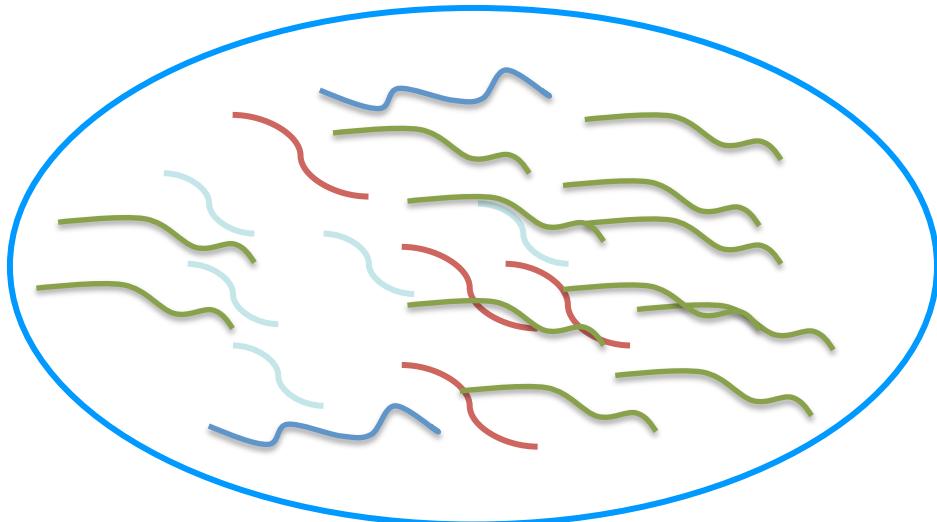


# RNA-seq Quantification

# Hubert Rehrauer



# RNA-seq: A census of RNA molecules



- RNA-seq aims at determining the transcriptional activity at each gene locus

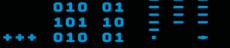
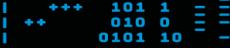
- A mammal cell has 20'000 – 30'000 gene loci
- They can give rise to >100'000 different RNA sequences
- In a typical cell at a given time there are between 100'000 to 500'000 mRNA molecules plus other RNA molecules
- This corresponds to 10-30 picograms
- For a bulk RNA-seq experiment we typically start with >= 100ng from 100'000 or more cells



10

01

101



01

10

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

01

01

10

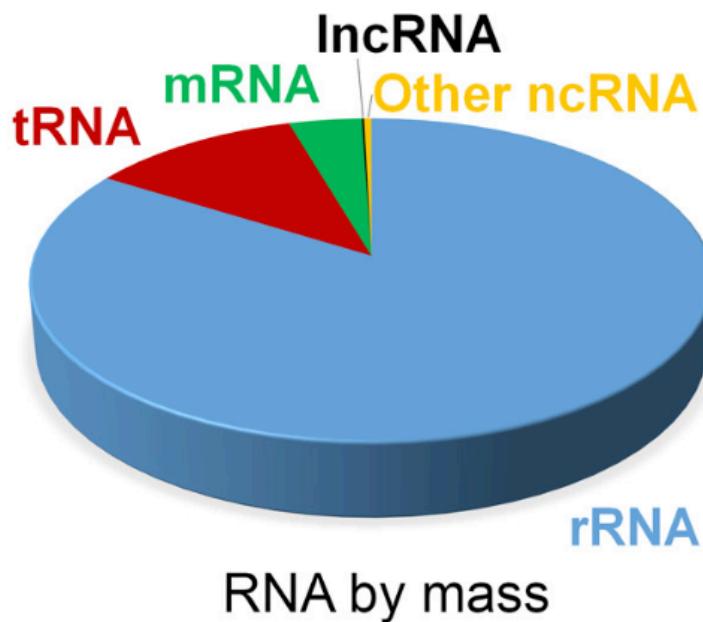
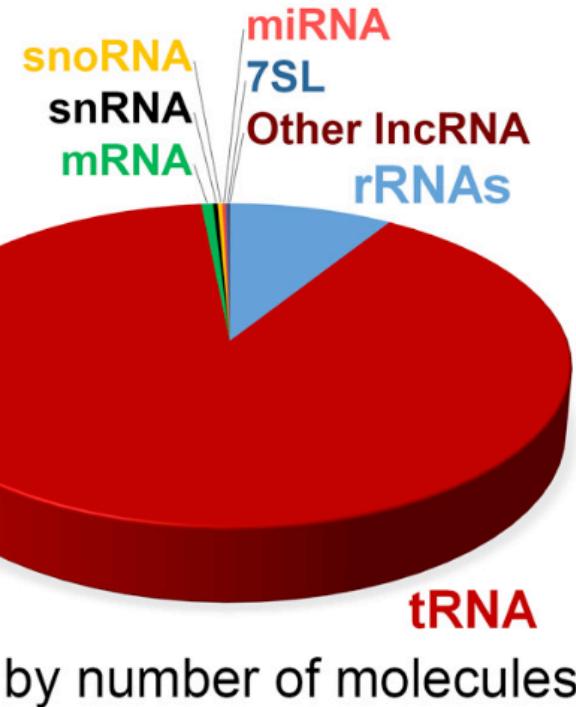
01

01

10

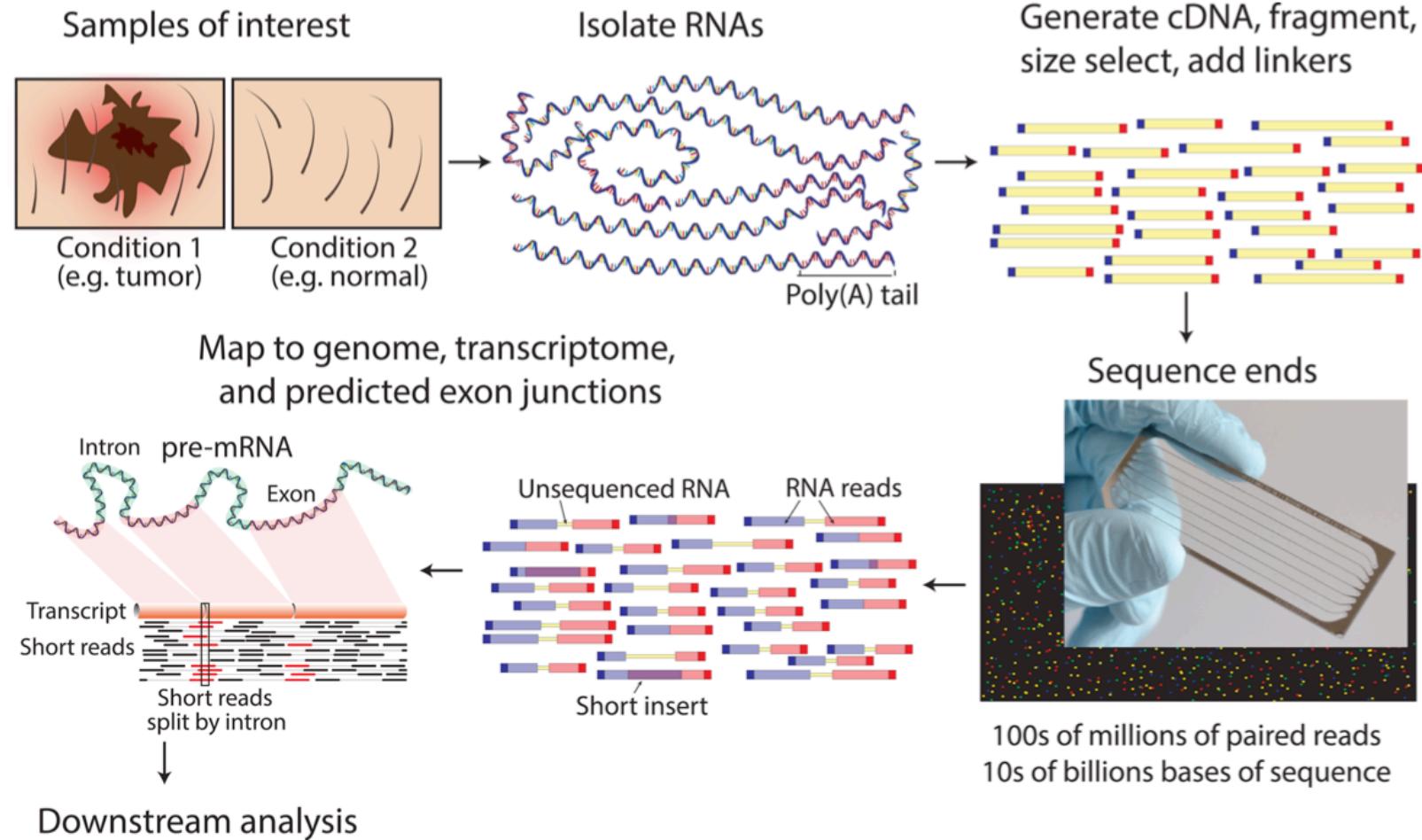
# RNA-seq: A census of RNA molecules

- library preparation includes fragmentation and amplification  
→ billions of RNA fragments
- The sequencing is then a sampling process that “randomly” selects ~ 20 – 50 Mio fragments and determines their nucleotide sequence
- The typical mRNA length is 1000 – 3000 nt
- fragments to be sequenced are in the range 150 – 400 nt

10  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
0110  
01  
01**A****B**

