

# Analysis of RNA-seq Data

Bernard Pereira



UNIVERSITY OF  
CAMBRIDGE



# The many faces of RNA-seq



AREAS OF INTEREST ▾ TECHNIQUES ▾ SYSTEM

## RNA Sequencing

Overview >

[Targeted RNA Sequencing](#)

[mRNA-Seq](#)

[Total RNA-Seq](#)

[Small RNA-Seq](#)

[Low-Quality/FFPE RNA-Seq](#)

[Ultra-Low-Input & Single-Cell RNA-Seq](#)

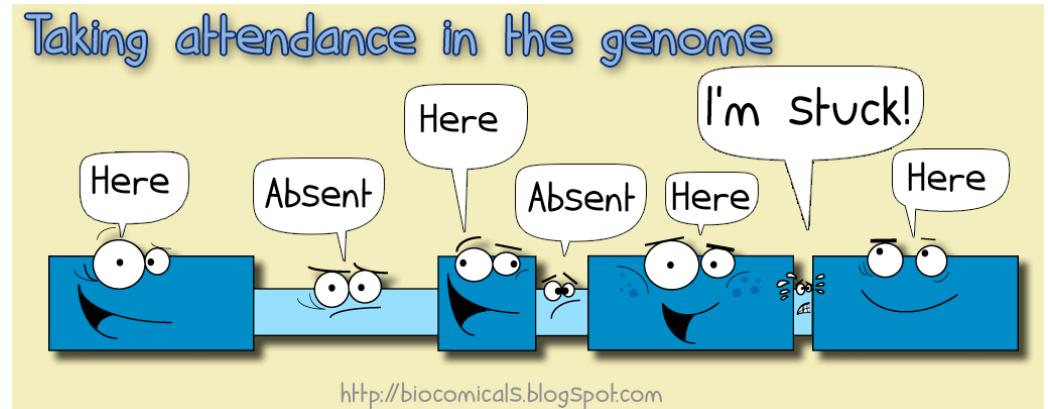
[Ribosome Profiling](#)

[RNA-Seq Data Analysis](#)

# Applications – discovery

## Discovery

- Find new transcripts
- Build gene models
- Mutation expression



# Applications – comparison

## Comparison

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

- Gene expression levels
- Isoform abundance ratios, splice patterns, transcript boundaries

# Examples

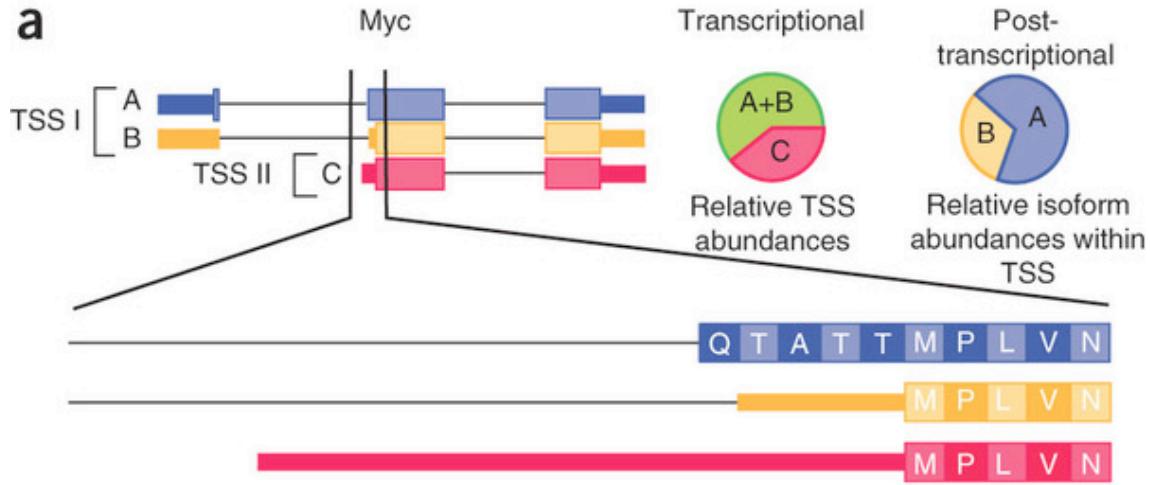
nature  
biotechnology

LETTERS

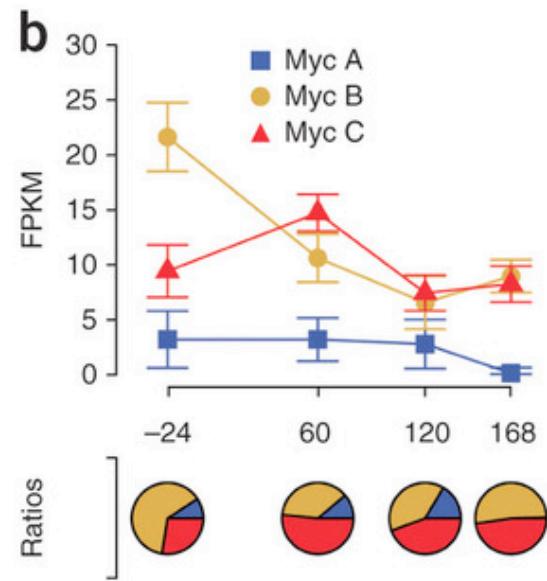
Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

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

a



b



# Examples

Journal of Pathology

J Pathol 2015; 235: 571–580

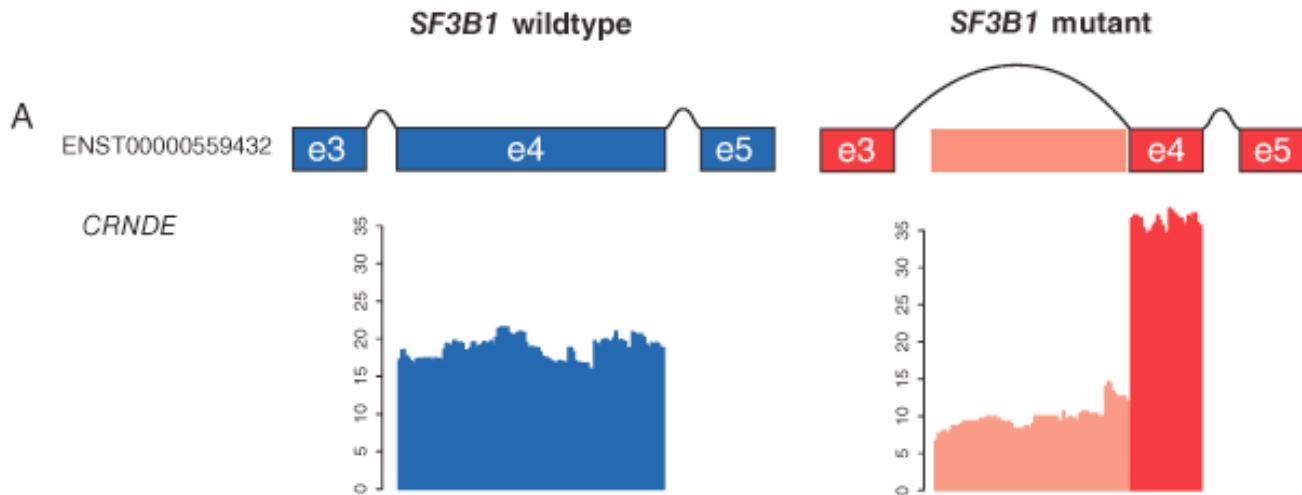
Published online 22 December 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/path.4483

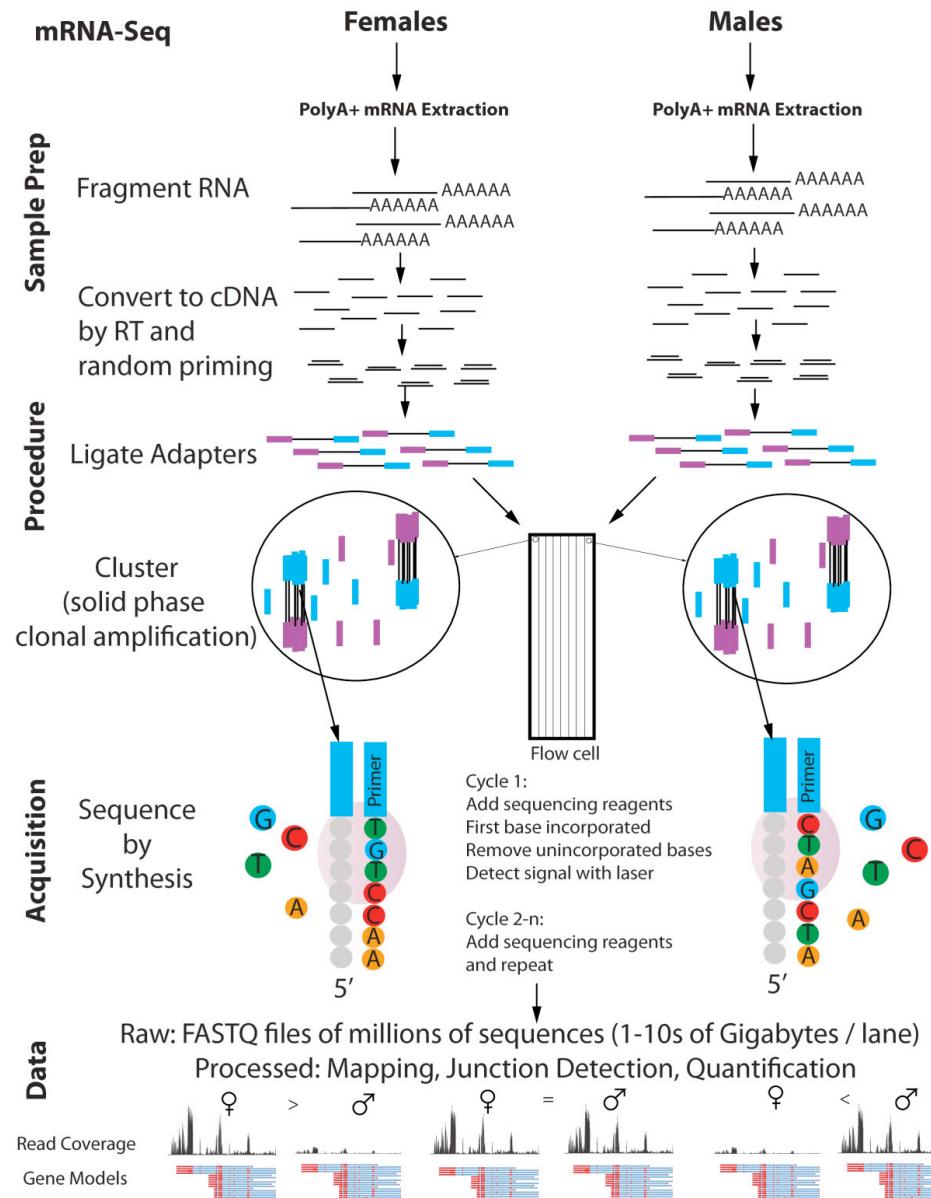
ORIGINAL PAPER

## *SF3B1* mutations constitute a novel therapeutic target in breast cancer

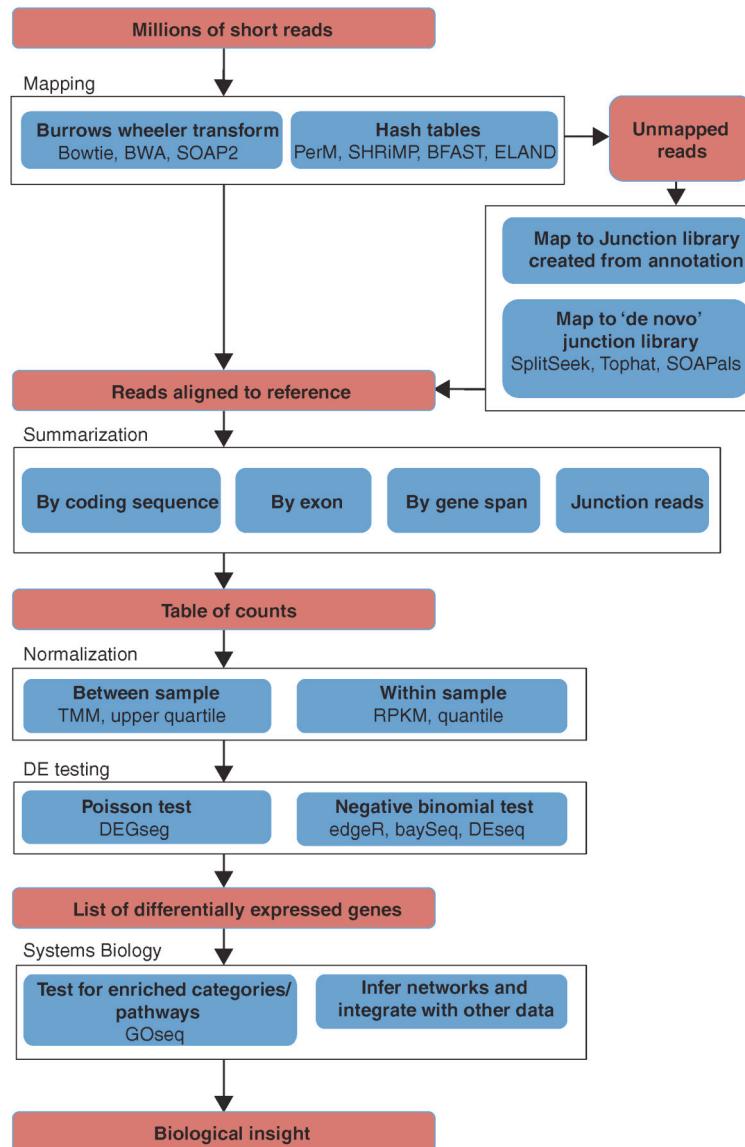
Sarah L Maguire,<sup>1,†</sup> Andri Leonidou,<sup>1,2,†</sup> Patty Wai,<sup>1,2,†</sup> Caterina Marchiò,<sup>2,3</sup> Charlotte KY Ng,<sup>3,4</sup> Anna Sapino,<sup>2</sup> Anne-Vincent Salomon,<sup>5,6</sup> Jorge S Reis-Filho,<sup>3,4</sup> Britta Weigelt<sup>3,4</sup> and Rachael C Natrajan<sup>1,2,\*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



