



INSTITUTO
GULBENKIAN
DE CIÊNCIA

The Gulbenkian Training Programme in Bioinformatics

NDARC16 - NGS Data Analysis, RNAseq, ChIPseq

Analysis of RNA-seq data

30 March 2016

Slides by Bernard Pereira (CRUK-CI, U. Cambridge) et al



Instituto
de Medicina
Molecular

Nuno Barbosa Morais



`nmorais@medicina.ulisboa.pt`



<http://imm.medicina.ulisboa.pt/group/compbio/>

The many faces of RNA-seq

<http://www.illumina.com/techniques/sequencing/rna-sequencing.html>

illumina®

[MyIllumina](#) [Quick Order](#)

[Contact Us](#) [English](#)

[AREAS OF INTEREST](#) ▾

[TECHNIQUES](#) ▾

[SYSTEMS](#) ▾

[PRODUCTS & SERVICES](#) ▾

[INFORMATICS](#) ▾

[SCIENTIFIC CONTENT](#) ▾

[COMPANY](#) ▾

[SUPPORT](#) ▾

[SEARCH](#)

[Techniques](#) / [Sequencing](#) / [RNA Sequencing](#)

Key RNA-Seq Methods

mRNA Sequencing

Accurately measure gene and transcript abundance and detect both known and novel features in the coding transcriptome.

[Learn More](#)

Total RNA Sequencing

Accurately measure gene and transcript abundance and detect both known and novel features in coding and multiple forms of noncoding RNA.

[Learn More](#)

Targeted RNA Sequencing

Measure the expression of transcripts or pathways of interest. Perform differential expression analysis, measurement of allele-specific expression, and detection of gene fusions.

[Learn More](#)

Small RNA Sequencing

Isolate and sequence small RNA species, such as microRNA, to understand the role of noncoding RNA in gene silencing and posttranscriptional regulation of gene expression.

[Learn More](#)

Ribosome Profiling

Deeply sequence ribosome-protected mRNA fragments to gain a complete view of all the ribosomes active in a cell at a specific time point and predict protein abundance.

[Learn More](#)

Ultra-Low-Input and Single-Cell RNA-Seq

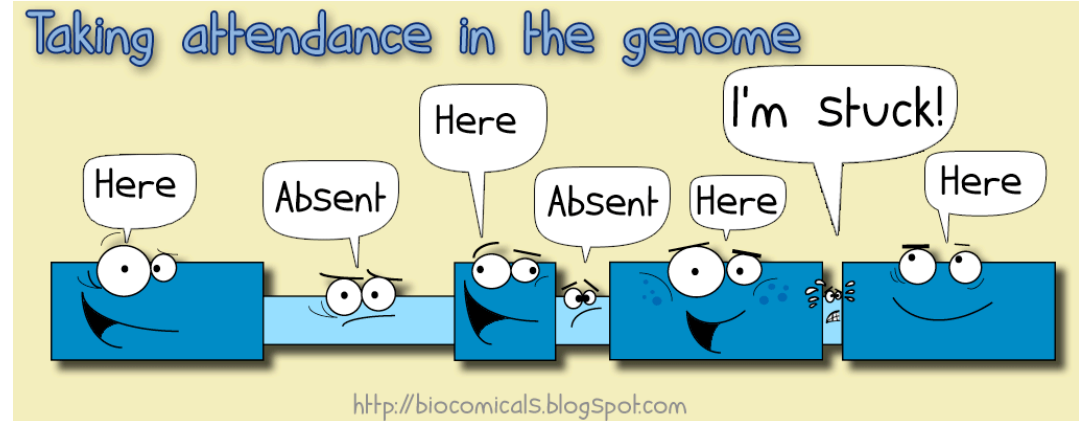
Use deep RNA-Seq to examine the signals and behavior of a cell in the context of its surrounding environment. This method is advantageous for biologists studying cell function in time-dependent processes such as differentiation, proliferation, and tumorigenesis.

[Learn More](#)

Applications

Discovery

- Find new transcripts
- Find transcript boundaries
- Find splice junctions
- Find gene fusions
- Find mutations (SNPs)
- Quantify allele specific expression



Comparison

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

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

Applications

Journal of Pathology

J Pathol 2015; **235**: 571–580

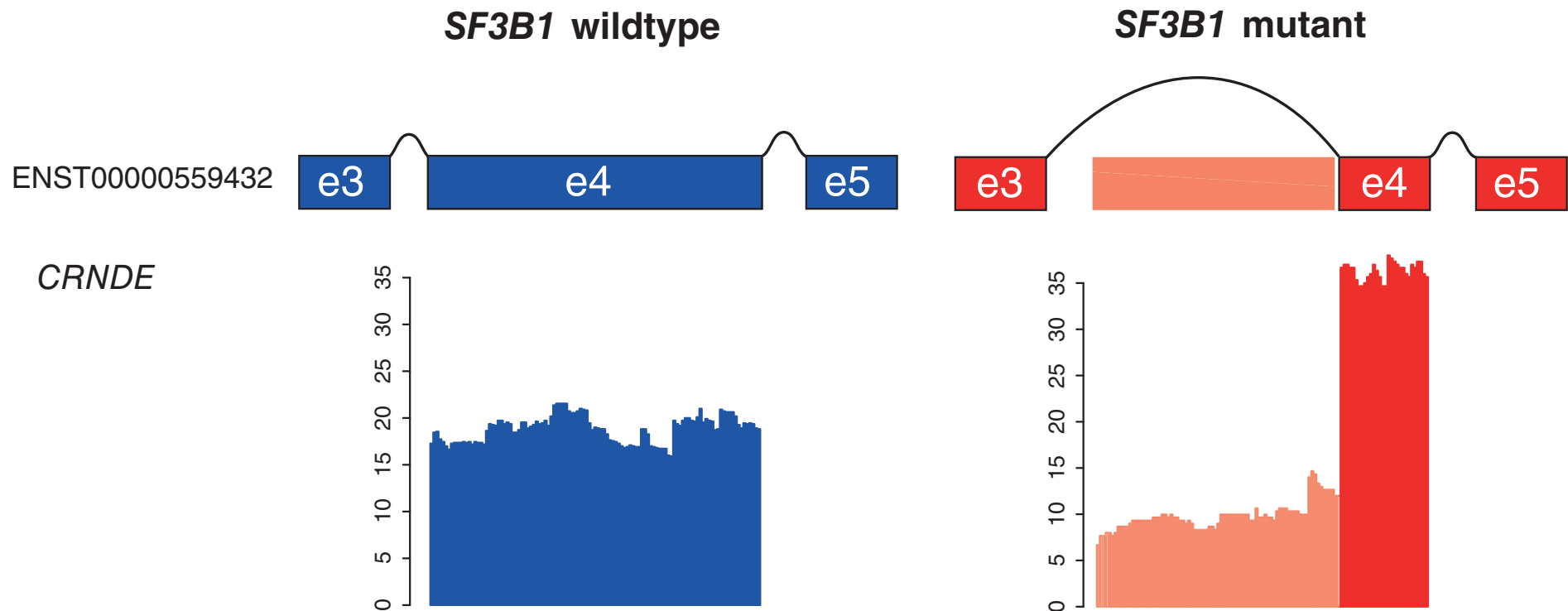
Published online 22 December 2014 in Wiley Online Library

(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

LETTERS

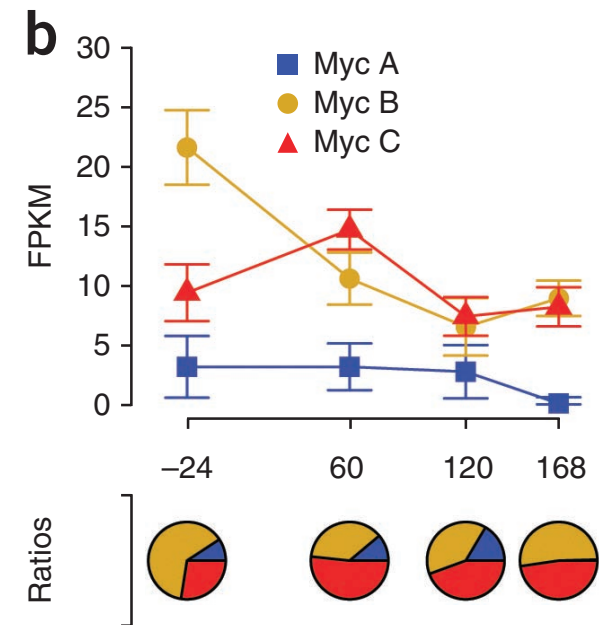
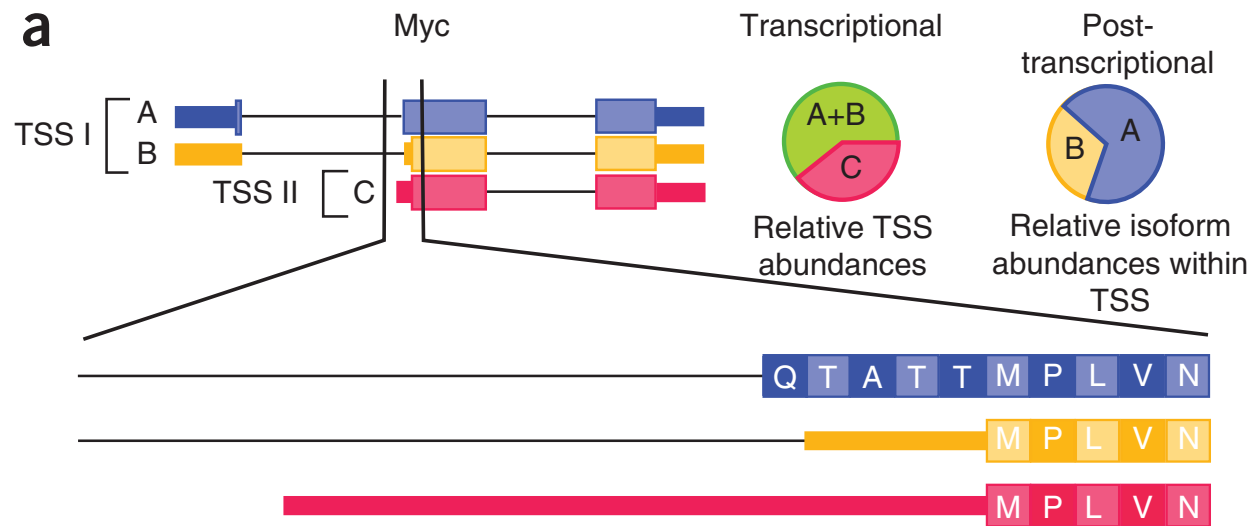
nature
biotechnology

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}

NATURE BIOTECHNOLOGY VOLUME 28 NUMBER 5 MAY 2010

511



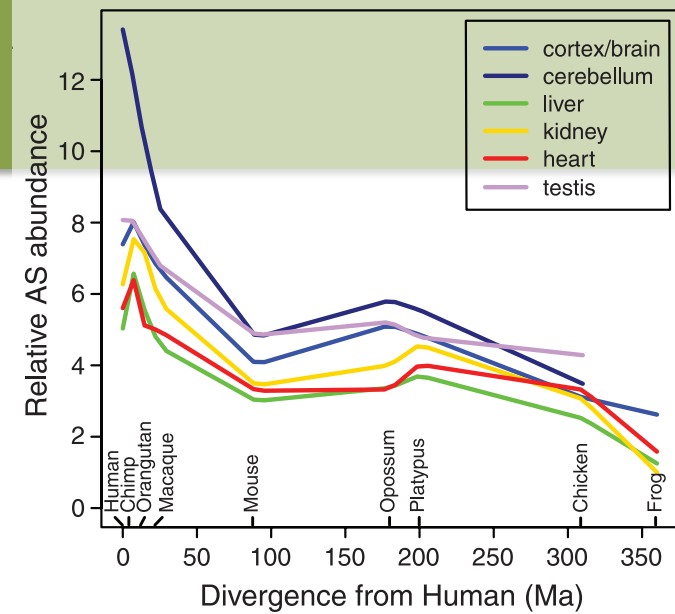
Applications

The Evolutionary Landscape of Alternative Splicing in Vertebrate Species

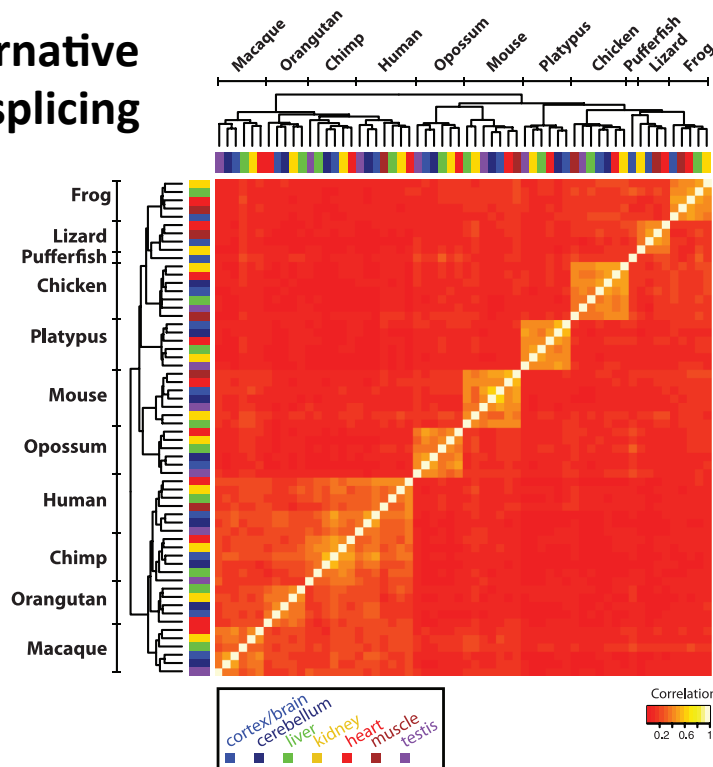
Nuno L. Barbosa-Morais,^{1,2} Manuel Irimia,^{1*} Qun Pan,^{1*} Hui Y. Xiong,^{3*} Serge Gueroussov,^{1,4*} Leo J. Lee,³ Valentina Slobodeniuc,¹ Claudia Kutter,⁵ Stephen Watt,⁵ Recep Çolak,^{1,6} TaeHyung Kim,^{1,7} Christine M. Misquitta-Ali,¹ Michael D. Wilson,^{4,5,7} Philip M. Kim,^{1,4,6} Duncan T. Odom,^{5,8} Brendan J. Frey,^{1,3} Benjamin J. Blencowe^{1,4†}

www.sciencemag.org SCIENCE VOL 338 21 DECEMBER 2012

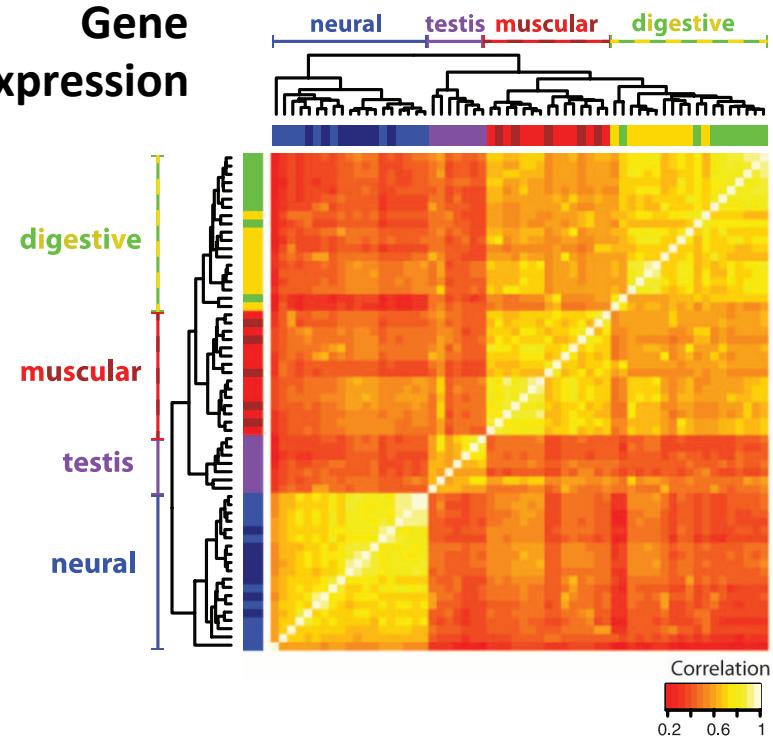
1587



Alternative splicing

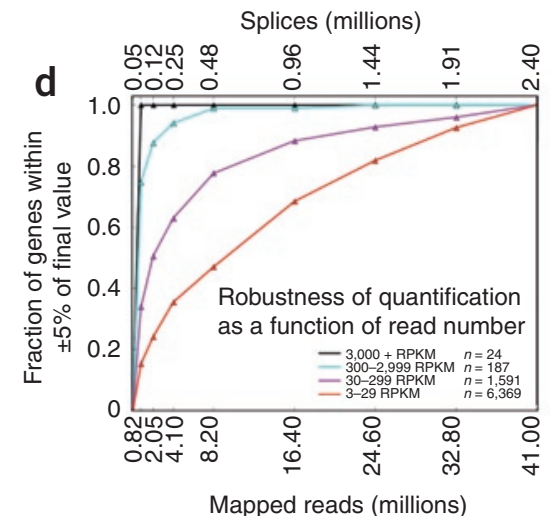
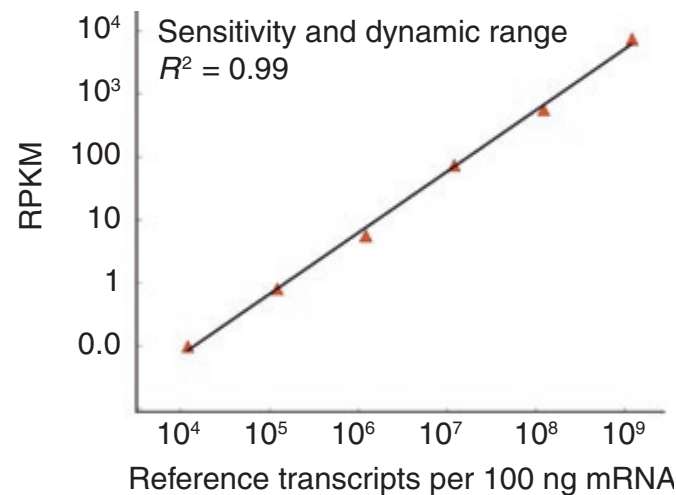
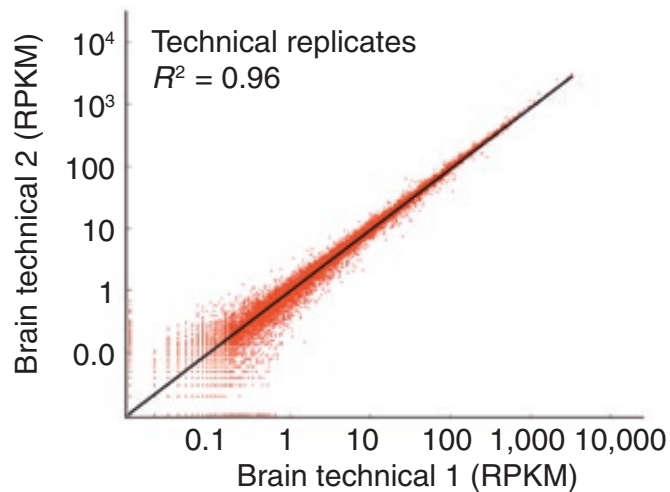