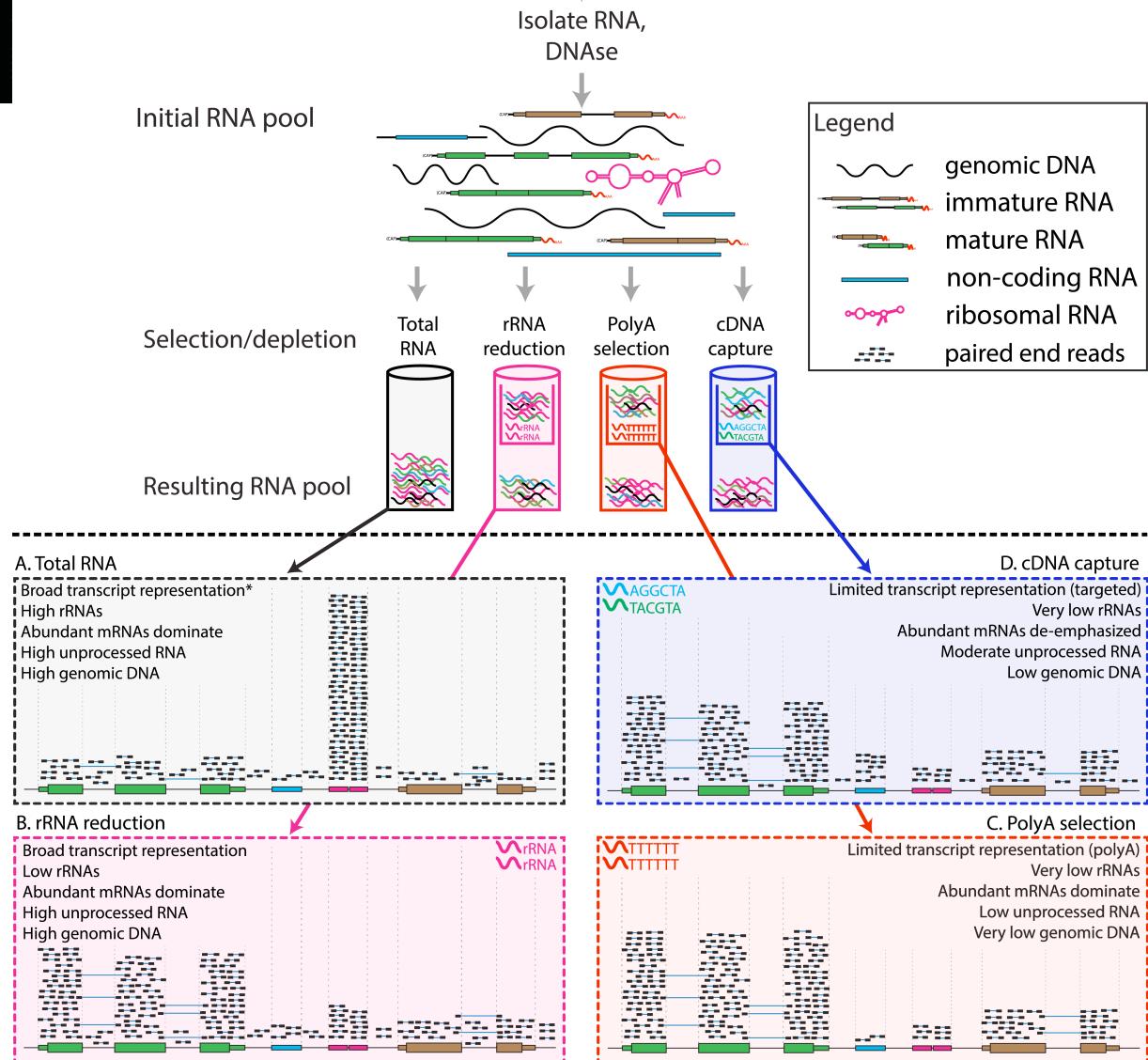
10  
01  
10110  
01  
101101 1  
010 0  
0101 10101 1  
010 10  
010 01101 1  
010 10  
010 01101 1  
010 10  
010 01101 1  
010 10  
010 01101 1  
010 10  
010 01

f g c z

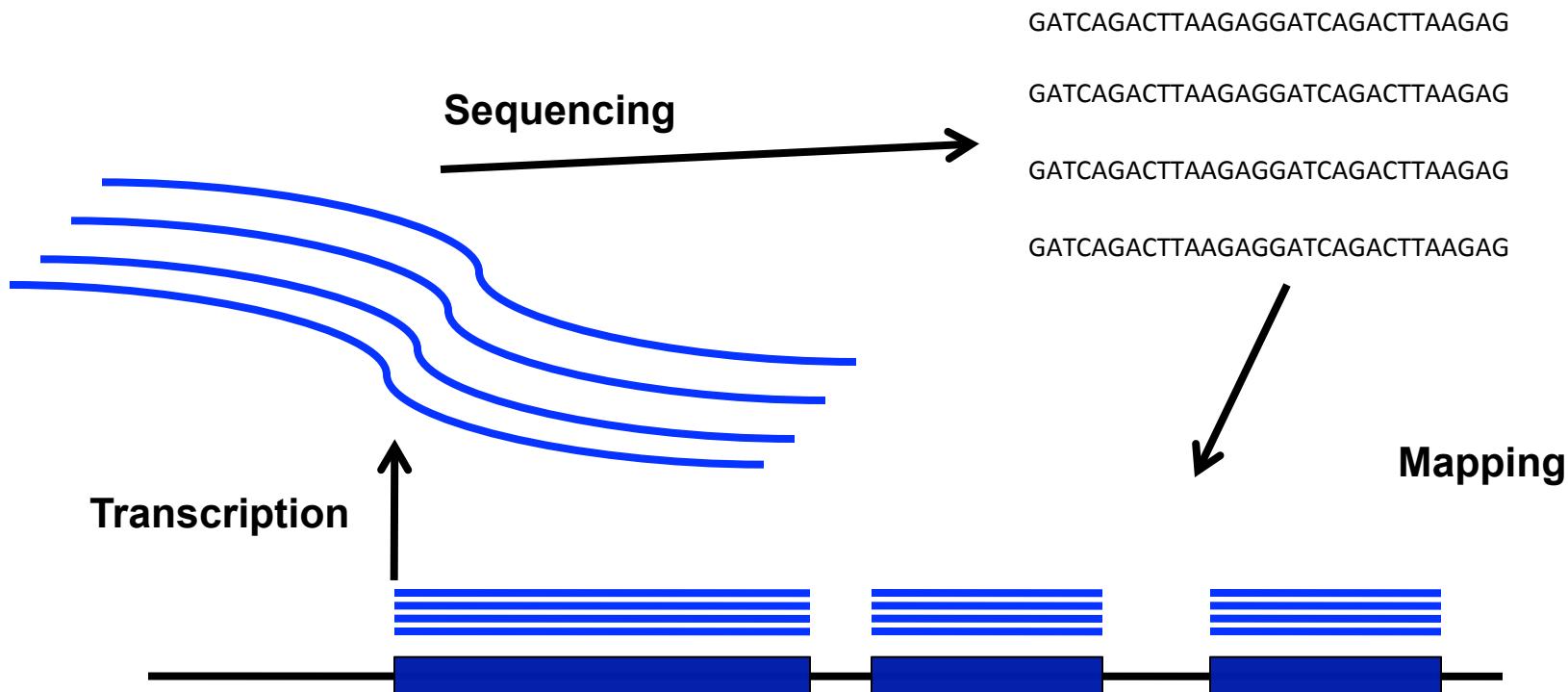
101 1  
010 10  
010 01

# RNA-seq Experiment

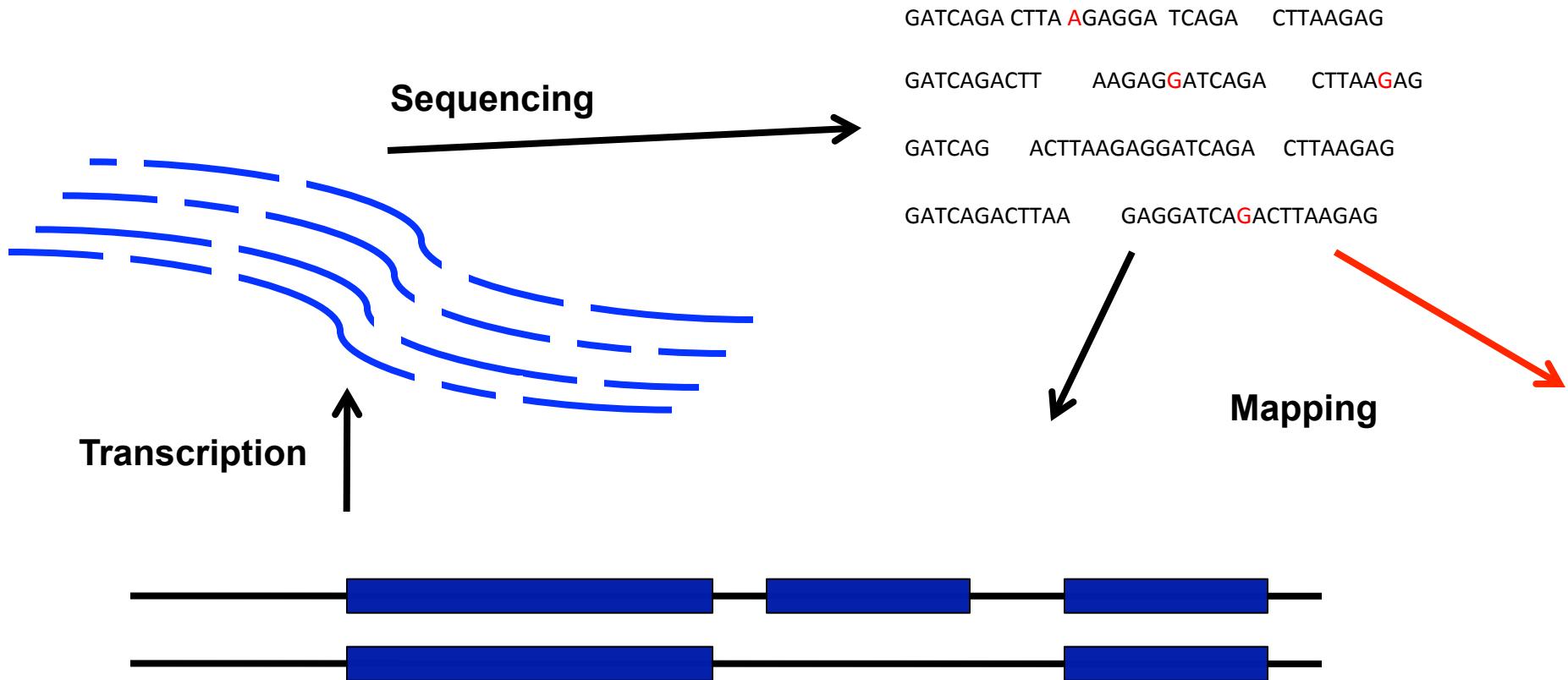
- Note: DNase eliminates the genomic DNA



