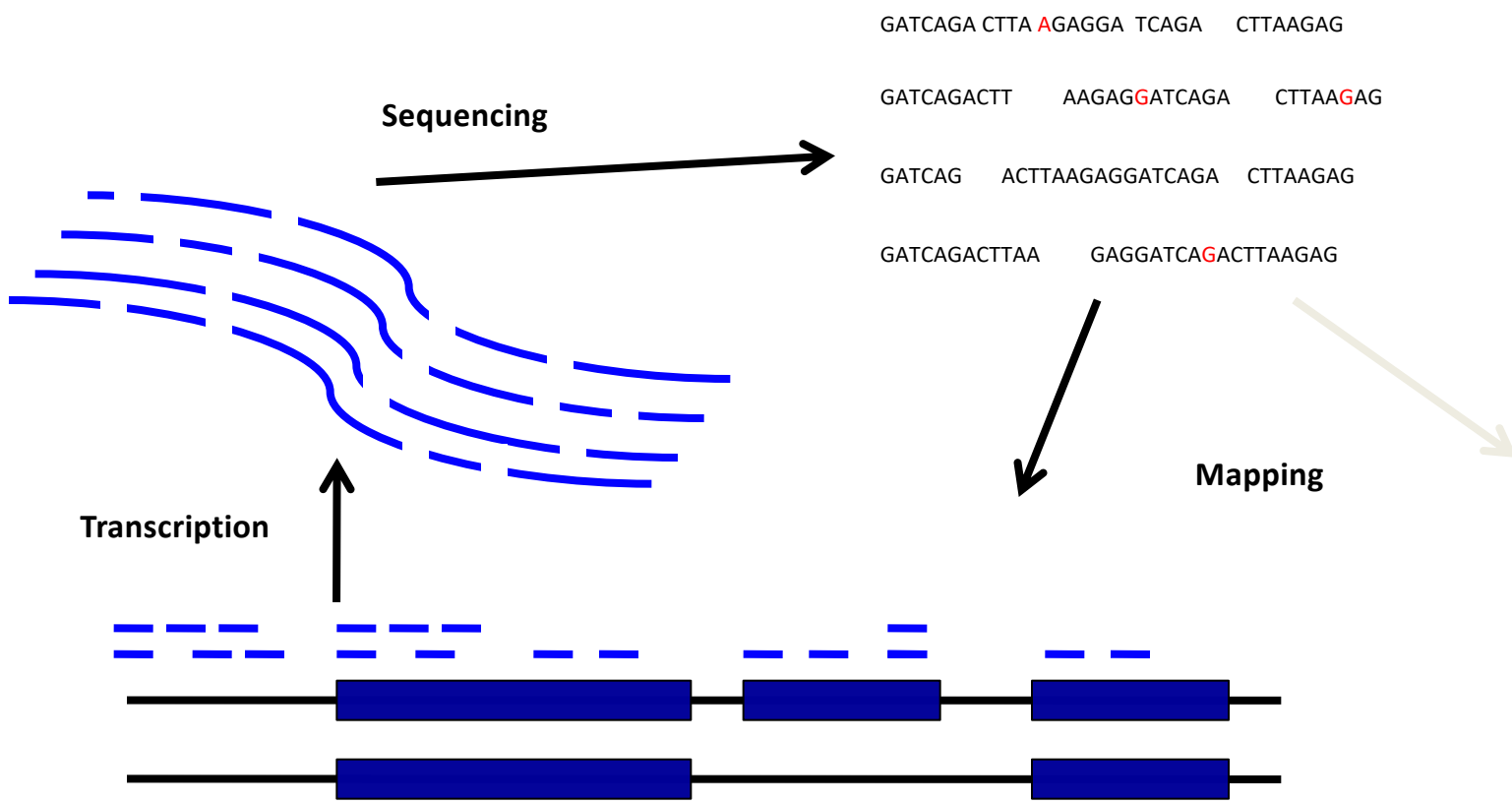


	0
01	1
10	0
01	1

# 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

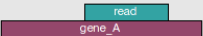
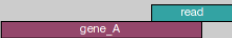


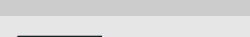

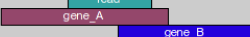
## Gene-level Read Counts

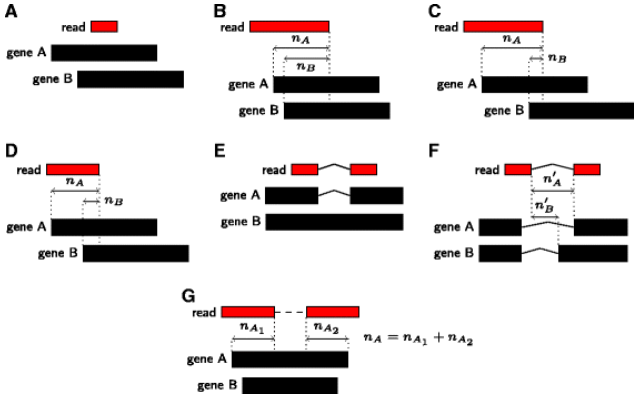
- rather straightforward to compute:
  - # reads that uniquely map to a gene locus
    - biased by length, discards information in multi-mappers
  - #reads that map to gene locus (including multi-mappers)
    - disambiguation is not possible if you do not have abundance estimates of the isoforms
   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

## *Rsubread::featureCounts* – assigning reads to genes

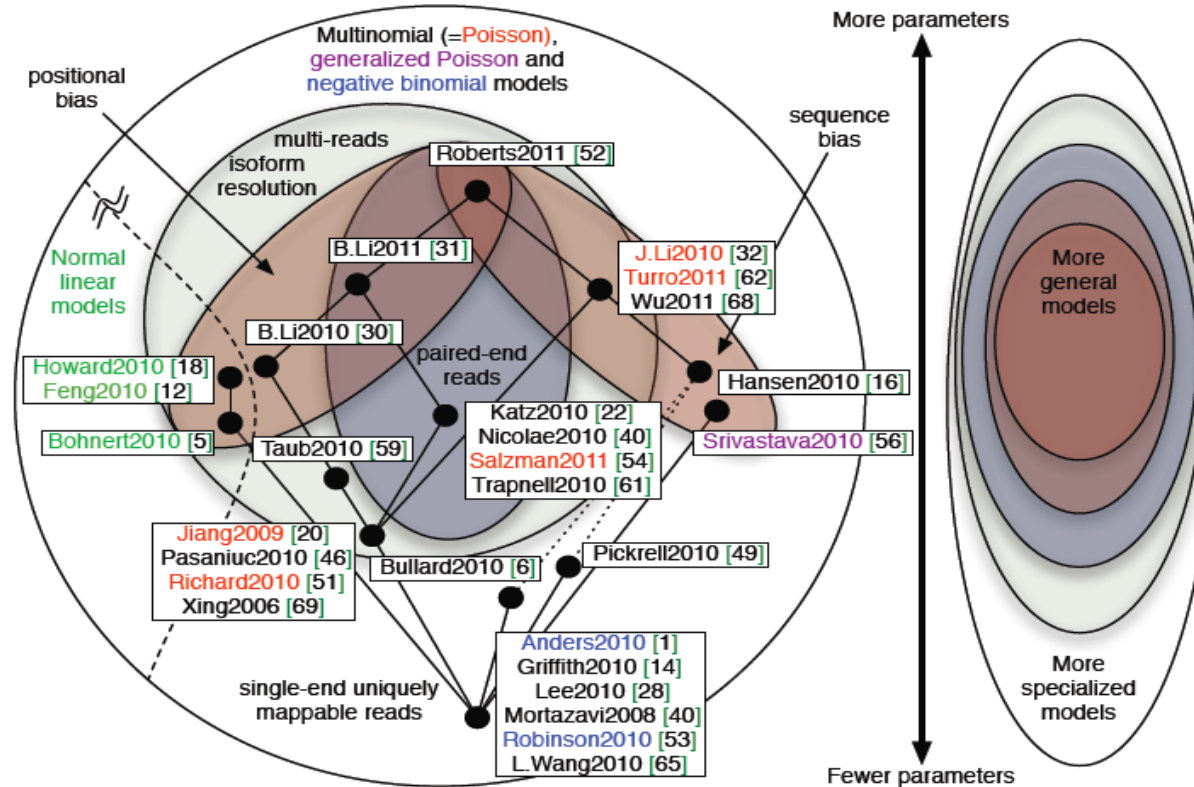
- versatile function to count reads towards genes
- implements many different counting modes
- covers different aspects of overlap situations
  - partial overlap
  - overlapping multiple features at the same alignment position
  - overlapping multiple features at different alignment positions
- Simple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



