

# Analysis of RNA-seq Data

Ashley Sawle  
based on slides by Bernard Pereira

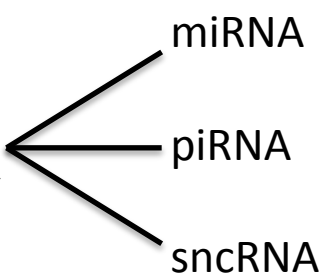


UNIVERSITY OF  
CAMBRIDGE



CANCER  
RESEARCH  
UK

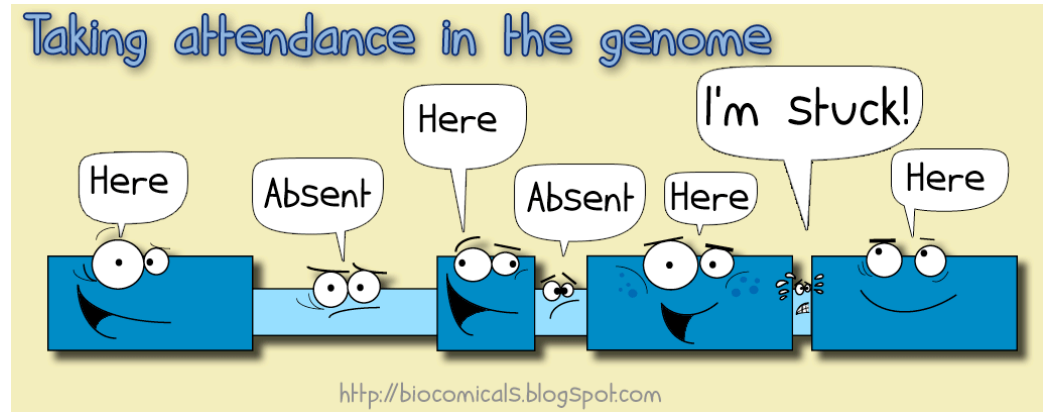
# The many faces of RNA-seq – Techniques

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

# The many faces of RNA-seq – Applications

## Discovery

- Transcripts
- Isoforms
- Splice junctions
- Fusion genes

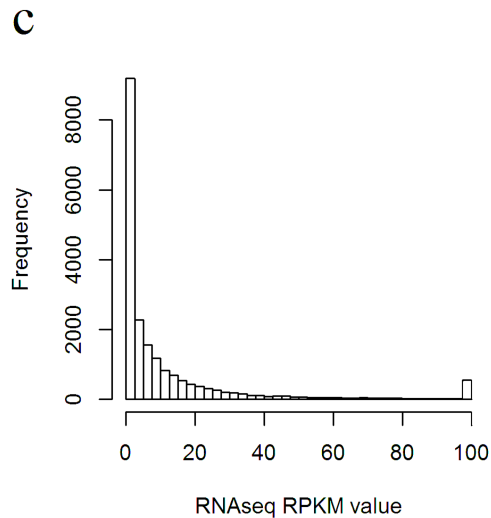
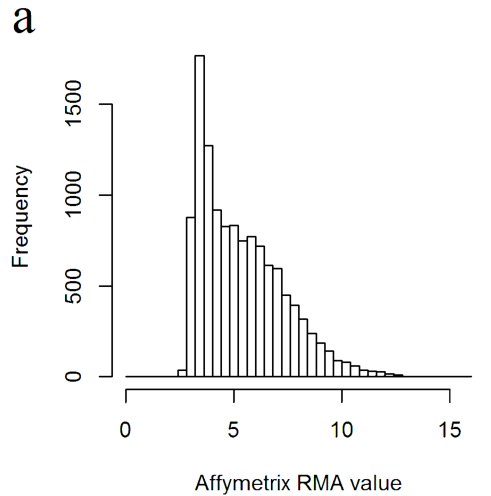


## Differential expression

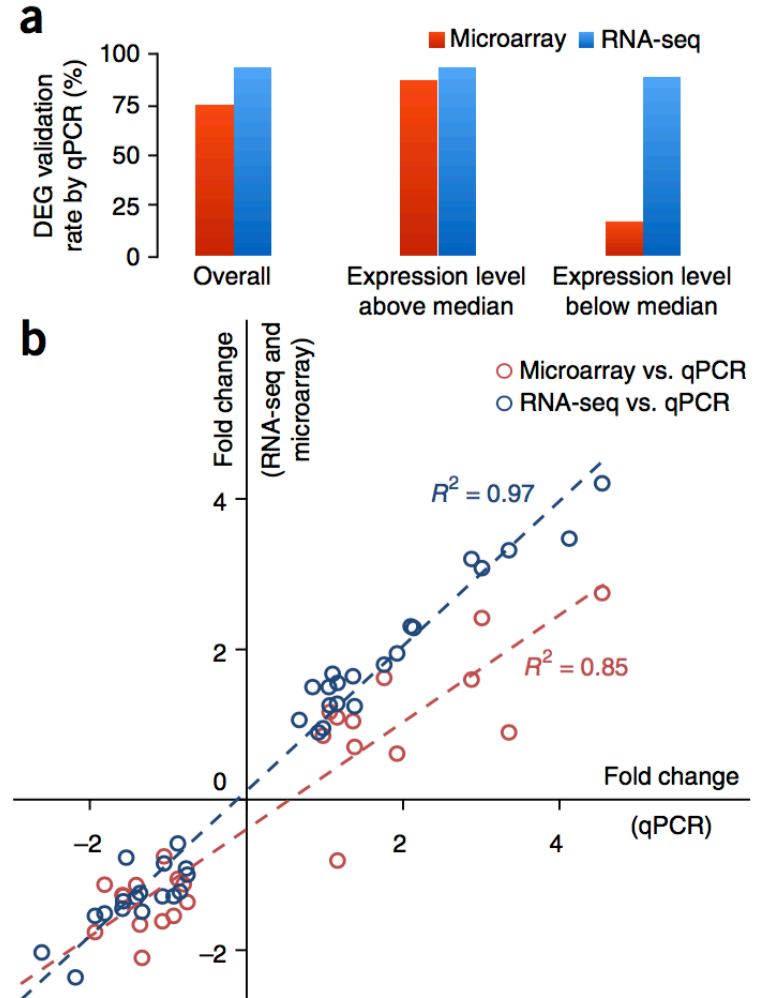
- **Gene level expression changes**
- Relative isoform abundance
- Splicing patterns

