



**University of
Zurich** ^{UZH}

Institute of Molecular Life Sciences

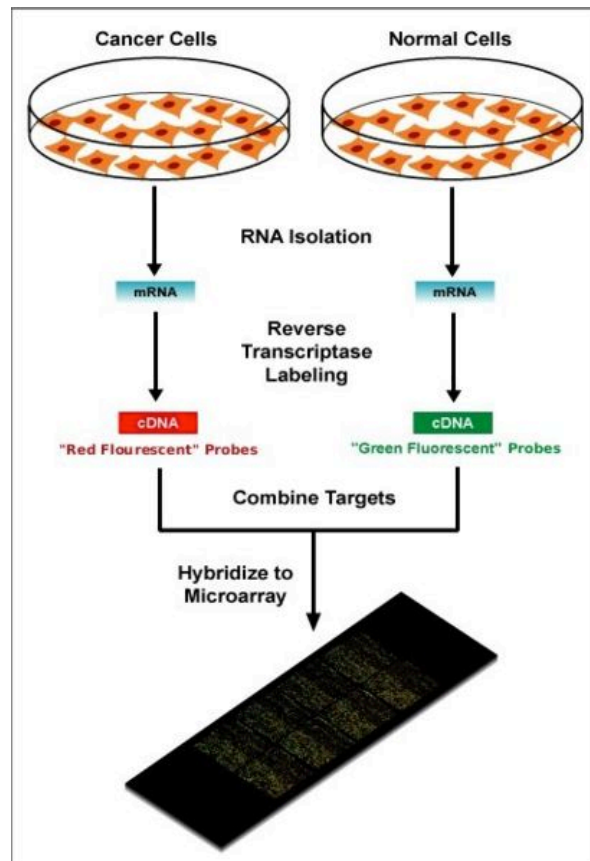
Lecture part: BioC2014

Differential gene- and exon-level expression analyses for RNA-seq data using edgeR, voom and featureCounts