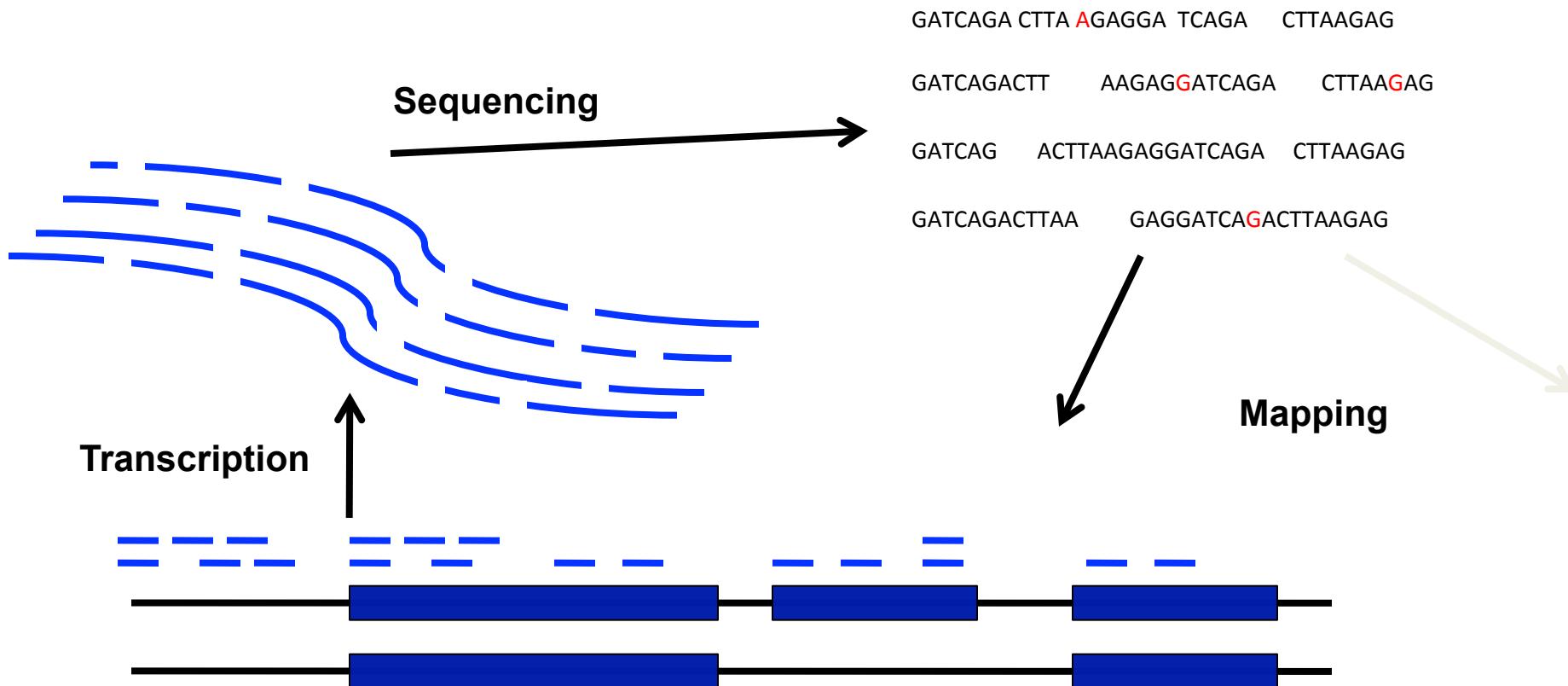
# Read mapping and counting (ideal)



# Read mapping and counting (today)



# Read mapping and counting (today)





10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

# 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**
- Naïve approach for abundance estimate:
  - # reads that uniquely map to a gene locus
    - biased by length, discards information in multi-mapping reads
- More elaborate approaches consider multi-mappers and isoform lengths

10  
01  
10101 10  
01 01  
101 10  
10 01  
01 01  
.. f g c z .. 01 10  
01 10

## RNA-seq comes with absolute counts but relative abundances

Gene	Sample 1 [Bn transcripts] Mio transcripts	Sample 1 [Mio sequenced reads]	Sample 2 [Bn transcripts] 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
<b>total</b>	<b>200</b>	<b>10</b>	<b>500</b>	<b>10</b>

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.

# Number of reads and expression level

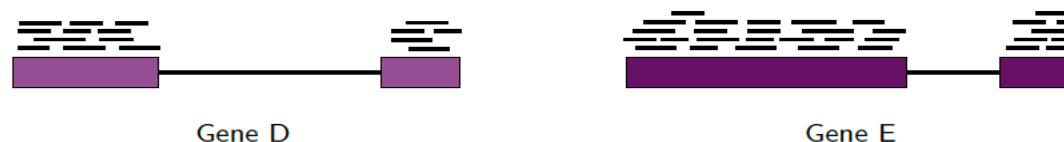
	Sample 1	Sample 2	Sample 3
Gene A	5	3	8
Gene B	17	23	42
Gene C	10	13	27
Gene D	752	615	1203
Gene E	1507	1225	2455

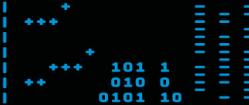
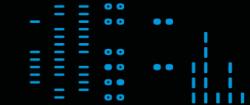
- Gene E has about twice as many reads aligned to it as Gene D
- What does it mean?

1) Gene E is expressed with twice as many transcripts as Gene D



2) Both genes are expressed with the same number of transcripts but Gene E is twice as long as Gene D



10  
01  
101functional genomics center zurich  
010 01  
101 10  
10 10  
010 01  
10 10  
.. f g c z ..  
01 10  
10 01

# Number of Reads per Gene

- The number of reads **sequenced** per gene depends on
  - relative gene expression (relative to the other genes)
  - gene length
  - sequencing depth
- The number of reads that can be **aligned** to a gene depends additionally
  - expression of other genes that have identical sequences
  - sequencing errors
  - allele sequences
  - mapper settings, e.g. STAR

```
--outFilterMultimapNmax 20
```

max number of multiple alignments allowed for a read: if exceeded, the read is considered unmapped



# 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

→ clear where to count the read

→ typically also counted

→ Read with intronic part  
 a.) counts towards expression, because in theory the pre-mRNA can be used as template

} overlapping gene loci

→ not clear where to count to

# Straightforward Counting Modes

- unique reads
  - unique reads + multi-reads randomly assigned
  - unique reads + multi-reads proportionally assigned
  - ...  
  - **Do not count multi-reads towards all genes they align to.**  
In that case the sum of the counts could be larger than the number of reads

# 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.

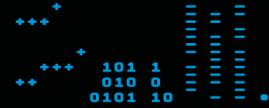
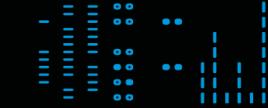
## RNA-seq model

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:

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

10  
01  
101functional genomics center zurich  
010 01  
101 10  
10 01  
010 01  
10 01  
f g c z  
01 10  
01 10

## Definition of expression levels

- Reads Per Kilobase per Million of mapped reads

Not very comparable across species & experiments because this number still depends on the total length of transcripts

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

- Transcripts Per Million Transcripts

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

↳ better choice since we use it to normalise

- Preferable is TPM because it is independent of the transcriptome

# Maximum Likelihood Estimation

- The estimated abundances represent unique MLE estimates

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

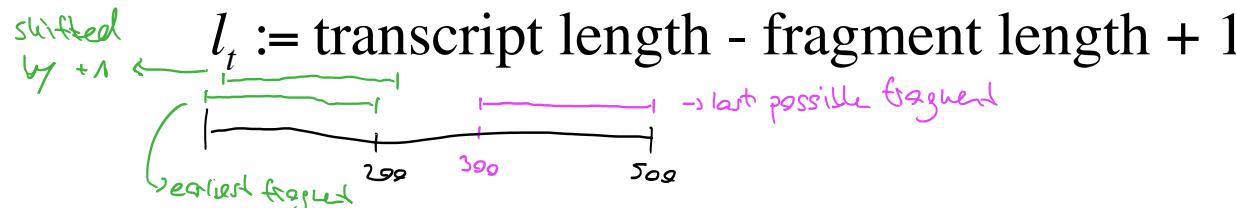
$$L[\alpha] = \prod_{t \in T} \prod_{f \in F_t} P[f \in t] \frac{1}{l_t}$$

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

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

# Effective Transcript Length

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



- 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, ...



