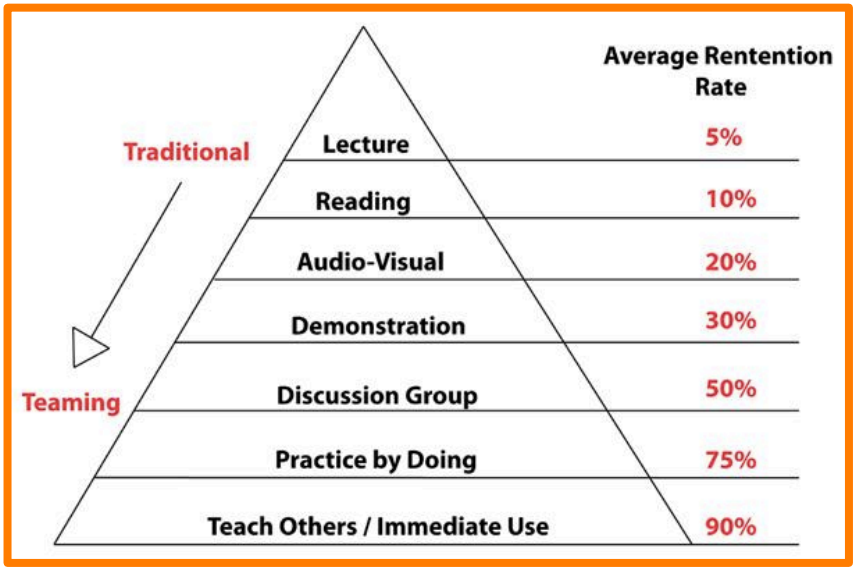
University of Zurich[UZH]

Institute of Molecular Life Sciences

Computational Genomics – A Hands-On Course on Data Analysis

RNA-seq – 22 Oct 2015

1. WHAT
2. HOW
3. WHY
4. DO



Mark D. Robinson, Statistical Genomics, IMLS

University of Zurich[UZH]

Institute of Molecular Life Sciences

ICEBREAKER

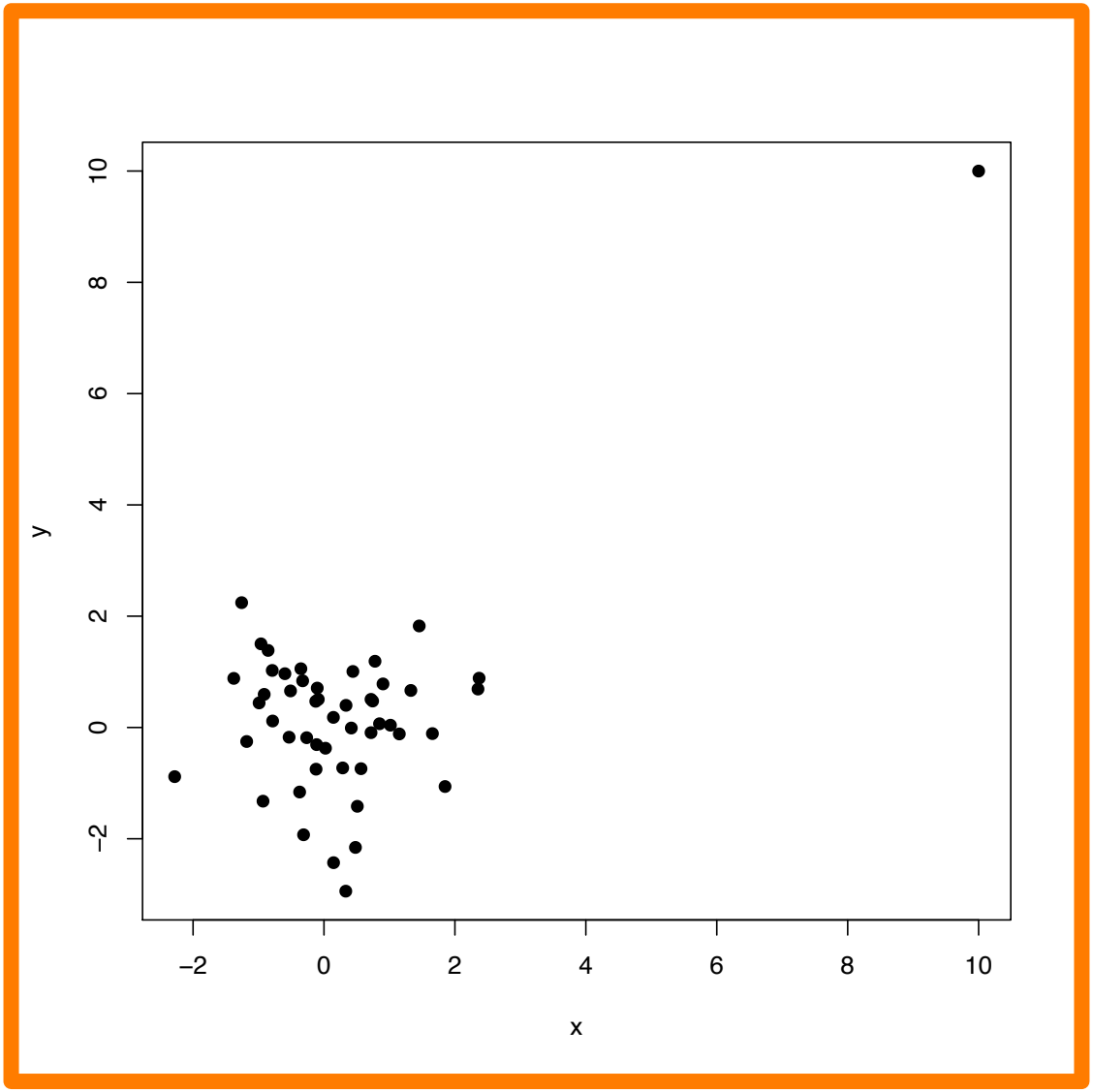
University of Zurich[UZH]

Institute of Molecular Life Sciences

movo.ch

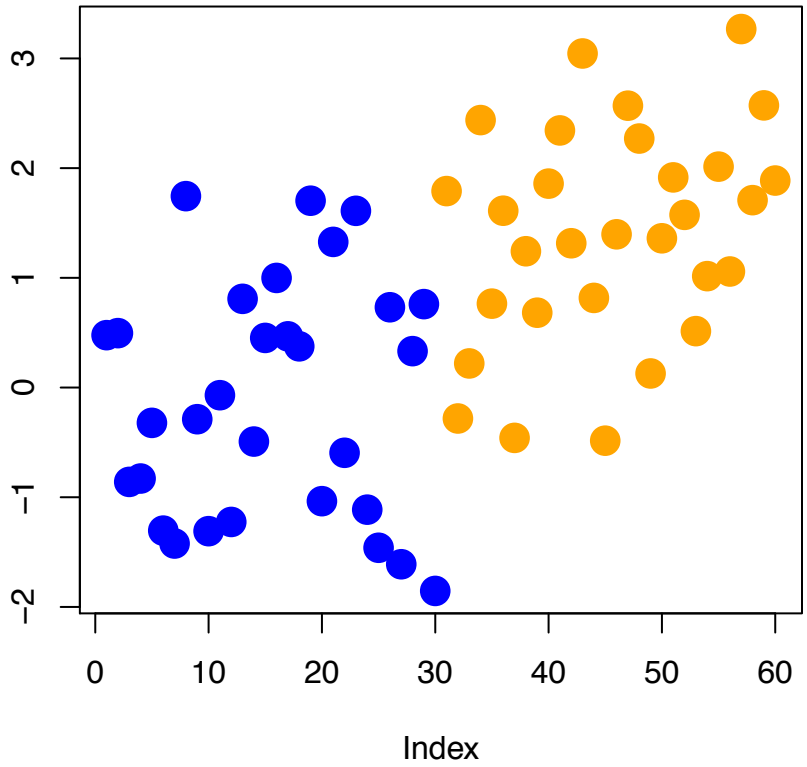
VI XA CO TE

In your view, what best describes the associations shown in the plot of 'x' and 'y' ?
(Pearson = linear correlation; Spearman = rank correlation)

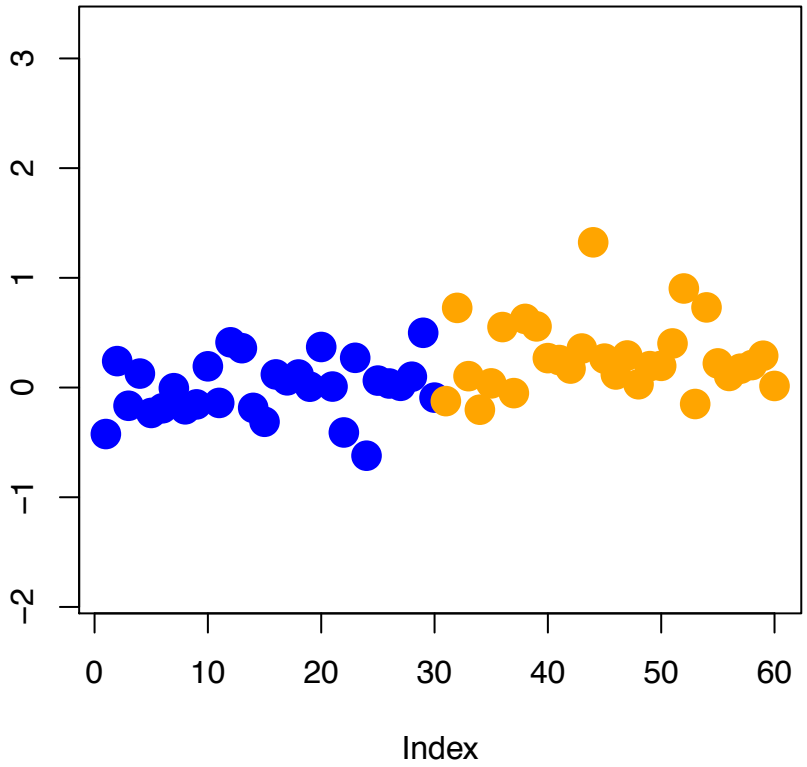


Which plot highlights more statistical evidence for a change in the population means (between orange and blue)?