## Variant calling

# Microarray → RNA-seq



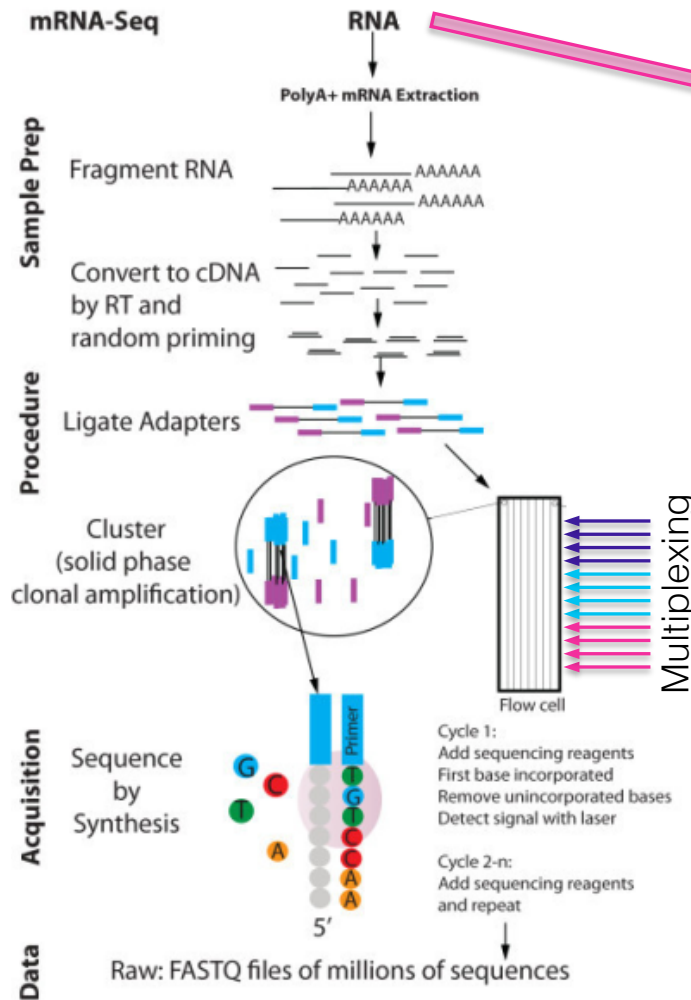
Guo et al. (2013) *Plos One*



Wang et al (2014) *Nature Biotech.*

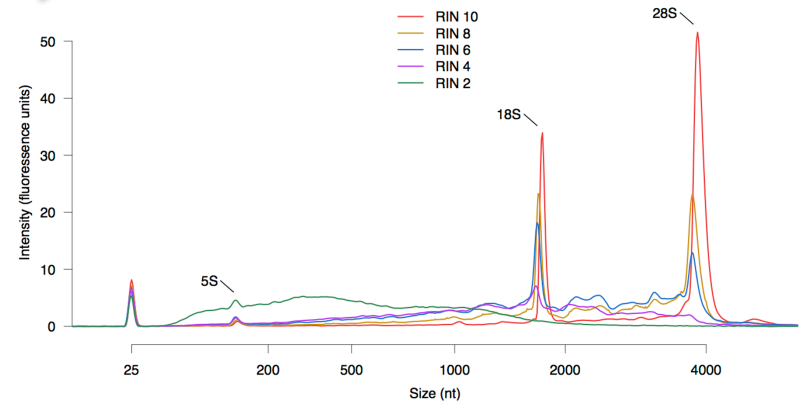


# Library Preparation & Sequencing



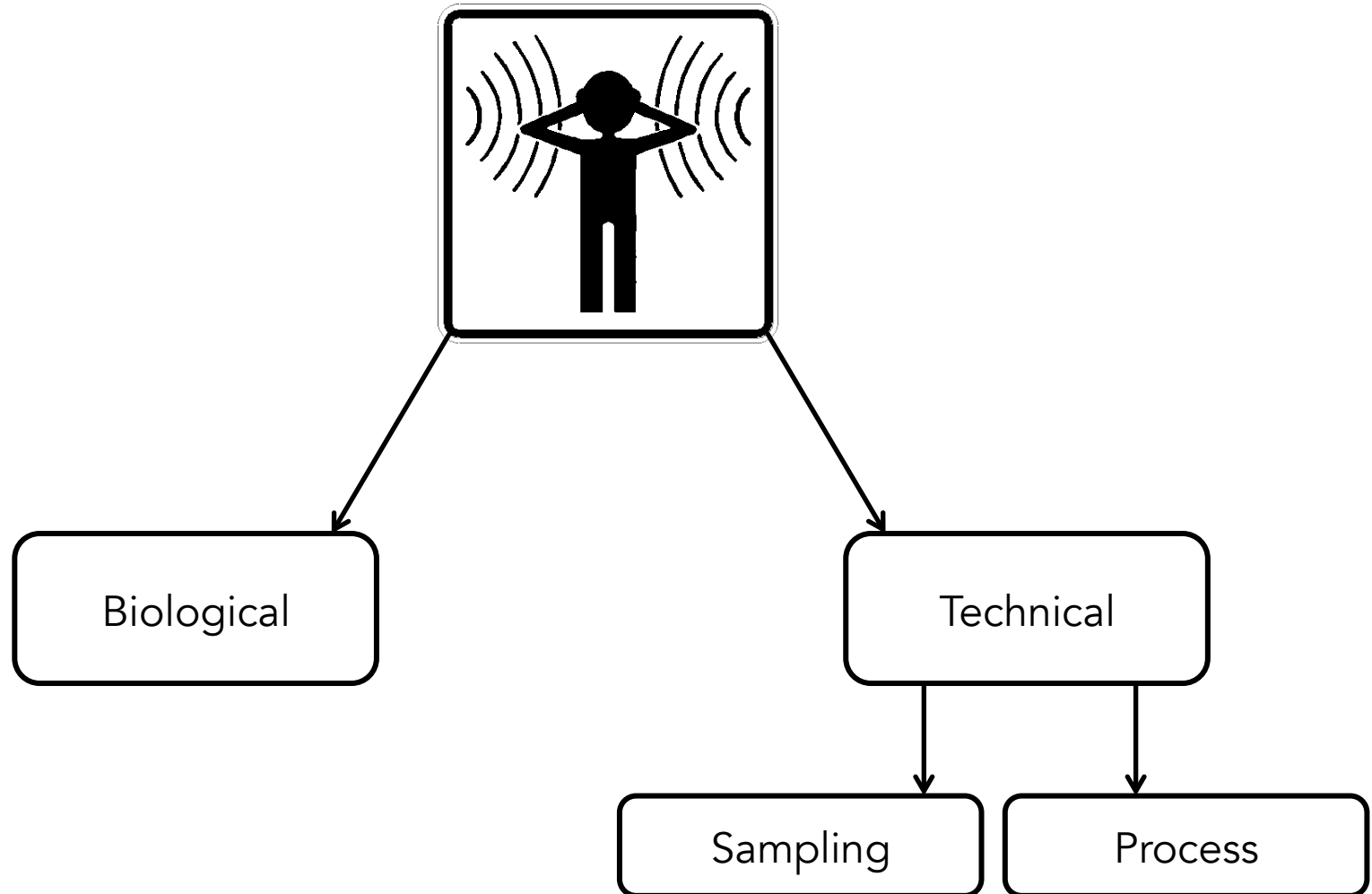
## QC - RIN number

Effects of degradation on RNA size distribution

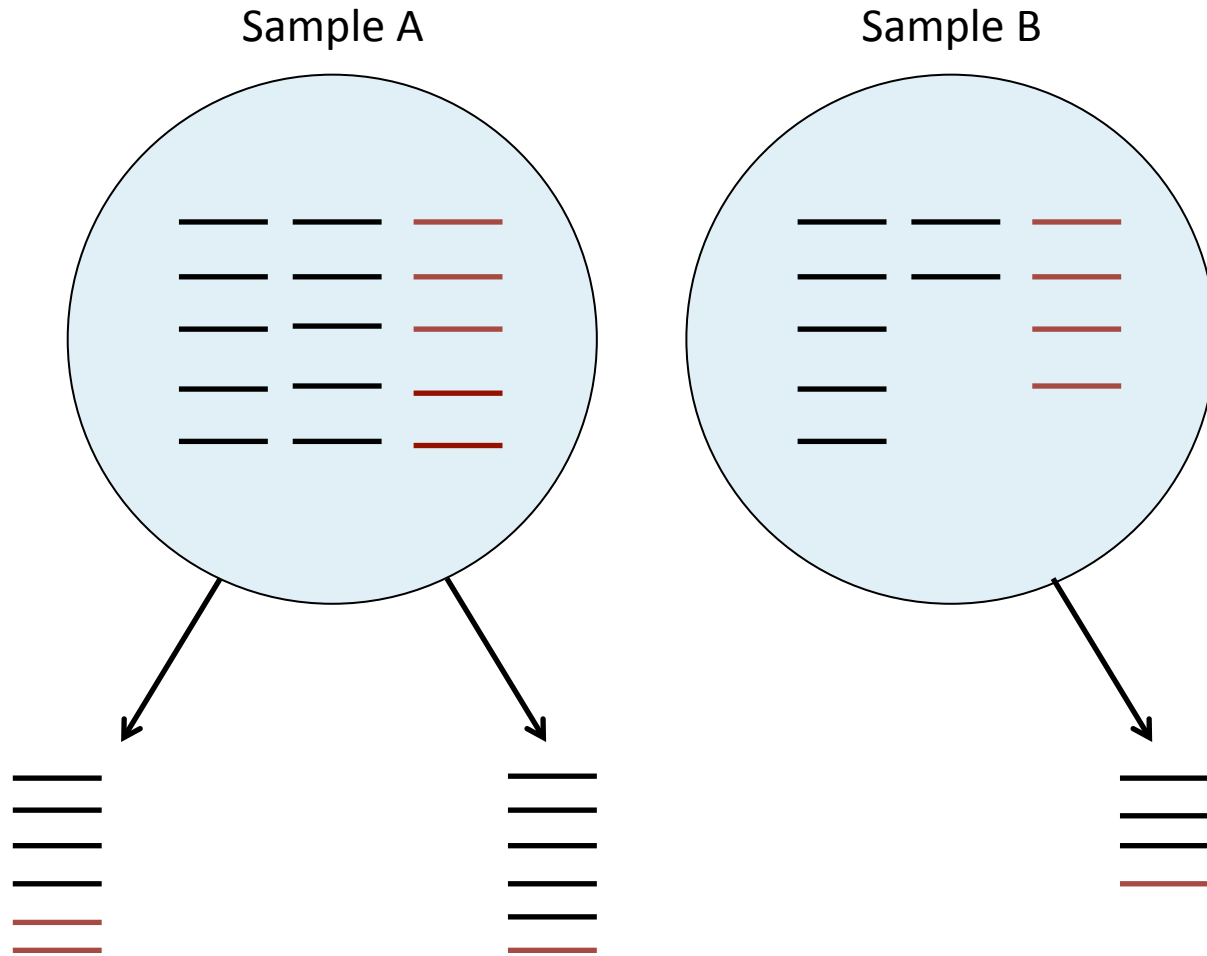


Sigurgeirsson, Emanuelsson & Lundberg (2014) PLOS ONE

# Sources of Noise

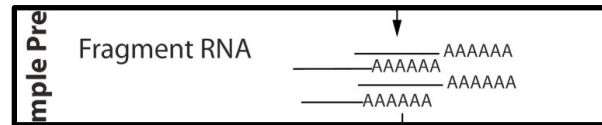


# Sources of Noise – Sampling Bias

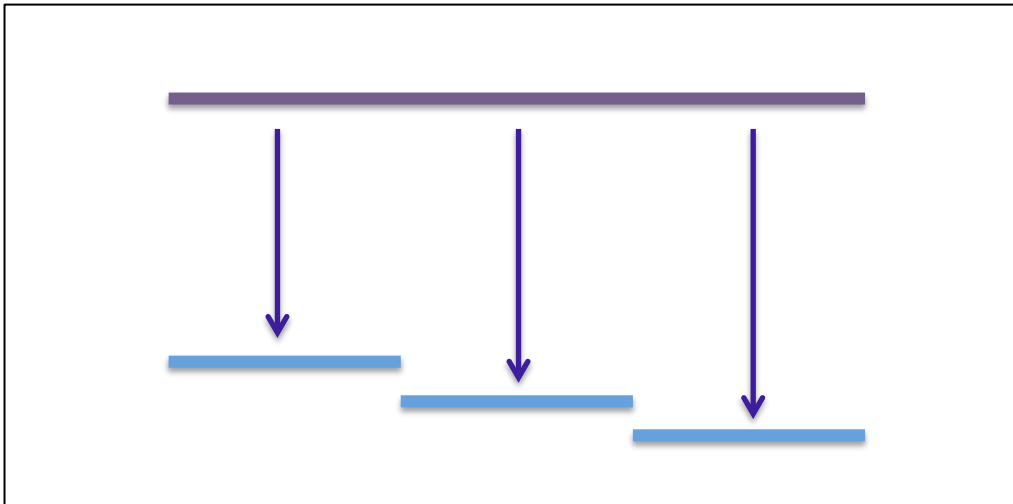


Subsampling a from a pool of RNAs

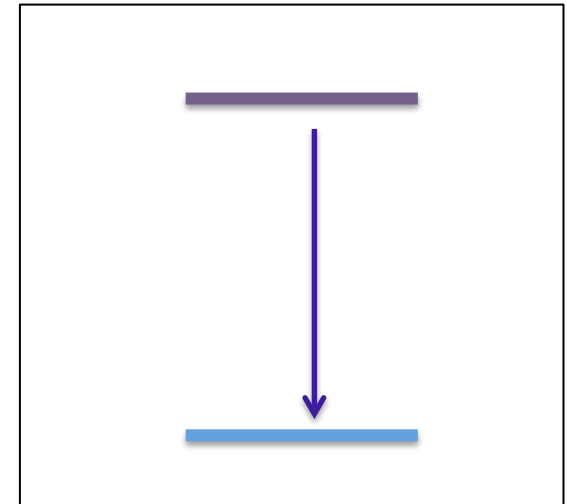
# Sources of Noise – Sampling Bias



Transcript A



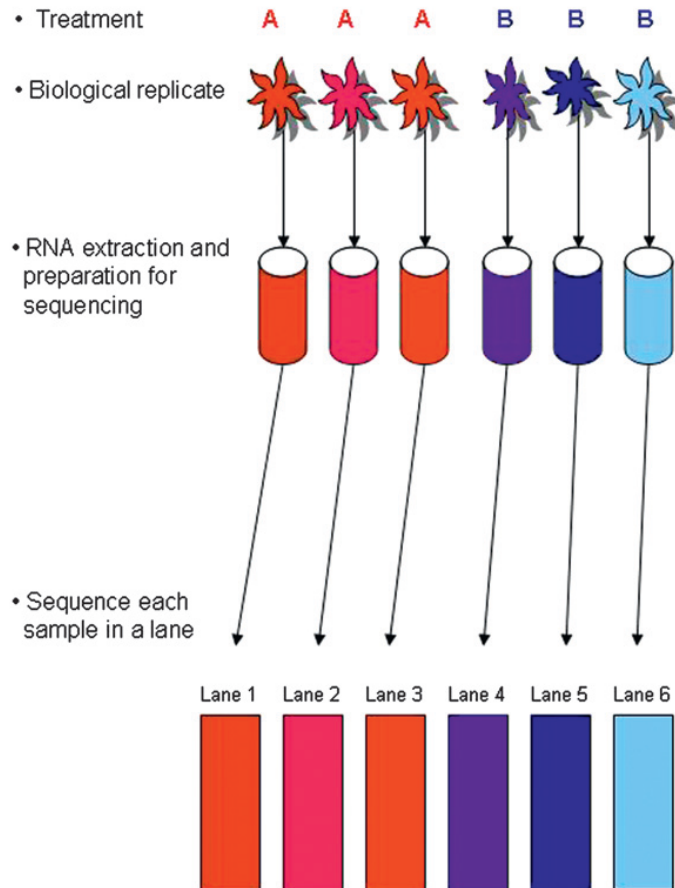
Transcript B



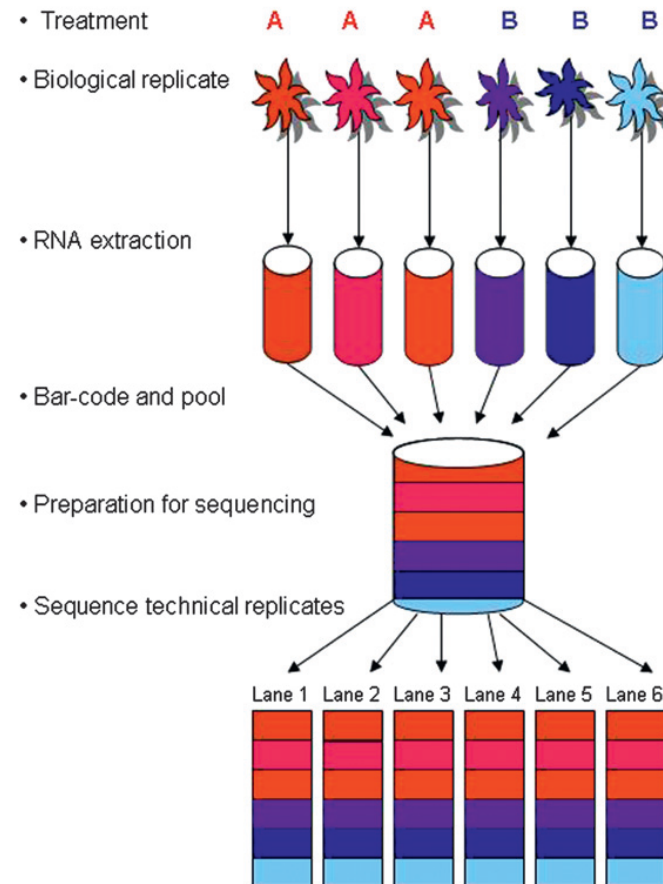
Transcript length affects the number of RNA fragments present in the library from that gene