Gene expression



Differential Expression

- Comparing feature abundance under different conditions
- Assumed linearity, reproducibility and sensitivity

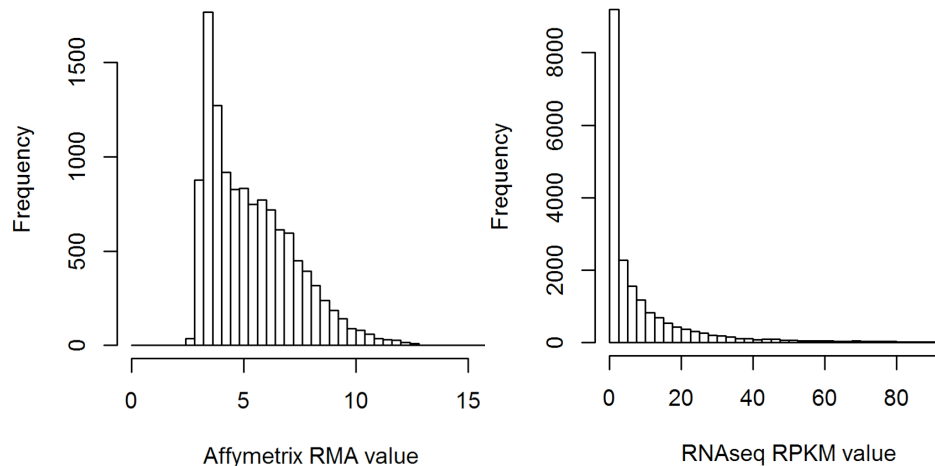


Mortazavi, A. et al (2008) *Nature Methods*

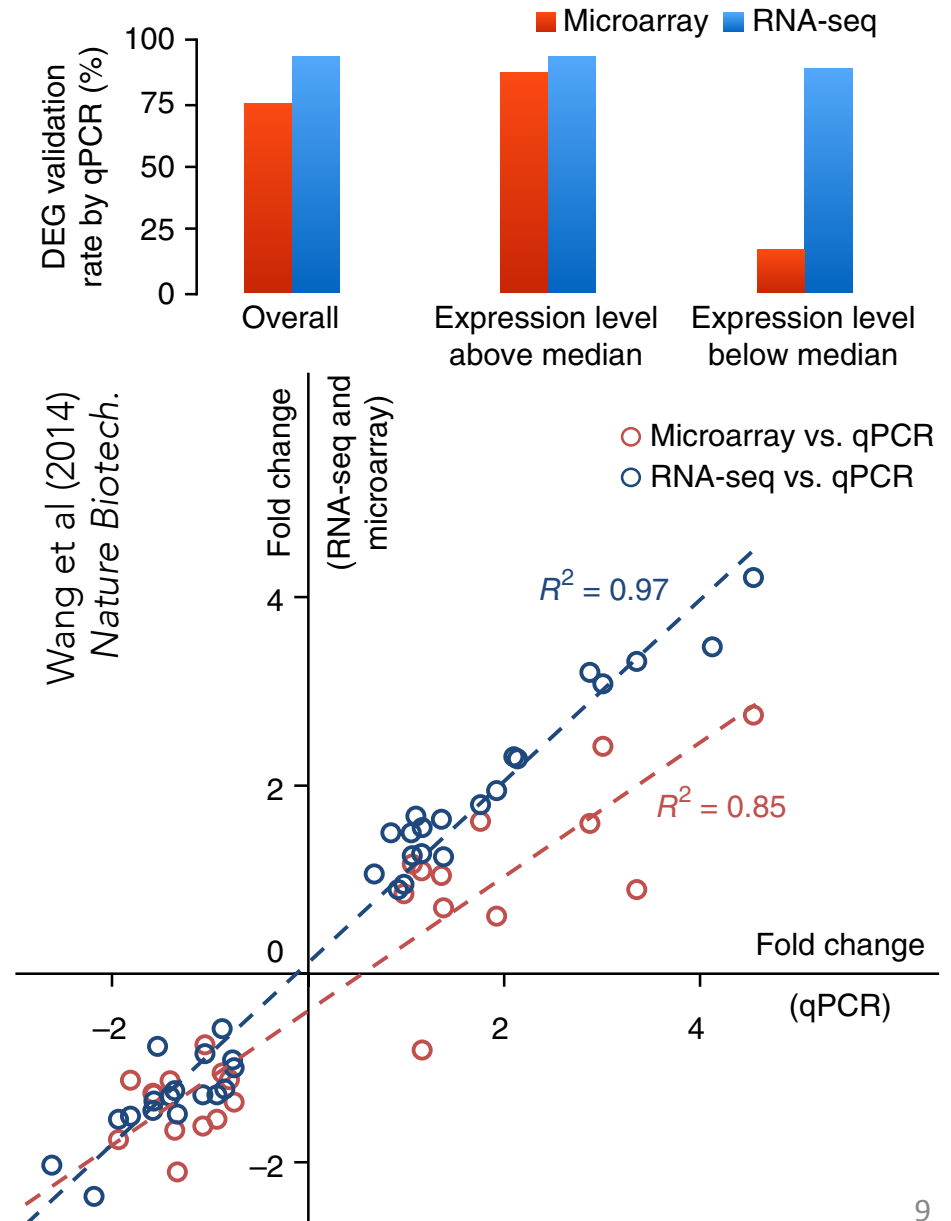
- When *feature=gene*, well-established pre- and post-analysis strategies exist (including those originally conceived for microarrays)

Better than microarrays?

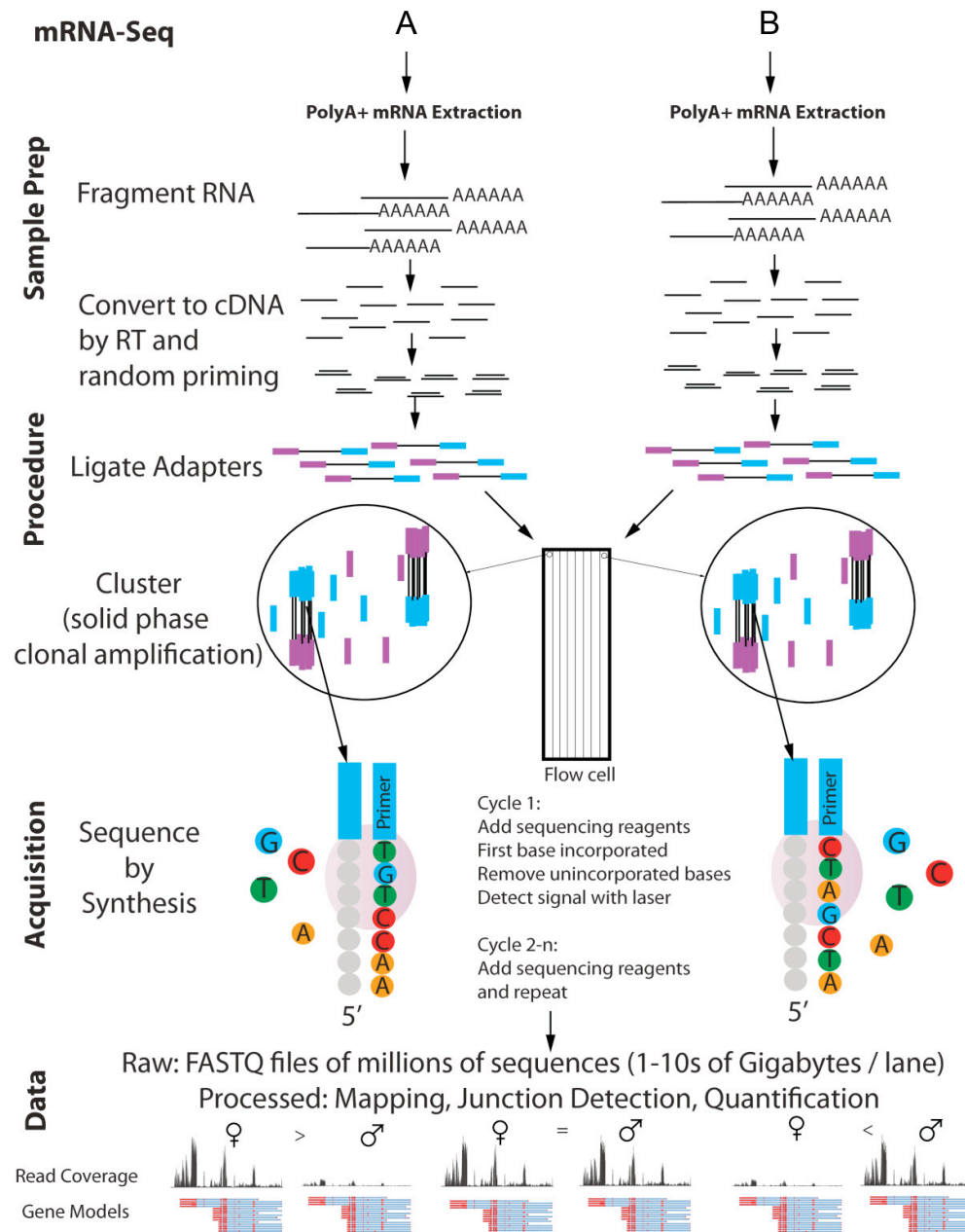
- Better dynamic range
- Not biased by probe design (specificity)
- More sensitive, no saturation
- Better validation
- More expensive (prices dropping)



