# RNA-seq Quantification

# Outline

- the RNA-seq setting
- normalizing fractions
- short-read RNA-seq
- quantification:
  - read counts
  - generative models; relative abundances
  - pseudo-mapping
- error models
- analysis at log-scale
  - additive background
  - more complex models

# Typical setting for bulk RNA-seq

Given a tissue or cell line:

- ~25’000 genes in the genome
- ~10 – 18’000 genes expressed at any given time; expression level of the different genes differs largely

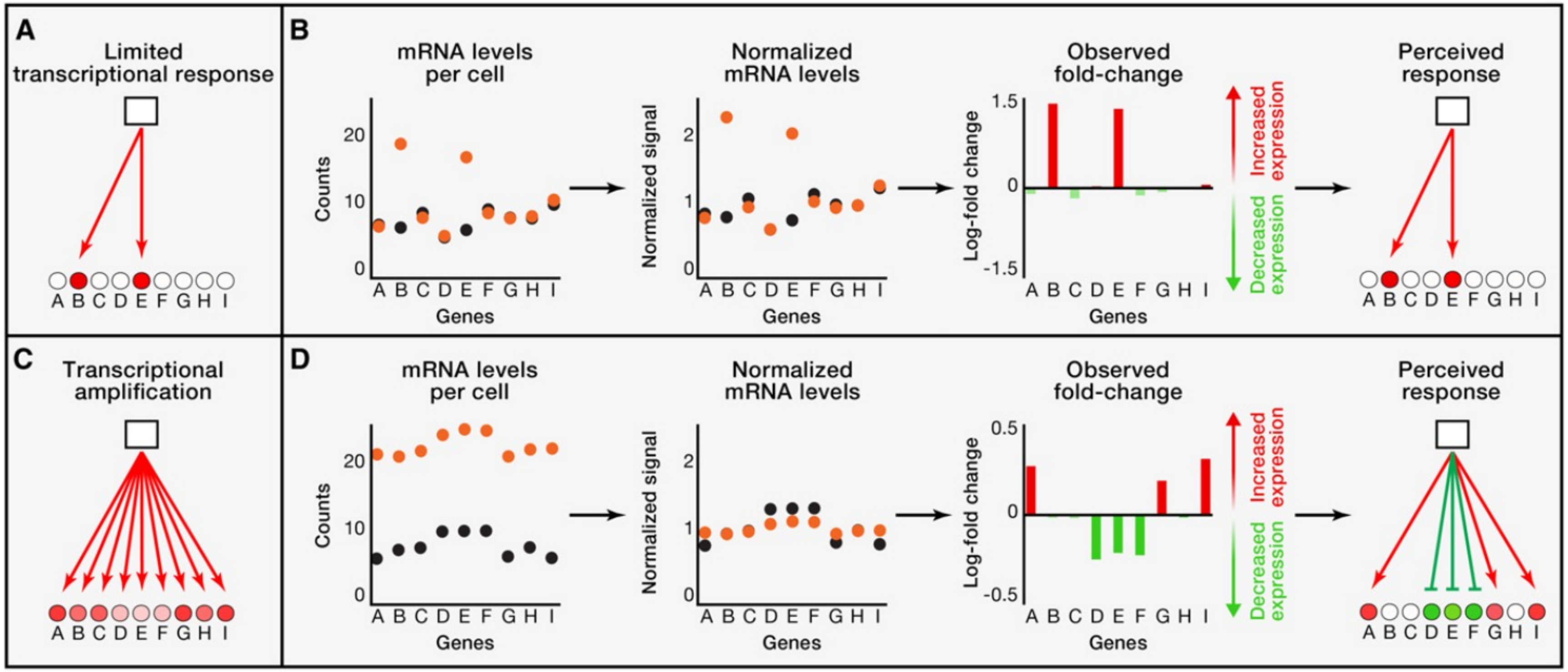
Working hypotheses:

- relative prevalence of a gene reflects activity of associated pathways; more prevalent, more active
- baseline prevalence is not known → **perturbation experiments**: treatment vs control
- informative quantity: relative abundance of the genes in a tissue

Limitations:

- genes vs protein
- alternative explanations …. cell size effects; other side effects