# Sources of Noise - Process

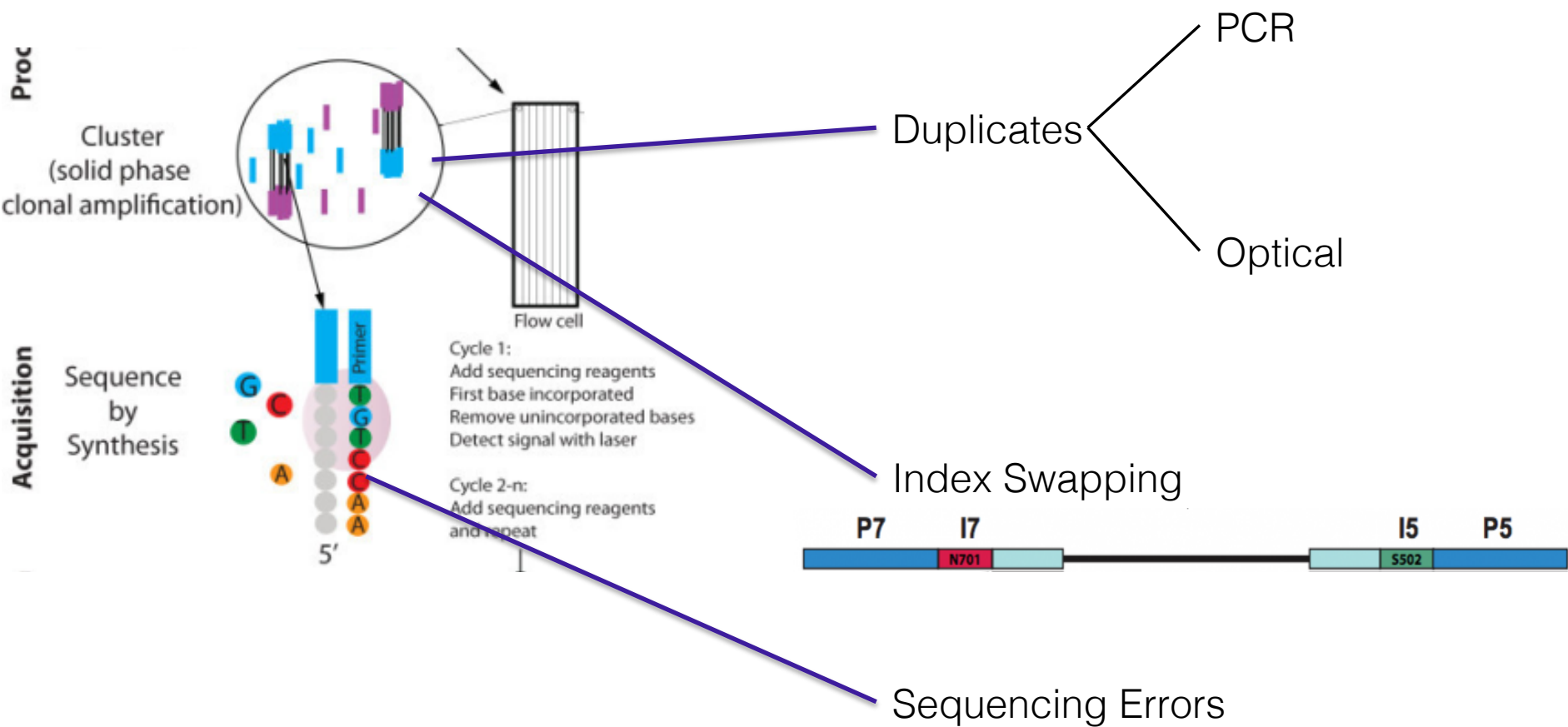
## Confounded Design



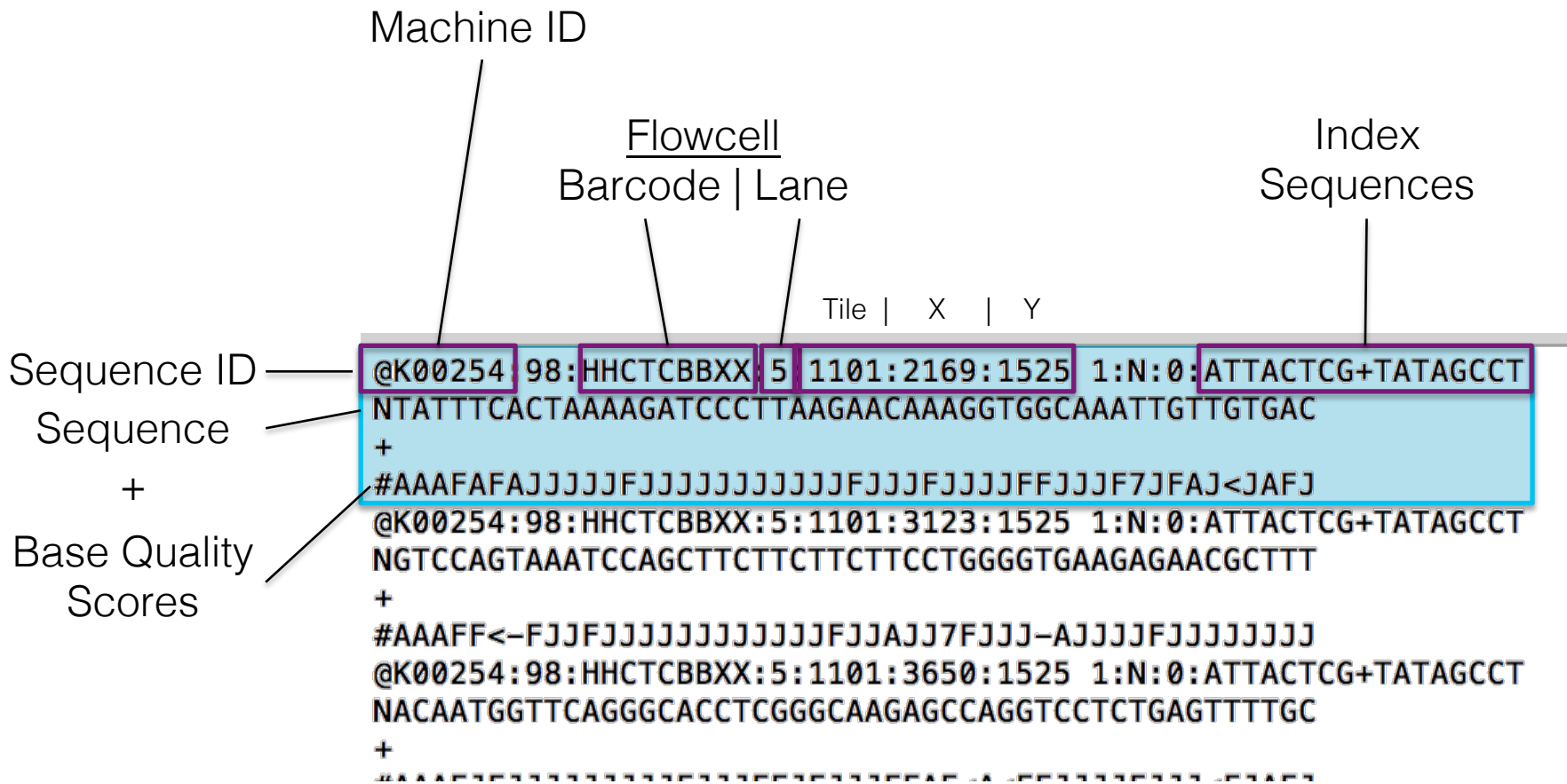
## Balanced Blocked Design



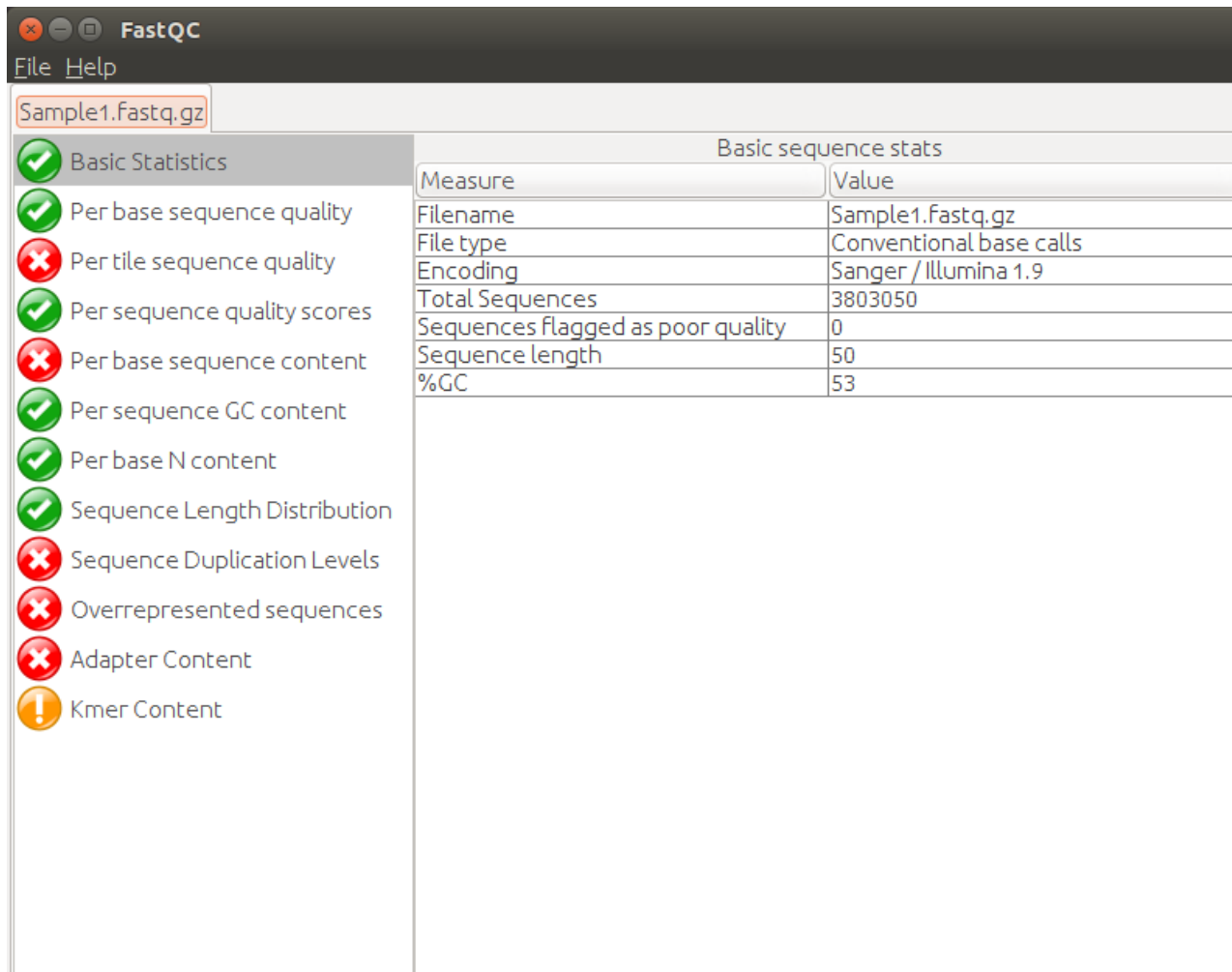
# Sources of Noise – Process



# Raw Sequence – FASTQ files



# Raw Sequence QC - FASTQC



FastQC

File Help

Sample1.fastq.gz

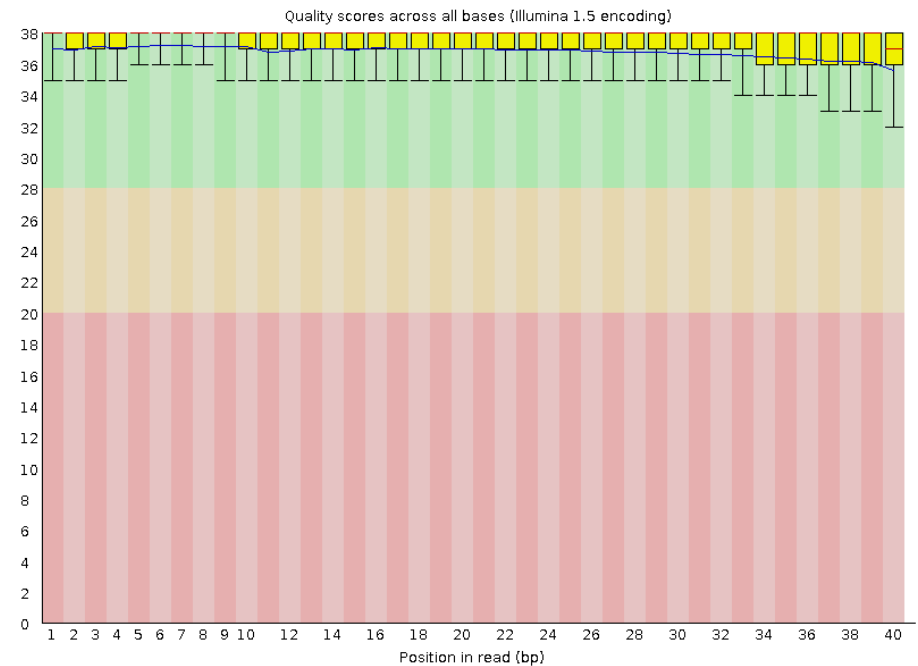
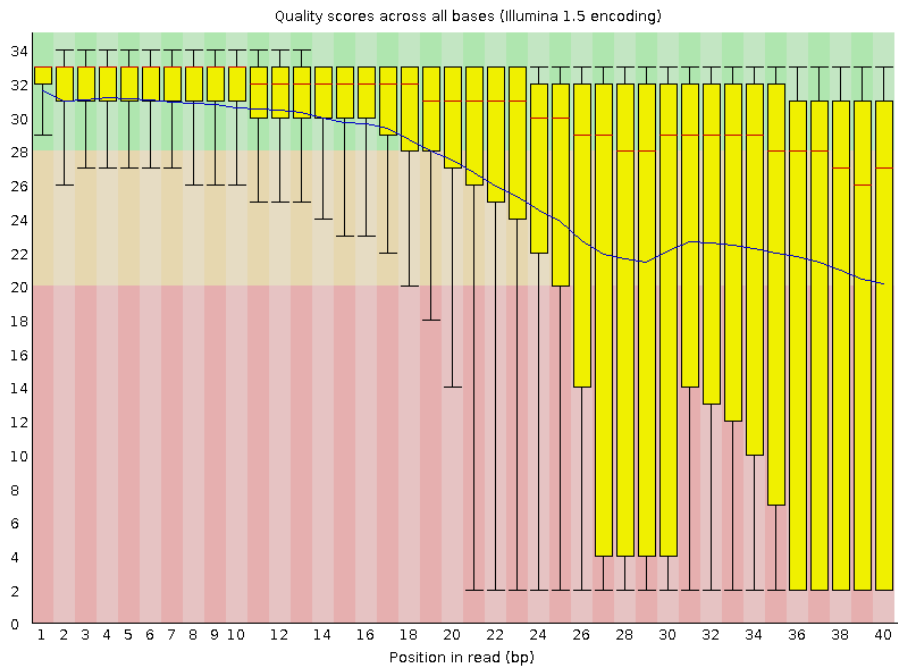
Basic sequence stats

Measure	Value
Filename	Sample1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	3803050
Sequences flagged as poor quality	0
Sequence length	50
%GC	53