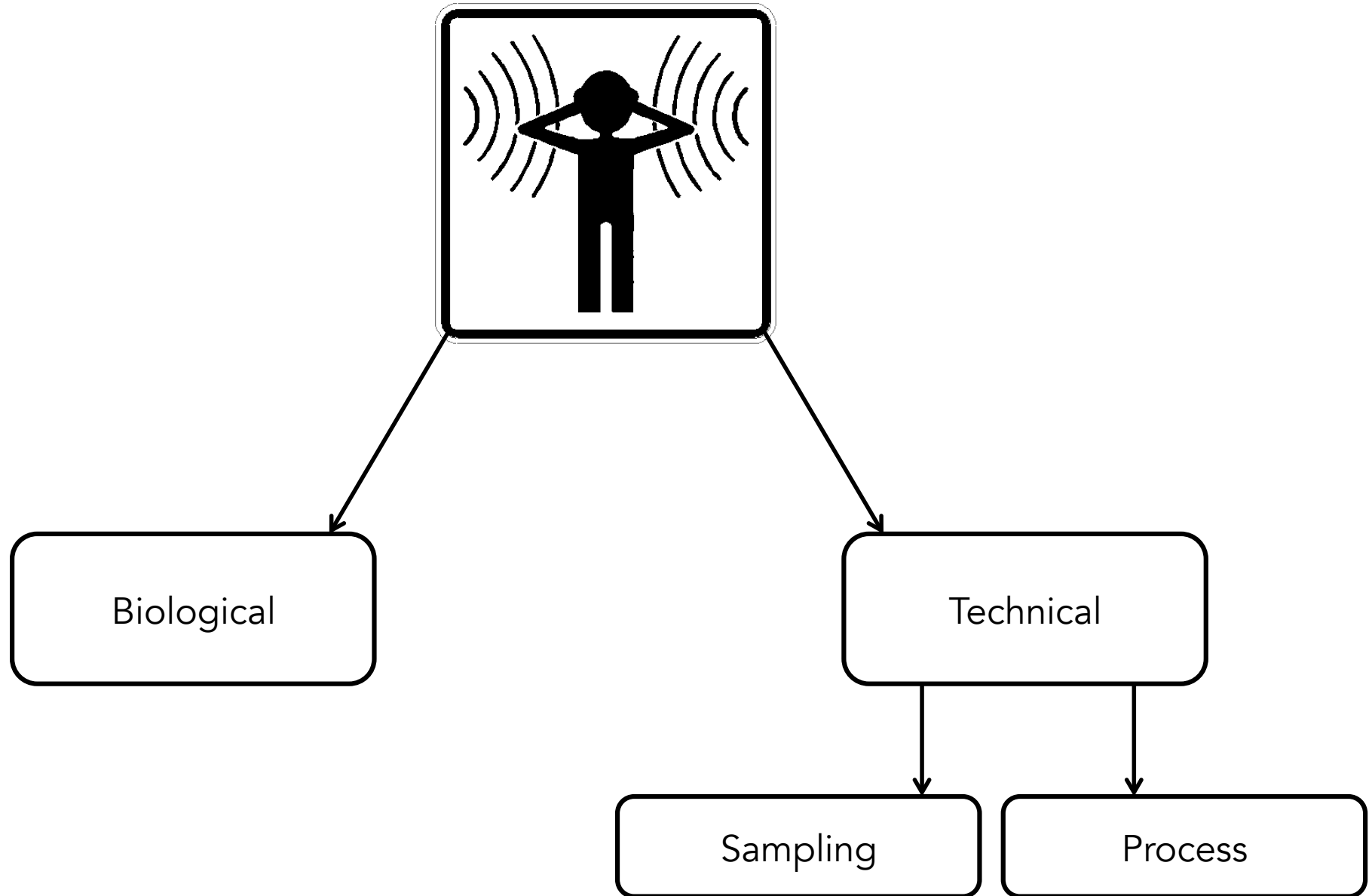
Guo et al. (2013) *Plos One*



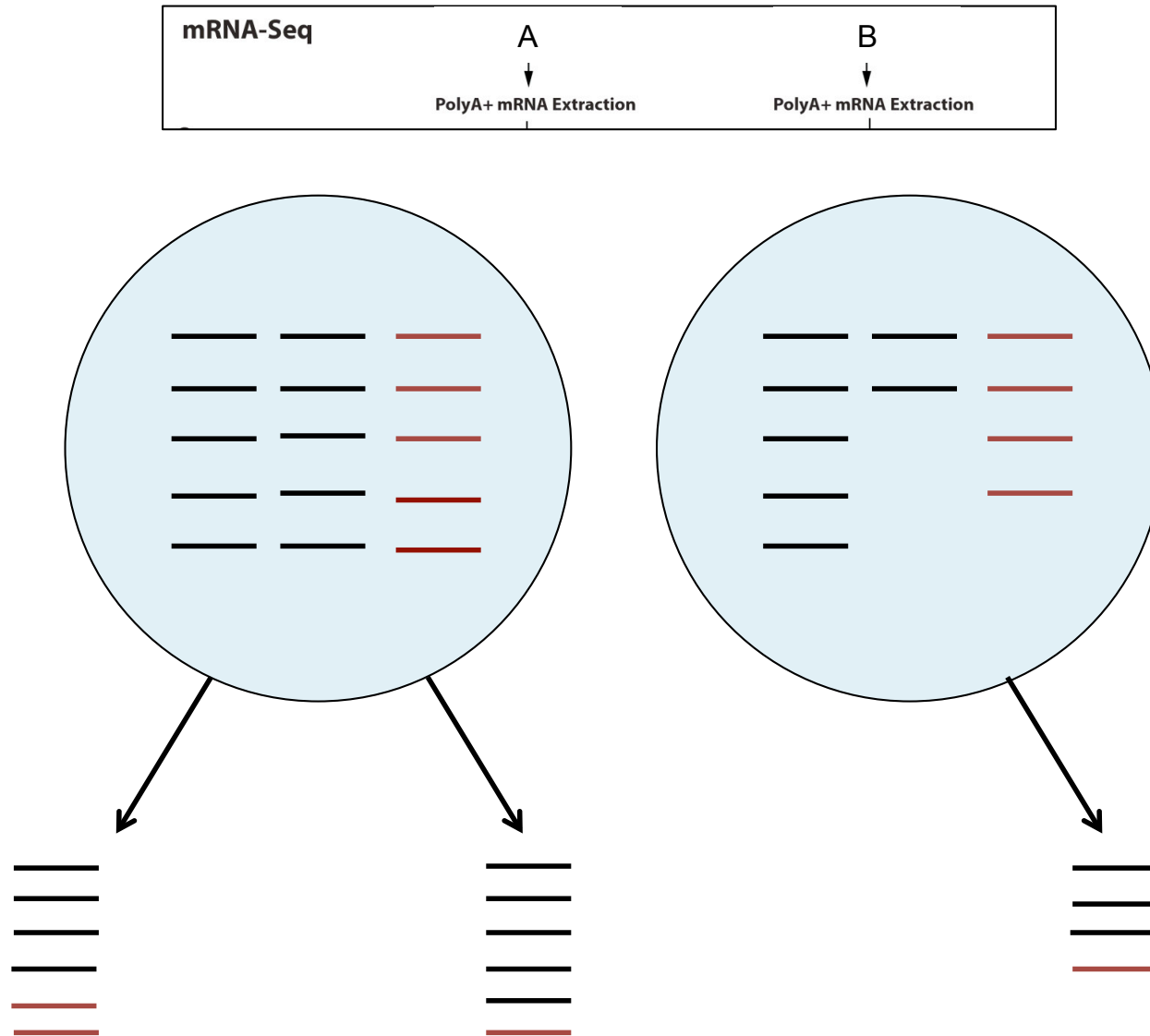
Library Prep i



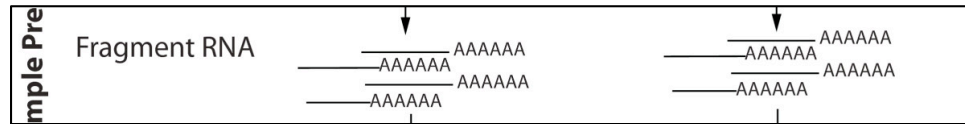
Library Prep ii



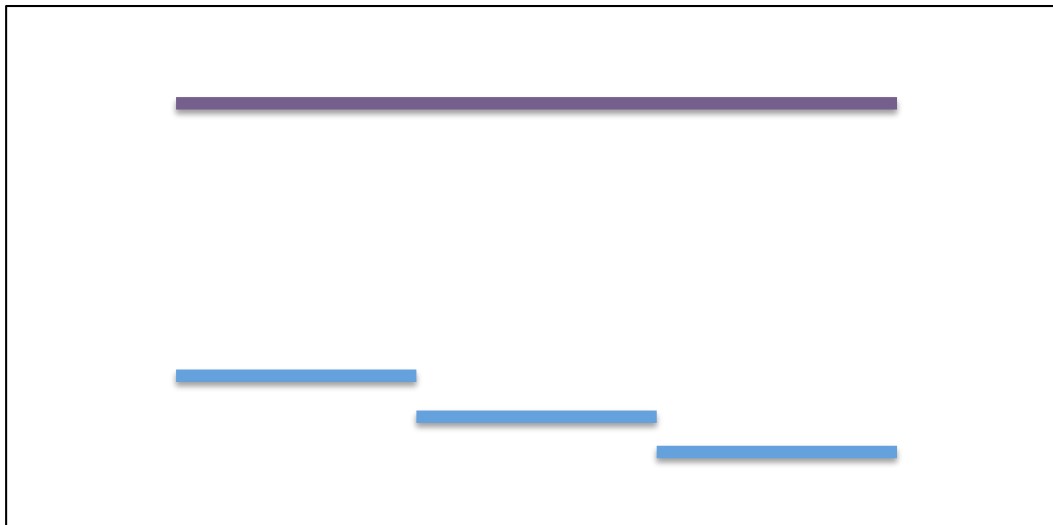
Library Prep iii



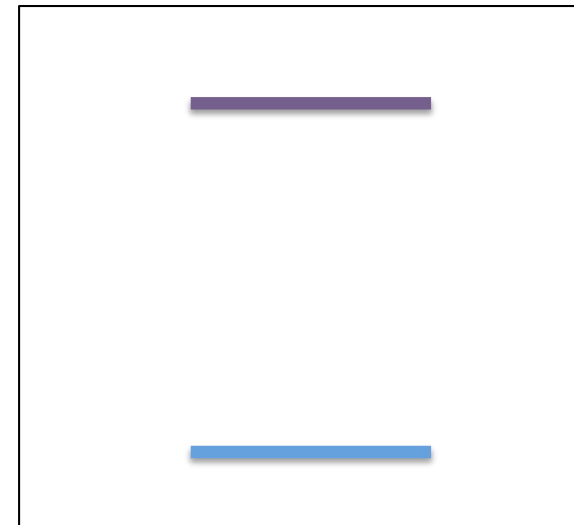
Library Prep iii



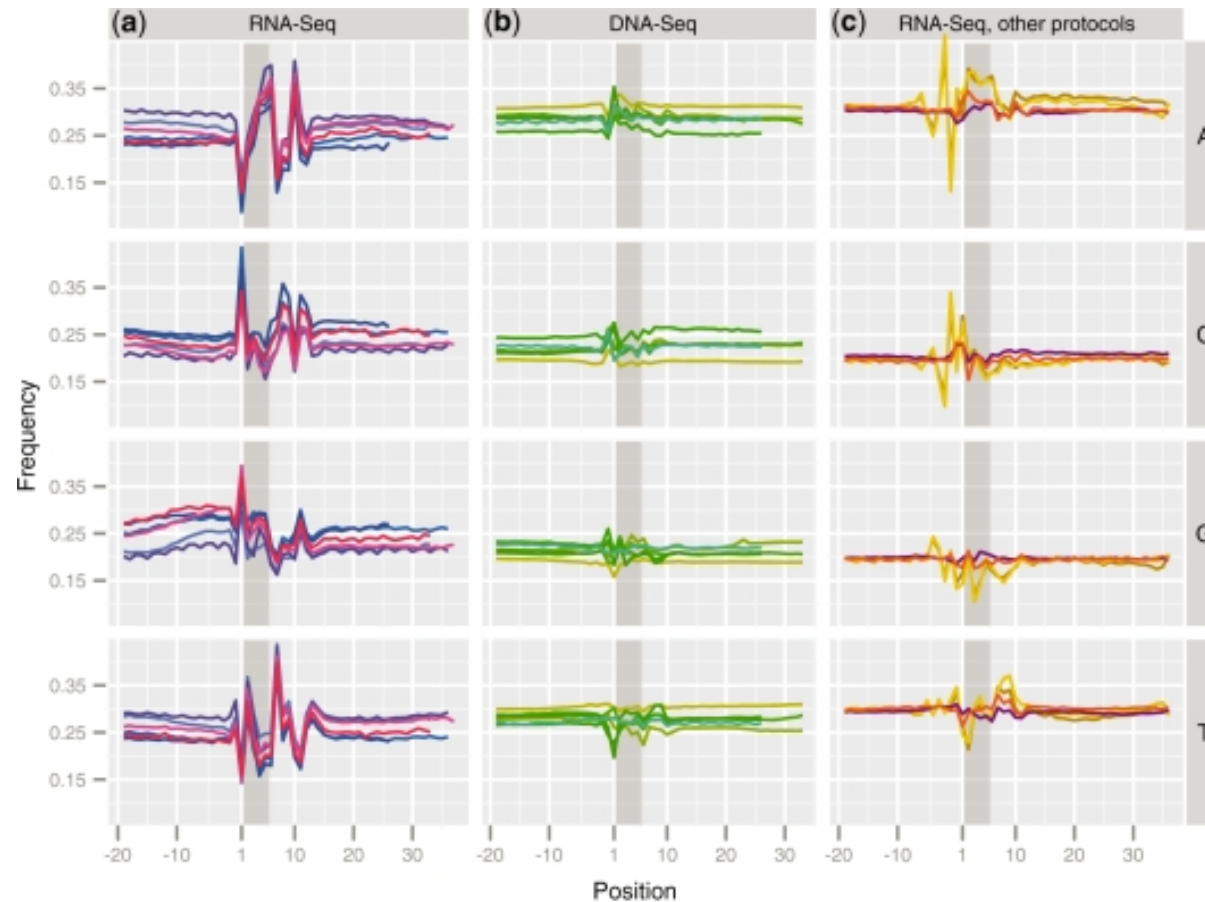
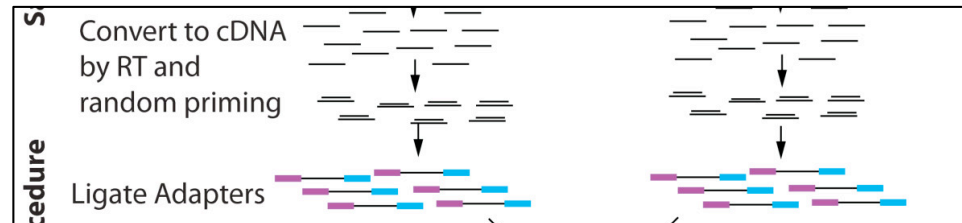
A



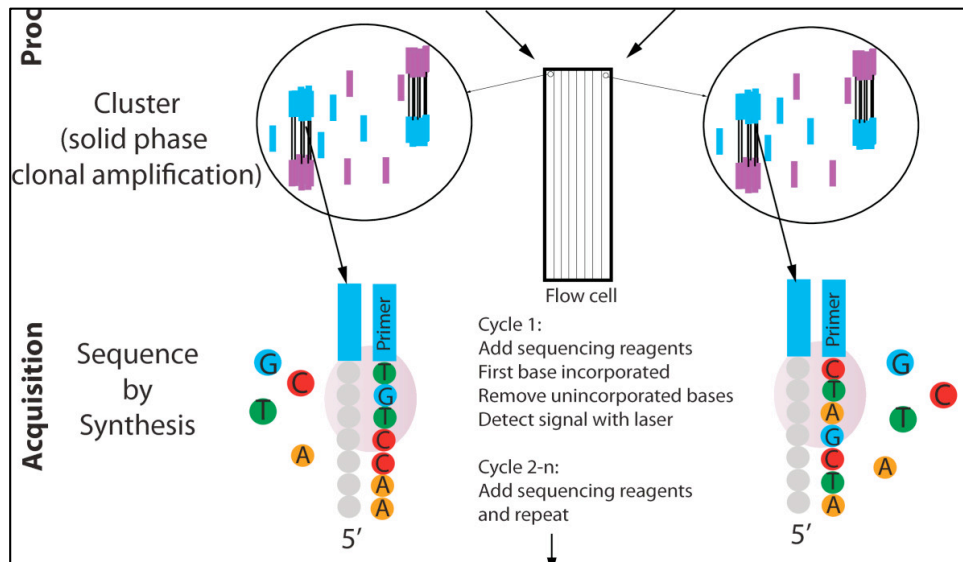
B



Library Prep iv

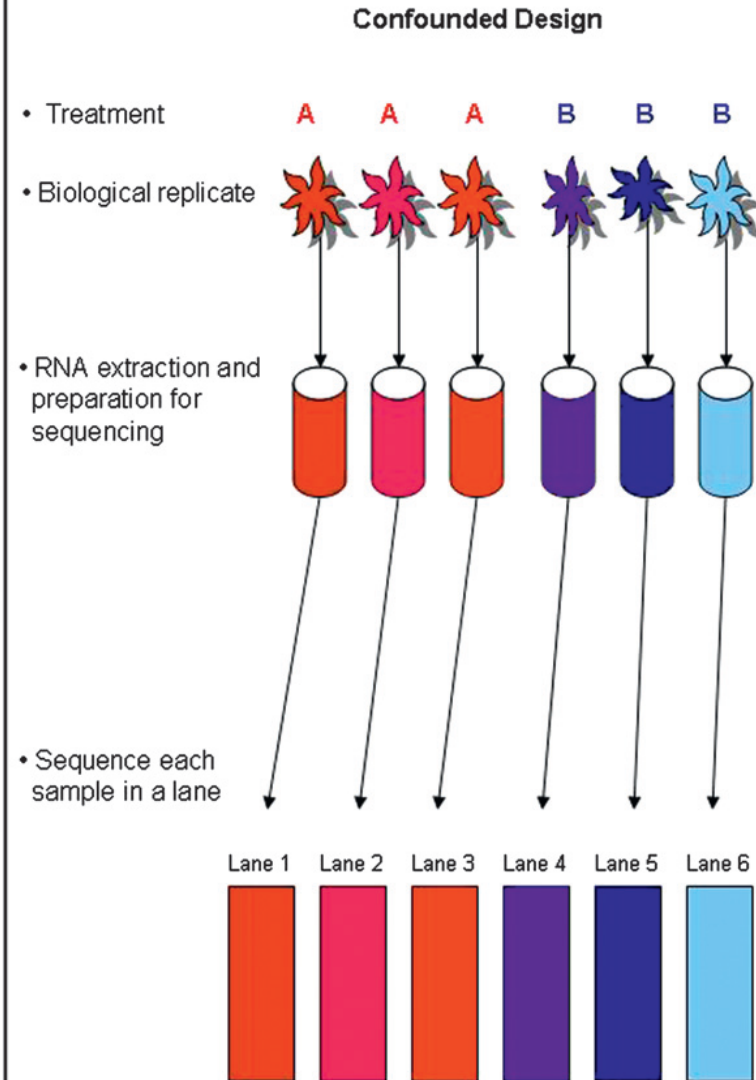
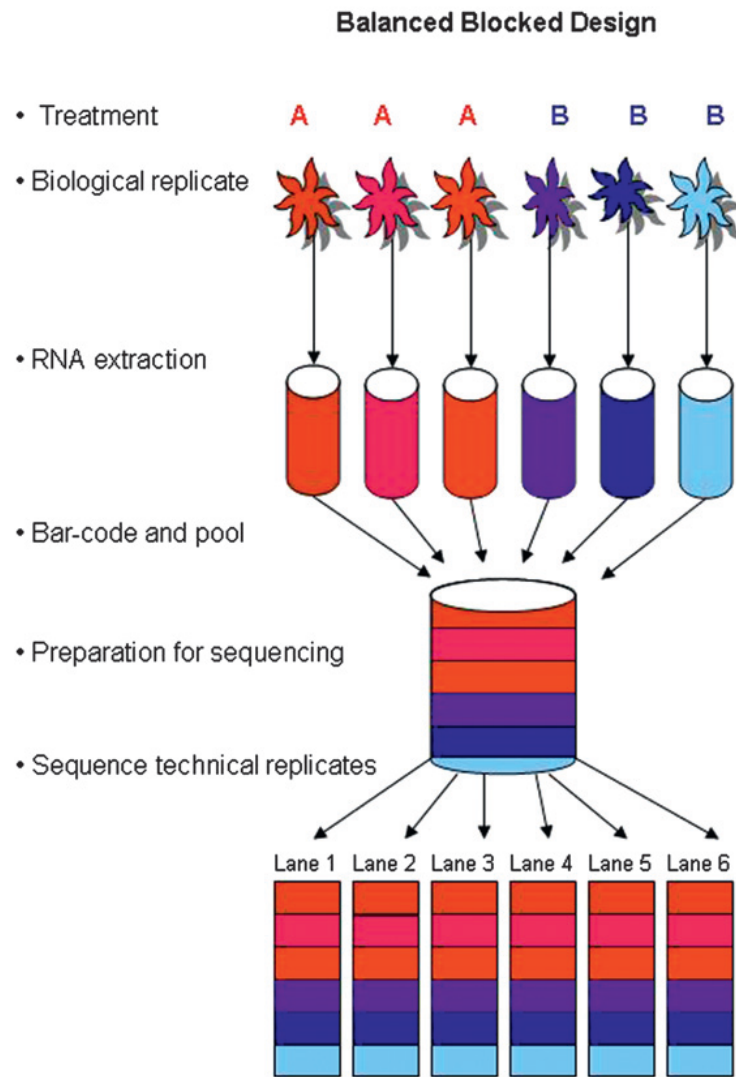


