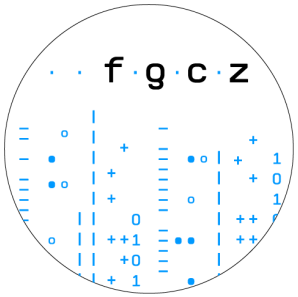
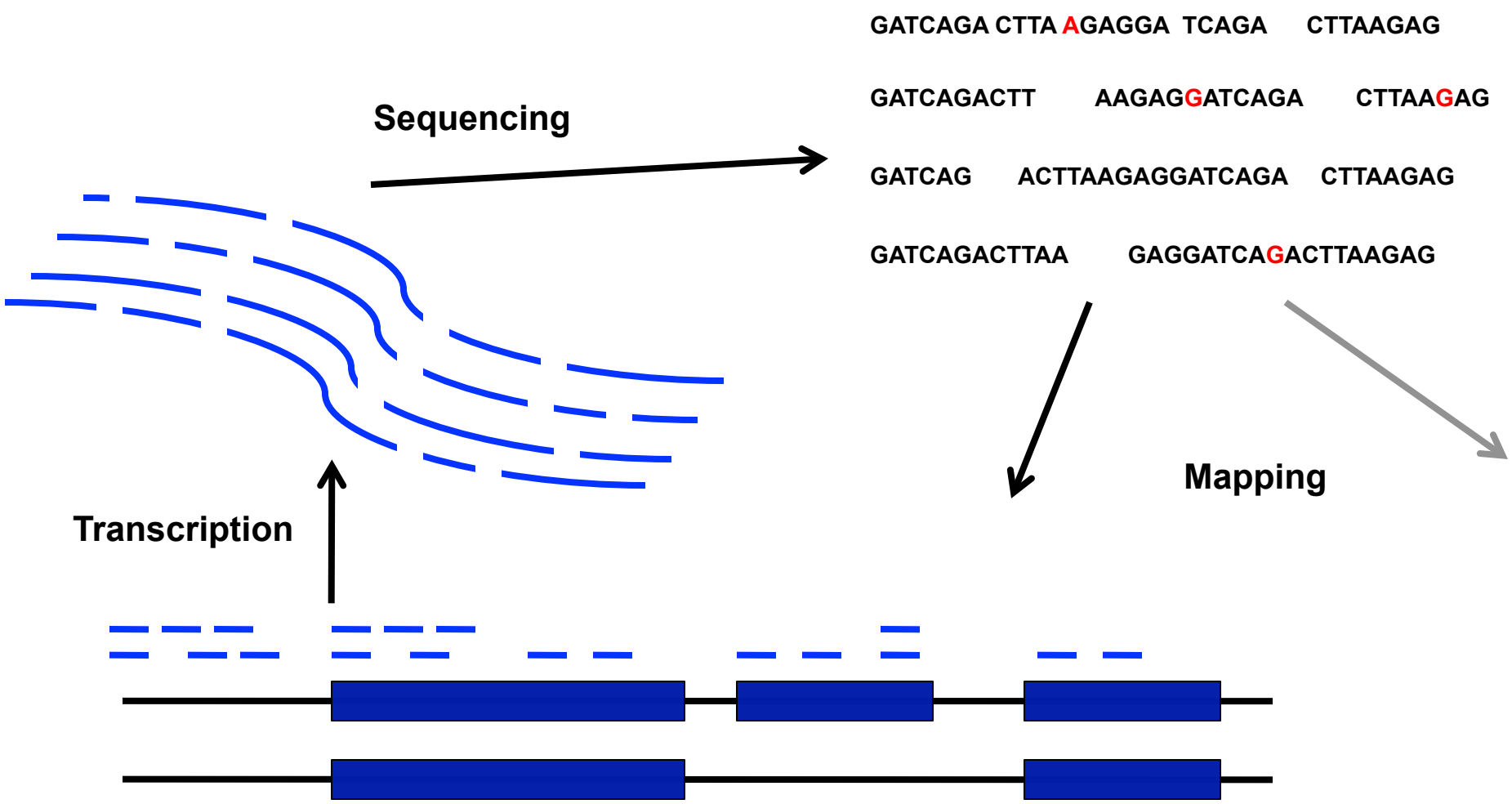
RNA-seq Quantification

Dr. Hubert Rehrauer



Hubert.Rehrauer@fgcz.ethz.ch

RNA-seq isoform quantification problem: How many transcripts?



RNA-seq comes with absolute counts but relative abundances

Gene	Sample 1 [Mio transcripts]	Sample 1 [Mio sequenced reads]	Sample 2 [Mio transcripts]	Sample 2 [Mio sequenced reads]
gene a	10	0.5	10	0.2
gene b	10	0.5	10	0.2
gene c	10	0.5	10	0.2
gene d	10	0.5	10	0.2
gene e	160	8.0	460	9.2
total	**200**	**10**	**500**	**10**

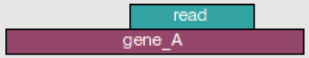
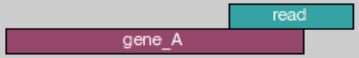
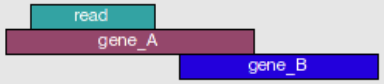
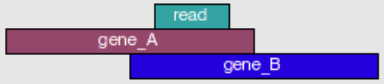
With RNA-seq different amounts of starting material will give the identical numbers of reads!
The read count for a gene is always relative to the counts for the other genes.

Abundance estimates

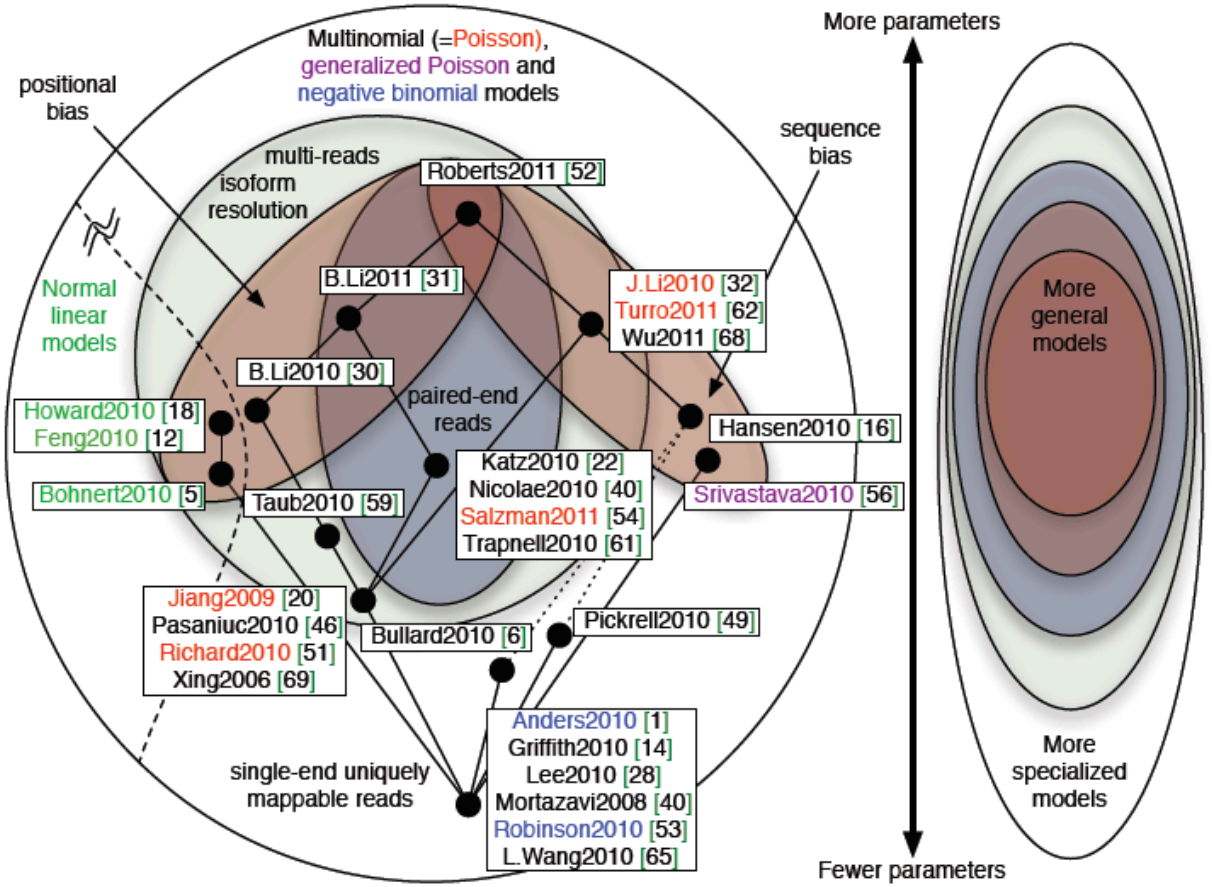
Abundance of what???