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

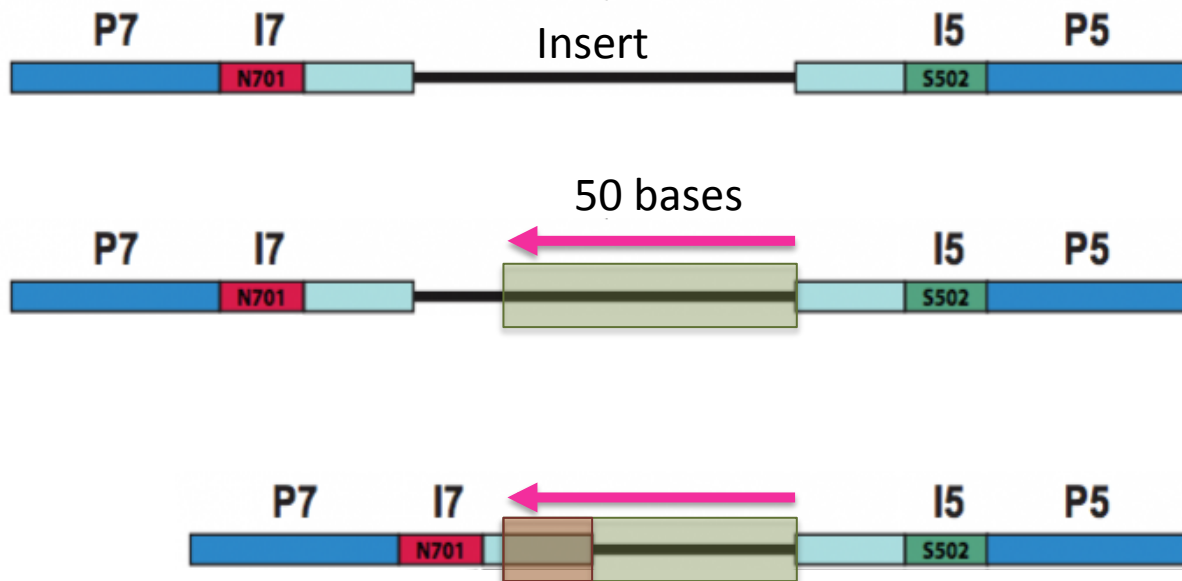


# Raw Sequence QC - FASTQC

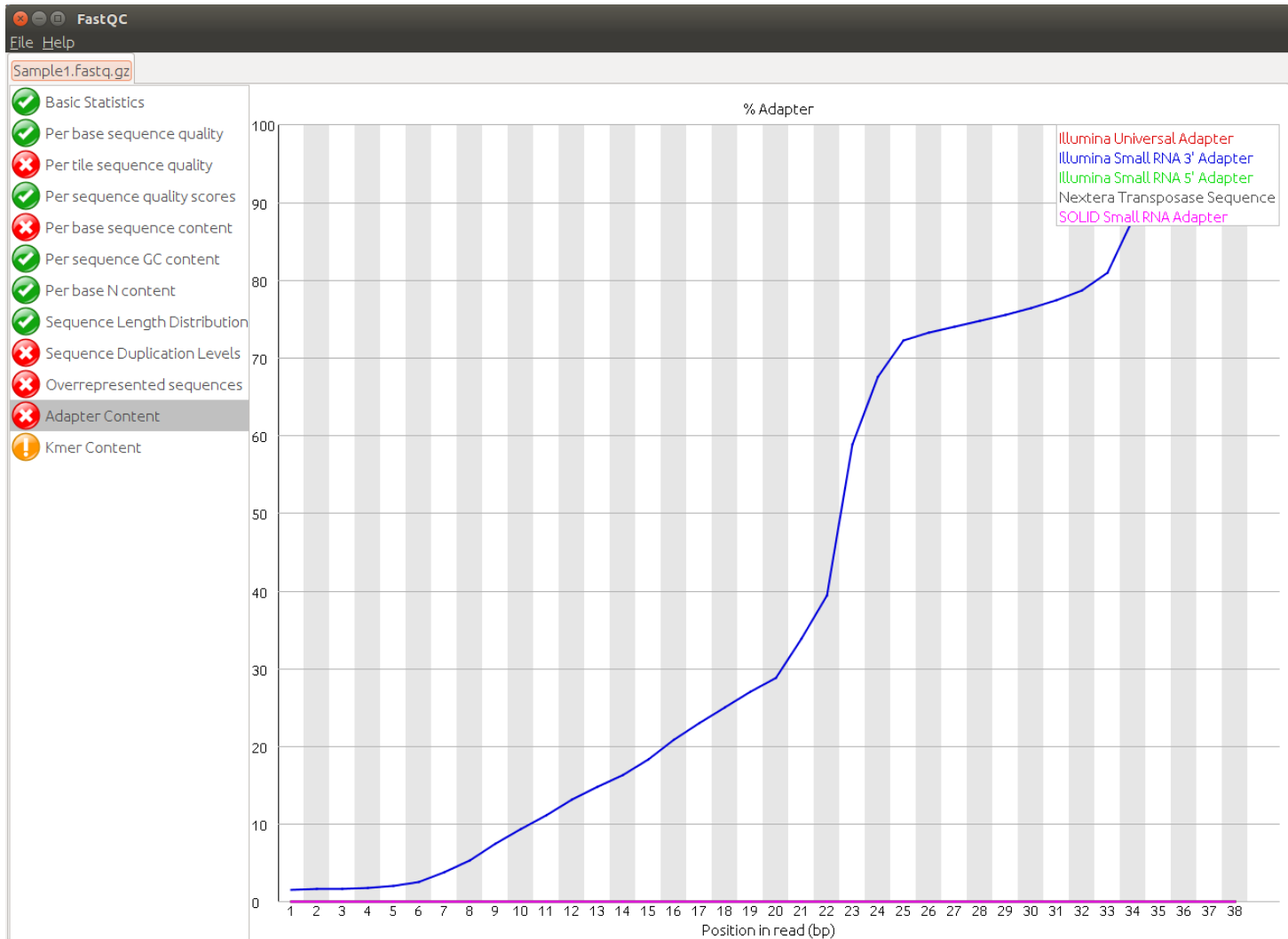


# Trimming

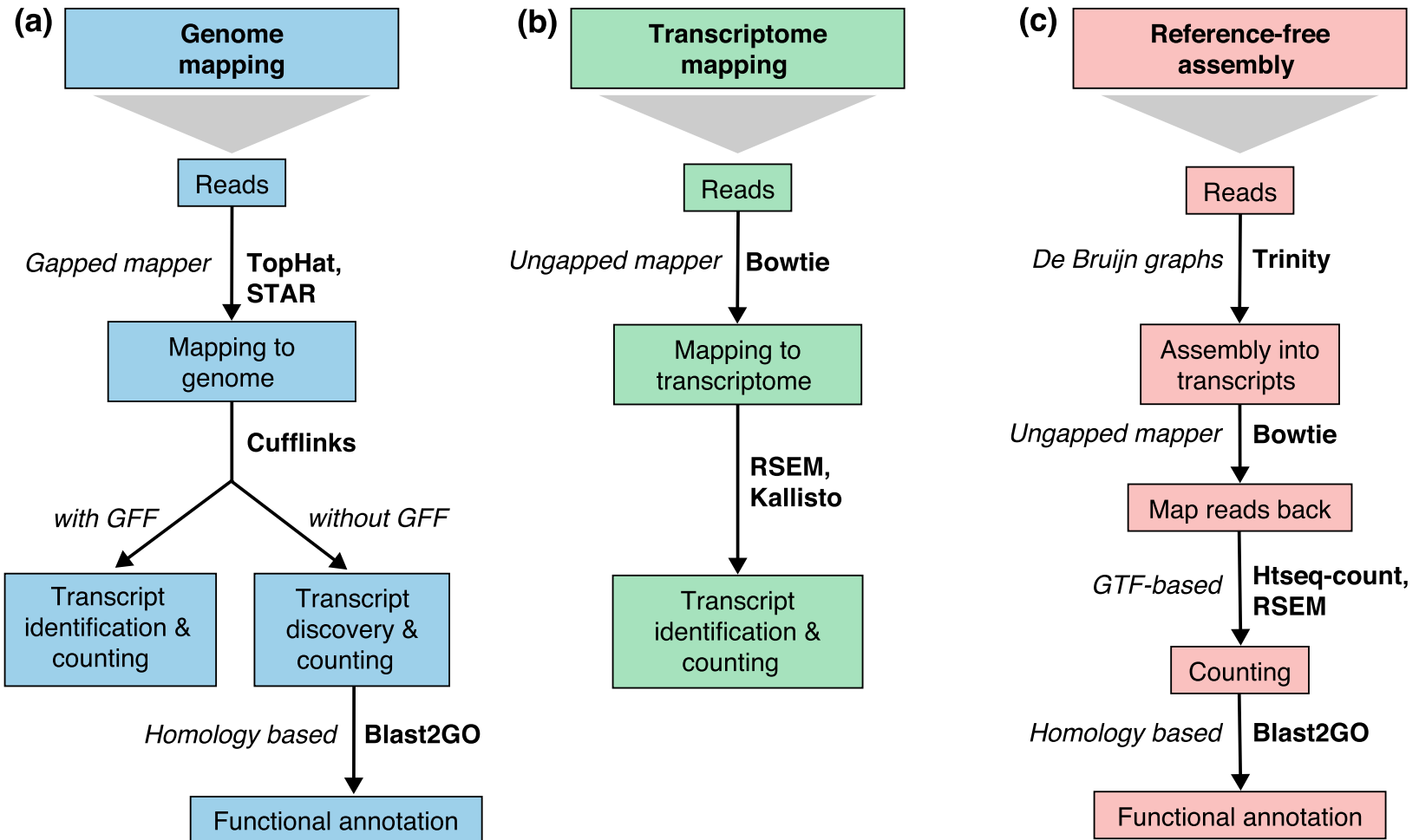
- Quality-based Trimming
- Adapter contamination