Library Prep v

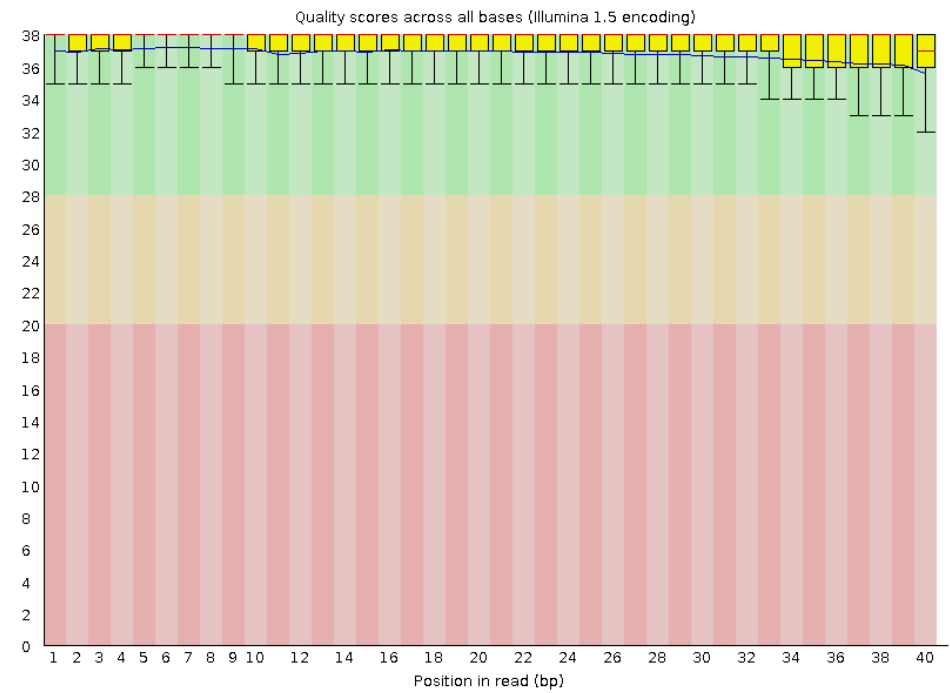
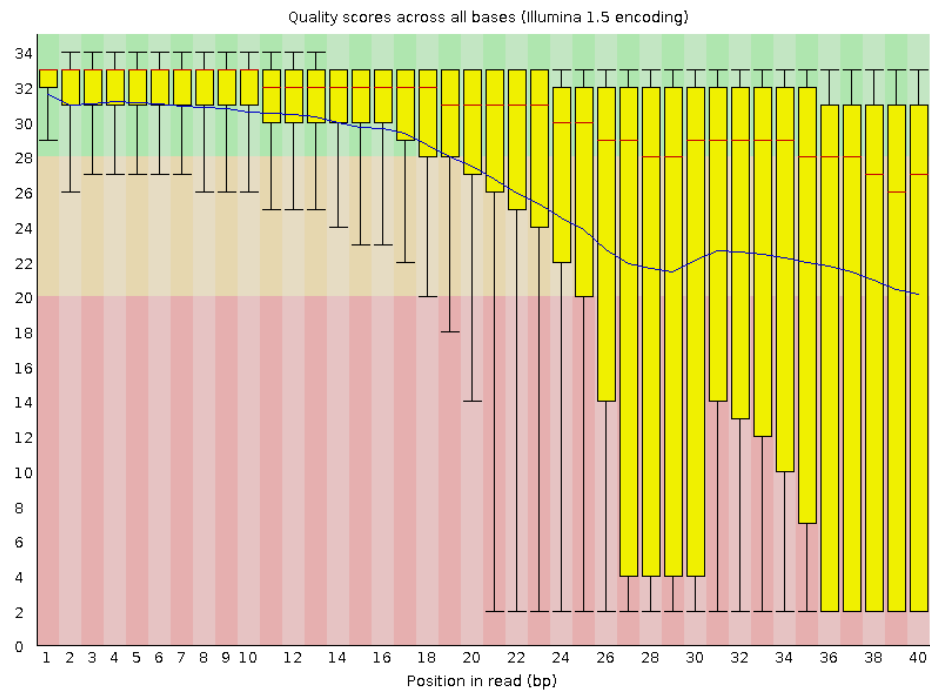
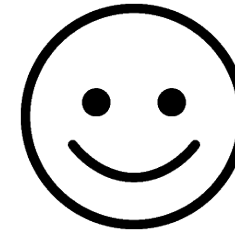


- Duplicates (optical & PCR)
- Sequence errors
- Indels
- Repetitive/problematic sequence

Hot off the sequencer...



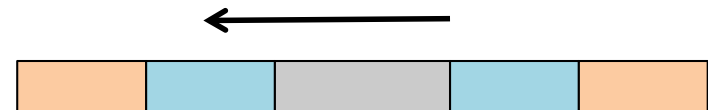
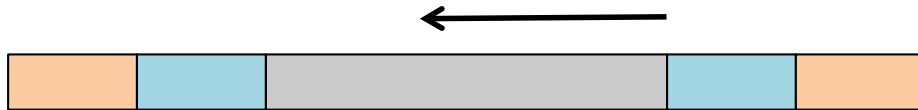
FASTQC



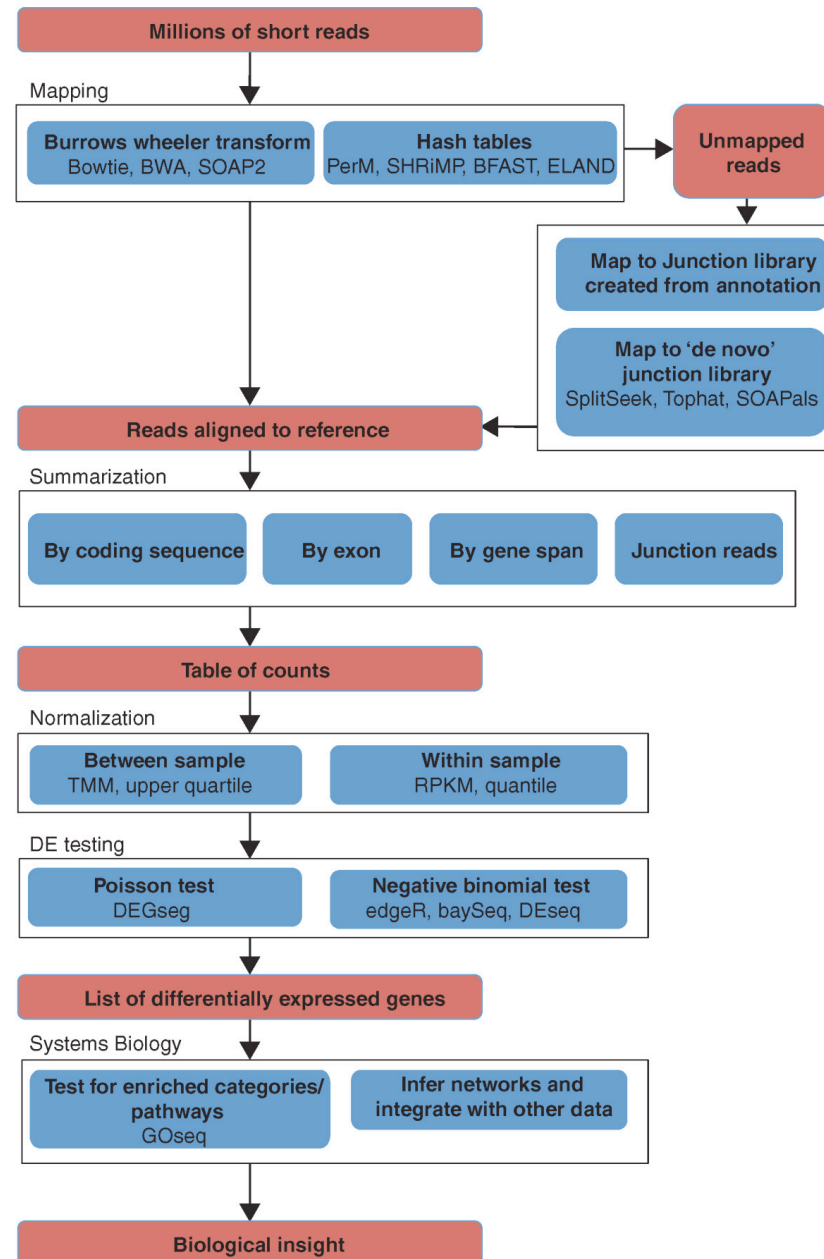
$$Q = -\log_{10}(P_{\text{error}})$$

Trimming

- Quality-based trimming
- Adapter 'contamination'

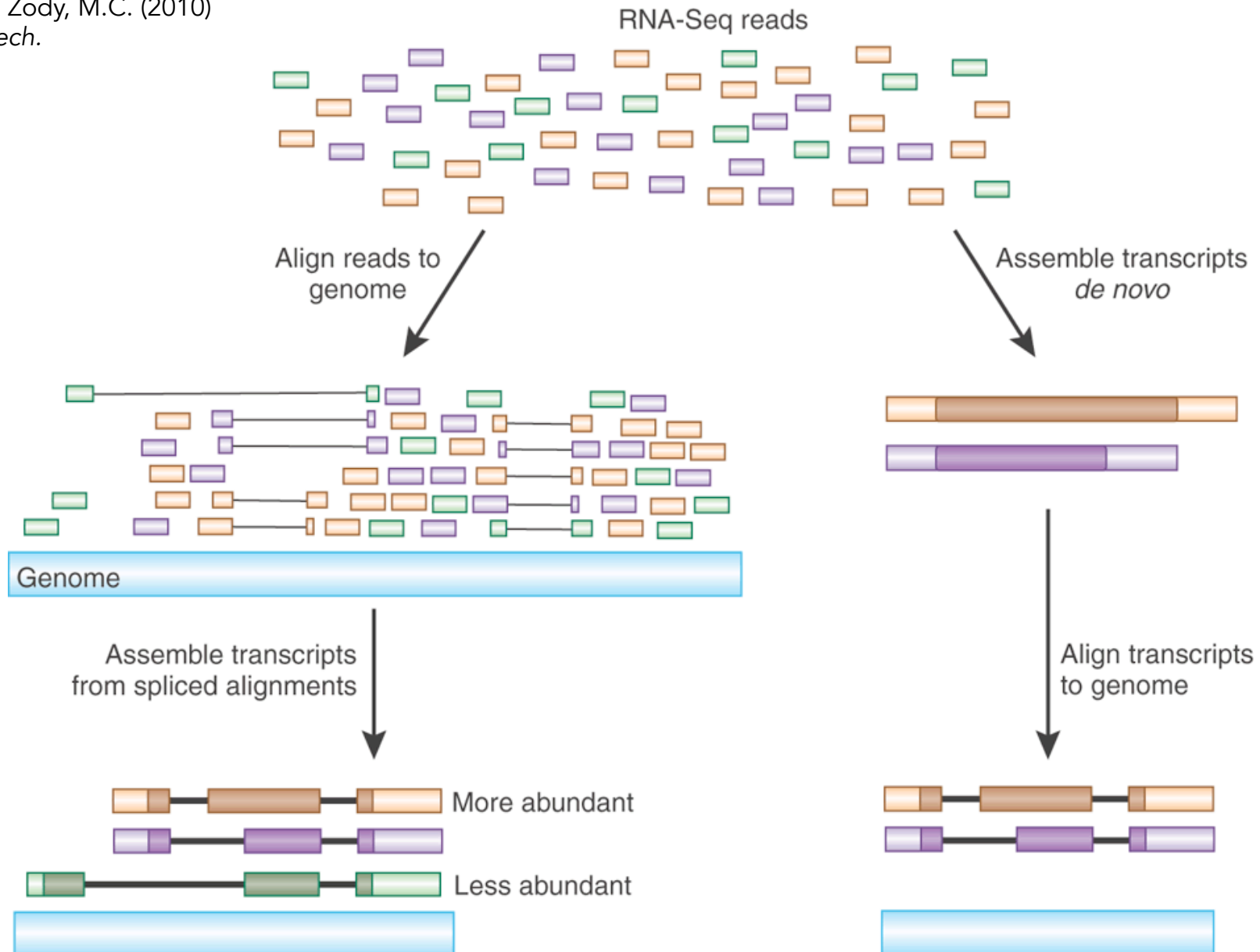


Analysis overview



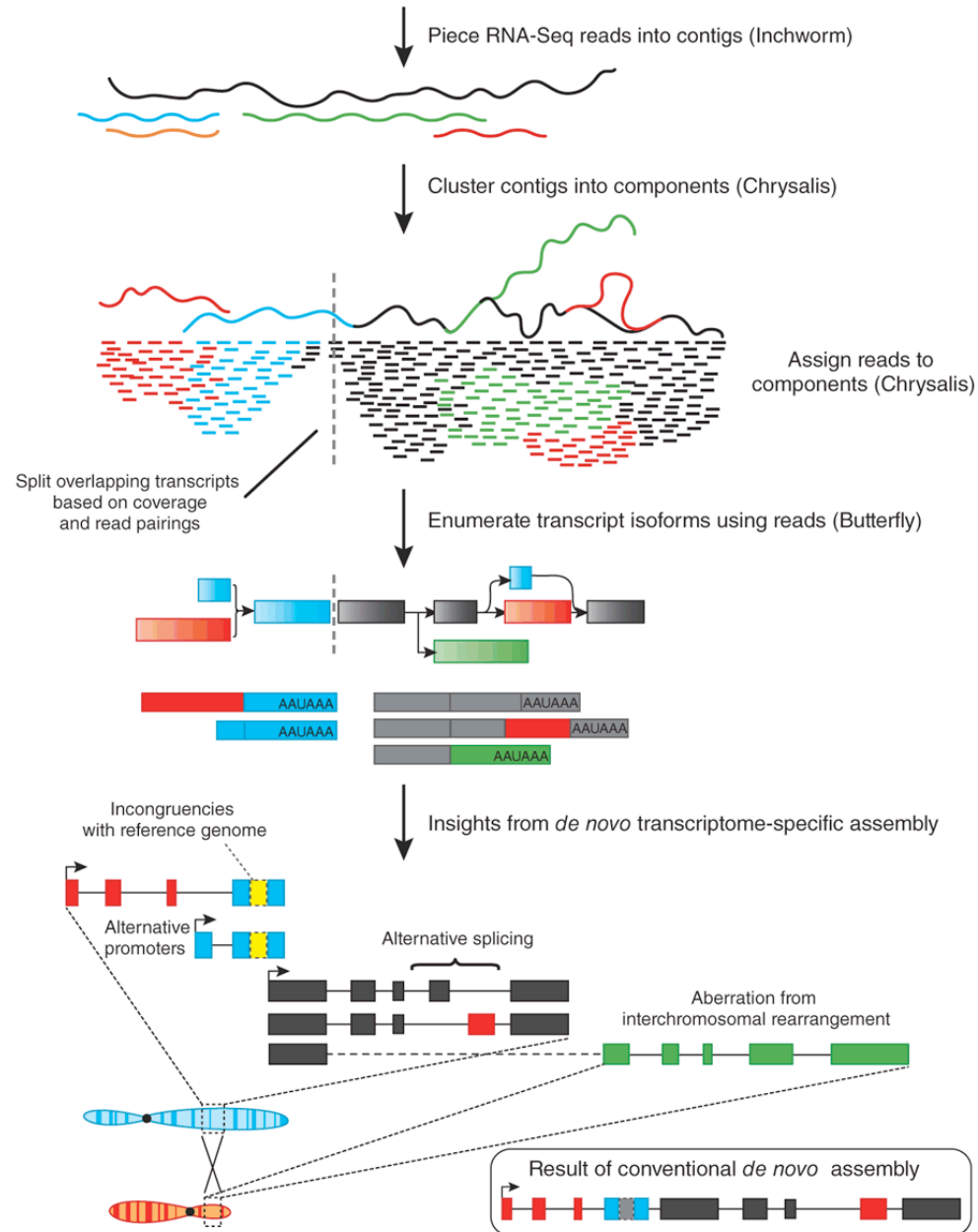
Sequence to sense

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

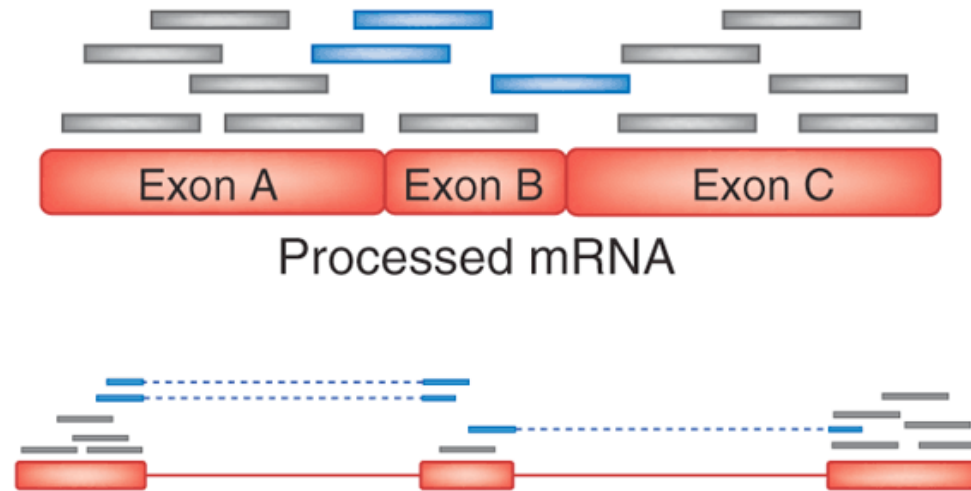


De novo assembly

- e.g. Trinity



Reference-based assembly



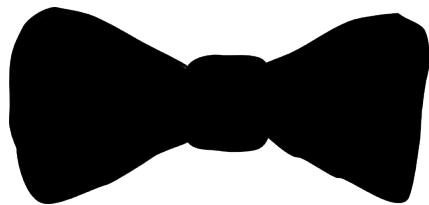
Genome mapping

- 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)