- Biologically relevant:
  - **gene level:**
    - # molecules transcribed from one gene locus
  - **isoform level:**
    - # molecules of a specific isoform transcribed from one gene
- Feasible with RNA-seq:
  - **relative fractions**
- Really easy to get but not that useful:
  - # reads that uniquely map to a gene locus
    → biased by length, discards information in multi-mappers
  - #reads that map to gene locus (including multi-mappers)
    → disambiguation is not possible if you do not have abundance estimates of the isoforms

Model-free Counting of Overlapping reads – Count Modes

	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

Model Hierarchy



Pachter 2011

RNA-seq model

$$\alpha_t = \mathrm{P}[\text{read from transcript } t] = \frac{1}{Z}\rho_t l_t$$

with:

ρ_t expression level / abundance / fraction

l_t transcript length

$Z = \sum_t \rho_t l_t$ normalization factor

The normalization factor is the weighted mean length of the transcripts.

RNA-seq model

Estimation of the probability that a read is from a specific transcript:

$$\hat{\alpha}_t = \frac{X_t}{N} = \frac{\#\text{reads mapping to transcript } t}{\#\text{mappable reads in total}}$$

Abundance estimates:

$$\hat{\rho}_t \propto \frac{\hat{\alpha}_t}{l_t}$$

Definition of expression levels

- Reads Per Kilobase per Million of mapped reads

$$\text{RPKM for transcript } t = 10^6 \times 10^3 \times \frac{X_t}{l_t N}$$

- Transcripts Per Million Transcripts

$$\text{TPM for transcript } t = 10^6 \times Z \times \frac{X_t}{l_t N}$$

- Preferable is TPM because it is independent of the transcriptome

Maximum Likelihood Estimation

- The estimated abundances represent unique MLE estimates

$$\text{with } \alpha = \{\alpha_t\}_{t \in T}$$

$$L[\alpha] = \prod_{t \in T} \prod_{f \in F_t} P[f \in t] \frac{1}{l_t}$$

$$= \prod_{t \in T} \prod_{f \in F_t} \alpha_t \frac{1}{l_t}$$

$$= \prod_{t \in T} \left(\frac{\alpha_t}{l_t} \right)^{X_t}$$

Effective Transcript Length

- Since fragments have a non-zero length the read probabilities depend actually on an effective length:

$$l_t := \text{transcript length - fragment length + 1}$$

- For simplicity we continue to use the symbol without tilde but will always assume it is the effective length
- The effective length represents the stretch of the transcript from which I can get a fragment that I can then map back to the transcript
- → The effective length must also consider mappability!
- → Mappability does depend on mapping algorithm, mutations, …

Multi-reads

- Reads that cannot be uniquely assigned to one transcript were ignored so far
- Multi-reads can occur
  - if a read aligns more than once in the genome
  - if at an alignment position there is more than one transcript defined
- Multi-reads do occur due to homology not due to pure chance

Considering Multi-reads

- Define a compatibility matrix

$$\mathbf{Y} = \left\{ y_{ft} \right\}_{f \in F, t \in T}$$

with

$$y_{ft} = \begin{cases} 1 \text{ if read } f \text{ aligns to transcript } t \\ 0 \text{ else} \end{cases}$$

- The likelihood is now:

$$\mathrm{L}[\alpha] = \prod_{f} \left(\sum_{t} y_{ft} \frac{\alpha_t}{l_t} \right)$$

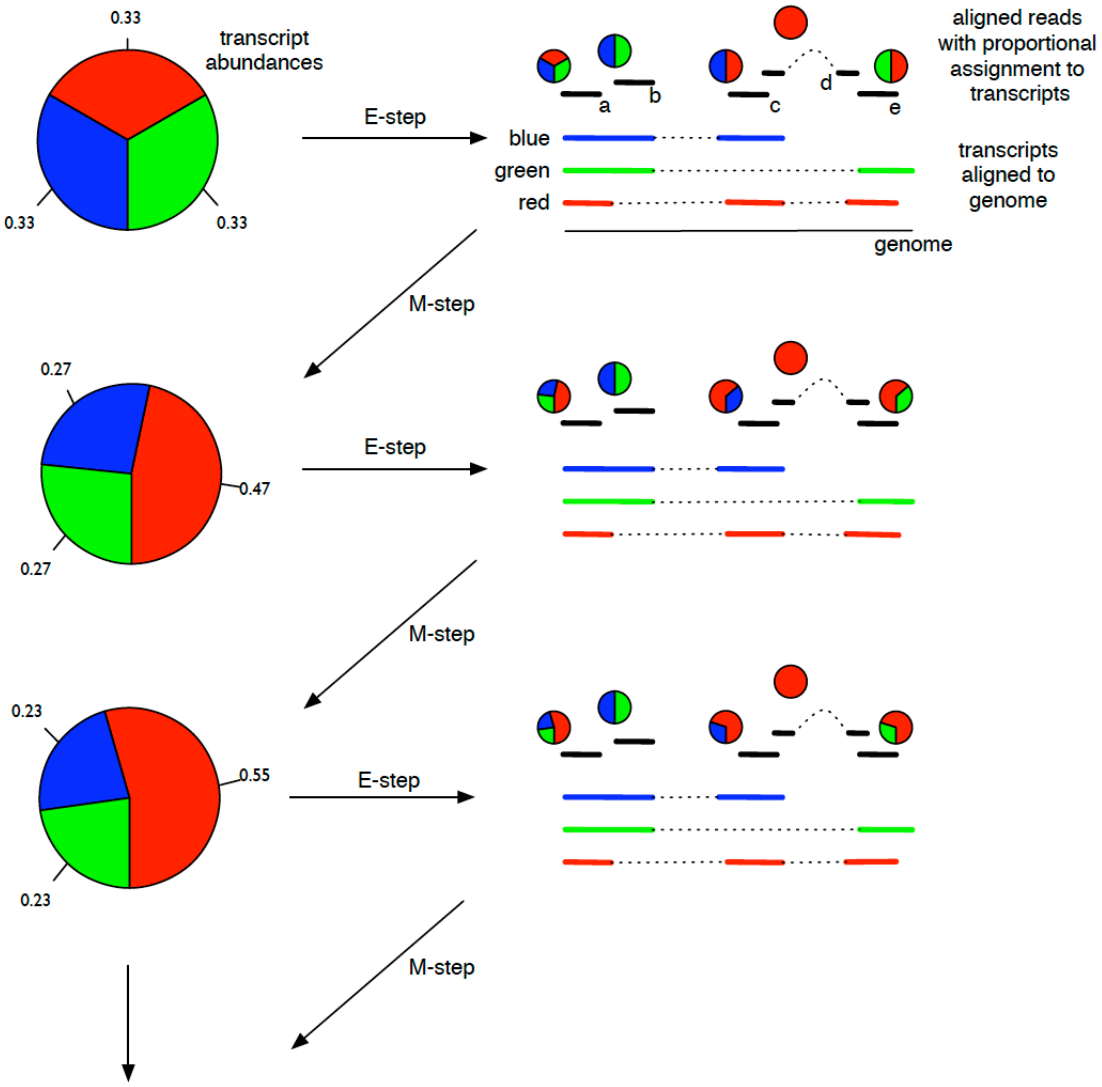
- but now abundances have to be estimated iteratively