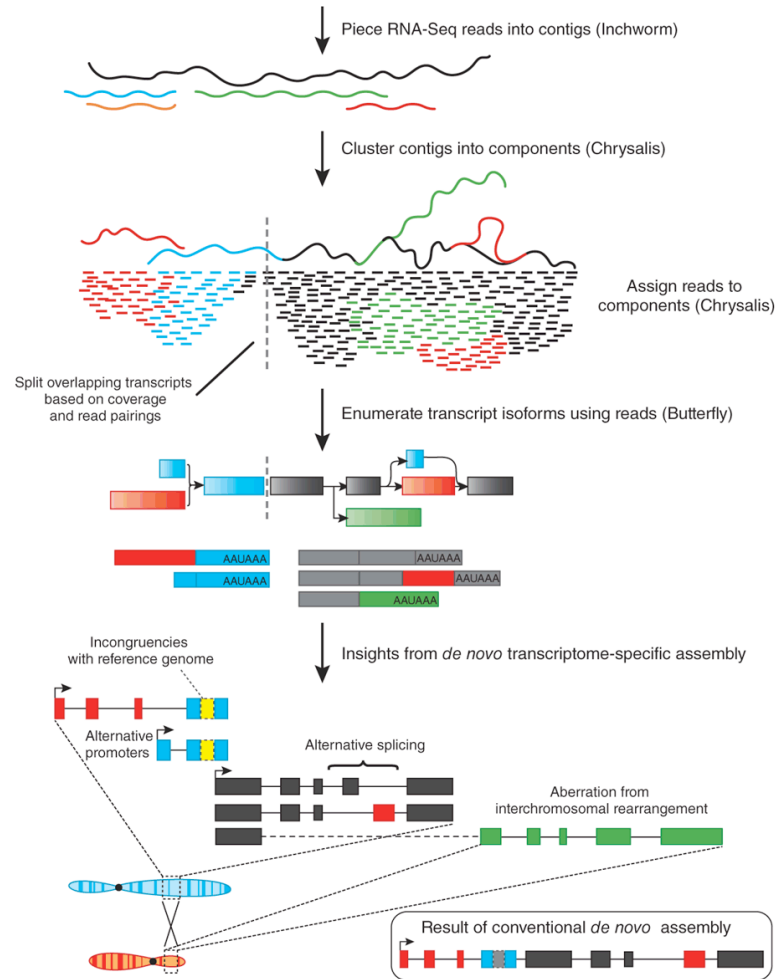
# Adapter contamination - FASTQC



# Sequence to Sense

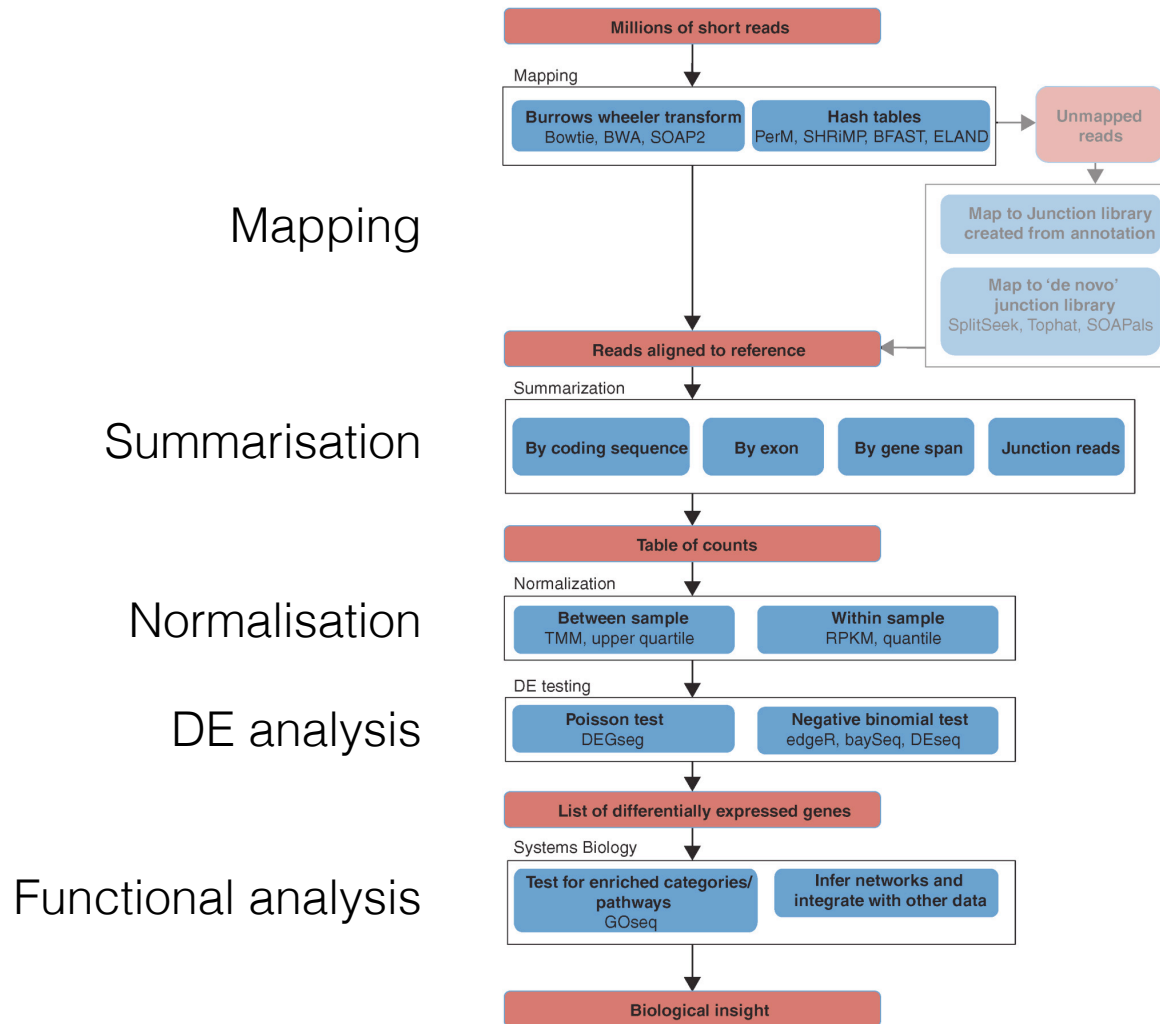


# De Novo assembly

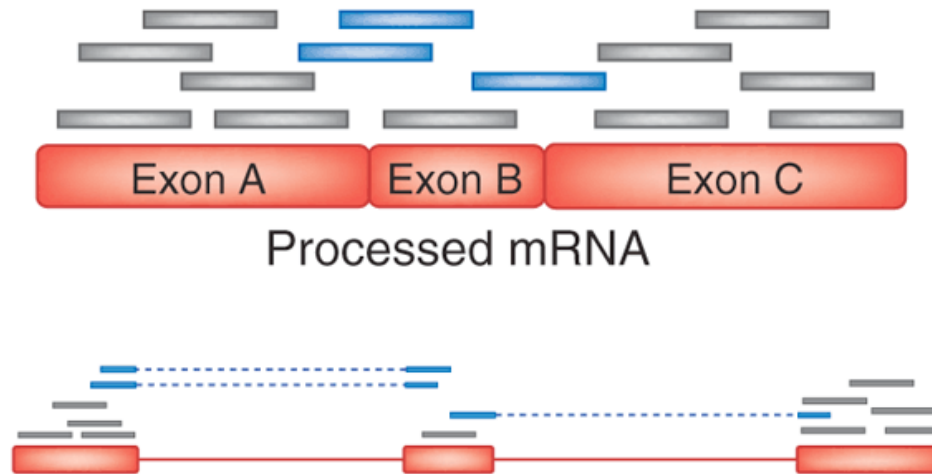


e.g. TRINITY

# Analysis Overview



# Reference-based assembly




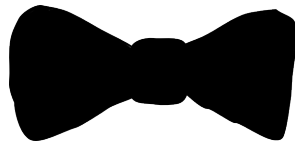
## Genome mapping

- 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?

# A smart suit(e) for RNA-seq analysis



**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



# Spliced Alignment

RNA T A T A C A A A C G T T G C T A C G G T G A A T G  
READ C A A A C G T T G C T A C G G T

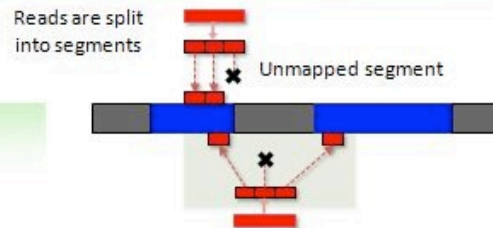
DNA T A T A C A A A C G T T T A INTRON C G G C T A C G G T G A A T G

Spliced Alignment  
C A A A C G T T ..... G C T A C G G T  
T A T A C A A A C G T T T A INTRON C G G C T A C G G T G A A T G

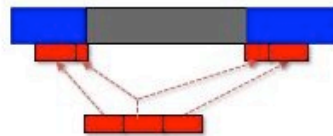
# Spliced Alignment with Tophat/Bowtie

## (3) Spliced alignment

(3-1) Segment alignment to genome



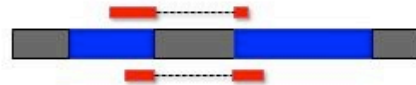
(3-2) Identification of splice sites (including indels and fusion break points)



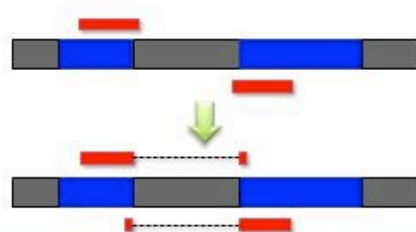
(3-3) Segments aligned to junction flanking sequences



(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.

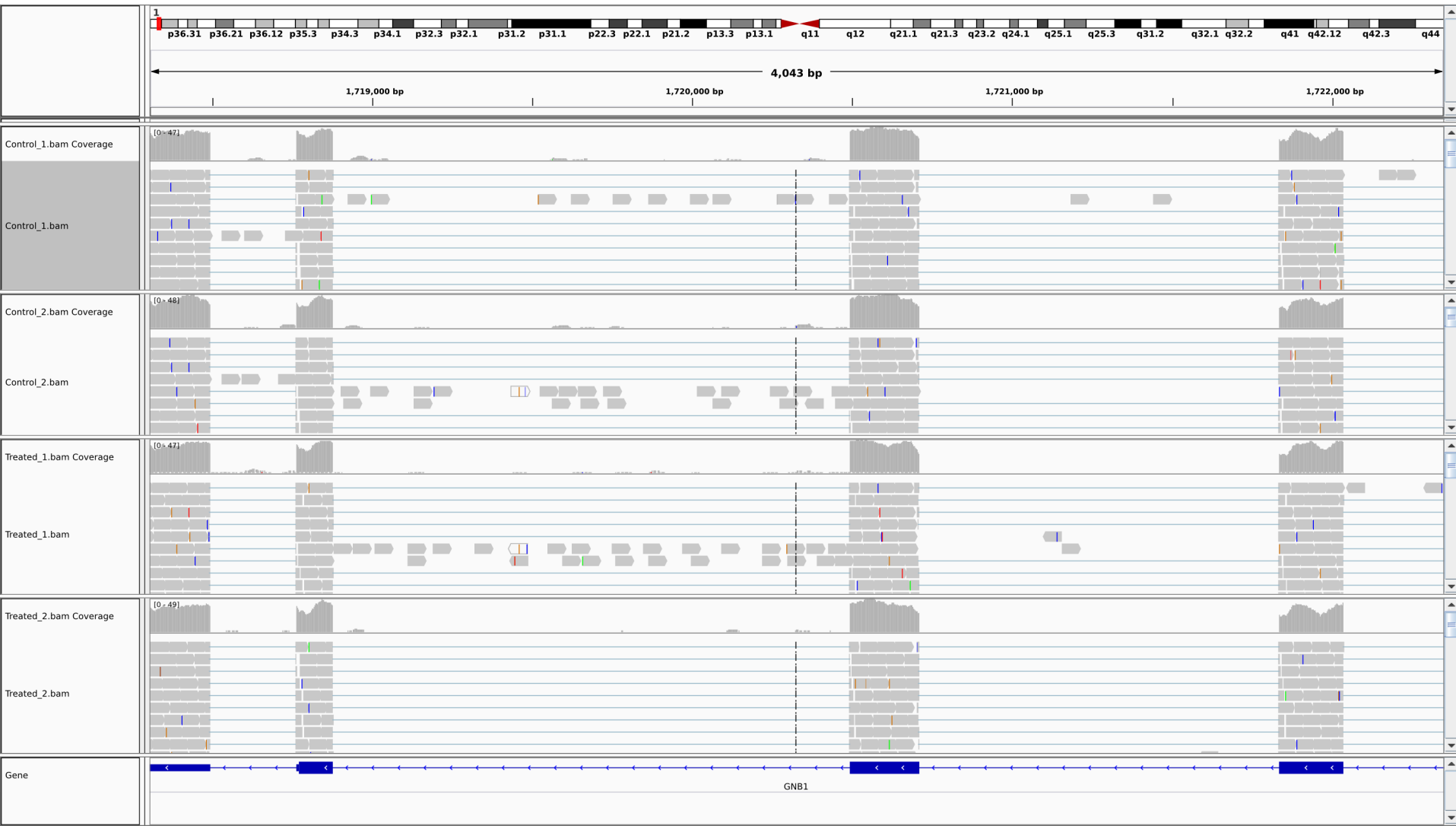
Sequences flanking a splice site are concatenated and segments are aligned to them.

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.

# Visualising Mapping Results – IGV



# Summarisation/Counting

(b)



## Genome-based features

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

## Transcript-based features

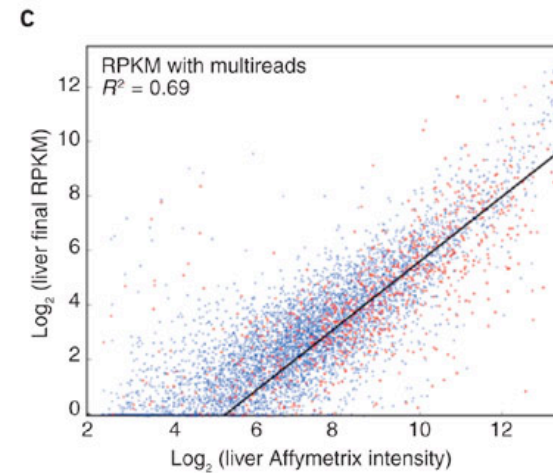
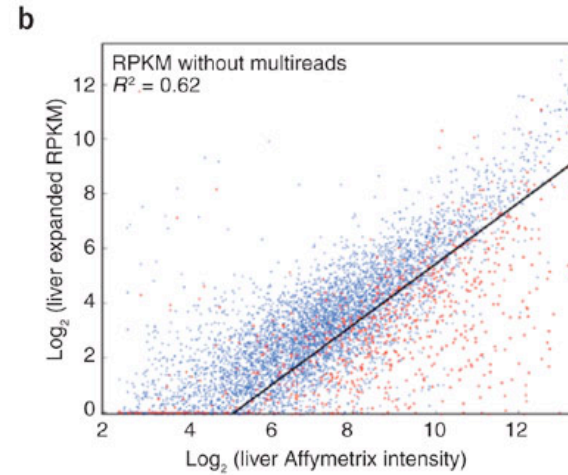
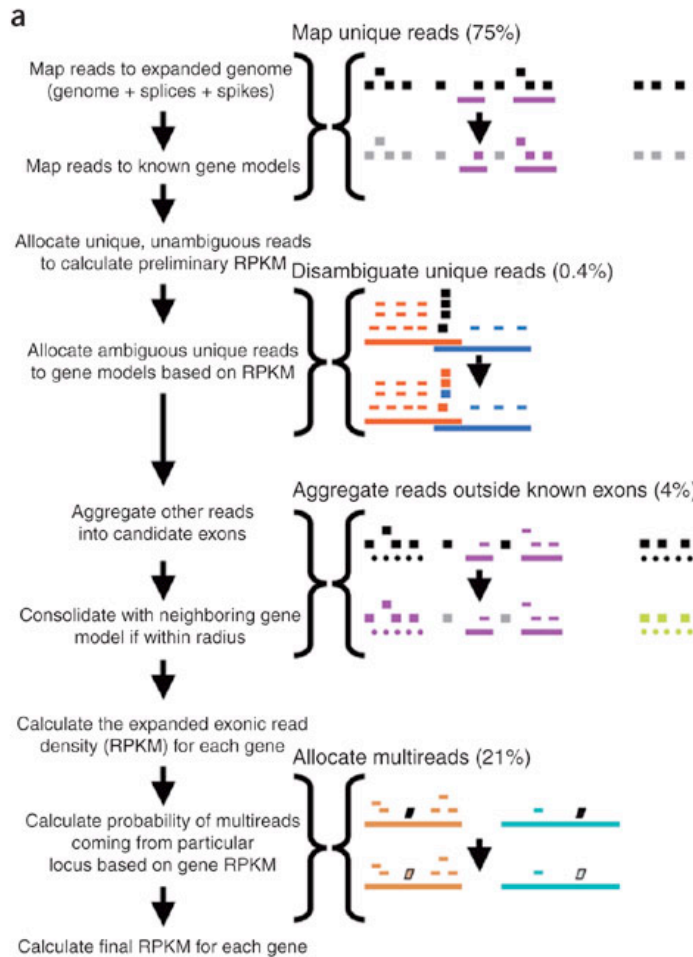
- Transcript assembly
- Novel structures
- Isoform multireads

# 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



# 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

# Normalisation

- Counting
  - estimate of *relative* counts for each gene

Does this accurately represent the original population?