The Tuxedo suite



Bowtie
Extremely fast, general purpose short read aligner



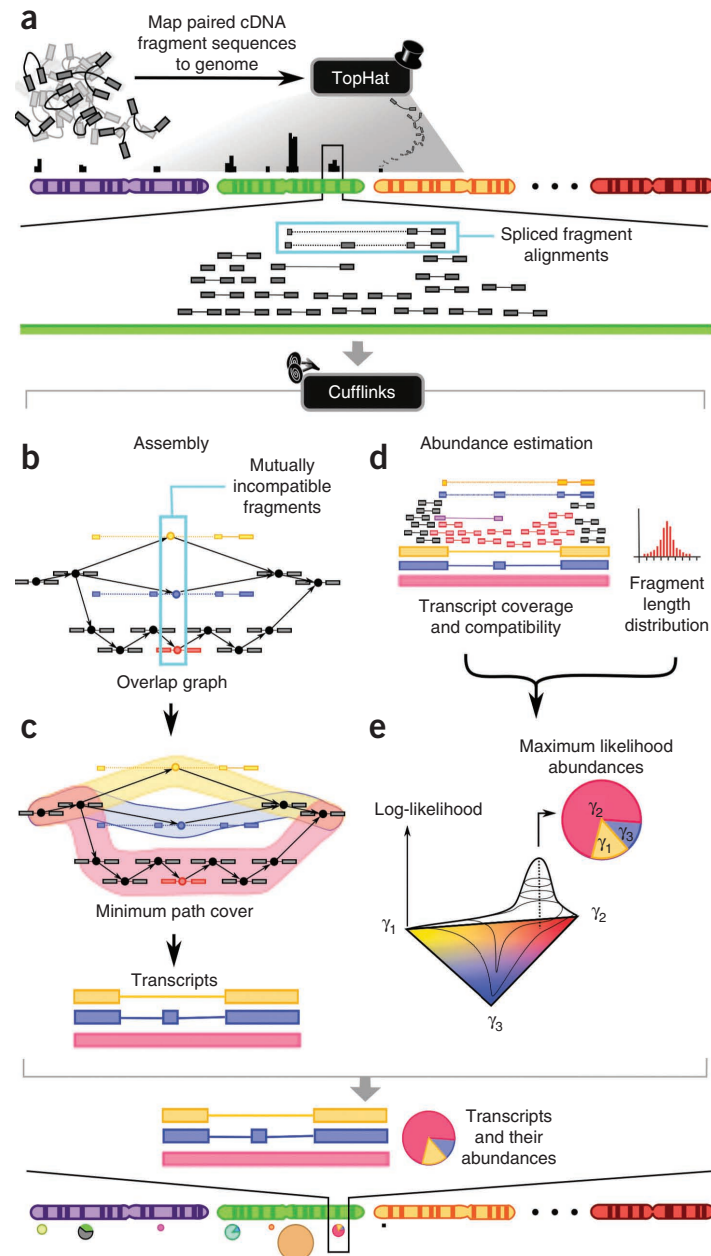
TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites



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



Tophat/Bowtie

(1) Transcriptome alignment (optional)

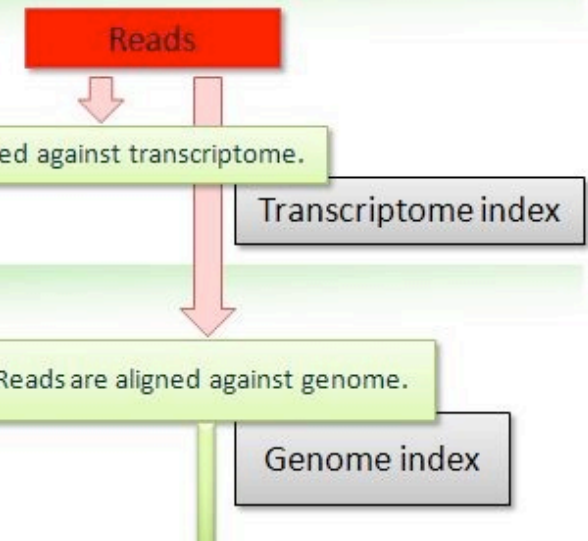
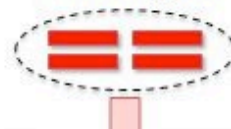


(2) Genome alignment

Reads spanning a single exon are **mapped**



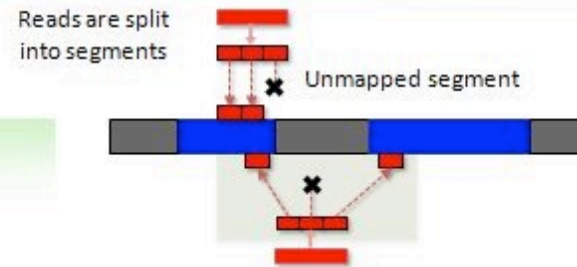
Multi-exon spanning reads are **unmapped**



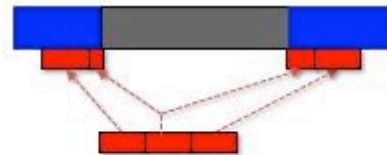
Tophat/Bowtie

(3) Spliced alignment

(3-1) Segment alignment to genome



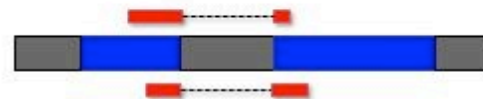
(3-2) Identification of splice sites (including indels and fusion break points)



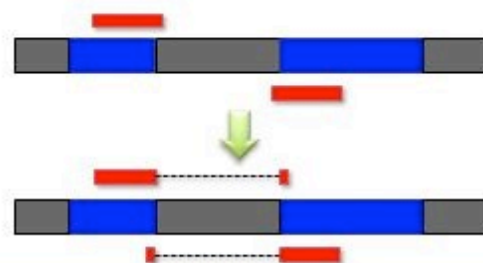
(3-3) Segments aligned to junction flanking sequences



(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.

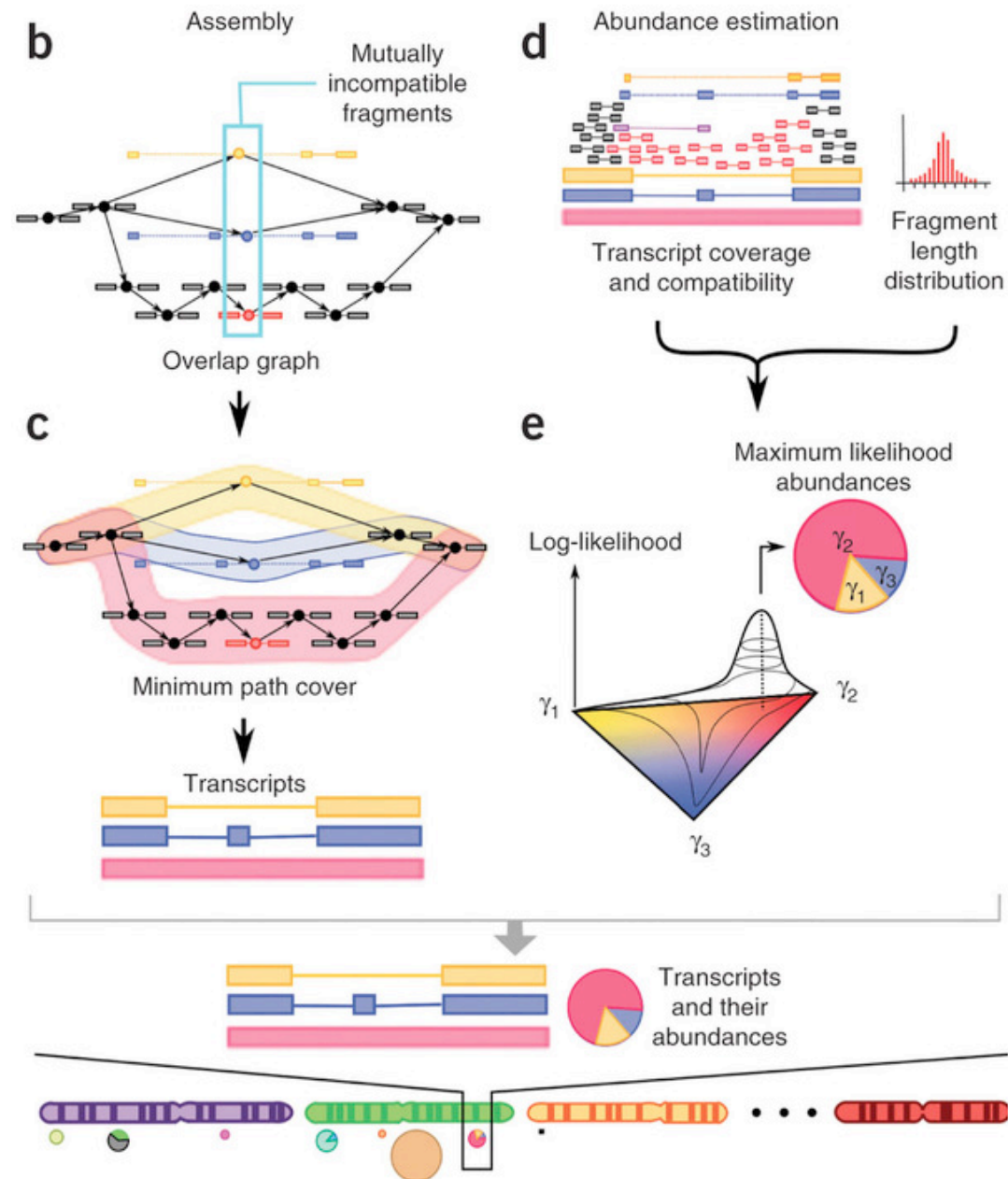
Sequences flanking a splice site are concatenated and segments are aligned to them.

Junction flanking index

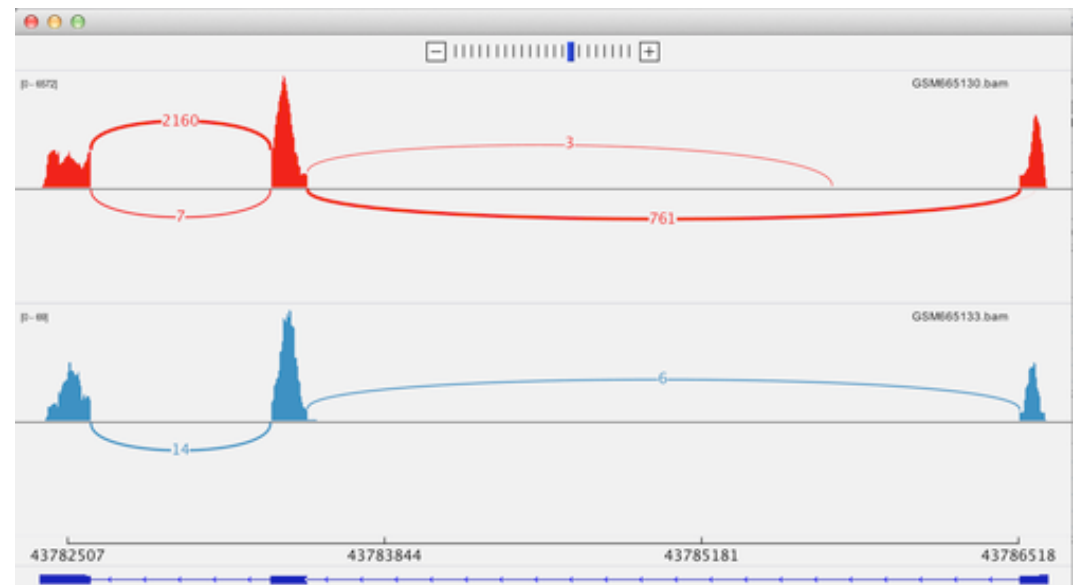
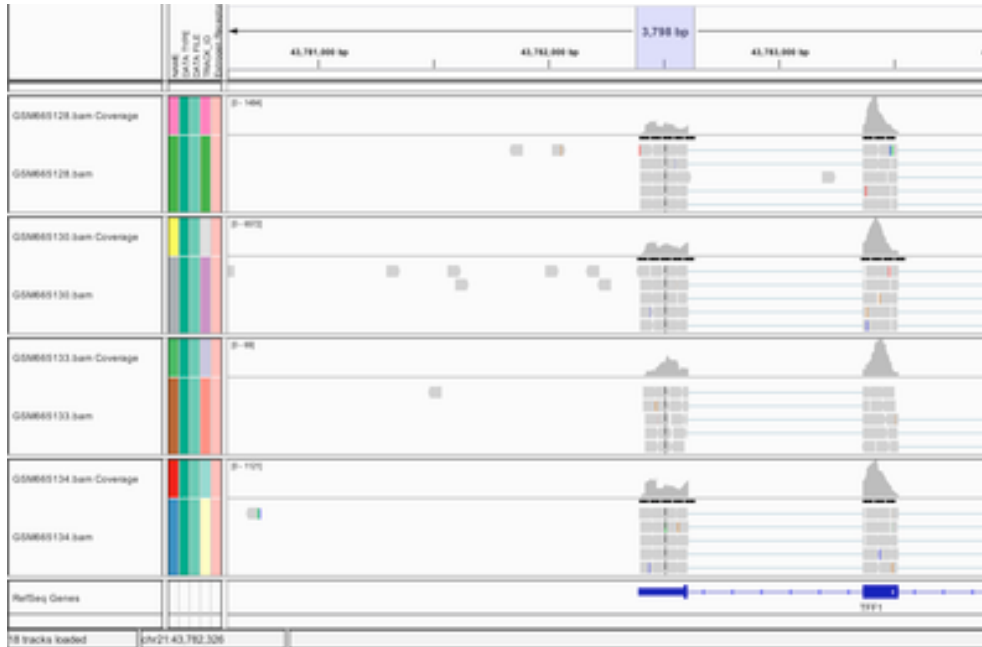
Mapped segments against either genome or flanking sequences are gathered to produce whole read alignments.

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

Platform/pipeline

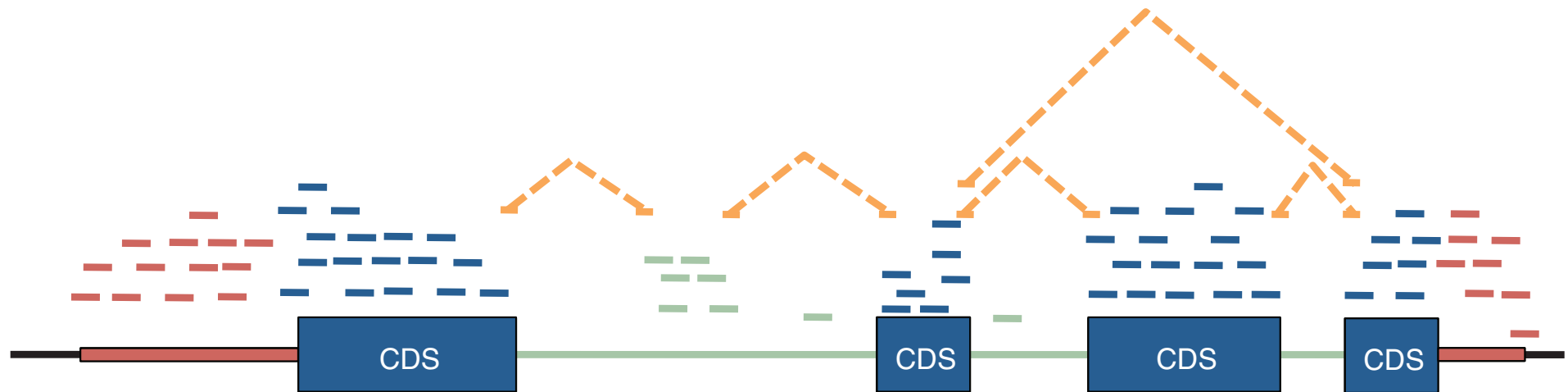
?

