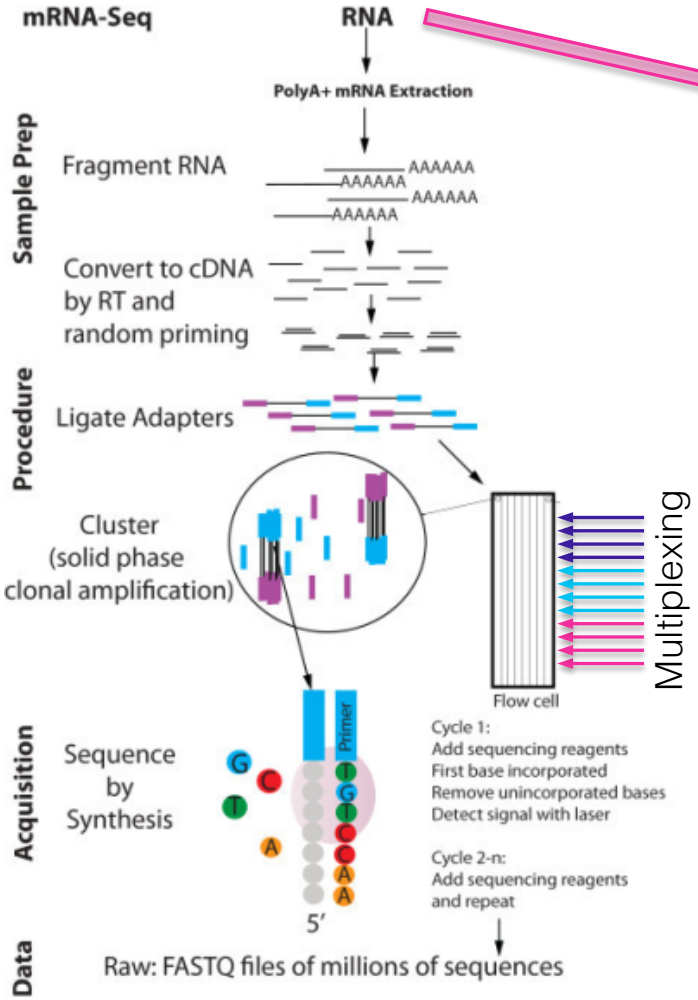


Differential expression

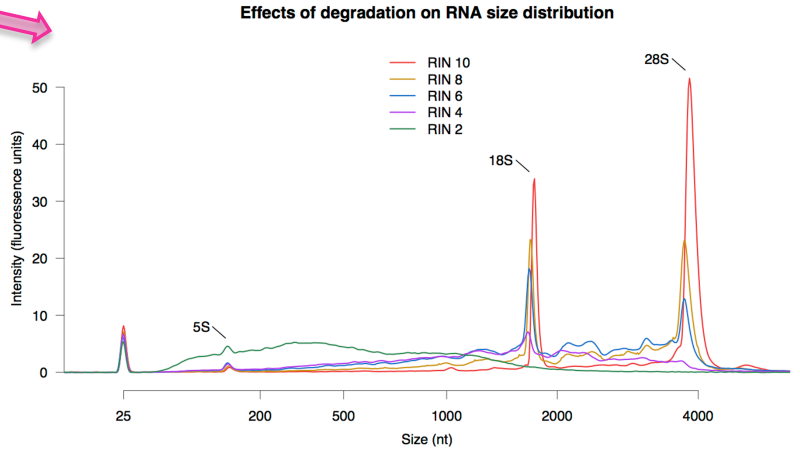
- **Gene level expression changes**
- Relative isoform abundance
- Splicing patterns

Variant calling

Library Preparation & Sequencing



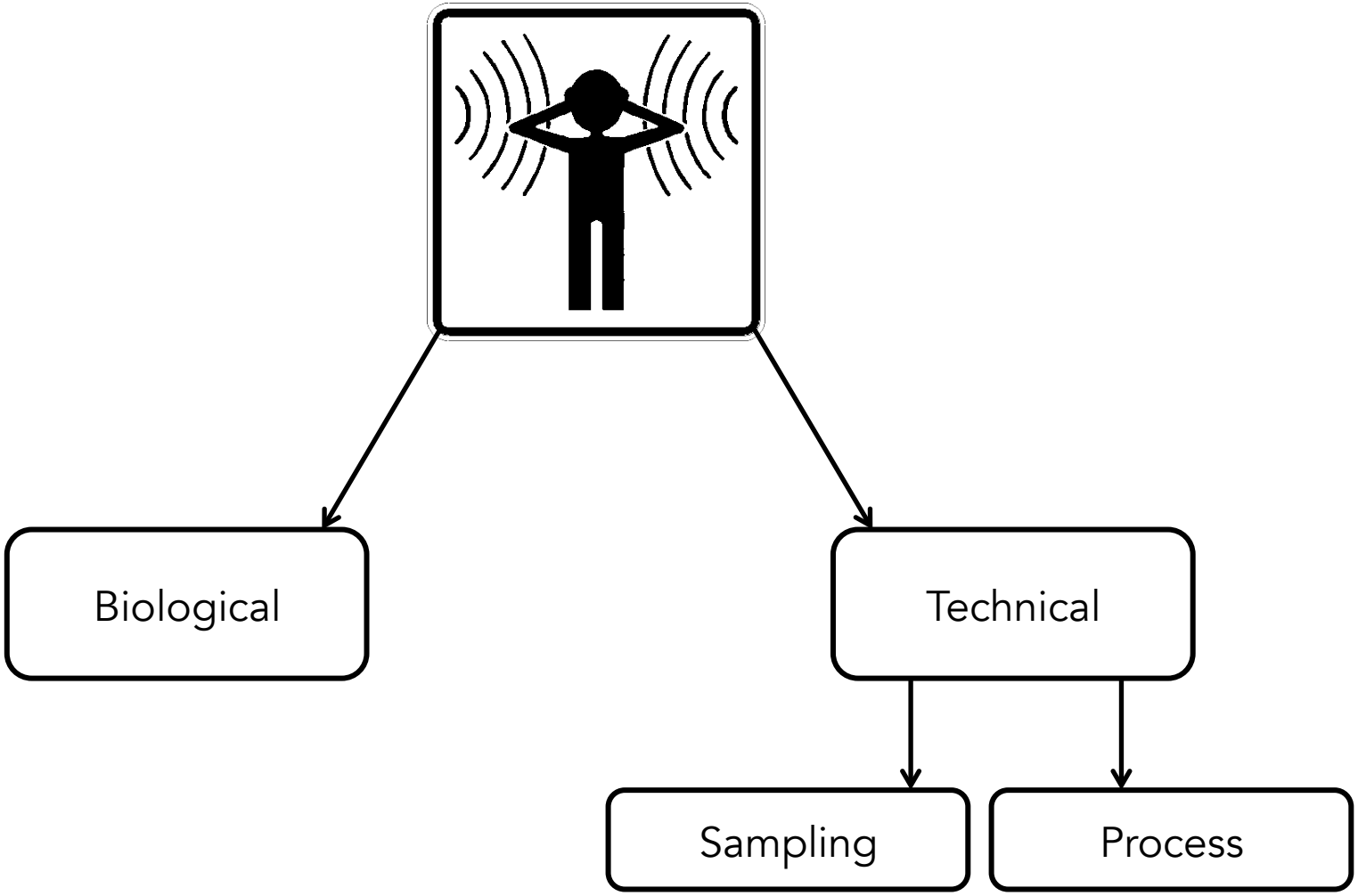
QC - RIN number



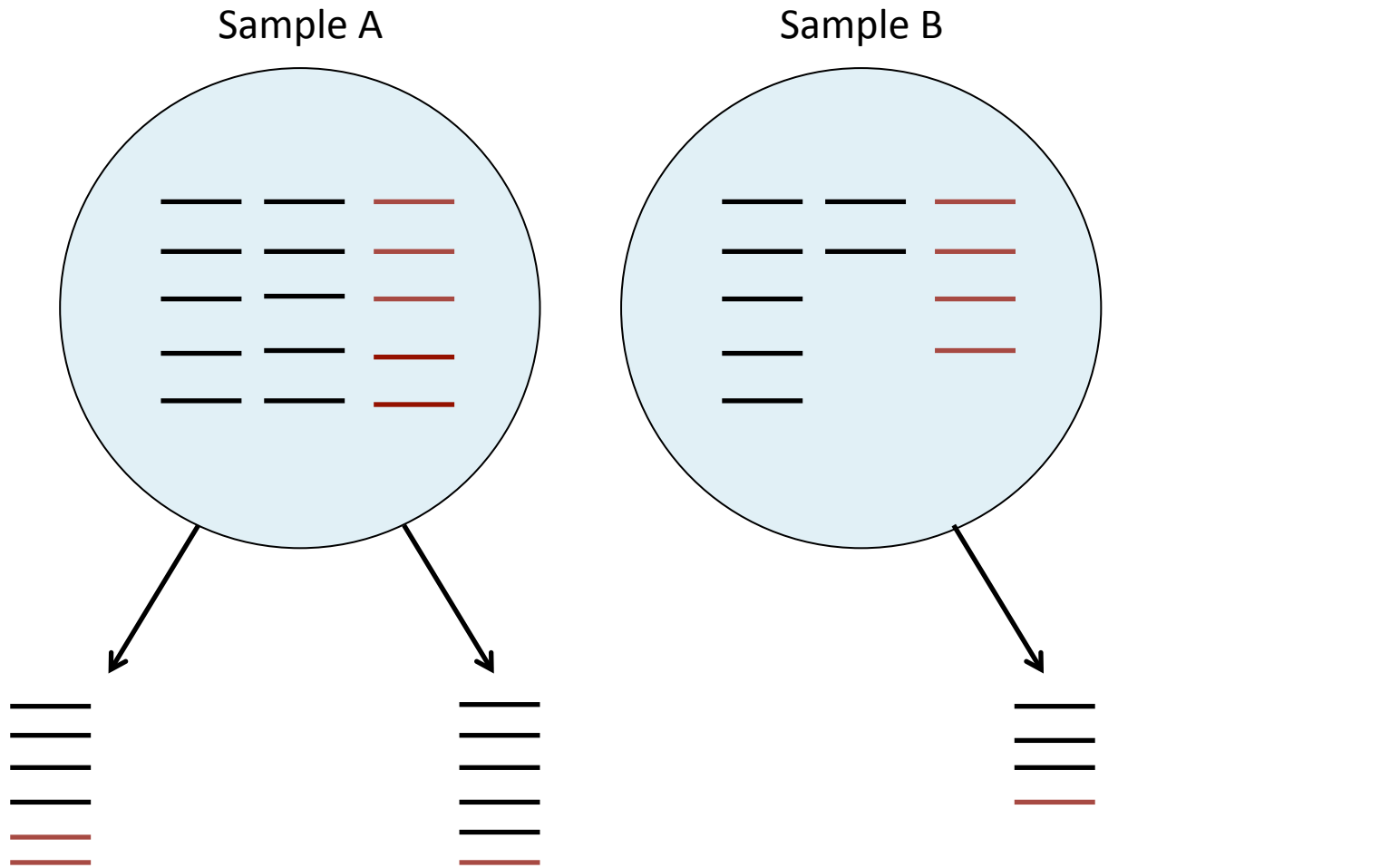
Sigurgeirsson, Emanuelsson & Lundeberg (2014) PLOS ONE

modified from Malone JH, Oliver B (2011) BMC Biol.