Mark D. Robinson, University of Zurich

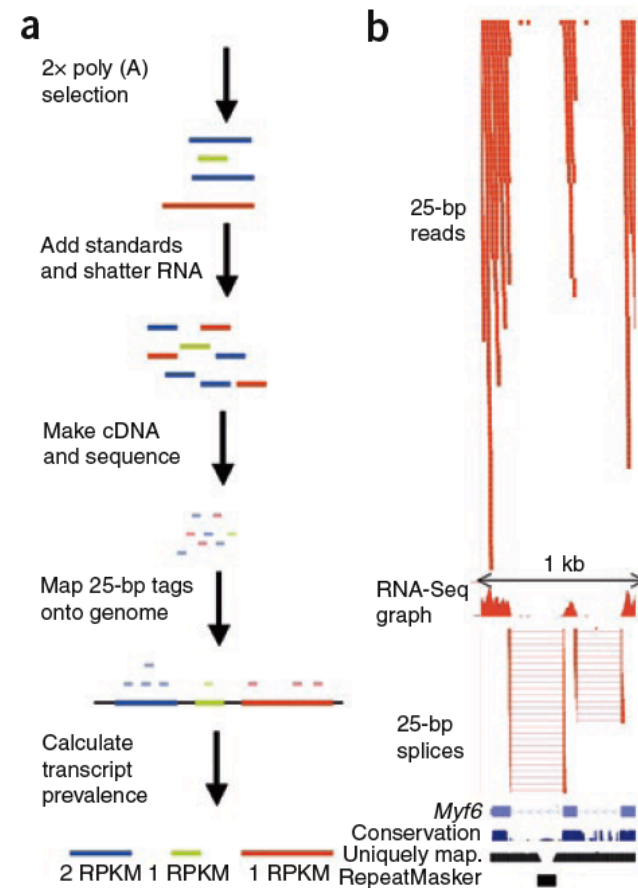


Abundance by Fluorescence Intensity



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

Abundance by Counting



Mortazavi et al., Nature Methods, 2008

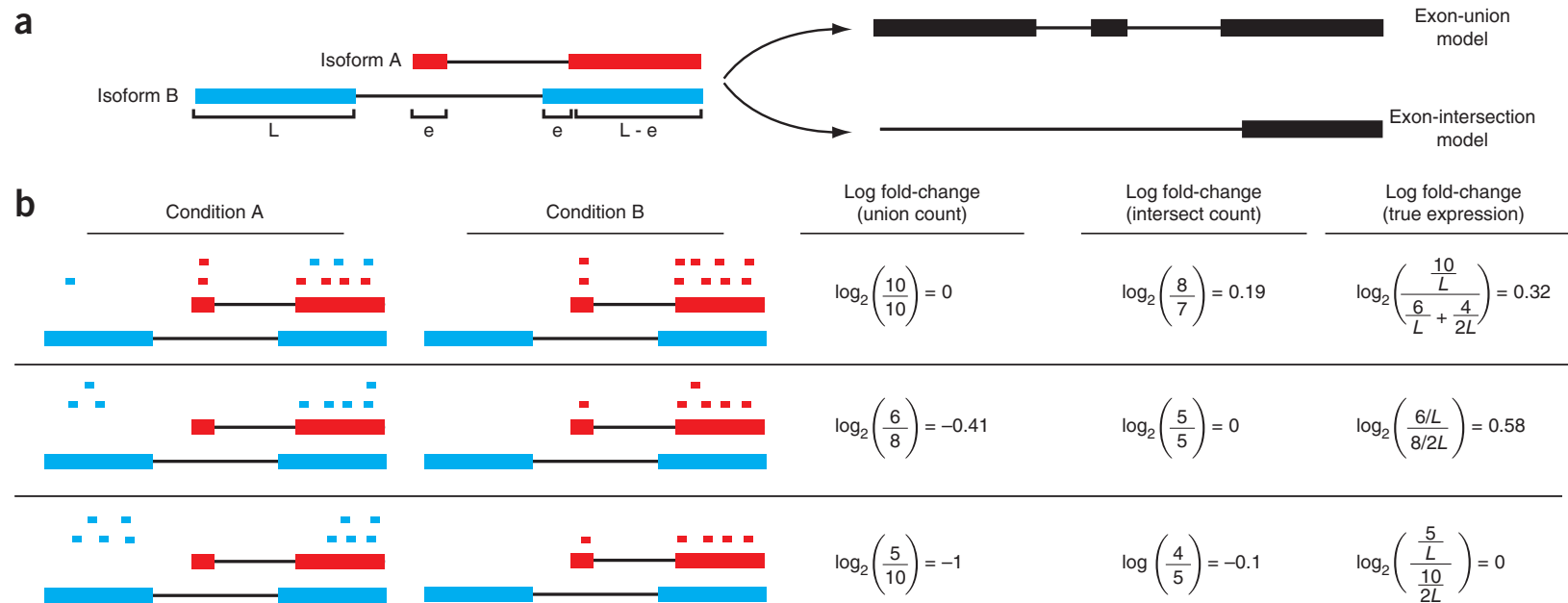


University of
Zurich^{UZH}

Institute of Molecular Life Sciences

Gene-level counting: issues can be dealt with at second step

Trapnell et al. 2013 Nat Biotech



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

Mar Gonzàlez-Porta¹, Adam Frankish², Johan Rung¹, Jennifer Harrow² and Alvis Brazma^{1*}



University of
Zurich^{UZH}

Data analysis pipelines for RNA-seq differential expression

Institute of Molecular Life Sciences

edgeR, DESeq

cufflinks, cuffdiff

Trinity

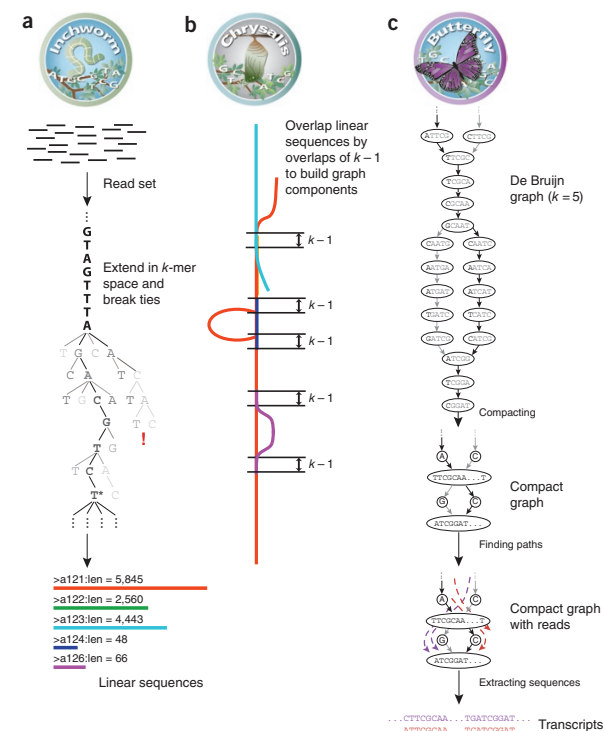
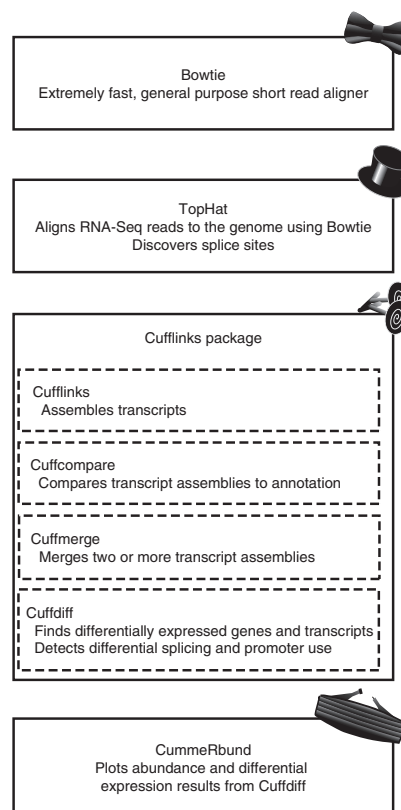
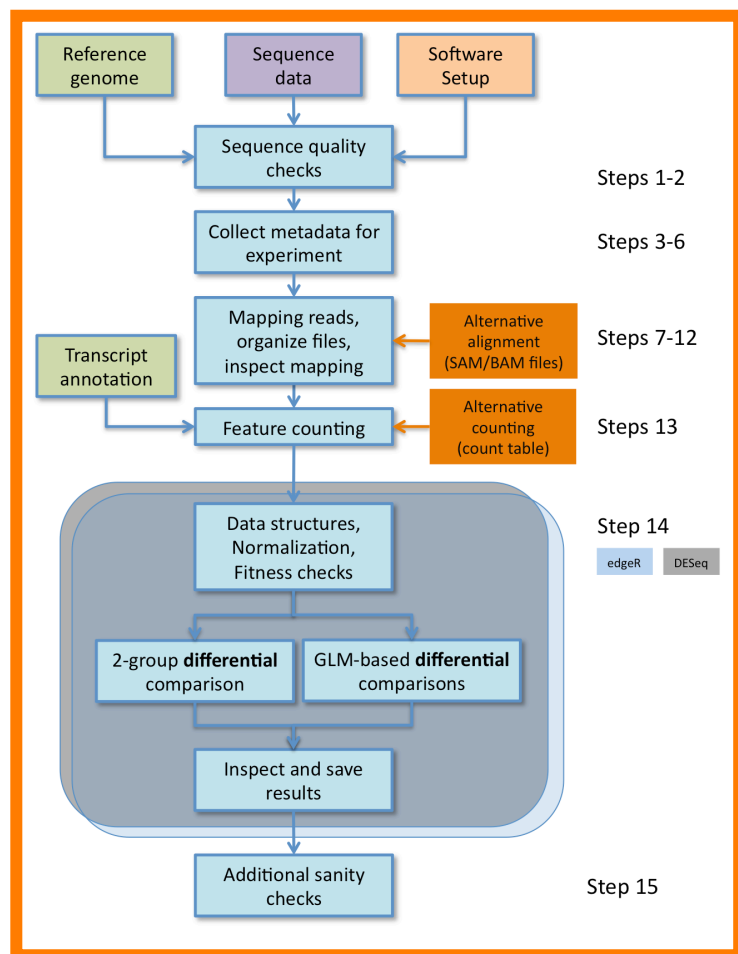


Figure 1 Overview of Trinity. (a) Inchworm assembles the read data set (short black lines, top) by greedily searching for paths in a k -mer graph (middle), resulting in a collection of linear contigs (color lines, bottom), with each k -mer present only once in the contigs. (b) Chrysalis pools contigs (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).