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

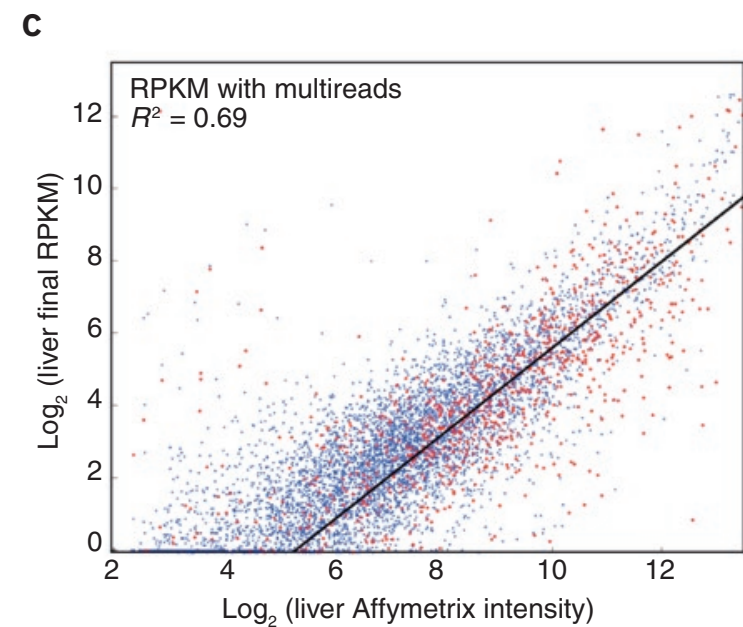
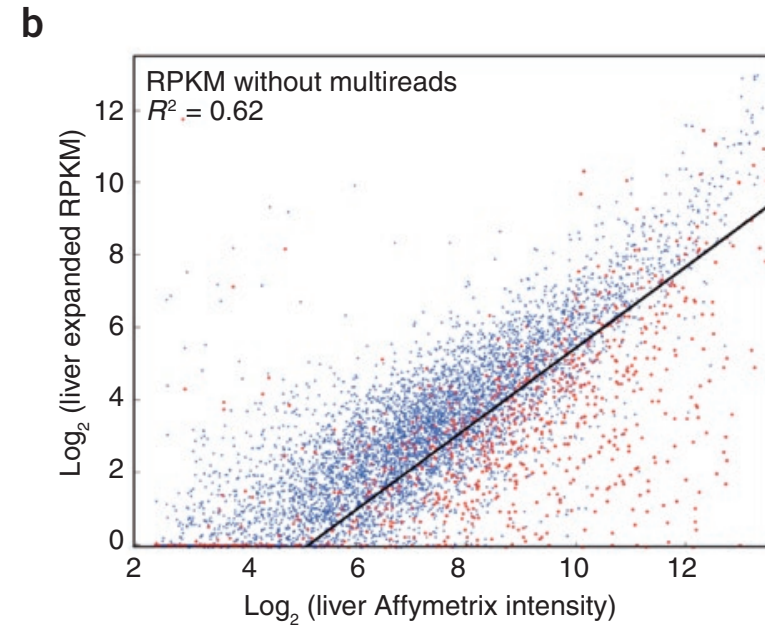
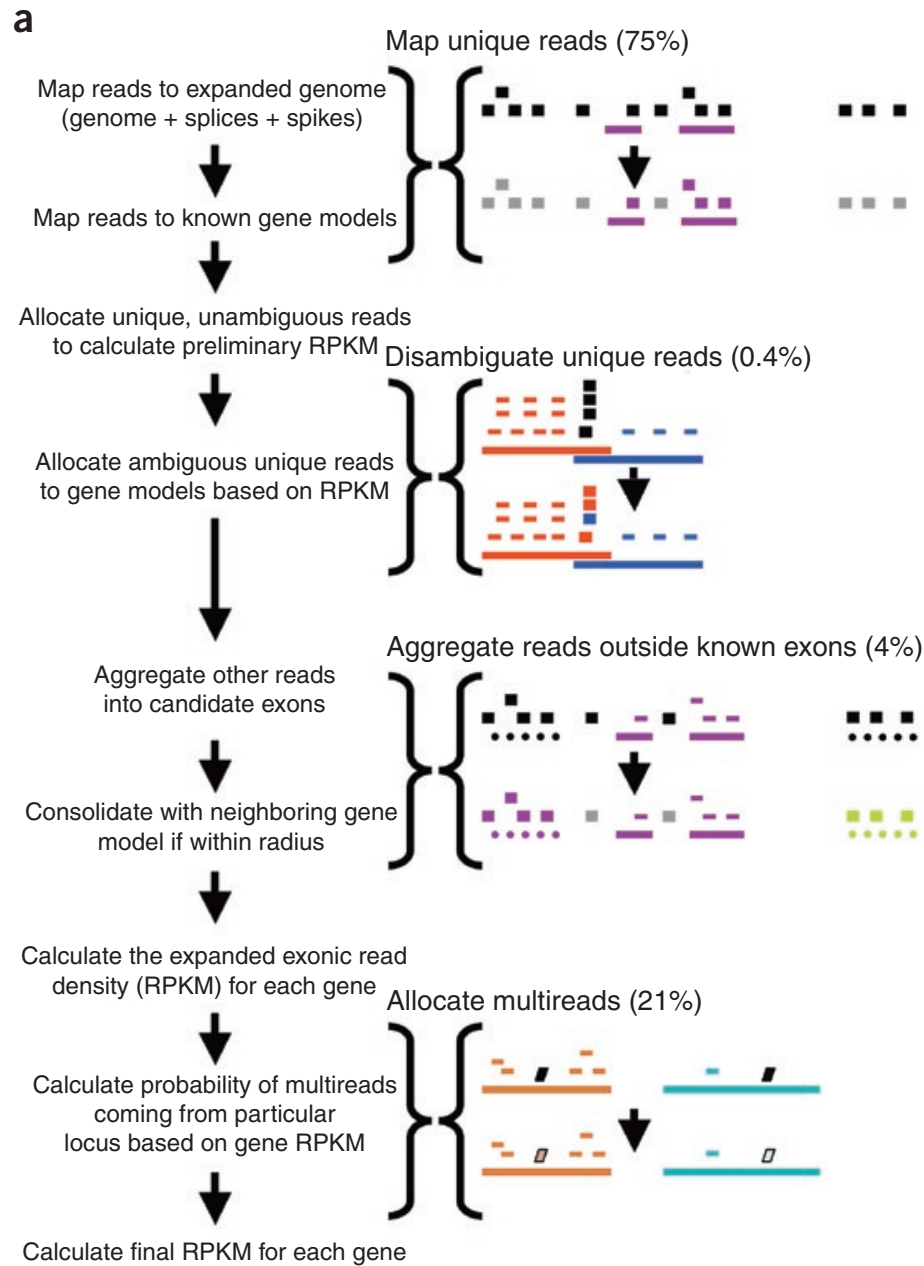
- Transcript assembly
- Novel structures
- Isoform multireads

Counting (e.g. Htseq)

<http://www-huber.embl.de/users/anders/HTSeq/doc/count.html>

	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 (e.g. ERANGE)



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

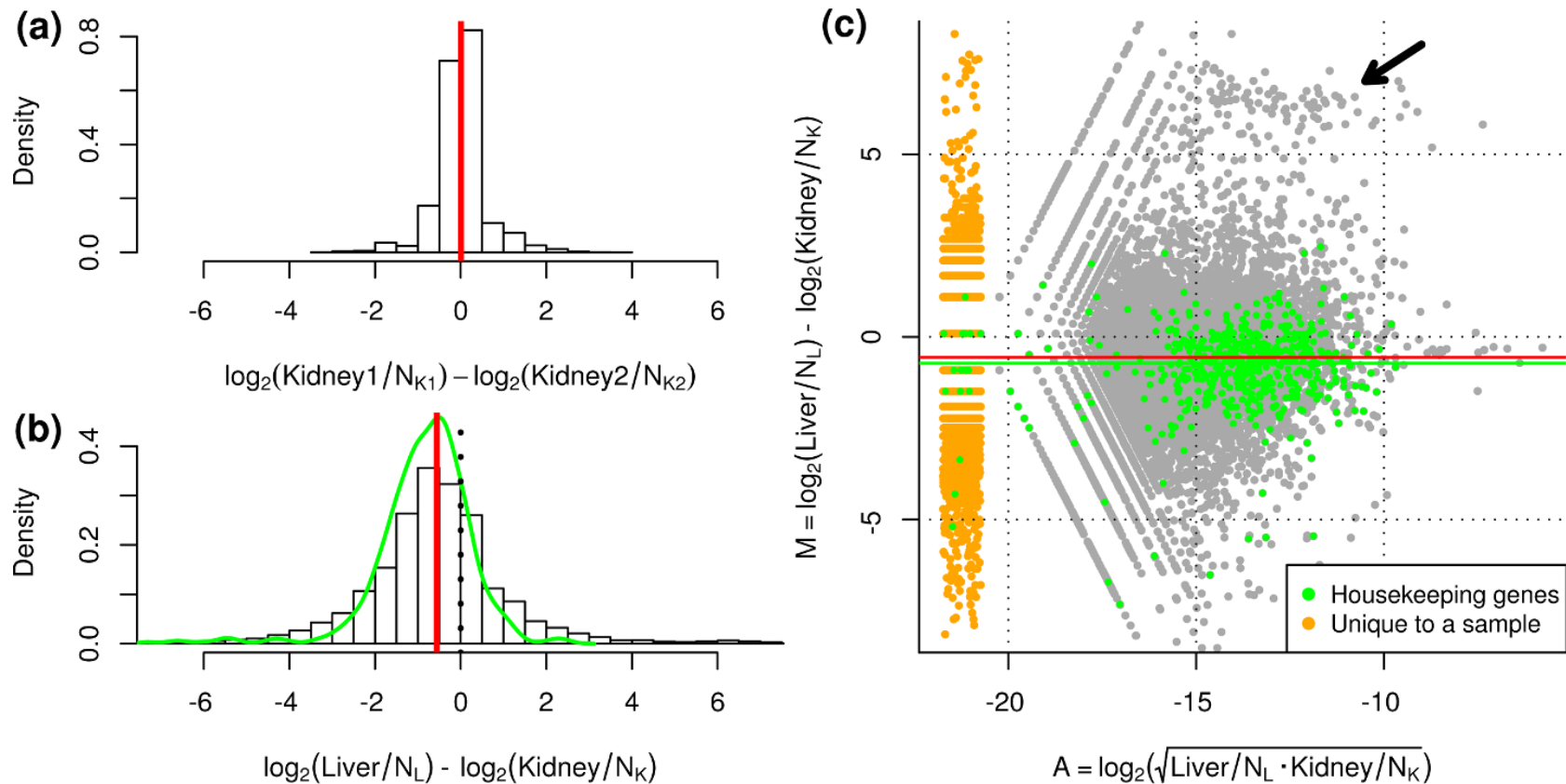
Library composition

- 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 (median, upper quartile)



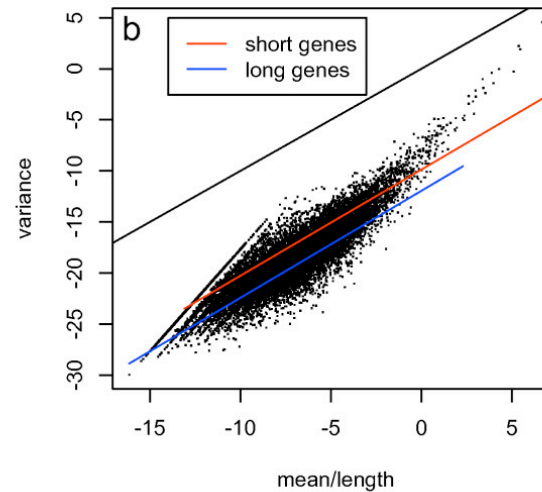
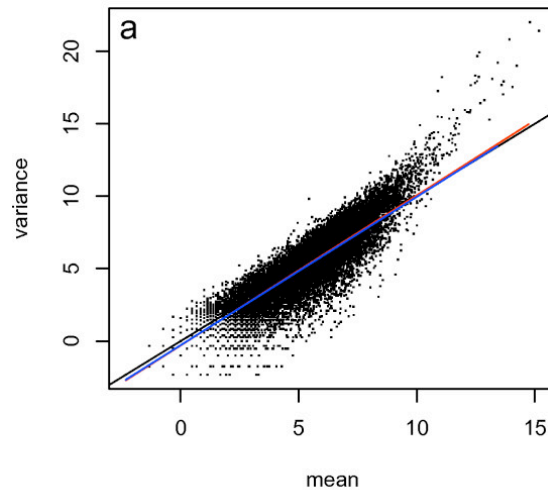
Normalisation ii

RPKM

- Reads per kilobase per million =

reads for gene A

length of gene A (kb) X Total number of reads (M)



Normalisation ii

cRPKM

- Corrected reads per kilobase per million =

reads for gene A

uniquely mappable positions in gene A (k) X Total # of mapped reads (M)

Dependent on read length:

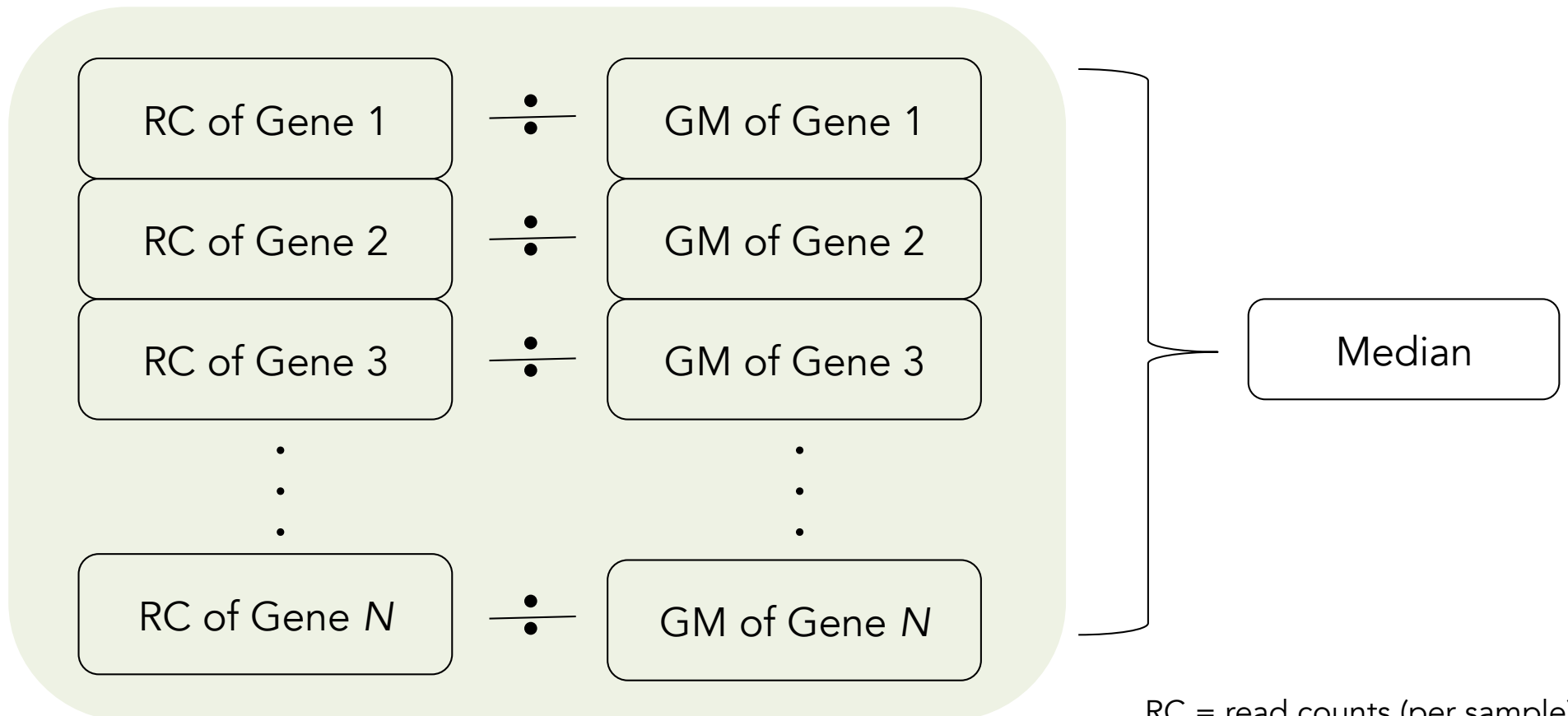
Maximum gene mappability = gene length – read length + 1

Normalisation iii

Geometric scaling factor

```
estimateSizeFactors()  
sizeFactors()
```

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

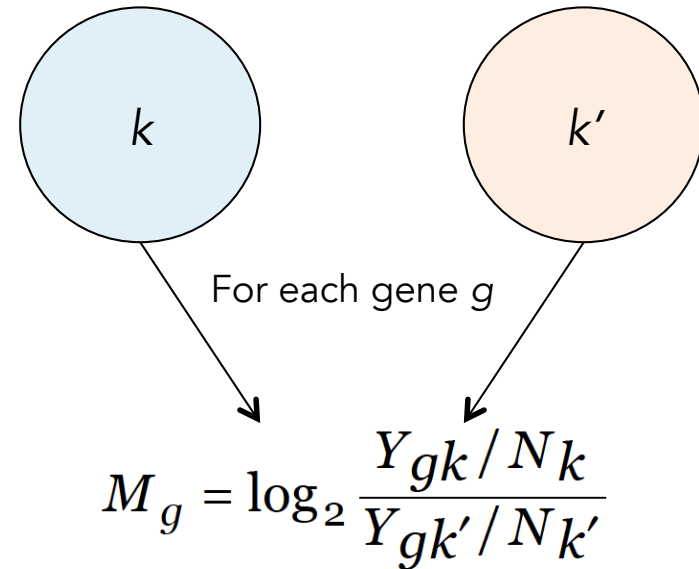


RC = read counts (per sample)
GM = geometric mean (all samples)

Normalisation iv

Trimmed mean of M

- Implemented in edgeR
`calcNormFactors()`
- Assumes most genes are not differentially expressed



$$A_g = \frac{1}{2} \log_2 (Y_{gk}/N_k \cdot Y_{gk'}/N_{k'}) \text{ for } Y_{g\bullet} \neq 0$$