Sources of Noise



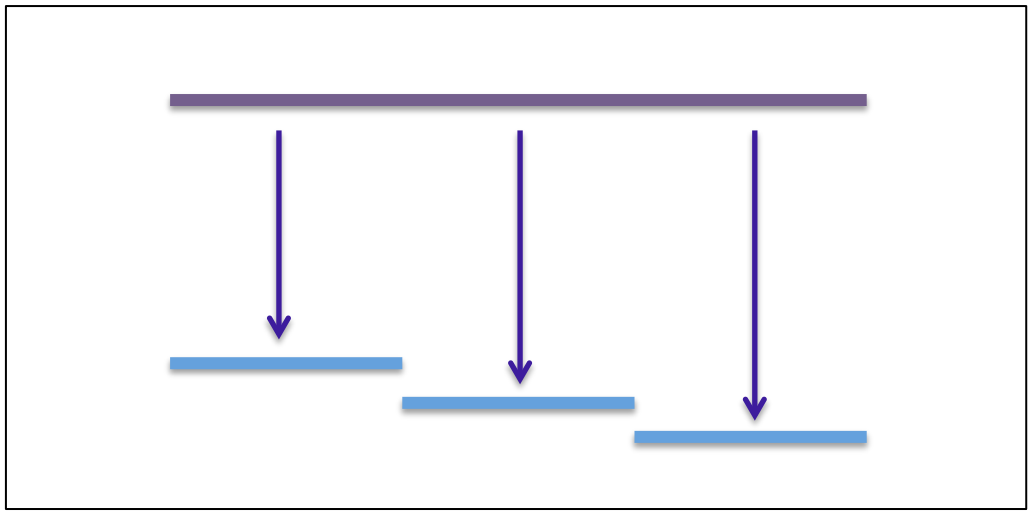
Sources of Noise – Sampling Bias



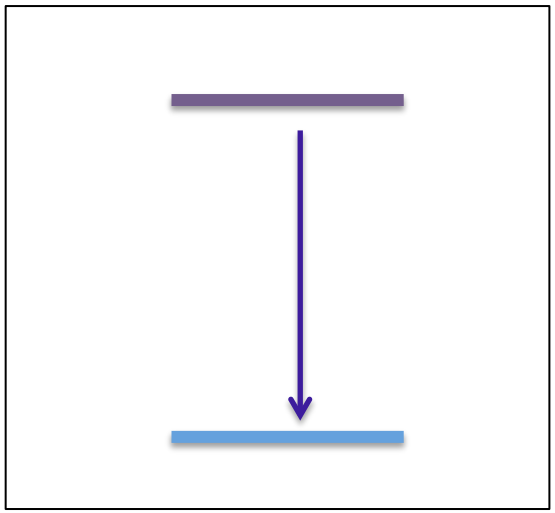
Subsampling a from a pool of RNAs

Sources of Noise – Sampling Bias

Transcript A

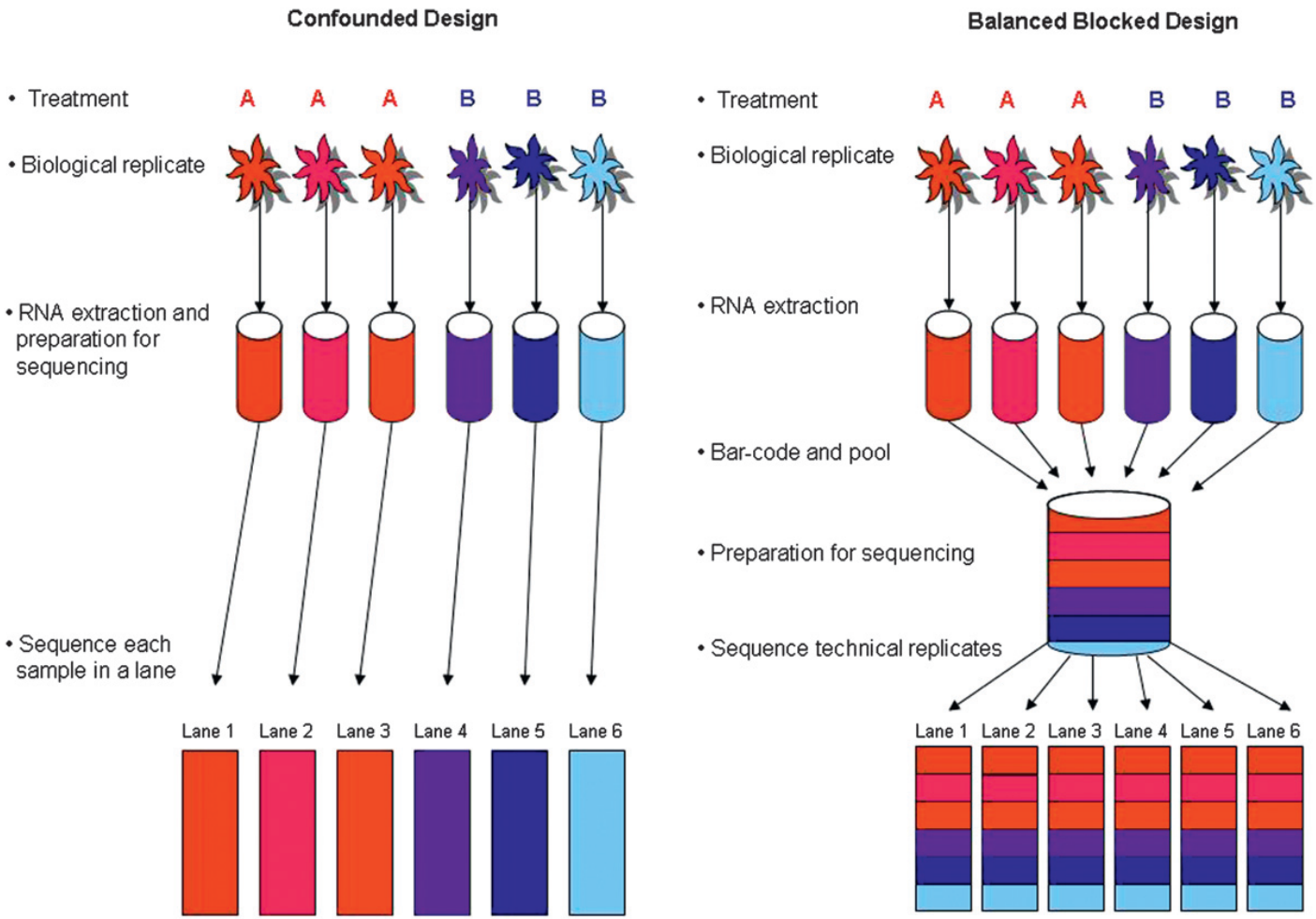


Transcript B

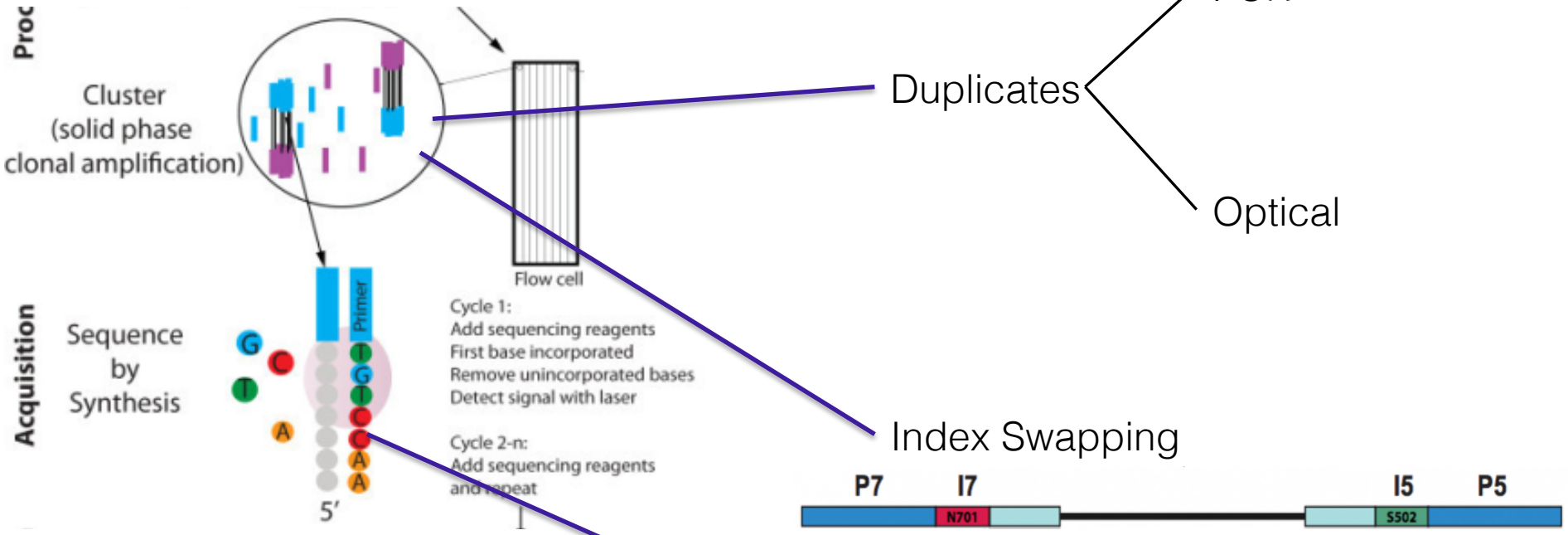


Transcript length affects the number of RNA fragments present in the library from that gene