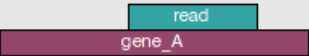
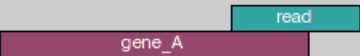


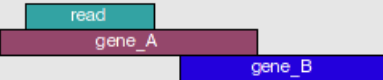

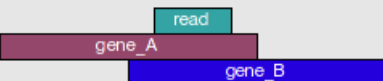
Nature Protocols September 2013 (preprint
at <http://arxiv.org/pdf/1302.3685v3.pdf>)



Counting: a few considerations (gene-level)

All the downstream statistical methods start with a count table.

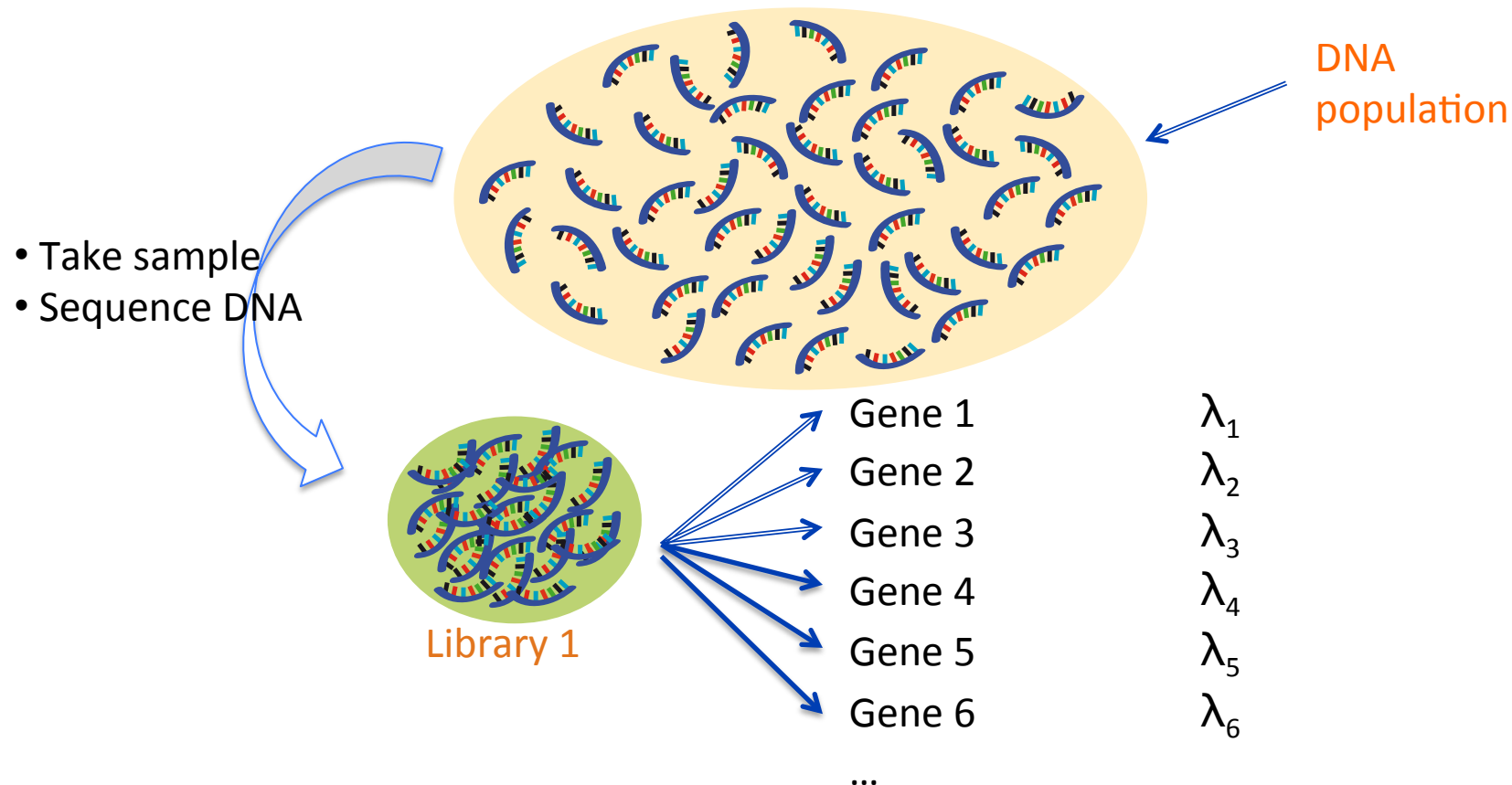
- annotation-based? What about novel genes?
- gene-level versus transcript-level? versus exon-level?
- ambiguities