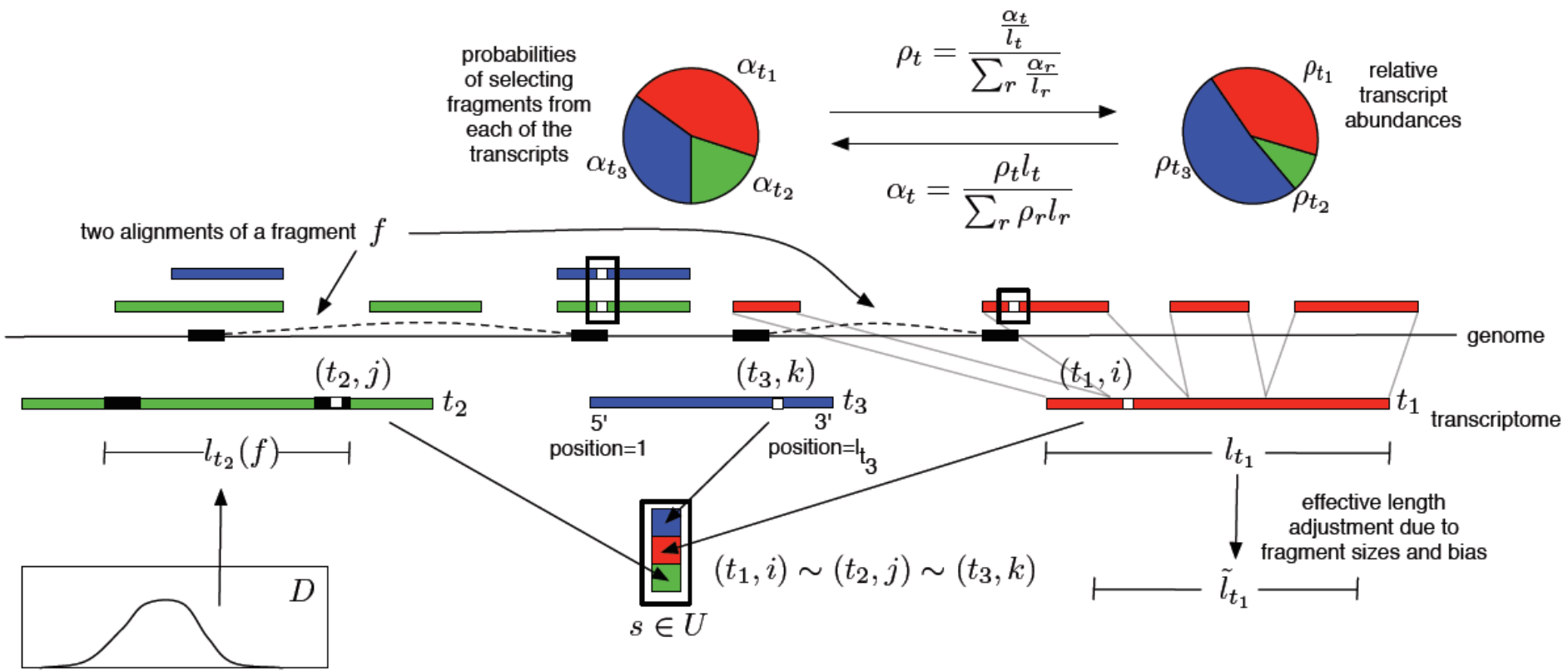
Iterative Estimation

Three step algorithm

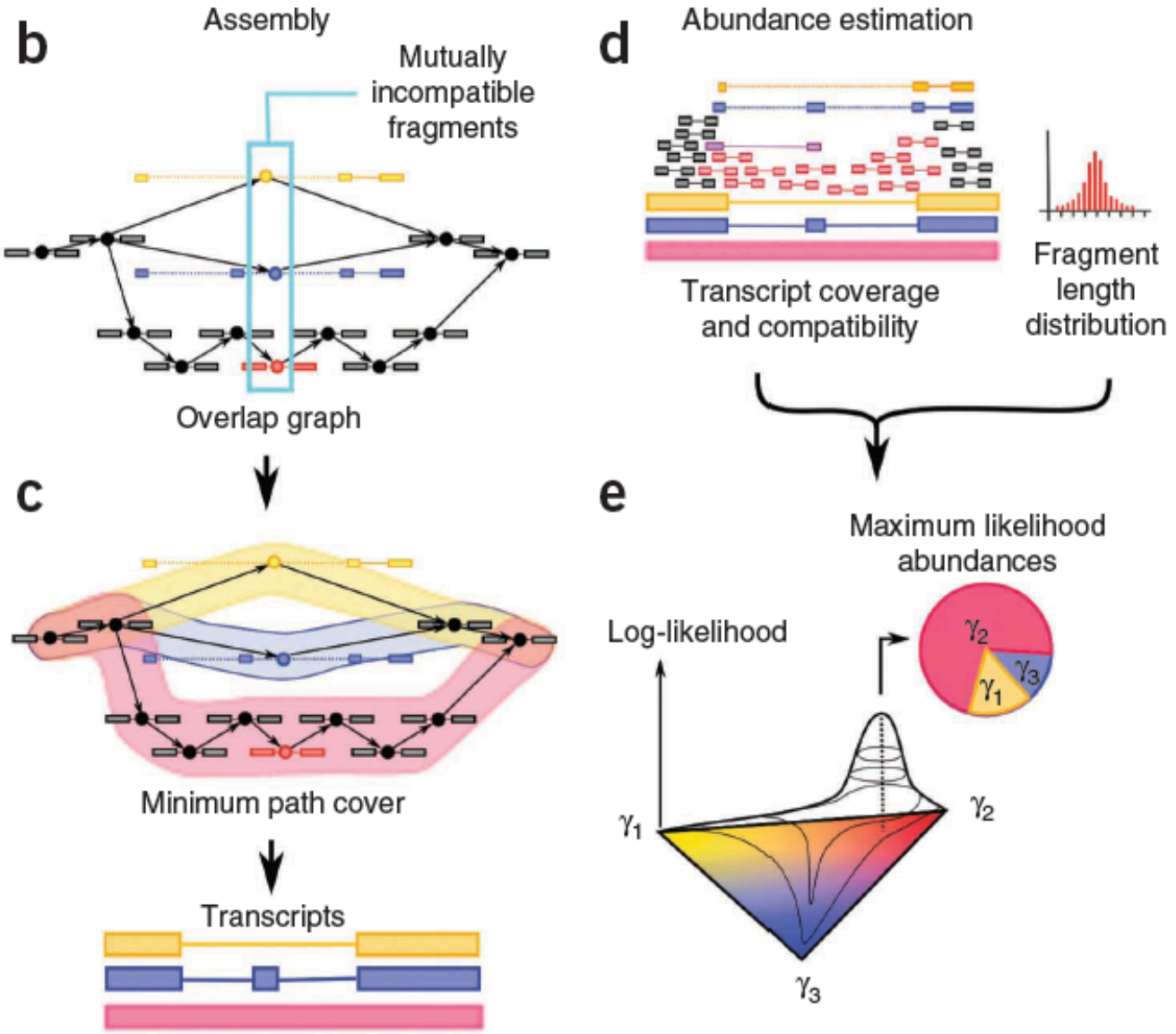
1. Estimate abundances based on uniquely mapping reads only
2. For each multi-read, divide it between the transcripts to which it maps, proportionally to their abundances estimated in the first step
3. Recompute abundances based on updated counts for each transcript
4. Continue with Step 2