10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

10

01

101

# 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
- Percentage of multi-reads is typically low < 5%
- Ignoring multi-reads leads to
  - loss of information
  - biased expression estimates (e.g. for genes in gene families)

# Considering Multi-reads

- Define a compatibility matrix

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

with

↑ transcript  
↳ read

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

- The likelihood is now:

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

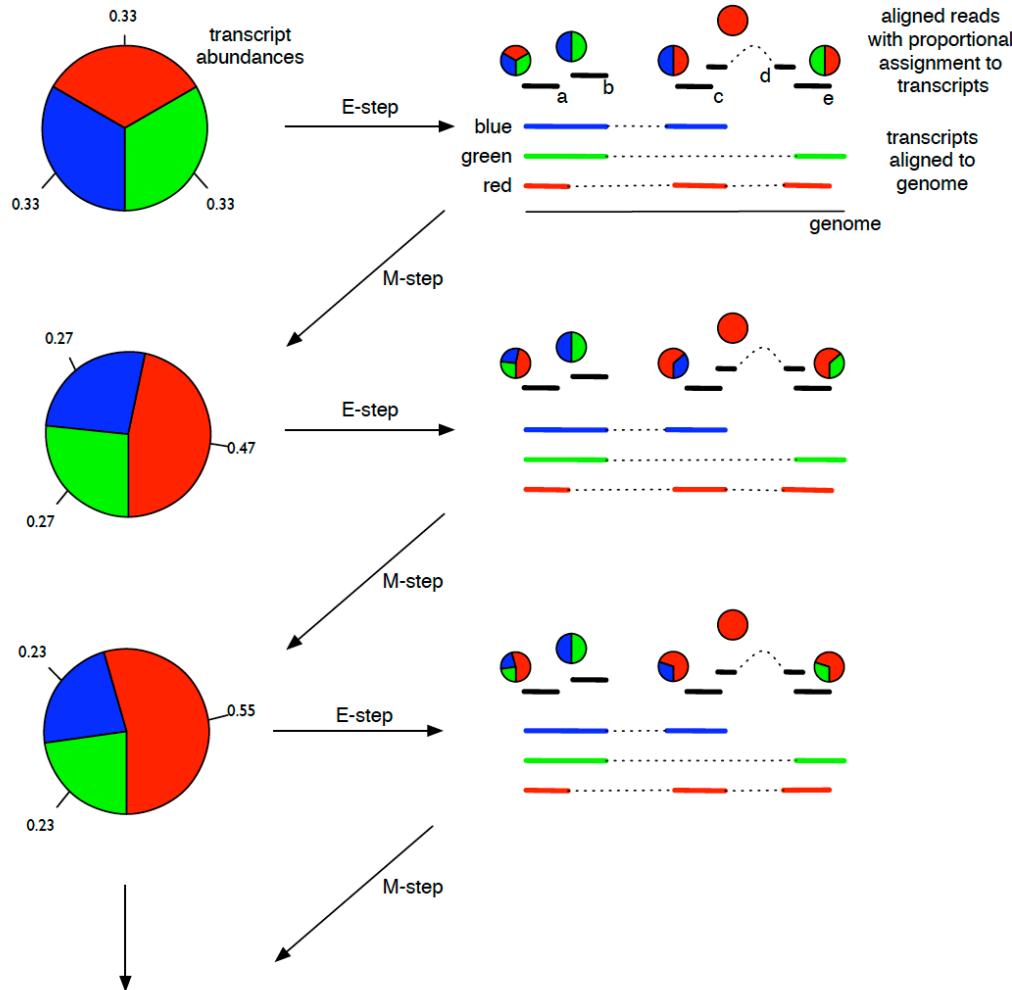
- but now abundances have to be estimated iteratively