A

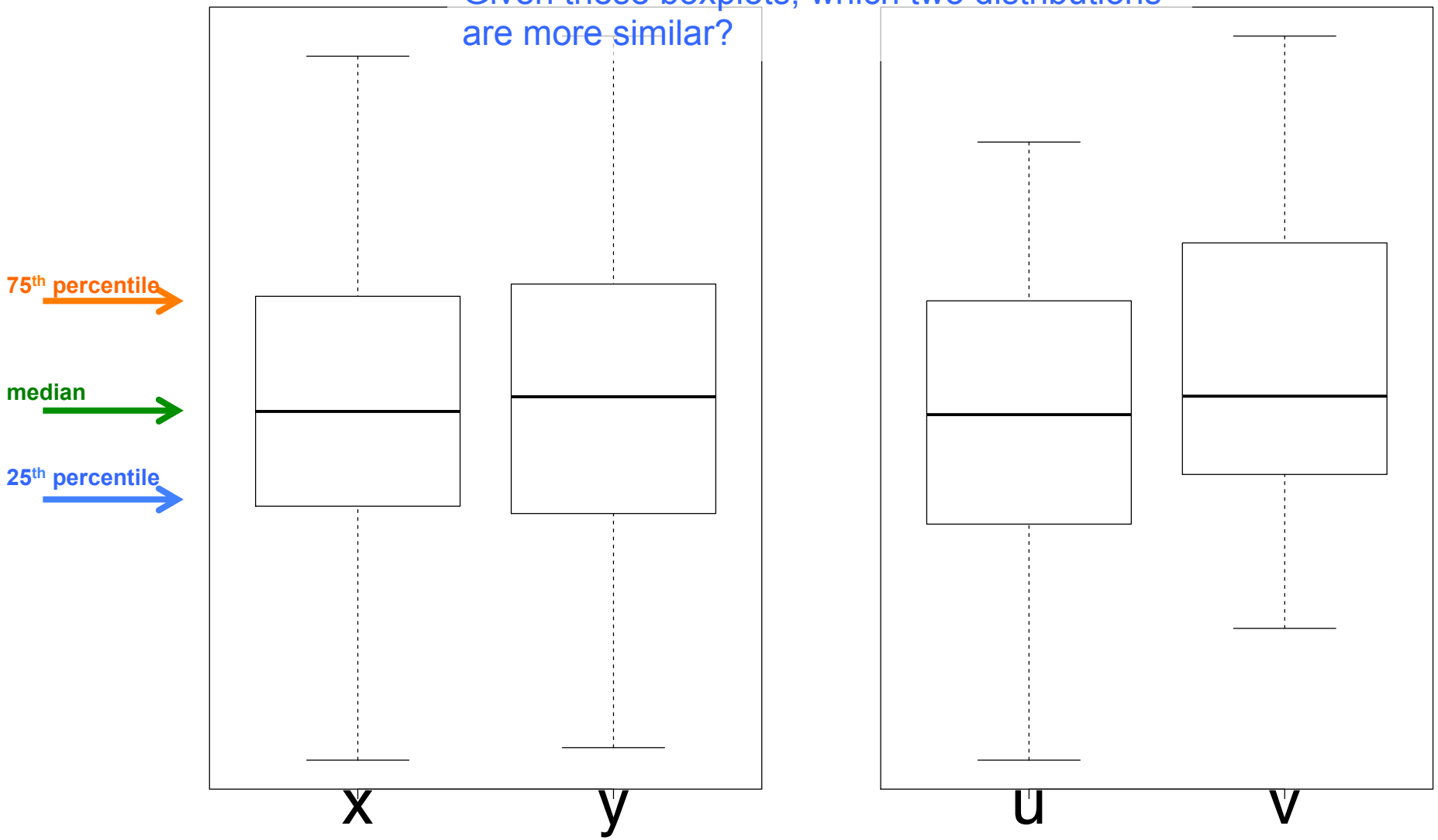


B



$$X = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Given these boxplots, which two distributions are more similar?



1 $$\frac{(\overline{x}_1 - \overline{x}_2) - d_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

2 $$\sum^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

3 $$\frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

http://en.wikipedia.org/wiki/Statistical_hypothesis_testing

University of Zurich[UZH]

Institute of Molecular Life Sciences

Etherpad

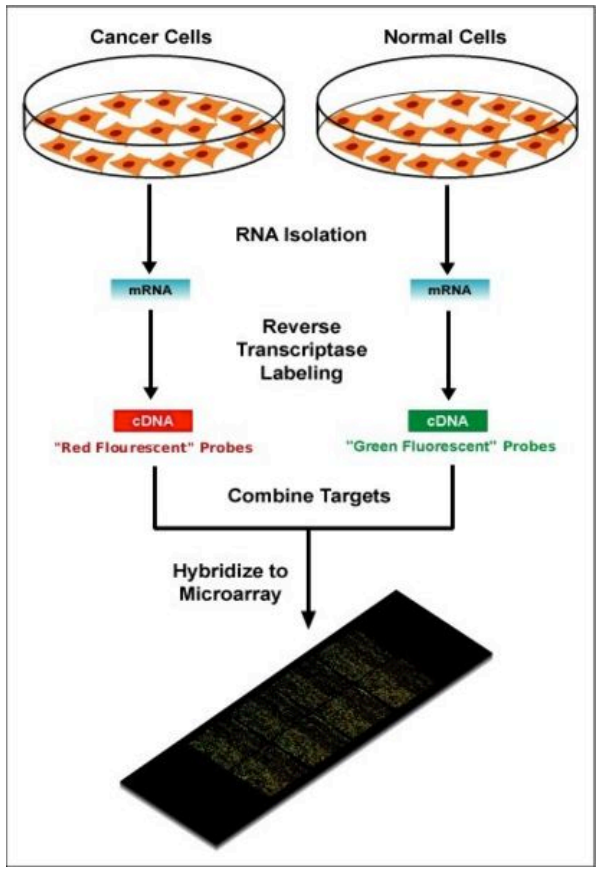
https://beta.etherpad.org/p/compgen_RNAseq

University of Zurich[UZH]

Institute of Molecular Life Sciences

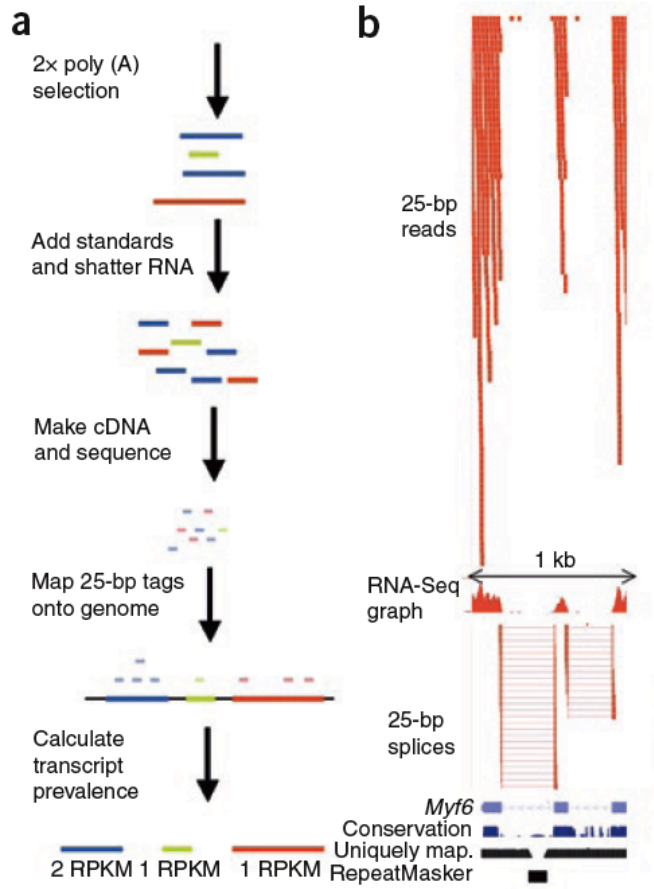
Quick intro

Abundance by Fluorescence Intensity

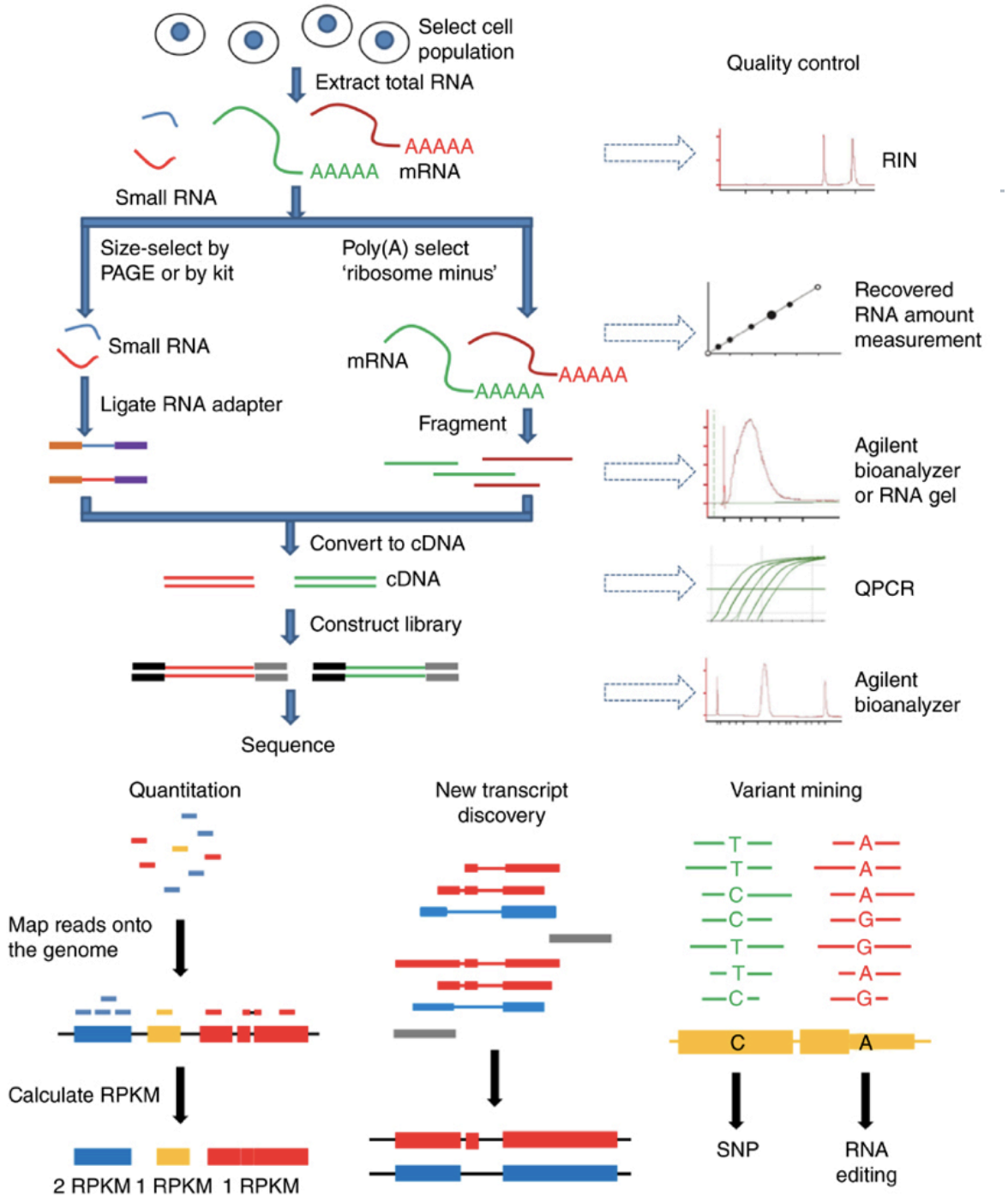


http://en.wikipedia.org/wiki/DNA_microarray

Abundance by Counting



Mortazavi et al., Nature Methods, 2008



Zeng & Mortazavi, Nature Immunology, 2012

Brainstorm

What all can we do with RNA-seq ?

20 minutes: Break into groups. Research/discuss your topic within your group. Find one slide/figure from the literature/web (or some discussion points). A nominated representative from your group can discuss.

