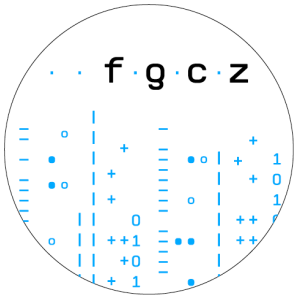
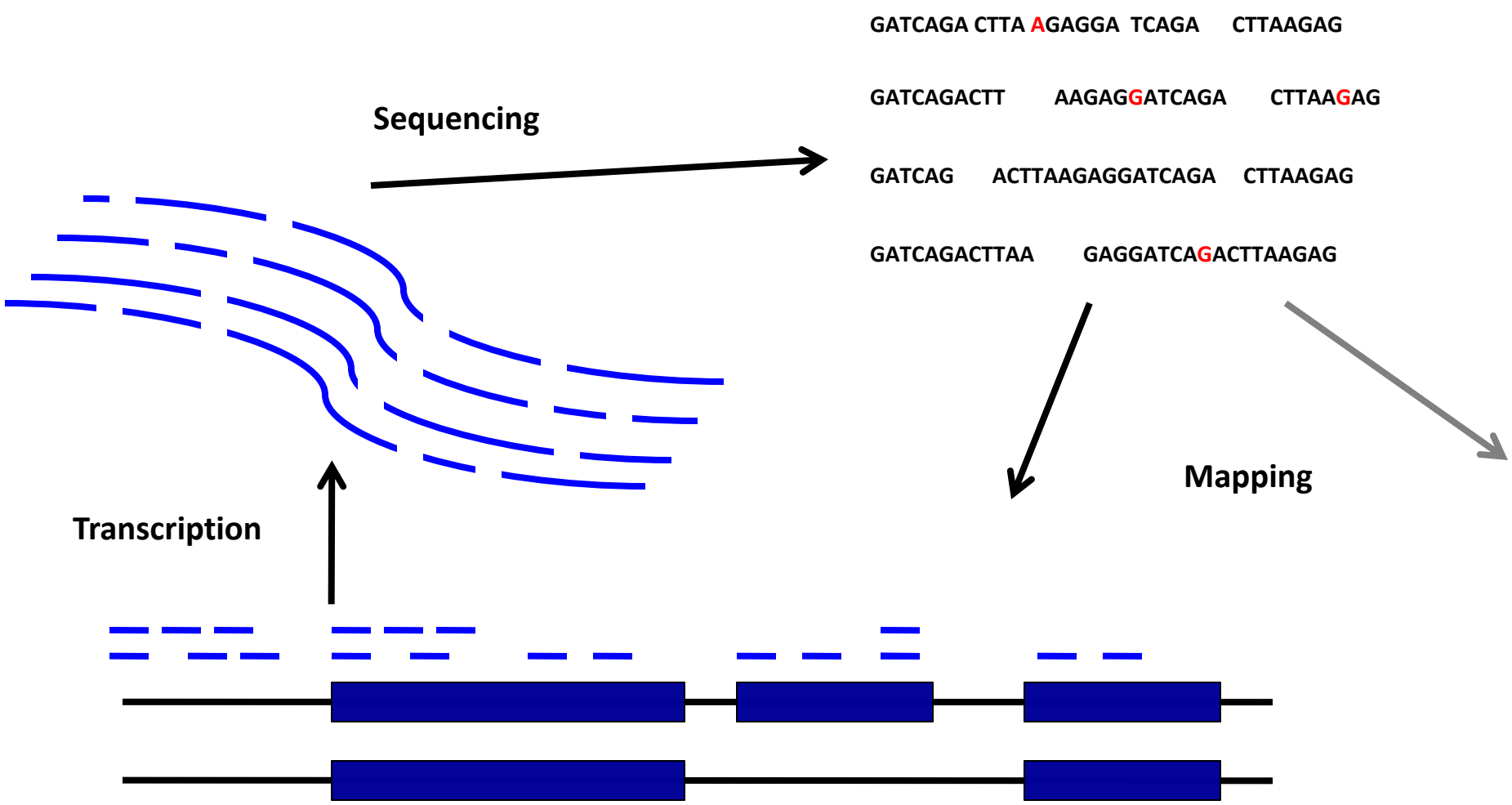
RNA-seq Quantification

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 (per cell)
  - **isoform level:**
    - # molecules of a specific isoform transcribed from one gene (per cell)
- Feasible with RNA-seq:
  - **relative fractions what indicate the abundance relative to all other genes/isoforms**

Gene-level Read Counts

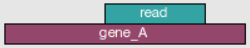
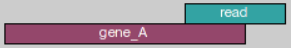
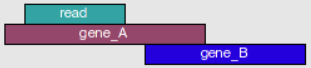
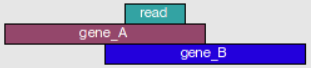
- rather straightforward to compute:
  - # reads that uniquely map to a gene locus
    → biased by length, discards information in multi-mappers
  - #reads that map to gene locus (including multi-mappers)
    → disambiguation is not possible if you do not have abundance estimates of the isoforms
    → needs to resort to heuristics to assign multi-mappers
    - randomly assign to one of the matching genes
    - do a fractional assignment with a with 1/#genes mapped

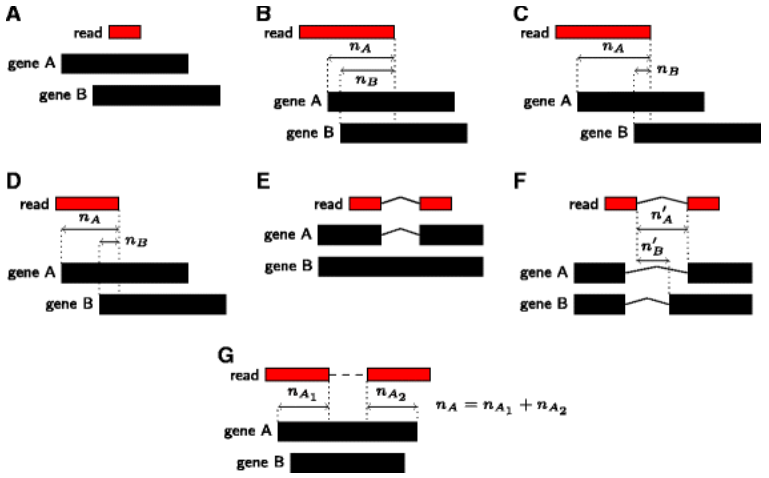
featureCounts – assigning reads to genes

- versatile function to count reads towards genes
- implements many different counting modes
- covers different aspects of overlap situations
  - partial overlap
  - overlapping multiple features at the same alignment position
  - overlapping multiple features at different alignment positions

- Simple overlap is not sufficient, read must be compatible with exon structure

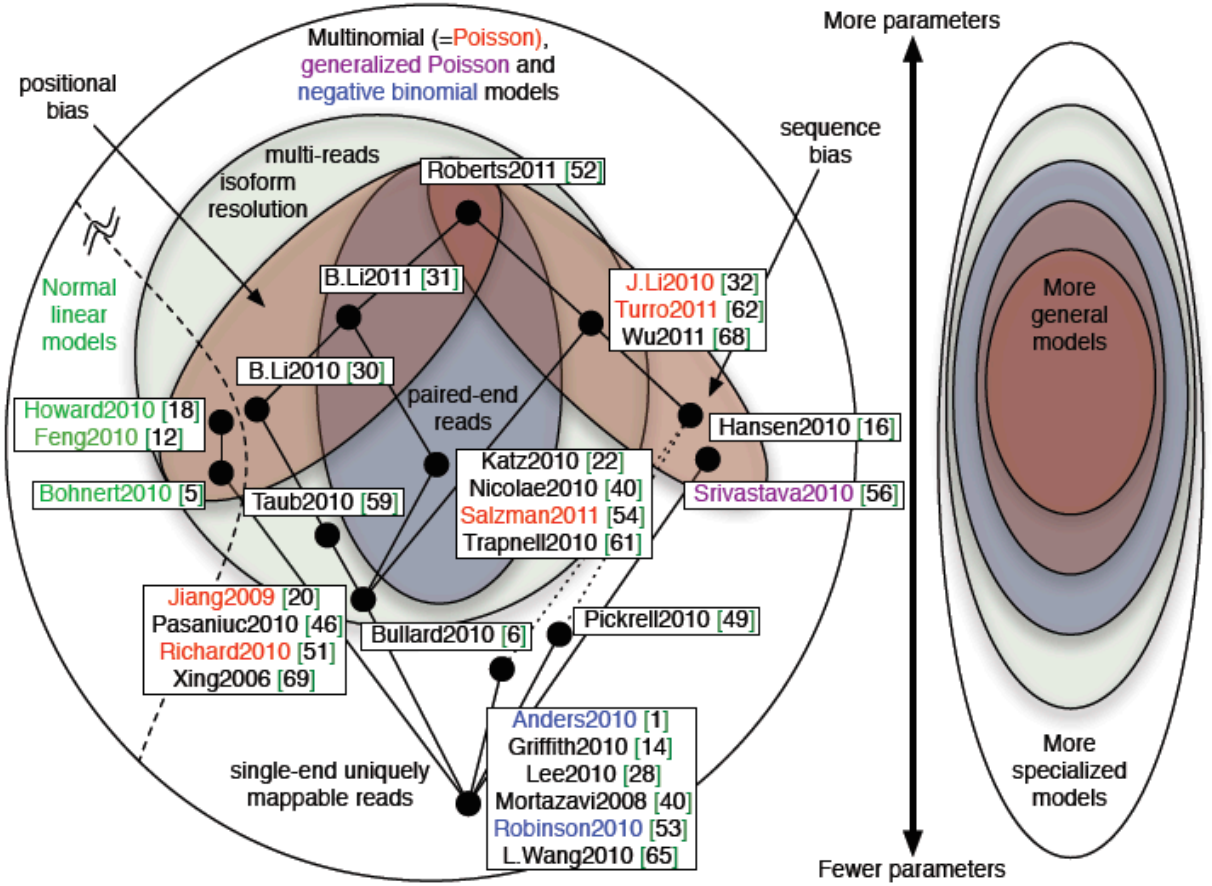
Model-free Counting of Overlapping reads – Count Modes