Y_{gk} - observed count for gene g in library k
 N_k - total number of reads for library k

$$\log_2(\text{TMM}_k^{(r)}) = \frac{\sum_{g \in G^*} w_{gk}^r M_{gk}^r}{\sum_{g \in G^*} w_{gk}^r} \text{ where } M_{gk}^r = \frac{\log_2 \left(\frac{Y_{gk}}{N_k} \right)}{\log_2 \left(\frac{Y_{gr}}{N_r} \right)} \text{ and } w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}};$$

$Y_{gk}, Y_{gr} > 0.$

r - reference sample
 G^* - not trimmed genes

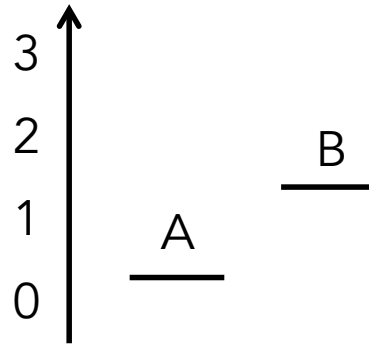
Weight each gene by inverse of its variance ('trimming'*)

[*typically 30% on M and 5% on A]

Mean weighted ratio

Differential expression

- Simple

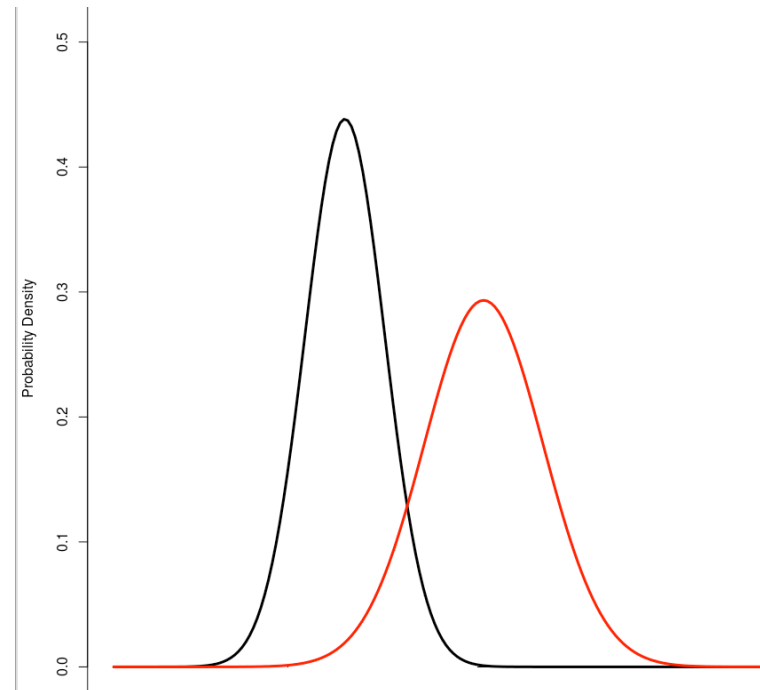
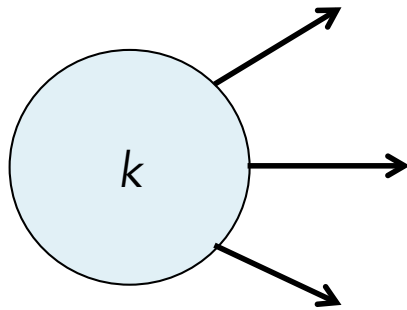


All we need

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

Modelling – old trends

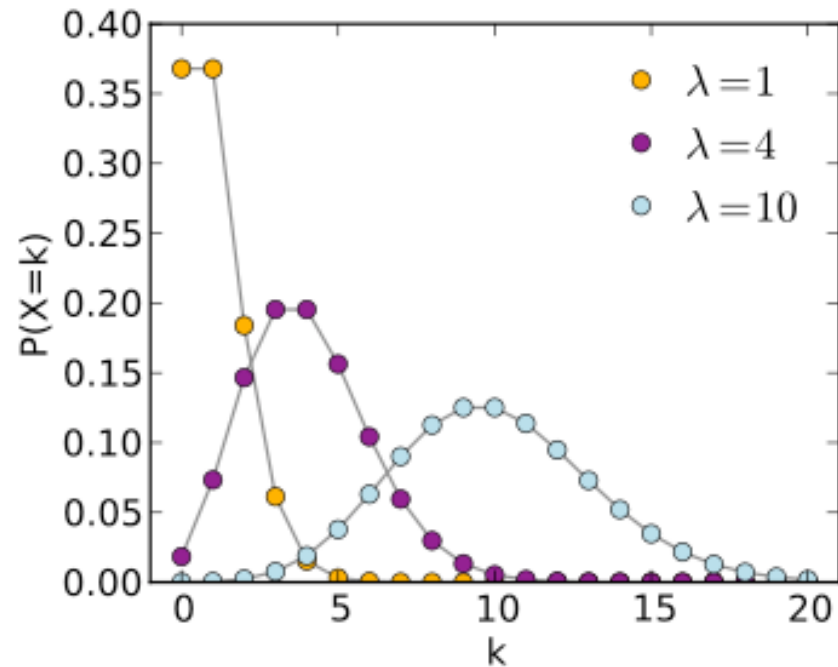
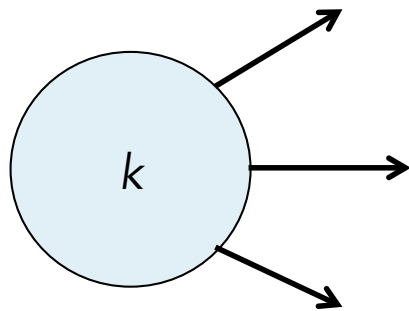
- Technical replicates introduce some variance



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

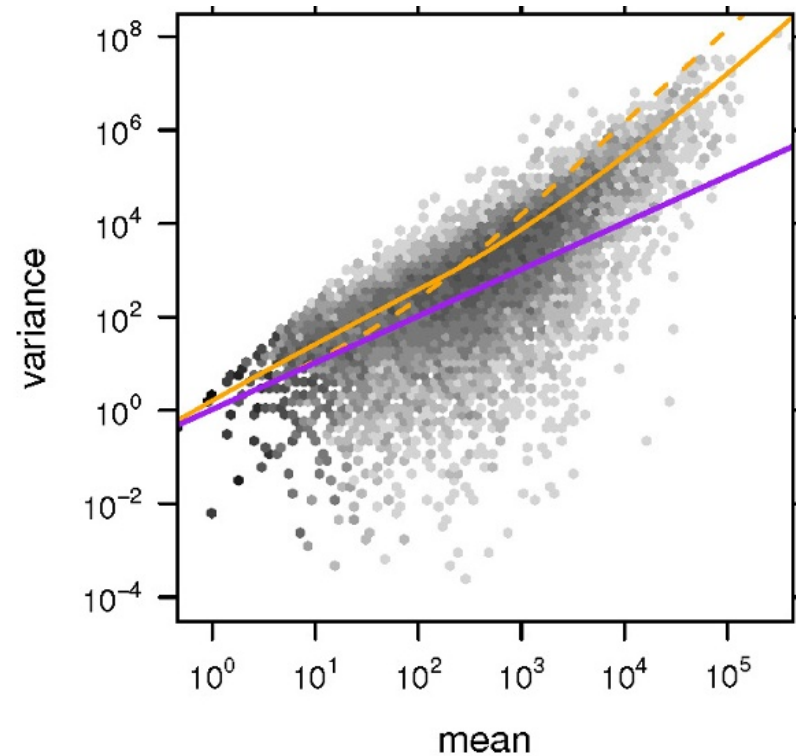
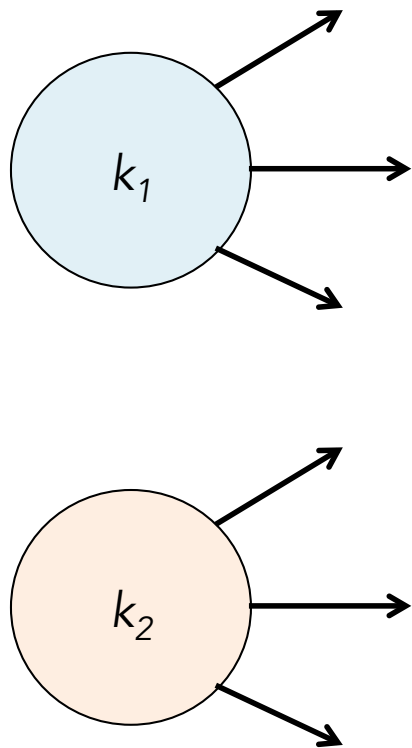
Modelling – in fashion

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



Modelling – in fashion

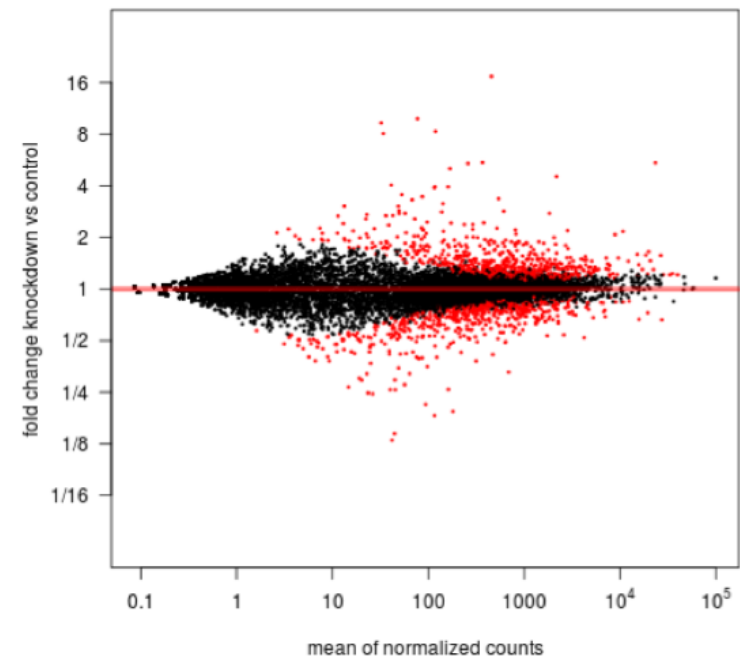
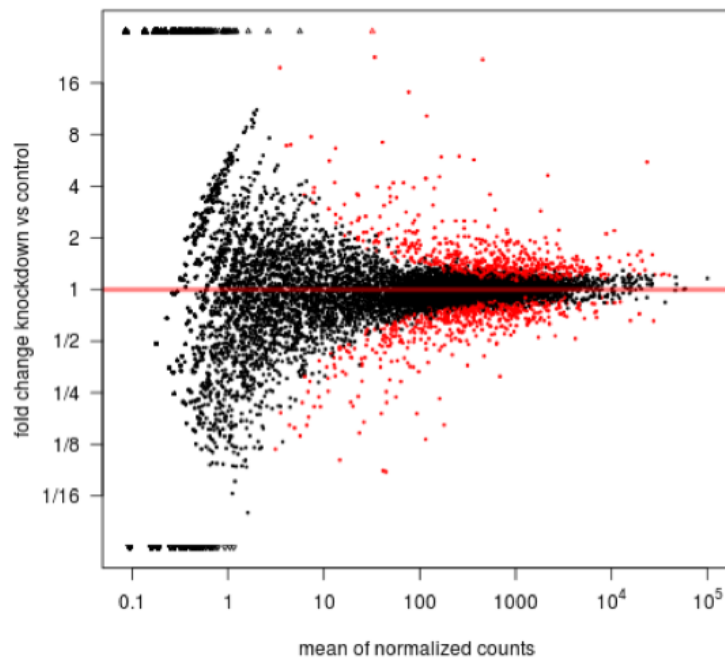
- Biology is never that simple...



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

Modelling – in fashion

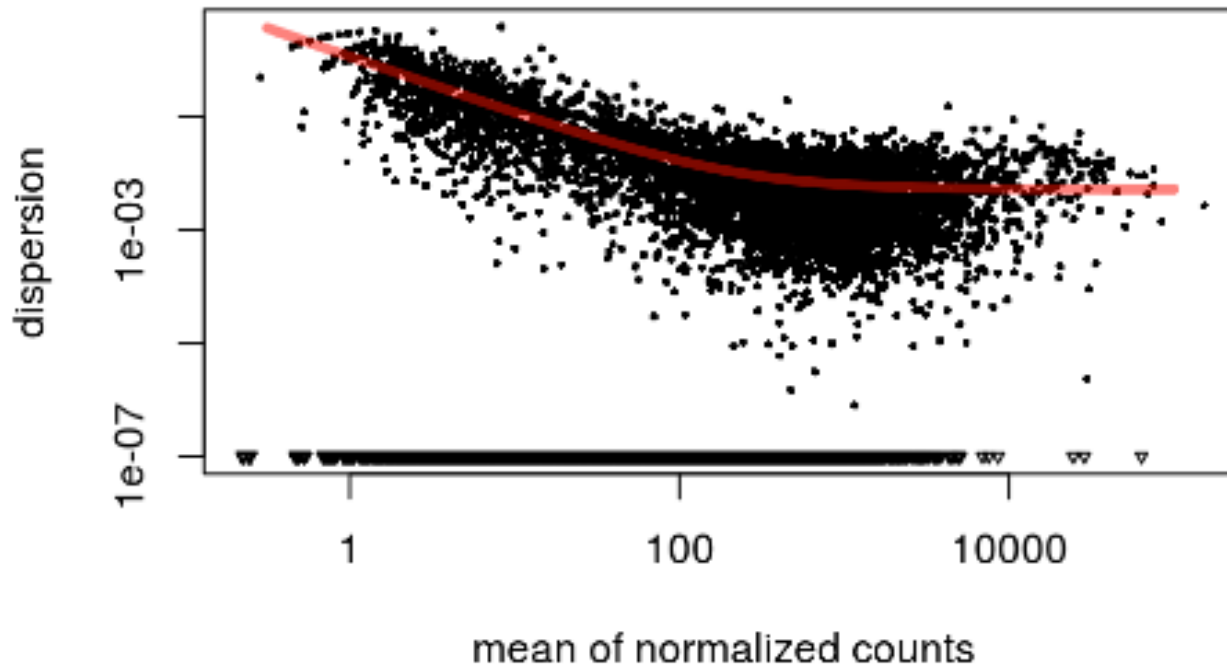
- Estimating the dispersion parameter can be difficult with a small number of samples
- edgeR: models the variance as the sum of technical and biological variance
- ‘Share’ information from all genes to obtain global estimate - *shrinkage*