Groups:

1. Differential expression
2. Differential splicing
3. allele-specific expression
4. RNA editing
5. *de novo* discovery/assembly
6. expression quantitative trait loci

Experimental design

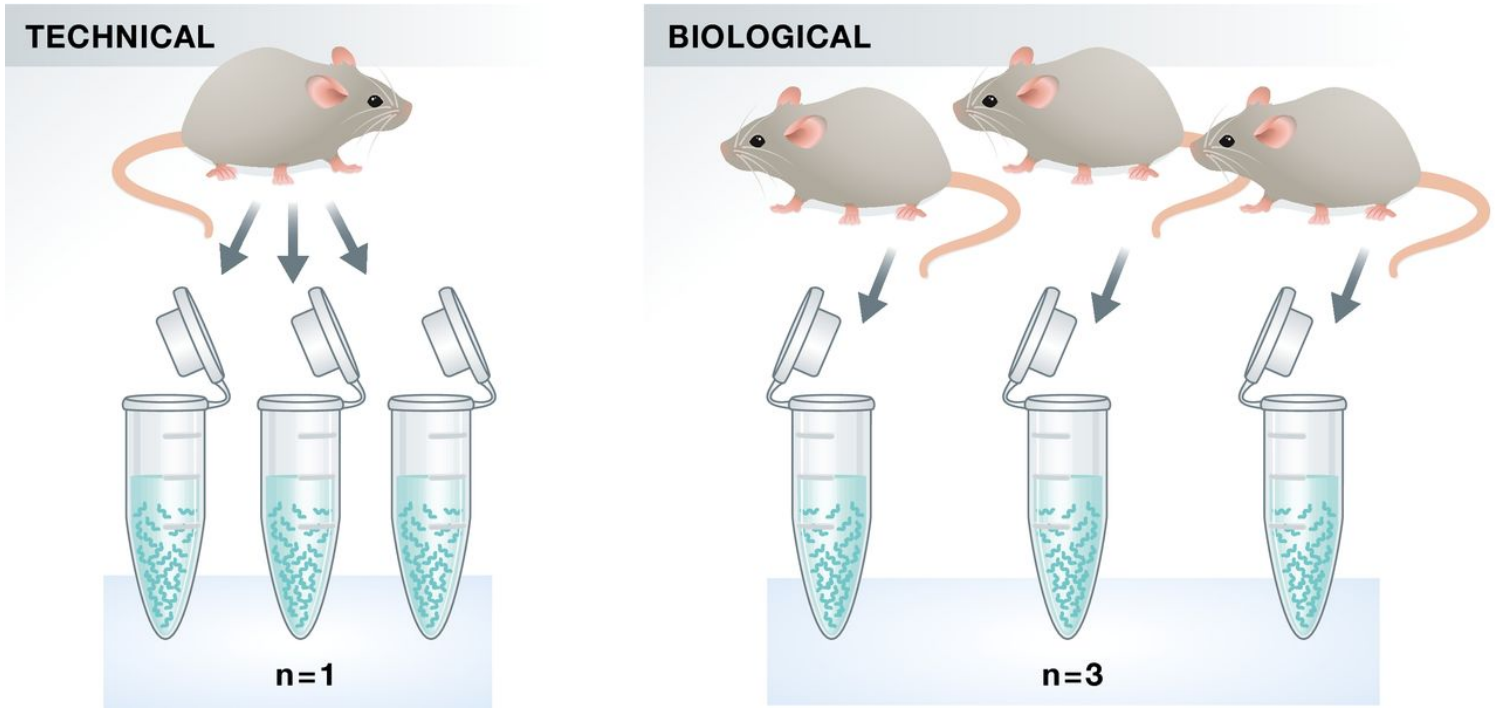
Replication

Randomization

Blocking

From http://emboj.embopress.org/content/early/2015/09/21/embj.201592958

Biological versus technical replication



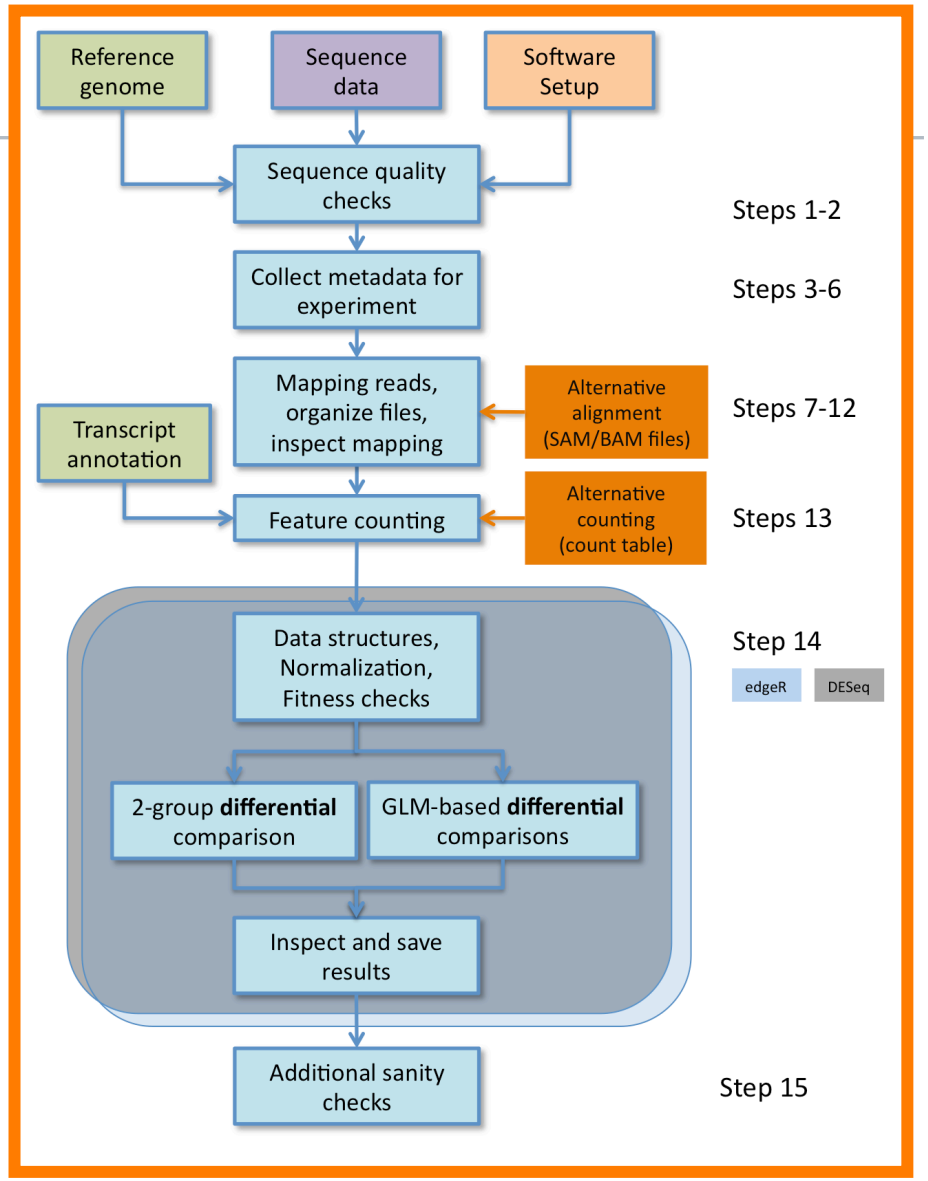
What do we want? Why?

Standard pipeline

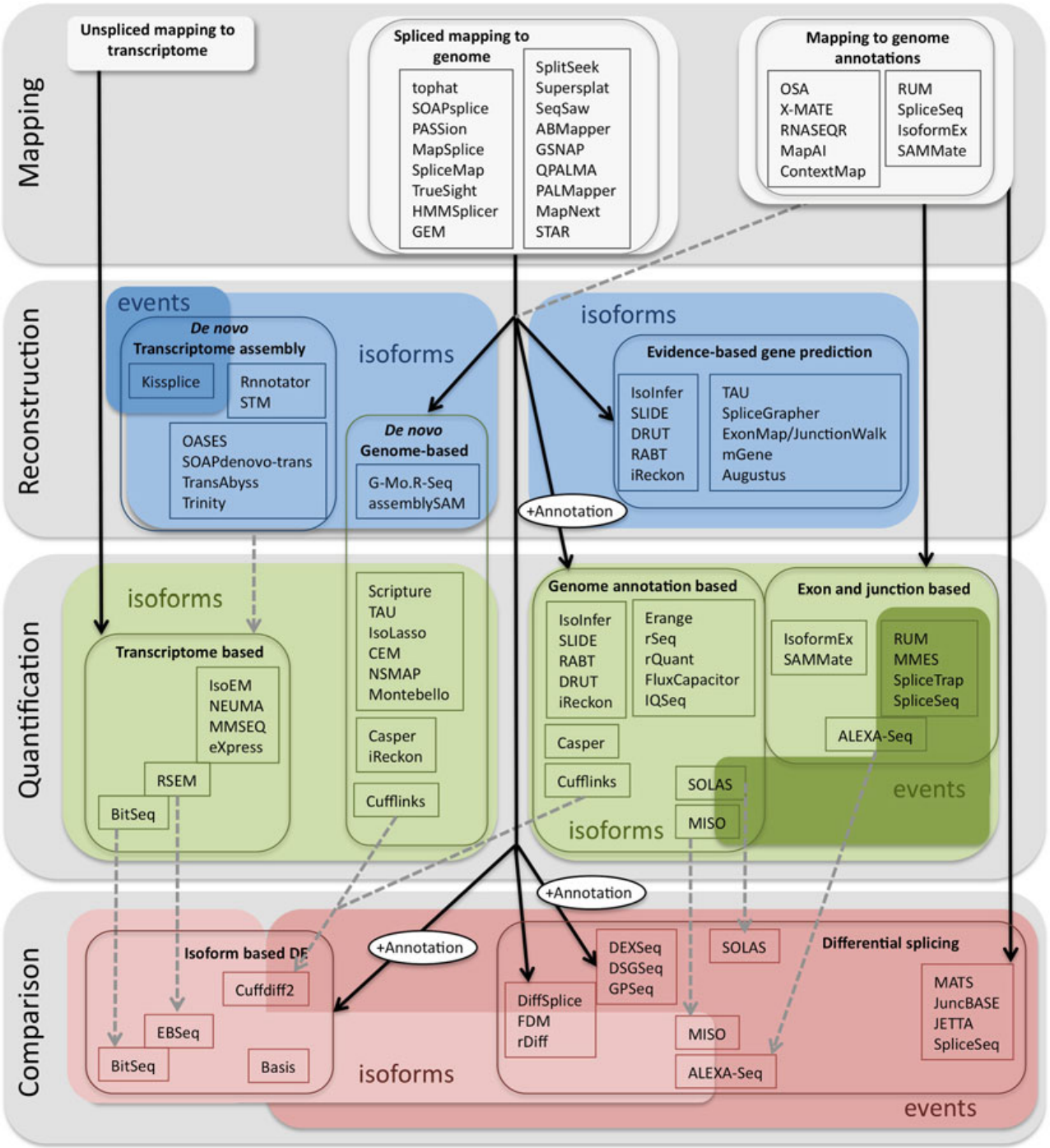
Issues/queries:

Alignment – to genome? to transcriptome ?
versus alignment-free or pseudo-alignment ?