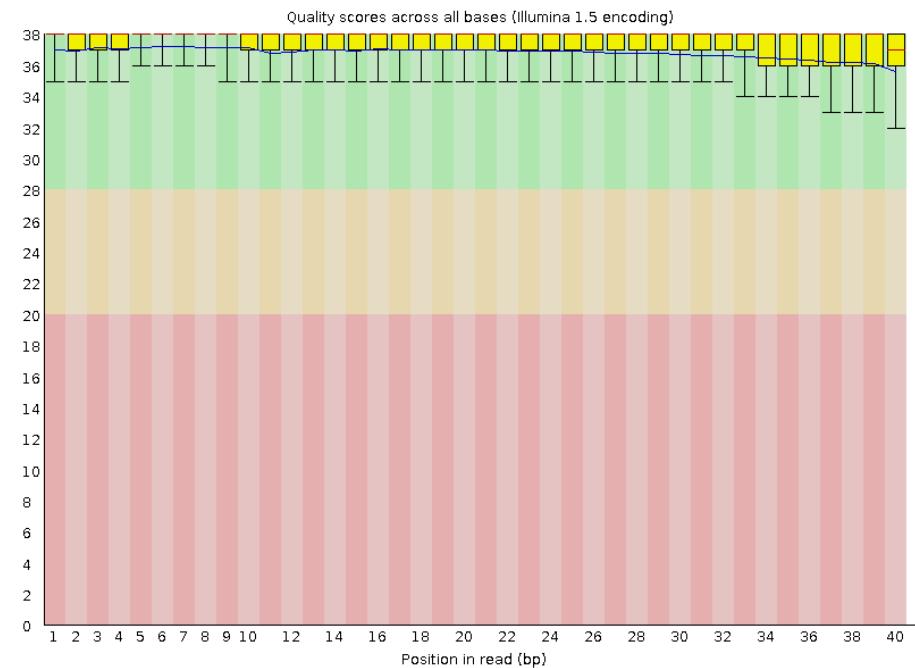
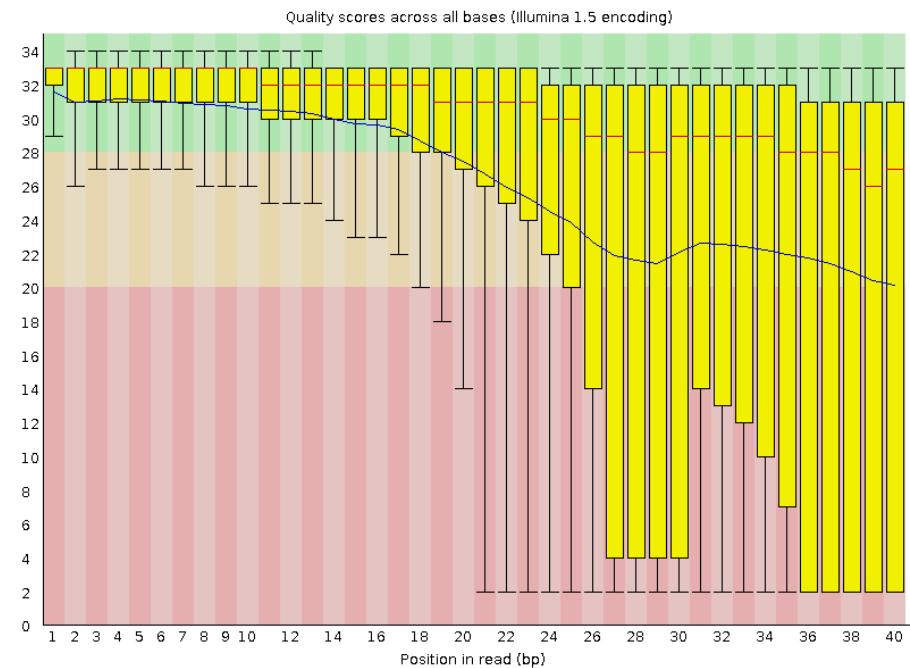
# Workflow – library prep



# Analysis pipeline

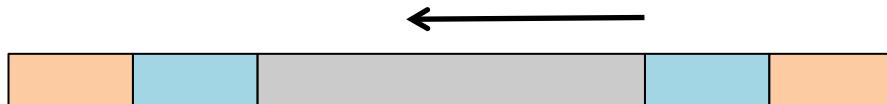


# FASTQC

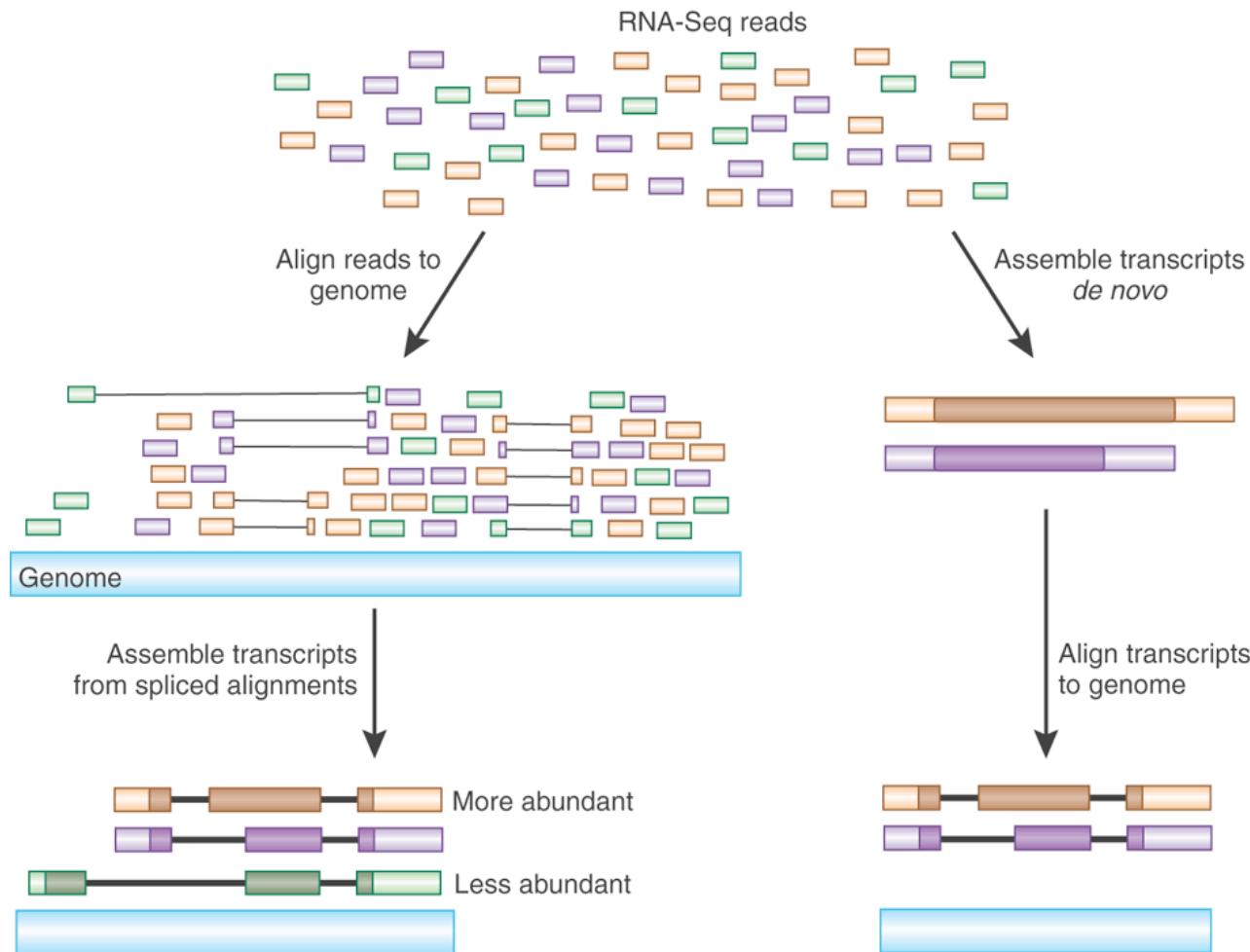


# Trimming

- Quality-based trimming
- Adapter 'contamination'

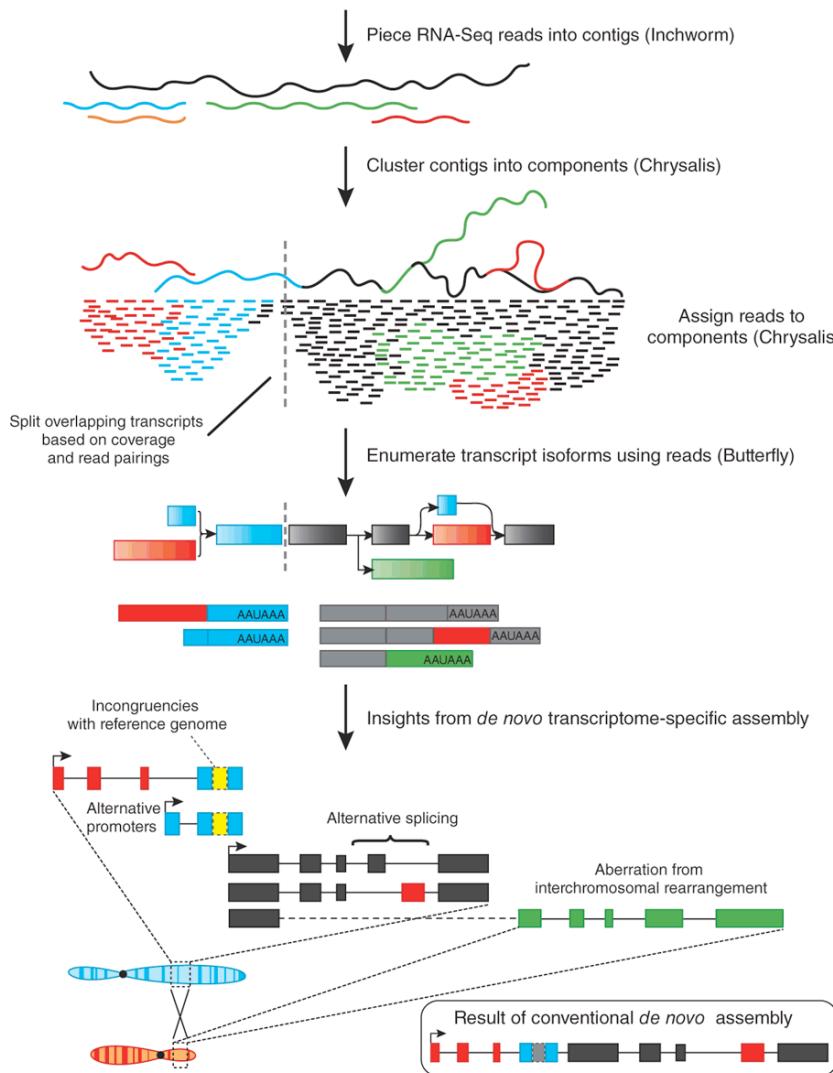


# Sequence to sense

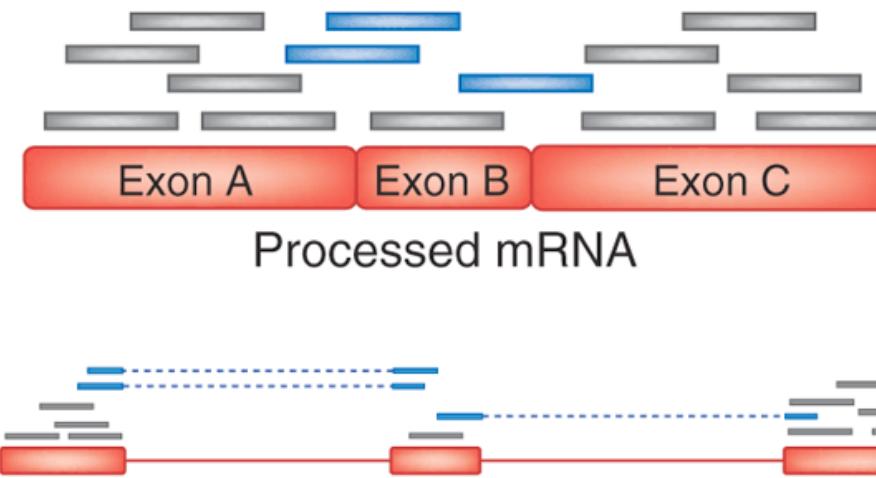


# *De novo* assembly

- eg. Trinity



# Reference-based assembly



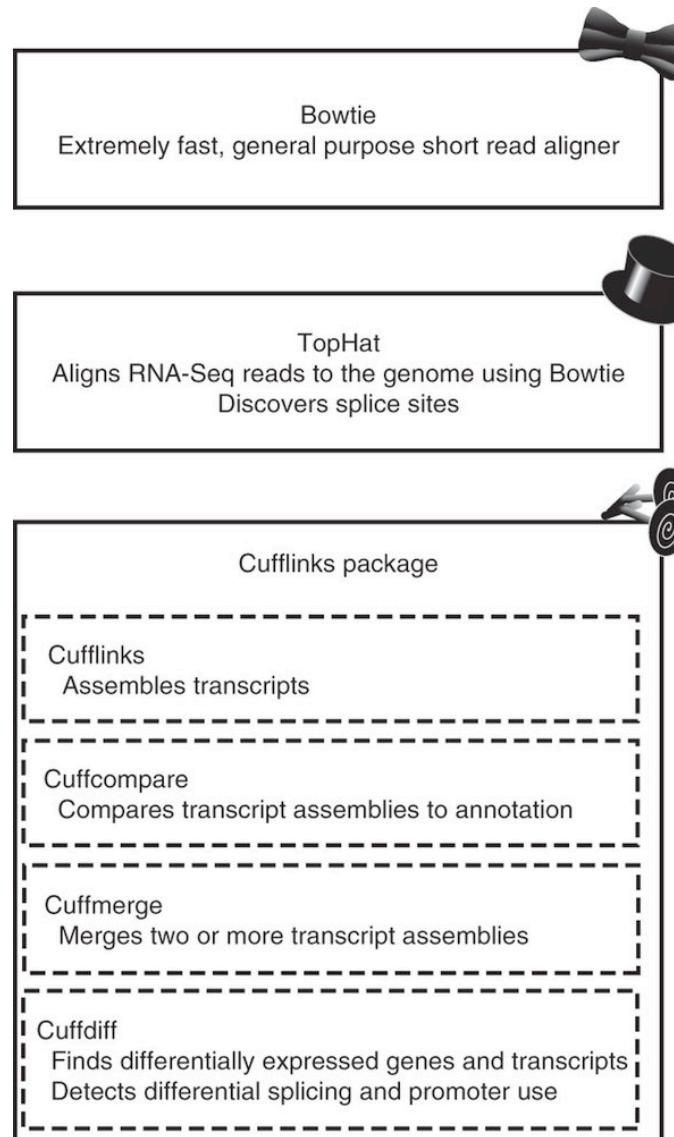
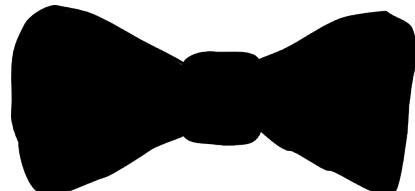
## Genome mapping