## Library size

Sequencing depth varies between samples

## Gene Properties

GC content, length, sequence

## Library composition

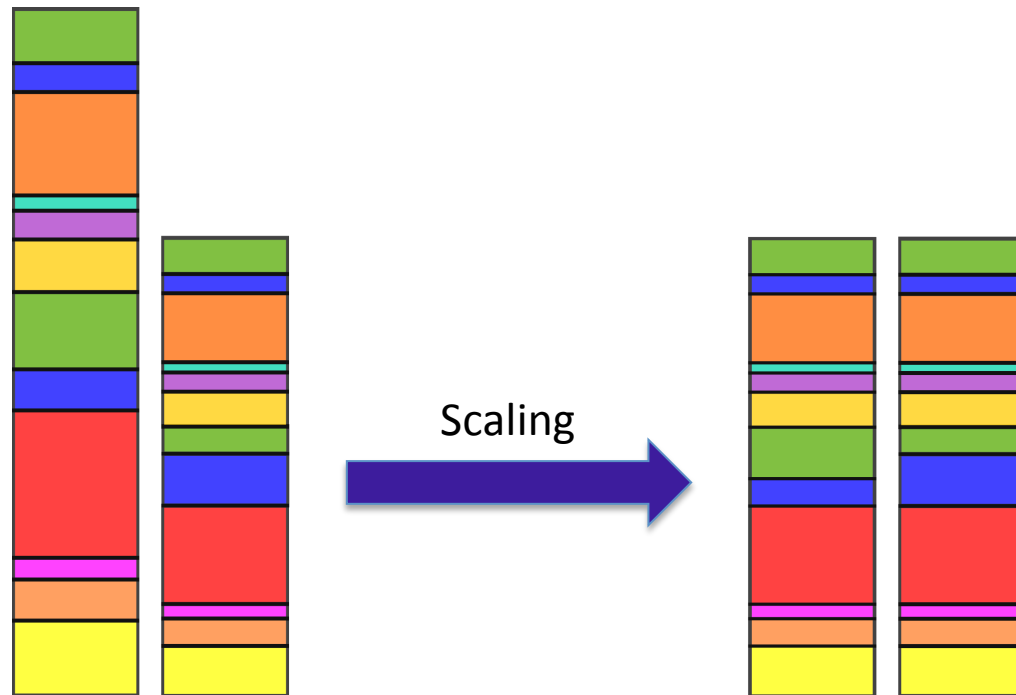
Highly expressed genes overrepresented at cost of lowly expressed genes



# Normalisation - 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

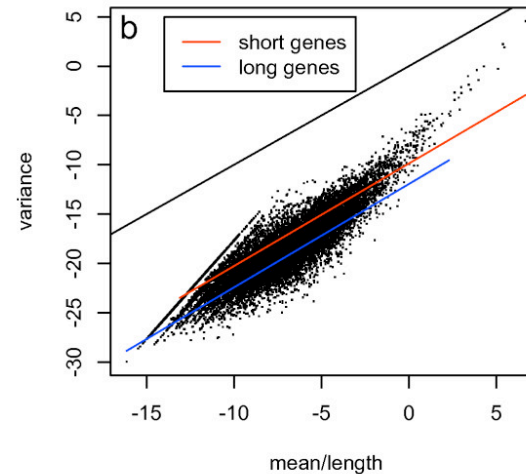
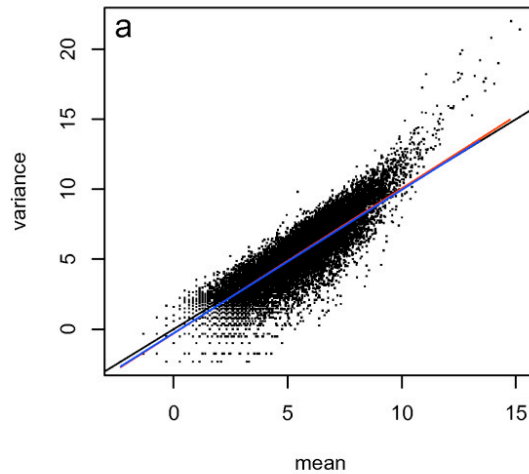


# Normalisation - RPKM

## RPKM

- Reads per kilobase per million =

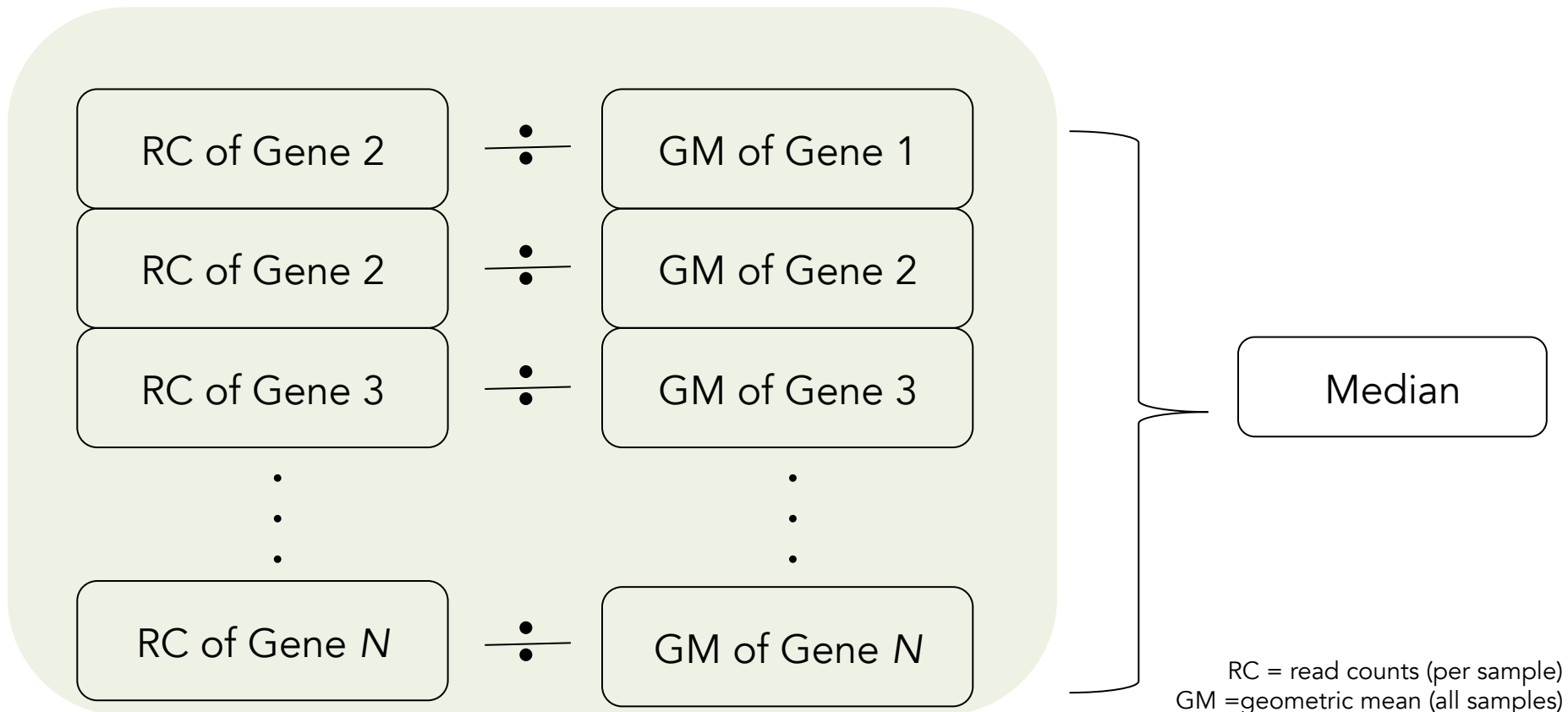
$$\frac{\text{reads for gene A}}{\text{length of gene A} \times \text{Total number of reads}}$$



# Normalisation – Geometric Scaling

## Geometric scaling factor

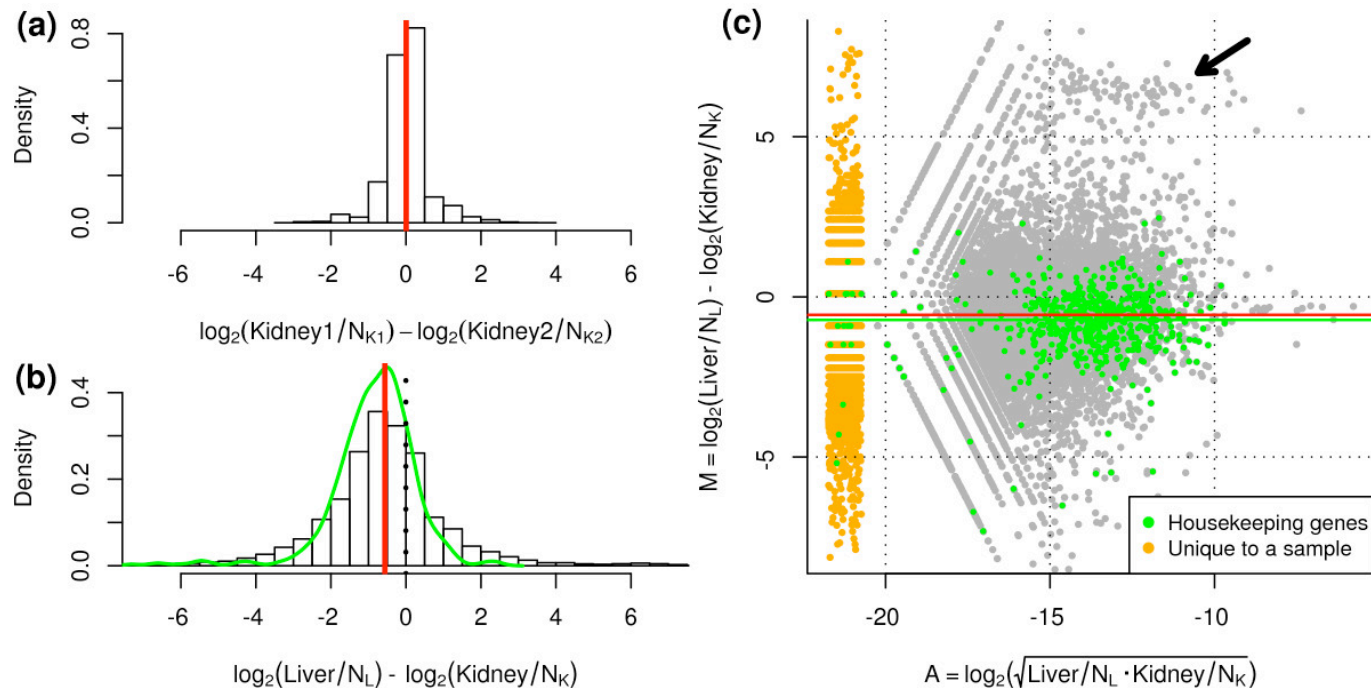
- Assumes that most genes are not differentially expressed



# Normalisation – Trimmed Mean of M

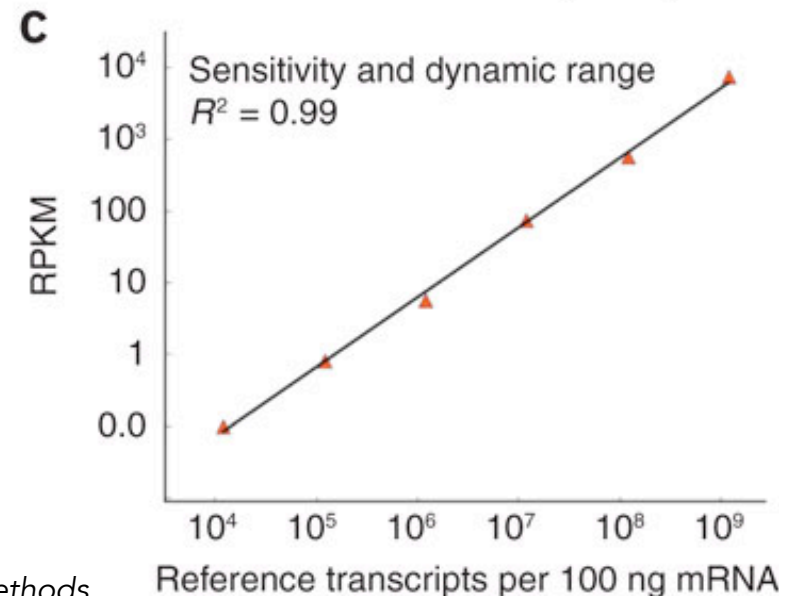
## Trimmed mean of $M$

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



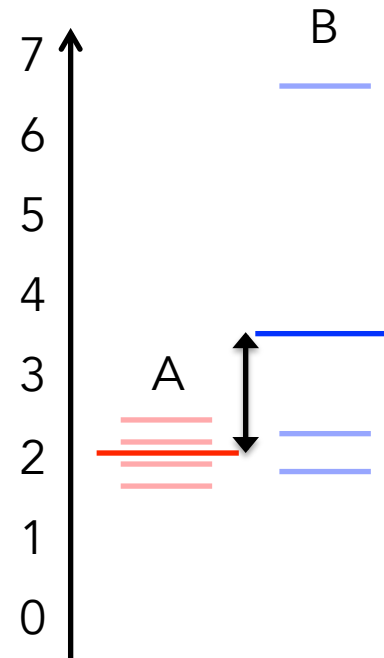
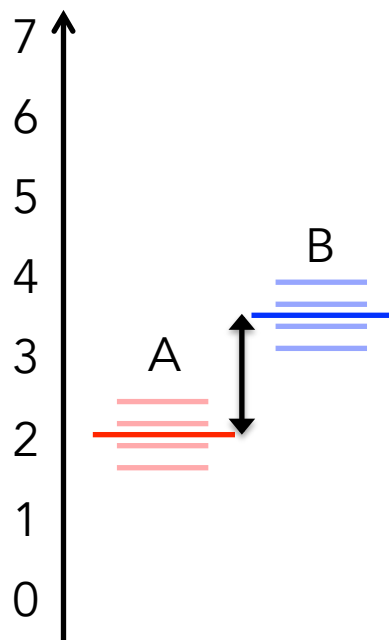
# Differential Expression