## Generative Models for RNA-seq quantification





## RNA-seq model

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

with:

$\rho_t$  expression level / abundance / fraction

$l_t$  transcript length

$$Z = \sum_t \rho_t l_t \quad \text{normalization factor}$$

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

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:

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

- 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

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

with

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

- The likelihood is now:

$$L[\alpha] = \prod_t \left( \sum_f y_{ft} \frac{\alpha_t}{l_t} \right)$$

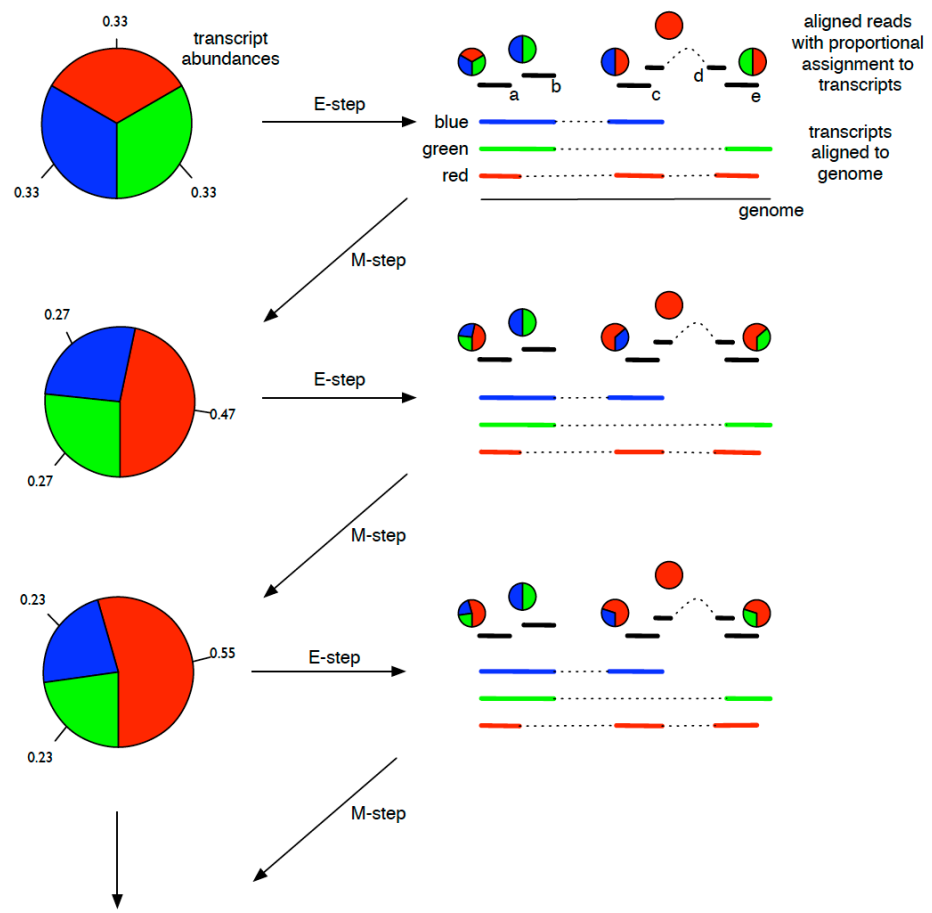
- but now abundances have to be estimated iteratively