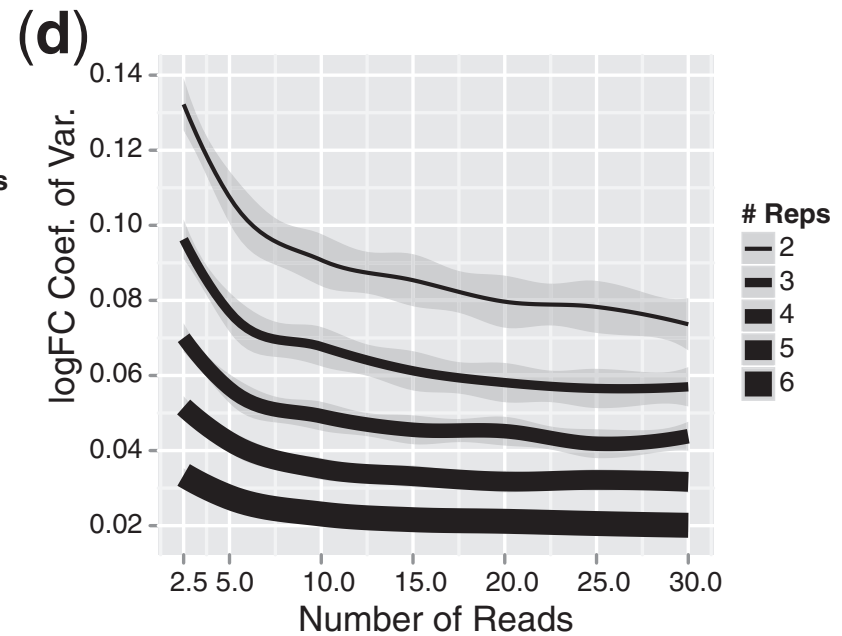
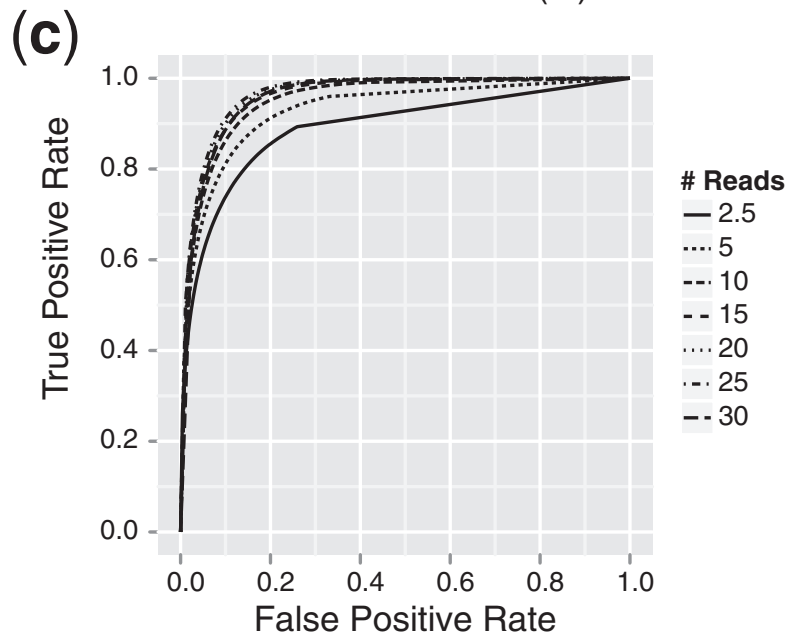
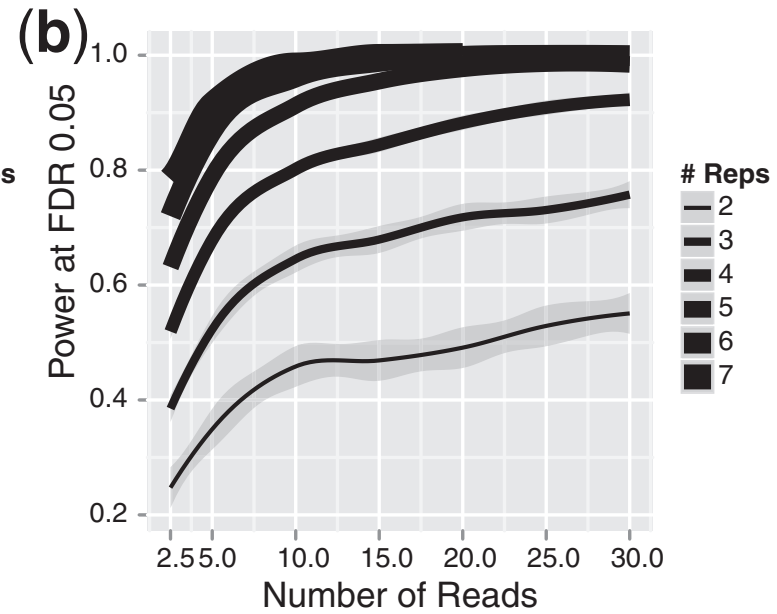
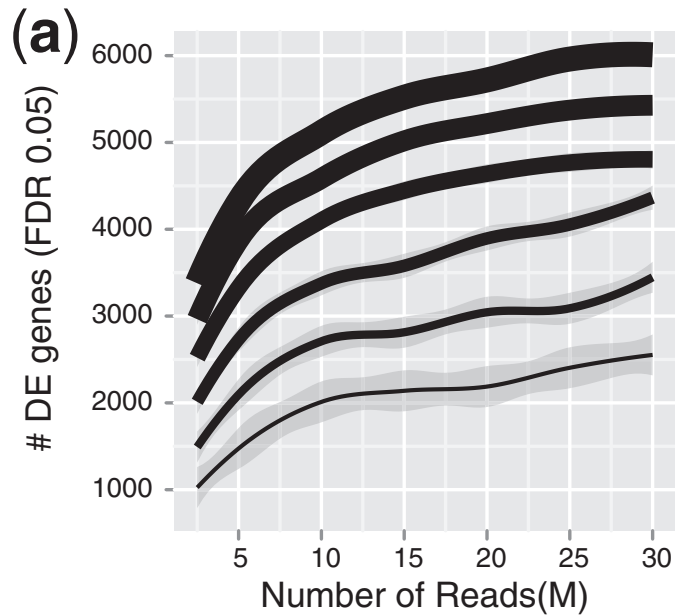
Modelling – in fashion

- DESeq uses a similar formulation of the variance term

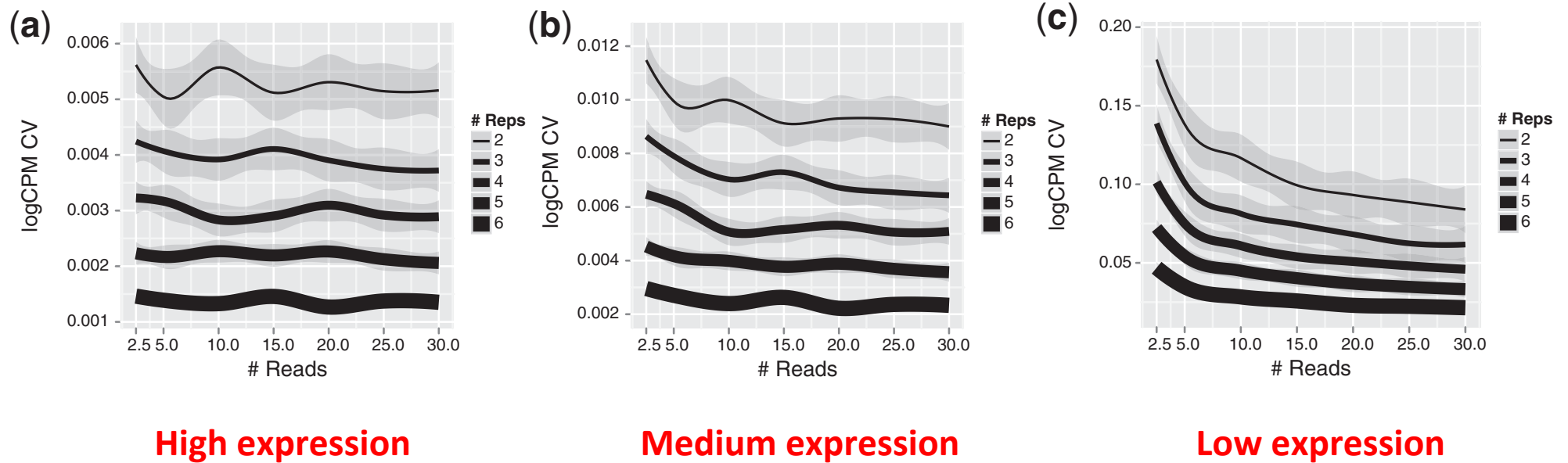
$$\sigma_{ij}^2 = \underbrace{\mu_{ij}}_{\text{shot noise}} + \underbrace{s_j^2 v_{i,\rho(j)}}_{\text{raw variance}}$$



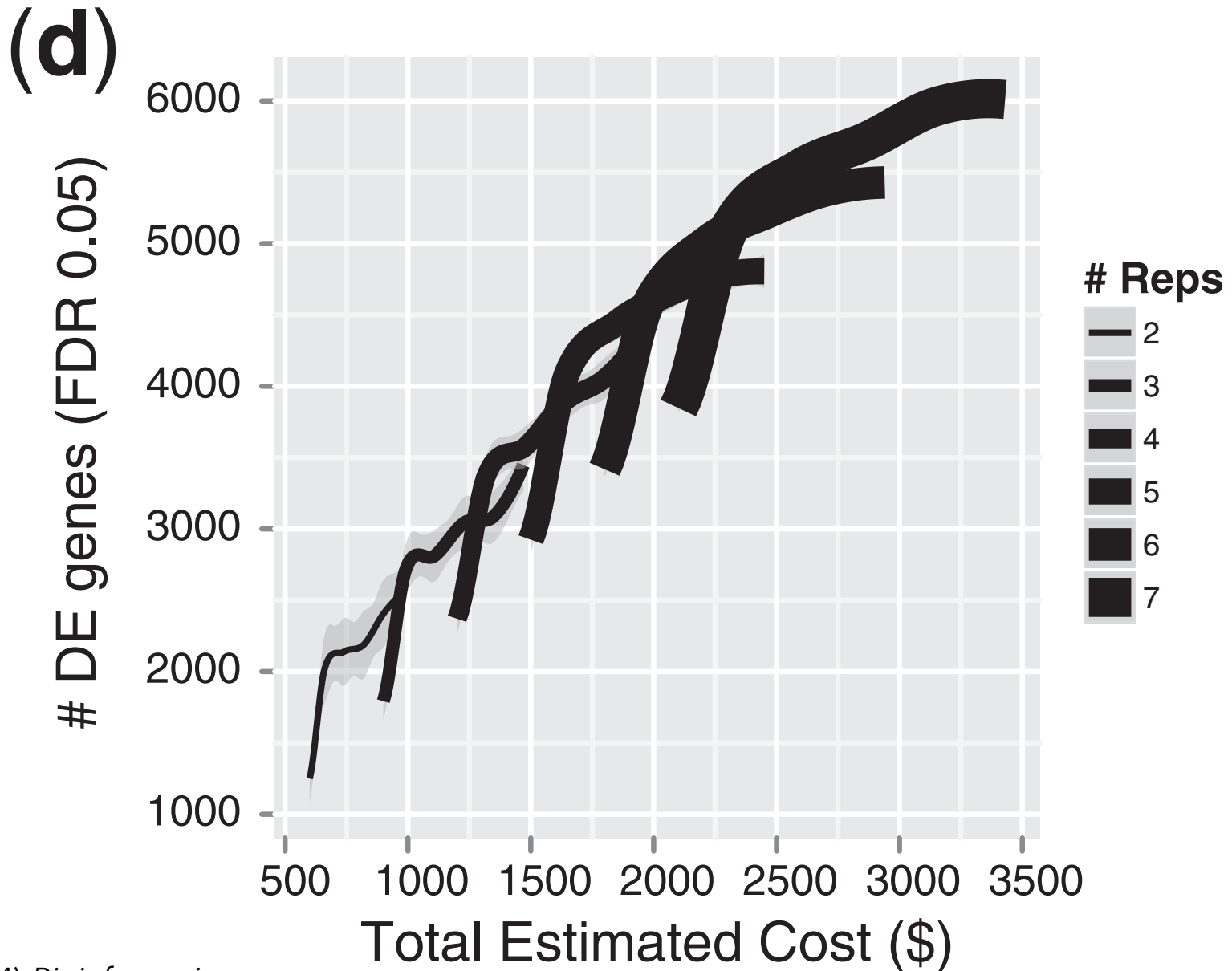
On replicates...



On replicates...

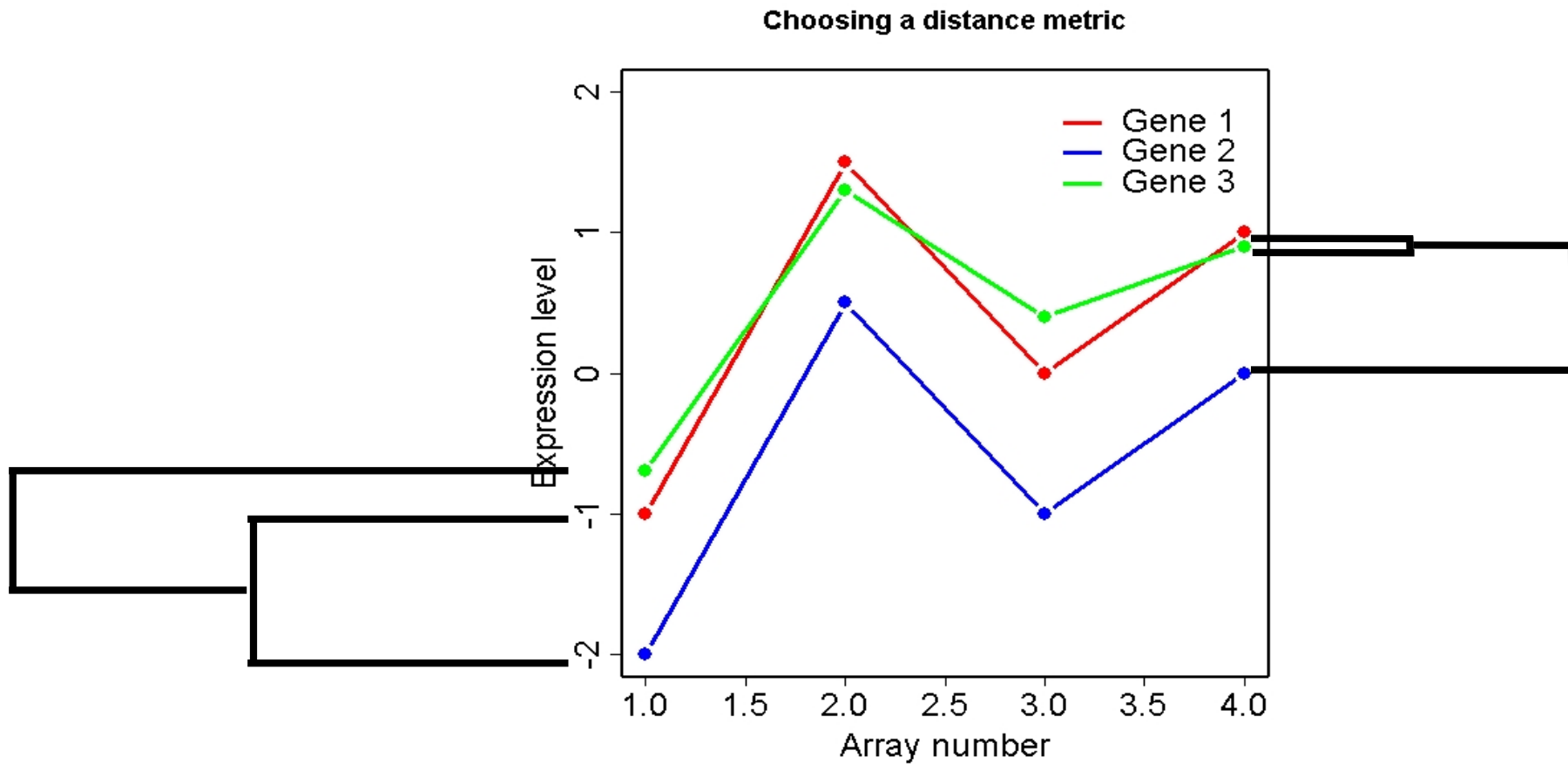


On replicates...



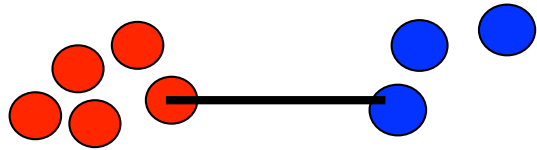
What next?

- Hierarchical clustering = define metric & look for similarities

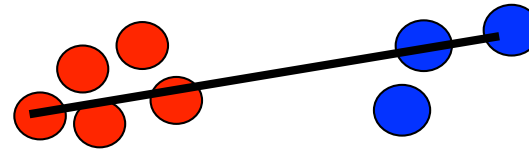


What next?

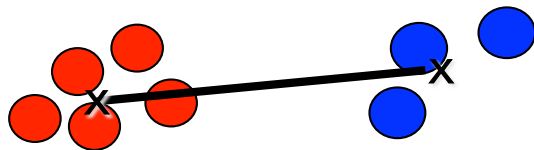
- Merging clusters according to a metric



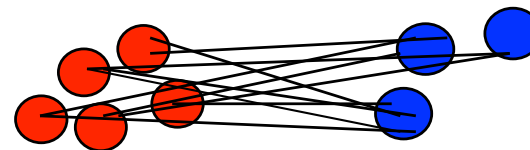
Single
(min. of pairwise distances)



Complete
(max. of pairwise distances)

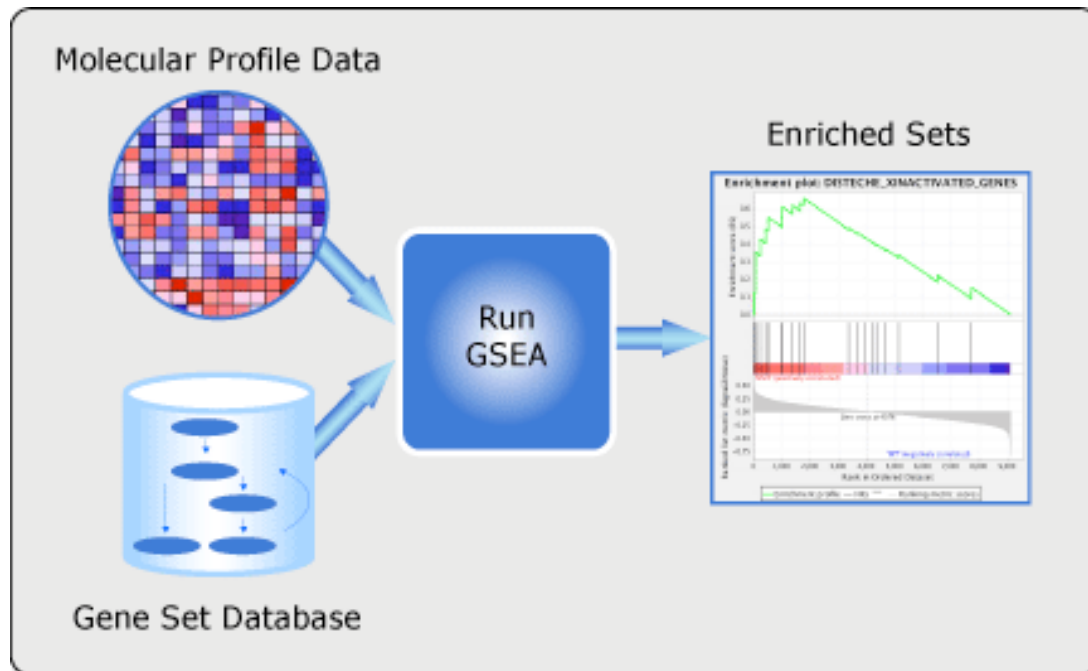


Distance between centroids



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) [?]