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

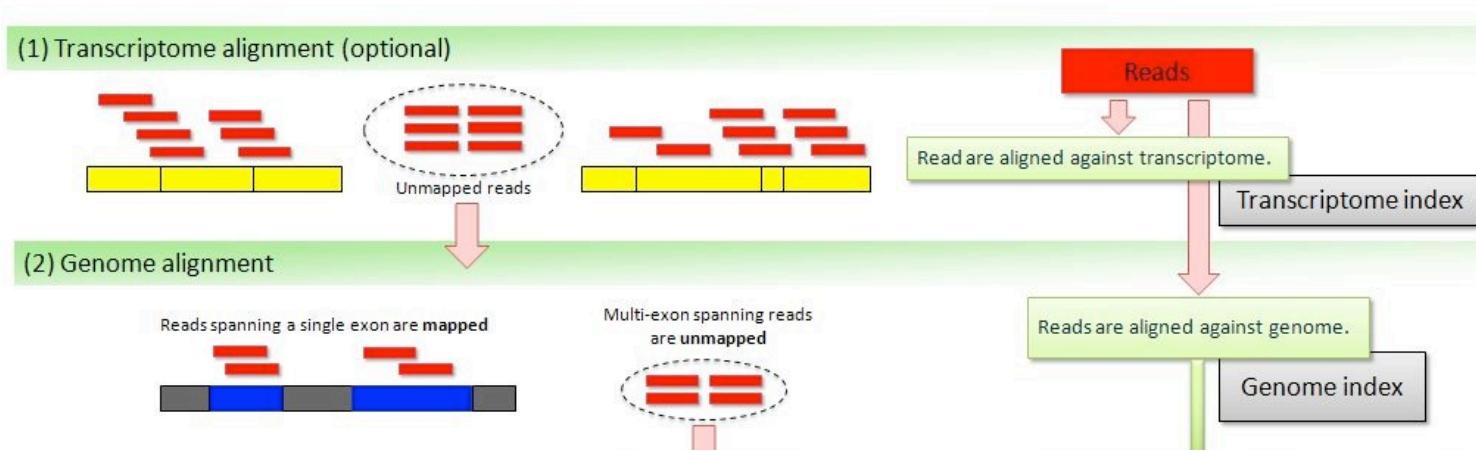
## Transcriptome mapping

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

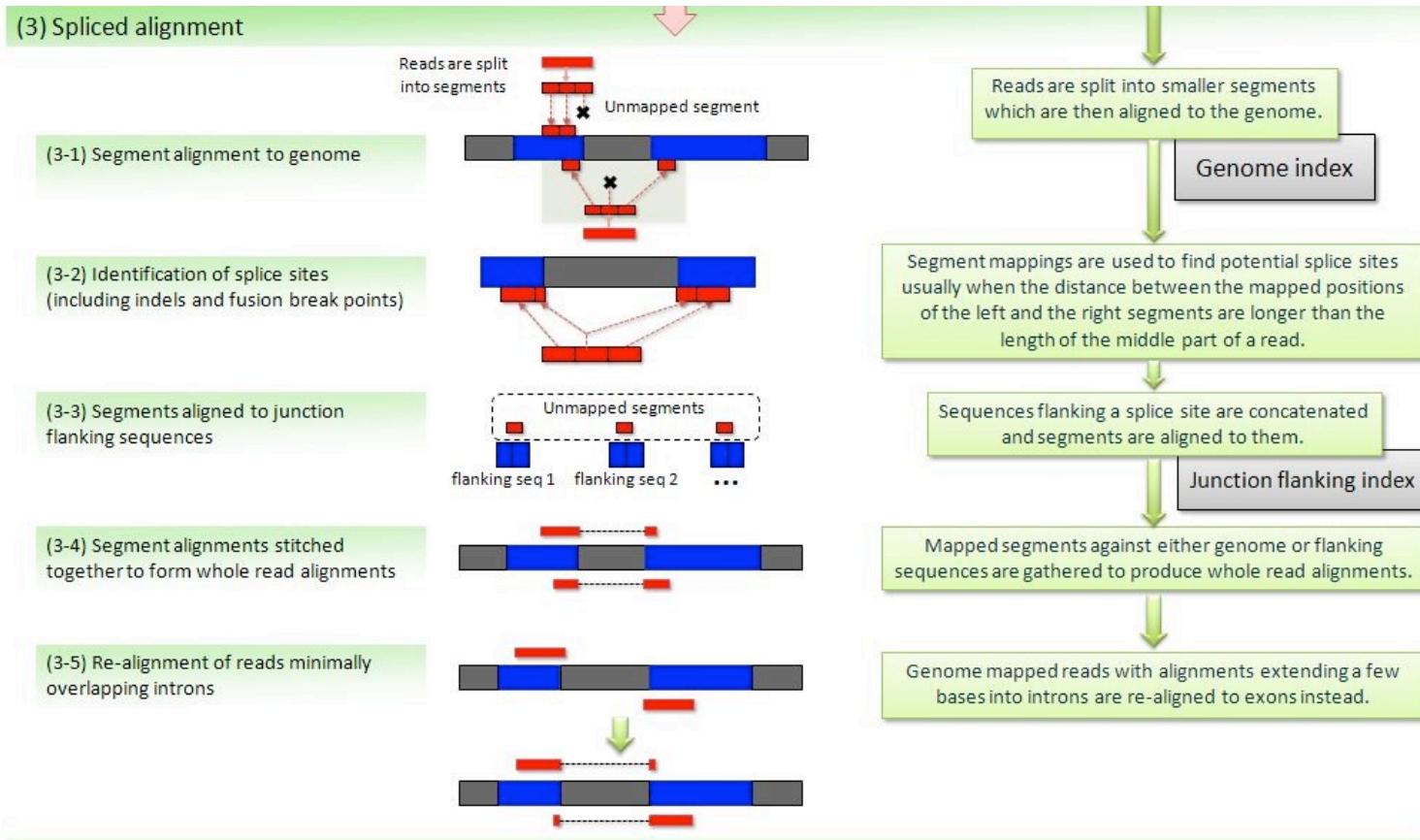
# A smart suit(e)



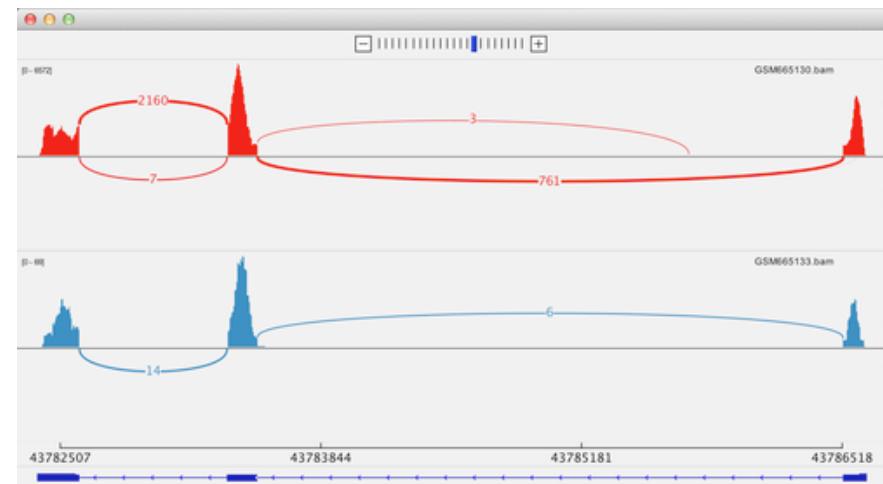
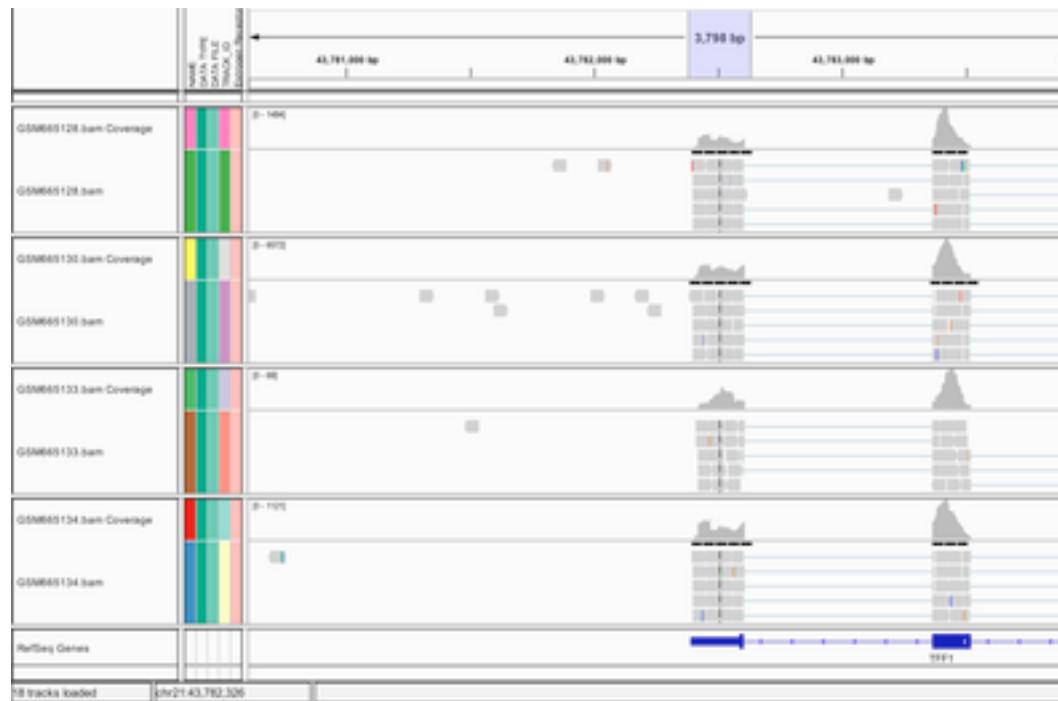
# Tophat/Bowtie