	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

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



Sampling reads from population of DNA fragment is multinomial





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

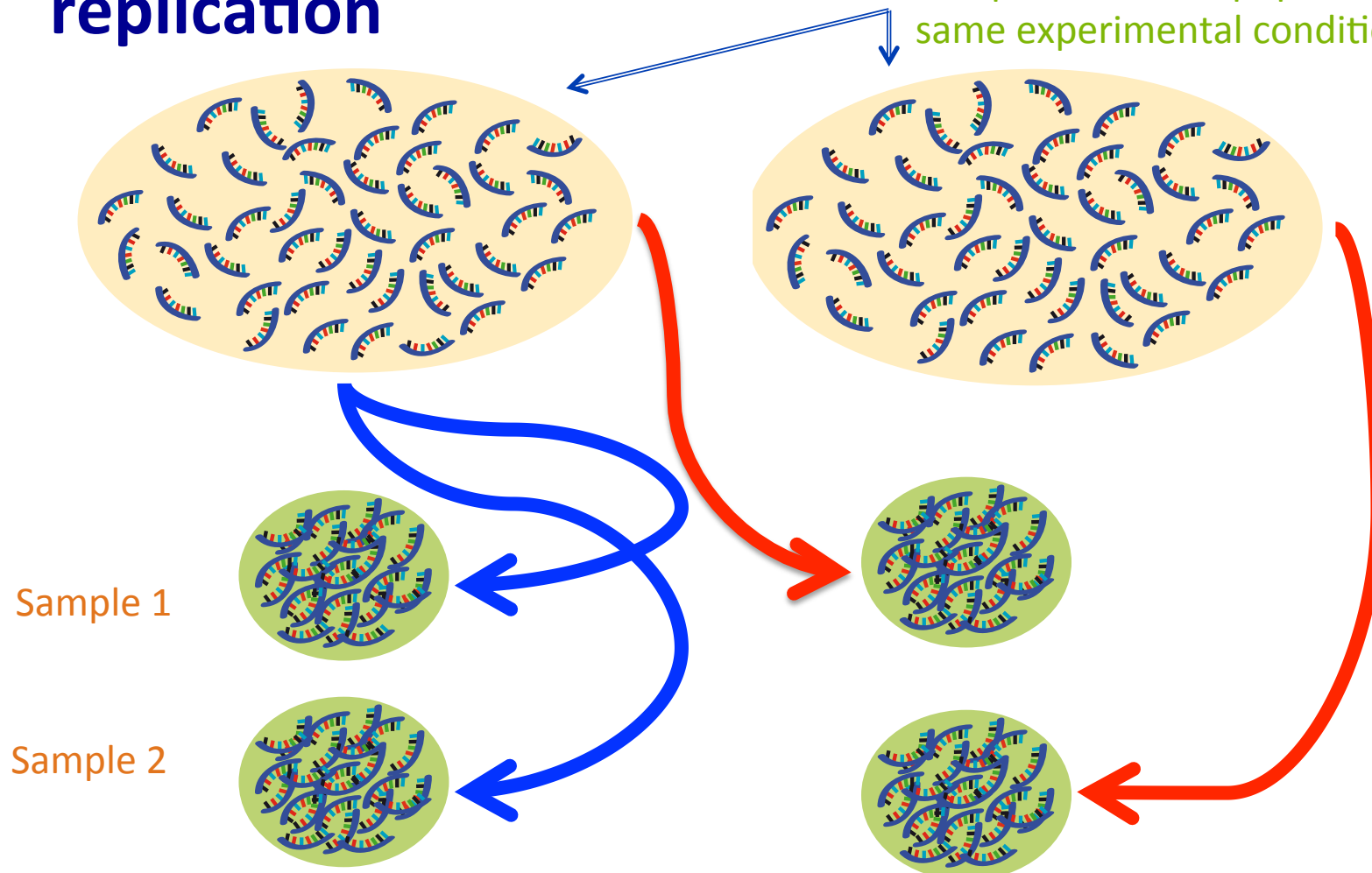
λ_i - proportion

Large M , small $\lambda_i \rightarrow$ approximated well by Poisson($\mu_i = M \cdot \lambda_i$)



Technical replication versus **biological** replication

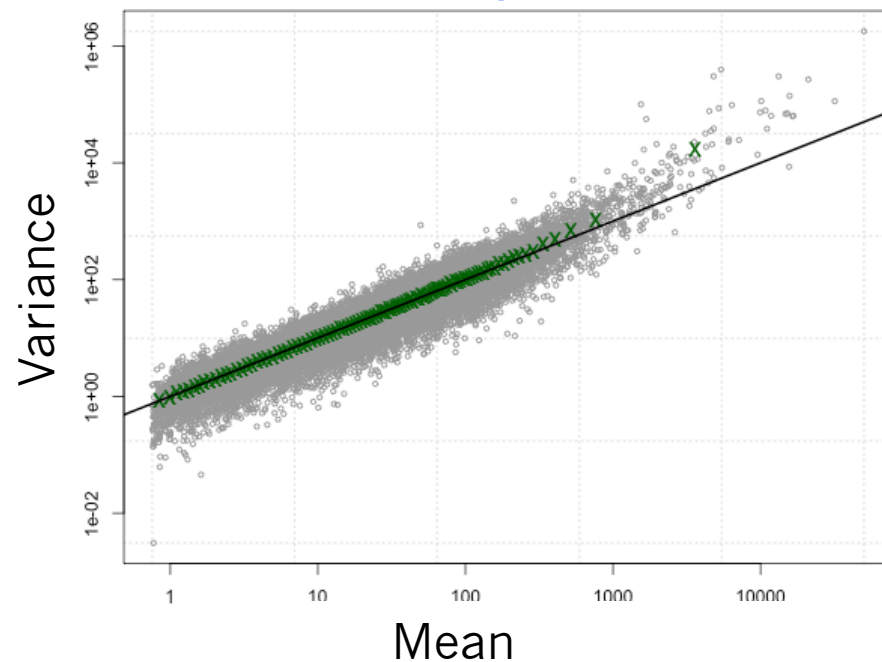
Independent DNA populations from
same experimental condition





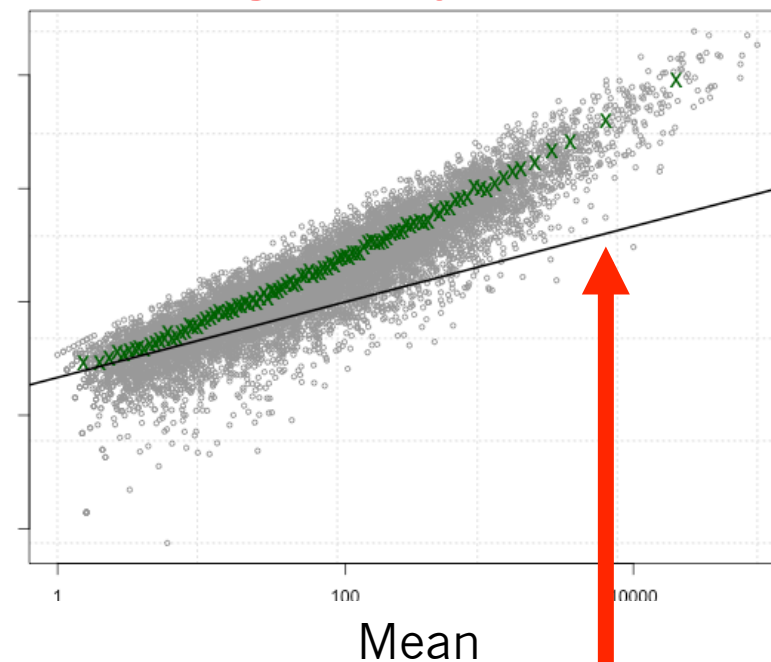
Mean-Variance plots: What we see in real data