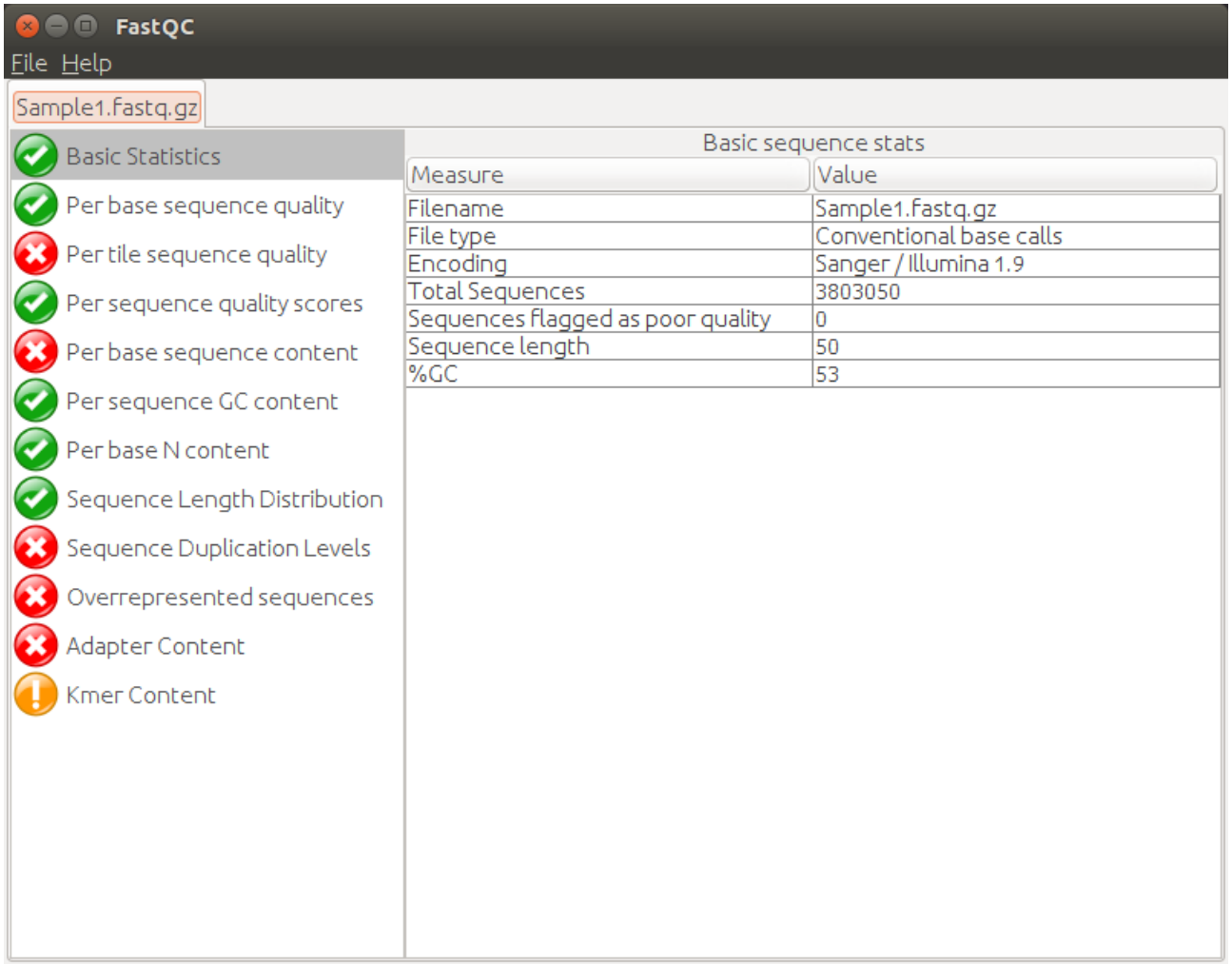
Sources of Noise - Process



Sources of Noise – Process

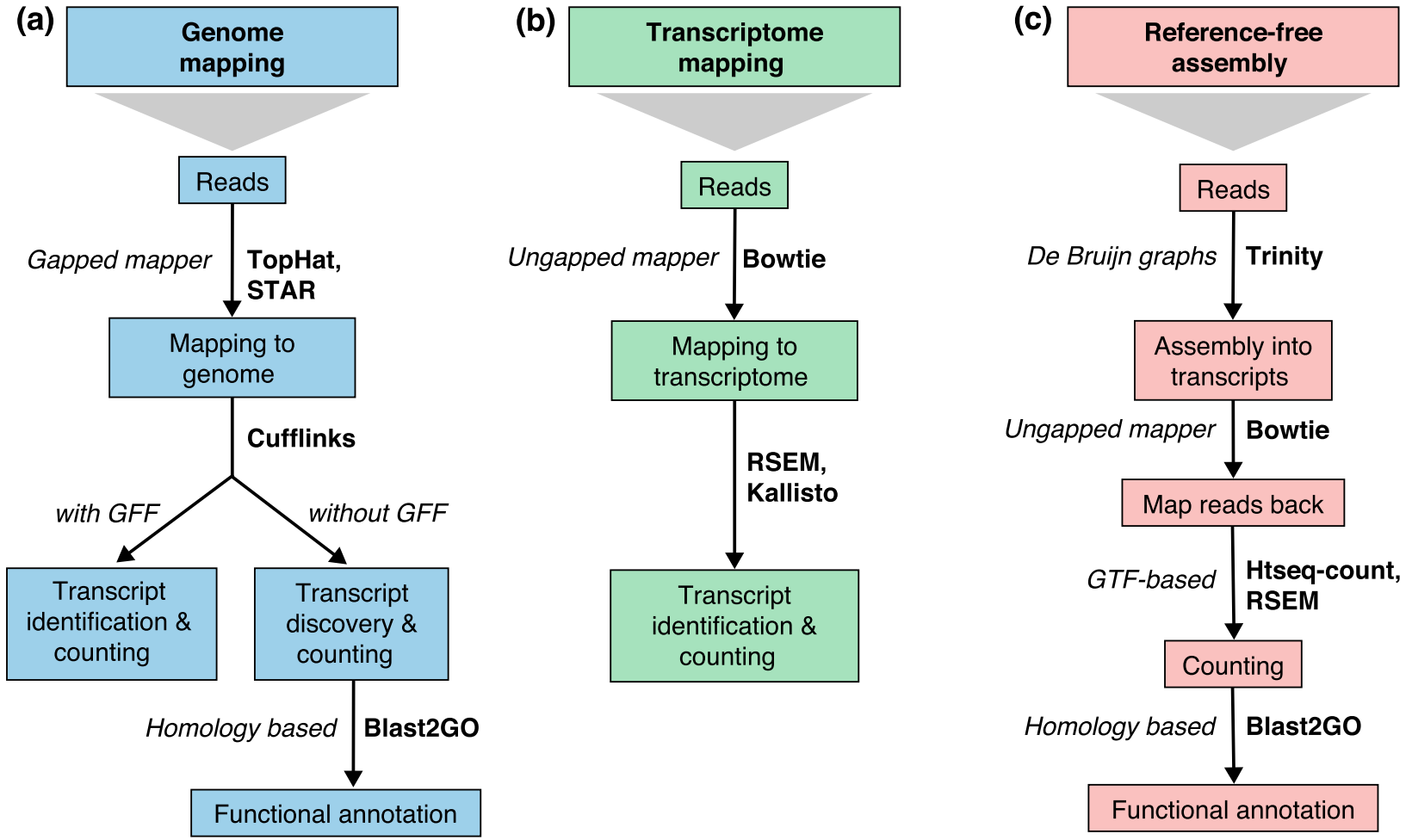


Raw Sequence QC - FASTQC



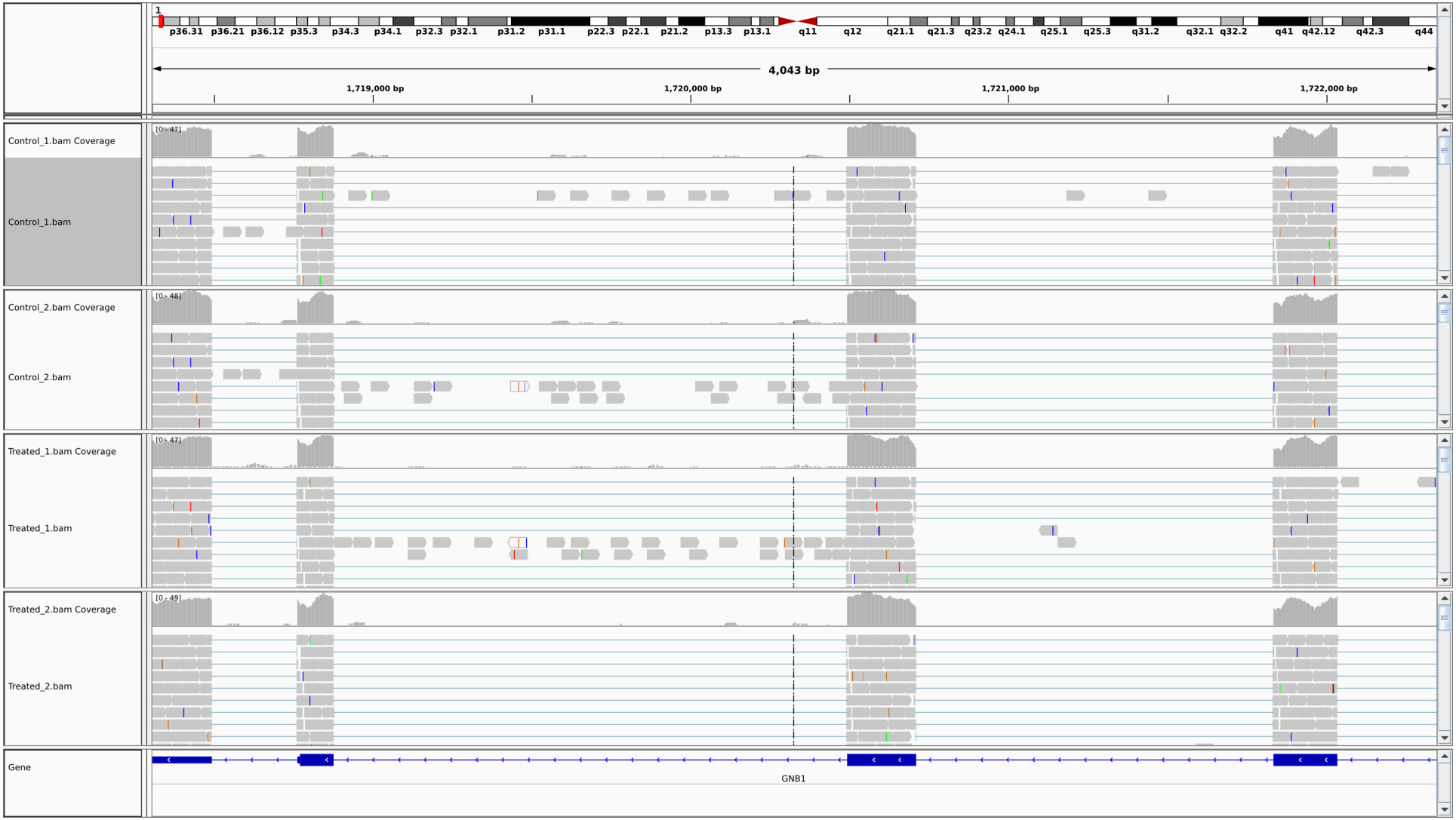
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Sequence to Sense



Conesa *et al.* (2016) Genome Biology

Visualising Mapping Results – IGV



Summarisation/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

GeneID	Sample_A	Sample_B	Sample_C	Sample_D	Sample_E	Sample_F	Sample_G	Sample_H	Sample_I
ENSG00000223972	23	11	31	9	11	13	17	17	22
ENSG00000227232	1000	828	1078	758	728	897	1075	793	1089
ENSG00000243485	8	6	2	2	3	4	2	5	6
ENSG00000237613	1	1	0	2	1	4	5	1	2
ENSG00000238009	107	69	85	66	87	64	89	55	81
ENSG00000233750	16	5	23	10	4	21	14	21	20
ENSG00000237683	1259	1025	1375	990	997	1109	1141	693	973
ENSG00000268903	3652	3422	2725	3274	3384	2154	2798	5761	6089
ENSG00000239906	25430	21022	13947	45938	47405	28038	8557	17889	16544
ENSG00000241860	194936	184076	162085	172115	164332	118233	146396	221478	262352
ENSG00000222623	49492	44102	41514	43487	43009	32654	40010	53883	65989
ENSG00000241599	4	10	3	6	5	2	5	9	6
ENSG00000228463	34074	32072	24434	41568	41246	27624	19095	39606	38636
ENSG00000237094	48499	45757	32395	77500	84031	57687	19371	32145	36202
ENSG00000250575	1	0	0	0	1	0	2	0	0
ENSG00000233653	0	1	3	1	0	2	0	0	0
ENSG00000235249	549	434	605	427	427	523	425	333	448
ENSG00000256186	599	591	842	683	724	843	700	391	478
ENSG00000236601	1	1	0	0	0	2	0	0	0
ENSG00000236743	91	57	85	59	58	70	82	57	70
ENSG00000236679	7	2	8	3	2	1	0	1	0
ENSG00000231709	266	213	297	191	210	300	299	174	274
ENSG00000235146	336	267	399	333	390	371	329	196	300
ENSG00000239664	25	14	30	30	29	23	16	13	12
ENSG00000230021	6	11	14	7	5	6	8	6	6
ENSG00000223659	4	7	10	5	12	12	7	4	7
ENSG00000225972	1	2	0	1	4	0	4	0	1
ENSG00000225630	98	99	120	92	92	101	95	59	105

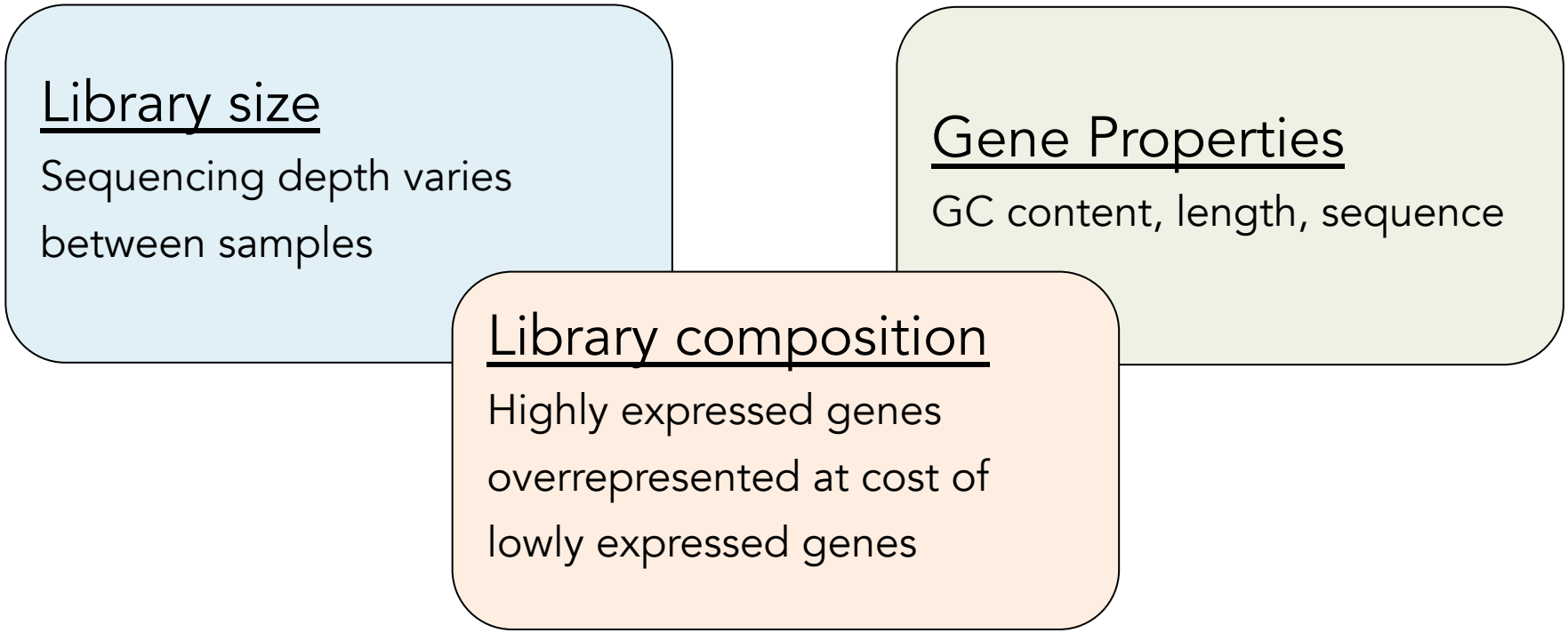
Downstream analysis

Normalisation

- Counting

  → estimate of *relative* counts for each gene

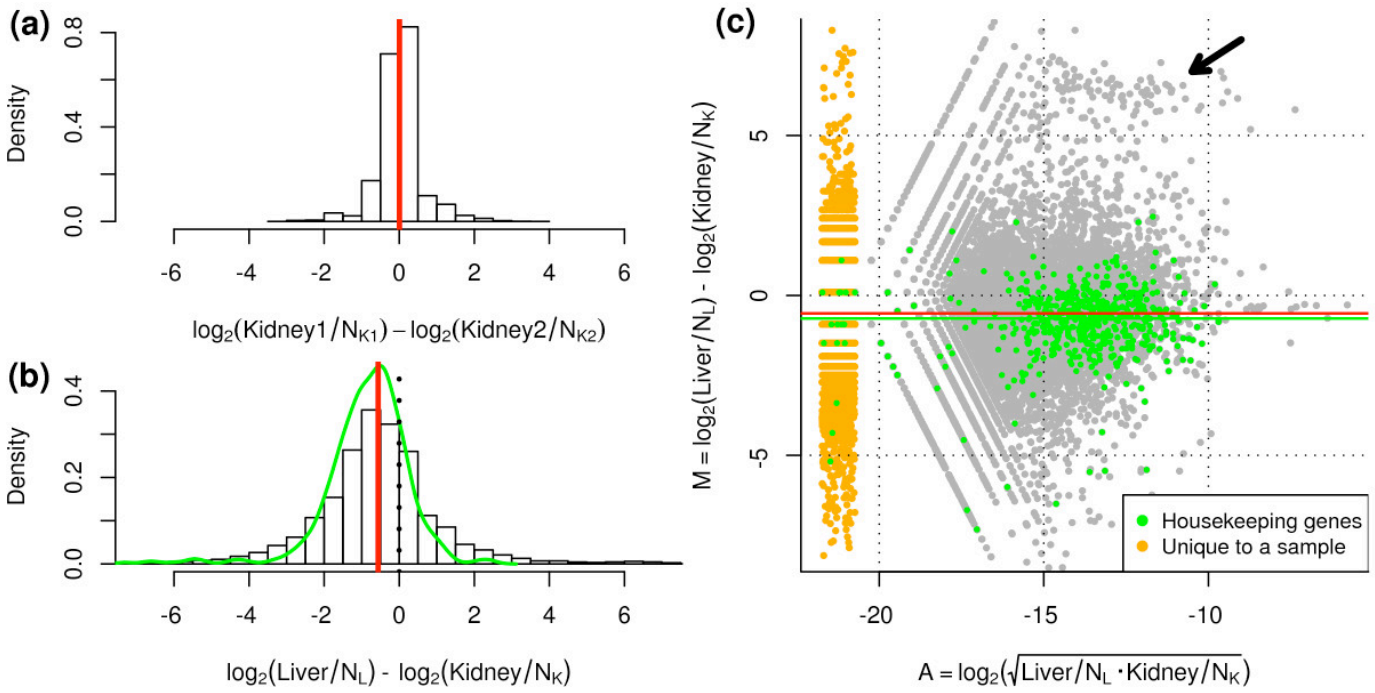
Does this accurately represent the original population?