# Tophat/Bowtie



# How do we look?



# Duplicates & RNA-seq

Intrinsically lower complexity

Highly expressed genes

Platform/pipeline

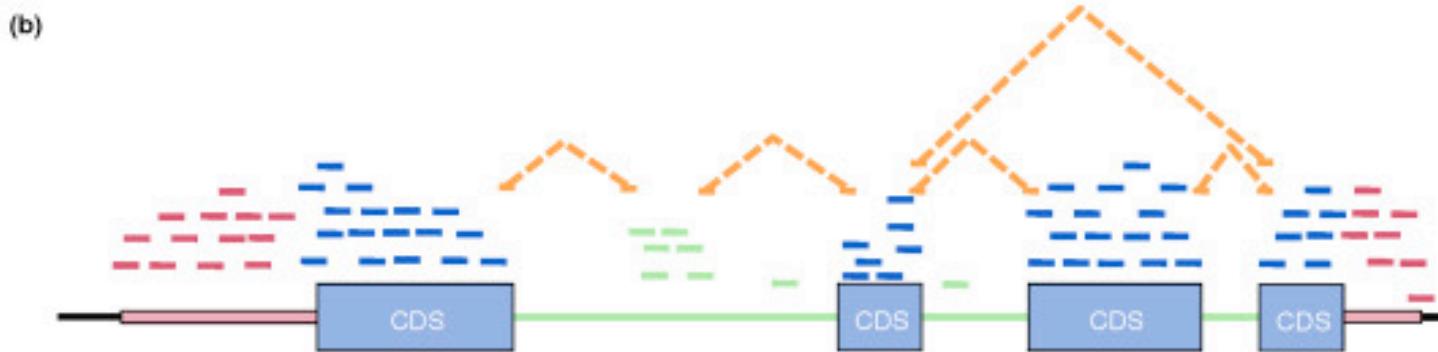
Variant calling vs  
DE analysis

Platform/pipeline

Single-end vs  
paired-end



# Counting



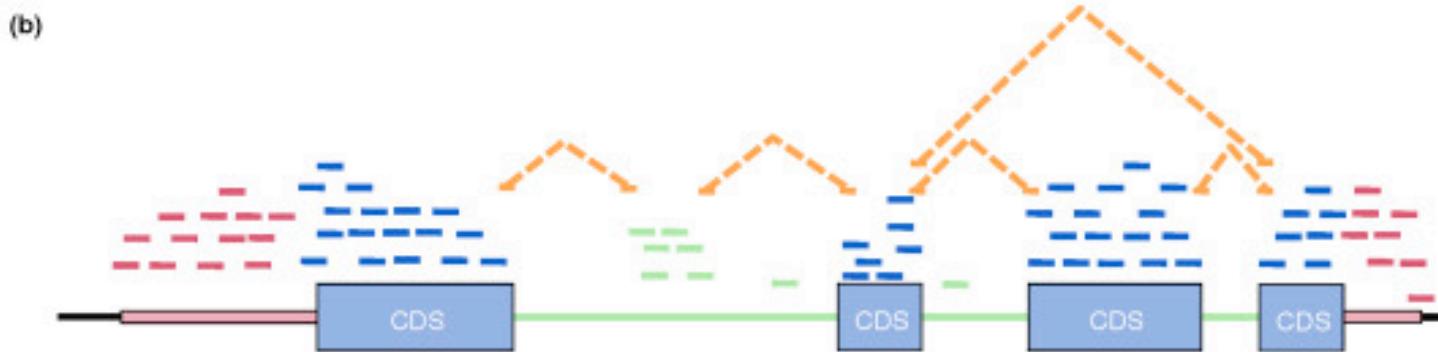
## Genome-based features

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

## Transcript-based features

- Transcript assembly
- Novel structures
- Isoform multireads

# Counting



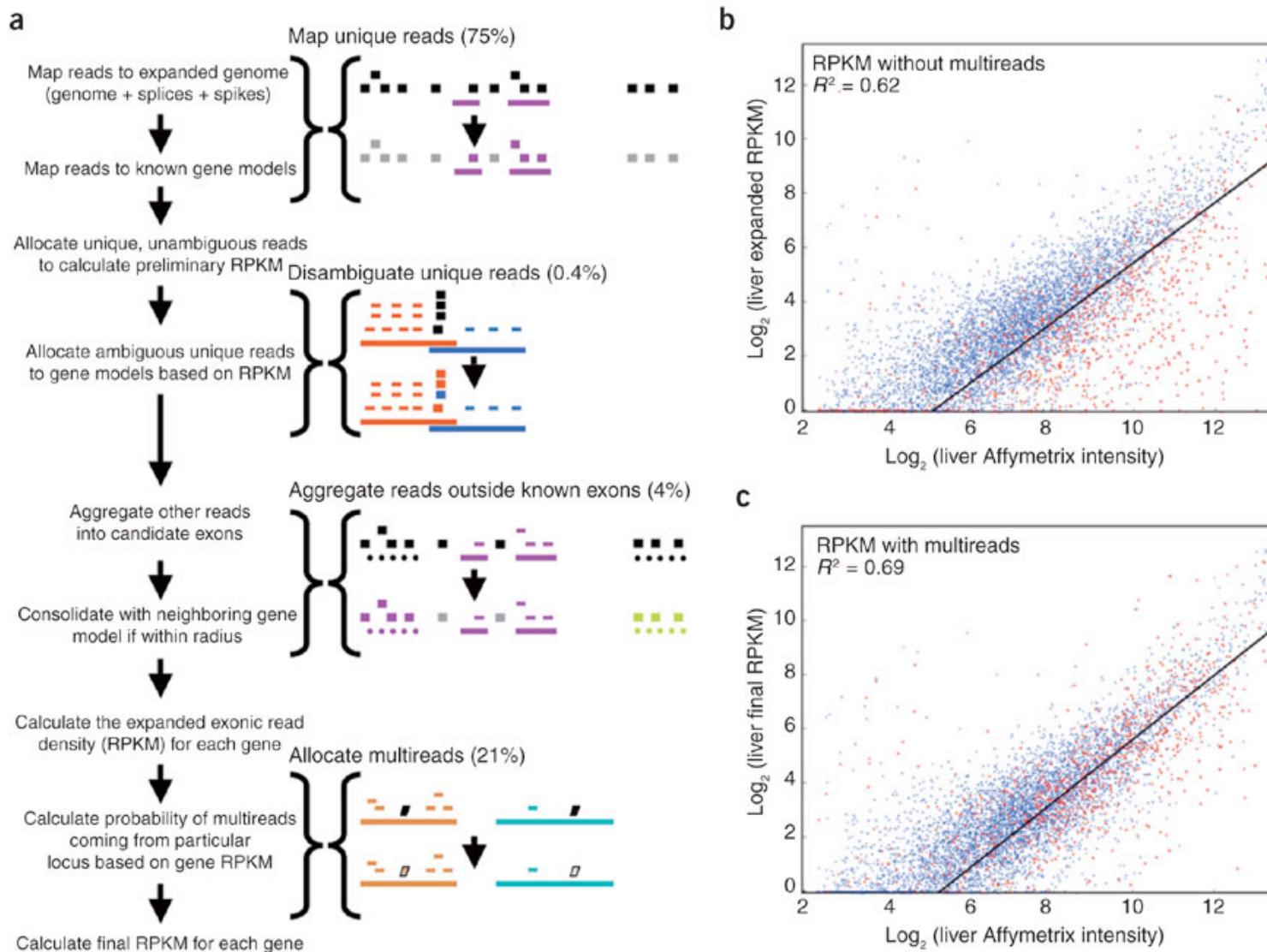
## Genome-based features

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

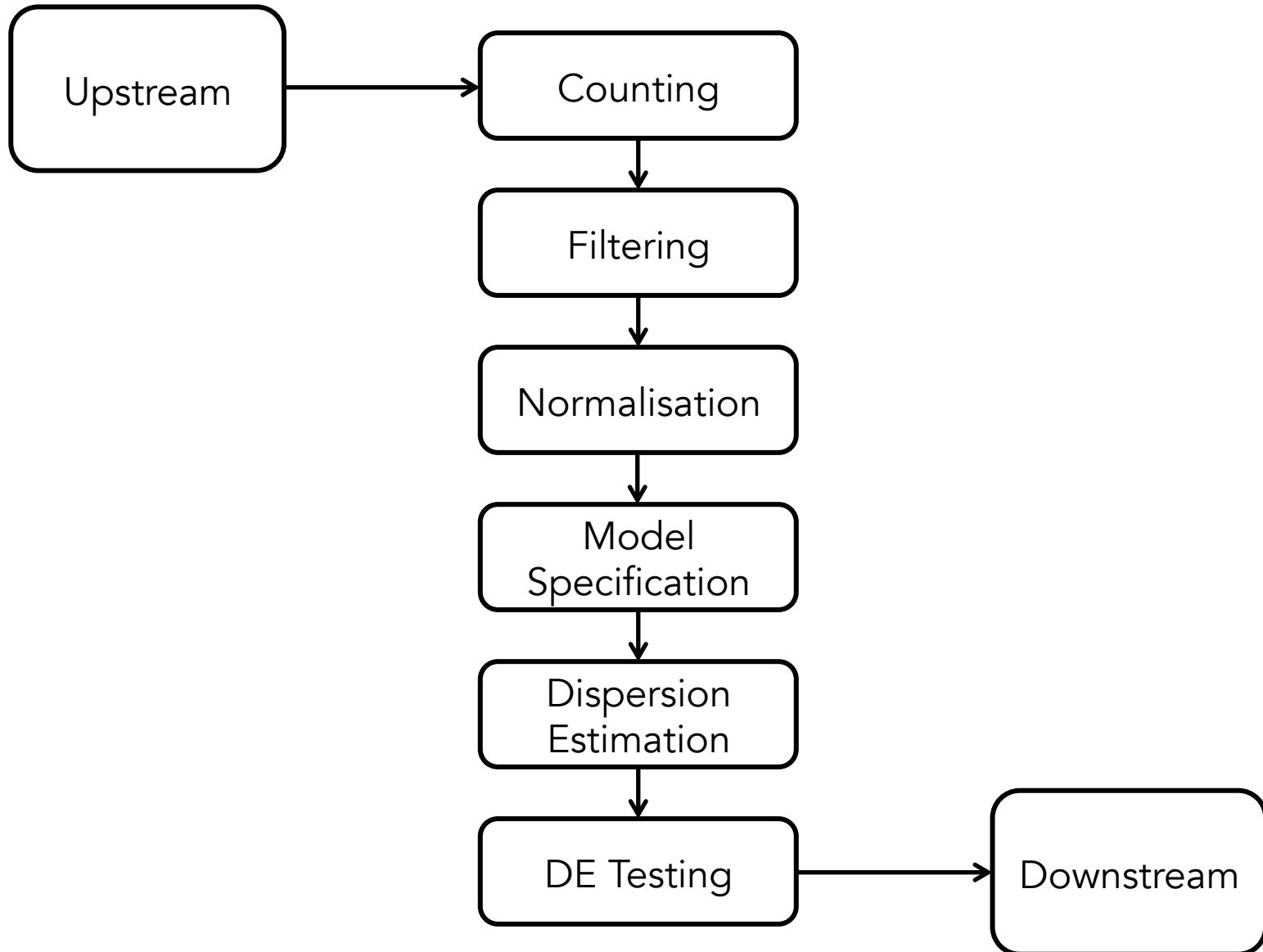
## Transcript-based features

- Transcript assembly
- Novel structures
- Isoform multireads

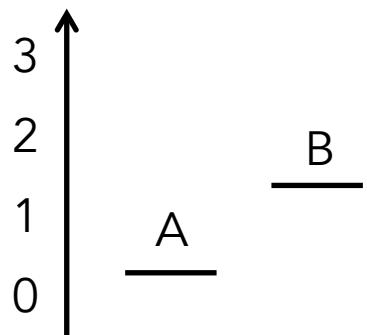
# Counting



# Differential expression



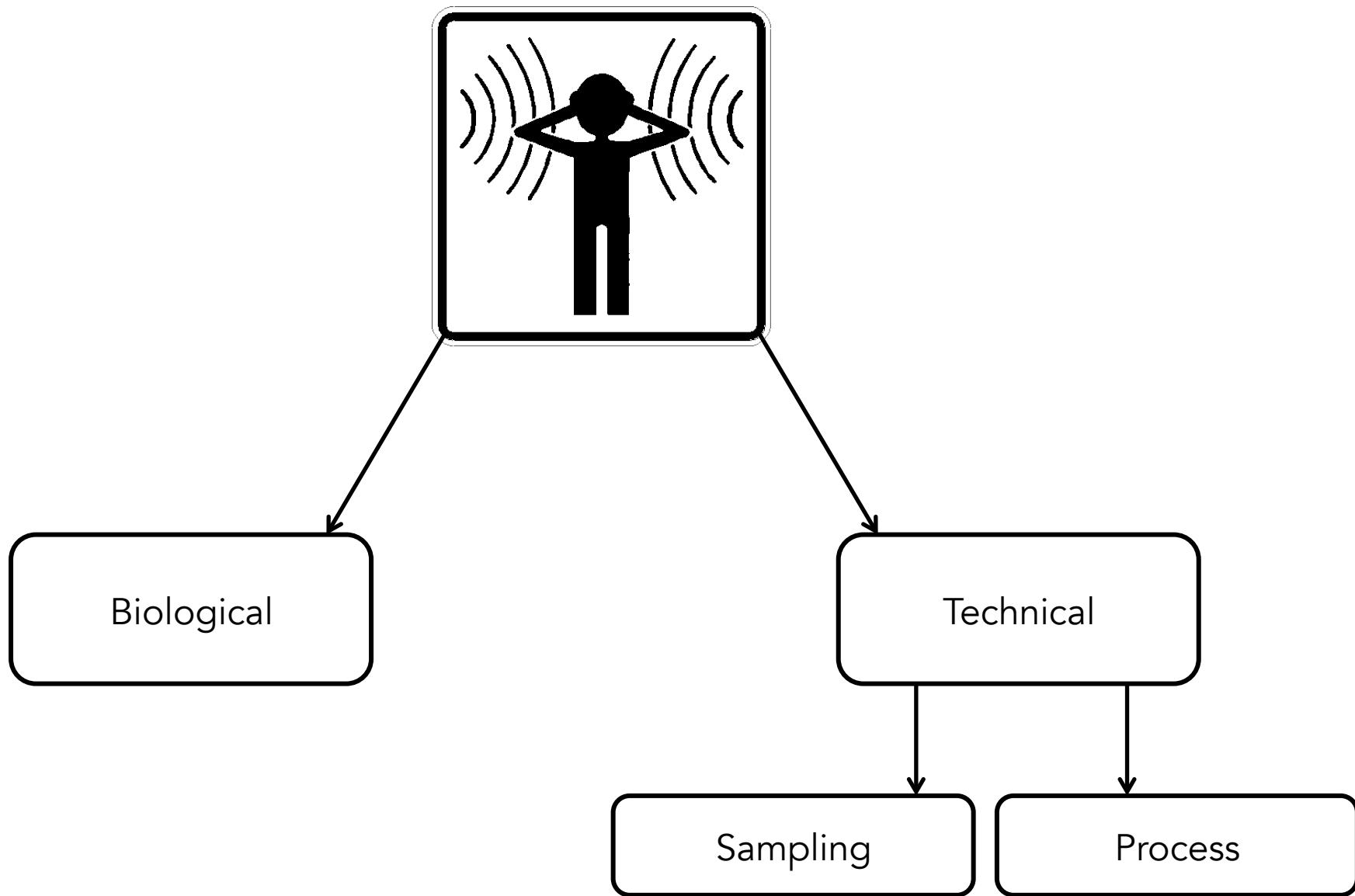
# Building models



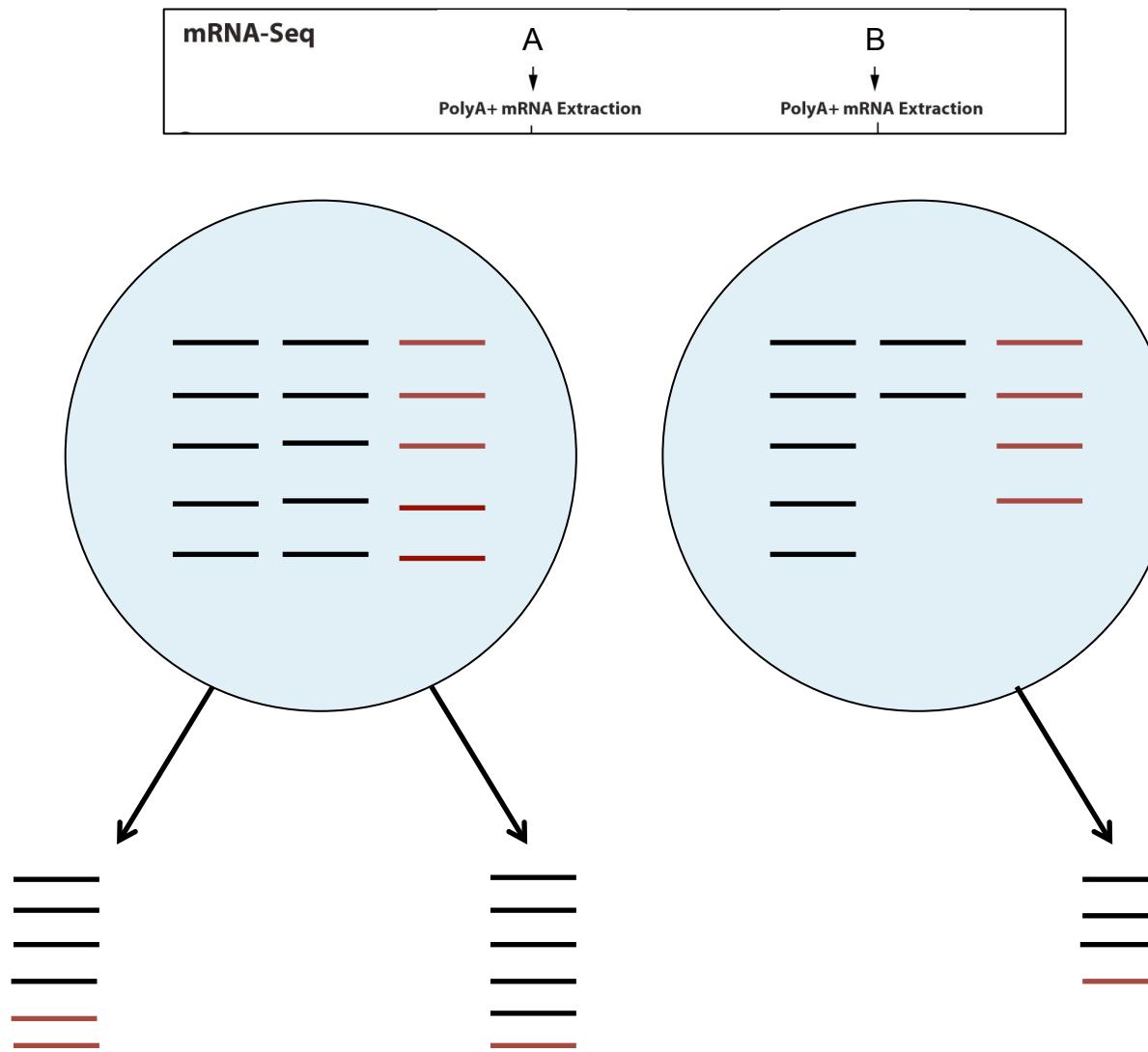
Need to understand the data so that we can build a statistical model that describes it.