Technical replicates



Data from Marioni et al. Genome Research 2008

Biological replicates



Data from Parikh et al.
Genome Biology 2010

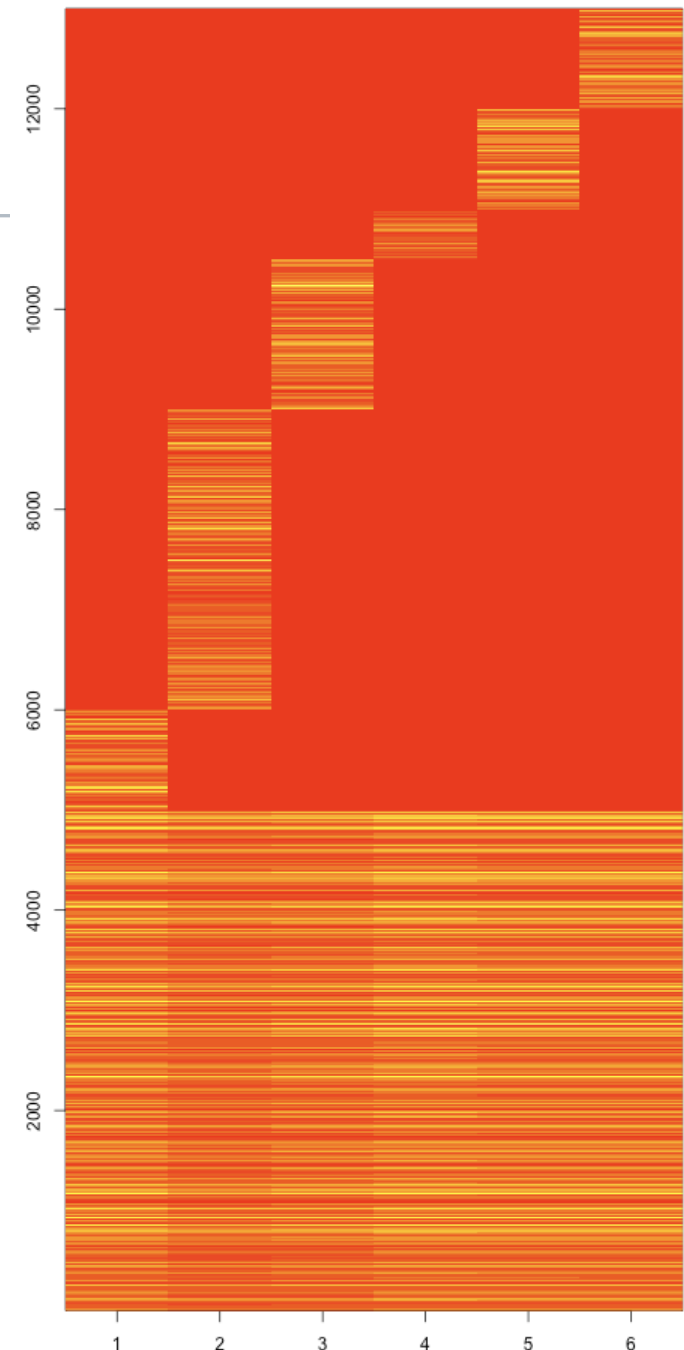
mean=variance
(Poisson assumption)



Normalization: “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





Model assumptions

M_j = library size
 λ_{ij} = relative abundance of
feature i

Poisson describes technical variation:

$$Y_{ij} \sim \text{Pois}(M_j * \lambda_{ij})$$

$$\text{mean}(Y_{ij}) = \text{variance}(Y_{ij}) = M_j * \lambda_{ij}$$

Negative binomial models **biological** variability using the dispersion parameter φ :

$$Y_{ij} \sim \text{NB}(\mu_{ij} = M_j * \lambda_{ij}, \varphi_i)$$

Same mean, variance is quadratic in the mean:

$$\text{variance}(Y_{ij}) = \mu_{ij} (1 + \mu_{ij} \varphi_i)$$

Critical parameter to estimate: dispersion

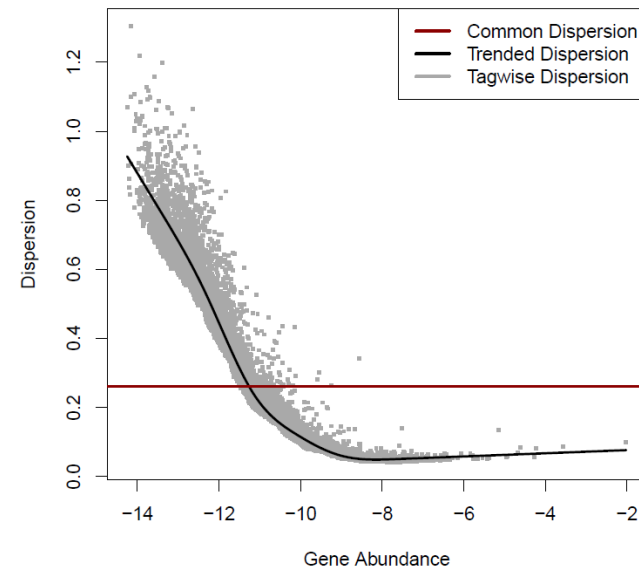
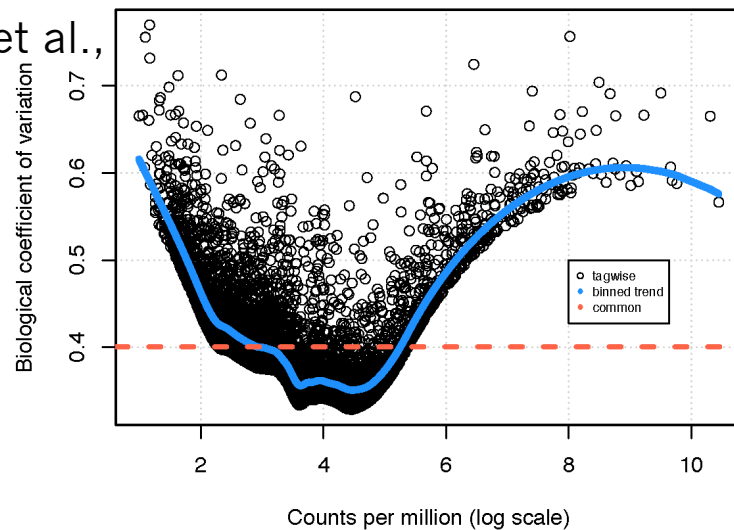