	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



RNA-seq quantification

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:

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

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:

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

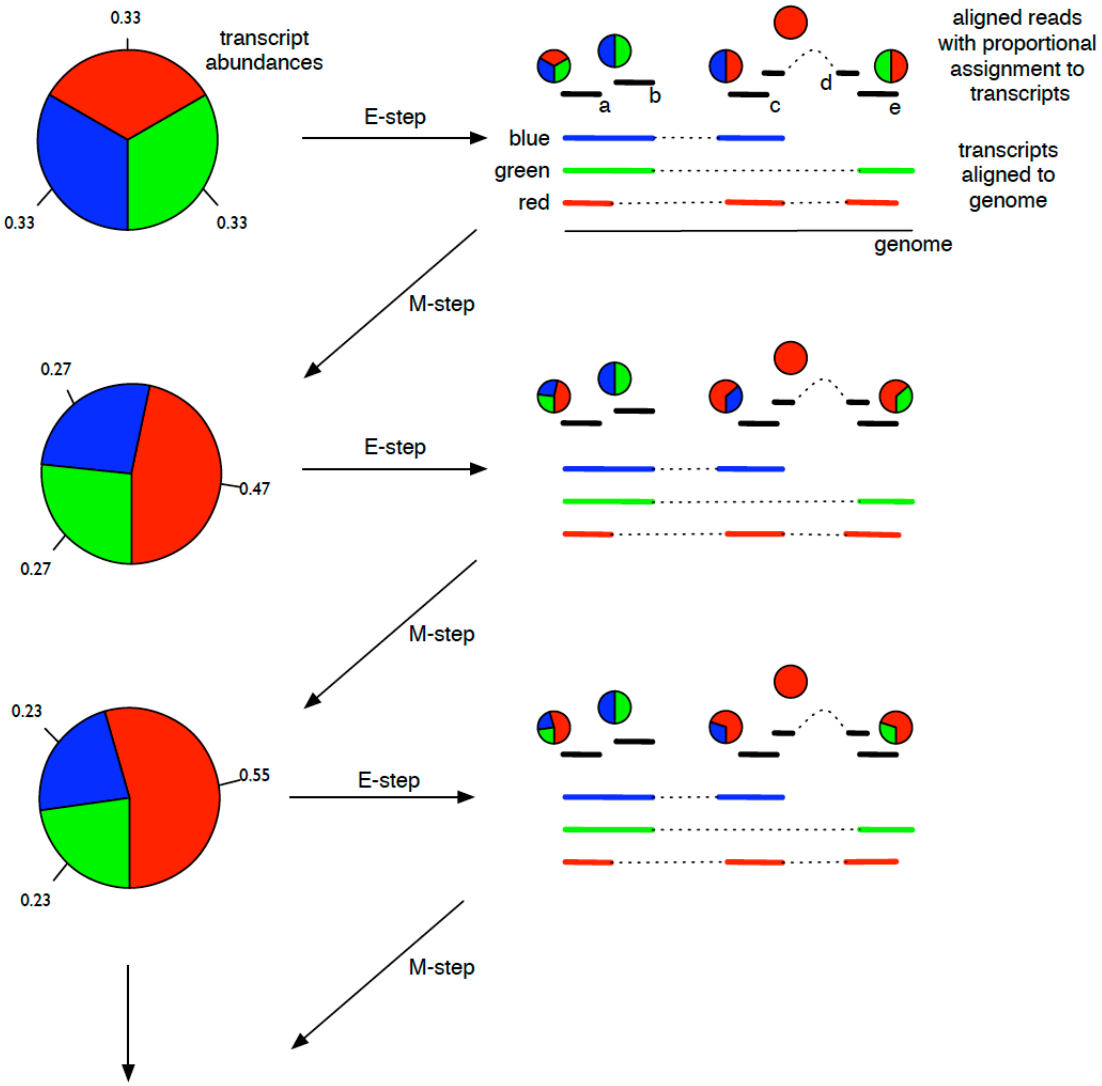
- but now abundances have to be estimated iteratively

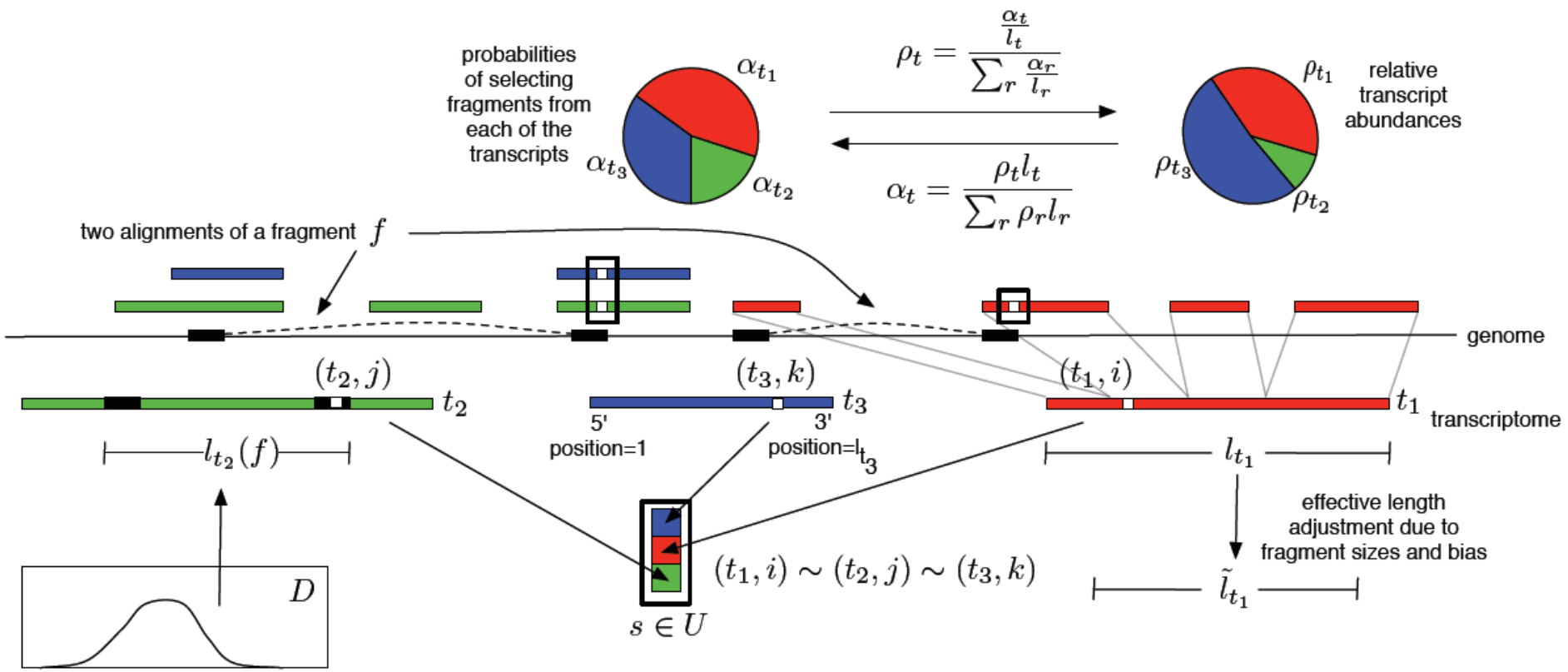
Iterative Estimation

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'
3'
position=1
position=l_{t_3}
$l_{t_2}(f)$
l_{t_1}
effective length adjustment due to fragment sizes and bias
D
$(t_1, i) \sim (t_2, j) \sim (t_3, k)$
$s \in U$
$\tilde{l}_{t_1}$

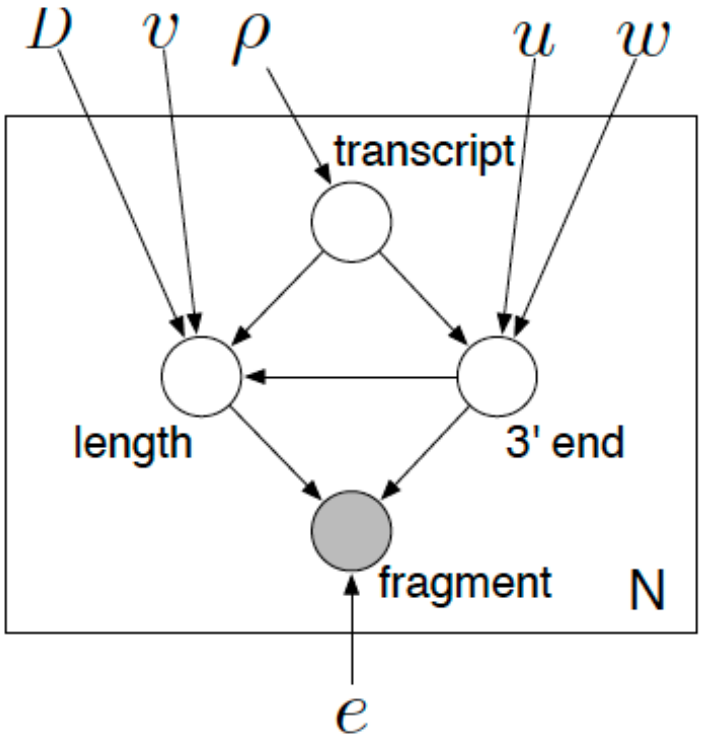
General Formulation of Abundance Estimation

A full model for the abundance estimation should consider:

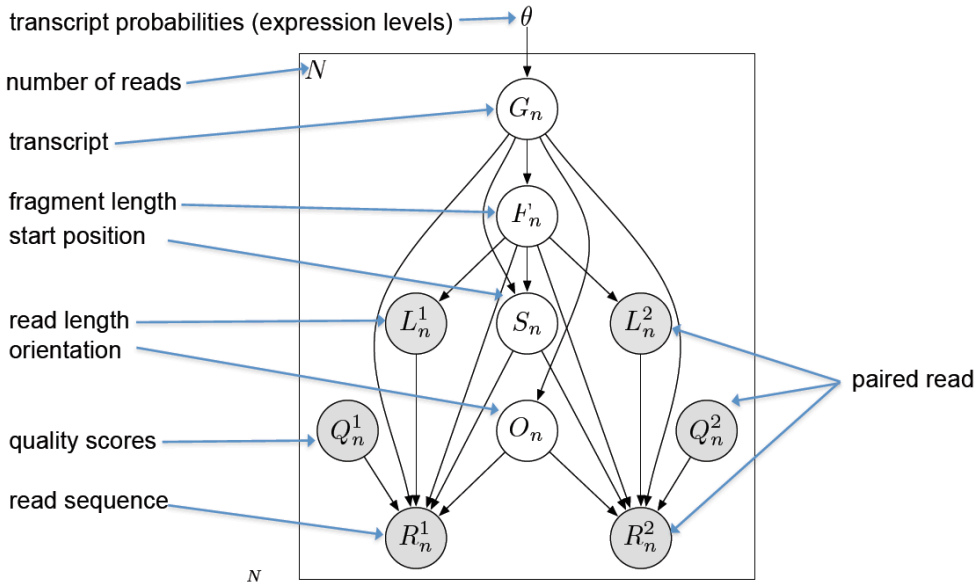
- position bias
- fragment-length distribution
- sequencing errors
- site-specific bias
- …

Example Implementations

Pachter: Cufflinks

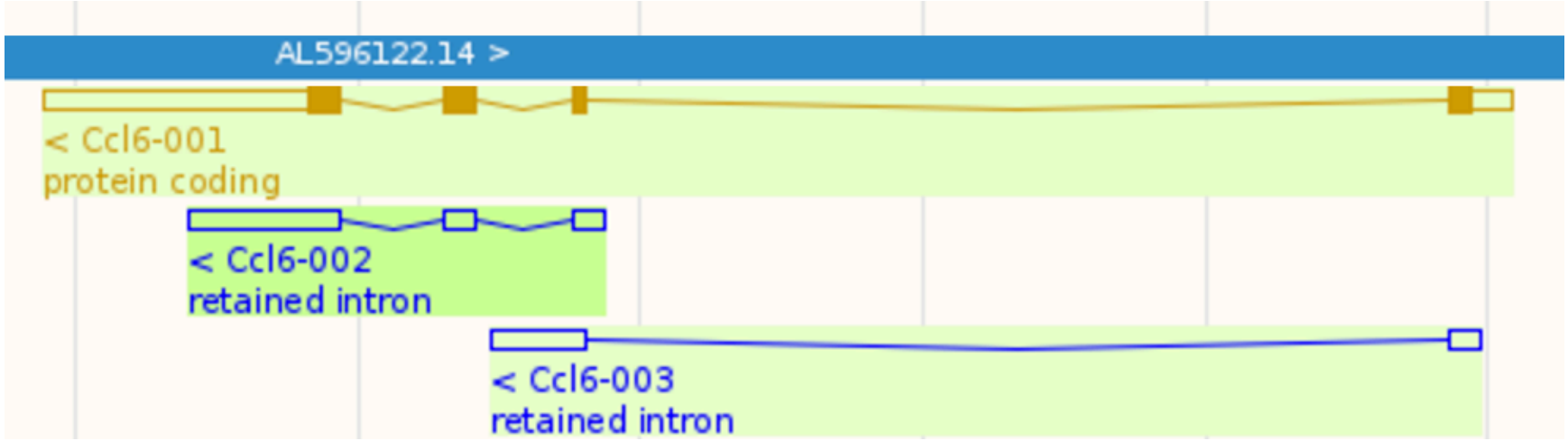


Dewey: RSEM



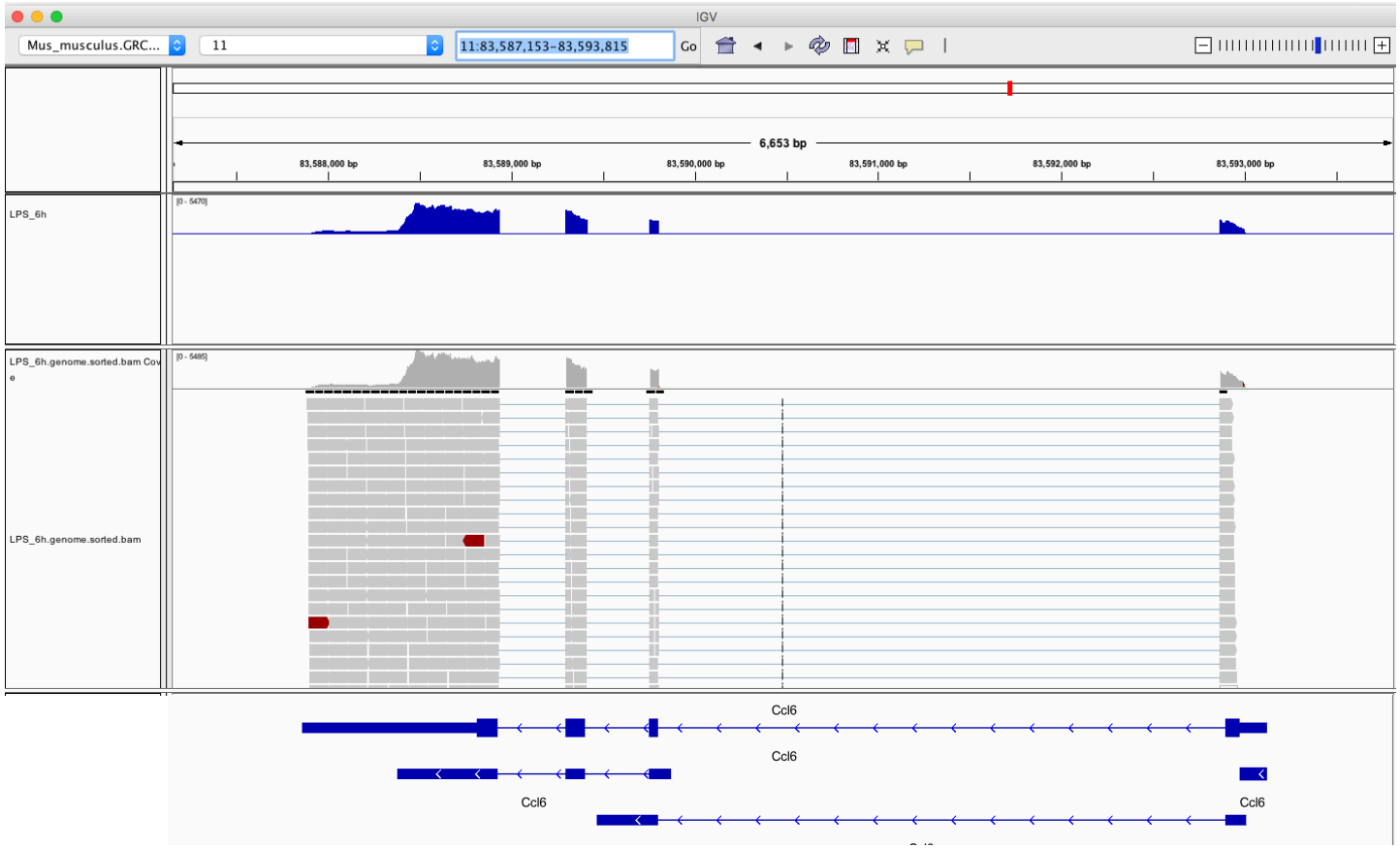
$$P(\mathbf{g},\mathbf{f},\mathbf{s},\mathbf{o},\ell,\mathbf{q},\mathbf{r}|\theta)=\prod_{n=1}^{N}P(g_n|\theta)P(f_n|g_n)P(s_n|f_n,g_n)P(o_n|g_n)P(q_n)P(\ell_n|f_n)P(r_n|g_n,f_n,s_n,o_n,\ell_n,q_n)$$

Example: RSEM



Ccl6 gene locus with 3 isoforms

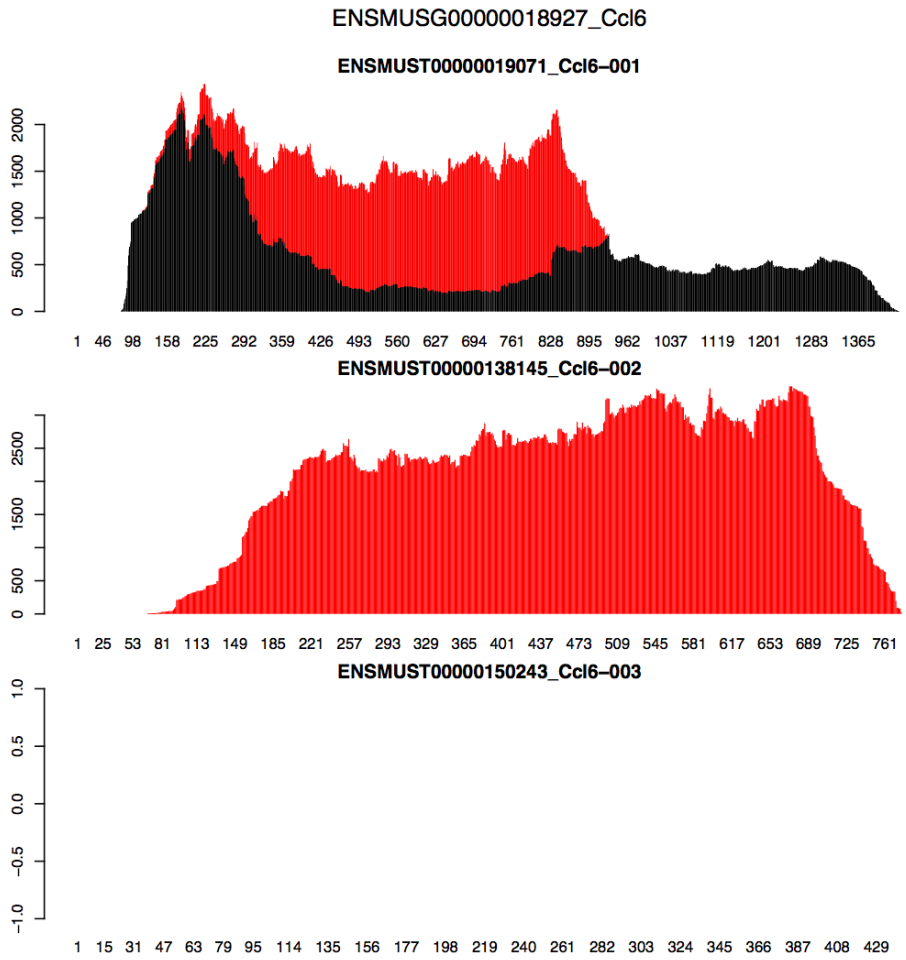
follows the example:
https://github.com/bli25broad/RSEM_tutorial



RSEM result:

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct		
ENSMUST00000019071_Ccl6-001	ENSMUSG00000018927_Ccl6	1440	1194.85	7805.95	8862.10	9334.46	31.00		
ENSMUST00000138145_Ccl6-002	ENSMUSG00000018927_Ccl6	776	530.94	7719.05	19721.39	20772.55	69.00		
ENSMUST00000150243_Ccl6-003	ENSMUSG00000018927_Ccl6	442	202.64	0.00	0.00	0.00	0.00		

Ccl6 coverage in transcript space



- orientation is flipped because gene is on negative strand
- black: unique alignments
- red: expected depth from multi-mapping reads

Limitations of Generative Models

- Estimates can not be correct if underlying model of transcripts are incorrect or incomplete
- Abundance estimates are fractions; these can be used to get estimates number of reads generated by a given gene; error distribution of estimated read counts may be unclear

Implementation of Generative Models

- **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).