## Iterative Estimation

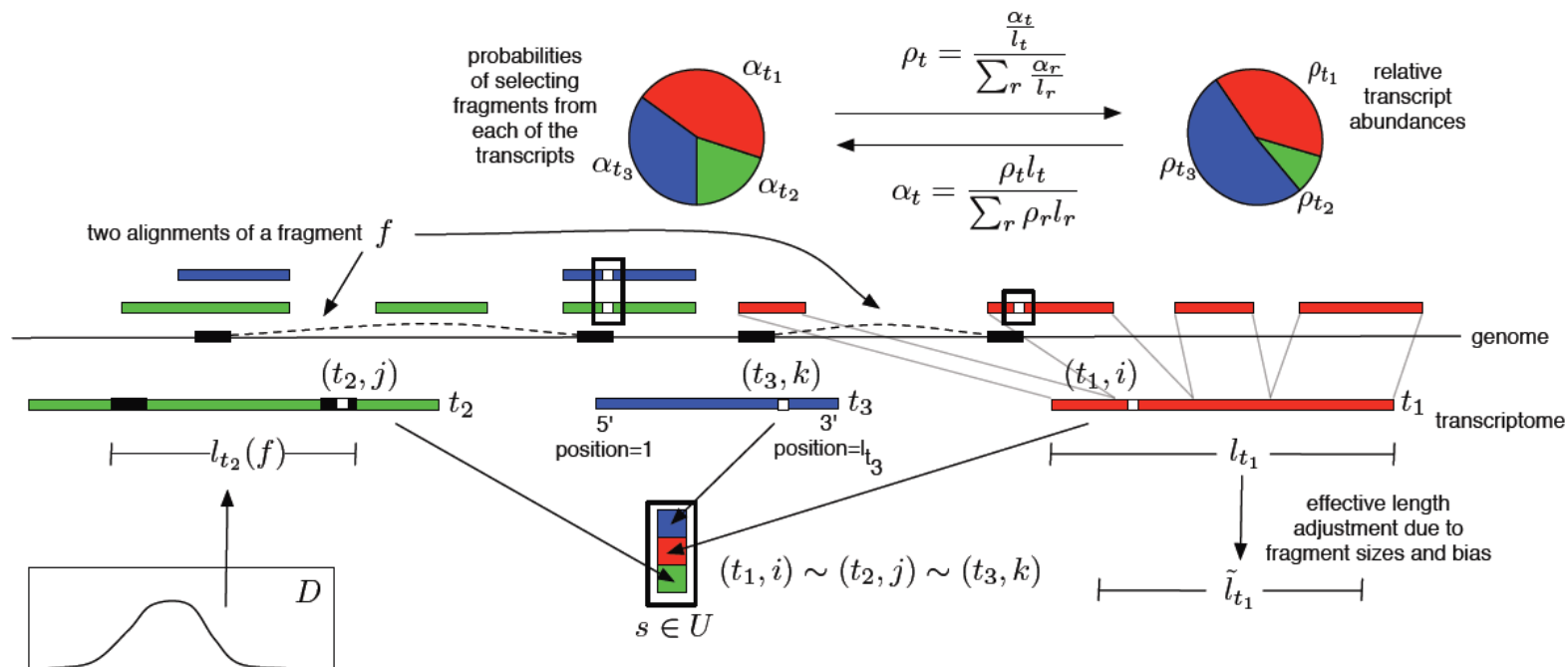
### Three step algorithm

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







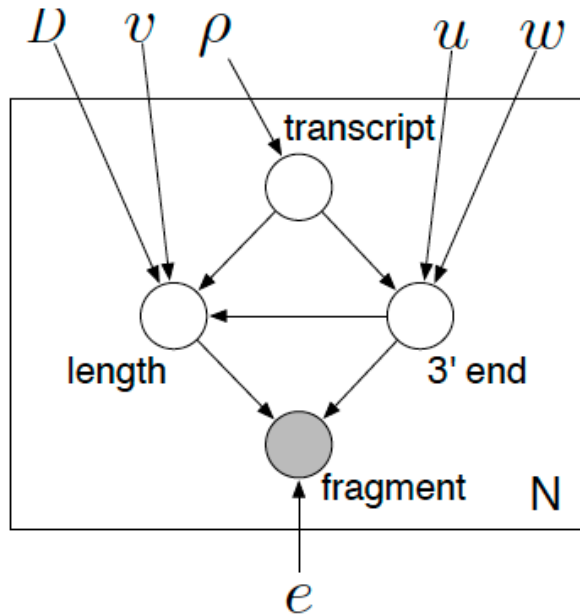
## General Formulation of Abundance Estimation

A full model for the abundance estimation should consider:

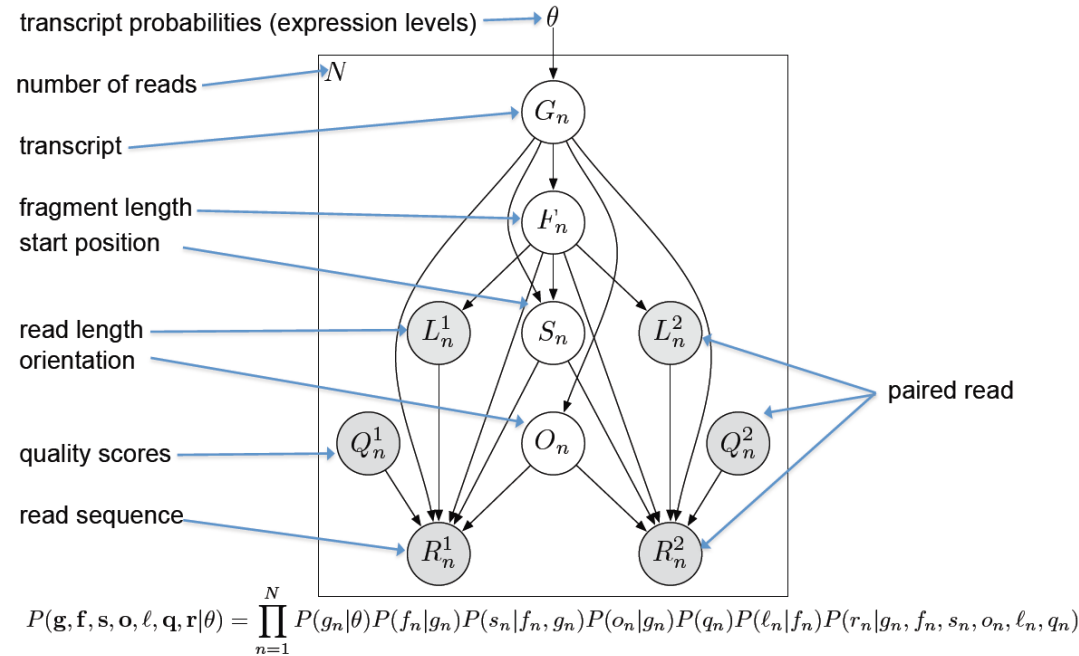
- position bias
- fragment-length distribution
- sequencing errors
- site-specific bias
- ...