- How is the data generated?
- What are the sources of variability we are interested in?
- What are the sources of variability we aren't interested in?

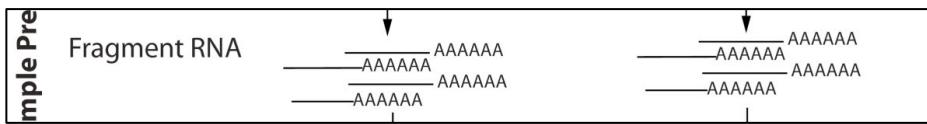
# Building models – noise



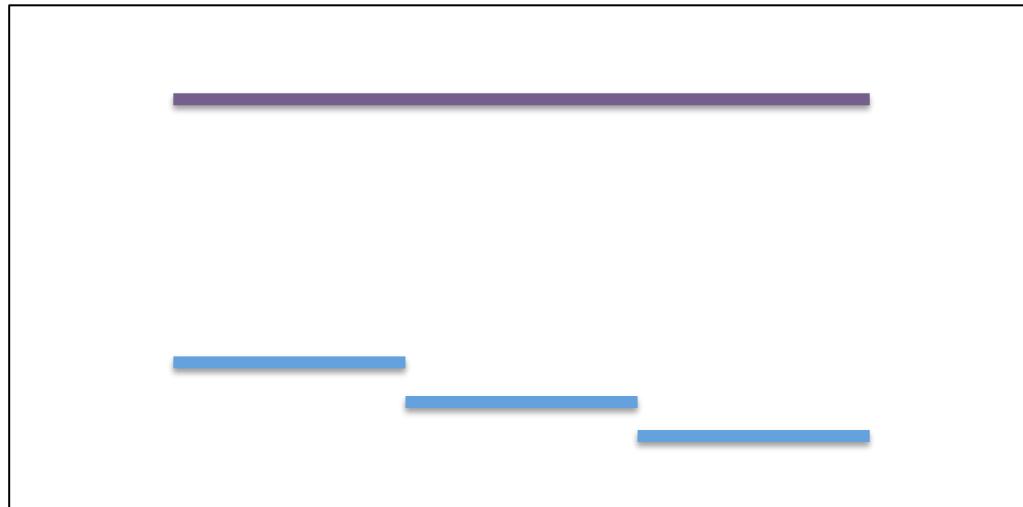
# Library Prep iii



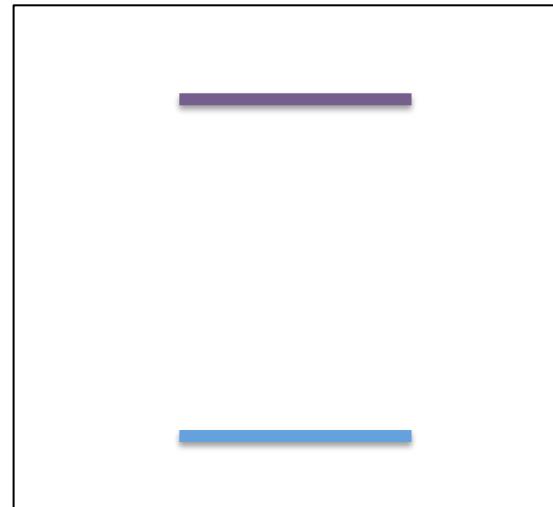
# Library Prep iii



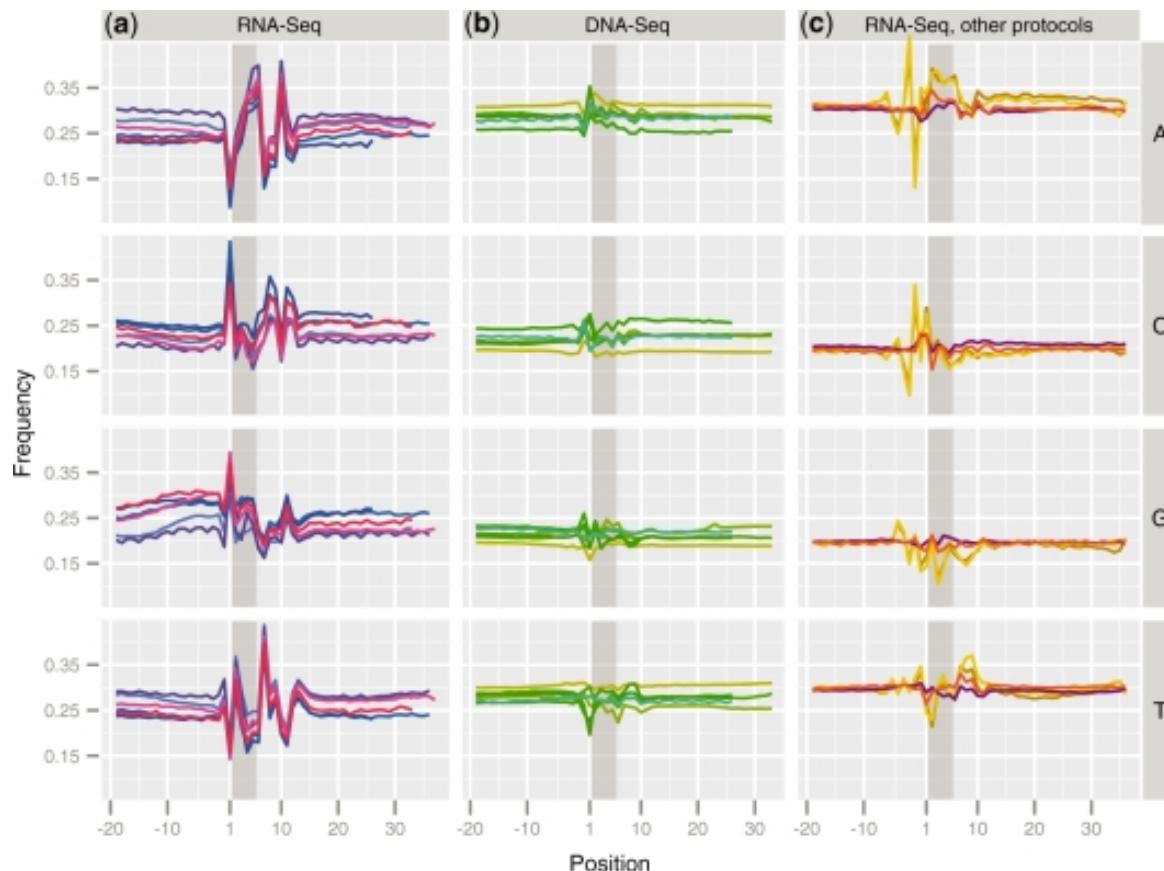
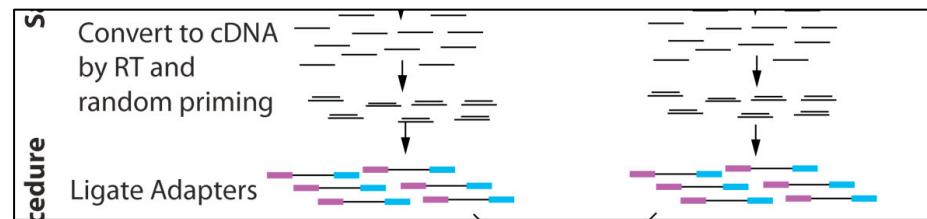
A



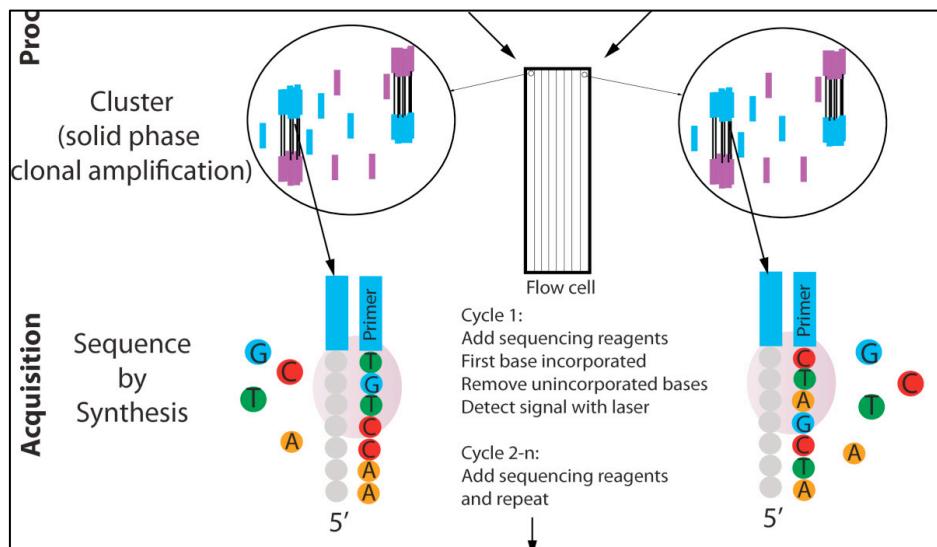
B



# Library Prep iv



# Library Prep v



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

# Counting & normalisation

- An estimate for the *relative* counts for each gene is obtained
- Assumed that this estimate is representative of the original population

## Library size

- Sequencing depth varies between samples

## Gene Properties

- GC content, length, sequence

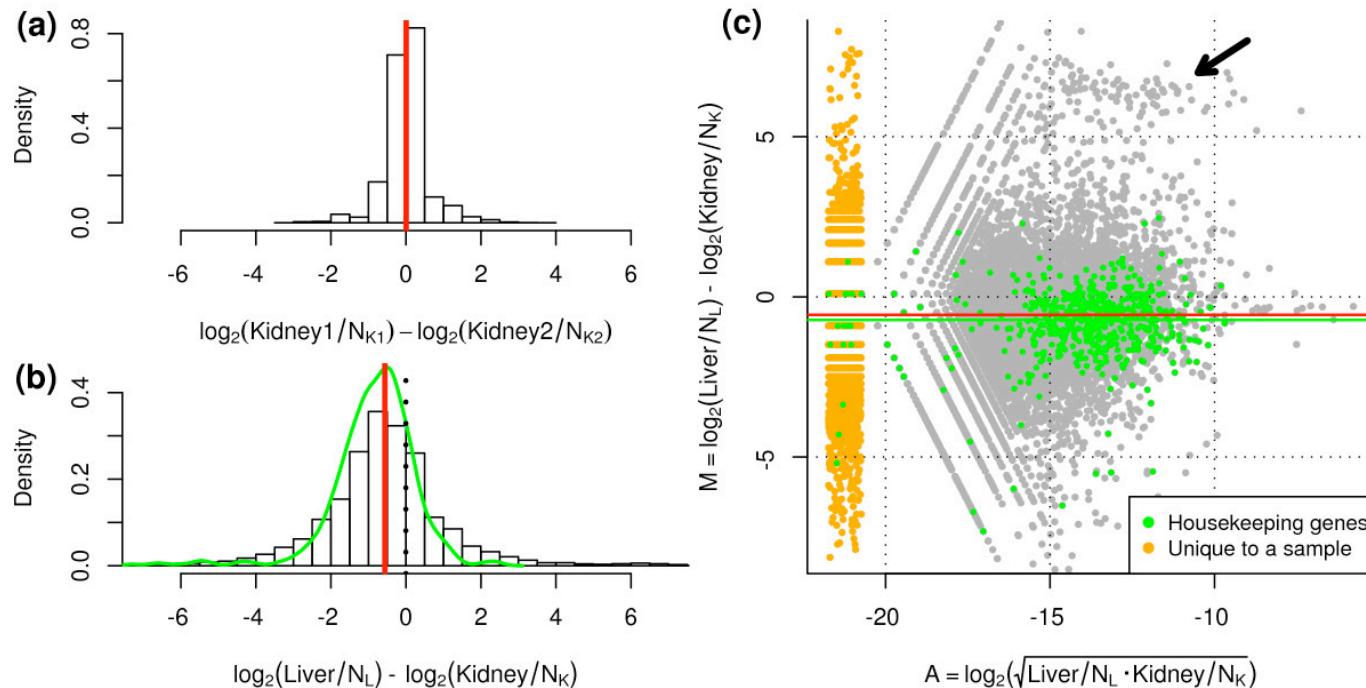
## Library composition

- Highly expressed genes overrepresented at cost of lowly expressed genes

# Normalisation i

## Total Count

- Normalise each sample by total number of reads sequenced.
- Can also use another statistic similar to total count; eg. median, upper quartile