Expectation-Maximization Estimation




probabilities of selecting fragments from each of the transcripts
α_{t_1}
α_{t_3}
α_{t_2}
$\rho_t = \frac{\frac{\alpha_t}{l_t}}{\sum_r \frac{\alpha_r}{l_r}}$
$\alpha_t = \frac{\rho_t l_t}{\sum_r \rho_r l_r}$
ρ_{t_1}
ρ_{t_3}
ρ_{t_2}
relative transcript abundances
two alignments of a fragment f
genome
(t_2, j)
t_2
(t_3, k)
t_3
(t_1, i)
t_1
transcriptome
5'
position=1
3'
position=l_{t_3}
$l_{t_2}(f)$
l_{t_1}
effective length adjustment due to fragment sizes and bias
D
$(t_1, i) \sim (t_2, j) \sim (t_3, k)$
$s \in U$
$\tilde{l}_{t_1}$

Transcript abundance estimation with Cufflinks



Trapnell 2010, Nature Biotechnology

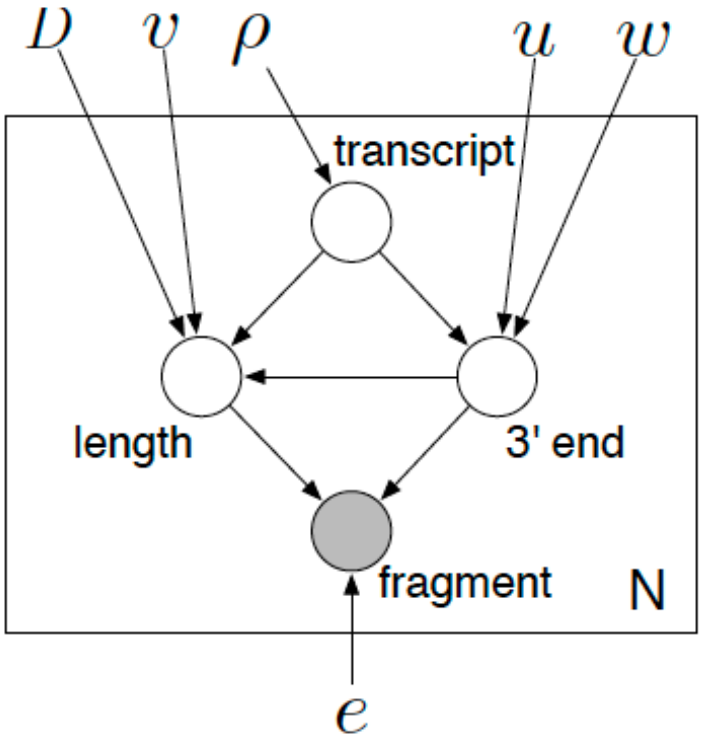
General Formulation of Abundance Estimation

A full model for the abundance estimation should consider:

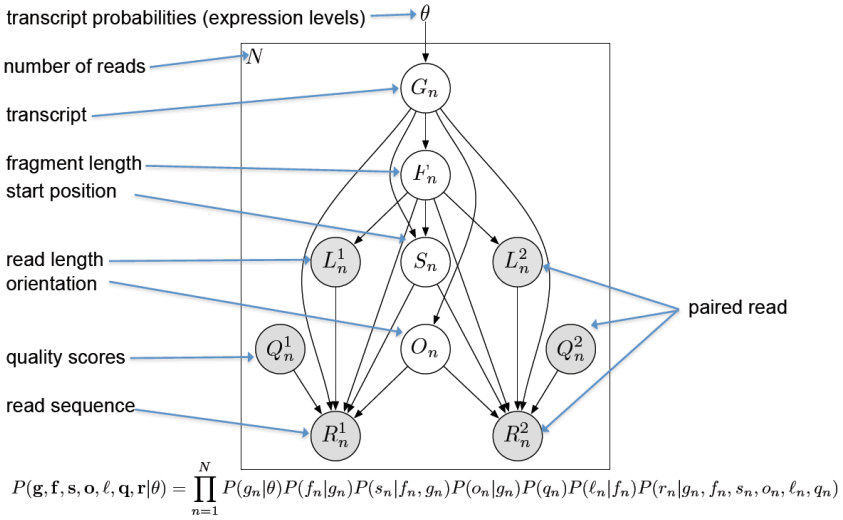
- position bias
- fragment-length distribution
- sequencing errors
- site-specific bias
- …