# Example Implementations

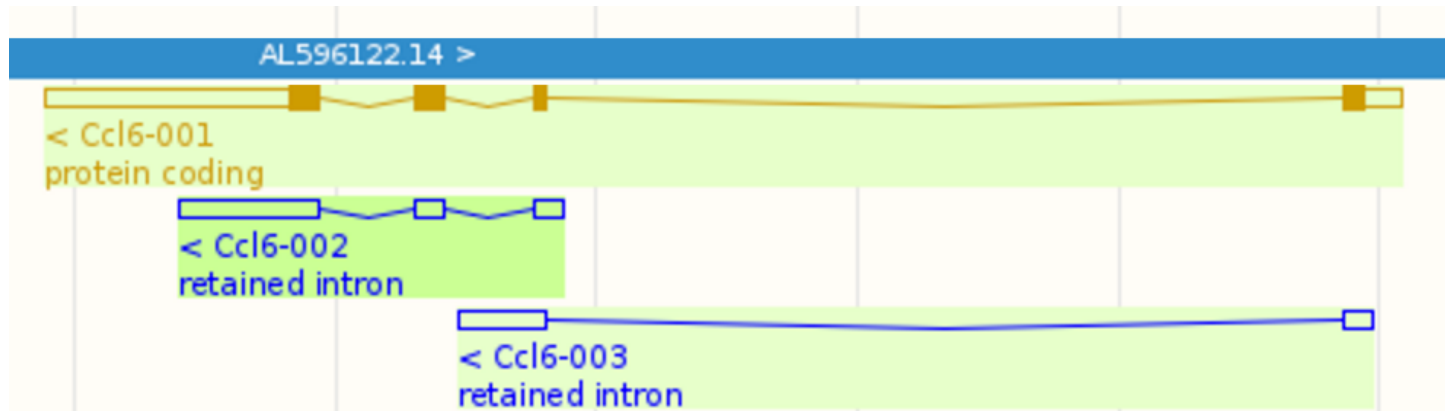
Pachter: Cufflinks



Dewey: RSEM



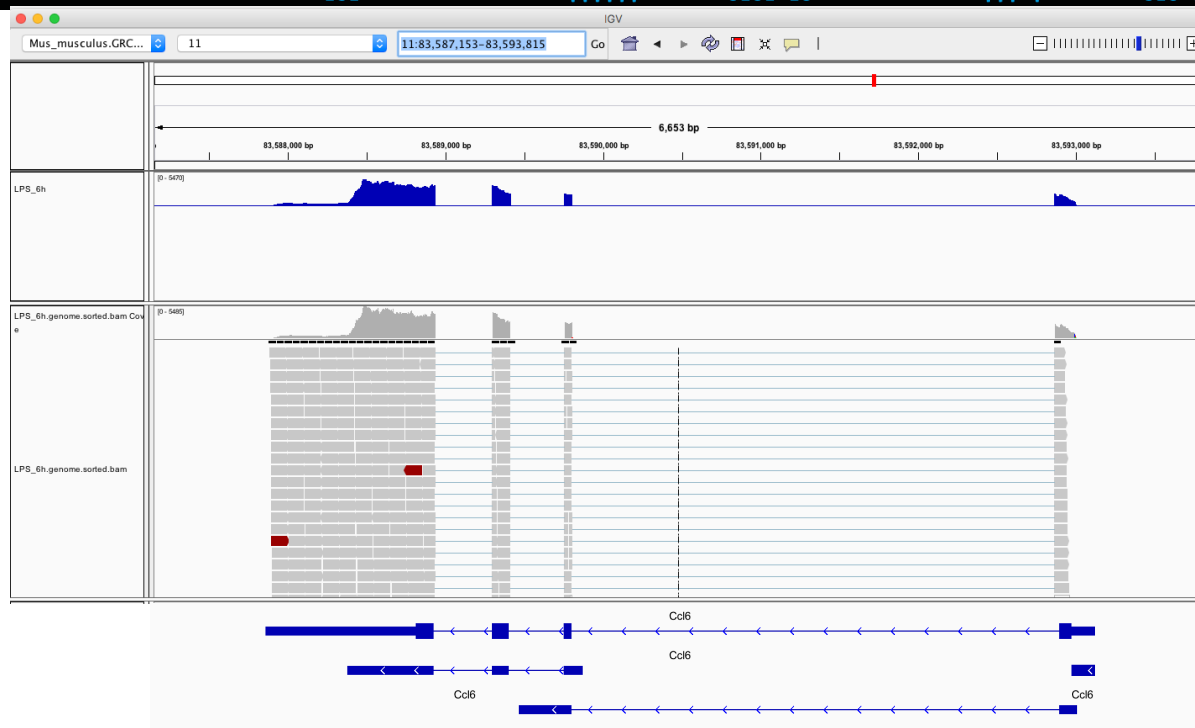
## Example: RSEM



Ccl6 gene locus with 3 isoforms

follows the example:

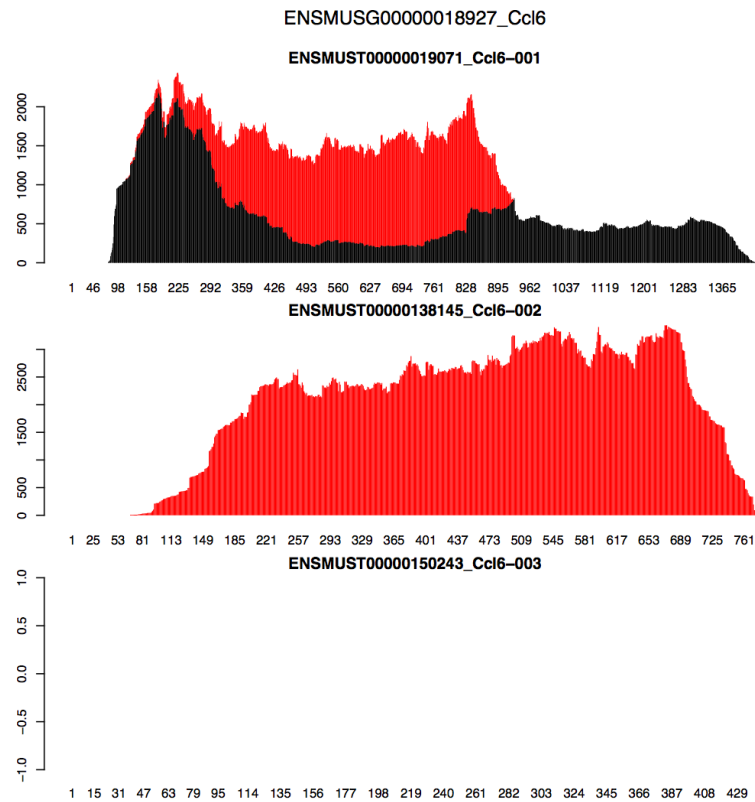
[https://github.com/bli25broad/RSEM\\_tutorial](https://github.com/bli25broad/RSEM_tutorial)



RSEM result:

transcript_id	gene_id	length	effective_length	expected_count	TPM	FPKM	IsoPct	
ENSMUST0000019071_Ccl6-001	ENSMUSG0000018927_Ccl6	1440	1194.85	7805.95	8862.10	9334.46	31.00	
ENSMUST00000138145_Ccl6-002	ENSMUSG0000018927_Ccl6	776	530.94	7719.05	19721.39	20772.55		69.00
ENSMUST00000150243_Ccl6-003	ENSMUSG0000018927_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 of the 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., Airoidi, 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