"Counting" – what does that really mean ?



Tools



Chapter 26

Methods to Study Splicing from High-Throughput RNA Sequencing Data

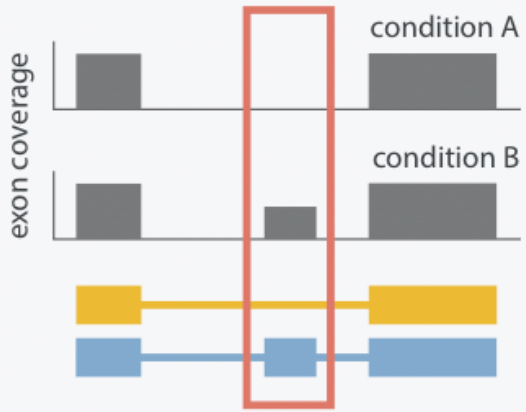
Gael P. Alamancos, Eneritz Agirre, and Eduardo Eyras

New: salmon, kallisto, etc.

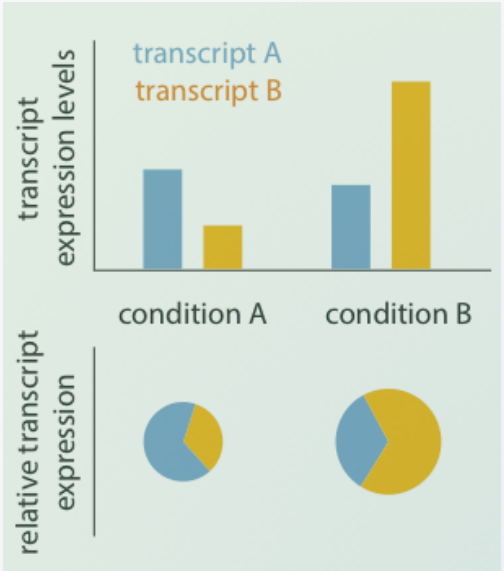
Some terms: DTE, DEU, DTU .. DGE

Differential transcript expression (DTE)

Differential exon usage (DEU)



Differential transcript usage (DTU)

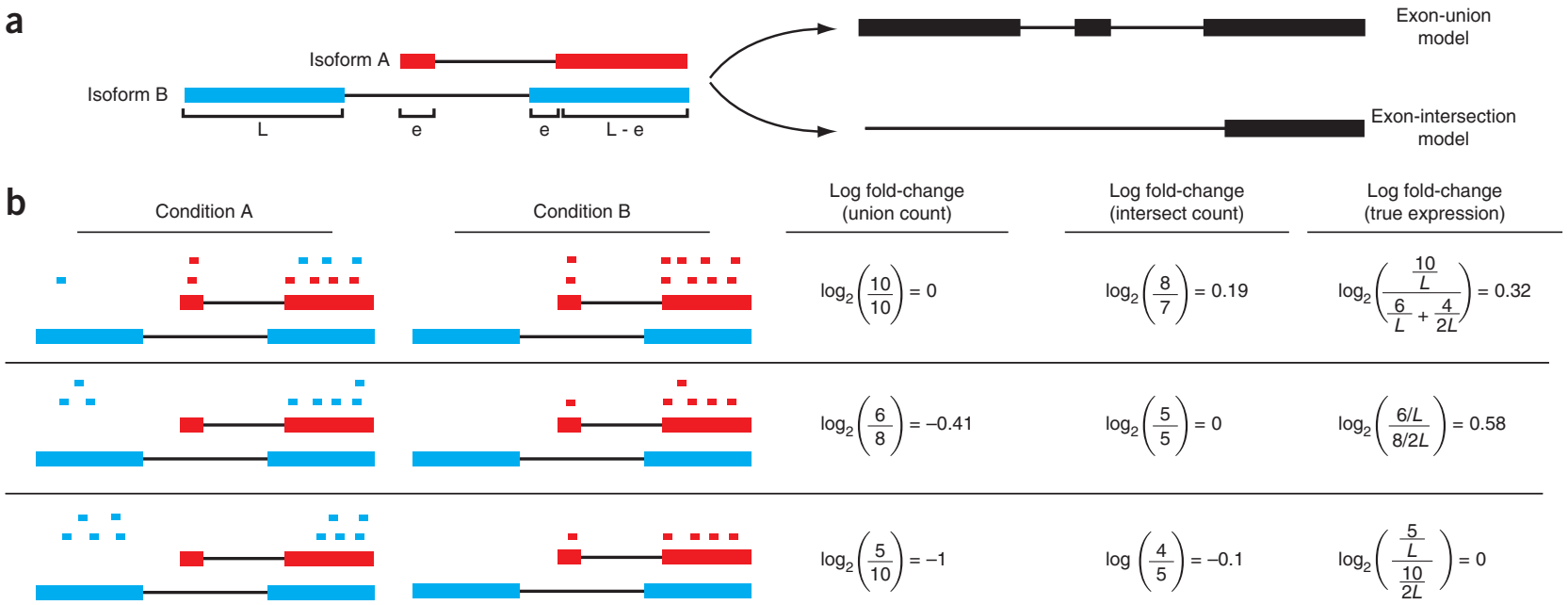


differential splicing

Slide adapted from Mar Gonzàlez-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.pdf

Caveat: gene-level counting can go wrong, but often not bad

Trapnell et al. 2013 Nat Biotech



Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene

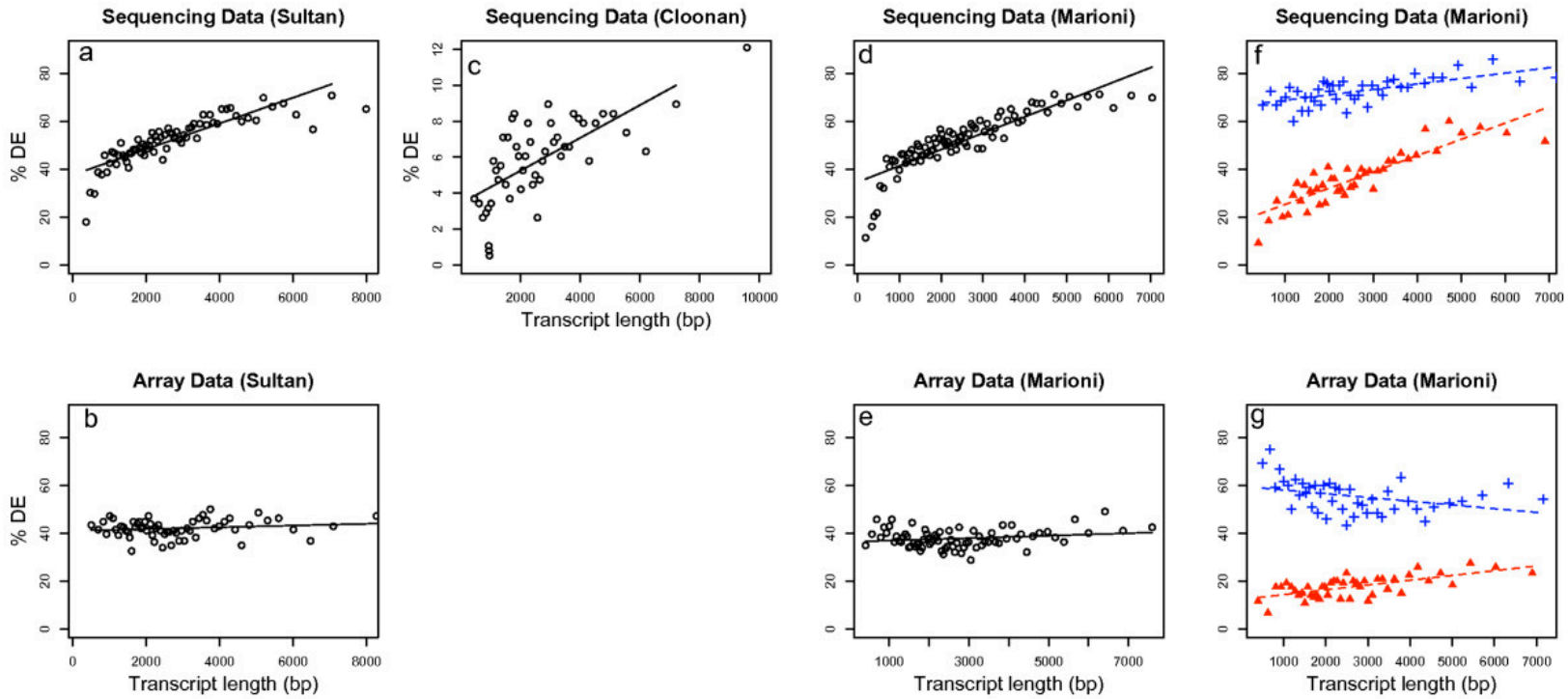
Mar Gonzàlez-Porta[1], Adam Frankish[2], Johan Rung[1], Jennifer Harrow[2] and Alvis Brazma[1*]

SOME OTHER CAVEATS, LIMITATIONS AND REMARKS

Multiple testing

- In genomics, typically P-values are corrected for multiple testing
- FWER = family-wise error rate (strong)
  - control probability of making 1 FP
- FDR = false discovery rate (weak)
  - control rate of FP amongst the set of positives

RNA-seq length bias



Oshlack and Wakefield, 2009

University of Zurich[UZH]

Institute of Molecular Life Sciences

Reconstructing transcripts from current generation RNA-seq?

Assessment of transcript reconstruction methods for RNA-seq

Tamara Steijger[1], Josep F Abril[2,11], Pär G Engström[1,10,11], Felix Kokocinski[3,11], The RGASP Consortium[4], Tim J Hubbard[3], Roderic Guigó[5,6], Jennifer Harrow[3] & Paul Bertone[1,7–9]

Nature Methods 2013

"Consequently, the complexity of higher eukaryotic genomes imposes severe limitations on transcript recall and splice product discrimination that are likely to remain limiting factors for the analysis of current-generation RNA-seq data."