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



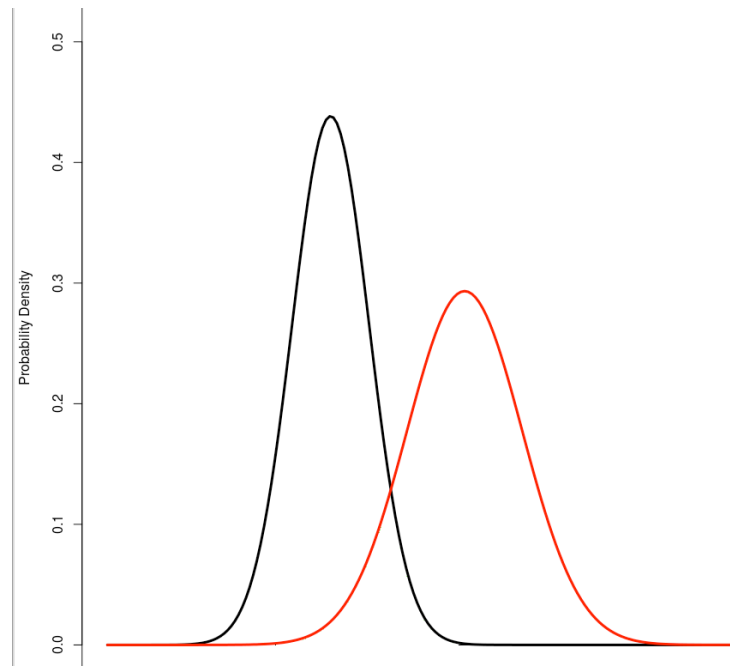
# Differential Expression

- Simple difference in means



- Replication introduces variance

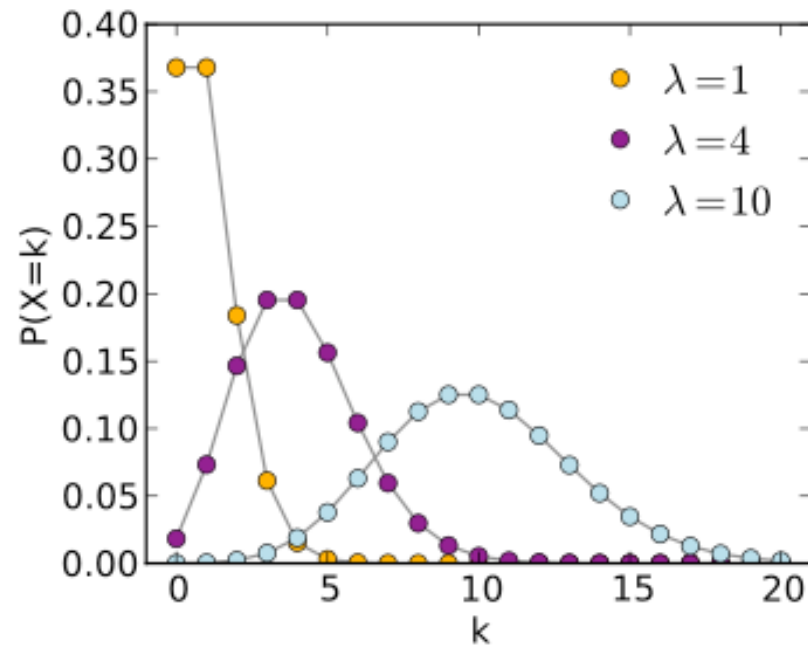
# Differential Expression - Modelling



Normal distribution → t-test

# Differential Expression- Modelling

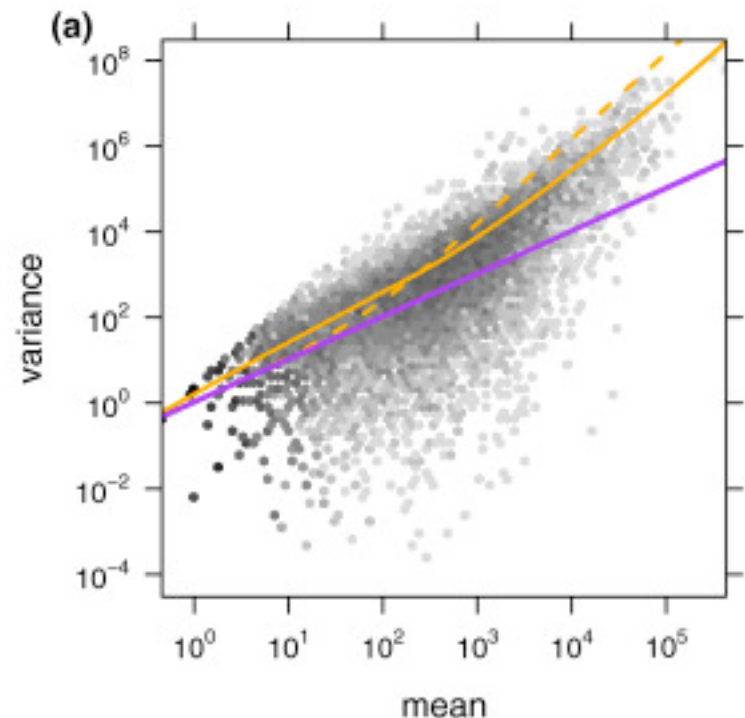
- Use the Poisson distribution for count data
- Just one parameter required – the mean





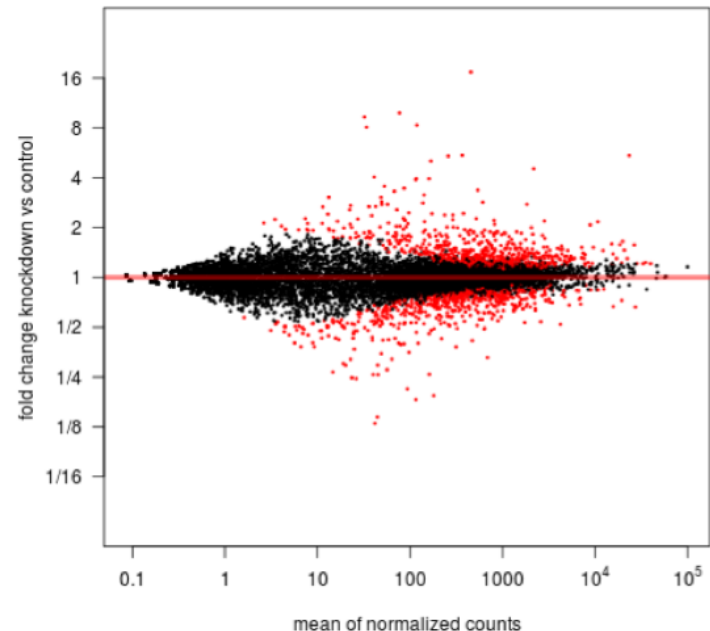
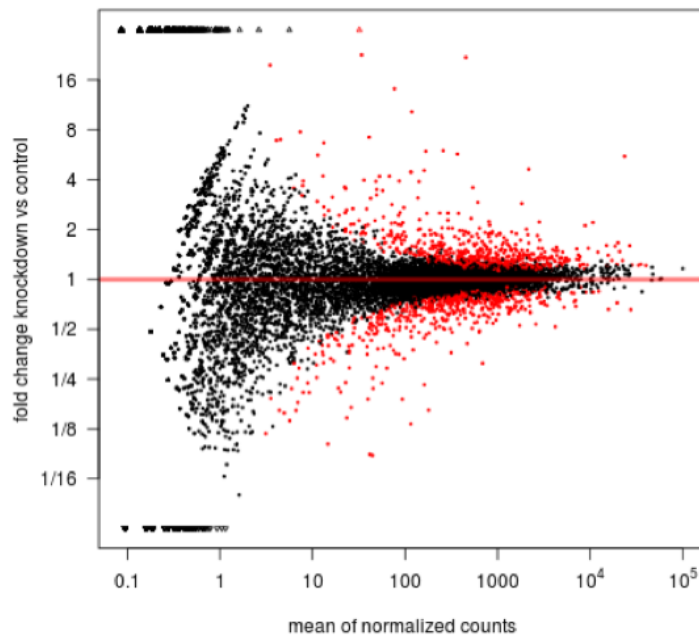
# Differential Expression- Modelling

- Biology is never that simple
- The negative binomial distribution represents an overdispersed Poisson distribution
- It has two parameters:  
    mean and (over)dispersion



# Differential Expression- Modelling

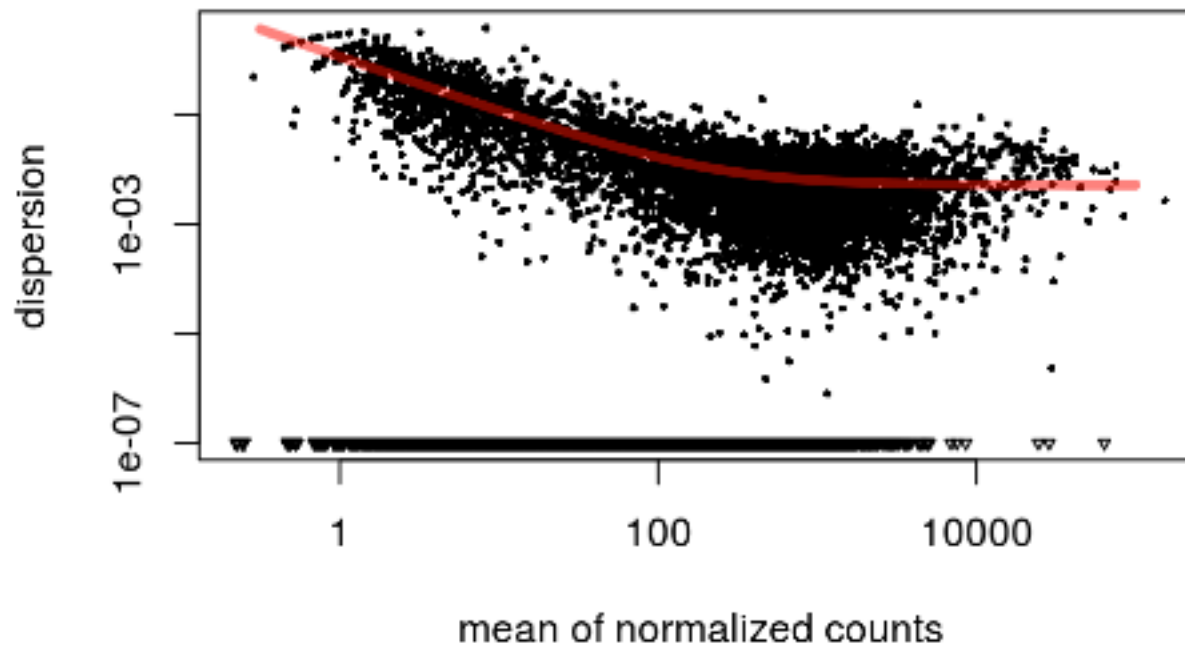
- Estimating the dispersion parameter can be difficult with a small number of samples
- edgeR: models the variance as the sum of technical and biological variance
- ‘Share’ information from all genes to obtain global estimate - shrinkage