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

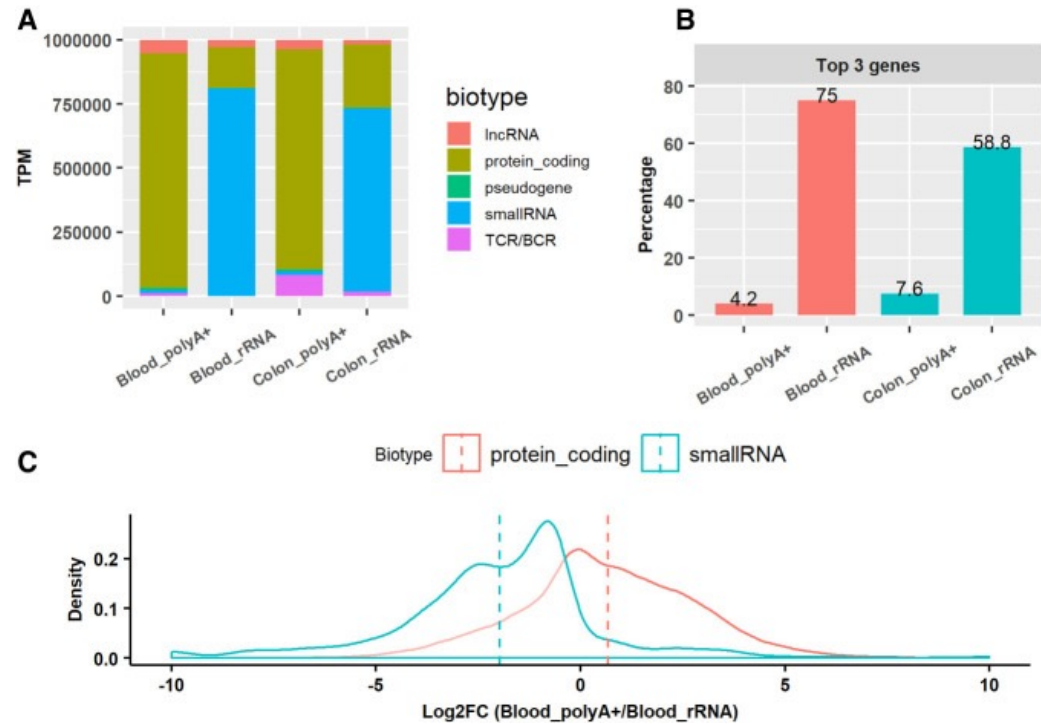
- Relationship

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

## Shortcomings of RPKM and TPM

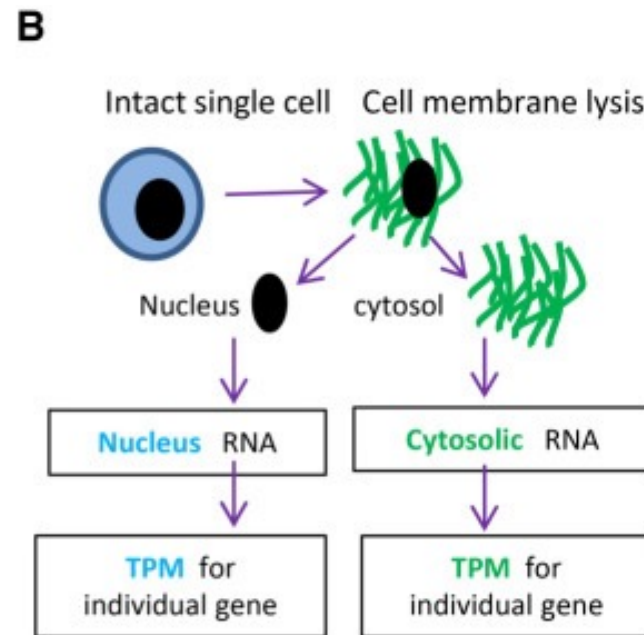
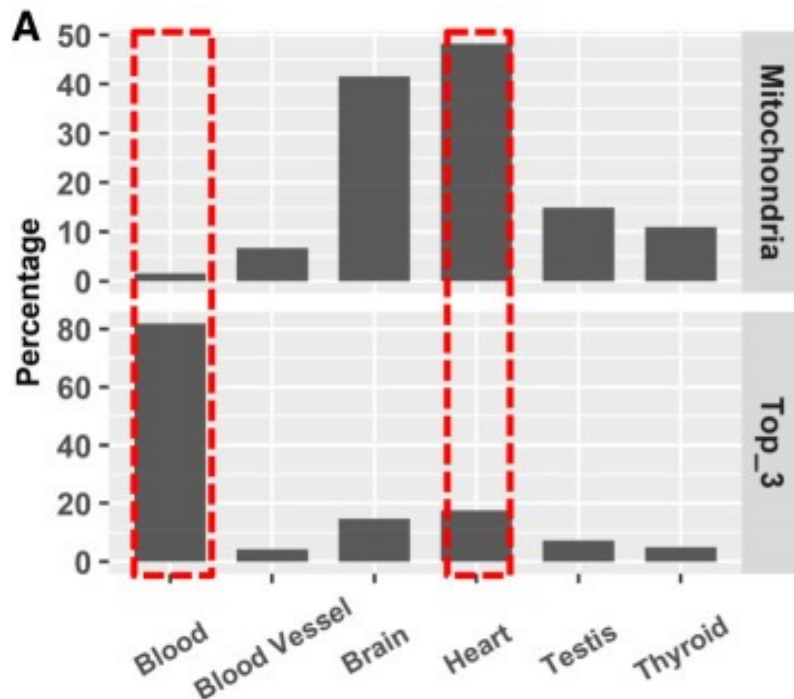
- Sum of RPKM varies from sample to sample, i.e. RPKM is not a measure of relative concentration because the measures of relative concentrations would sum up to constant
- TPM is unitless and satisfies this requirement
- **Only TPM should be used!**
- **But: even TPM is not a suitably normalized measure that can be used to compare samples from**
  - different tissues
  - different protocols

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



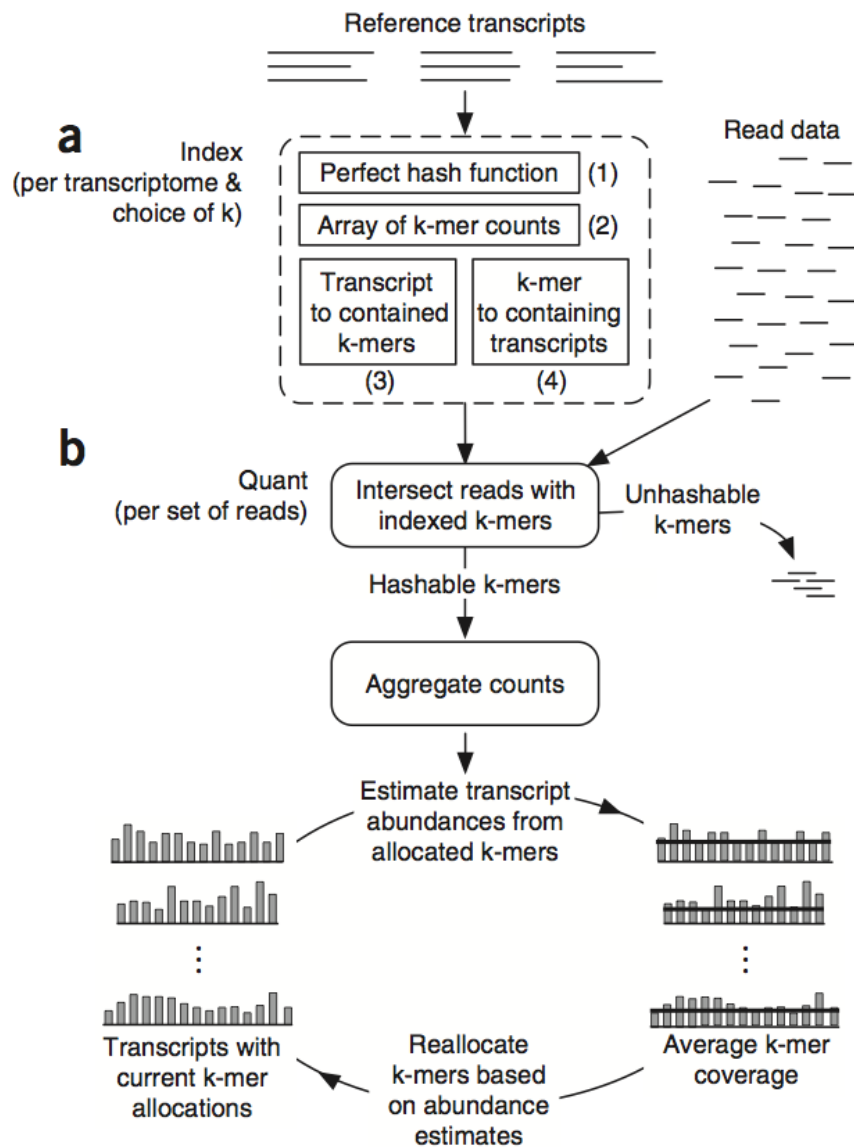
Different tissues may have different populations of genes expressed