BJP British Journal of Pharmacology

INTERNATIONAL UNION OF BASIC AND CLINICAL PHARMACOLOGY REVIEW

Should pharmacologists care about alternative splicing? IUPHAR Review 4

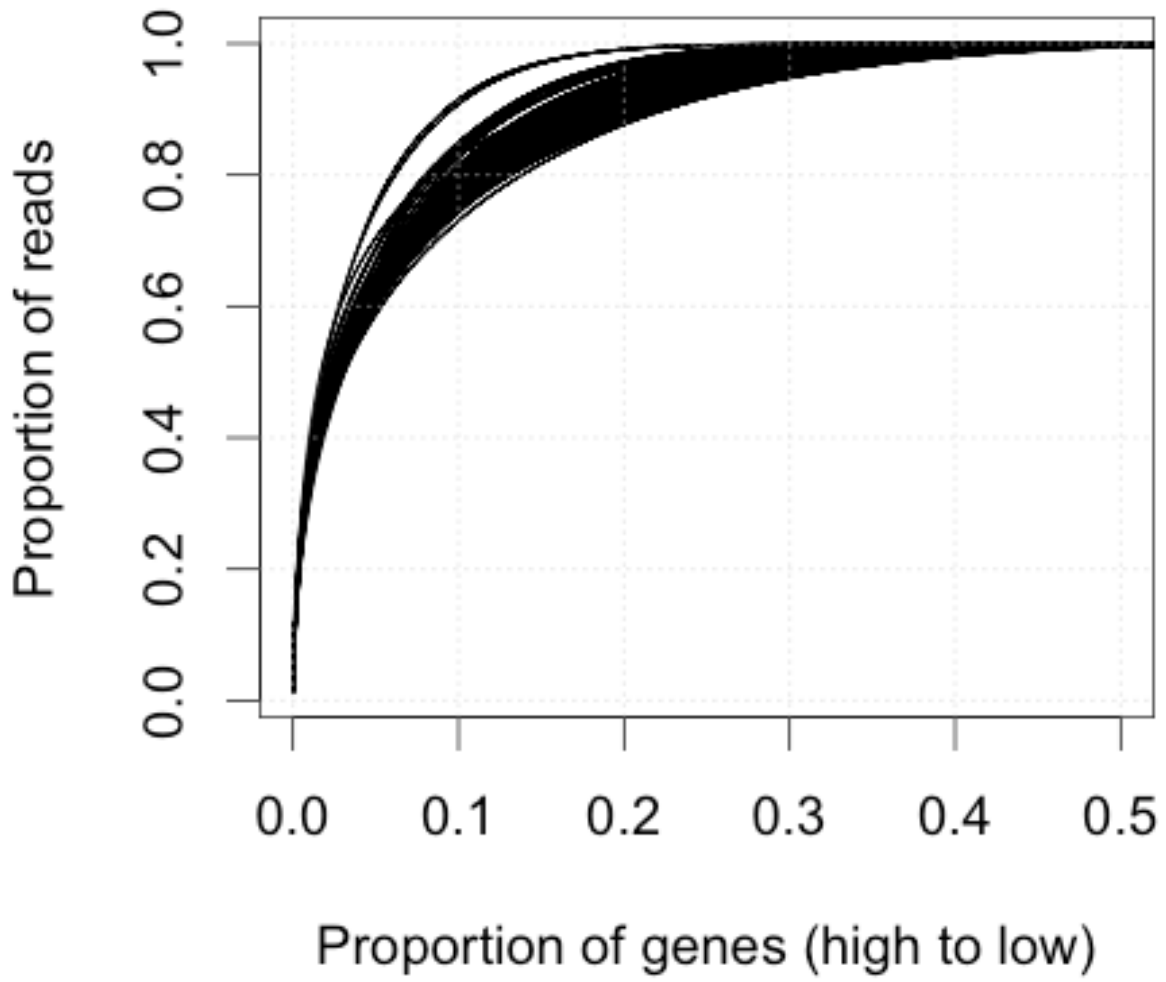
T I Bonner

National Institute of Mental Health, Bethesda, MD, USA

Genome-wide surveys: are we seeing events of interest ?

Alternative splicing of mRNAs occurs in the majority of human genes, and most differential splicing results in different protein isoforms with possibly different functional properties. However, there are many reported splicing variations that may be quite rare, and not all combinatorially possible variants of a given gene are expressed at significant levels. Genes of interest to pharmacologists are frequently expressed at such low levels that they are not adequately represented in genome-wide studies of transcription. In single-gene studies, data are commonly available on the relative abundance and functional significance of individual alternatively spliced exons, but there are rarely data that quantitate the relative abundance of full-length transcripts and define which combinations of exons are significant. A number of criteria for judging the significance of splice variants and suggestions for their nomenclature are discussed.

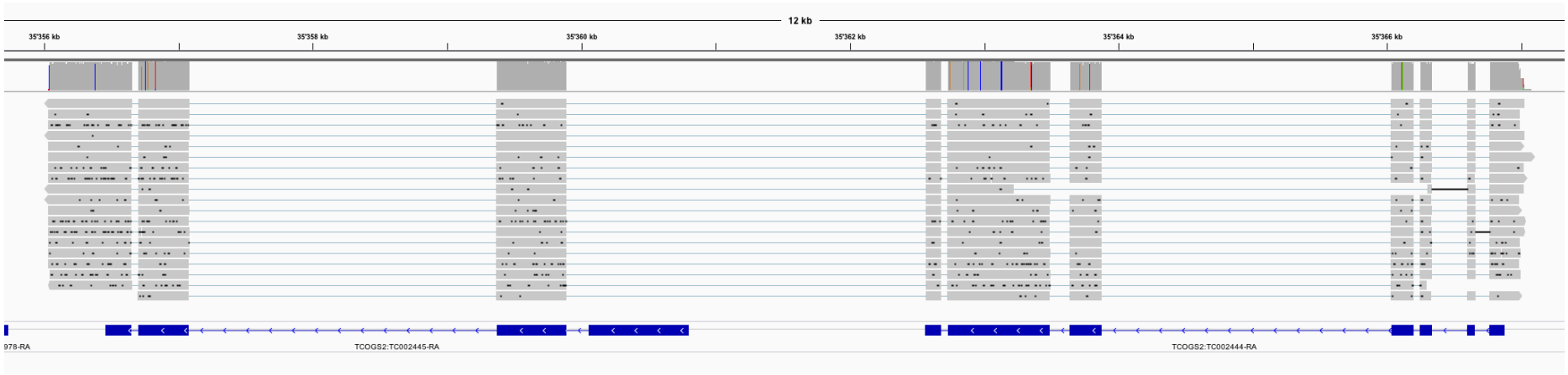
How much sequencing goes to highly expressed genes?



Soon: counting full-length transcripts?

→ PacBio, Oxford Nanopore, 10x Genomics + Illumina

PacBio IsoSeq data collected
(with D. Bopp and Functional Genomics Centre Zurich)



University of Zurich[UZH]

Institute of Molecular Life Sciences

THEORY

Differential expression, small sample inference

- Table of data (e.g., microarray gene expression data OR RNA-seq gene expression counts with replicates of each of condition A, condition B)
  - *rows* = features (e.g., genes), *columns* = experimental units (samples)
- Most common problem in statistical bioinformatics: want to infer whether there is a **change in the response** → a statistical test for each row of the table.

What test might you use? Why is this hard? What issues arise? How much statistical power is there [1] ?

```
> head(y)
        group0     group0      group0     group1     group1      group1
gene1 -0.1874854  0.2584037 -0.05550717 -0.4617966 -0.3563024 -0.03271432
gene2 -3.5418798 -2.4540999  0.11750996 -4.3270442 -5.3462622 -5.54049106
gene3 -0.1226303  0.9354707 -1.10537767 -0.1037990  0.5221678 -1.72360854
gene4 -2.3394536 -0.3495697 -3.47742610 -3.2287093  6.1376670 -2.23871974
gene5 -3.7978820  1.4545702 -7.14796503 -4.0500796  4.7235714 10.00033769
gene6  1.4627078 -0.3096070 -0.26230124 -0.7903434  0.8398769 -0.96822312
```

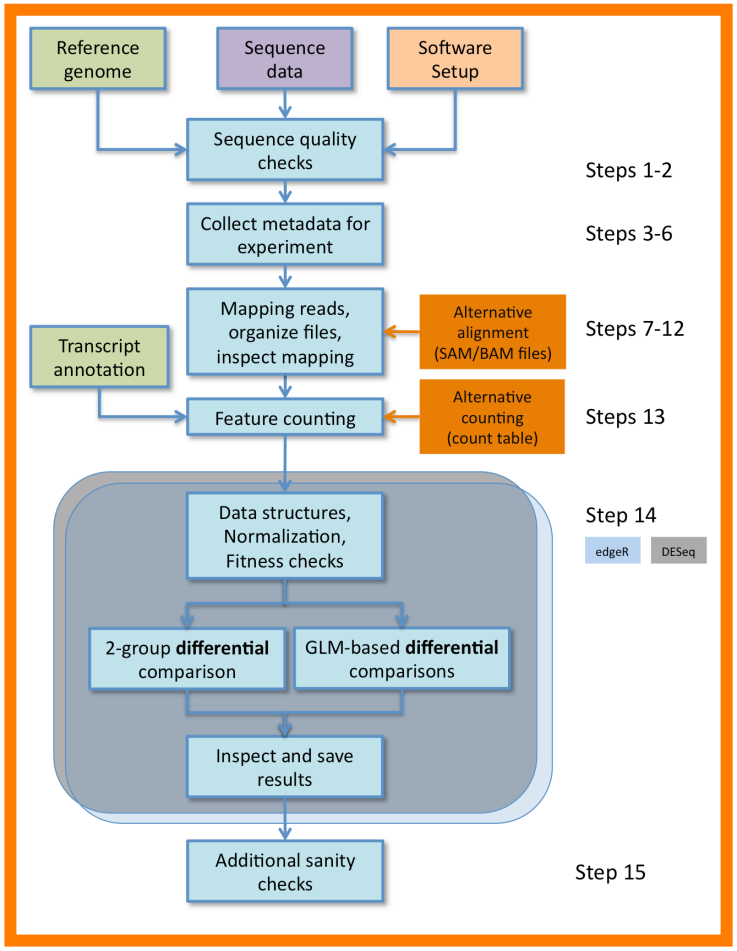
[1] http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html

University of Zurich[UZH]

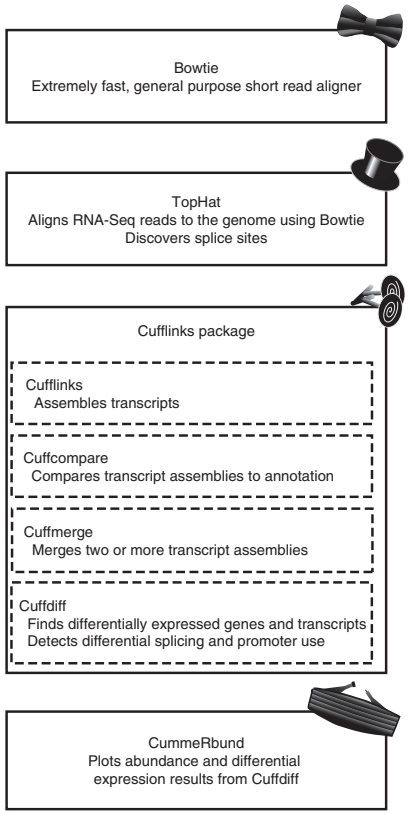
Data analysis pipelines for RNA-seq differential expression

Institute of Molecular Life Sciences

edgeR, DESeq



cufflinks, cuffdiff



Trinity

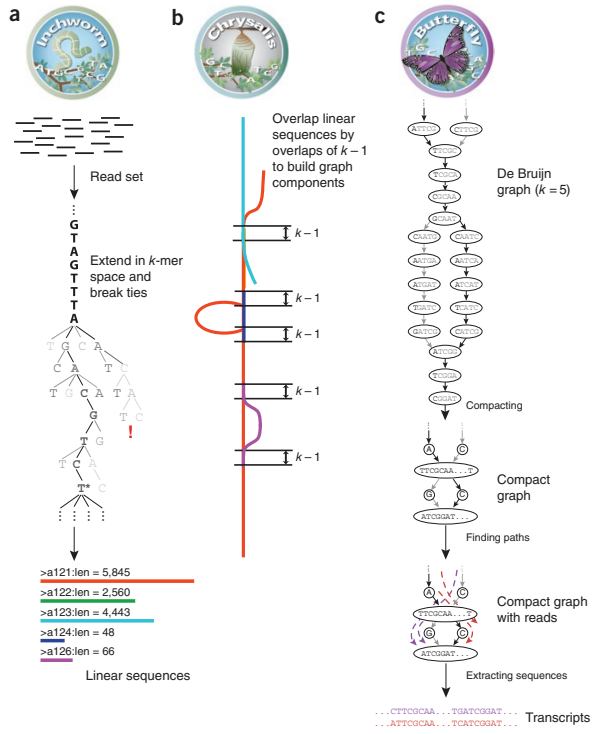


Figure 1 Overview of Trinity. (**a**) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a *k*-mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each *k*-mer present only once in the contigs. (**b**) Chrysalis pools contigs (colored lines) if they share at least one $k-1$-mer and if reads span the junction between contigs, and then it builds individual de Bruijn graphs from each pool. (**c**) Butterfly takes each de Bruijn graph from Chrysalis (top), and trims spurious edges and compacts linear paths (middle). It then reconciles the graph with reads (dashed colored arrows, bottom) and pairs (not shown), and outputs one linear sequence for each splice form and/or paralogous transcript represented in the graph (bottom, colored sequences).