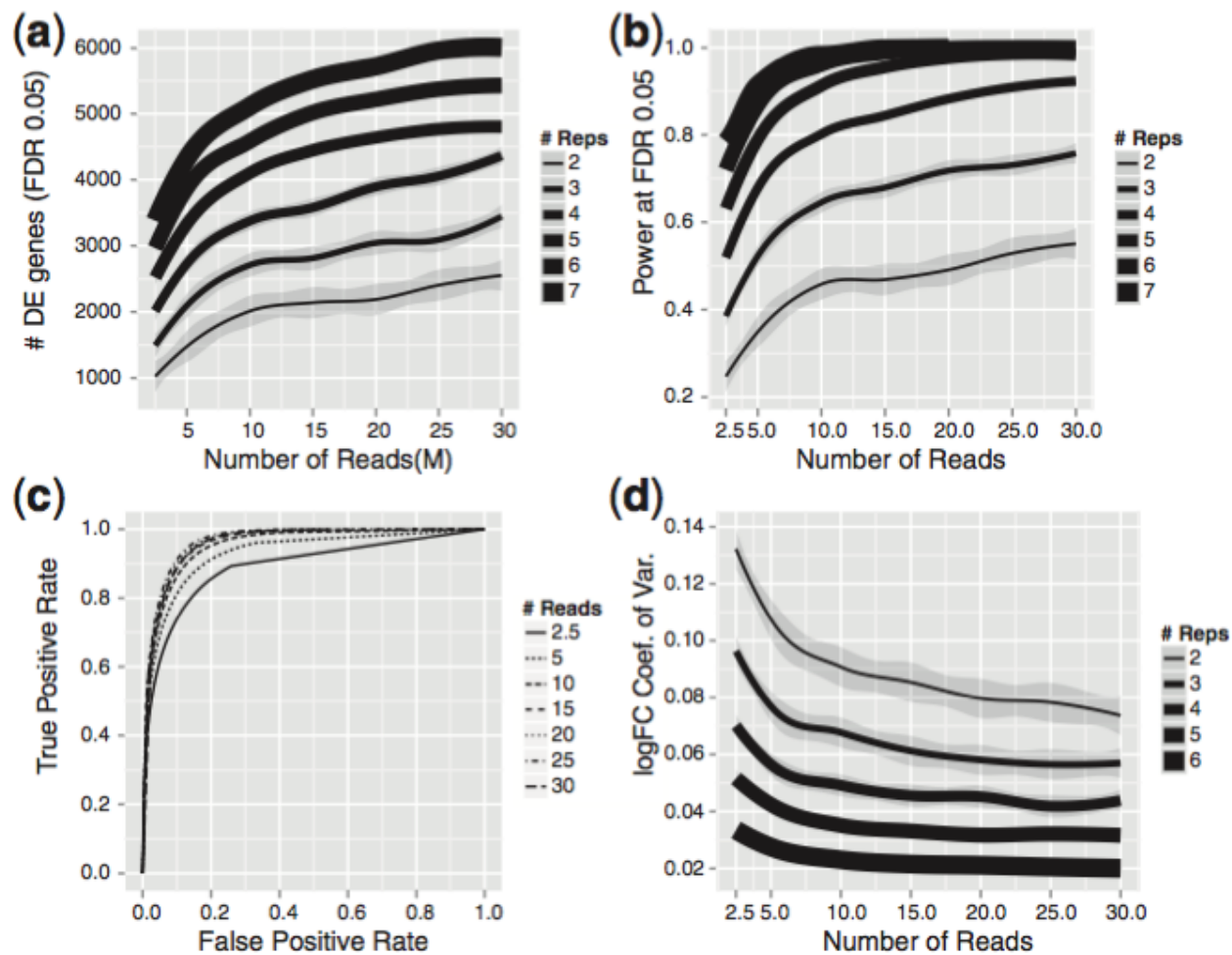
# Modelling – in fashion

- DESeq uses a similar formulation of the variance term

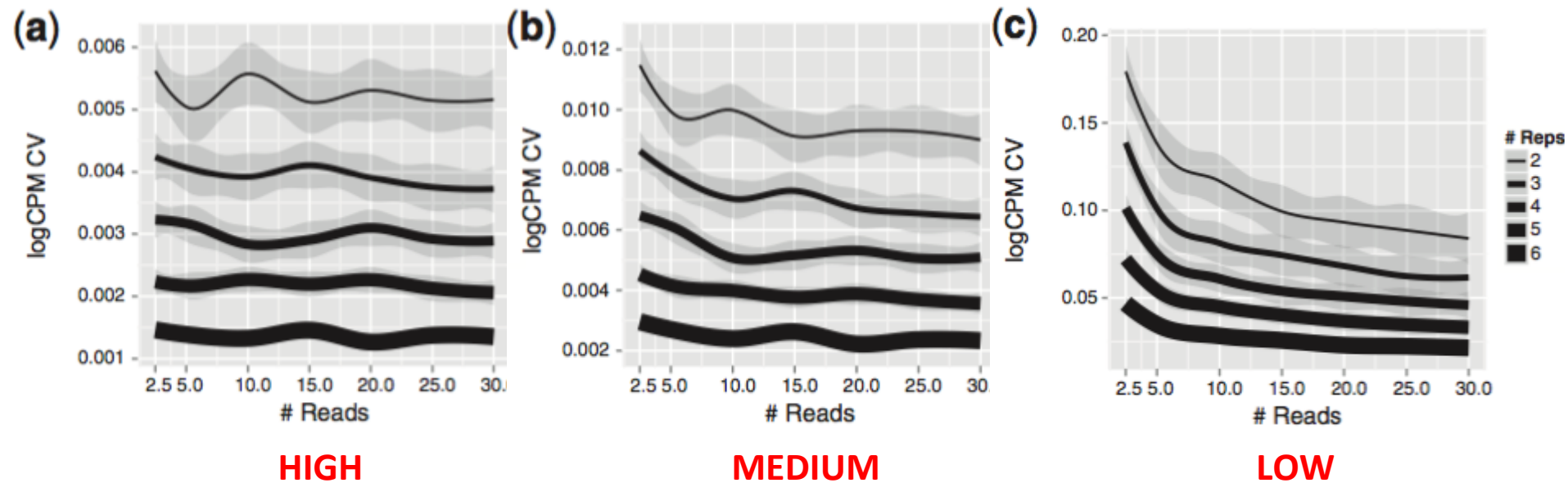
$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v(\mu_{ij})}_{\text{raw variance}} .$$



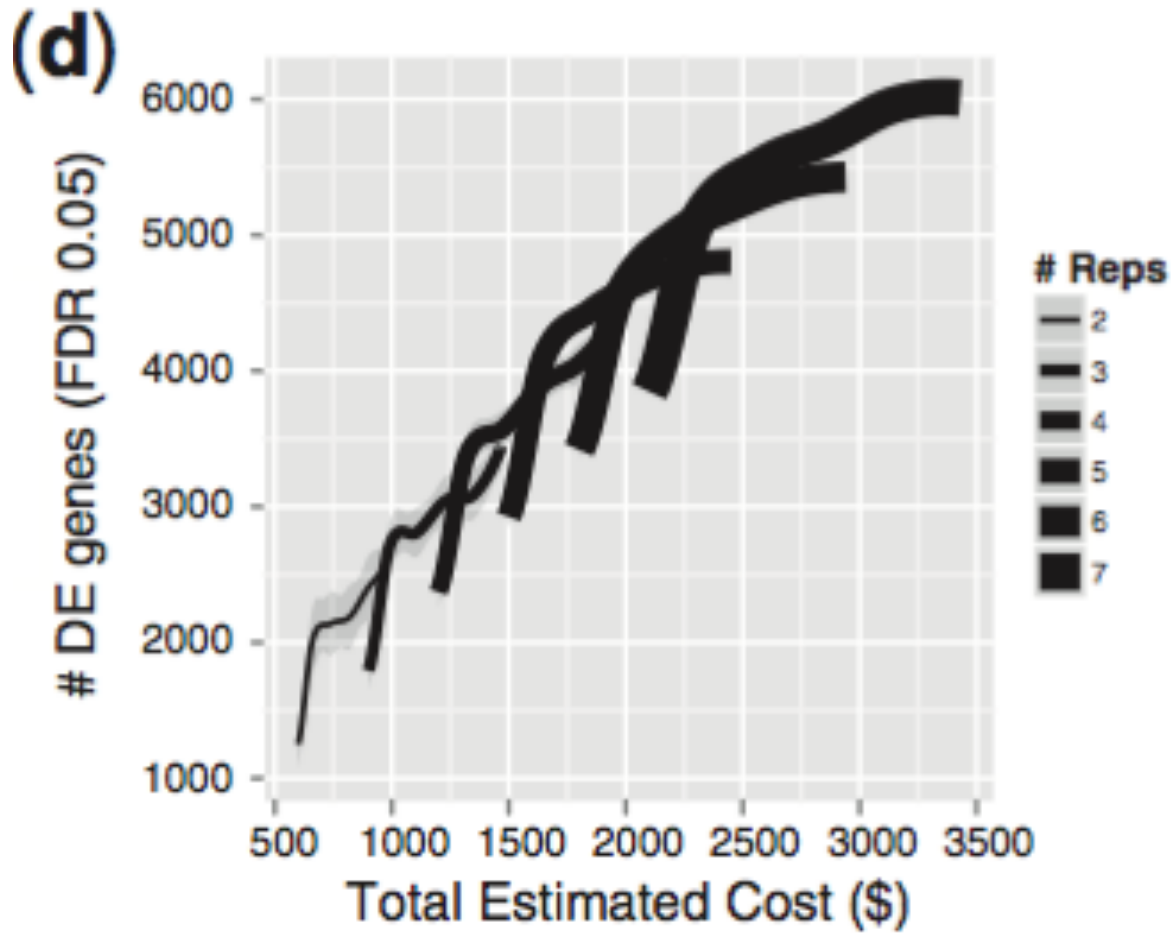
# Replicates v Sequencing Depth



# Replicates v Sequencing Depth

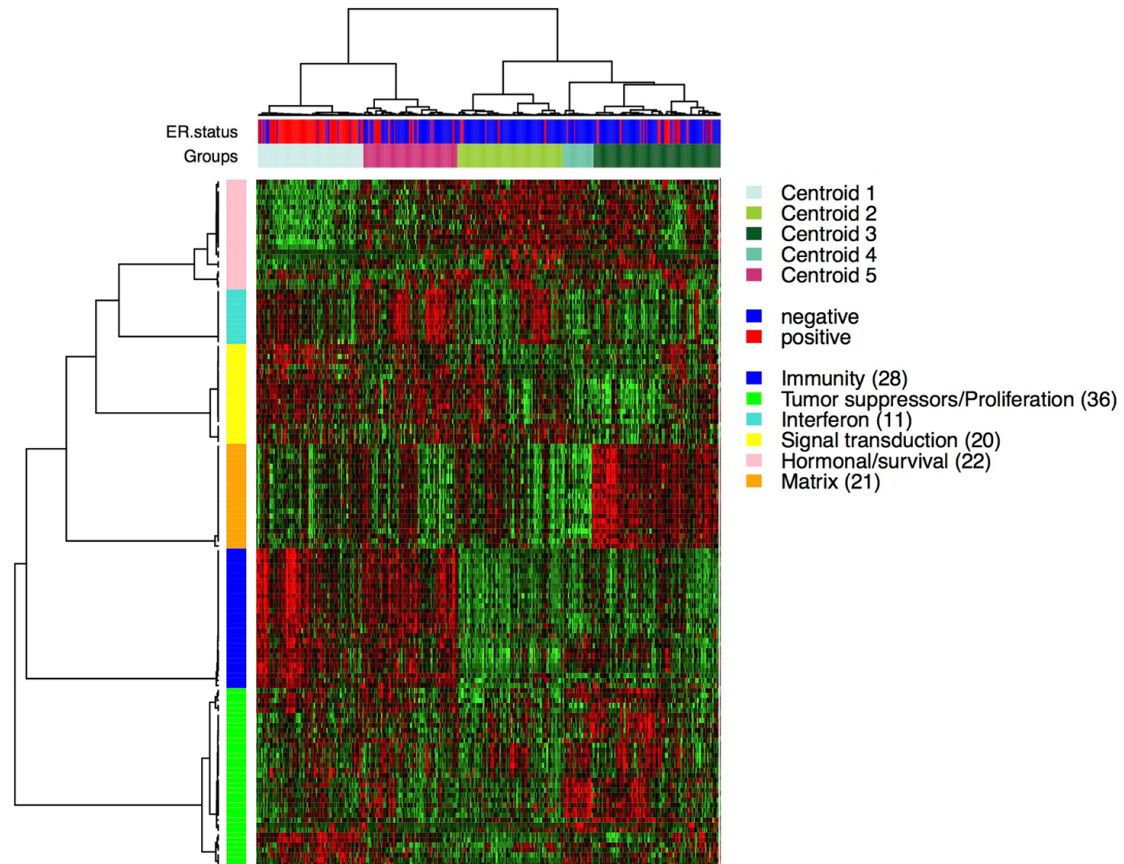


# Replicates v Sequencing Depth



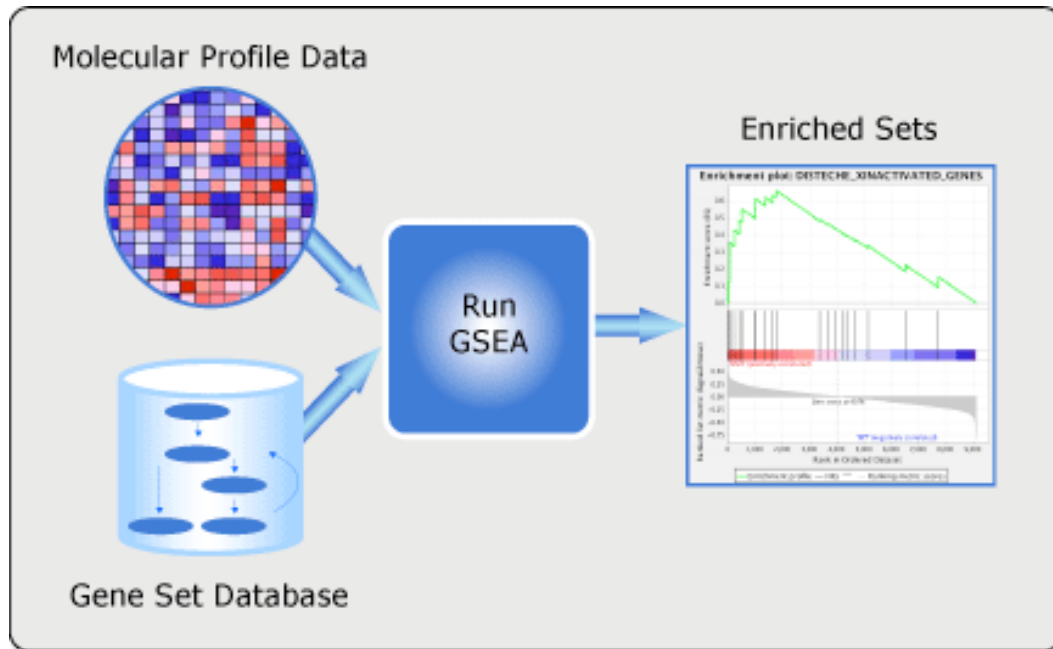
# Towards Biological Meaning

- Clustering



# Towards Biological Meaning

- Gene Set Enrichment Analysis



- ▶ **H** (hallmark gene sets, 50 gene sets) <sup>?</sup>
- ▶ **C1** (positional gene sets, 326 gene sets) <sup>?</sup>
  - ▶ 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) <sup>?</sup>
  - ▶ **CGP** (chemical and genetic perturbations, 3395 gene sets) <sup>?</sup>
  - ▶ **CP** (Canonical pathways, 1330 gene sets) <sup>?</sup>
  - ▶ **CP:BIOCARTA** (BioCarta gene sets, 217 gene sets) <sup>?</sup>
  - ▶ **CP:KEGG** (KEGG gene sets, 186 gene sets) <sup>?</sup>
  - ▶ **CP:REACTOME** (Reactome gene sets, 674 gene sets) <sup>?</sup>
- ▶ **C3** (motif gene sets, 836 gene sets) <sup>?</sup>
  - ▶ **MIR** (microRNA targets, 221 gene sets) <sup>?</sup>
  - ▶ **TFT** (transcription factor targets, 615 gene sets) <sup>?</sup>
- ▶ **C4** (computational gene sets, 858 gene sets) <sup>?</sup>
  - ▶ **CGN** (cancer gene neighborhoods, 427 gene sets) <sup>?</sup>
  - ▶ **CM** (cancer modules, 431 gene sets) <sup>?</sup>
- ▶ **C5** (GO gene sets, 1454 gene sets) <sup>?</sup>
  - ▶ **BP** (GO biological process, 825 gene sets) <sup>?</sup>
  - ▶ **CC** (GO cellular component, 233 gene sets) <sup>?</sup>
  - ▶ **MF** (GO molecular function, 396 gene sets) <sup>?</sup>
- ▶ **C6** (oncogenic signatures, 189 gene sets) <sup>?</sup>
- ▶ **C7** (immunologic signatures, 1910 gene sets) <sup>?</sup>



# Towards Biological Meaning

- Network analysis