# Fast approaches to get the Read-Transcript Compatibility Matrix

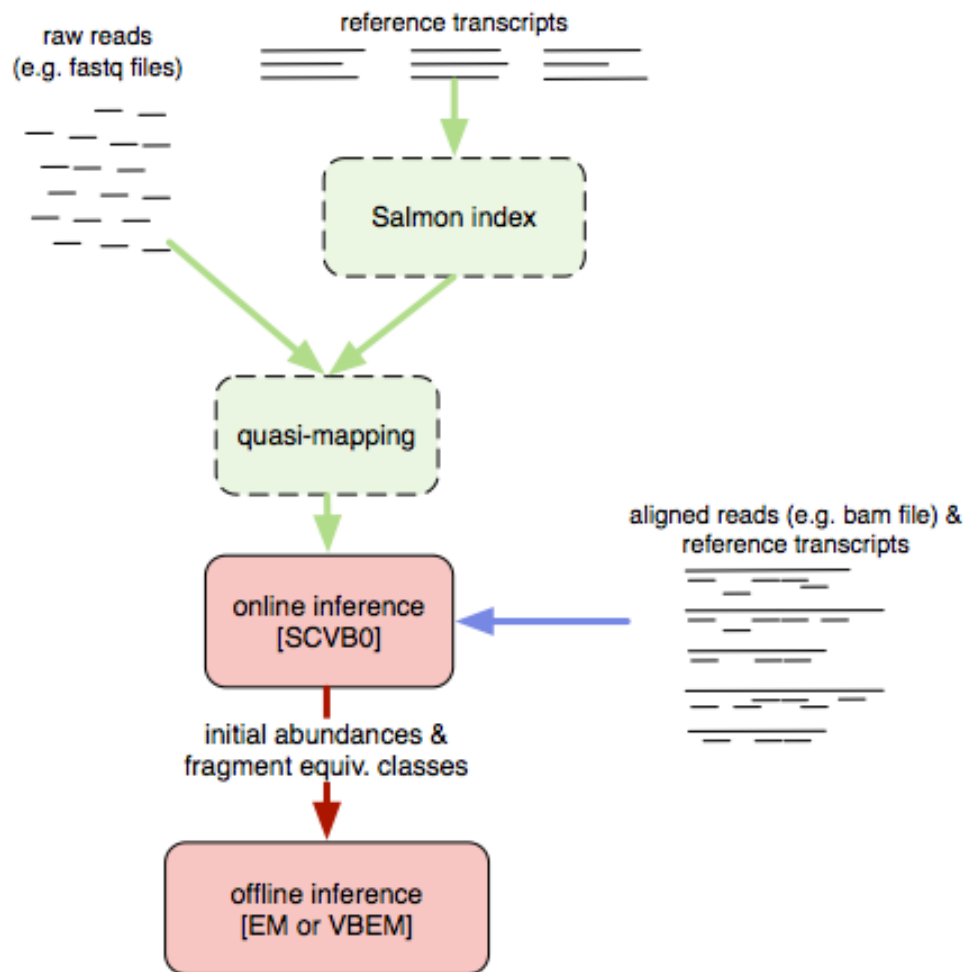
- Sailfish: lightweight alignment
- Salmon: improvement of sailfish
- kallisto: pseudo-alignments

## 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



# Salmon







# Quasi-Mapping

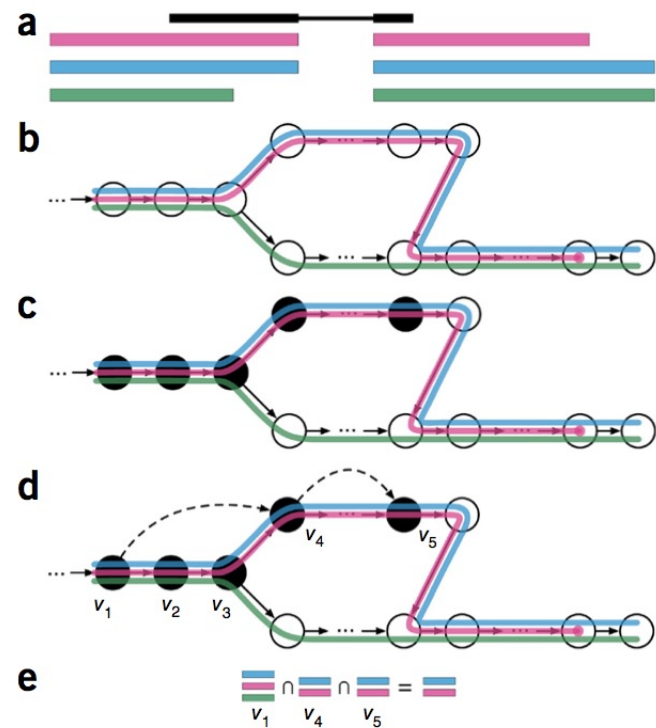
1. The read is scanned from left to right until a k-mer that appears in the hash table is discovered.
2. The k-mer is looked up in the hash table and the SA intervals are retrieved, giving all suffixes containing that k-mer
3. Similar to STAR, the maximal matching prefix (MMP) is identified by finding 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

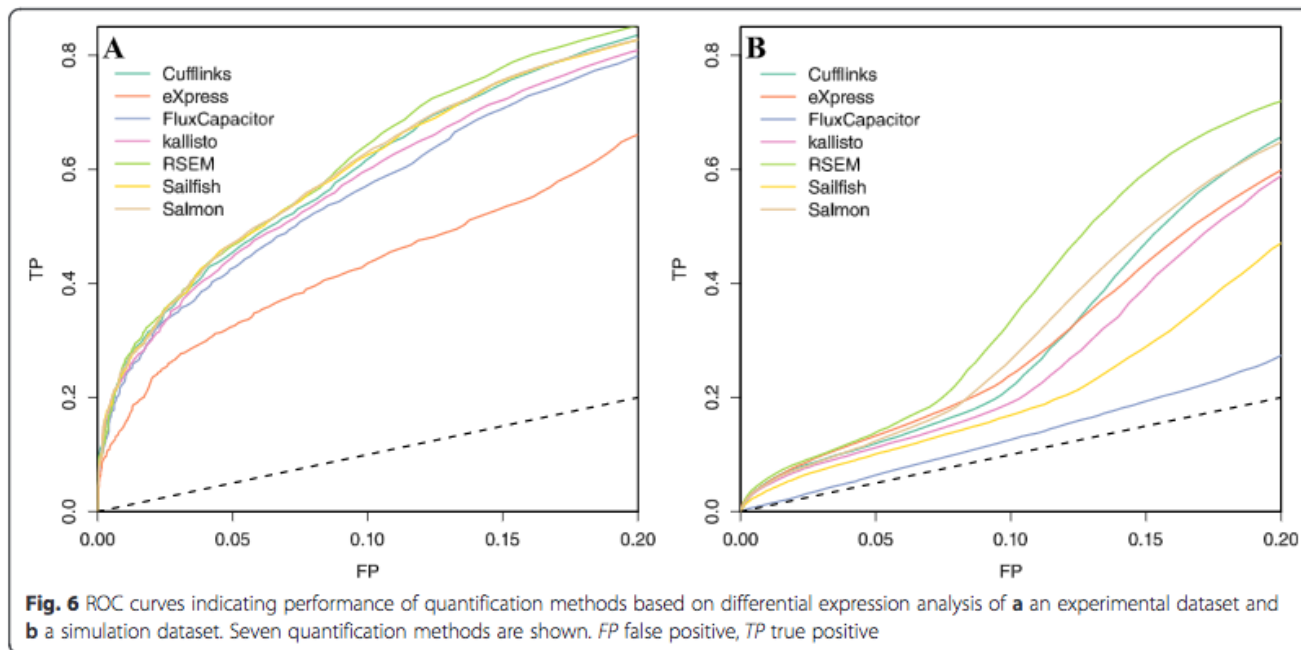
- Result: Read-Transcript compatibility matrix
- Only based on compatibility of short k-mers
- Has an optional step to ***validate mappings***:
  - goes through all the read-transcript associations and validates if the entire read is compatible with the transcripts by doing a base-by-base comparison