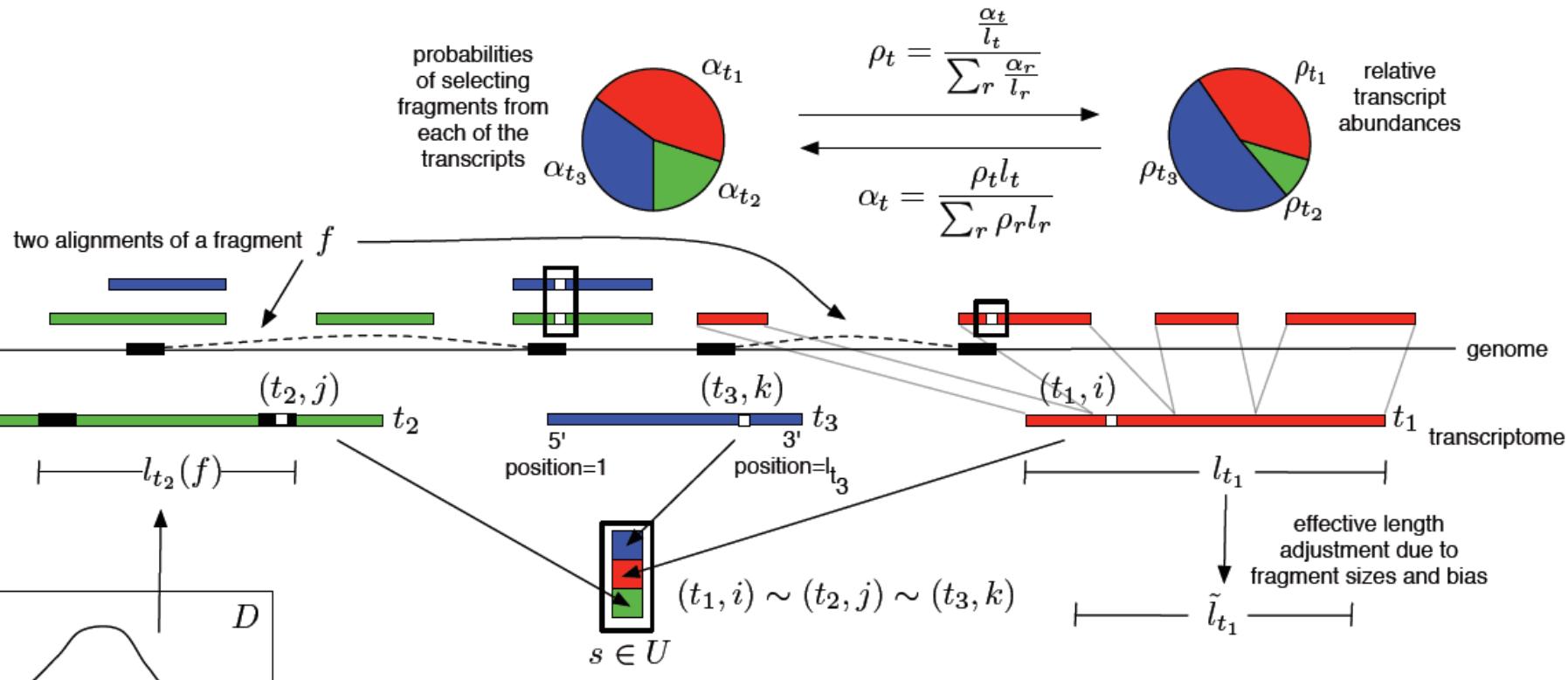
# Iterative Estimation

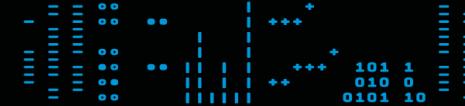
Three step algorithm

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

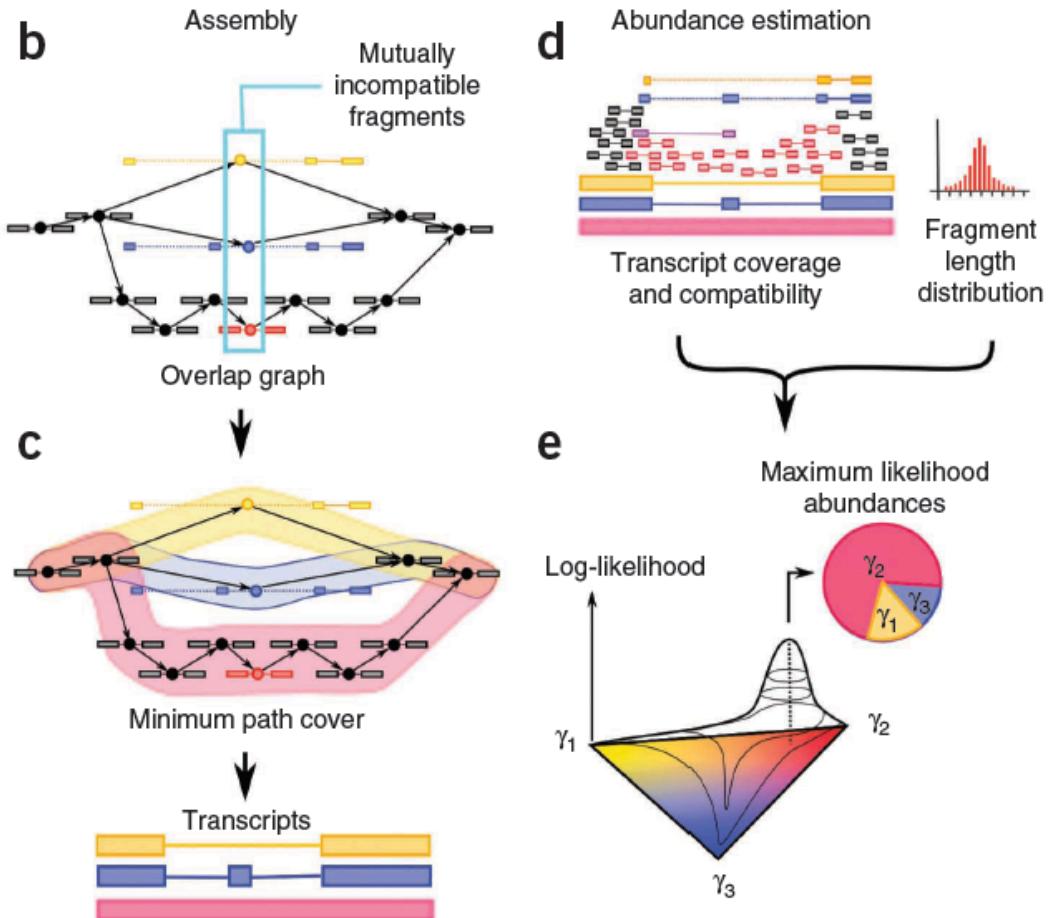
# Expectation Maximization Estimation





10  
01  
0101 10  
10 01  
10 10  
01 01  
01 01  
f g c z  
01 10  
01 10  
01 01  
01 01

# Transcript abundance estimation with Cufflinks





10

01

101

010

01

101

10

01

10

01

10

01

10

01

10

01

10

01

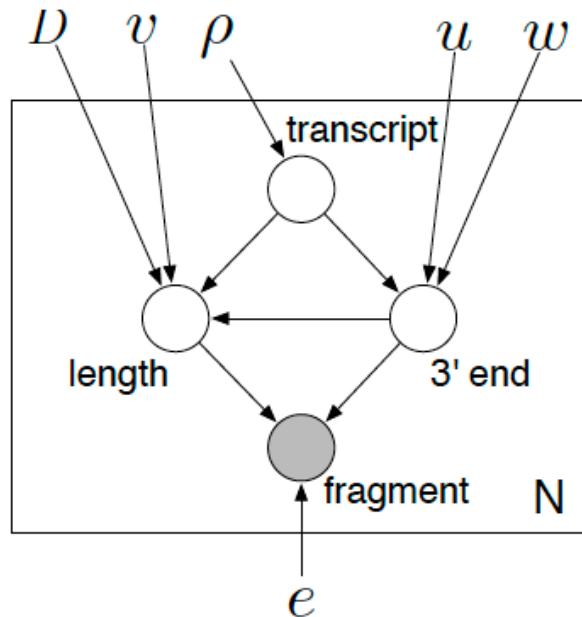
# General Formulation of Abundance Estimation

A full model for the abundance estimation should consider:

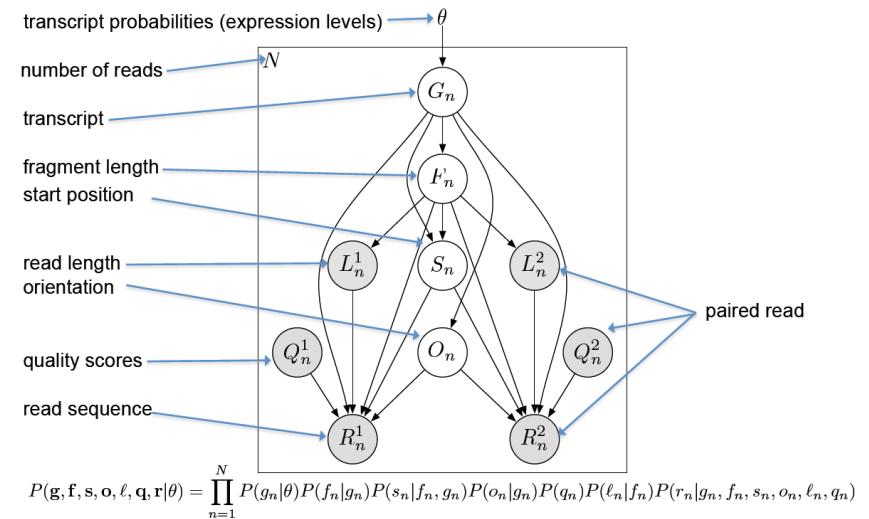
- position bias
- fragment-length distribution
- sequencing errors
- site-specific bias
- ...

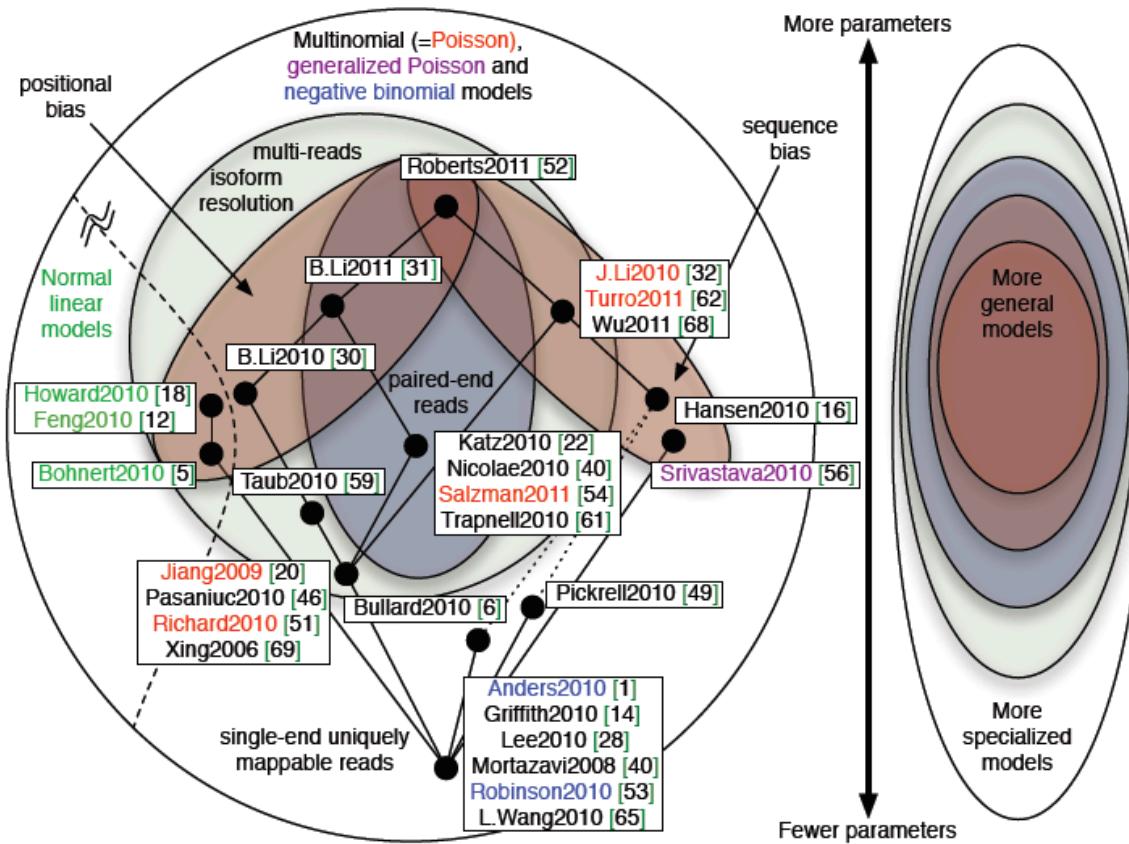
# Example Implementations

## Pachter: Cufflinks



## Dewey: RSEM





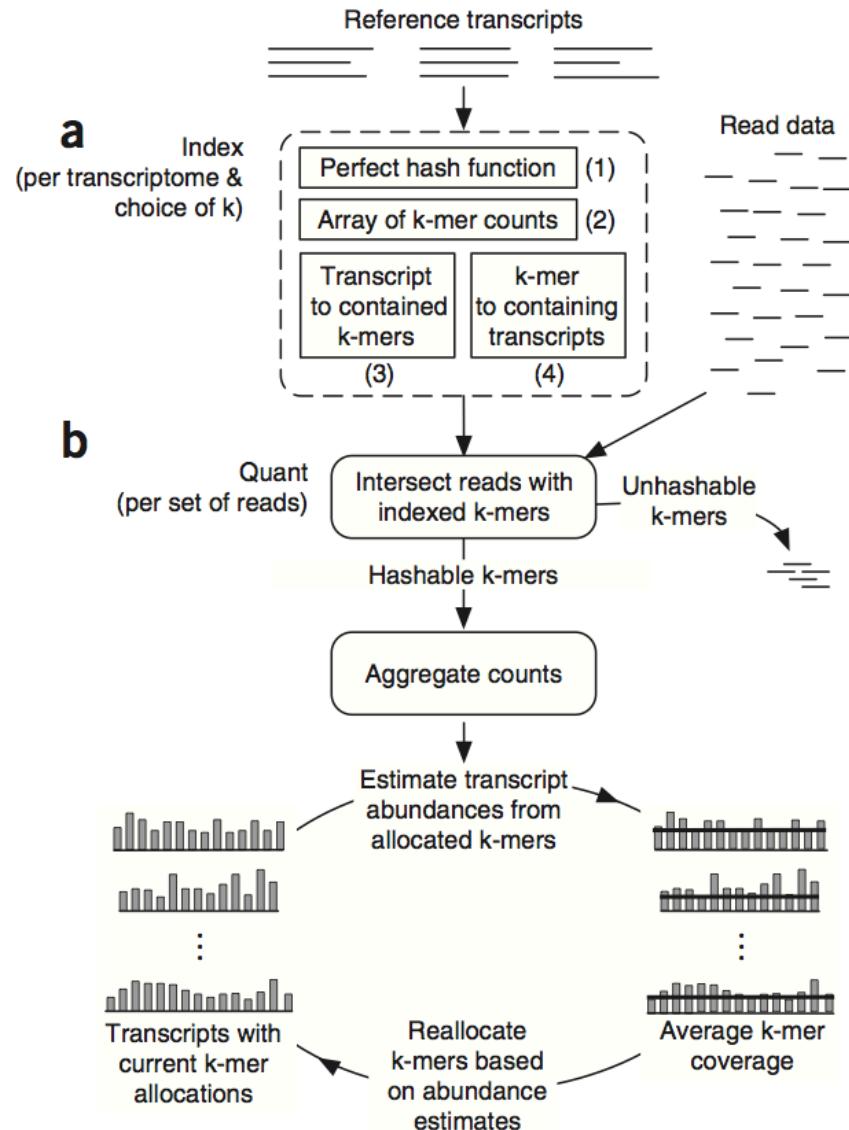
10  
01  
101101 1  
010 0  
0101 10101 01  
101 10  
10 10010 01  
101 10  
10 1010 10  
10 10  
01 0110 10  
10 10  
01 0110 10  
10 10  
01 0110 10  
10 10  
01 0110 10  
10 10  
01 0110 10  
10 10  
01 01