University of
Zurich^{UZH}

Institute of Molecular Life Sciences

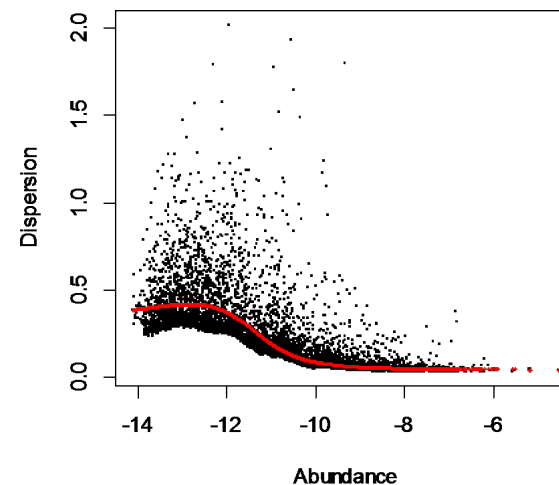
edgeR dispersion estimation: moderate towards trend

Data:
Tuch et al.,
2008



Mouse
hemapoeitic
stem cells,
(Samir Taoudi)

Advantage: share
information, but genes
are allowed to have
their own variance.



Mouse lymphomas
(Stan Lee)

Davis McCarthy



Response: negative binomial with dispersion fixed (to make it in the exponential family).

Link function (relate mean of response to linear combination of parameters)

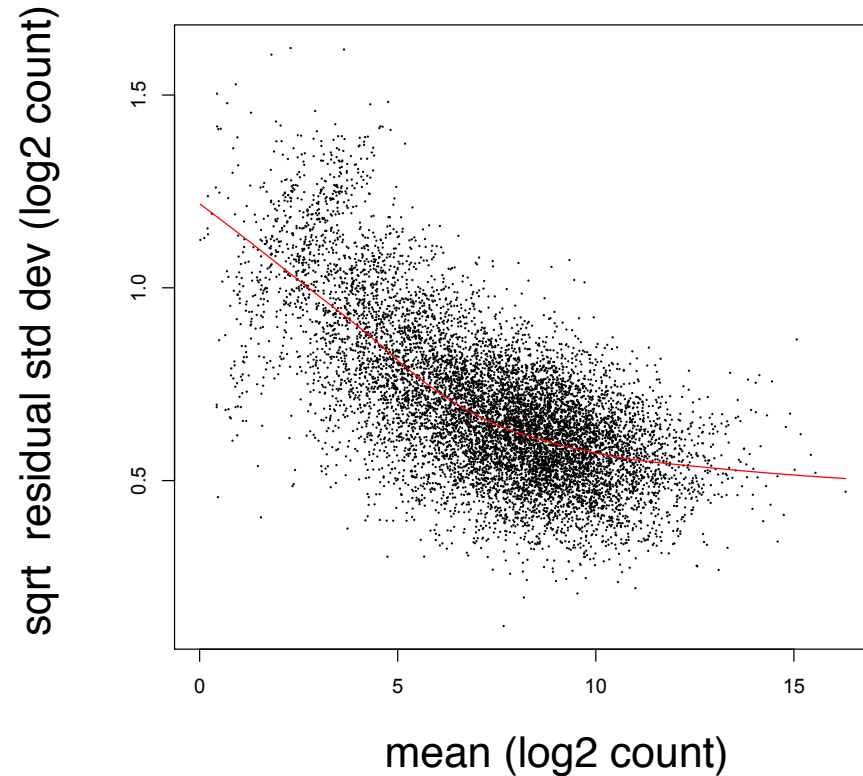
For example:

$Y_i \sim \text{NB}(\mu_i, \phi)$	X	– design matrix
	$\ln()$	– link function
$X\beta = \ln(\mu)$	β	– parameters

Applicability to a wide range of designs



- Converts discrete counts to (log-cpm)
- Removes trend in the variance of counts
- Estimate variances and use inverse as weight



Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Law et al. 2014. Genome Biology. 2014, 15:R29.