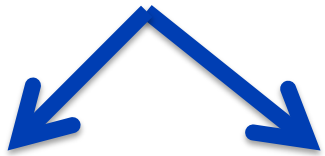
Nature Protocols September 2013 (preprint at http://arxiv.org/pdf/1302.3685v3.pdf)

Differential expression: why not use methods developed for microarrays?

Count data is discrete, not continuous.

Methods designed for microarrays are not directly applicable and suboptimal **(more on this later)**

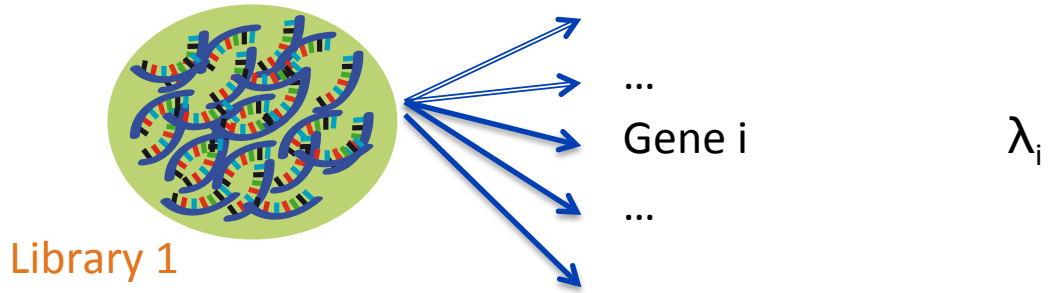
Two options:



Transform count data and apply standard methodology

Analyze using models for count data

For a single gene, it's a coin toss, i.e. Binomial



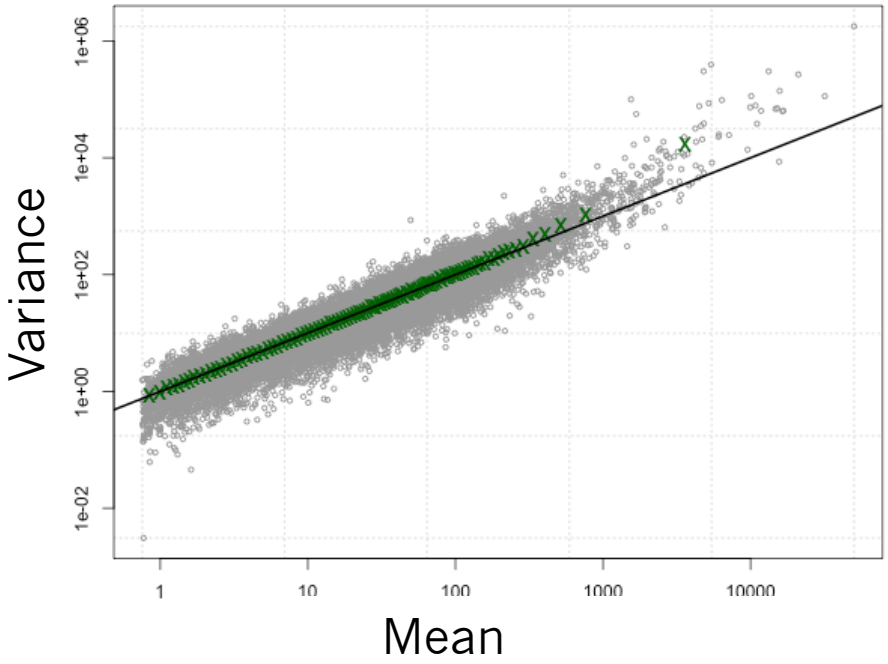
$Y_i \sim \text{Binomial}(M, \lambda_i)$

Y_i - observed number of reads for gene i
M - total number of sequences
λ_i - proportion

Large M, small λ_i → approximated well by Poisson($\mu_i = M \bullet \lambda_i$)

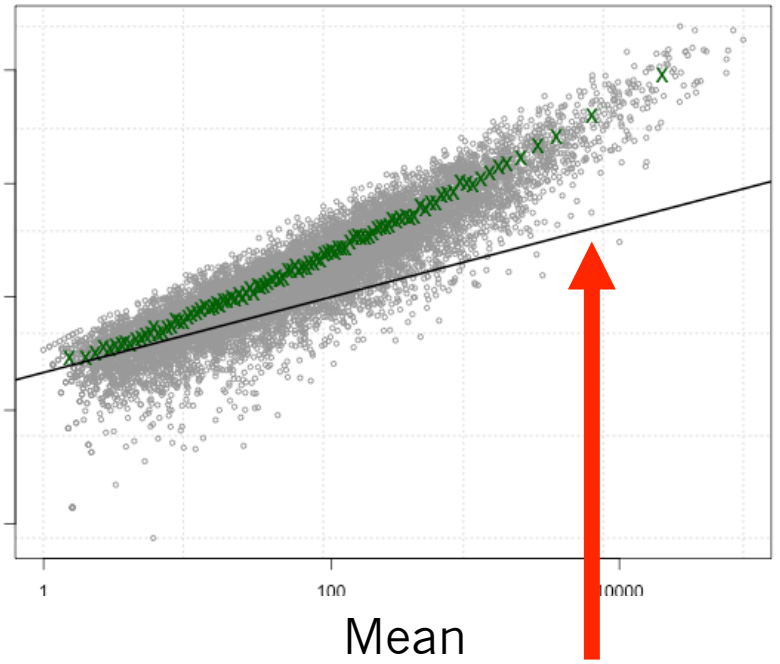
Mean-Variance plots: What we see in real data

Technical replicates



Data from Marioni et al. Genome Research 2008

Biological replicates



Data from Parikh et al. *Genome Biology* 2010

mean=variance
(Poisson assumption)

Count data modeling assumptions

Poisson adequately describes technical variation

$$Y_i \sim \text{Pois}(M * \lambda_i)$$

$$\text{mean}(Y_i) = \text{variance}(Y_i) = M * \lambda_i$$

Negative binomial (gamma-Poisson) model is a natural extension that allows **biological** variability:

$$Y_i \sim \text{NB}(\mu_i = M * \lambda_i , \phi_i)$$

Same mean, variance is quadratic in the mean:

$$\text{variance}(Y_i) = \mu_i (1 + \mu_i \phi_i)$$

M = library size

λ_i = relative contribution of gene i

Analogy to t-tests (microarray setting)

$$t_g = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{s_g\, c}$$

Feature-specific

$$\tilde{t}_g = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{\tilde{s}_g\, u}$$

Moderated

$$t_{g,\text{pooled}} = \frac{\overline{y}_{\text{mu}} - \overline{y}_{\text{wt}}}{s_0\, c}$$

Common

Why did moderated-t work well for microarrays ?

Moderated estimates: let's try the same strategy with counts

At one extreme, assume all genes have same dispersion (too strong)

At other extreme, estimate dispersion separately/ independently for each gene (poor estimates)

Shrink individual estimates toward common/trend (how?)

No hierarchical model (e.g. limma) to do this:
approximations, weighted likelihood

No t-distribution theory to formulate statistical tests.

Second challenge: Moderate dispersion estimate

Weighted likelihood -- individual log-likelihood plus a weighted version of the **common log**-likelihood:

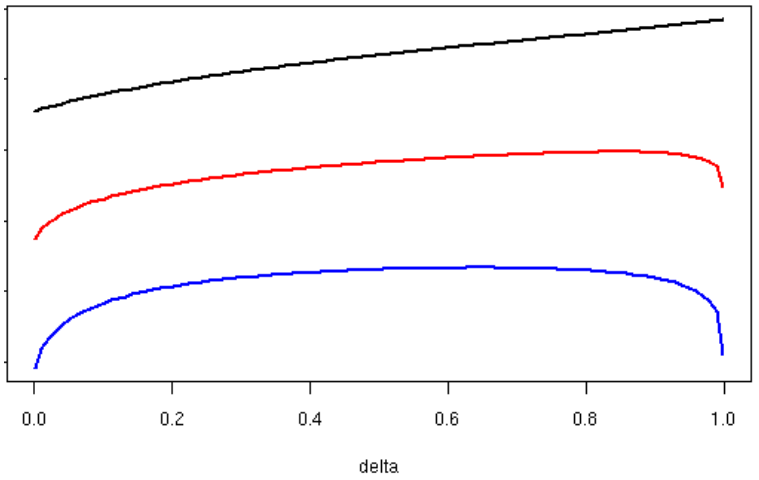
$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha\, l_C(\phi_g)$$

$(1\text{-}\alpha)$

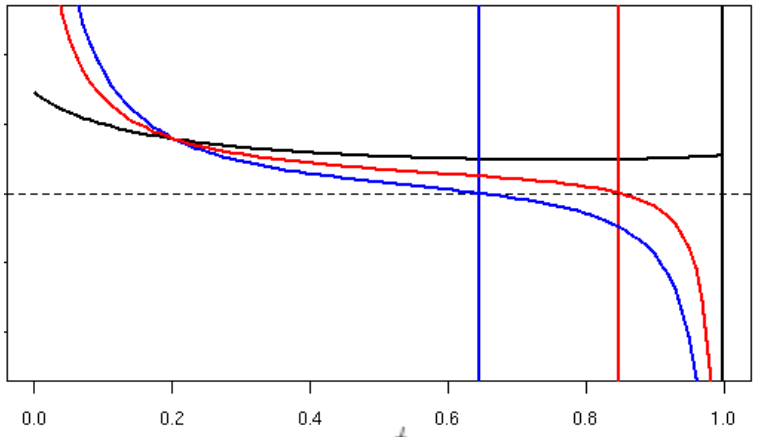
L_g - quantile-adjusted conditional likelihood