# Fast approaches to get the Read-Transcript Compatibility Matrix

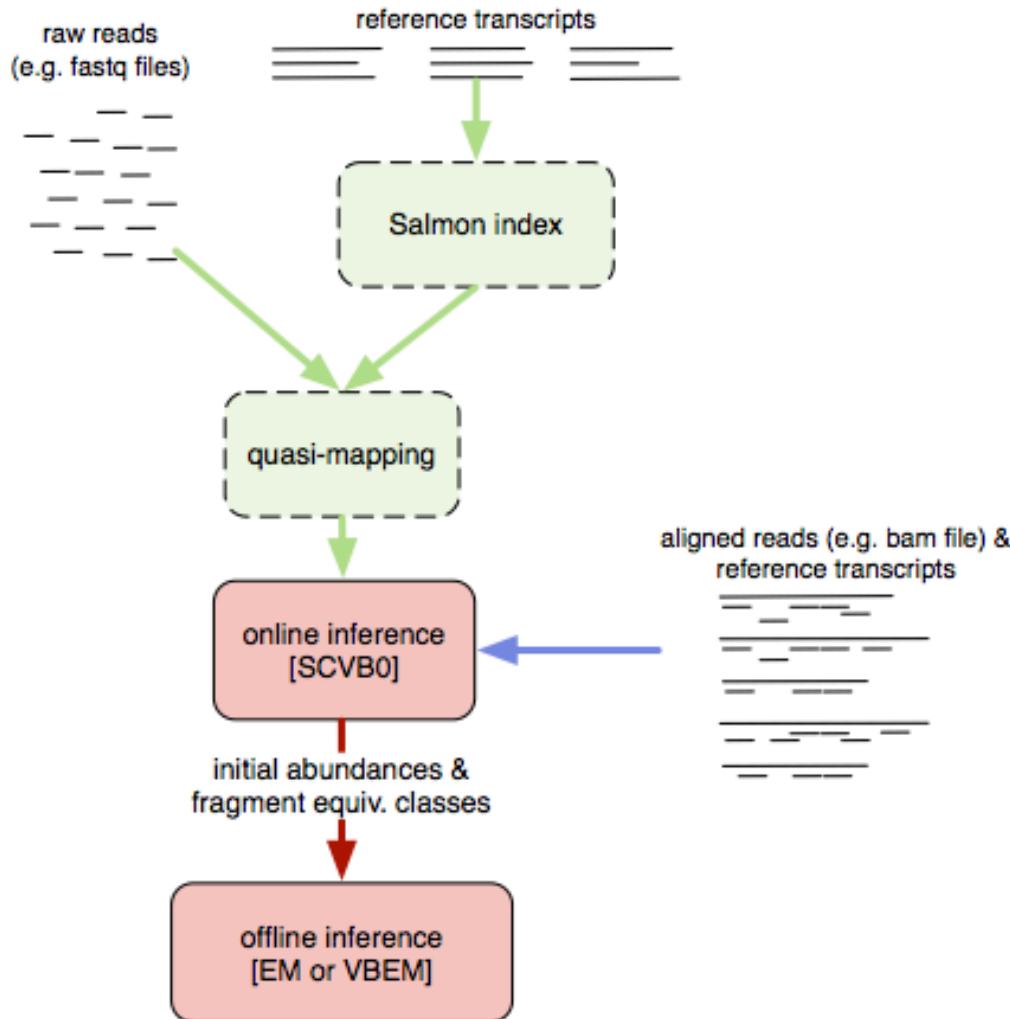
- Sailfish: lightweight alignment
- Salmon: improvement of sailfish
- kallisto: pseudo-alignments