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

Definition of expression levels

- Goal: Start from read counts and define a quantity that indicates **relative molar concentration of a transcript**

- 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

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

- Relationship

$$\text{TPM} = 10^6 * \frac{\text{RPKM}}{\text{Sum(RPKM)}}$$

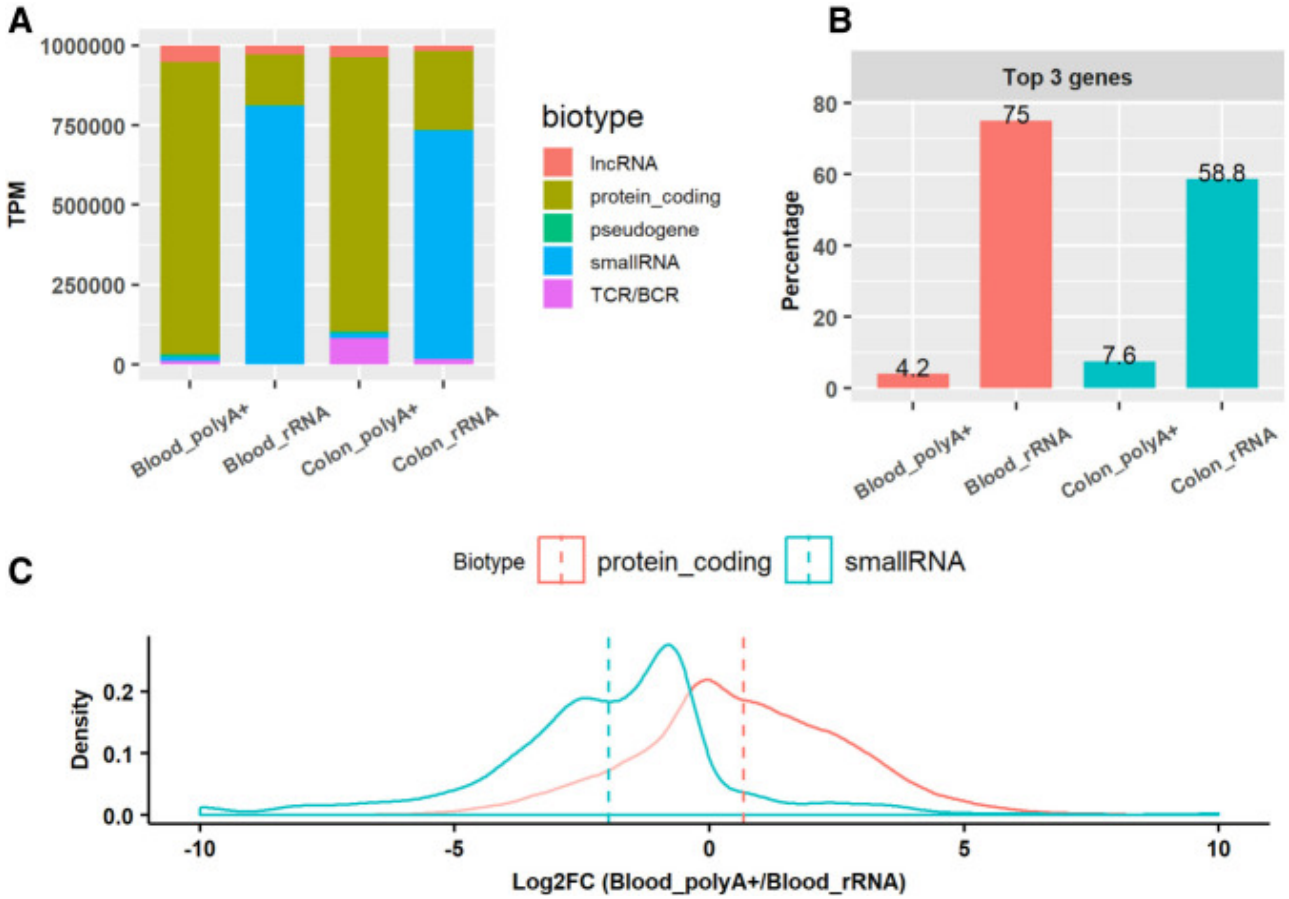
Shortcomings of RPKM and TPM

- Sum of RPKM varies from sample to sample, i.e. RPKM is not a measure of relative concentration because the measures of relative concentrations would sump up to constant
- TPM is unitless and satisfies this requirement

- **Only TPM should be used!**
- **But: even TPM is not a suitably normalized measure that can be used to compare samples from**
  - **different tissues**
  - **different protocols**

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7373998/

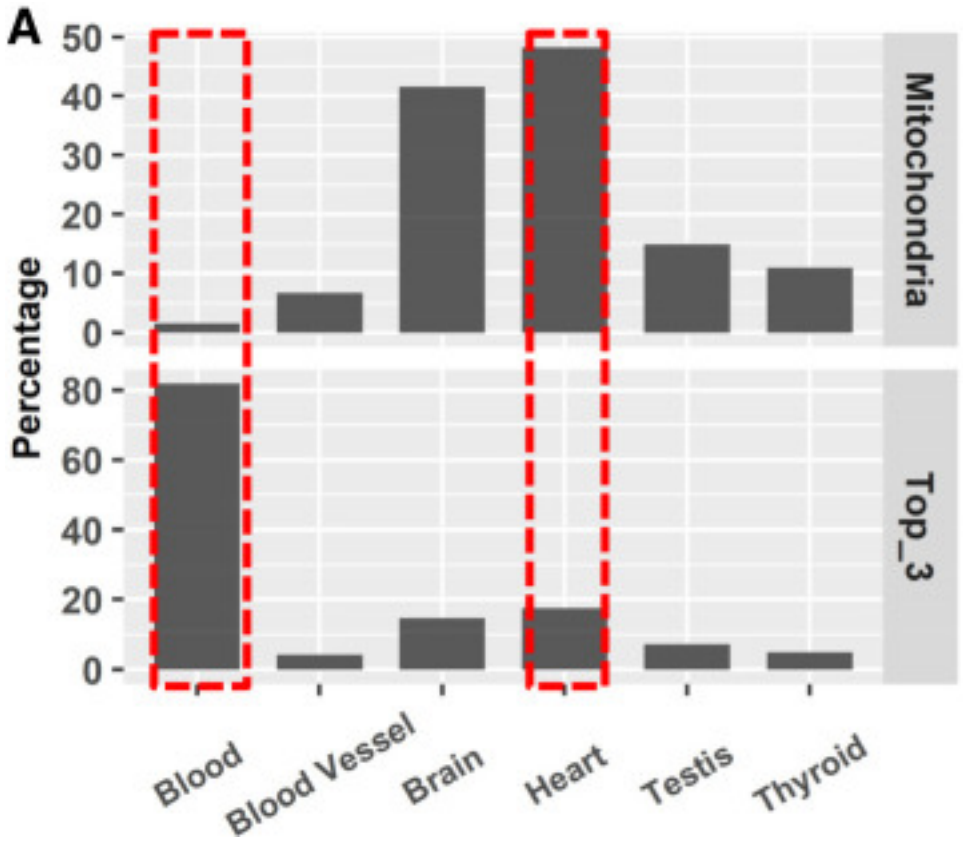
Comparing samples across protocols

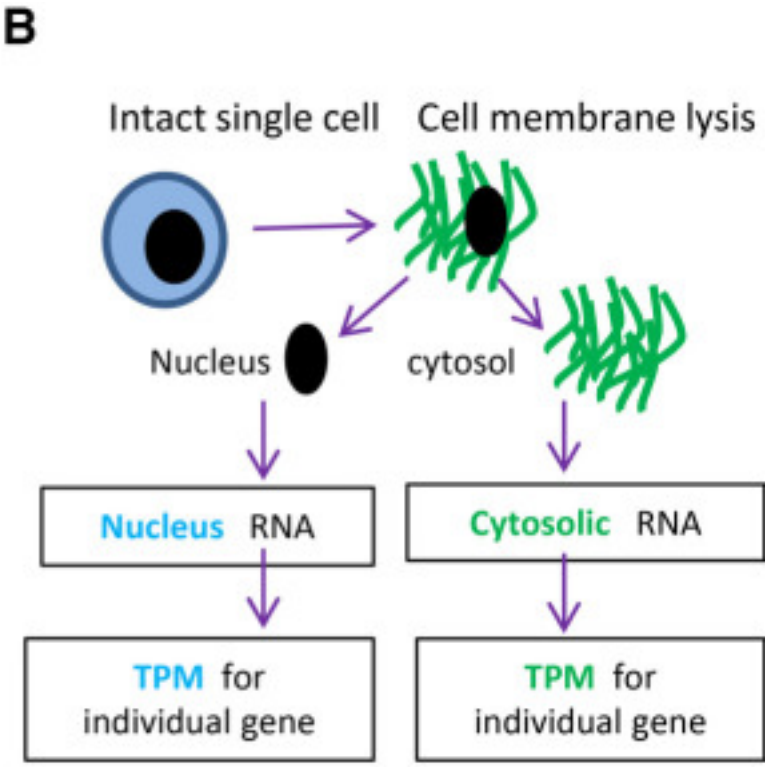


issues:

- surveyed populations are not comparable
- expression of top 3 genes will drive the TPM normalization (because it has a major influence on the sum of all reads)

Comparing samples across tissues





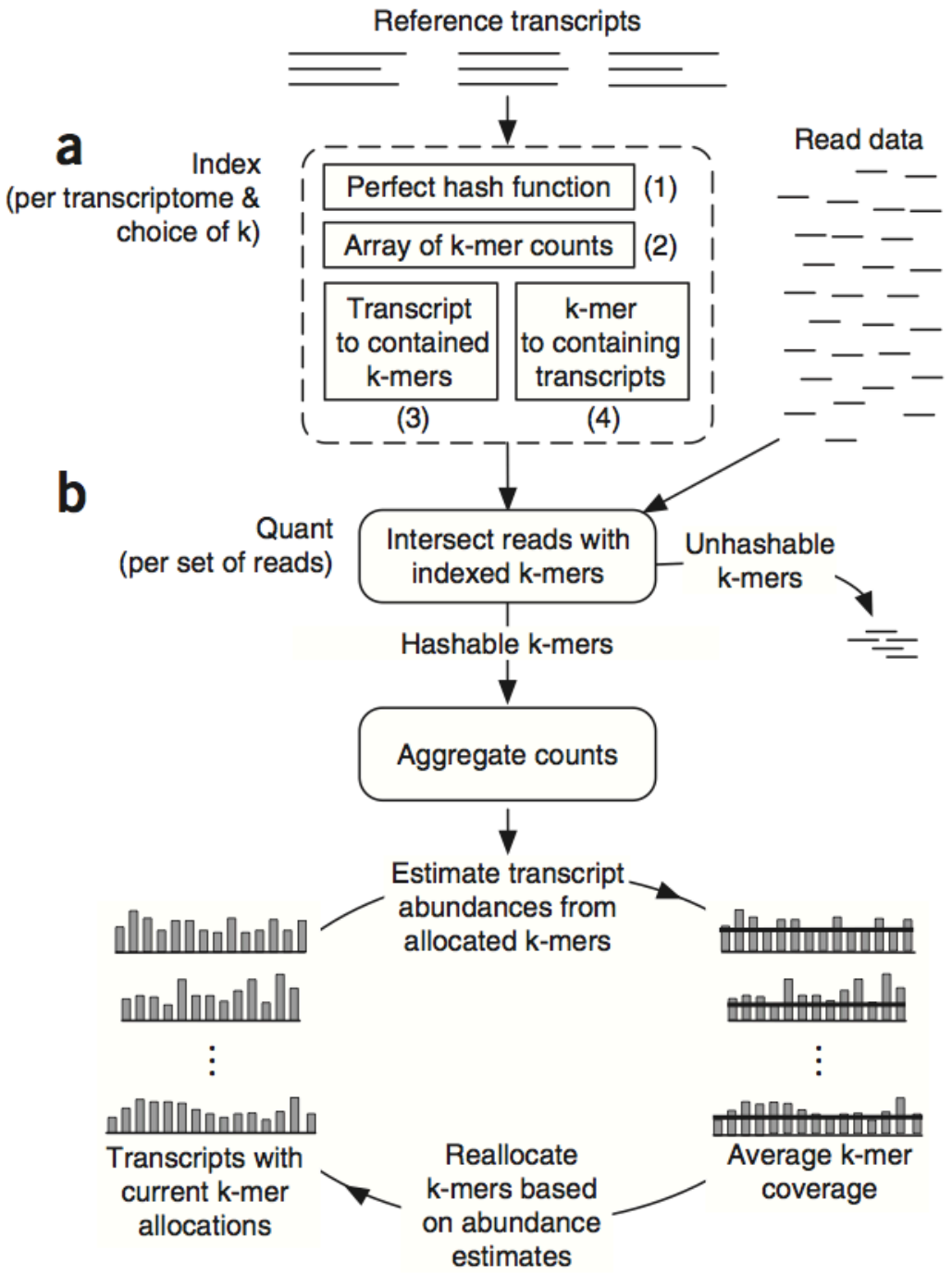
Different tissues may have different populations of genes expressed