Normalisation – Trimmed Mean of M

Trimmed mean of *M*

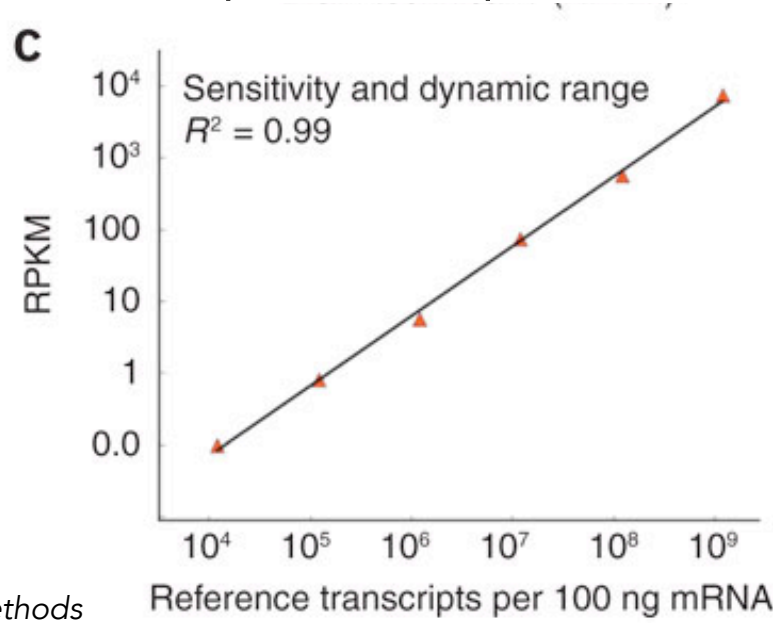
- Implemented in edgeR
- Assumes most genes are not differentially expressed



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

Differential Expression

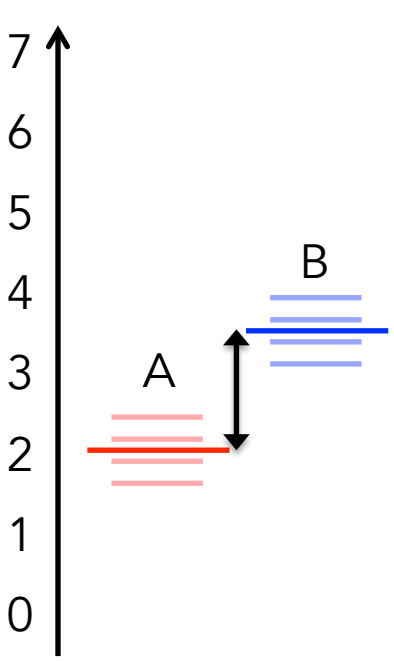
- Comparing feature abundance under different conditions
- Assumes linearity of signal
- When *feature=gene*, well-established pre- and post-analysis strategies exist

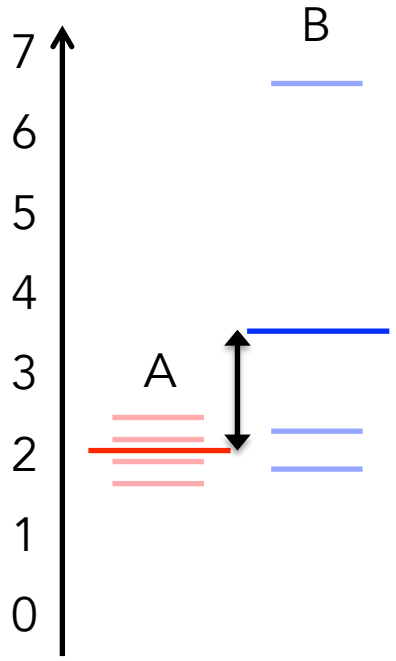


Mortazavi, A. et al (2008) *Nature Methods*

Differential Expression

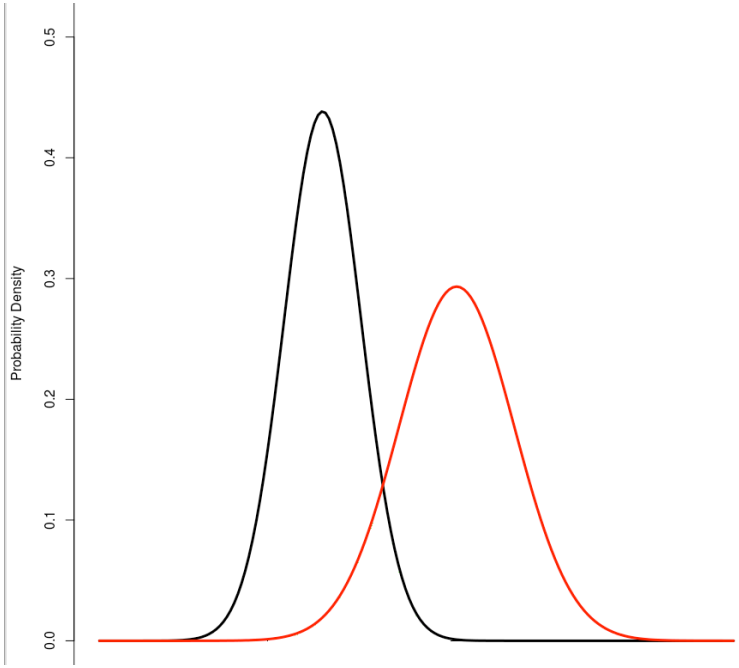
- Simple difference in means





- Replication introduces variance

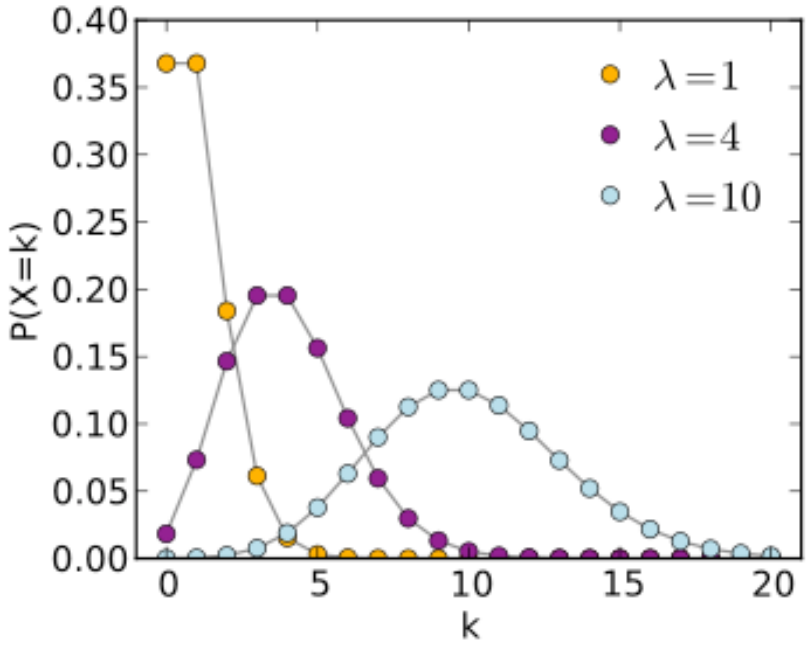
Differential Expression - Modelling



Normal distribution → t-test

Differential Expression- Modelling

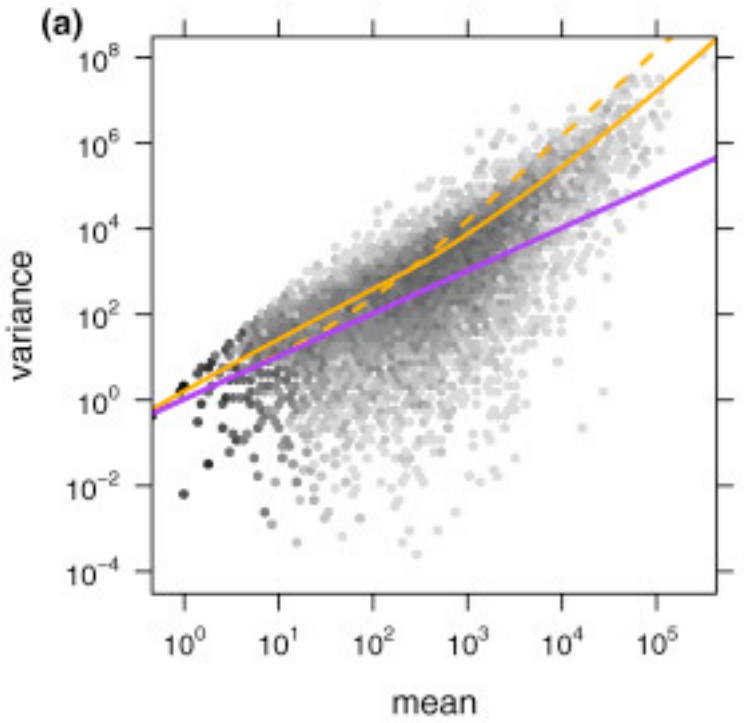
- Use the Poisson distribution for count data
- Just one parameter required – the mean



Differential Expression- Modelling

- Biology is never that simple
- The negative binomial distribution represents an overdispersed Poisson distribution
- It has two parameters:

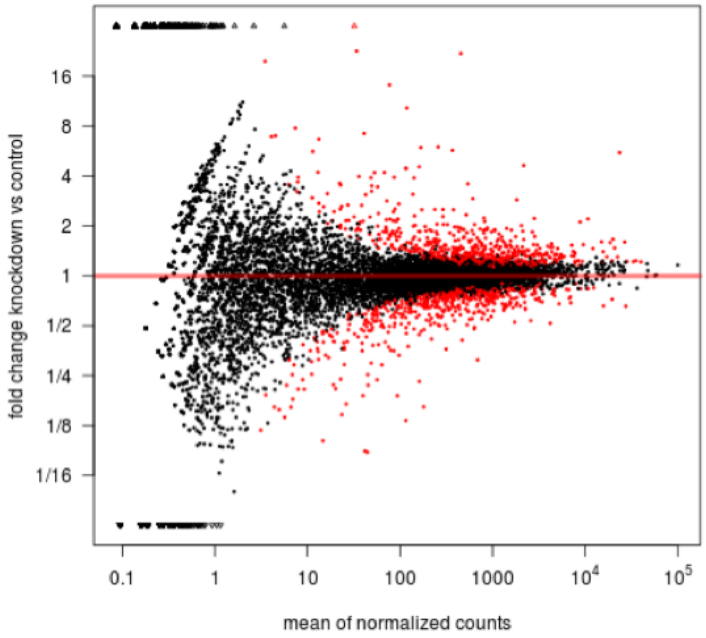
  mean and (over)dispersion

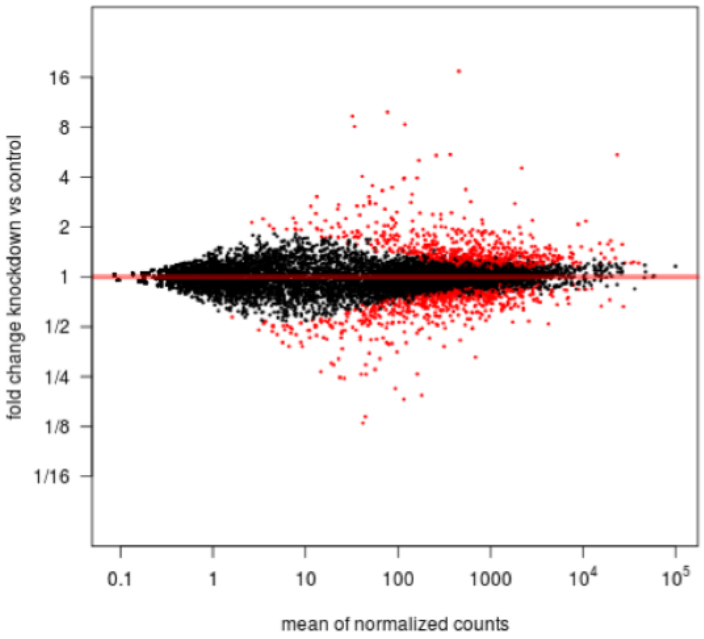


Anders, S. & Huber, W. (2010) *Genome Biology*

Differential Expression- Modelling

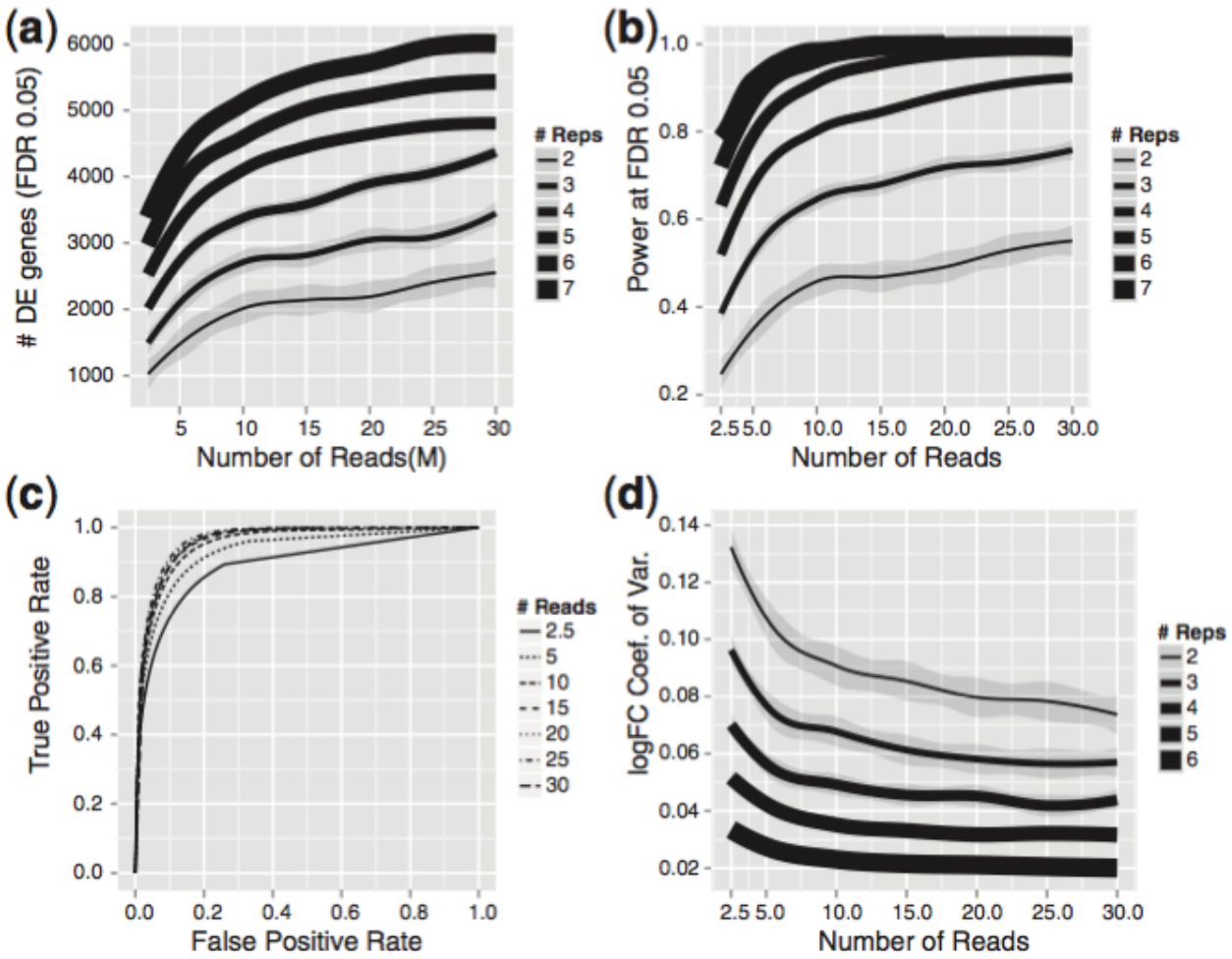
- Estimating the dispersion parameter can be difficult with a small number of samples
- edgeR: models the variance as the sum of technical and biological variance
- ‘Share’ information from all genes to obtain global estimate - shrinkage