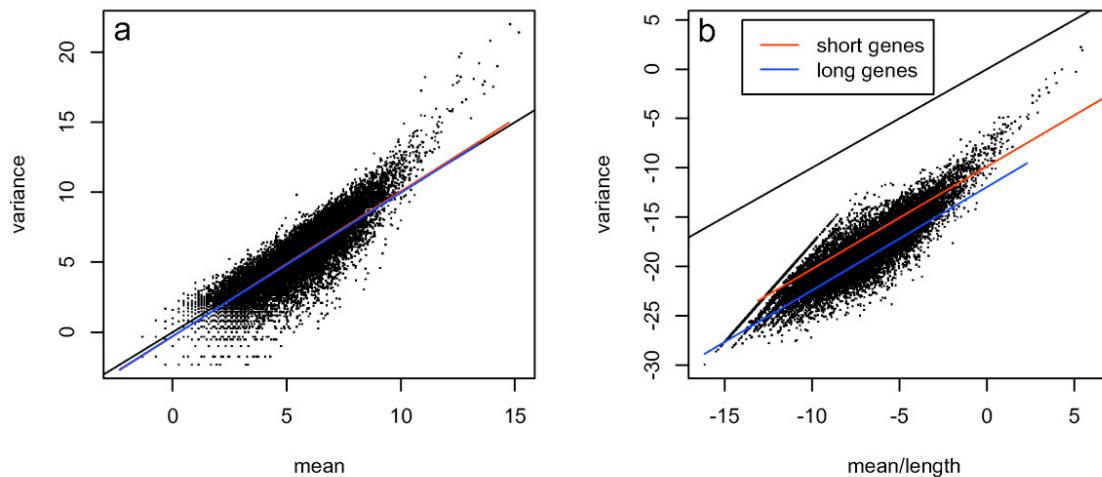


# Normalisation ii

## RPKM

- Reads per kilobase per million =

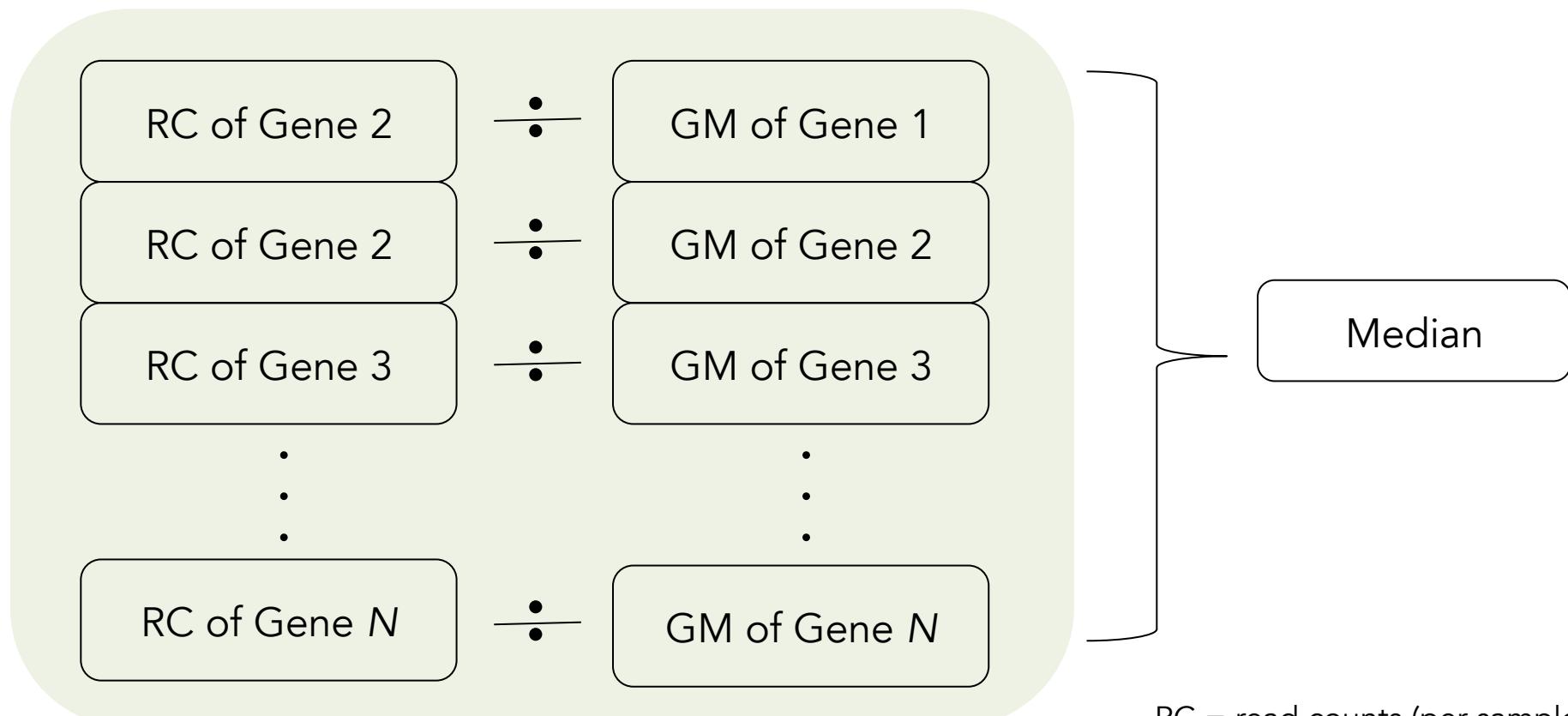
$$\frac{\text{reads for gene A}}{\text{length of gene A} \times \text{Total number of reads}}$$



# Normalisation iii

## Geometric scaling factor

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



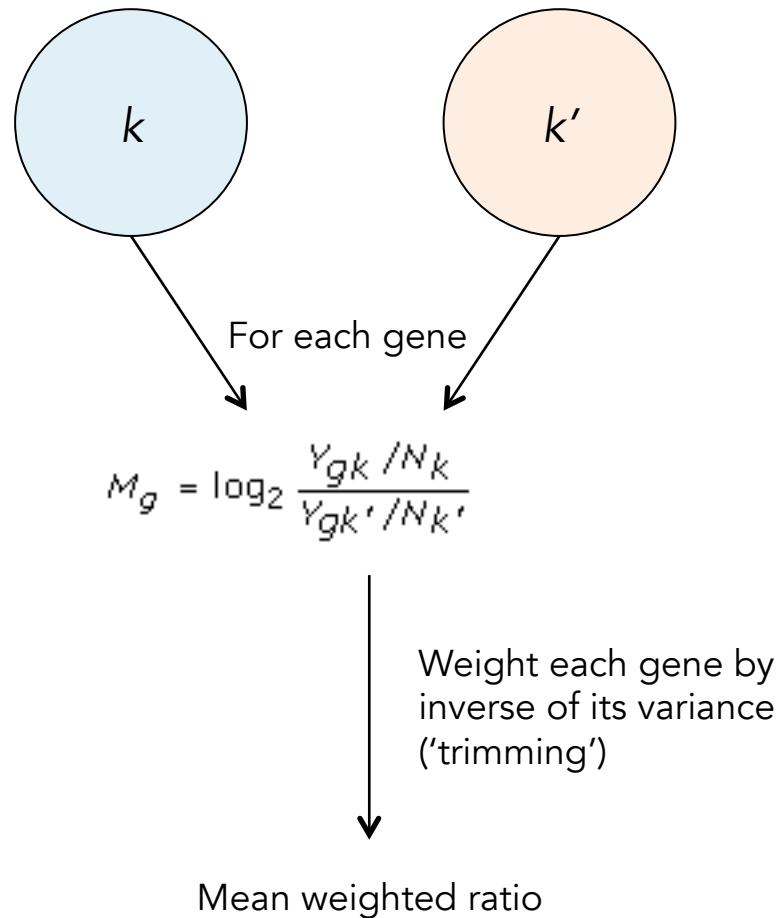
RC = read counts (per sample)

GM =geometric mean (all samples)

# Normalisation iv

## Trimmed mean of M

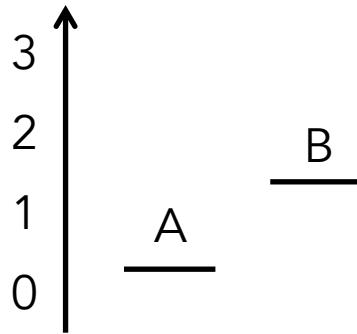
- Implemented in edgeR
- Assumes most genes are not differentially expressed



$g$  = each gene

# Differential expression

- Simple

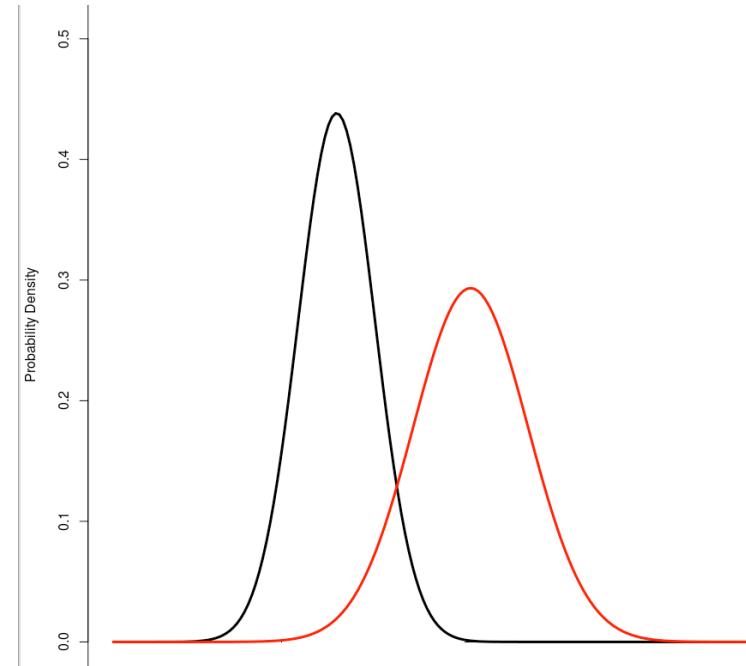
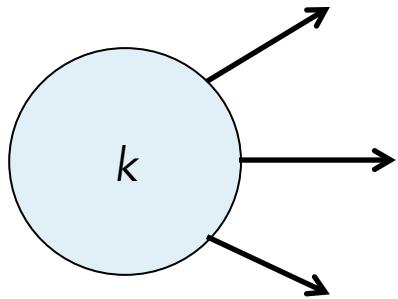


## All we need

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

# Modelling – old trends

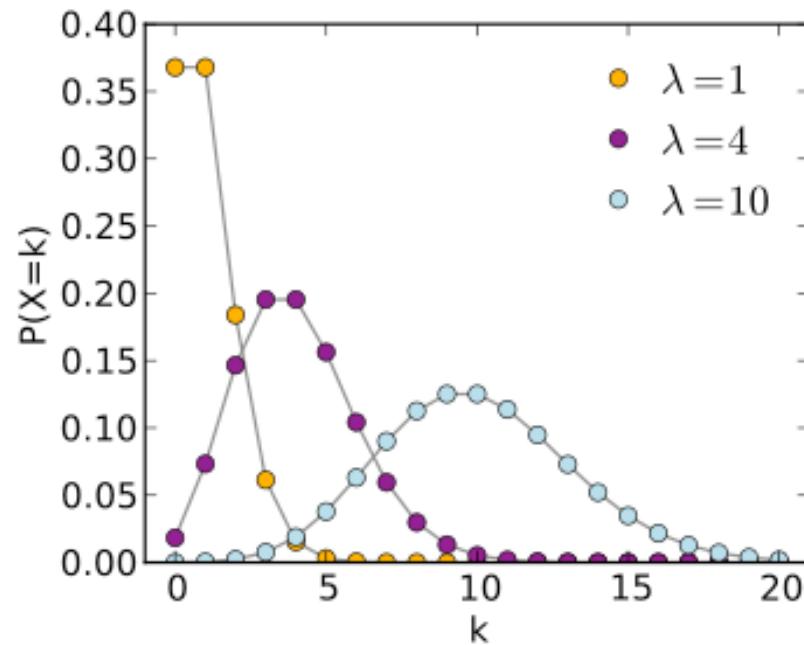
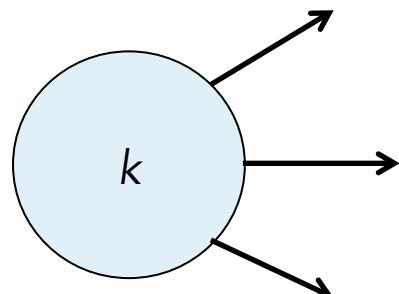
- Technical replicates introduce some variance



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

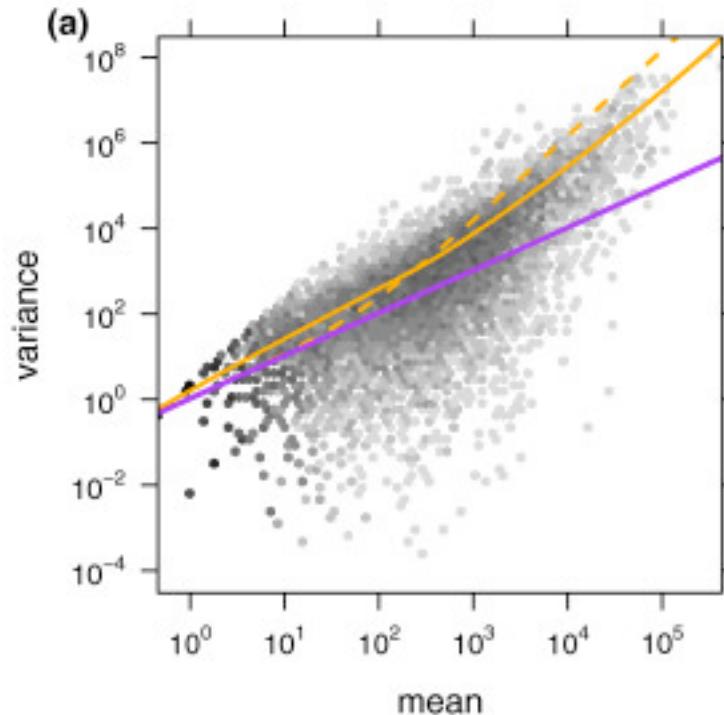
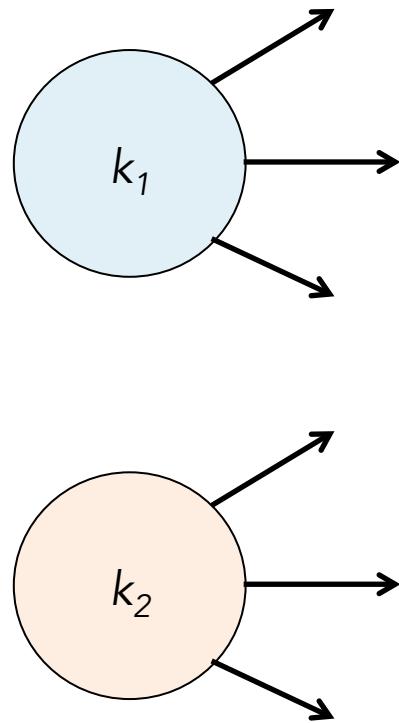
# Modelling – in fashion

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



# Modelling – in fashion

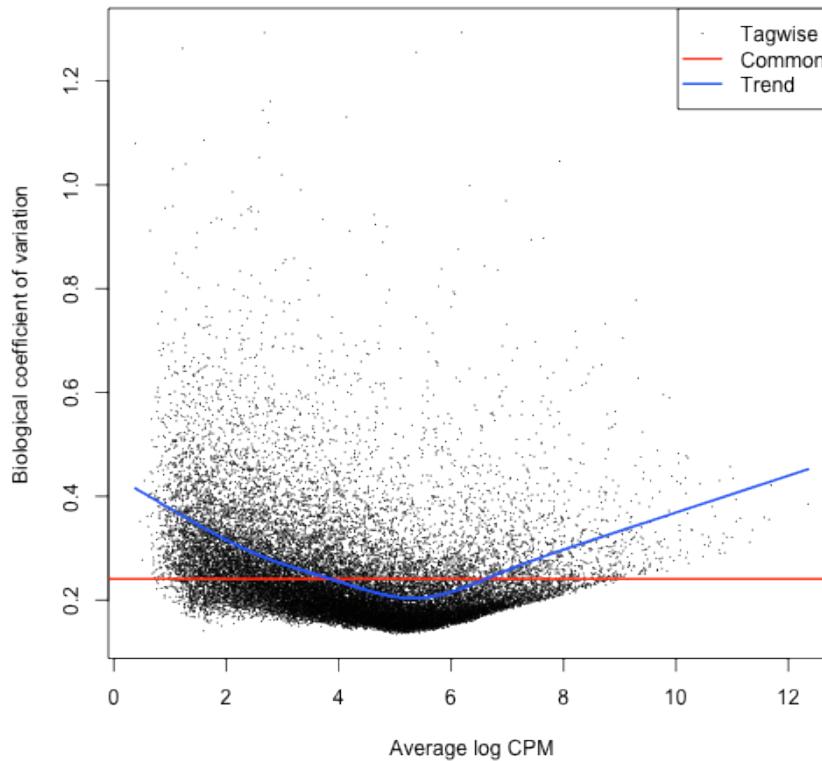
- Biology is never that simple...