# Quantification with pseudo-alignments

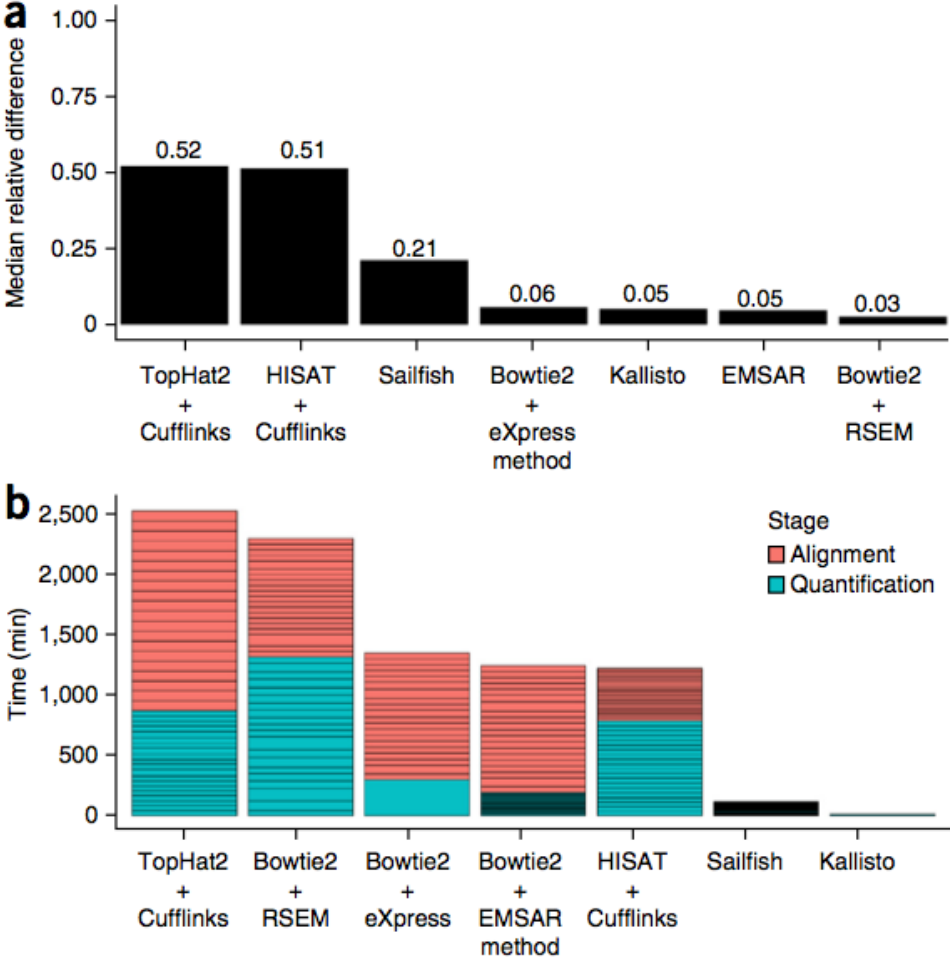
- Instead of hashing the transcriptome build a de Bruijn graph
- Find k-mer hits in the de Bruijn graph
- Identifies only transcripts that are consistent with all k-mer hits