Simon Anders

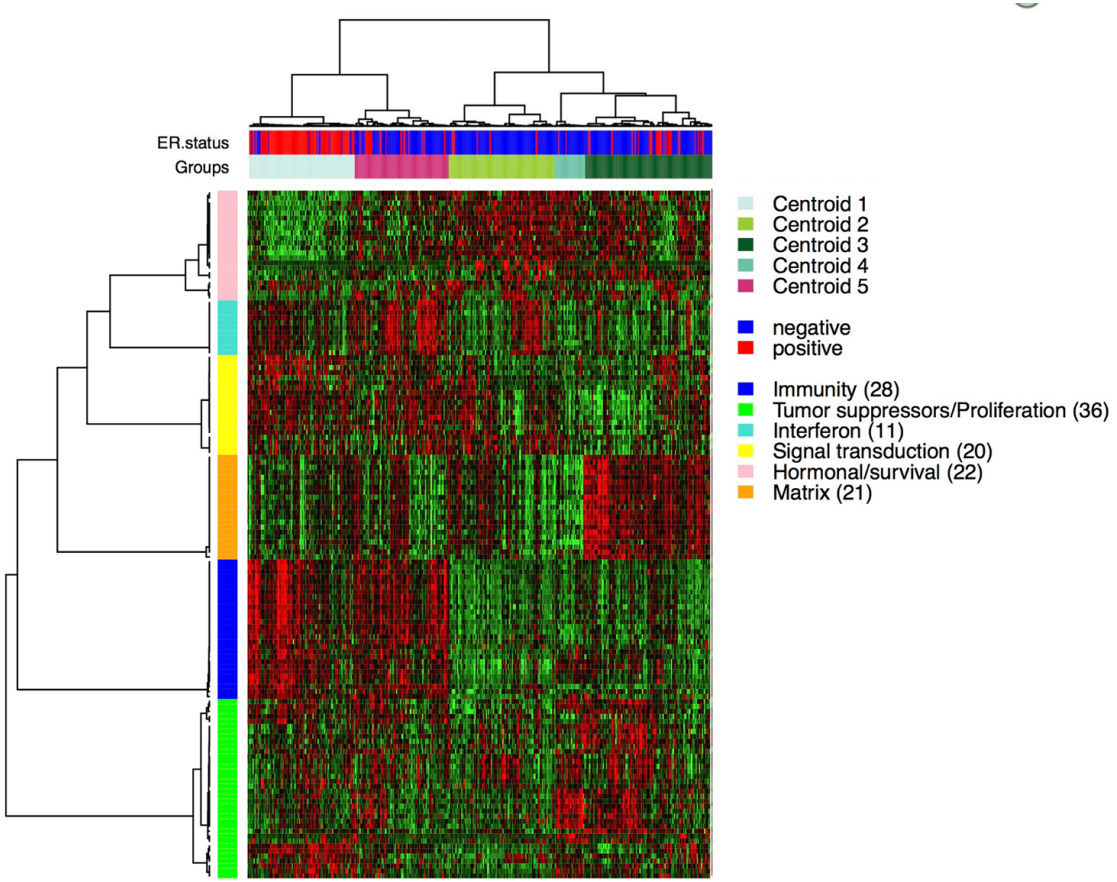
Replicates v Sequencing Depth



Liu et al. (2014) *Bioinformatics*

Towards Biological Meaning

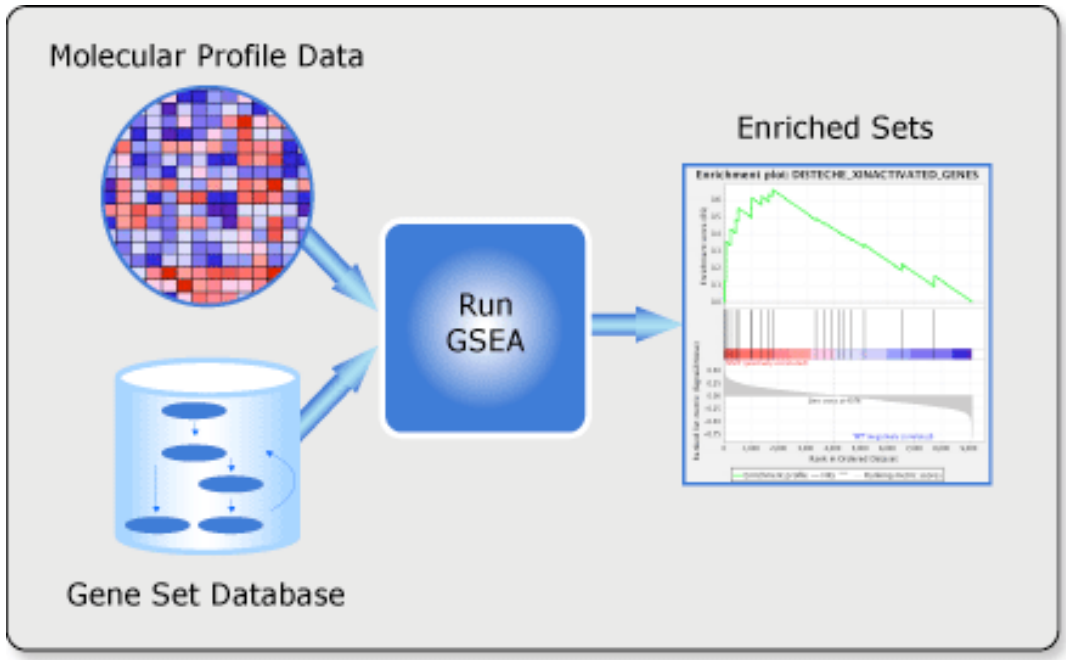
- Clustering



Hamy et al. (2016) *PLOS One*

Towards Biological Meaning

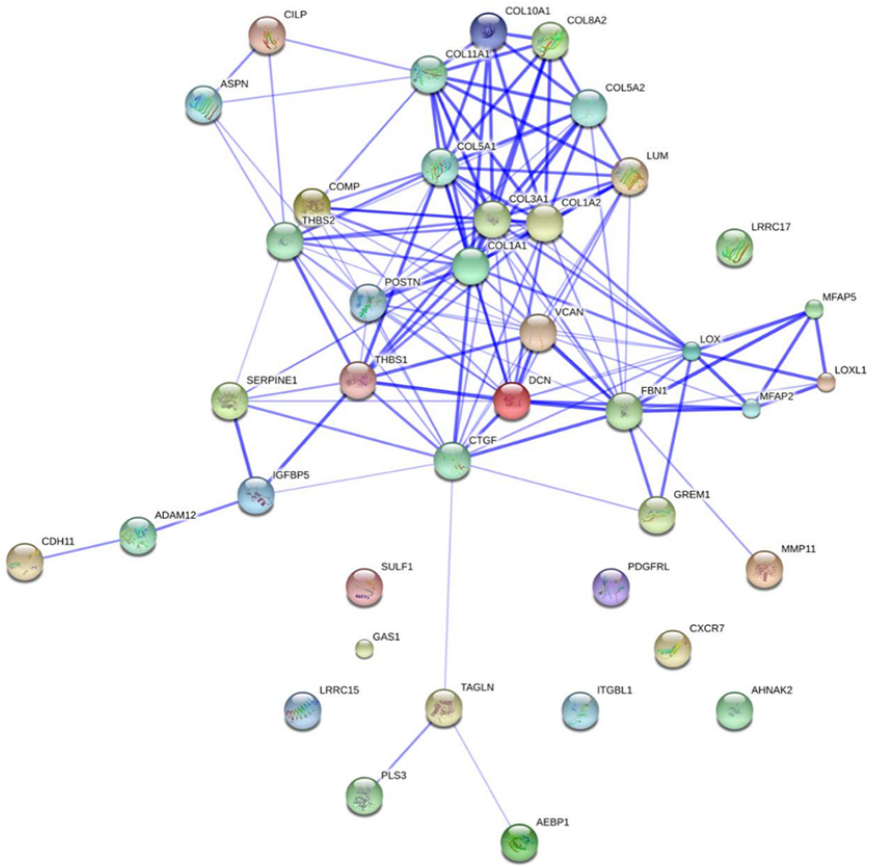
- Gene Set Enrichment Analysis



- **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)

Towards Biological Meaning

- Network analysis



Hamy et al. (2016) *PLOS One*

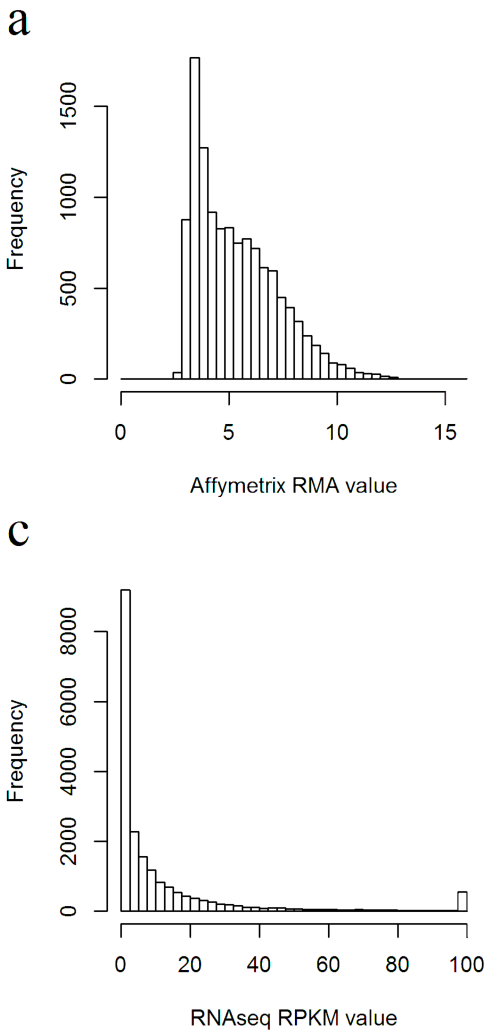
Thank you

cut out slides

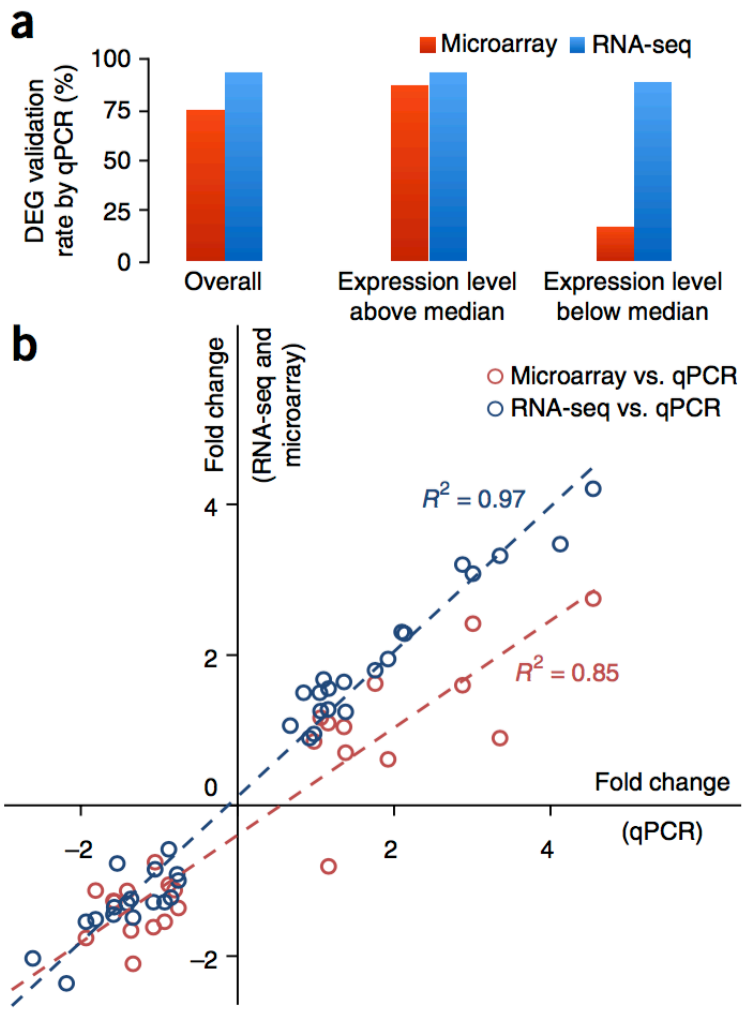
The many faces of RNA-seq – Techniques

- mRNA-seq
- Exome capture
- Targeted
- Small RNA
  - miRNA
  - piRNA
  - sncRNA
- Total RNA
- Ribosome profiling
- Single Cell RNA-Seq