Example Implementations

Pachter: Cufflinks



Dewey: RSEM



Implementations of Generative Models

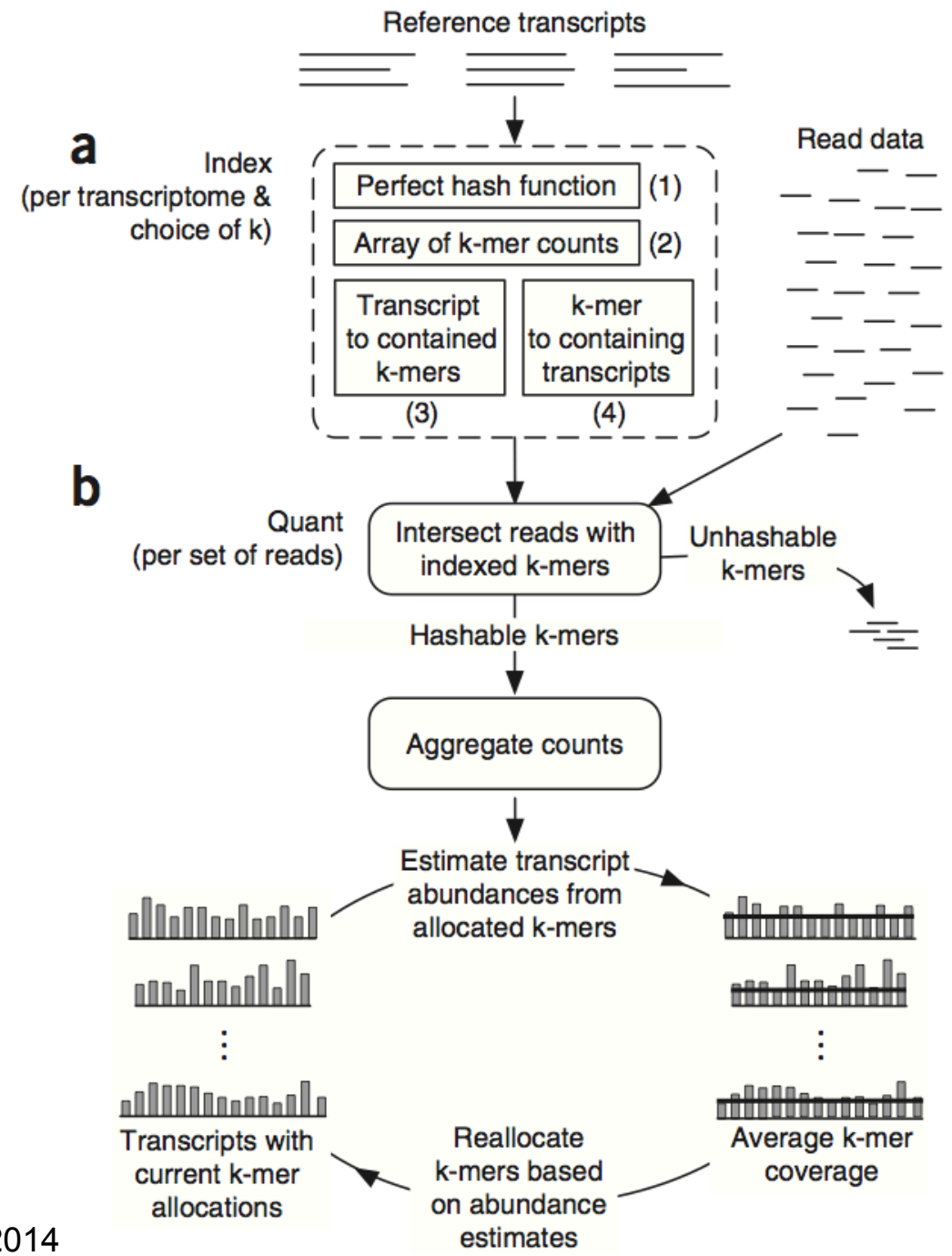
- RSEM
- Cufflinks
- NSMAP
- IsoEM
- MISO
- MMSEQ

Fast approaches to get the Read-Transcript Compatibility Matrix

- Sailfish: lightweight alignment
- Salmon: improvement of sailfish
- kallisto: pseudo-alignments

Sailfish

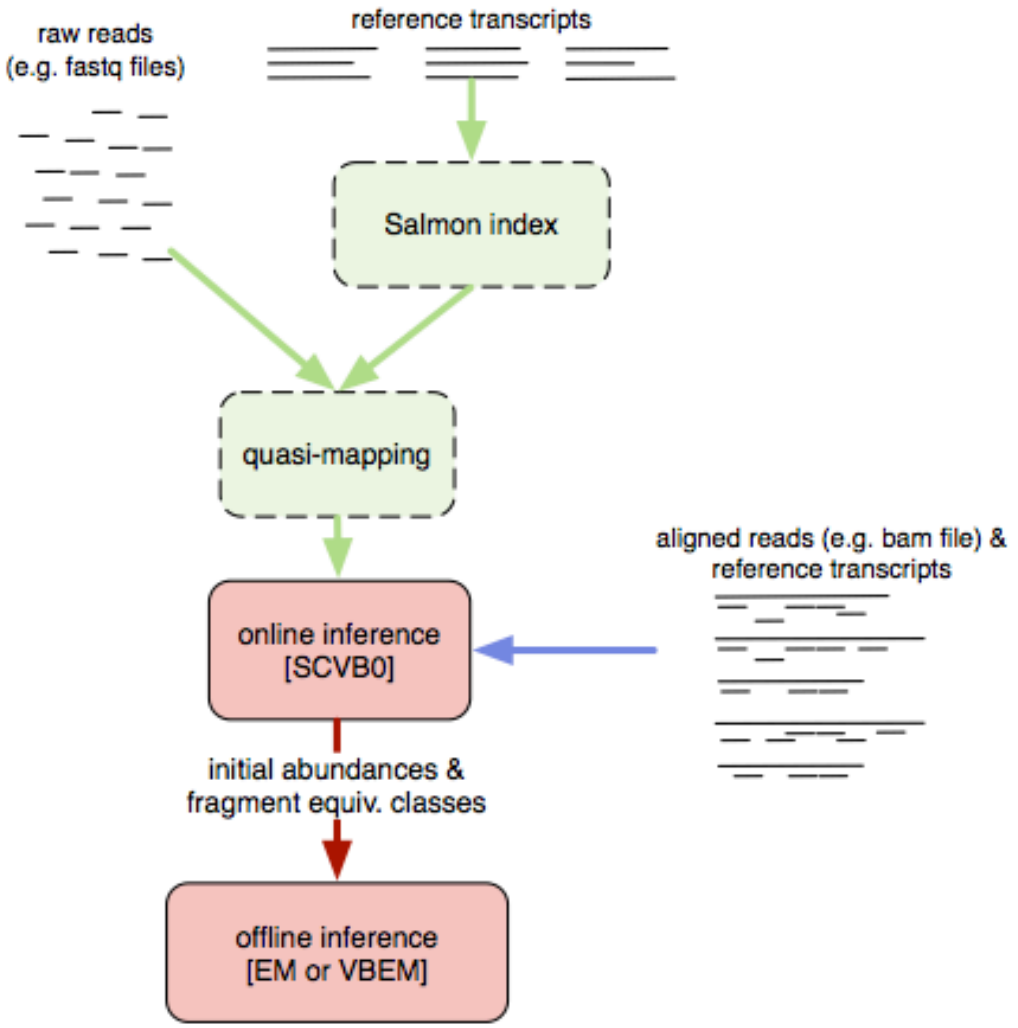
- No read alignment only k-mer lookup (very fast)
- Iterative resolution of ambiguous k-mers
- Original version treated k-mers of a read as independent