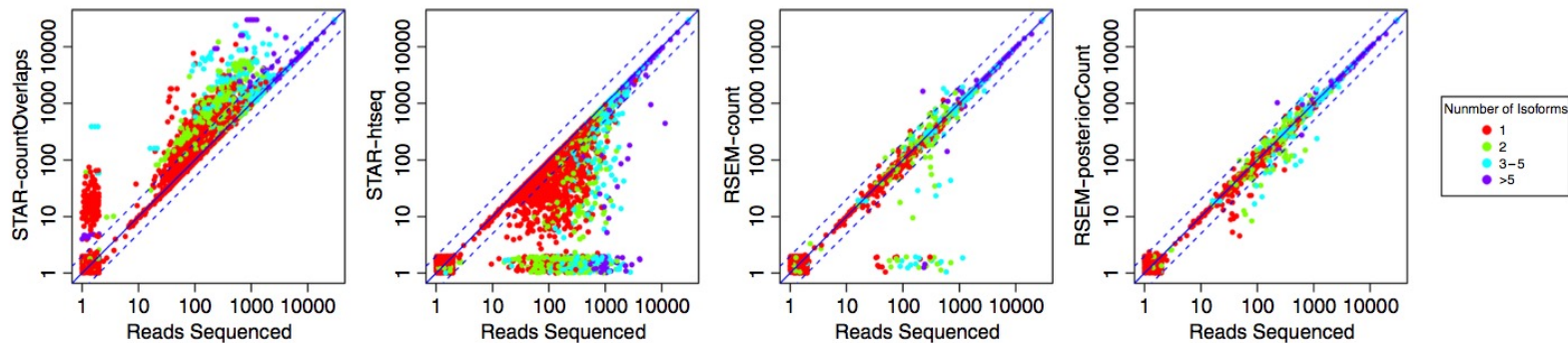
# Performance comparison



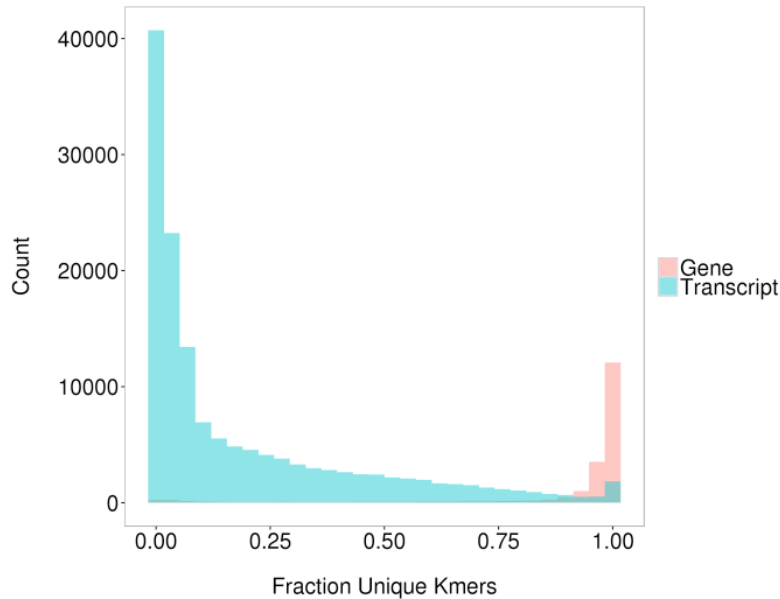
# Performance Comparison



# Read Counting Accuracy



# Uniqueness: Isoform-level vs gene-level



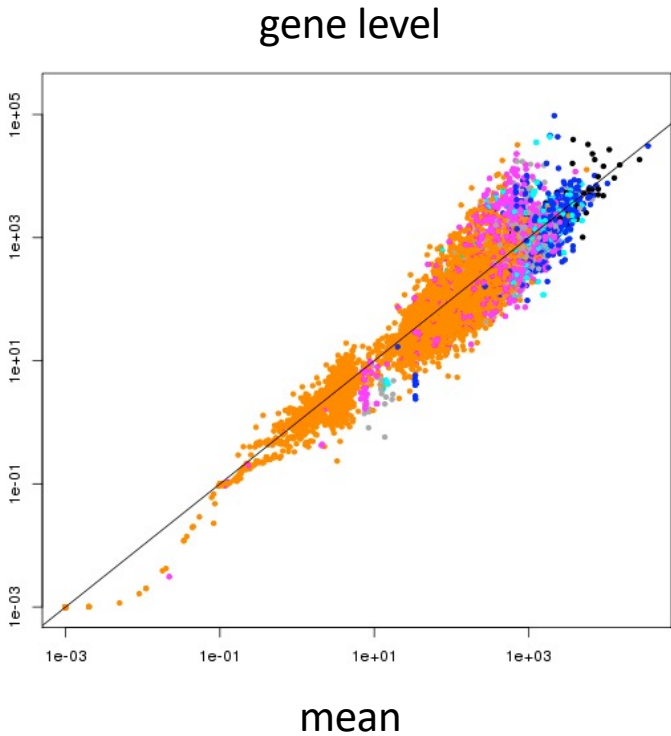
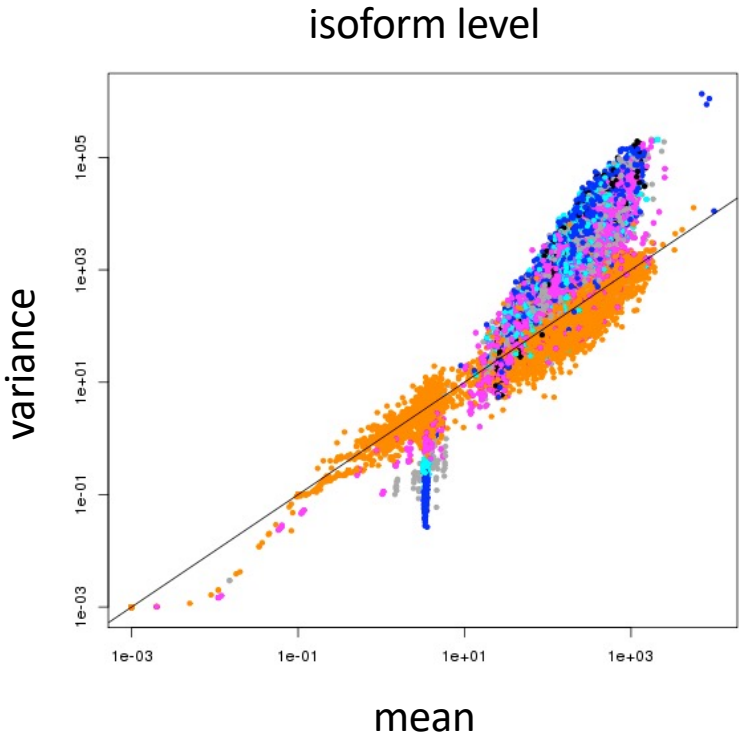
- Fraction of unique k-mer sequences for genes and transcripts
- Ambiguity is mainly between alternative transcripts from the same gene locus







Isoform level has higher variability

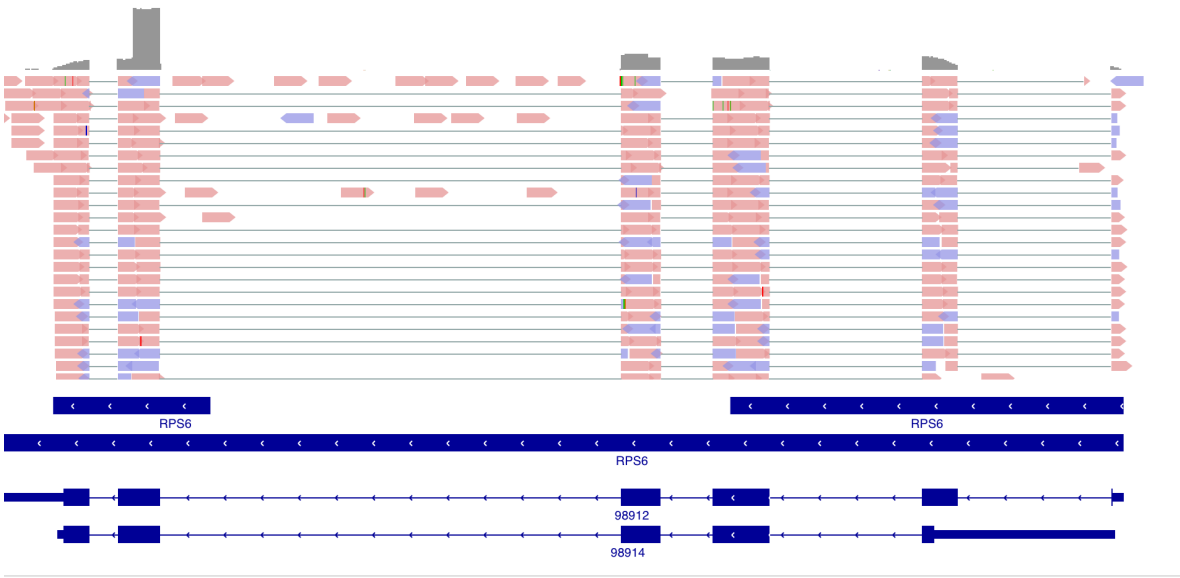


Positional bias of read distributions



# Highly Multiplicated Reads

- Mainly a concern for low starting amounts
- <1ng of total RNA

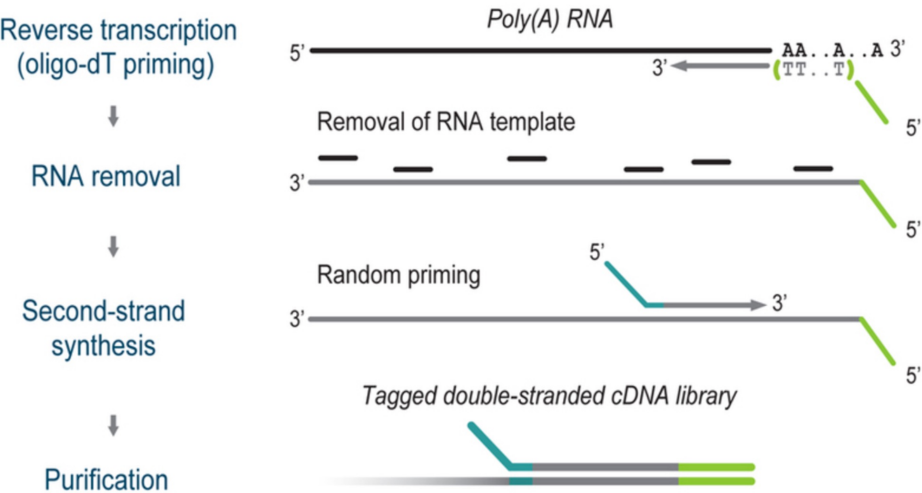


## Unspliced transcripts

- Isoform quantification assumes that only spliced transcripts have been measured
- But: unspliced transcripts are also present:
  - these are transcripts from the nucleus that are not yet spliced
  - limited capturing with poly-A based protocols
  - fully captured by random-priming protocols (1 – 10% of mRNA is in the nucleus)

# 3'-Tagging

- Reads are only generated near the 3'-end
- Isoforms can not be resolved
- Allows counting of the reads at the 3'-end



## Summary

- Pseudo-alignment is reliable and fast but needs as input the accurate and complete set of transcripts