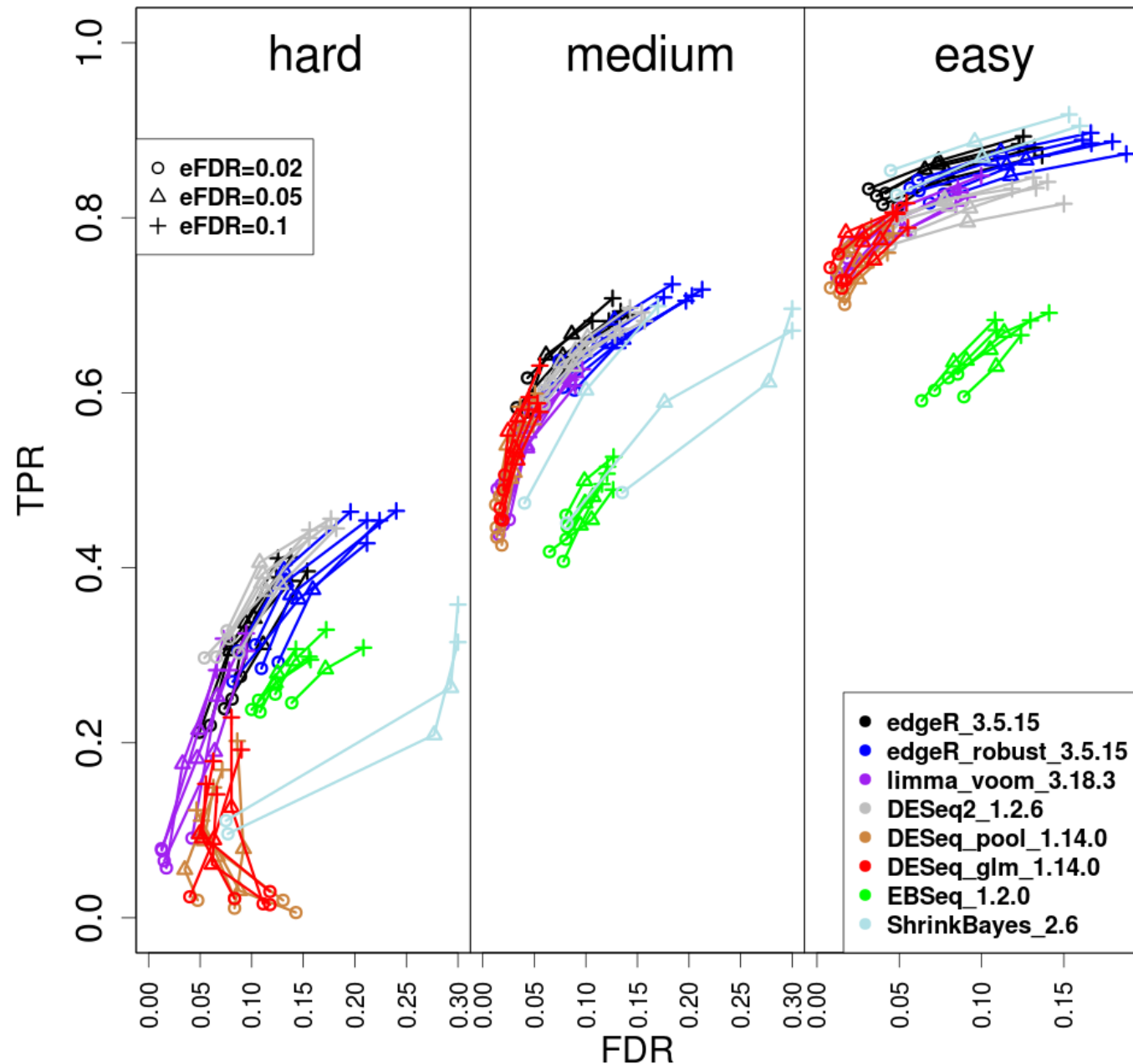
University of
Zurich^{UZH}

Institute of Mole

Simulation-
based
comparisons:
do methods
achieve their
FDRs?

Zhou et al., NAR, 2014

(a) No outliers/pickrell/5vs5





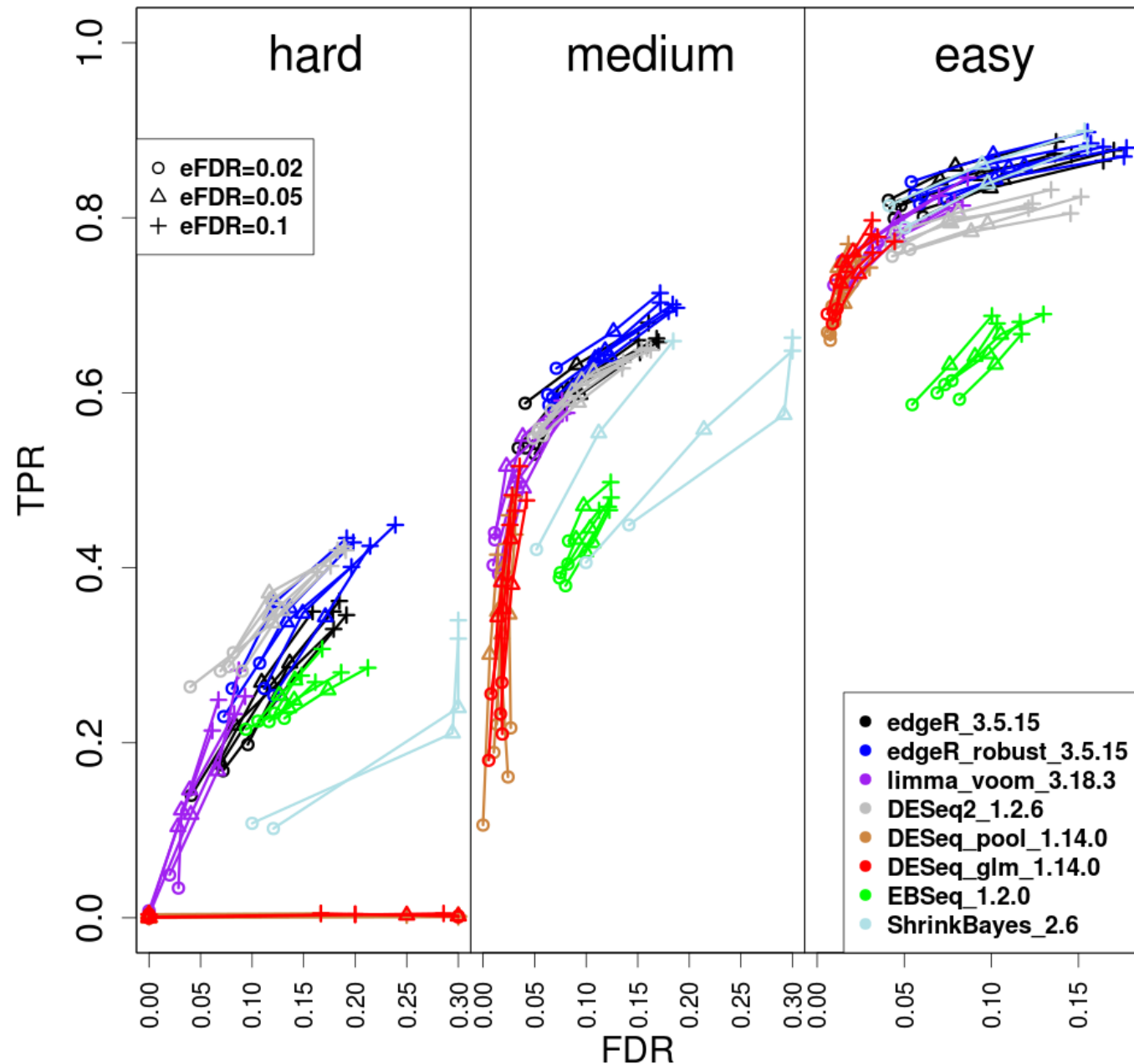
University of
Zurich^{UZH}

Institute of Mole

Simulation-
based
comparisons:
do methods
achieve their
FDRs?