Fast approaches to get the Read-Transcript Compatibility Matrix

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

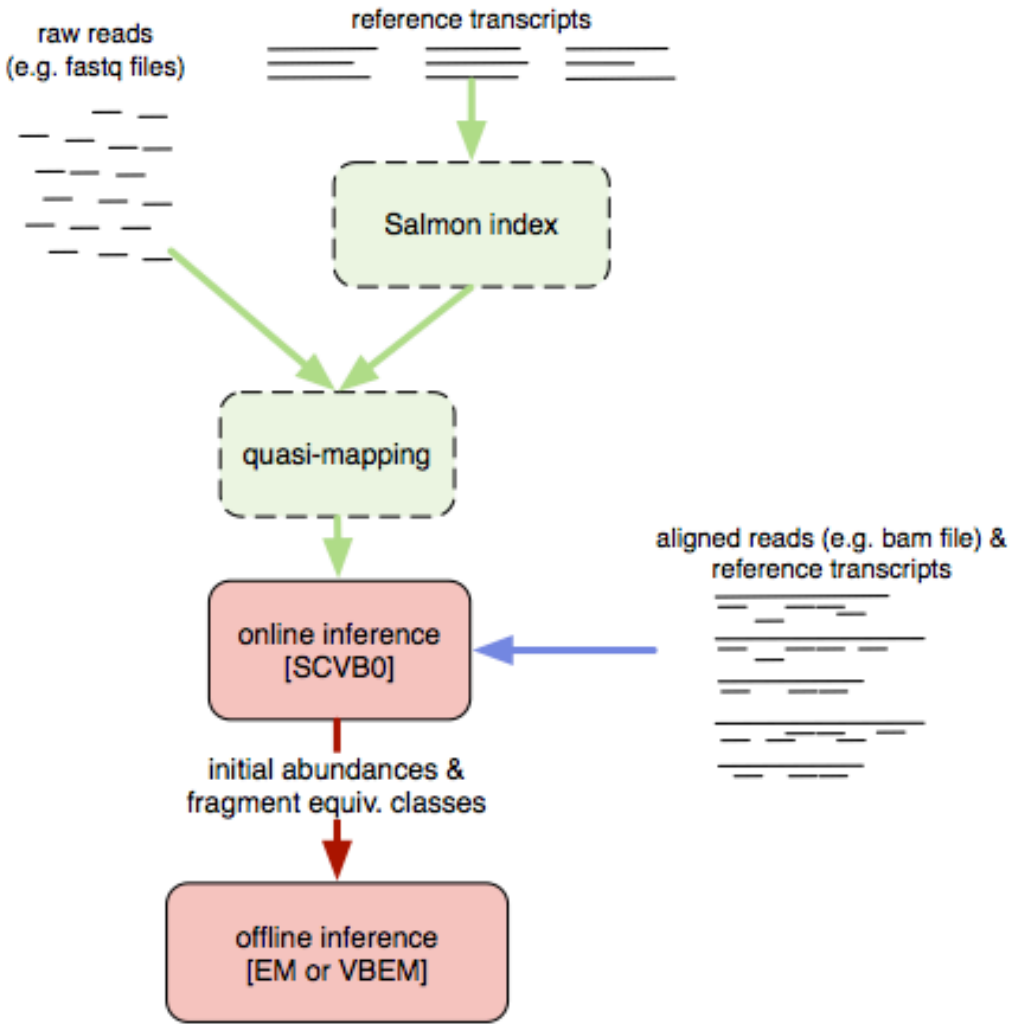
Sailfish

- 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

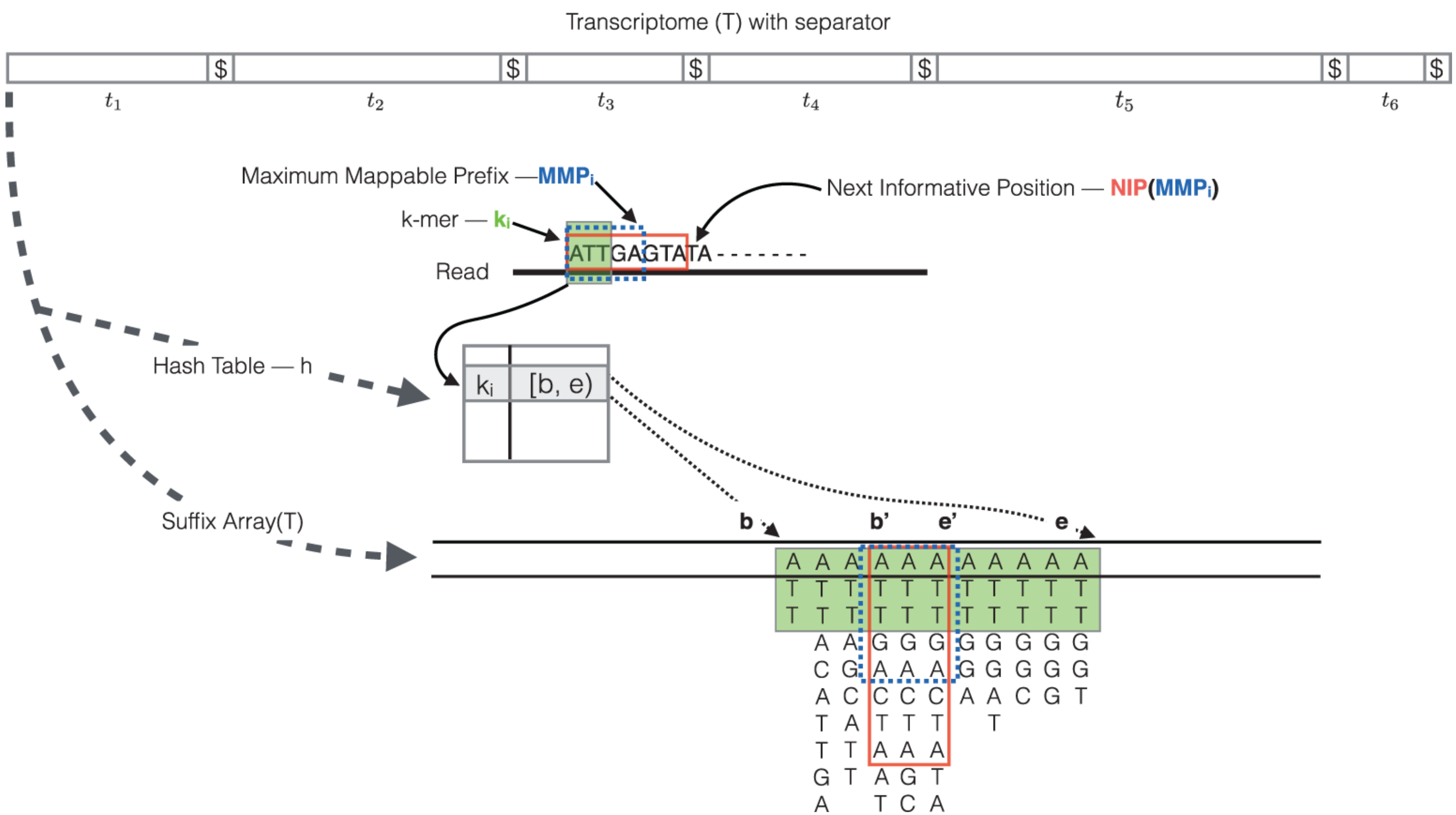


Patro et al. Sailfish …. Nature Biotechnology, 2014

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]

Quasi-mapping



Quasi-Mapping

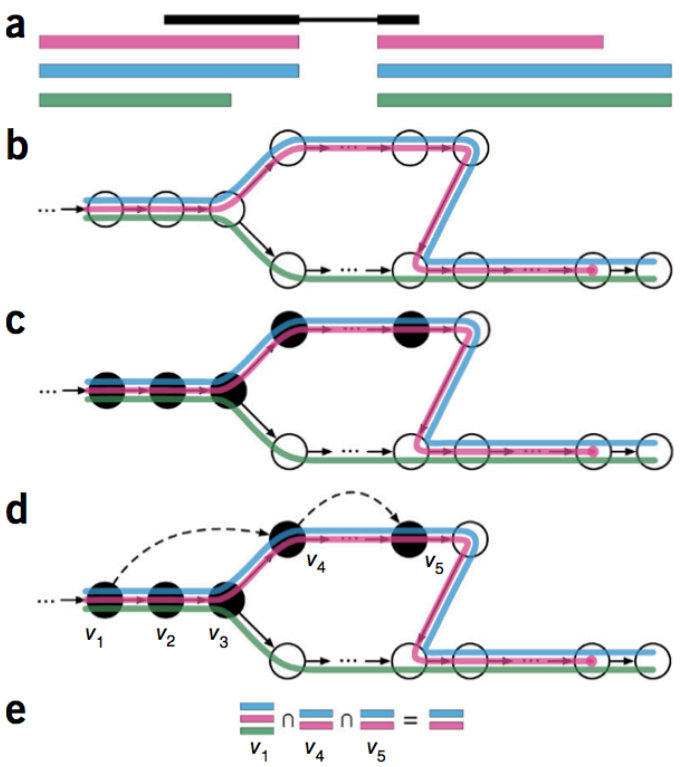
1. The read is scanned from left to right until a k-mer that appears in the hash table is discovered.
2. The k-mer is looked up in the hash table and the SA intervals are retrieved, giving all suffixes containing that k-mer
3. Similar to STAR, the maximal matching prefix (MMP) is identified by finding the longest read sequence that exactly matches the reference suffixes.
4. Salmon identifies the next informative position (NIP), by skipping ahead 1 k-mer (speedup)
5. Repeat above until the end of the read.
6. The final mappings are generated by determining the set of transcripts appearing in all MMPs for the read. The transcripts, orientation and transcript location are output for each read.

Quasi-Mapping

- Result: Read-Transcript compatibility matrix
- Only based on compatibility of short k-mers
- Has an optional step to ***validate mappings***:
  - goes through all the read-transcript associations and validates if the entire read is compatible with the transcripts by doing a base-by-base comparison

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

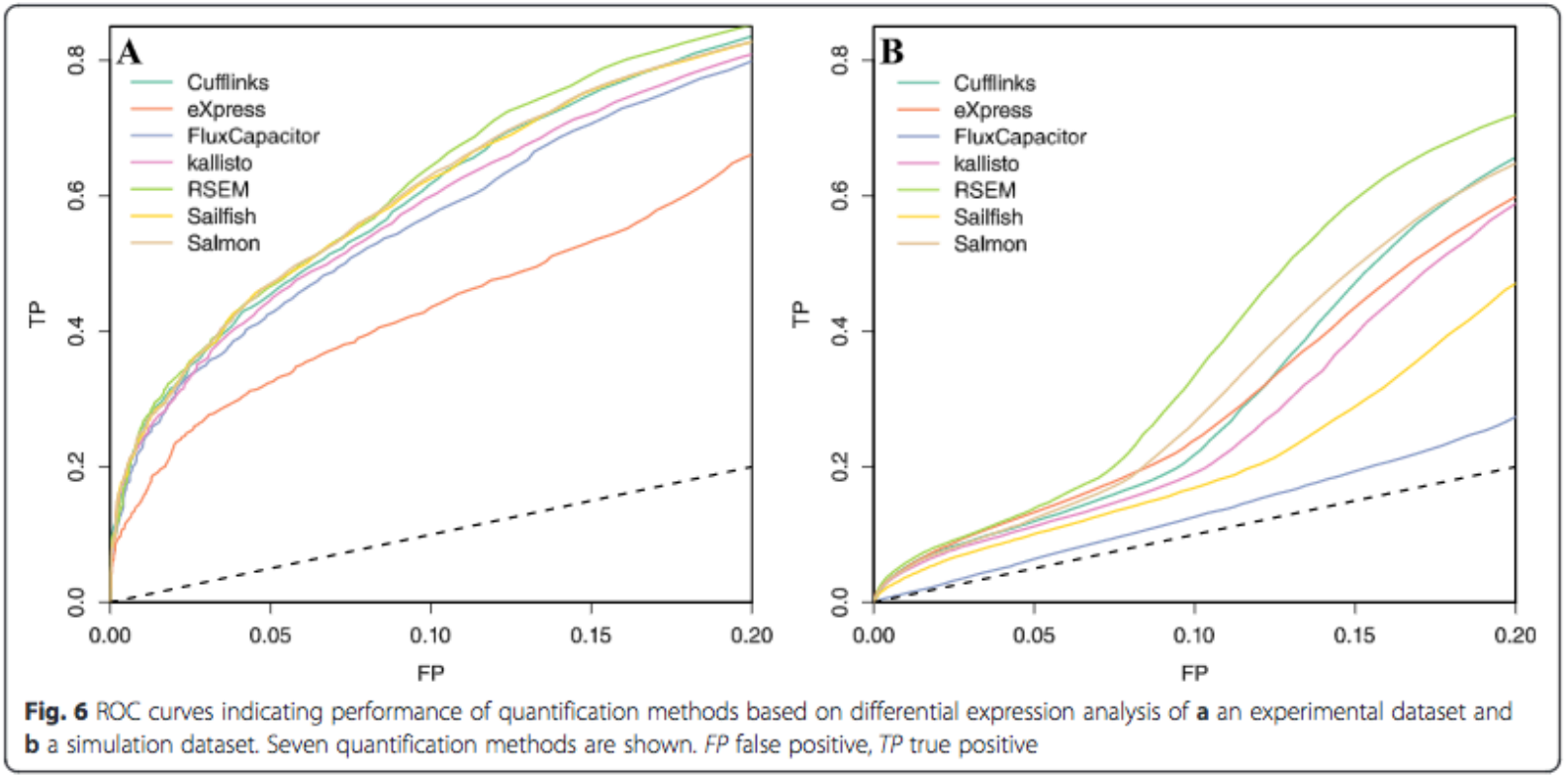
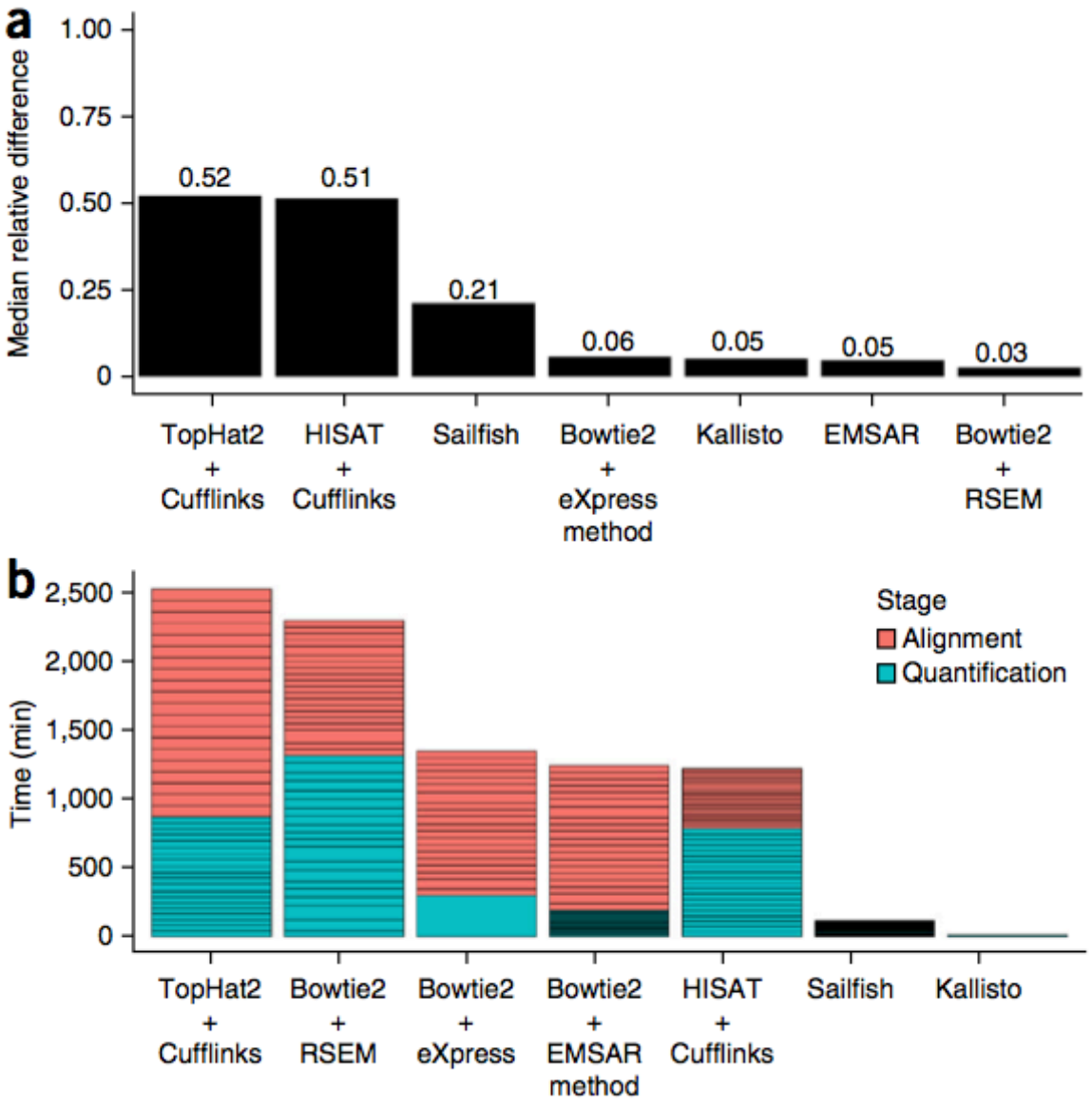