Microarray → RNA-seq



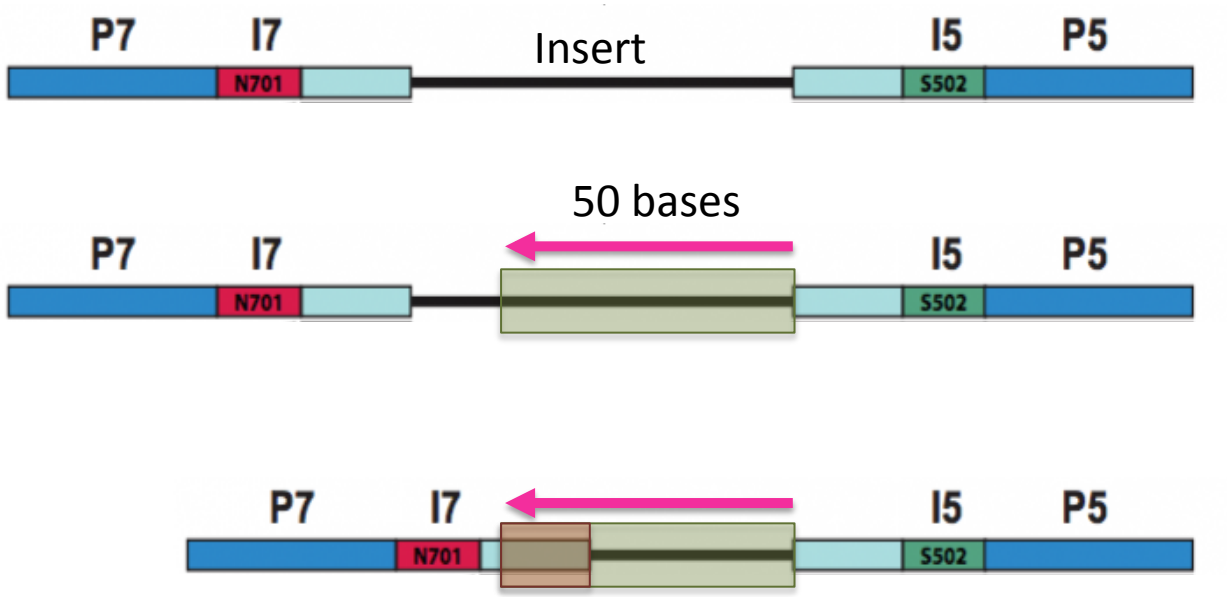
Guo et al. (2013) *Plos One*



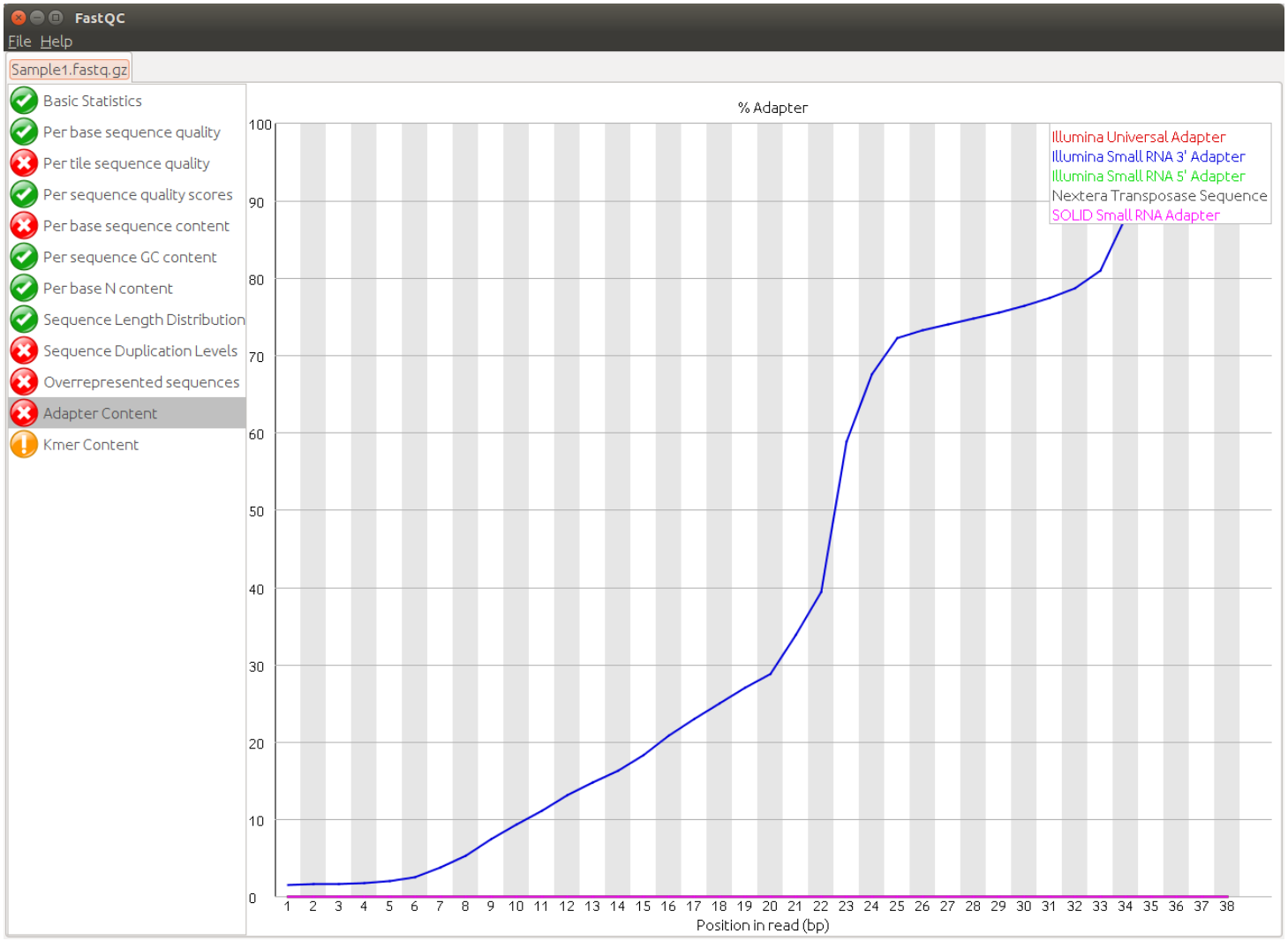
Wang et al (2014) *Nature Biotech.*

Trimming

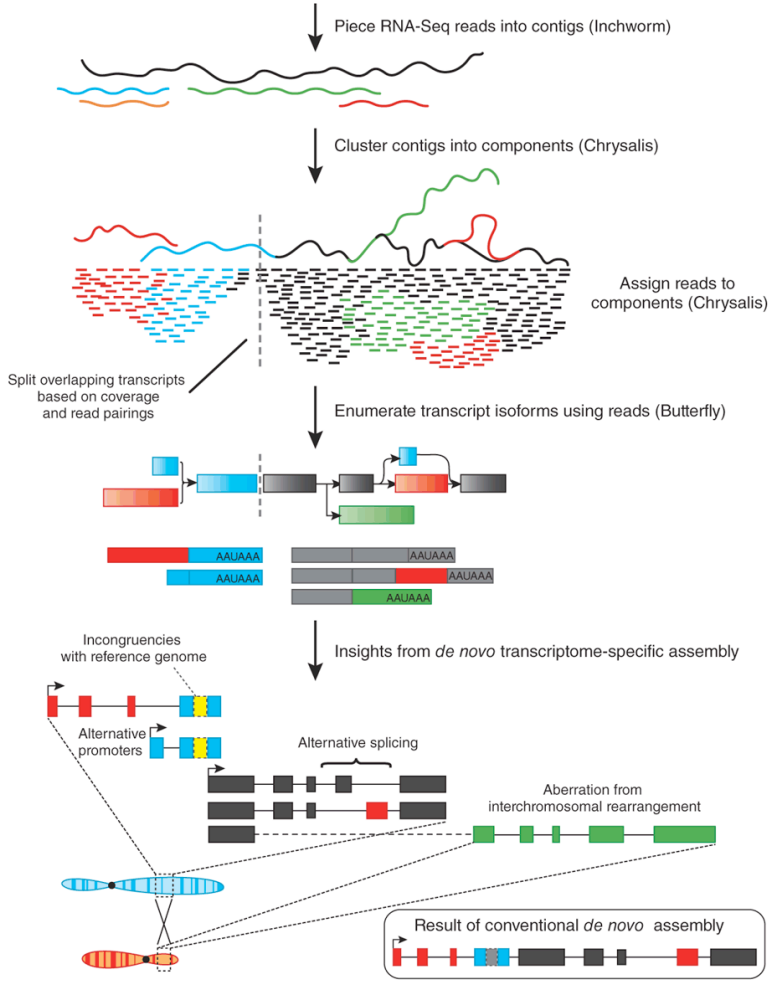
- Quality-based Trimming
- Adapter contamination



Adapter contamination - FASTQC



De Novo assembly



e.g. TRINITY

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

Analysis Overview

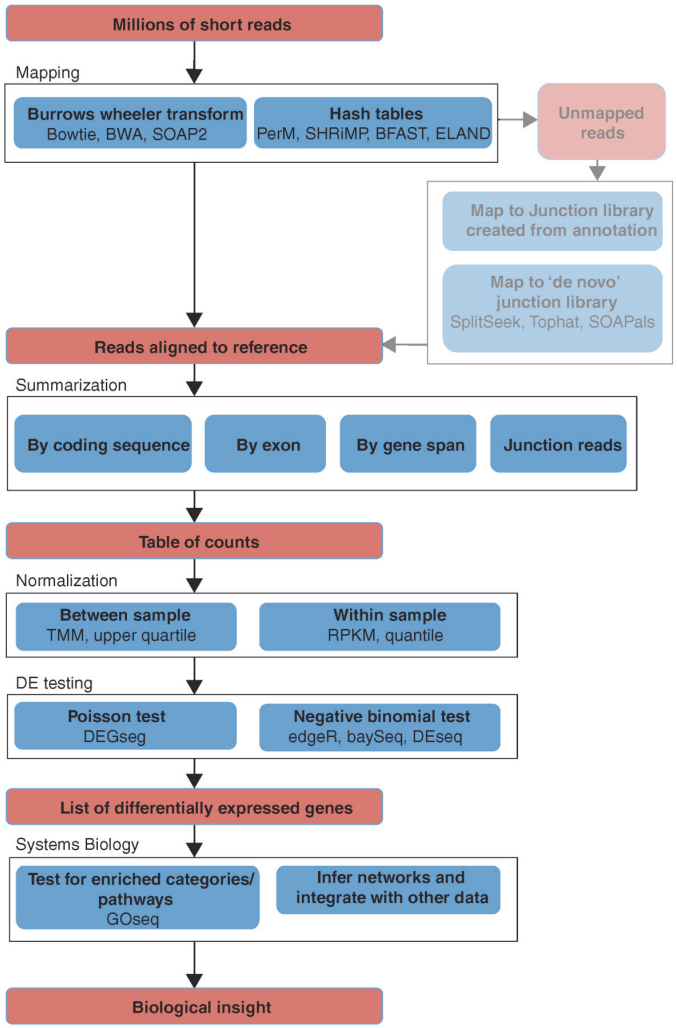
Mapping

Summarisation

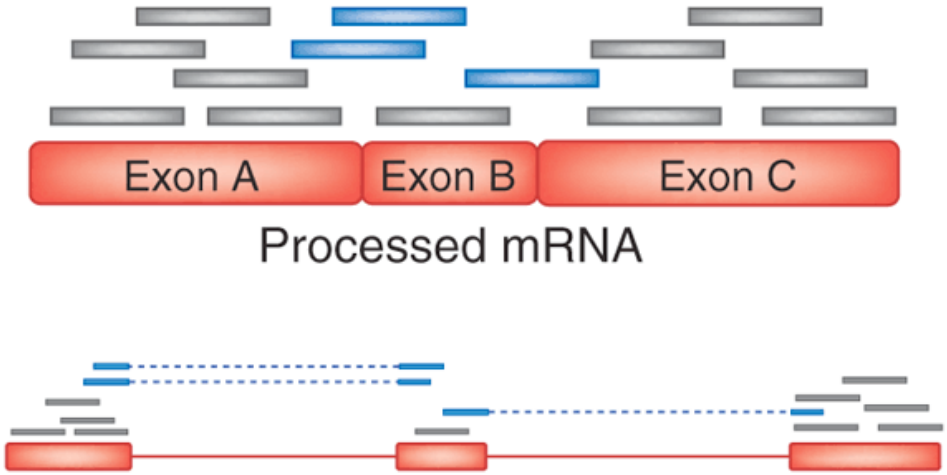
Normalisation

DE analysis

Functional analysis



Reference-based assembly



Genome mapping

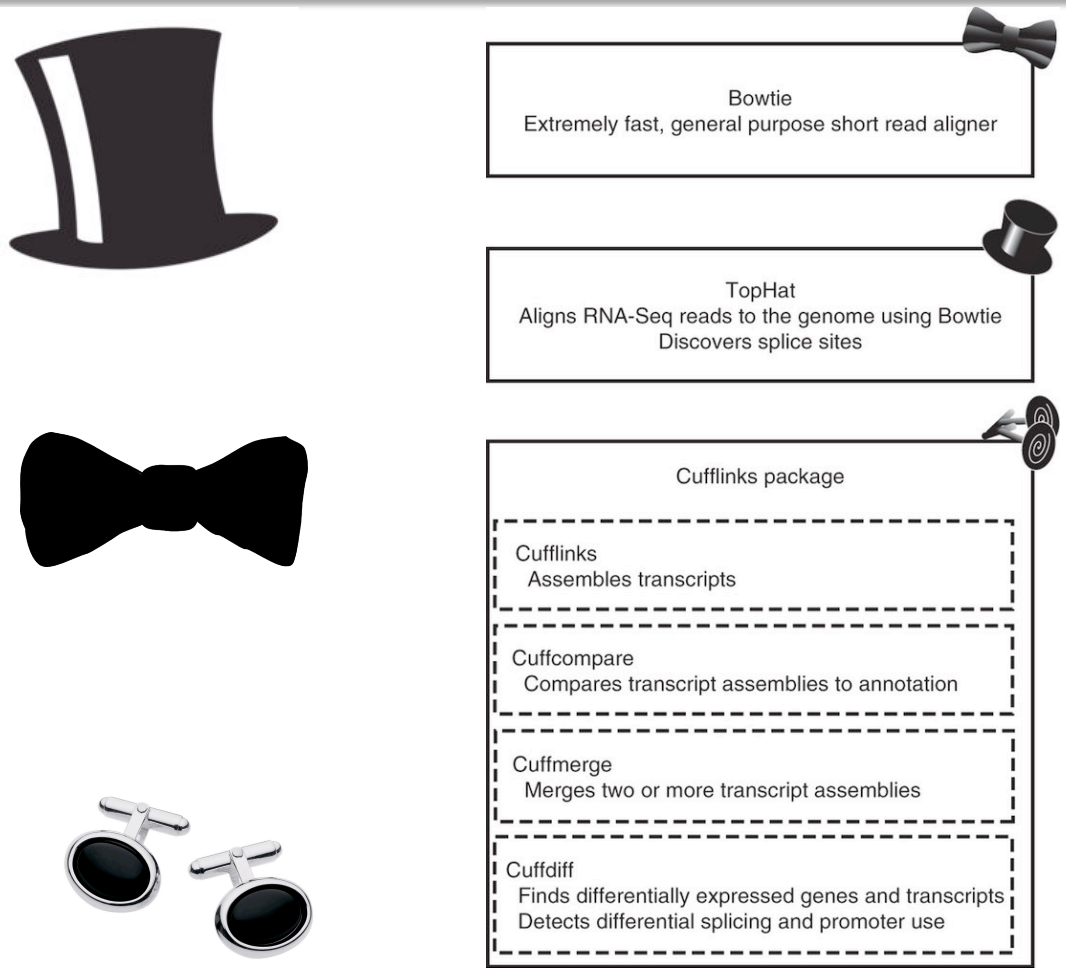
- Can identify novel features
- Splice aware?
- Can be difficult to reconstruct isoform and gene structures

Transcriptome mapping

- No repetitive reference
- Novel features?
- How reliable is the transcriptome?