Black: single tag
Blue: common dispersion
Red: Linear combination of the two

Log-Likelihood



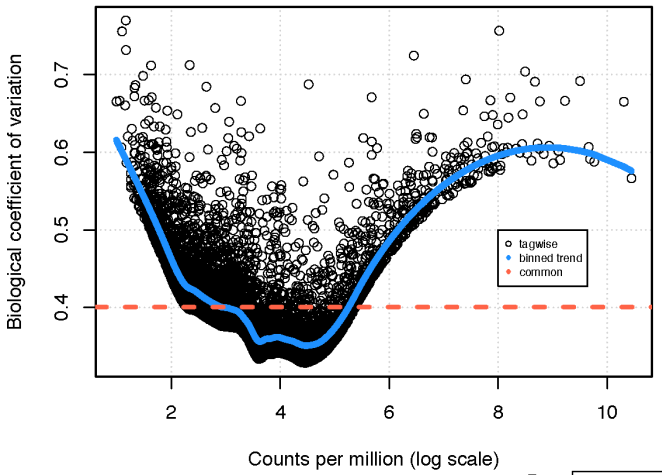
Score (1st derivative of LL)

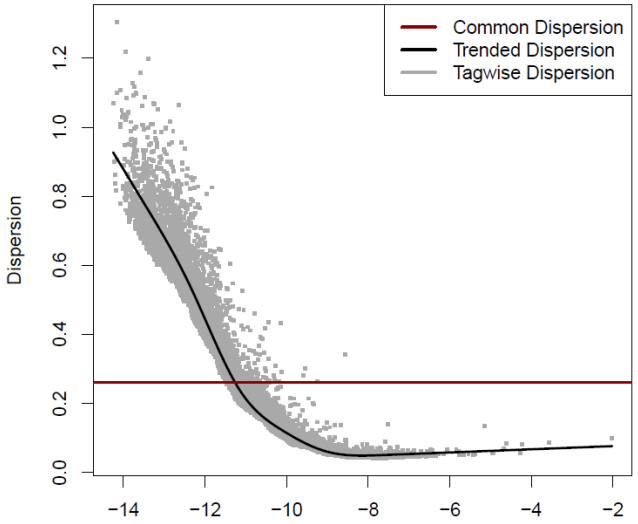


$$\delta = \frac{\phi}{\phi+1}$$

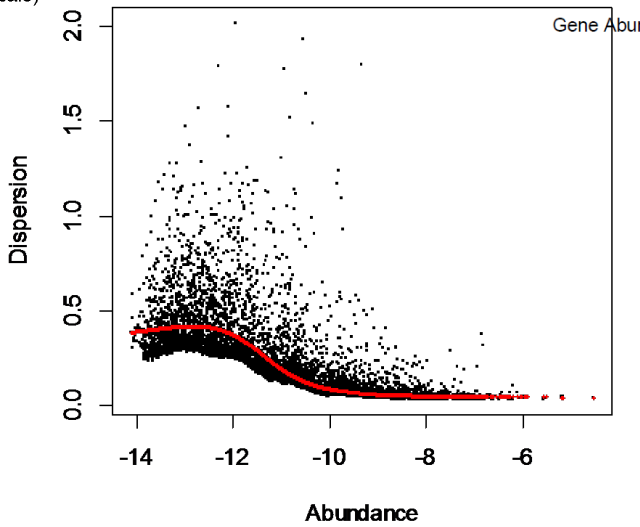
Dispersion varies with mean: moderate dispersion towards trend

Data:
Tuch et al., 2008





Mouse hemapoeitic stem cells

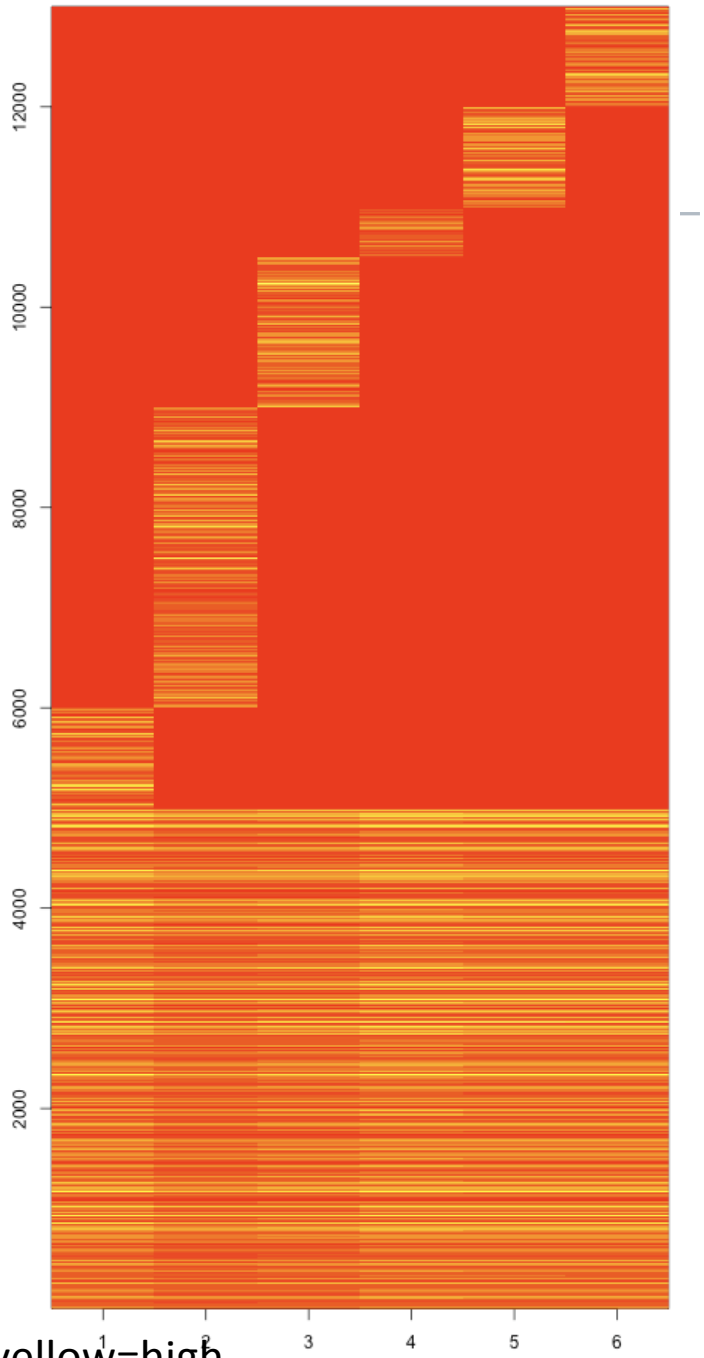


Mouse lymphomas

Advantage: genes are allowed to have their own variance.

"Composition" or "Diversity" can affect read depth

- Hypothetical example: Sequence 6 libraries to the **same** depth, with varying levels of *unique-to-sample* counts
- Read depth is affected not only by expression (and length), but also expression levels of other genes
- Composition can induce (sometimes significant) differences in counts



Red=low, goldenyellow=high

Kidney and Liver RNA have very different composition

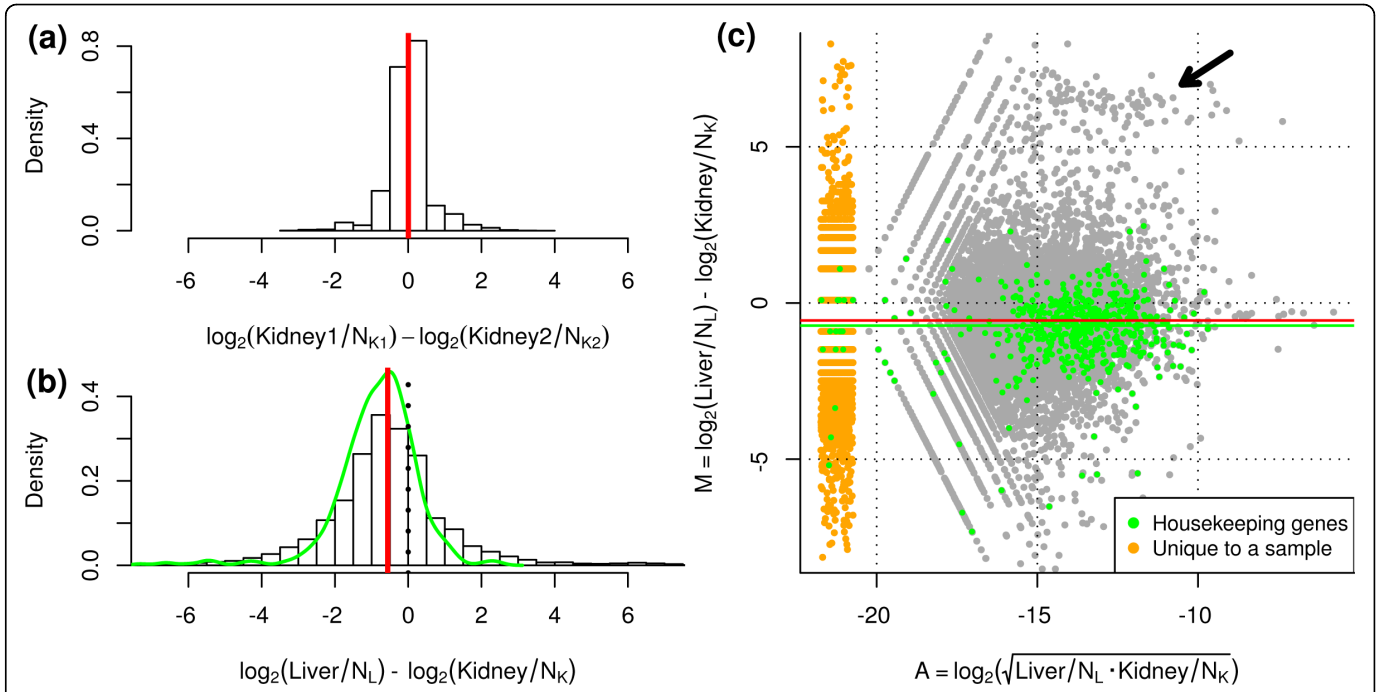


Figure 1 Normalization is required for RNA-seq data. Data from [6] comparing log ratios of **(a)** technical replicates and **(b)** liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. **(c)** An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney tissues. The black arrow highlights the set of prominent genes that are largely attributable for the overall bias in log-fold-changes.

Robinson and Oshlack (2010) Genome Biology

What does transformation do to M-V relationship?

For Poisson data, square-root should stabilize

Logarithm is too strong – variance decreases to asymptote (dispersion; Neg Bin) or 0 (Poisson)

How to pick? Doesn't matter … voom

voom: mean-variance modeling at the observational level

```
voom                        package:limma                        R Documentation

Transform RNA-Seq Data Ready for Linear Modelling

Description:

     Transform count data to log2-counts per million, estimate the
     mean-variance relationship and use this to compute appropriate
     observational-level weights. The data are then ready for linear
     modeling.
```

Model log counts per million

log counts per million:

$$z_{gi} = \log_2\left(1e6\frac{\text{count}_{gi}+0.5}{\text{libsize}_{gi}+1.0}\right) = \log_2\left(1e6\frac{y_{gi}+0.5}{M_{gi}+1.0}\right)$$

normalize libsize in advance or normalize z_{gi} as for microarrays.