Patro et al. Sailfish …. Nature Biotechnology, 2014

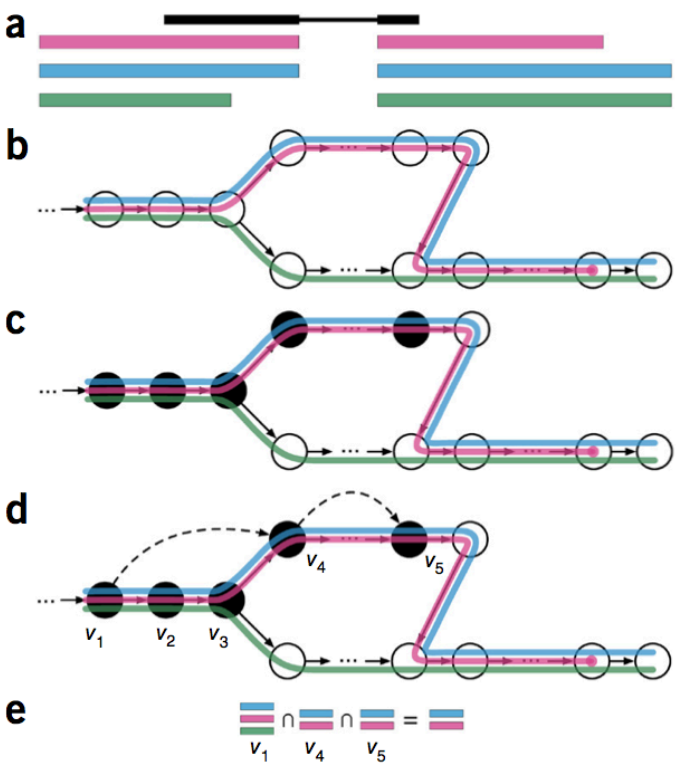
http://www.bluespheremedia.com/2012/03/isla-mujeres-sailfish-2012/

Salmon


raw reads
(e.g. fastq files)
reference transcripts
Salmon index
quasi-mapping
aligned reads (e.g. bam file) &
reference transcripts
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
offline inference
[EM or VBEM]

Quantification with pseudo-alignments

- Instead of hashing the transcriptome build a de Bruijn graph
- Find k-mer hits in the de Bruijn graph
- Identifies only transcripts that are consistent with all k-mer hits



Bray et al. Nature Biotechnology, 2016

Performance comparison

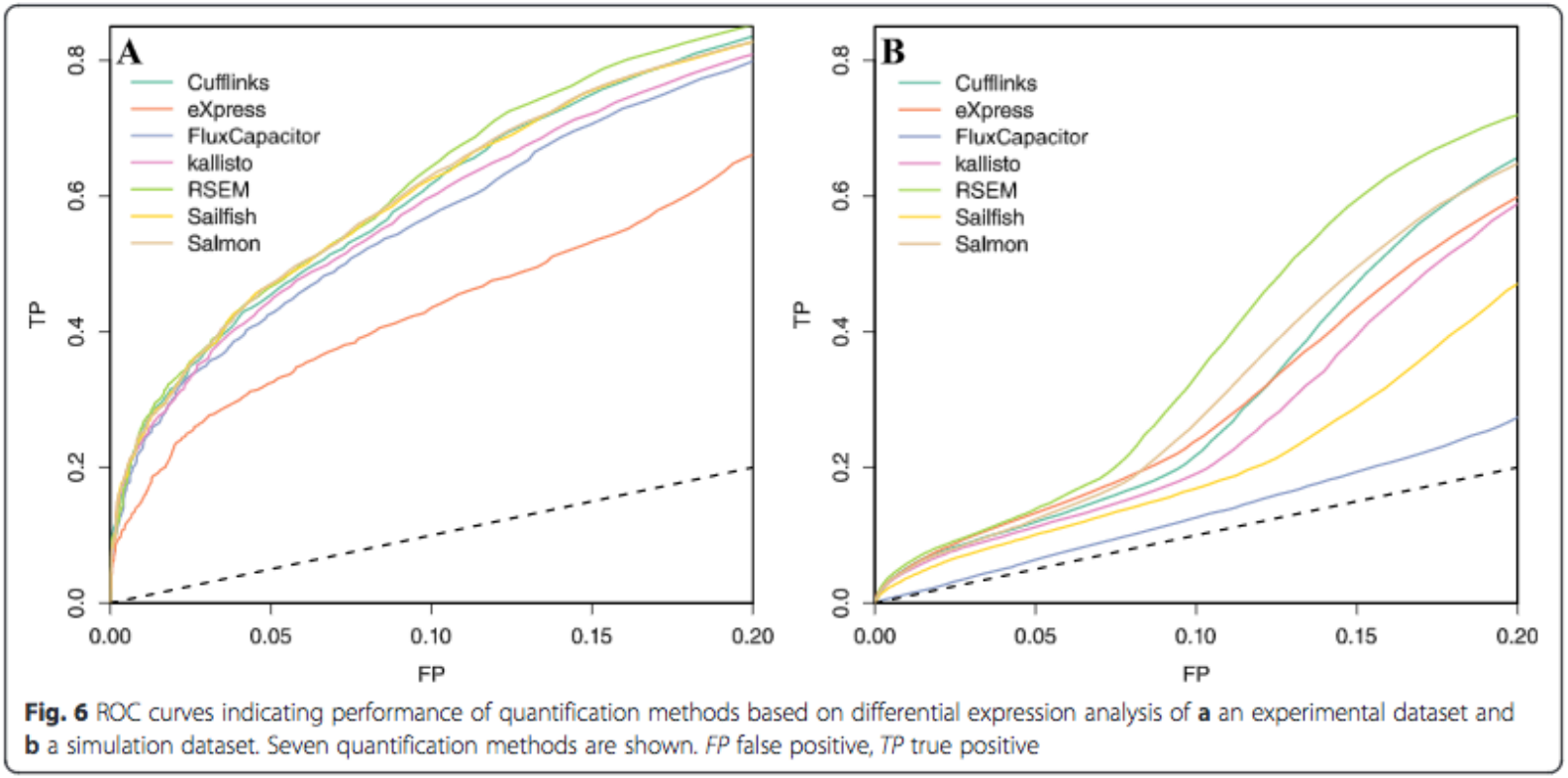
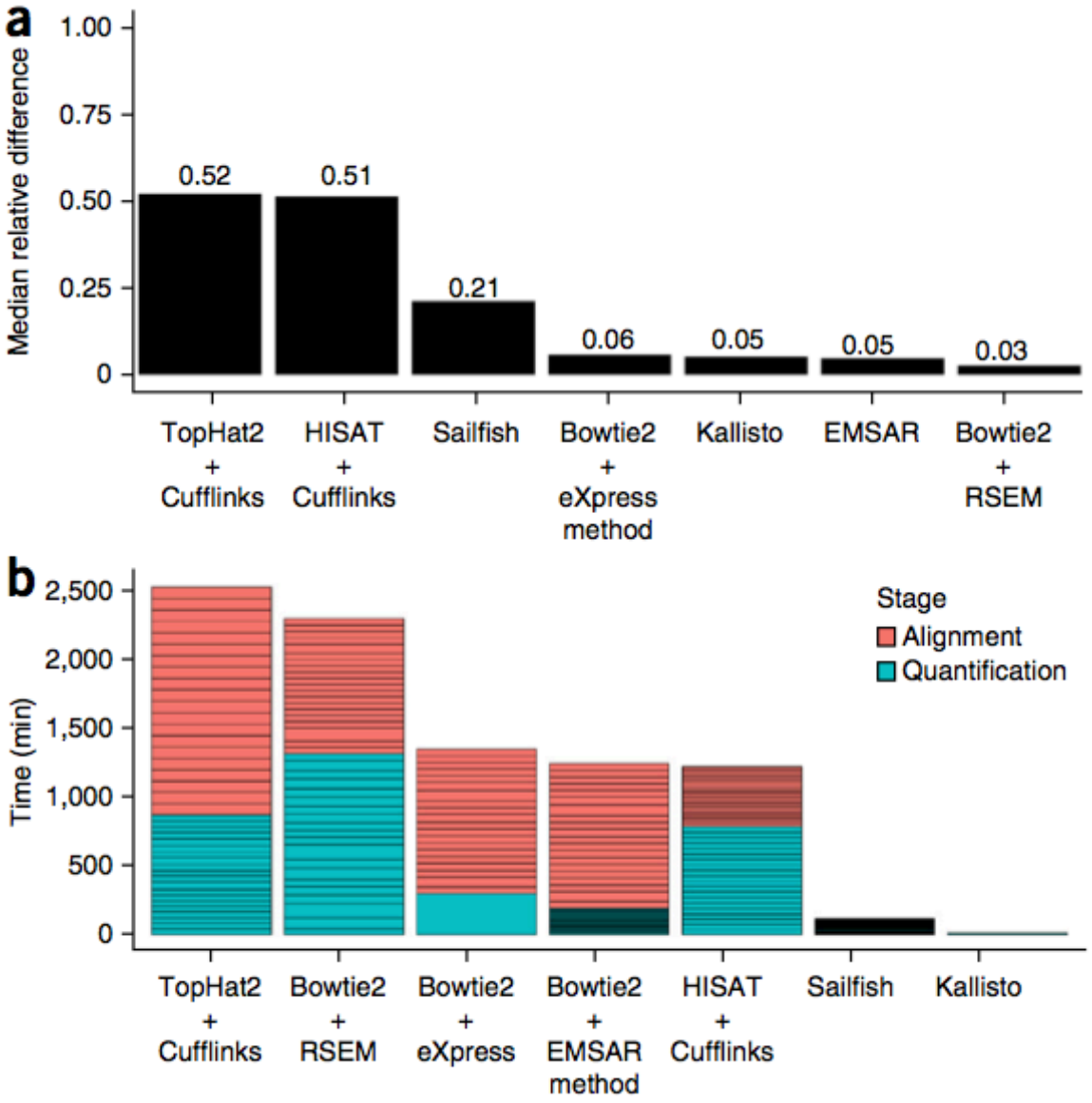