Zhou et al., NAR, 2014

(b) 10% outliers/S/pickrell/5vs5





Beyond differential expression: differential splicing

Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments

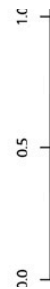
Hugues Richard^{1,*}, Marcel H. Schulz^{1,2}, Marc Sultan³, Asja Nürnberg³, Sabine Schrinner³, Daniela Balzereit³, Emilie Dagand³, Axel Rasche³, Hans Lehrach³, Martin Vingron¹, Stefan A. Haas¹ and Marie-Laure Yaspo³

¹Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73,

²International Max Planck Research School for Computational Biology and Scientific Computing, and

³Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany

Relative exon expression li

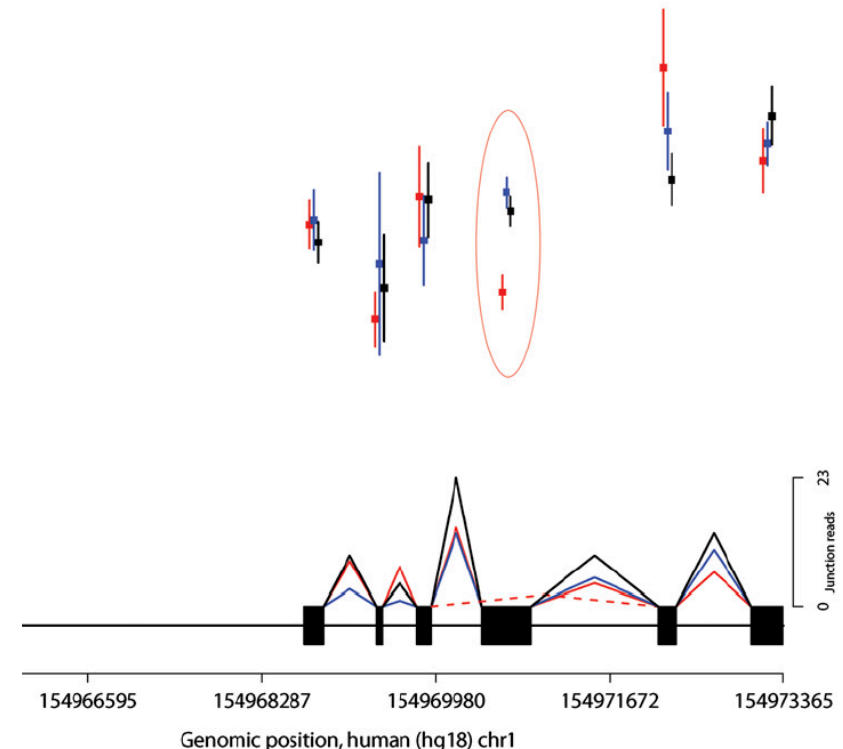


Sex-specific and lineage-specific alternative splicing in primates

Ran Blekman,^{1,4,5} John C. Marioni,^{1,4,5} Paul Zumbo,² Matthew Stephens,^{1,3,5} and Yoav Gilad^{1,5}

¹Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; ²Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; ³Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA

CGI-41





Counting: a few considerations (exon-level)

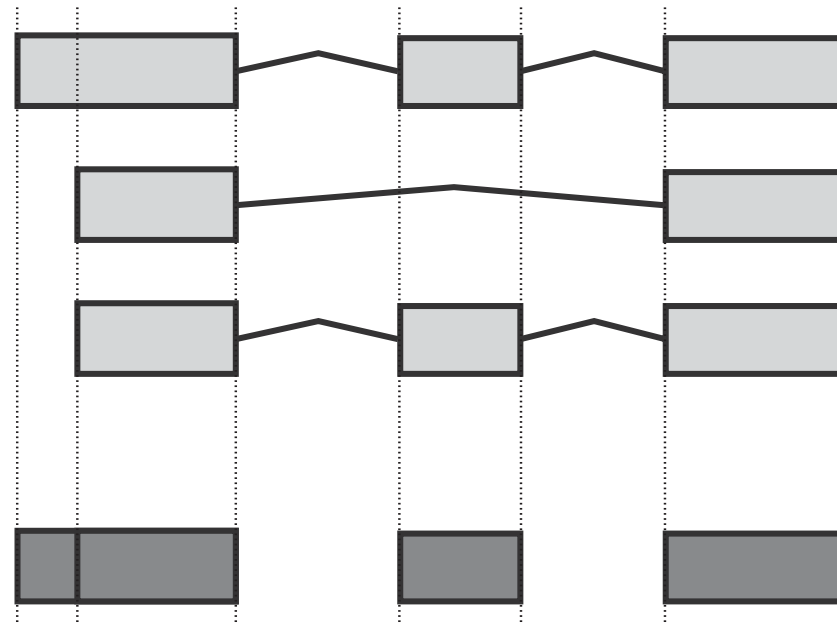


Figure 1. Flattening of gene models: This (fictional) gene has three annotated transcripts involving three exons (light shading), one of which has alternative boundaries. We form counting bins (dark shaded boxes) from the exons as depicted; the exon of variable length gets split into two bins.



DEXSeq – general structure

We use generalized linear models (GLMs) (McCullagh and Nelder 1989) to model read counts. Specifically, we assume K_{ijl} to follow a negative binomial (NB) distribution:

$$K_{ijl} \sim NB\left(\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}\right), \quad (1)$$

where α_{il} is the dispersion parameter (a measure of the distribution's spread; see below) for counting bin (i, l) , and the mean is predicted via a log-linear model as

$$\log \mu_{ijl} = \beta_i^G + \beta_{il}^E + \beta_{i\rho_j}^C + \beta_{i\rho_j l}^{EC}. \quad (2)$$

i – gene

j – sample ... ρ_j is condition (categorical)

l – bin

β_i^G – baseline “expression strength”

β_{il}^E – “exon” (bin) effect

$\beta_{i\rho_j}^C$ – condition effect

$\beta_{i\rho_j l}^{EC}$ – condition x “exon” interaction

Method

Detecting differential usage of exons from RNA-seq data

Simon Anders,^{1,2} Alejandro Reyes,¹ and Wolfgang Huber

European Molecular Biology Laboratory, 69111 Heidelberg, Germany