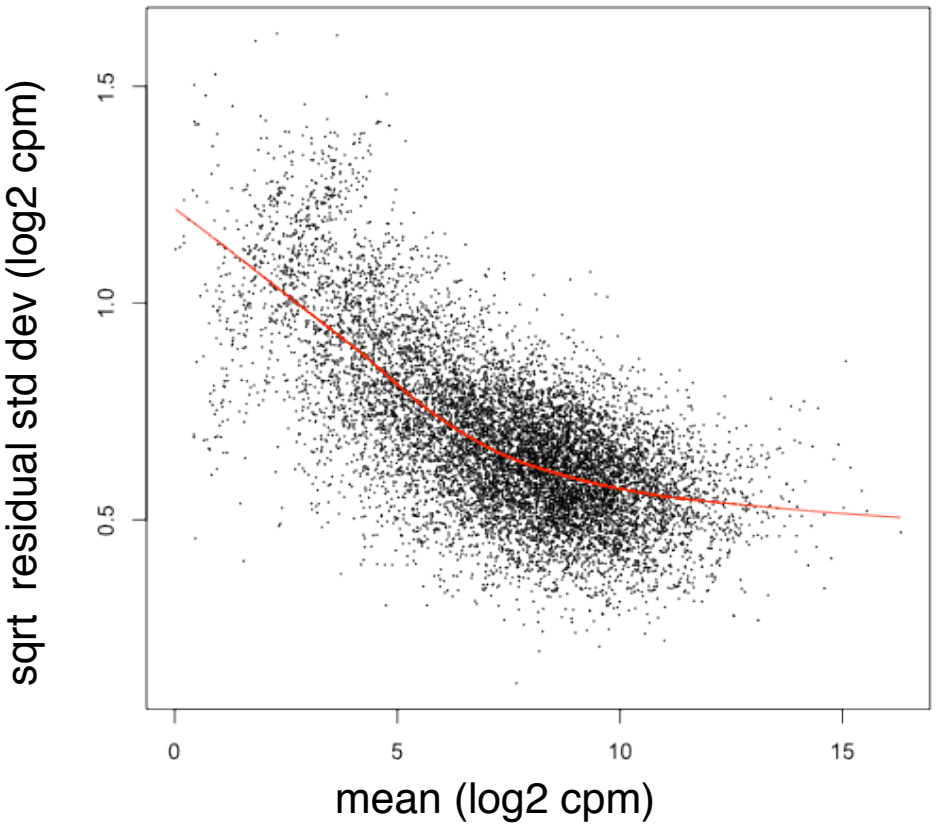
Linear modelling:

$$E\left(z_{gi}\right) = \mu_{gi} = x_i^T\beta_g$$

$$\text{var}\left(z_{gi}\right) = s(\mu_{gi})\sigma_g^2$$

Smooth function of mean

voom fits a lowess trend to the mean-variance relationship …



→ Use weights (1/var) in limma analysis

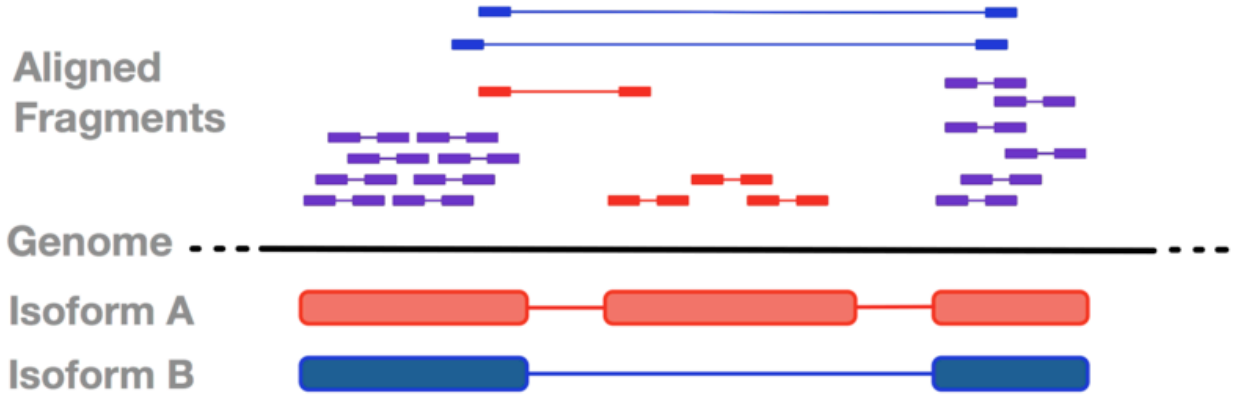
TRANSCRIPT-LEVEL ESTIMATION

Transcript-level expression estimation

BitSeq
CEM
Cufflinks
eXpress
IsoEM
MMSEQ
RSEM
rSEQ
Sailfish
Scripture
TIGAR2

salmon
kallisto

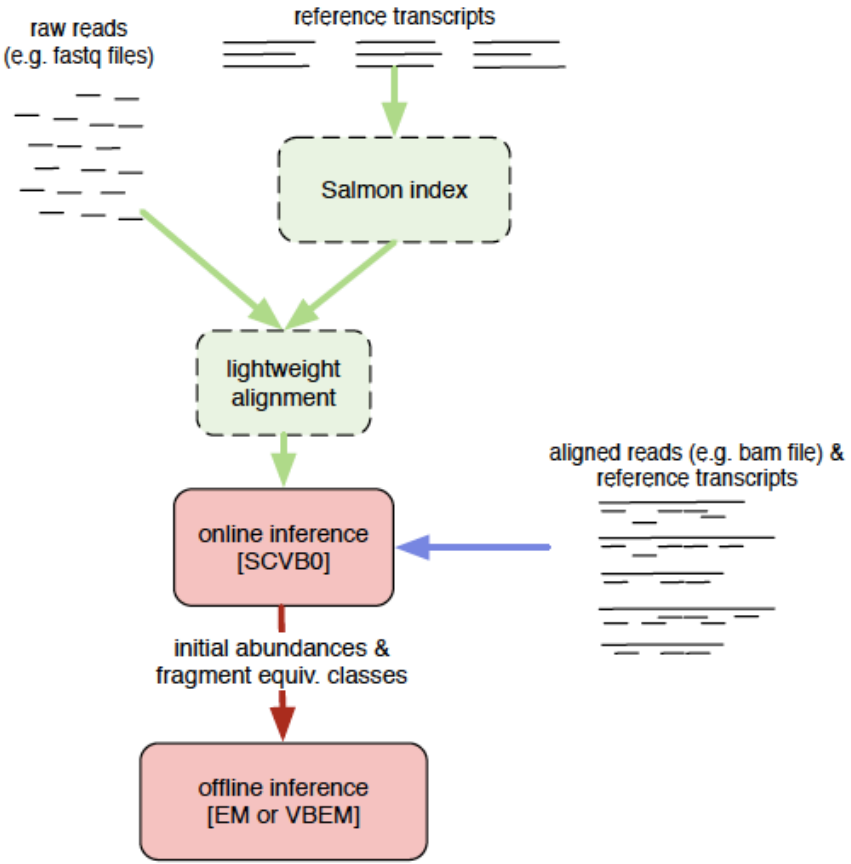
..



Open question: can you use “estimated counts” into a method that models counts?

Short answer: yes, though improvements that properly take estimation uncertainty into account may be possible.

WHAT ABOUT ALL THIS EXCITEMENT AROUND PSEUDOALIGNMENT?



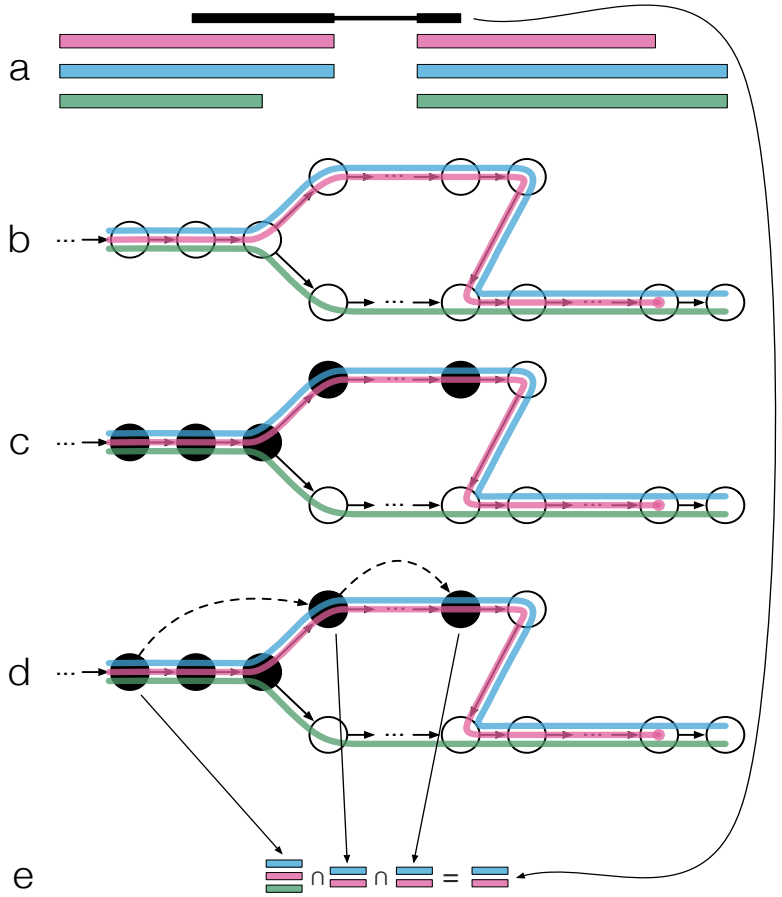


Figure 1: Overview of kallisto. (a) The input consists of a reference transcriptome and reads from an RNA-Seq experiment. (b) An index is constructed by creating the Transcriptome de Bruijn Graph (T-DBG) where nodes are *k*-mers, each transcript corresponds to a path and the path cover of the transcriptome induces a *k*-compatibility class for each *k*-mer. (c) Conceptually, the *k*-mers of a read are hashed (black nodes) to find the *k*-compatibility class of a read. (d) Skipping uses the information stored in the T-DBG to skip *k*-mers that are redundant due to having the same *k*-compatibility class. (e) The *k*-compatibility class of the read is determined by taking the intersection of the *k*-compatibility classes of its constituent *k*-mers.

Other useful resources

RNA-Seq Methods and Algorithms (Part I – Intro and overview of RNA-Seq) 2015 UC Davis Workshop
https://www.youtube.com/watch?v=96yBPM8lEt8

RNA-Seq Methods and Algorithms (Part II – Alignment Algorithms) 2015 UC Davis Workshop
https://www.youtube.com/watch?v=b4tVokh6Law

RNA-Seq Methods and Algorithms (Part III – Quantification) 2015 UC Davis Workshop
https://www.youtube.com/watch?v=ztyjiCCt_lM

RNA-Seq Methods and Algorithms (Part IV – Differential Expression) 2015 UC Davis Workshop
https://www.youtube.com/watch?v=BRWj6re9iGc