Fig. 6 ROC curves indicating performance of quantification methods based on differential expression analysis of **a** an experimental dataset and **b** a simulation dataset. Seven quantification methods are shown. *FP* false positive, *TP* true positive

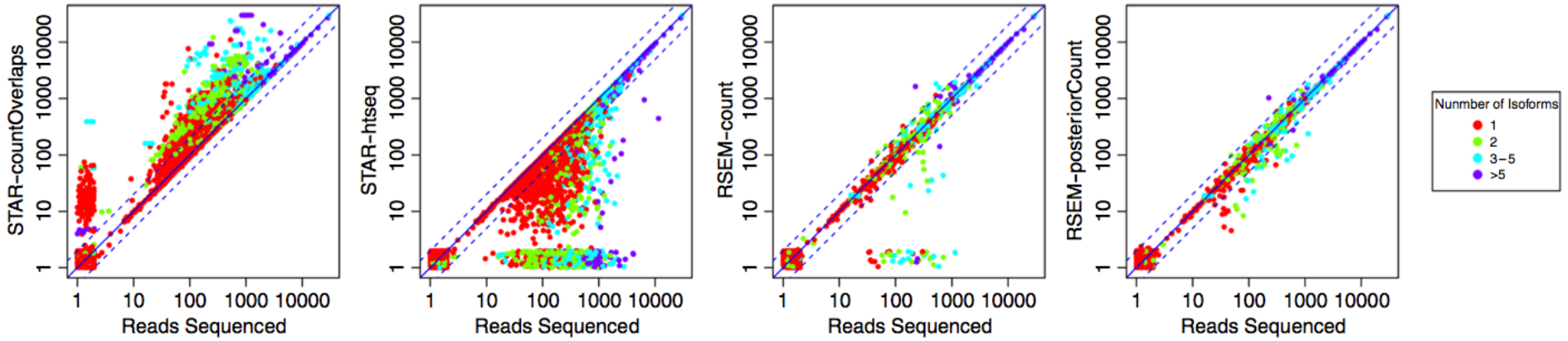
Teng et al. Genome Biology, 2016

Performance Comparison



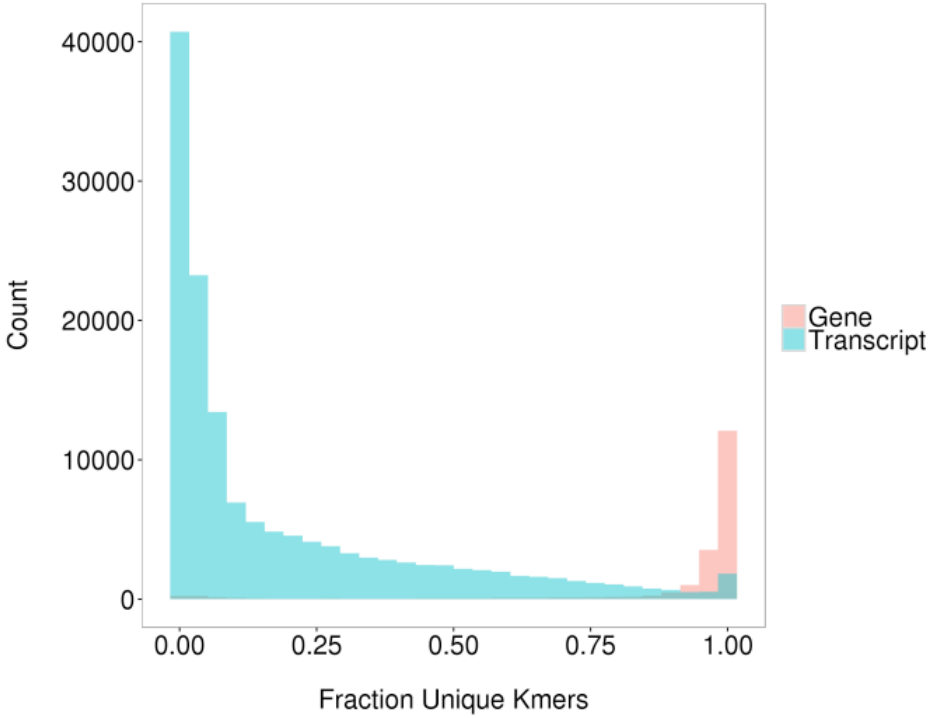
Bray et al. Nature Biotechnology, 2016

Read Counting Accuracy

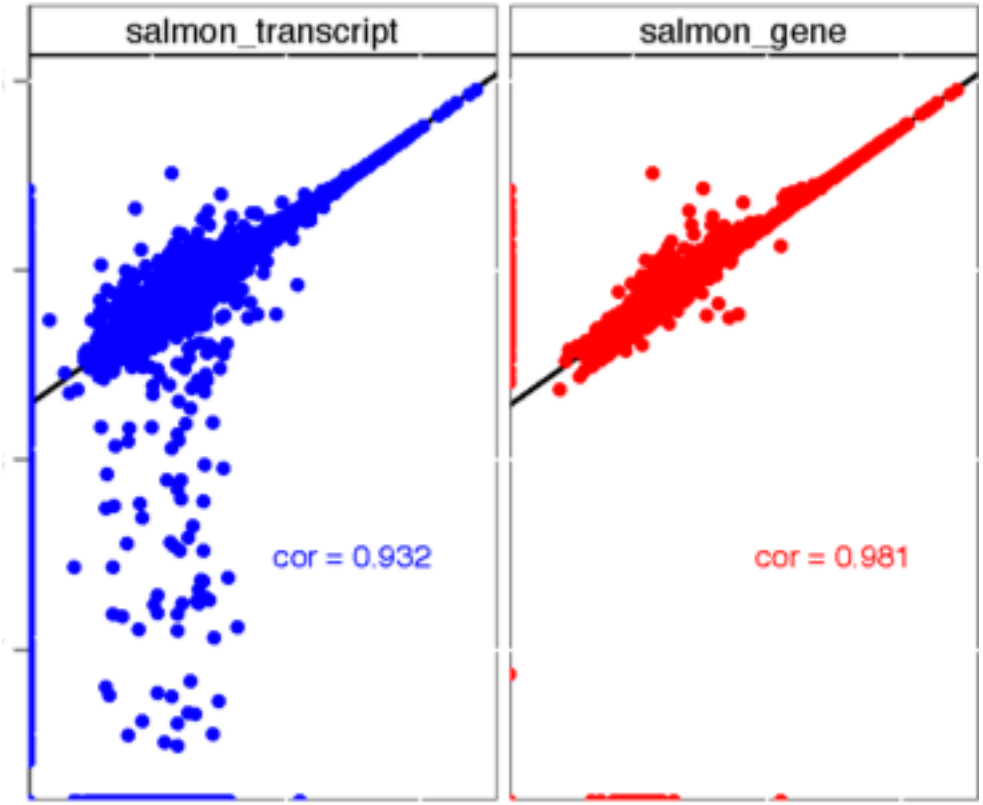


Rehrauer et al.BMC Bioinformatics, 2016

Uniqueness: Isoform-level vs gene-level

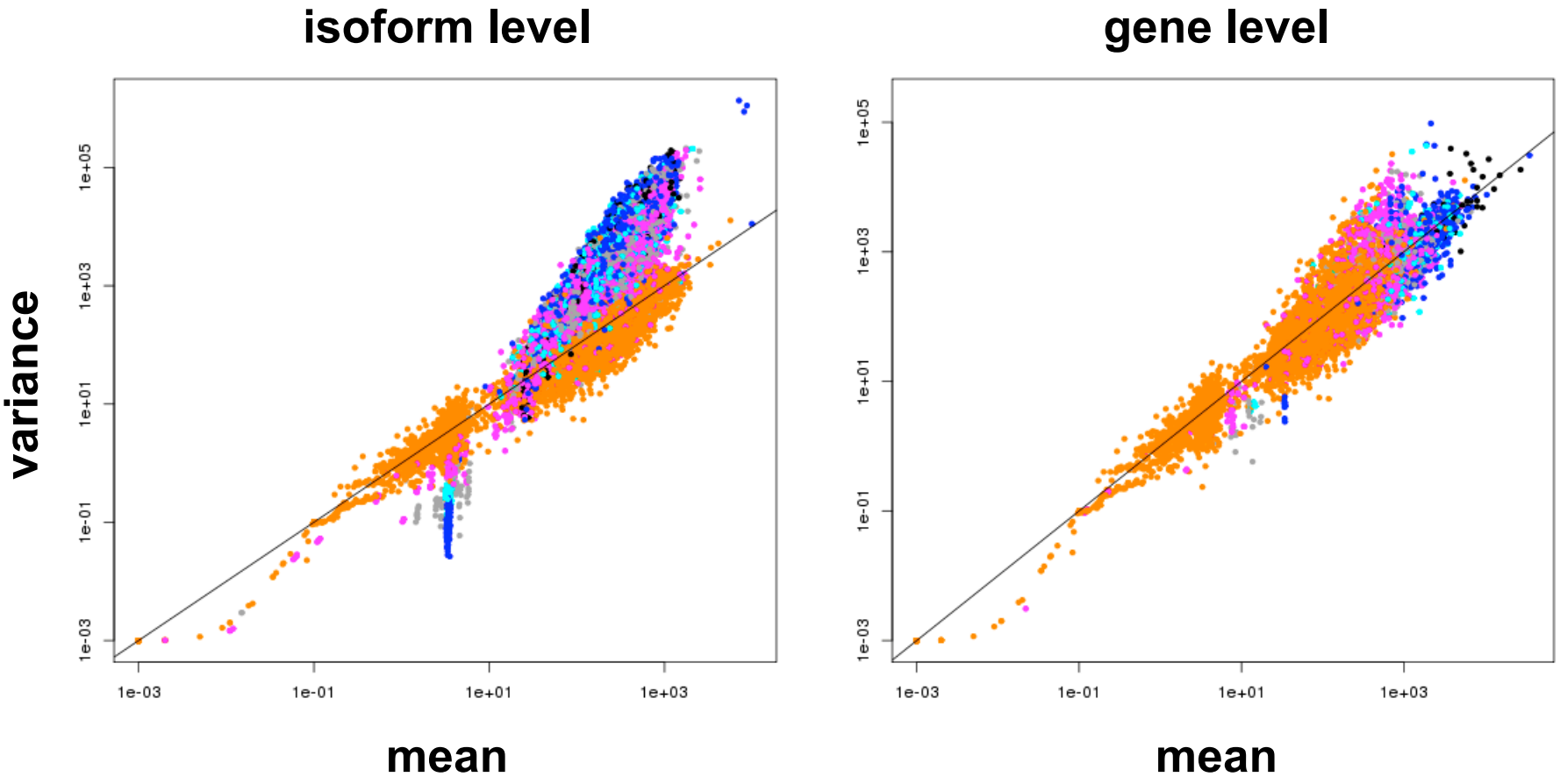


Accuracy: Isoform-level vs gene-level



Soneson et al. F1000, 2015

Isoform level has higher variability



Cufflinks and Related

- Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889* (2011).
- Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation** *Nature Biotechnology* doi: 10.1038/nbt.1621
- Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. **Improving RNA-Seq expression estimates by correcting for fragment bias** *Genome Biology* doi:10.1186/gb-2011-12-3-r22
- Roberts A, Pimentel H, Trapnell C, Pachter L. **Identification of novel transcripts in annotated genomes using RNA-Seq** *Bioinformatics* doi:10.1093/bioinformatics/btr355

- **RSEM:**
  Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
- **MISO:**
  Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7**, 1009–1015 (2010)
- **MMSEQ:**
  Turro, E. *et al.* Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol* **12**, R13 (2011).
- **NSMAP:**
  Xia, Z., Wen, J., Chang, C.-C. & Zhou, X. NSMAP: a method for spliced isoforms identification and quantification from RNA-Seq. *BMC Bioinformatics* **12**, 162 (2011).