# 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



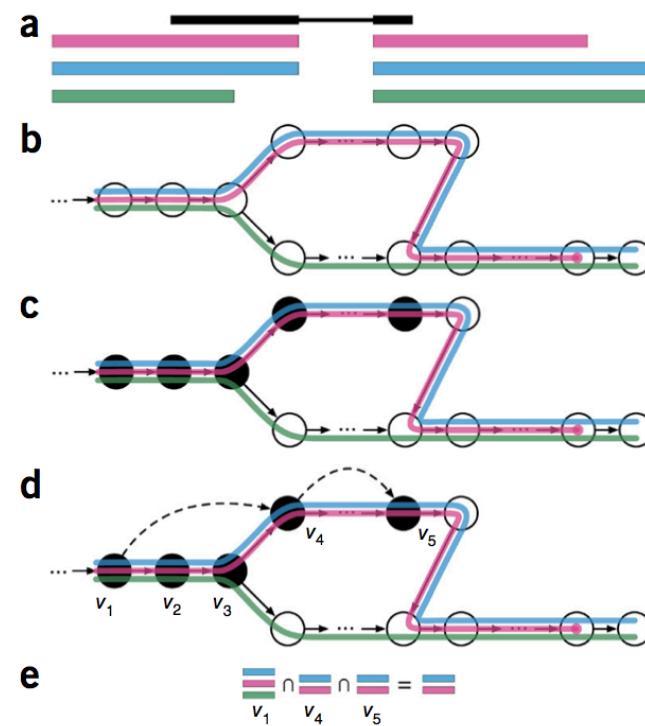
# Salmon



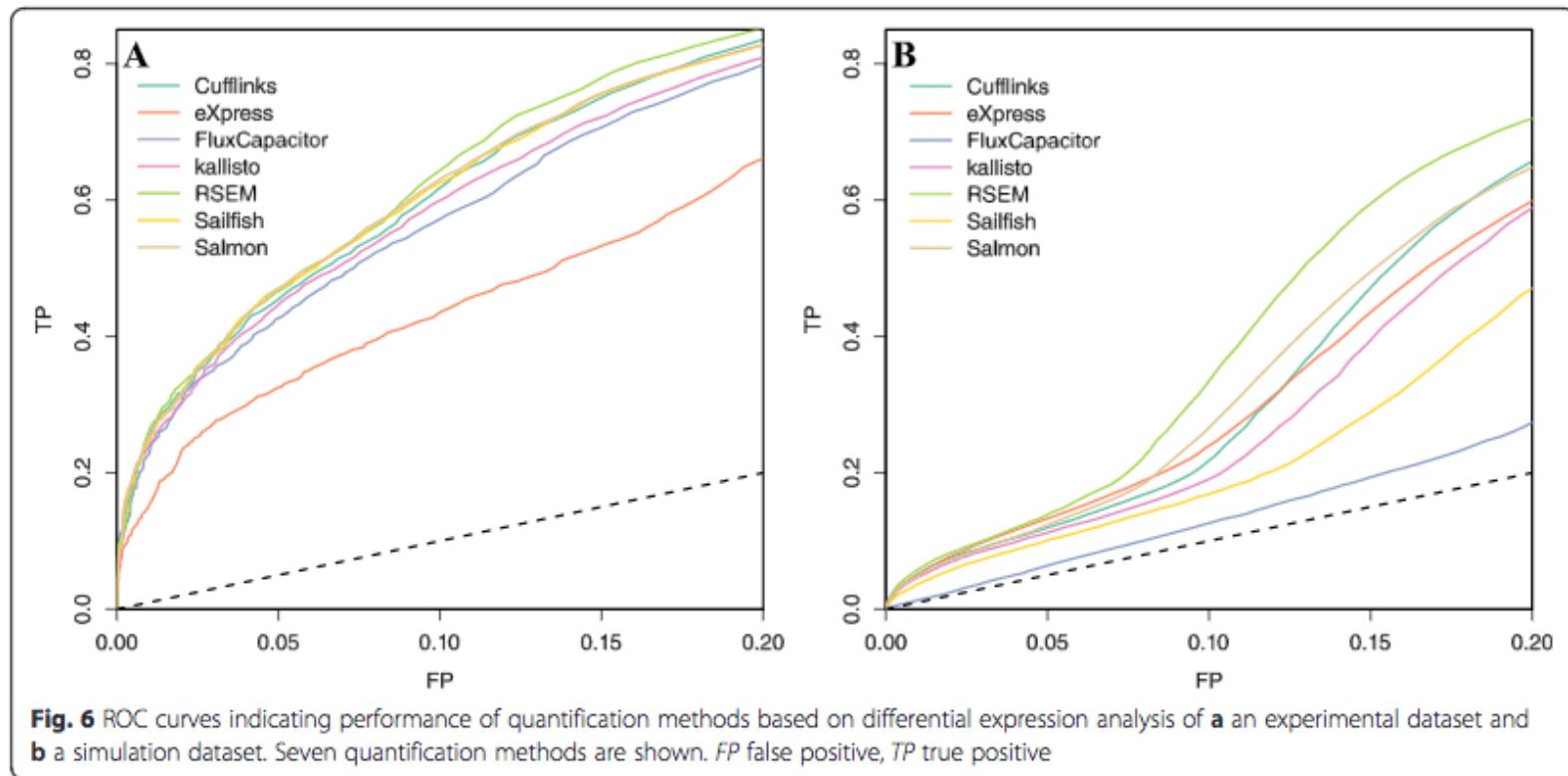
10  
01  
101

# 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

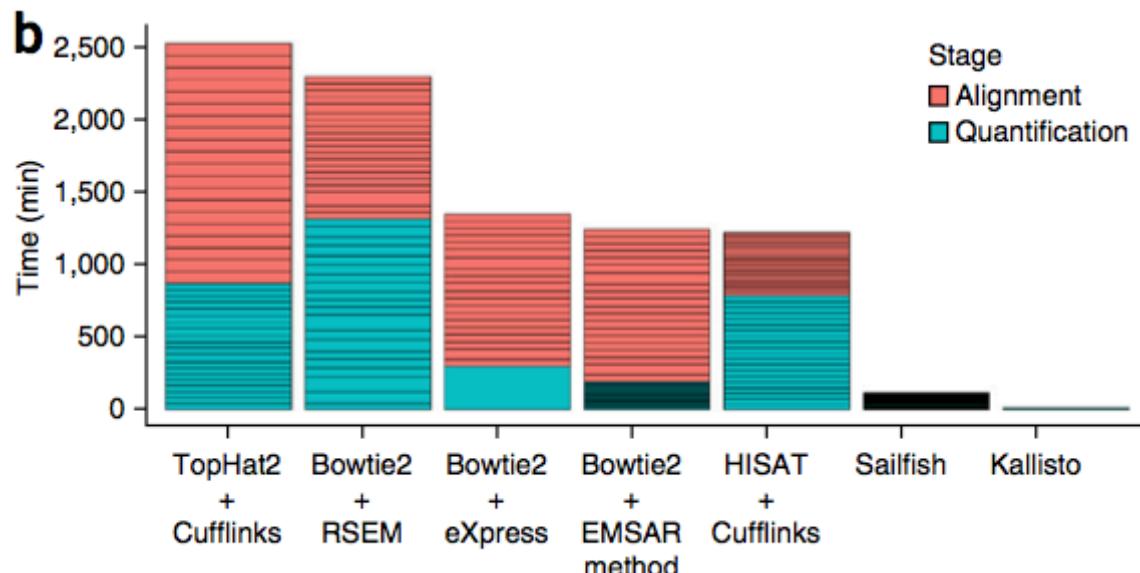
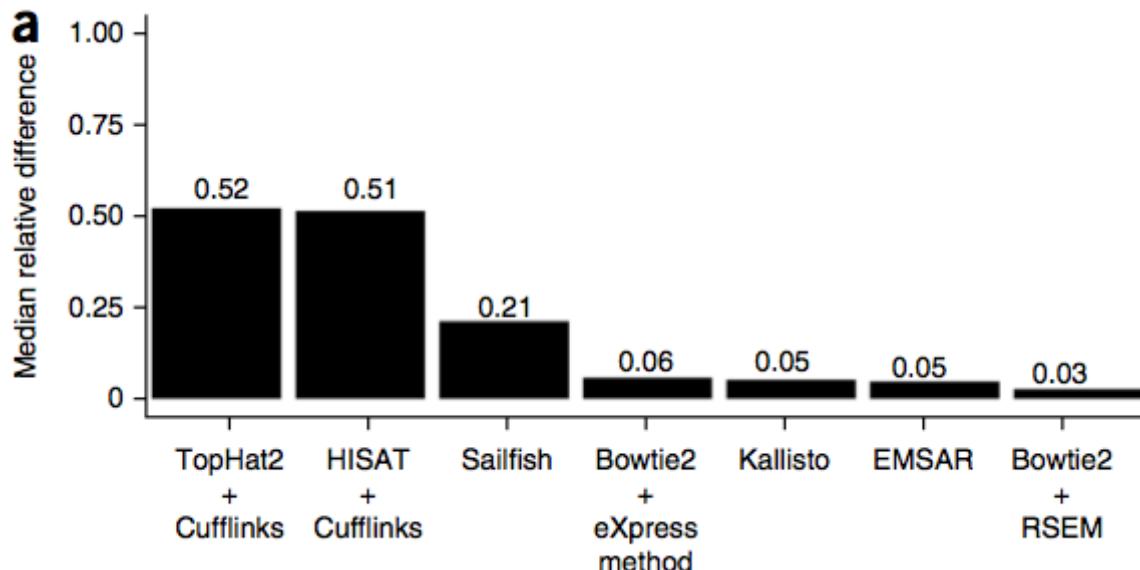


# Performance comparison

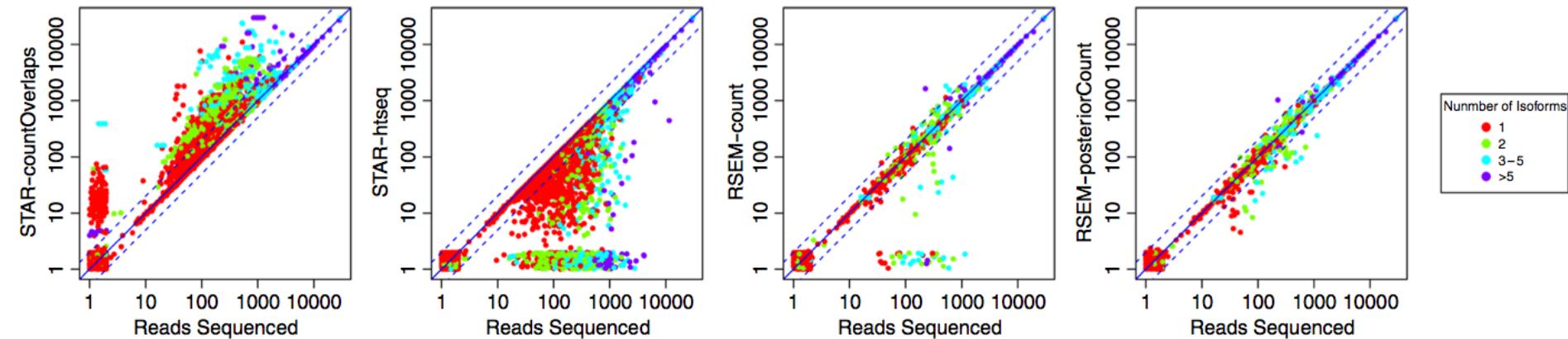


**Fig. 6** ROC curves indicating performance of quantification methods based on differential expression analysis of **a** an experimental dataset and **b** a simulation dataset. Seven quantification methods are shown. *FP* false positive, *TP* true positive