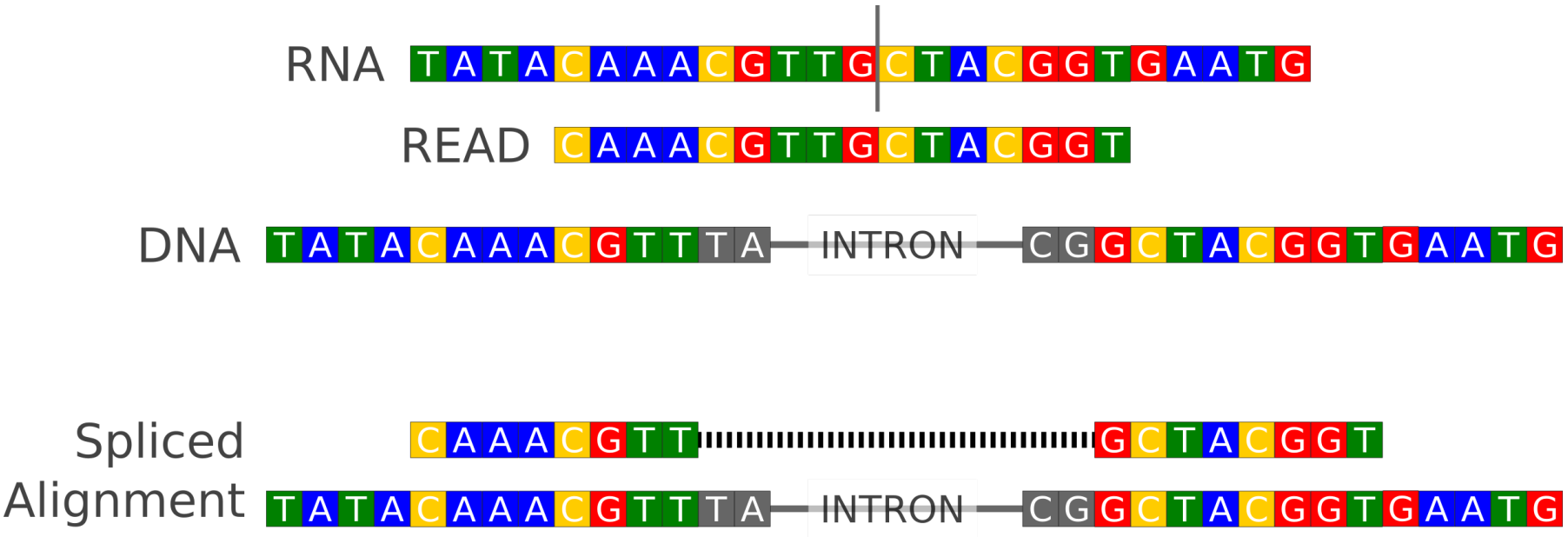
Trapnell & Salzberg (2009) Nature Biotech

A smart suit(e) for RNA-seq analysis

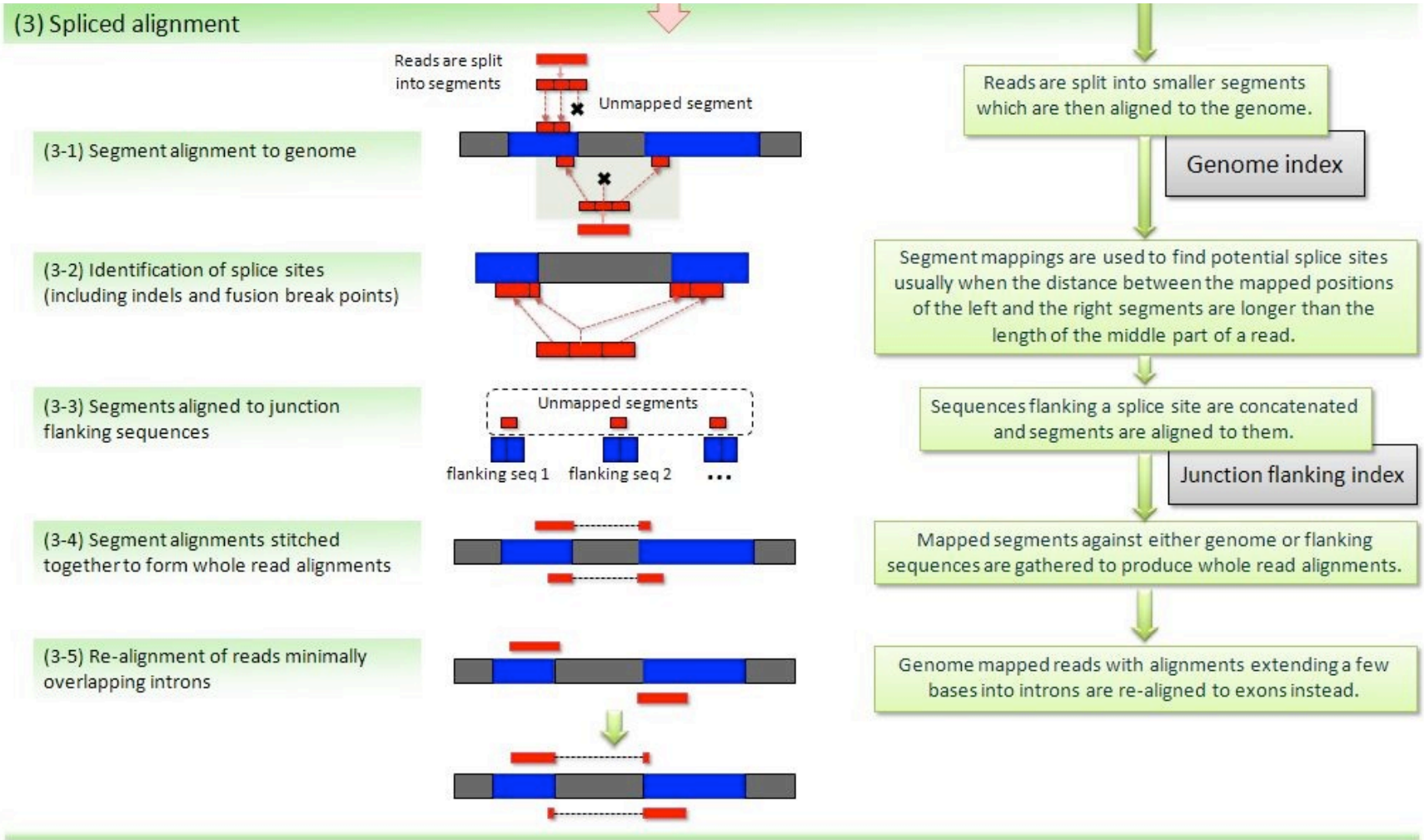


Trapnell, C. et al (2012) Nature Protocols

Spliced Alignment

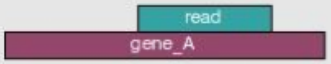
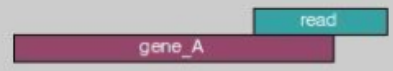
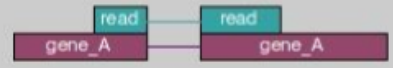
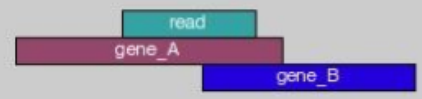


Spliced Alignment with Tophat/Bowtie



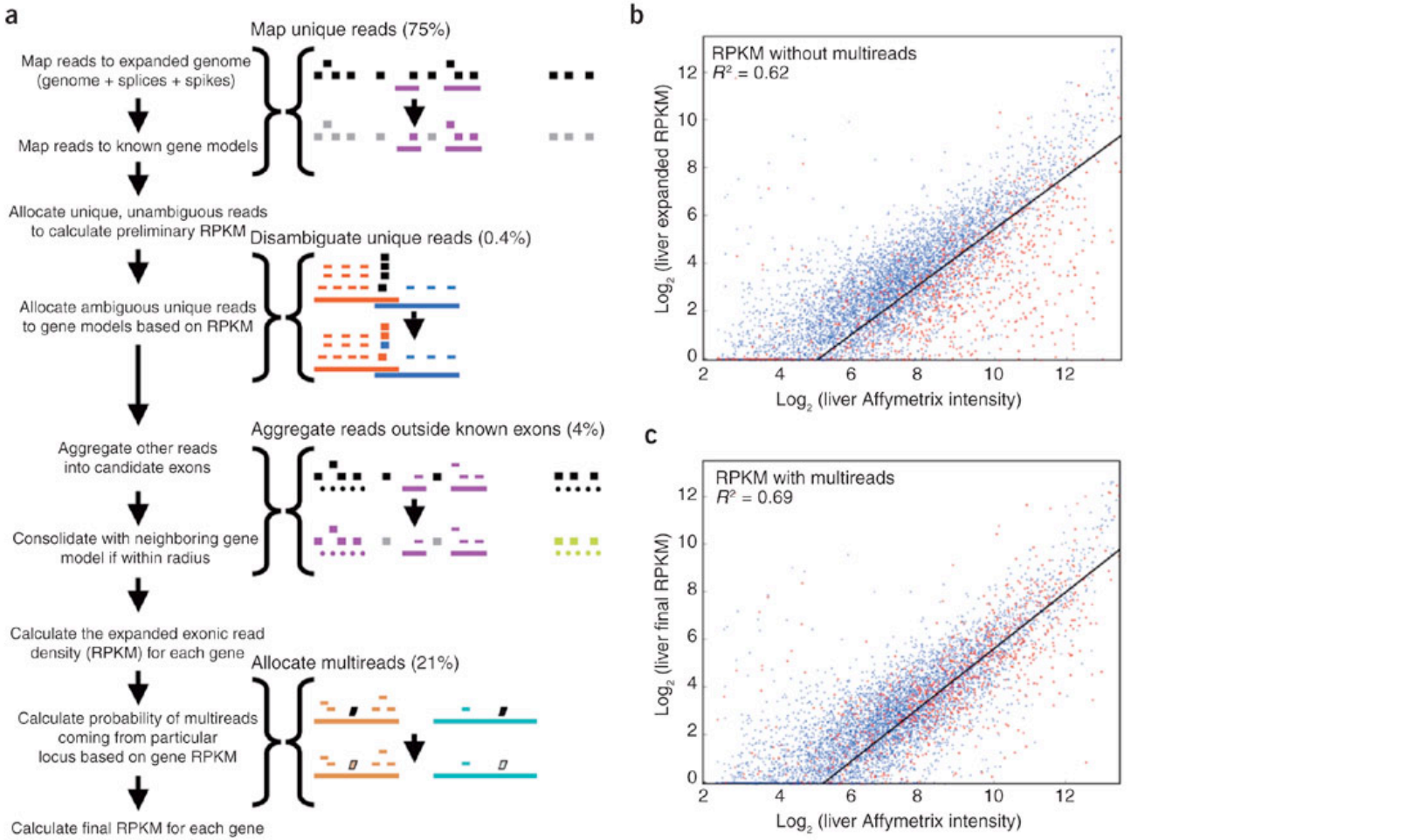
Kim, D. et al (2012) *Genome Biology*

Summarisation/Counting

	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

e.g. Htseq or Subread

Summarisation/Counting

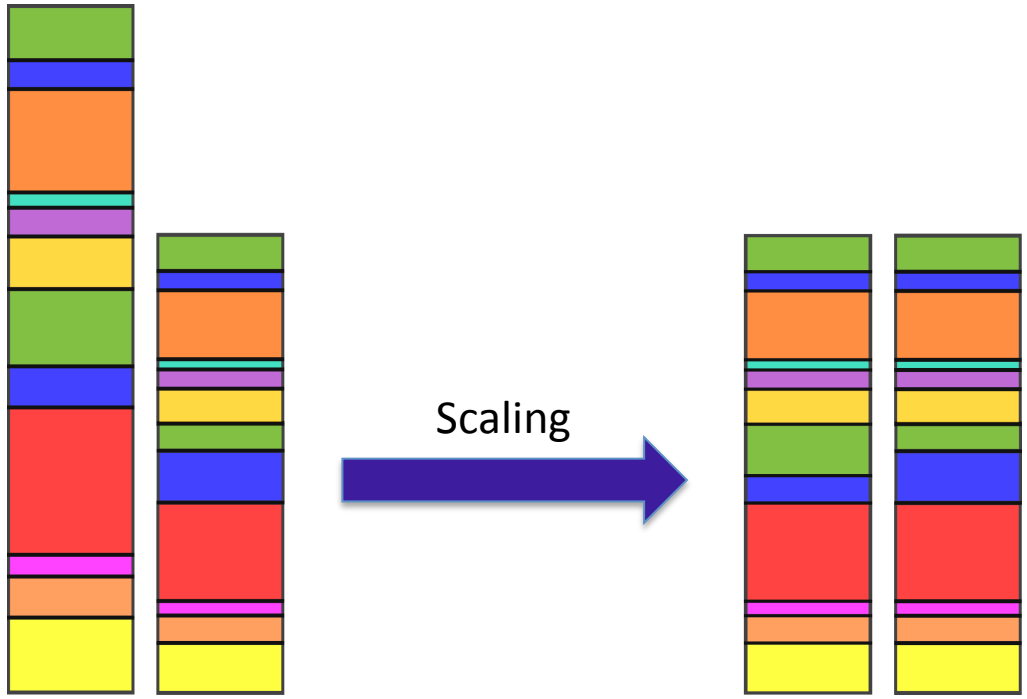


Mortazavi, A. et al (2008) *Nature Methods*

Normalisation - Scaling

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 - TPM

$$\frac{\text{reads for gene } A}{\text{length of gene } A \div 1000} \rightarrow \text{RPK for gene A}$$