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

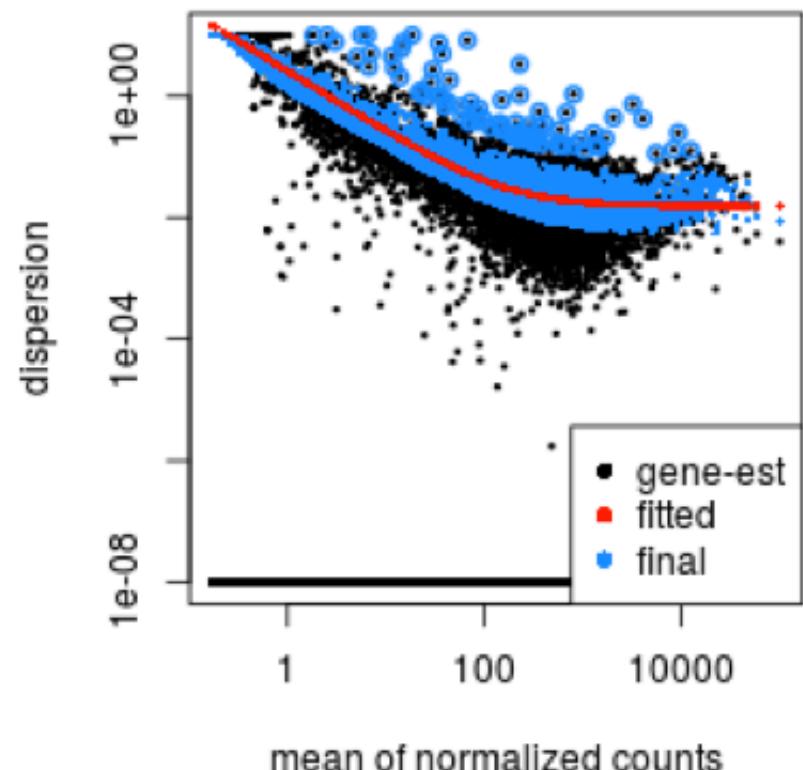
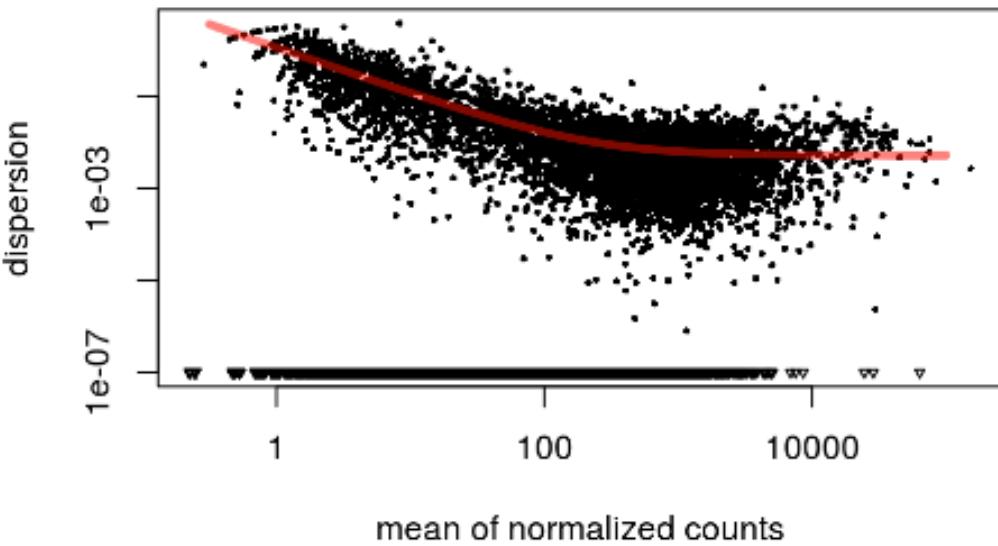
# Estimating dispersions

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

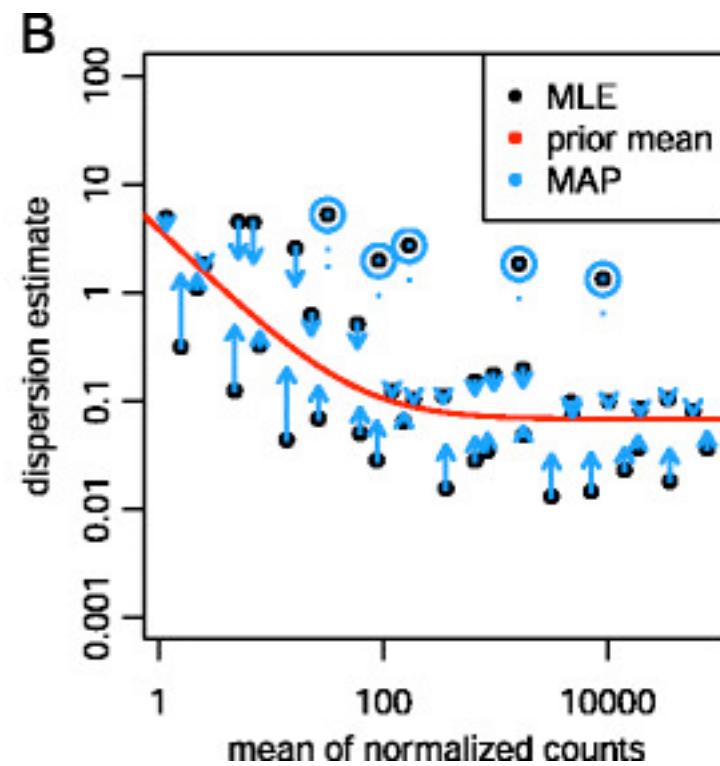
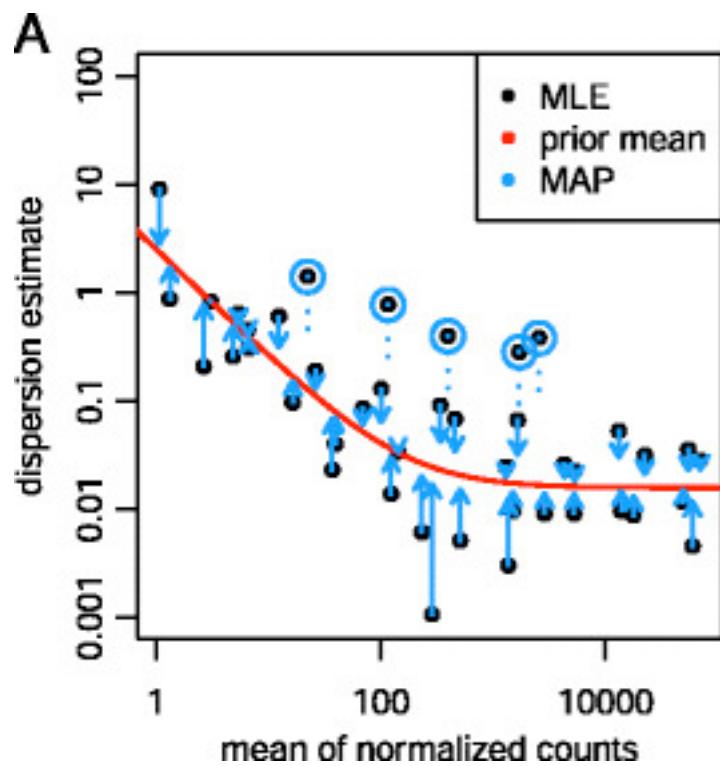


# Estimating dispersions

- An empirical Bayes method is used by edgeR and DESeq2 to 'shrink' per-gene estimates towards a trended estimate

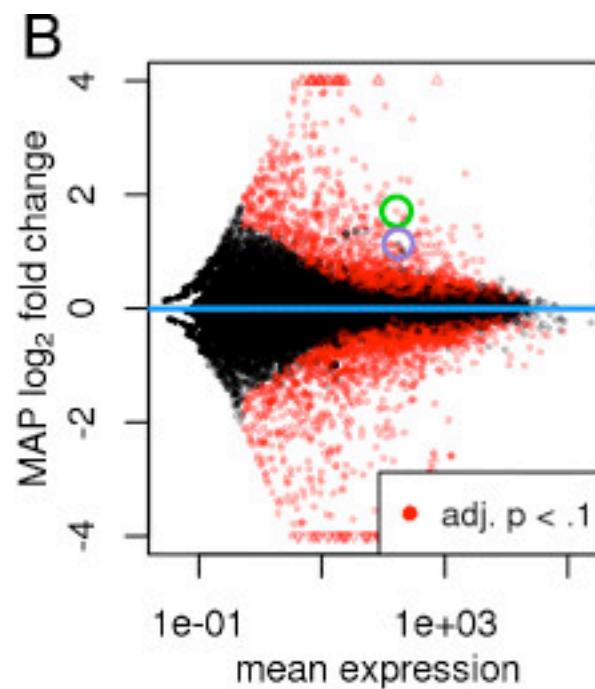
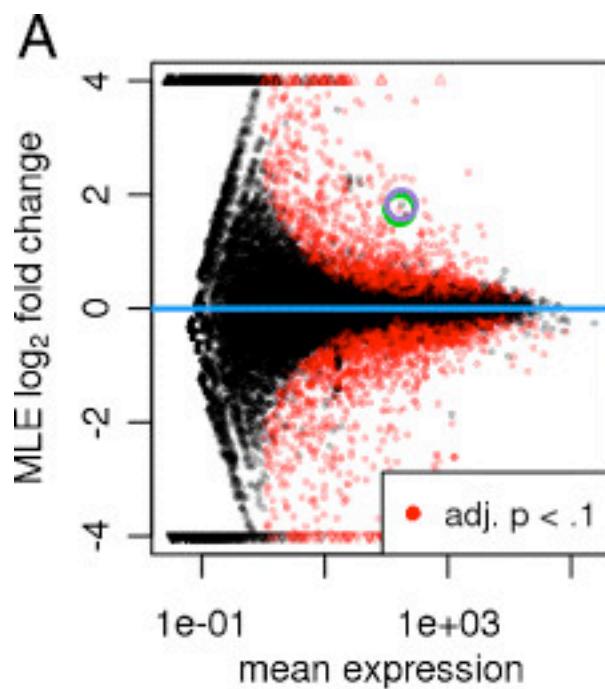