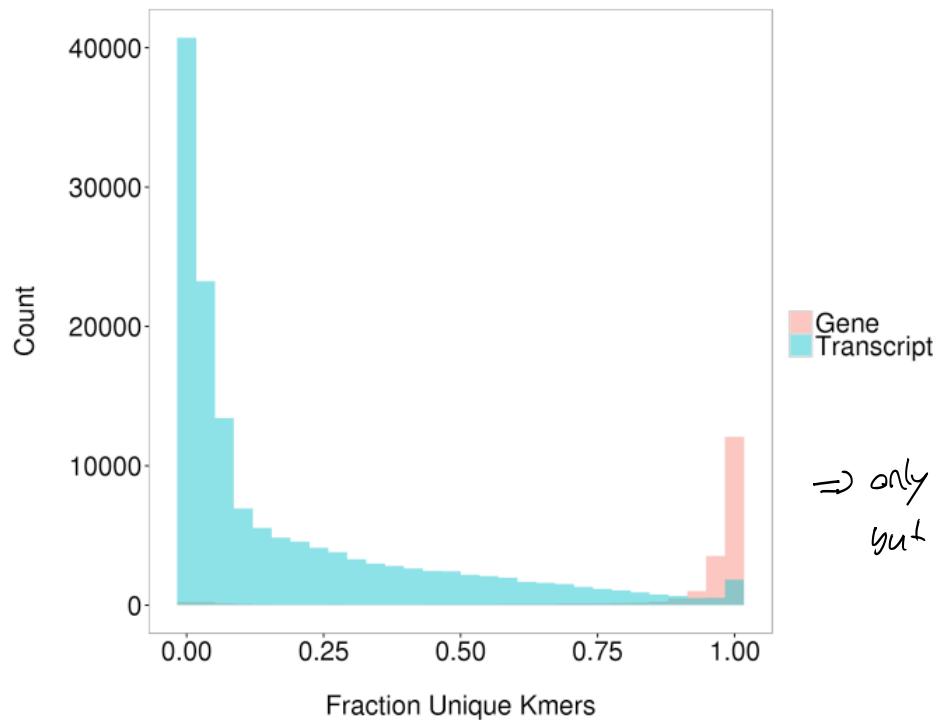
# Performance Comparison



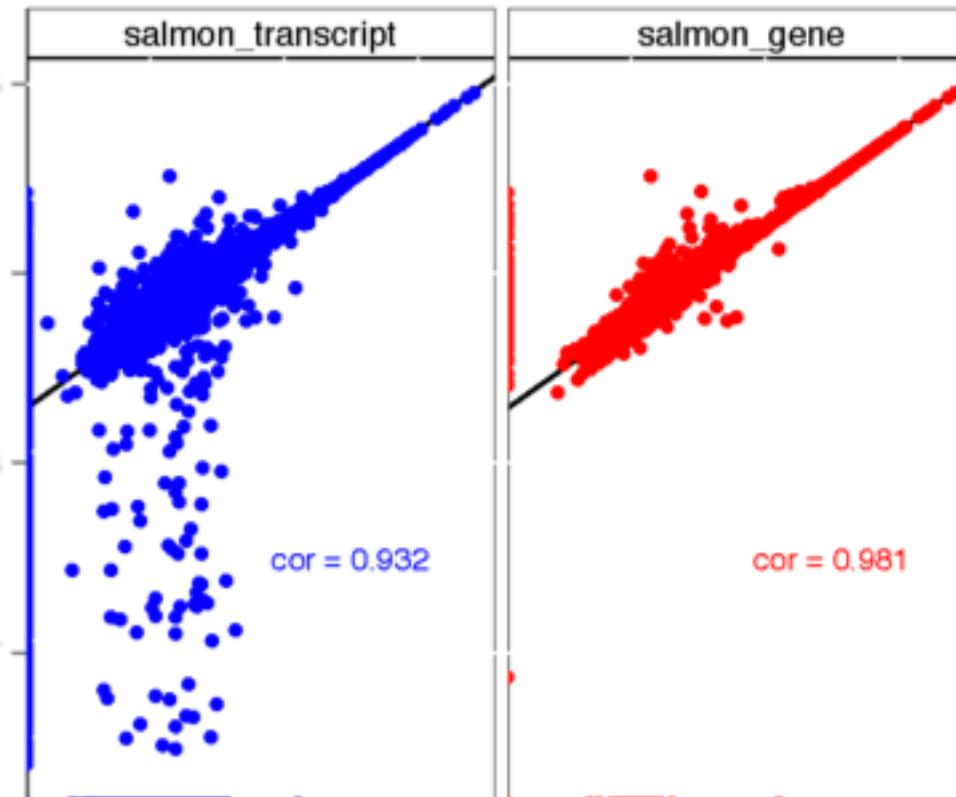
# Read Counting Accuracy



## Uniqueness: Isoform-level vs gene-level

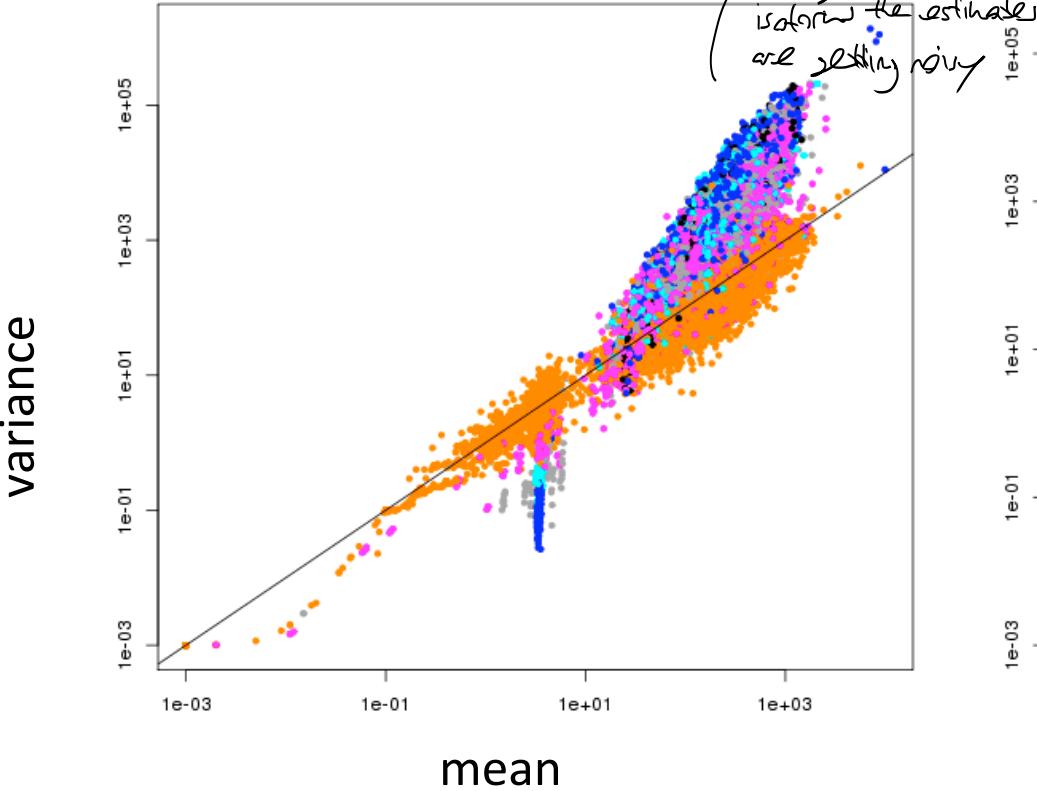


# Accuracy: Isoform-level vs gene-level

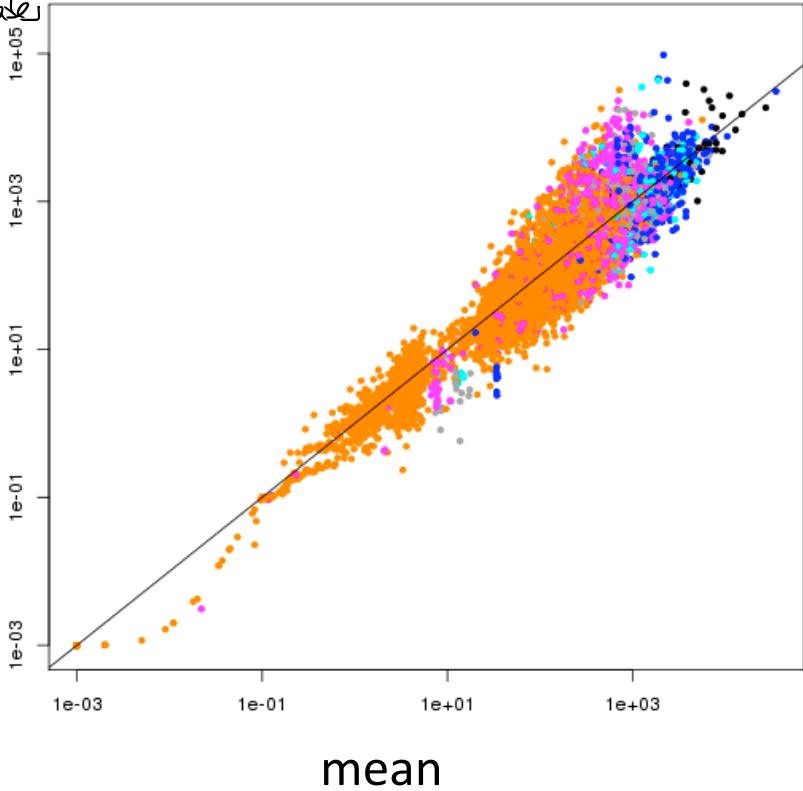


# Isoform level has higher variability

isoform level



gene level





10

01

101

10

01

0101



01

01

010

01

01

101

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

10

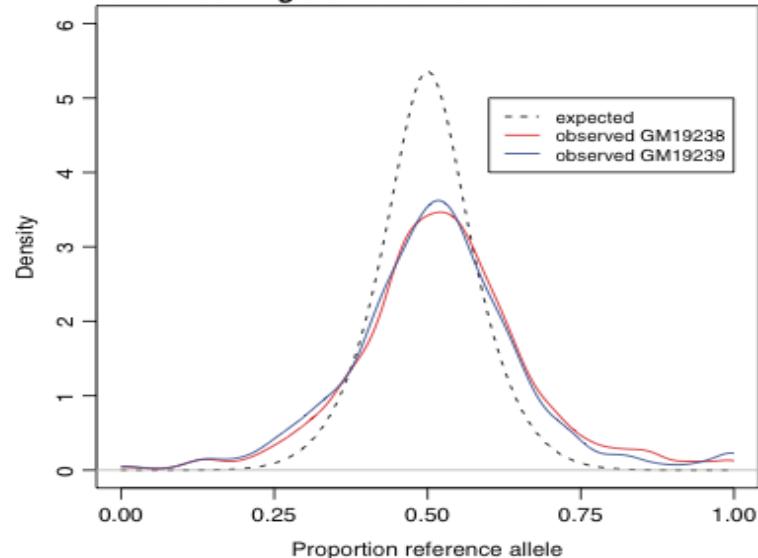
10

# Biological Issue: SNPs

- SNPs may lead to false positives in differential expression because fewer reads map to non-reference alleles
- Note: Sequencing enables allele-specific expression values

A

*Distribution across SNPs of the proportion of reads matching the reference allele*





# Cufflinks and Related

→ Worth reading

- 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

10  
01  
101

- **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).