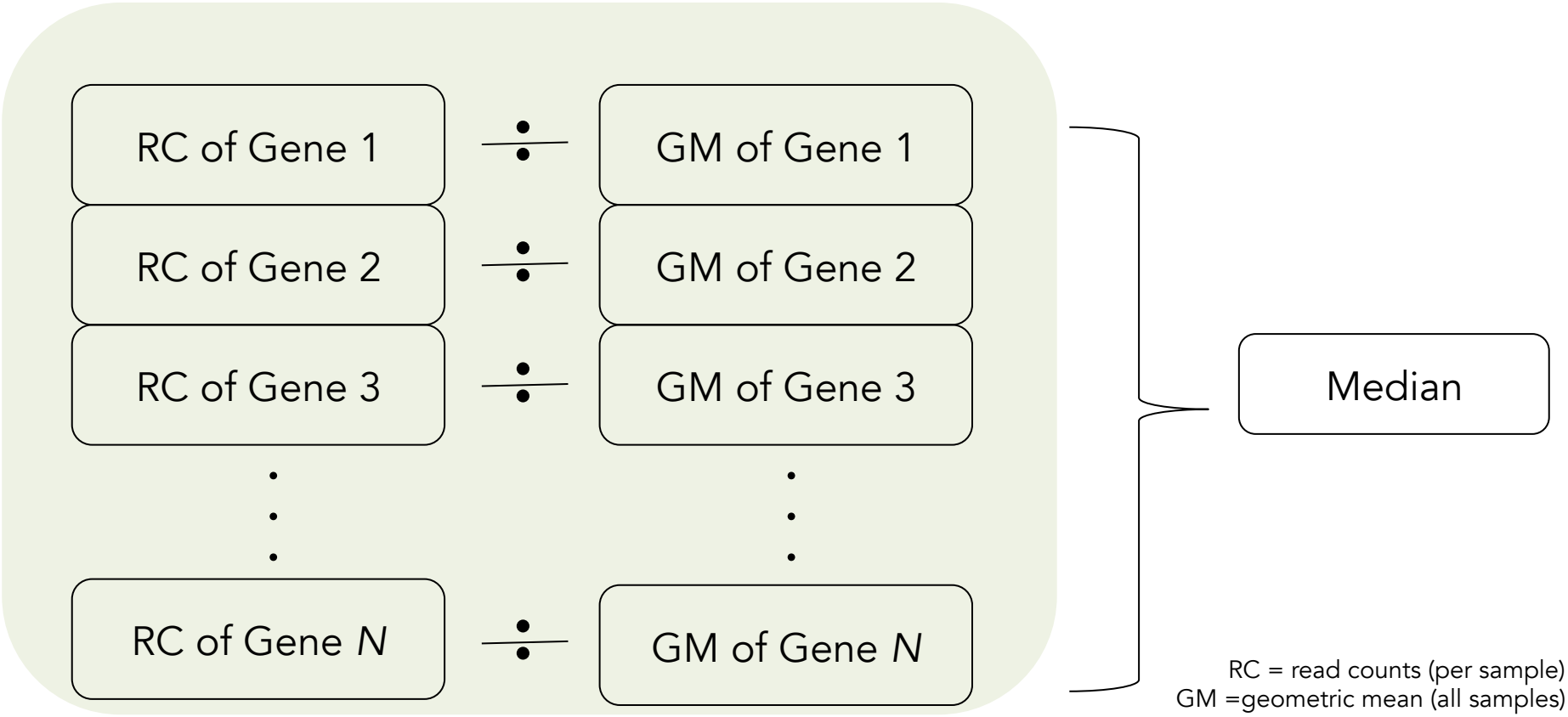
$$\frac{\text{sum of all RPKs}}{1{,}000{,}000} \rightarrow \text{Scaling factor}$$

$$\frac{\text{RPK for gene } A}{\text{Scaling factor}} \rightarrow \text{TPM for gene A}$$

Normalisation – Geometric Scaling

Geometric scaling factor

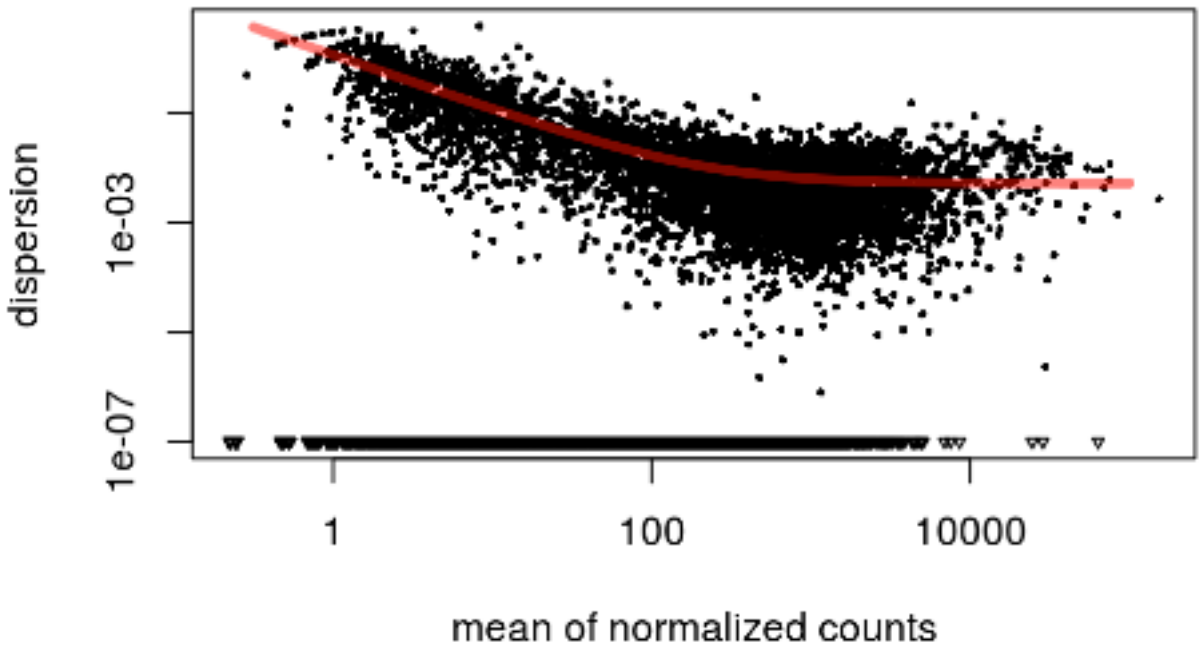
- Assumes that most genes are not differentially expressed



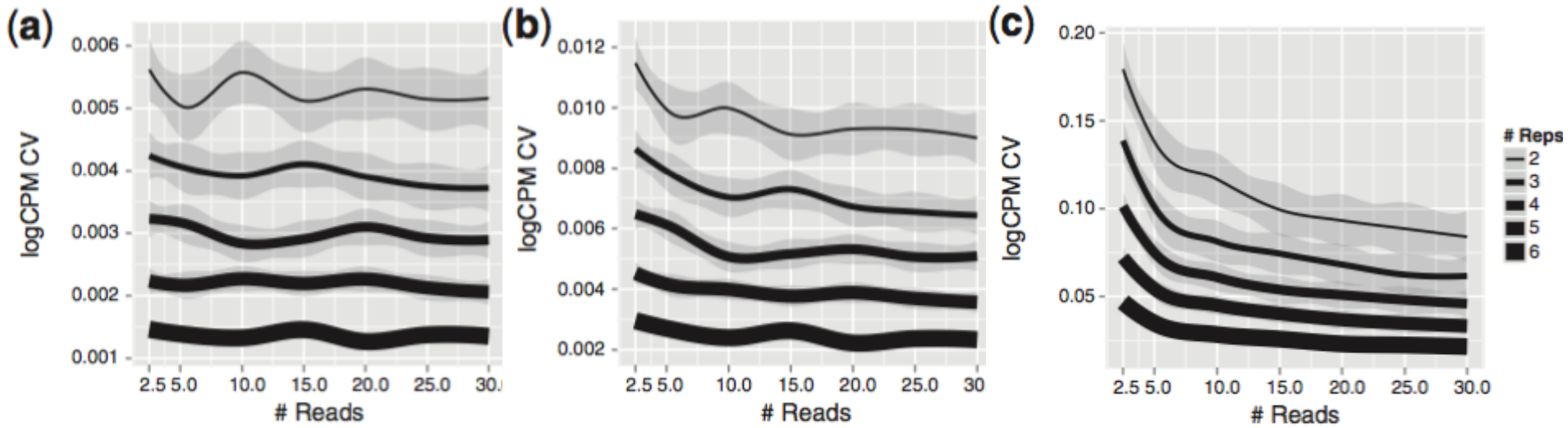
Modelling – in fashion

- DESeq uses a similar formulation of the variance term

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



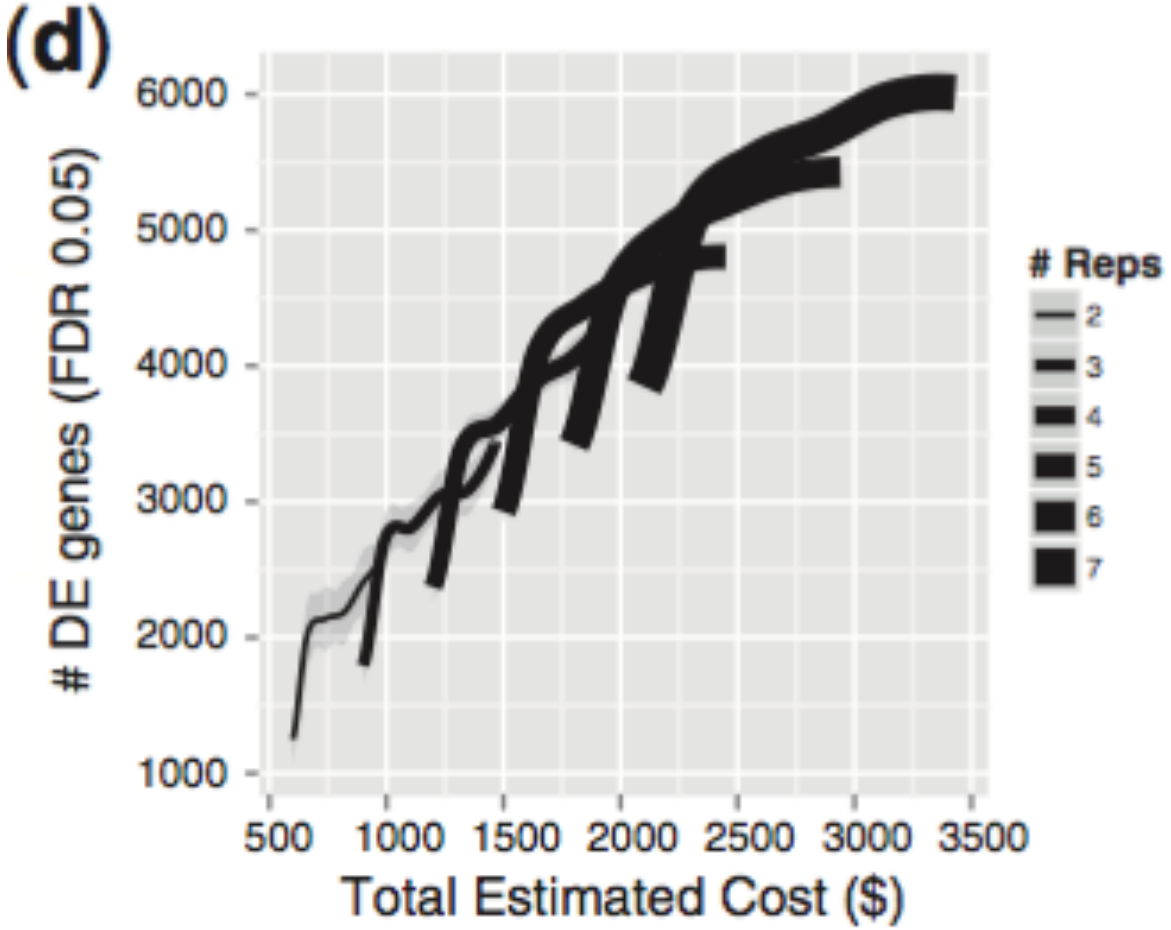
Replicates v Sequencing Depth



HIGH **MEDIUM** **LOW**

Liu et al. (2014) *Bioinformatics*

Replicates v Sequencing Depth



Liu et al. (2014) *Bioinformatics*