Fig. 6 ROC curves indicating performance of quantification methods based on differential expression analysis of **a** an experimental dataset and **b** a simulation dataset. Seven quantification methods are shown. *FP* false positive, *TP* true positive

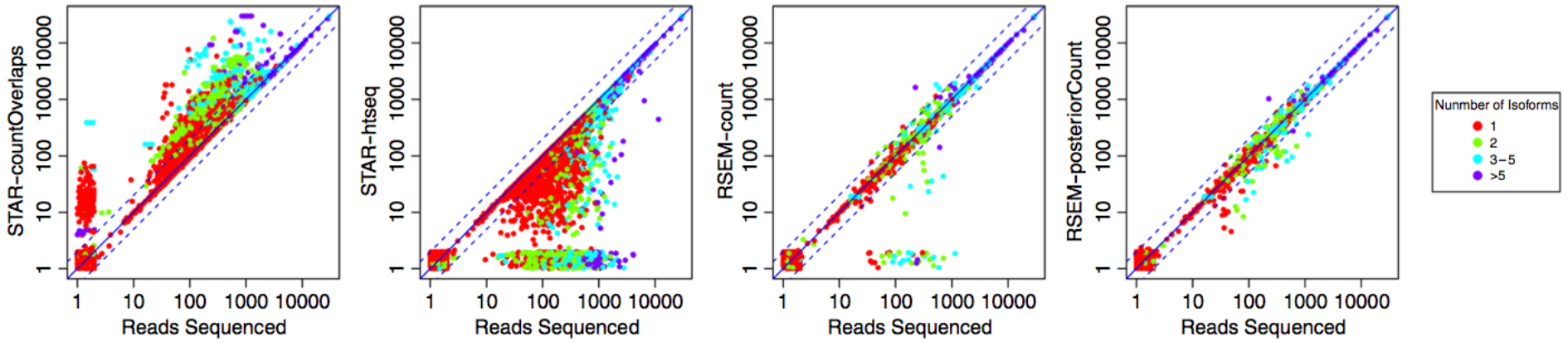
Teng et al. Genome Biology, 2016

Performance Comparison



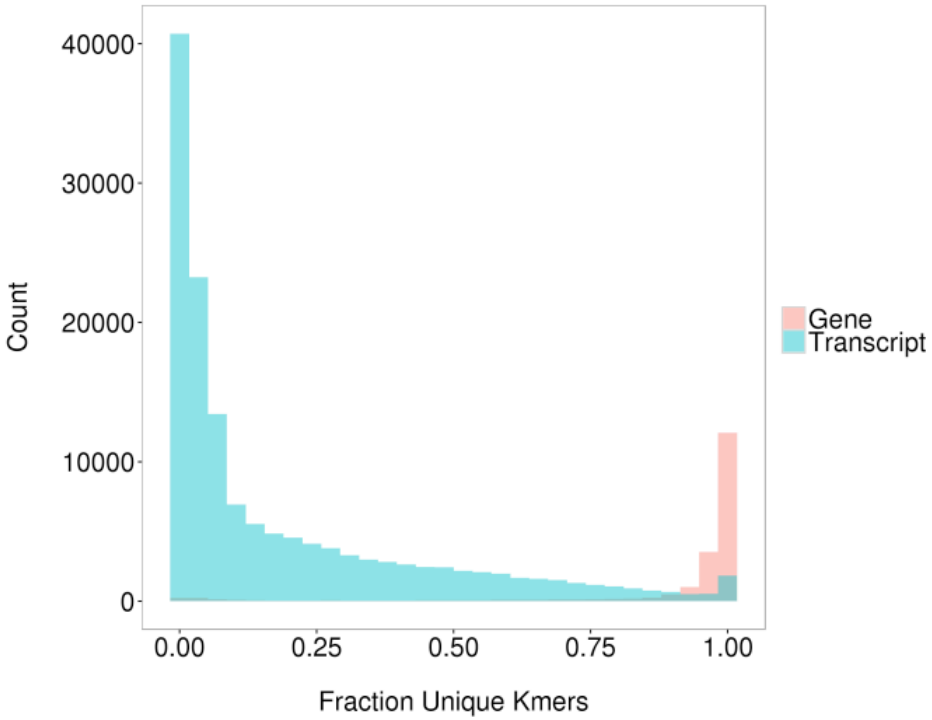
Bray et al. Nature Biotechnology, 2016

Read Counting Accuracy

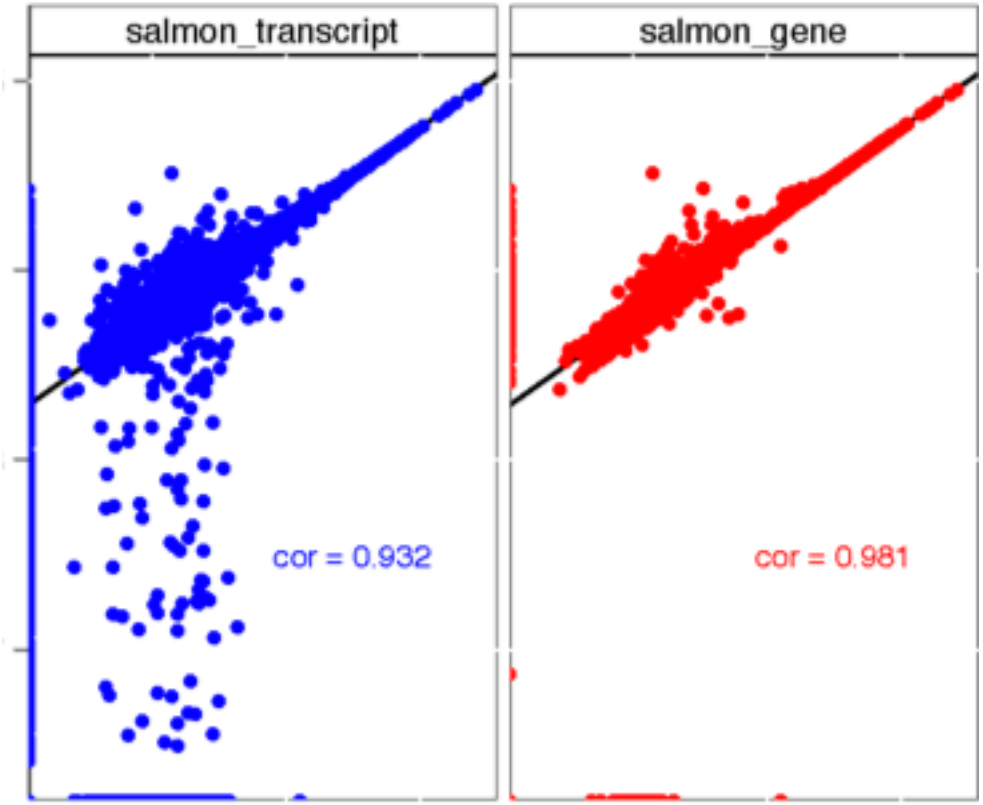


Rehrauer et al.BMC Bioinformatics, 2014

Uniqueness: Isoform-level vs gene-level

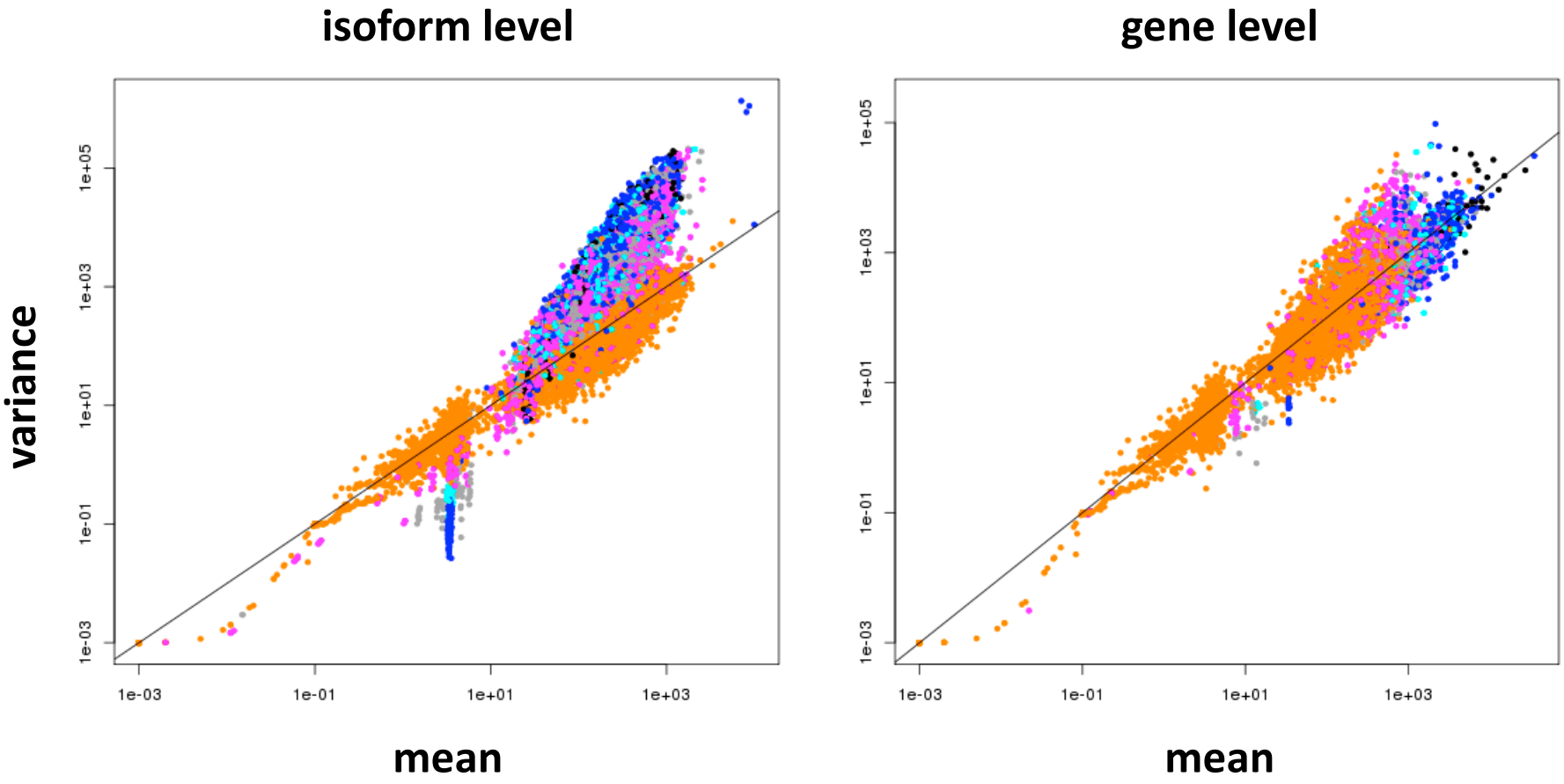


Accuracy: Isoform-level vs gene-level