# Shrinkage

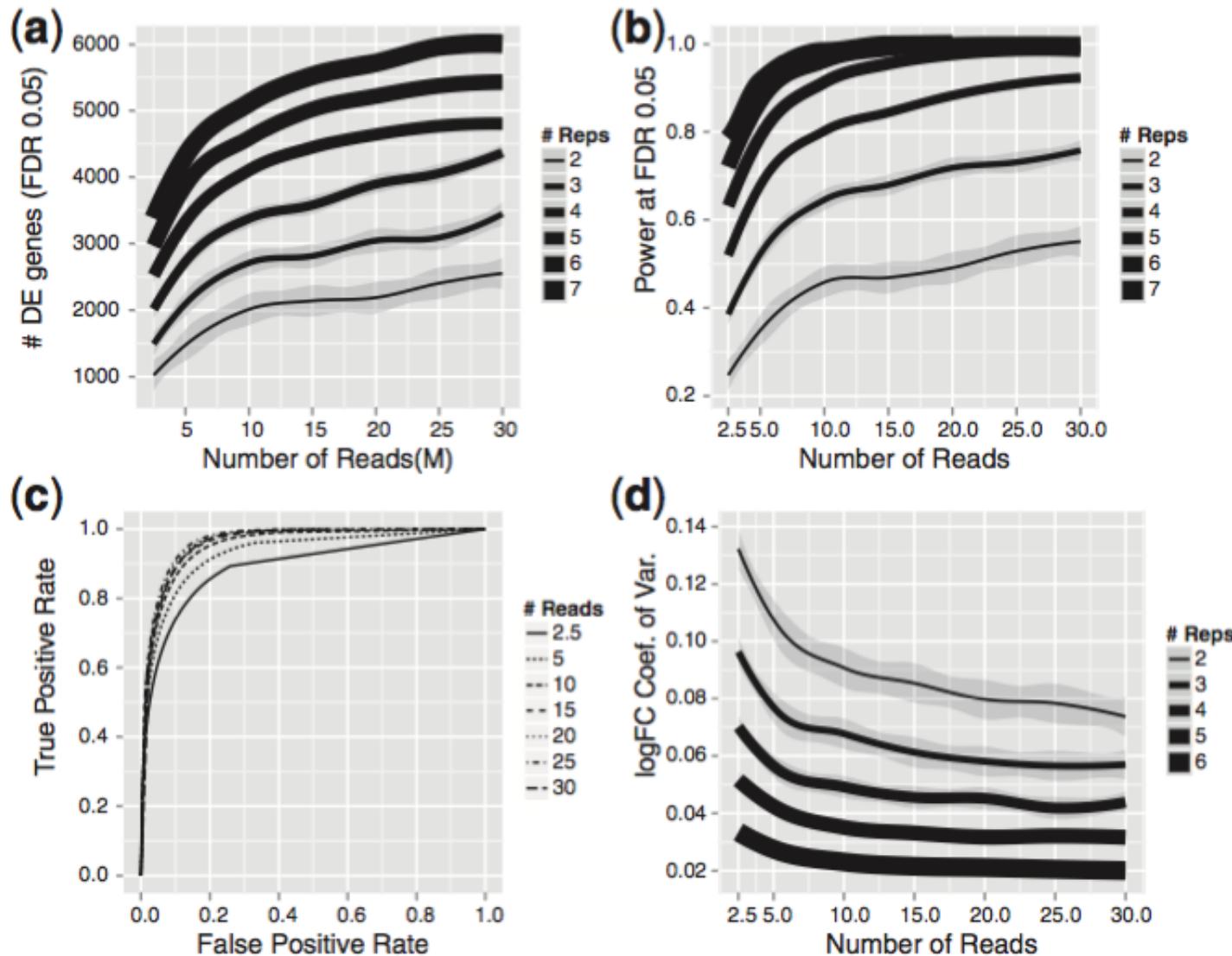


# DESeq2

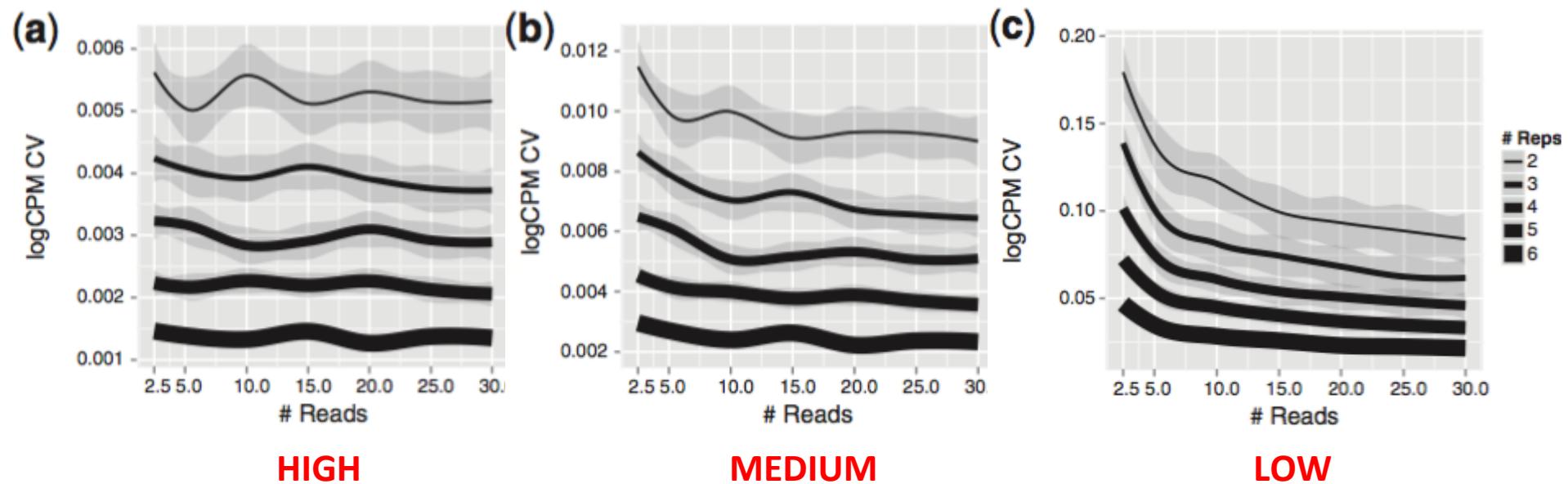
- Log fold change shrinkage in DESeq2



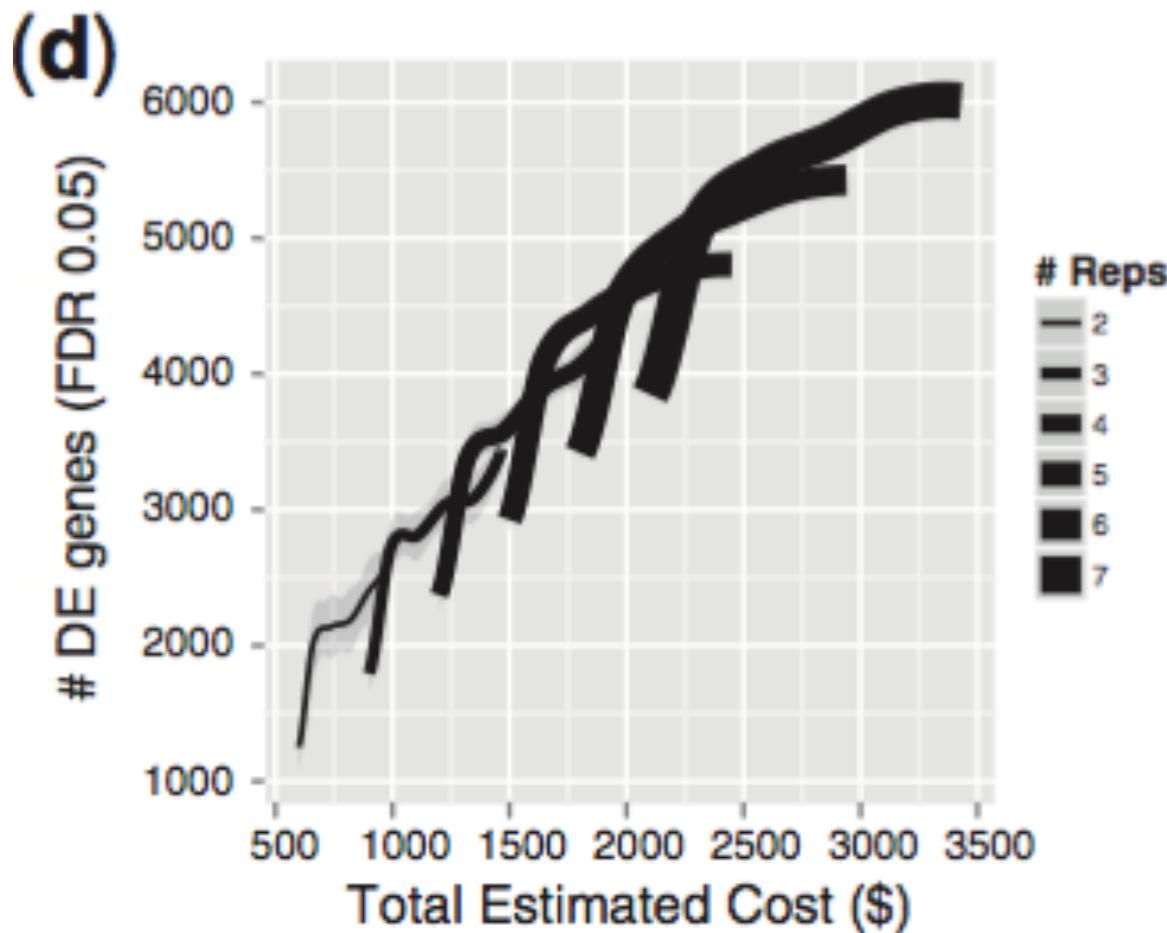
# On replicates...



# On replicates...

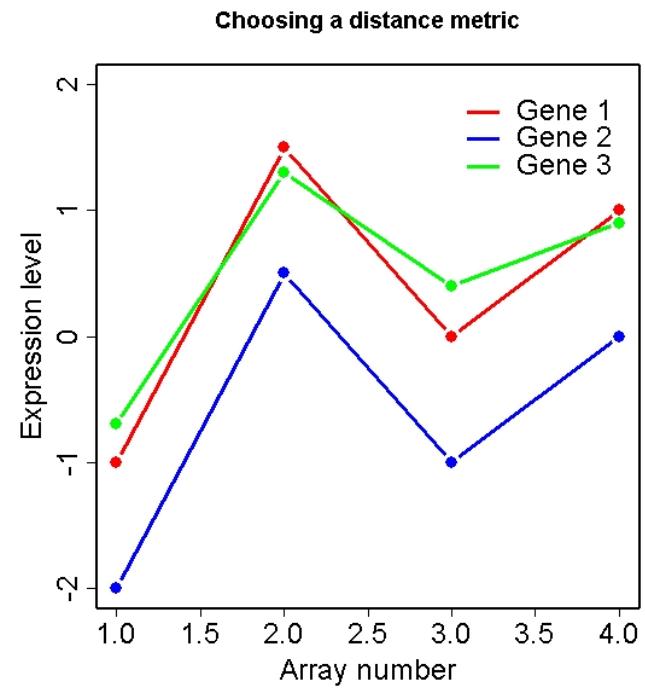
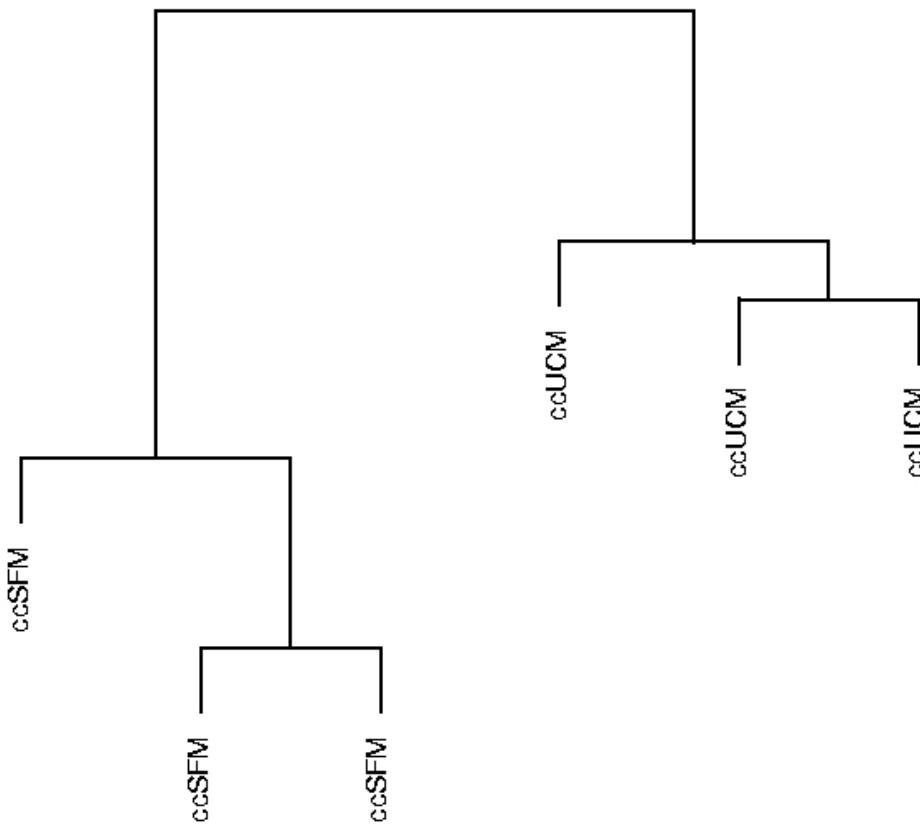


# On replicates...



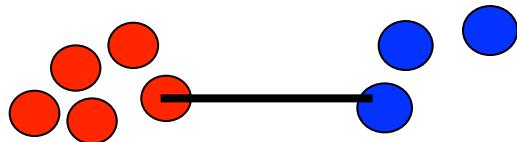
# What next?

- Hierarchical clustering = define metric & look for similarities

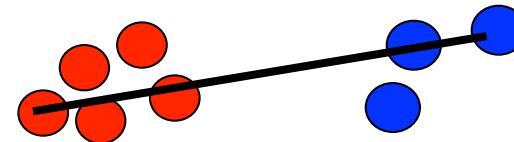


# What next?

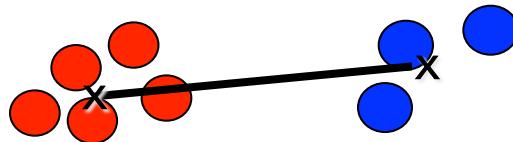
- Merging clusters according to a metric



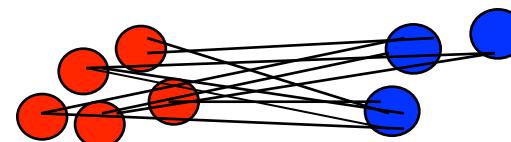
Single  
(min. of pairwise distances)



Complete  
(max. of pairwise distances)

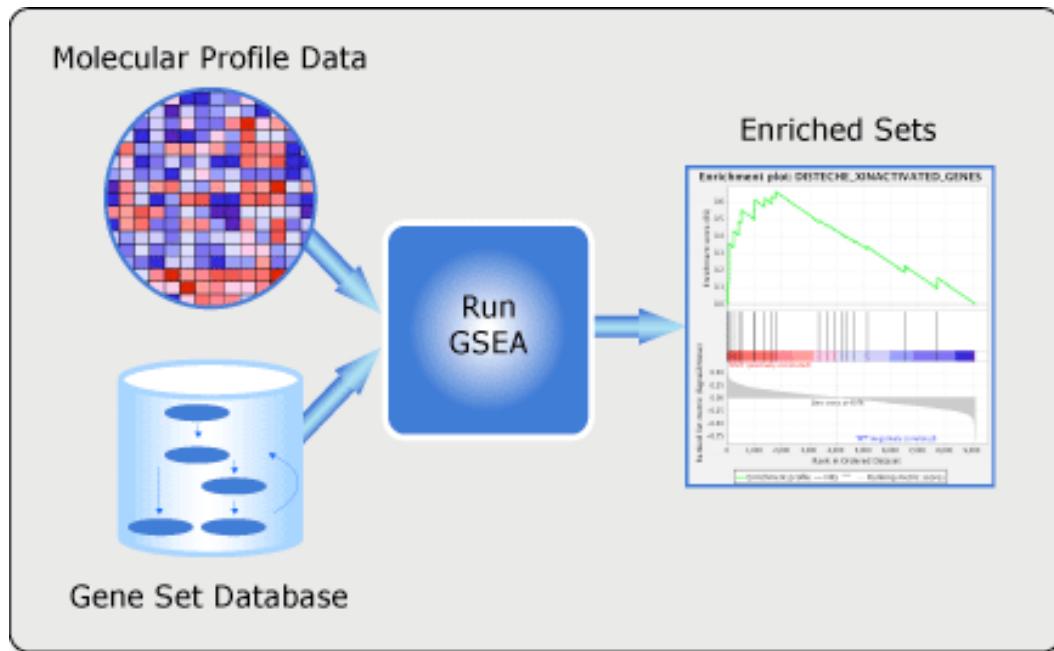


Distance between centroids



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) [?](#)