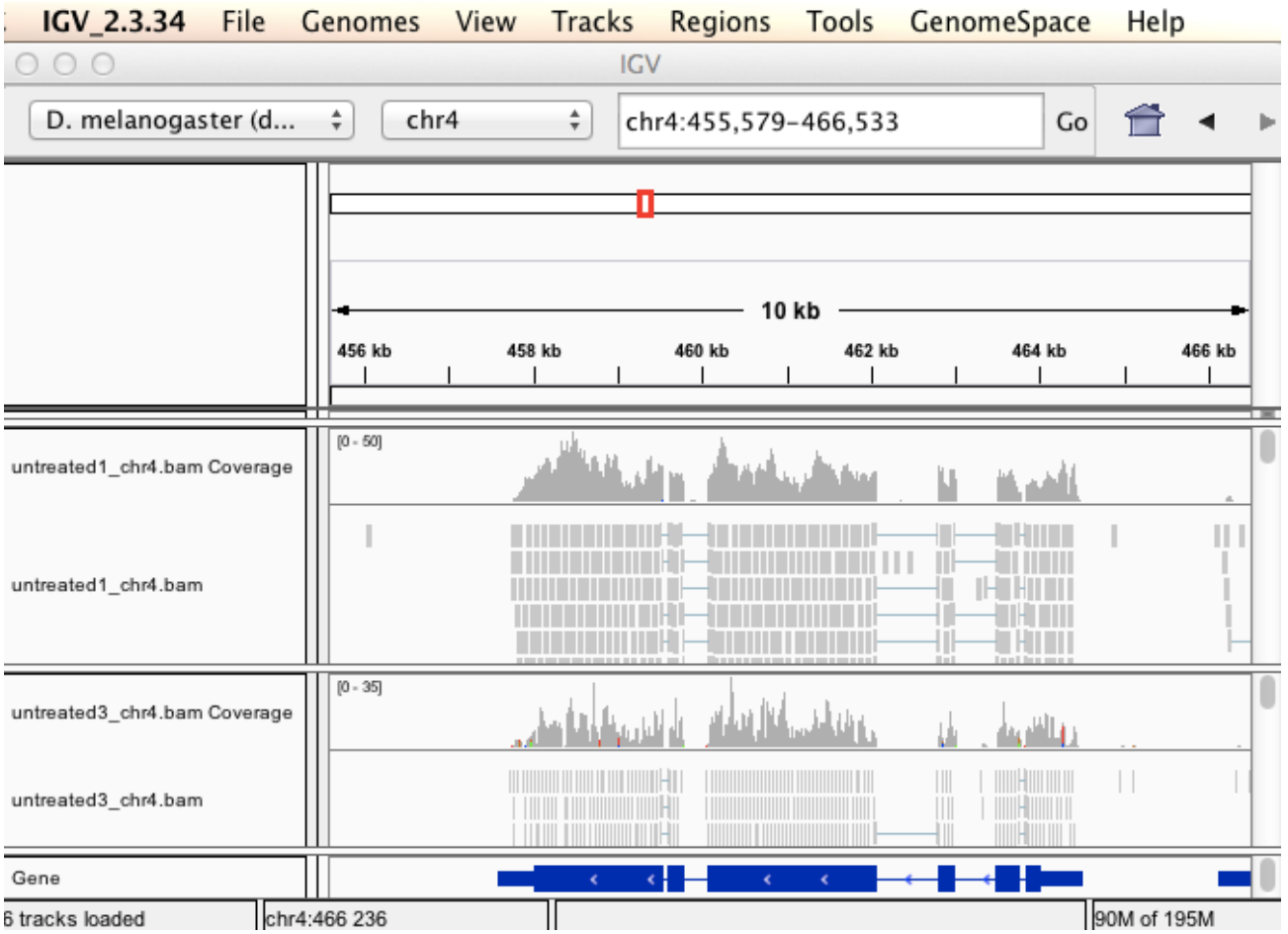


Soneson et al. F1000, 2015

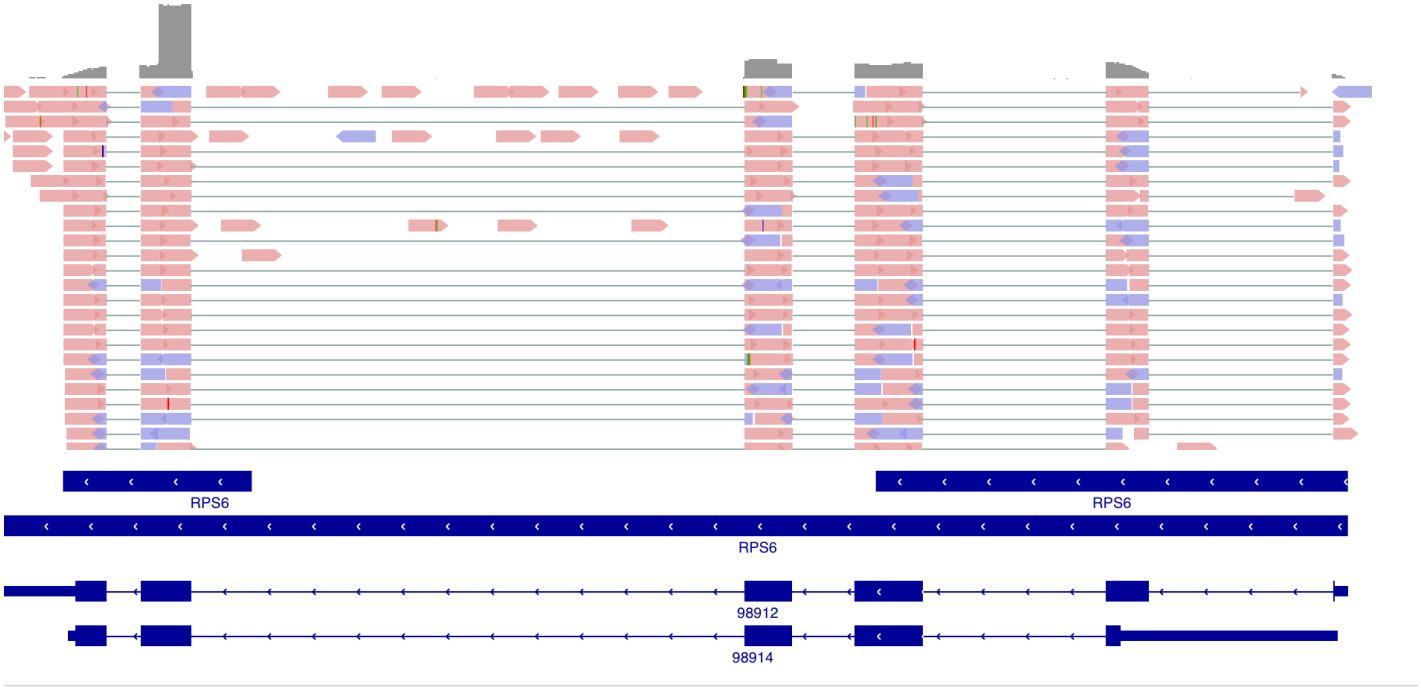
Isoform level has higher variability



Positional bias of read distributions



Duplicated Reads



Unspliced transcripts

- Isoform quantification assumes that only spliced transcripts have been measured
- Depending on the protocol unspliced transcripts might also be sequenced
  - transcripts that were newly generated in the nucleus but not yet spliced
  - usually unspliced transcripts are only a minor fraction

3’-Tagging

Reads are only generated near the 3’-end

Isoforms can not be resolved

Allows counting of the reads at the 3’-end