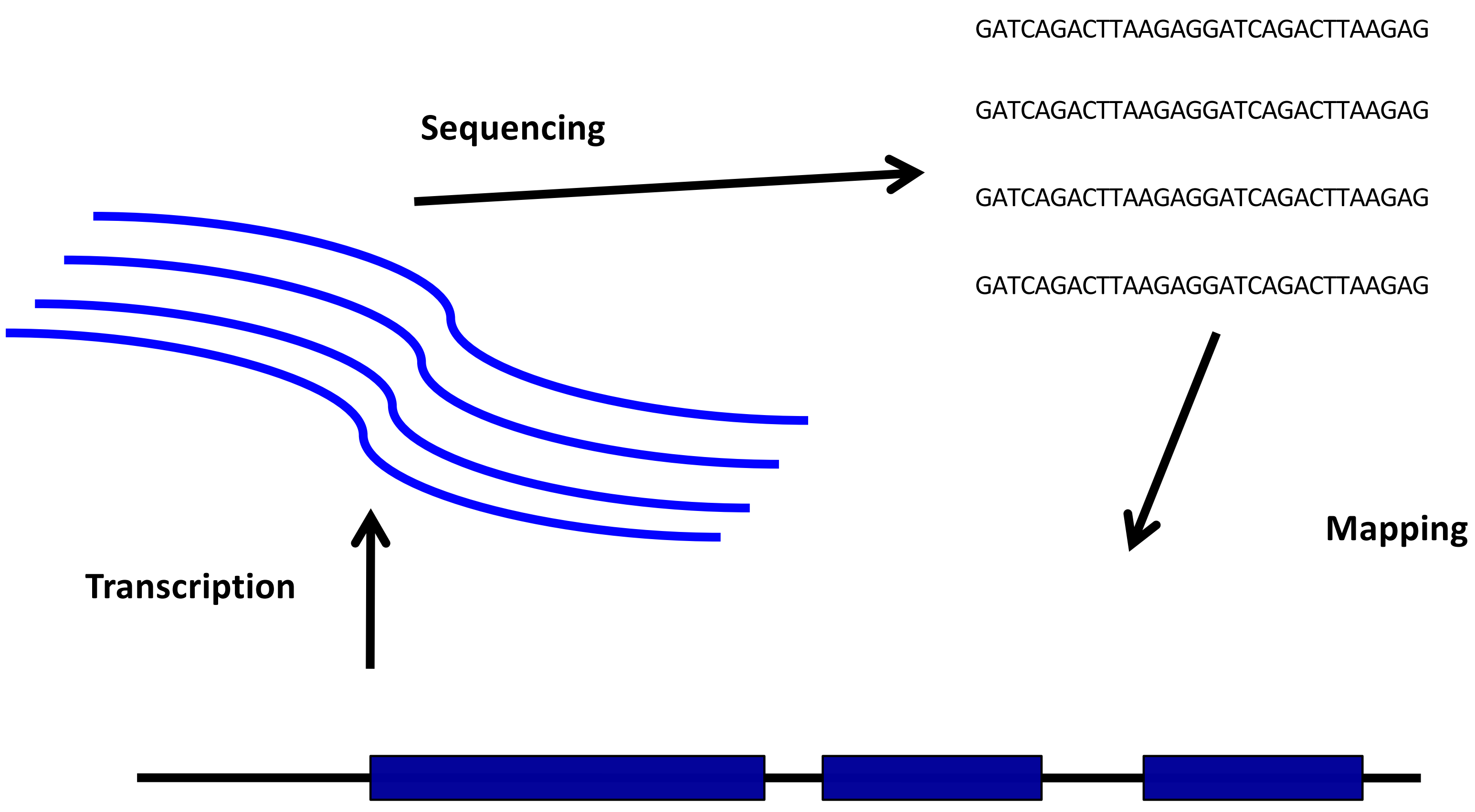
# Perturbation Experiments



- The top row shows the assumption how a baseline expression profile (black) changes as a response (red) to a perturbation
- Note: expression counts are all around ~10, in practice they range from 0 – 1 Mio

Lovén et al, 2012, https://doi.org/10.1016/j.cell.2012.10.012

# Transcript Expression with NGS



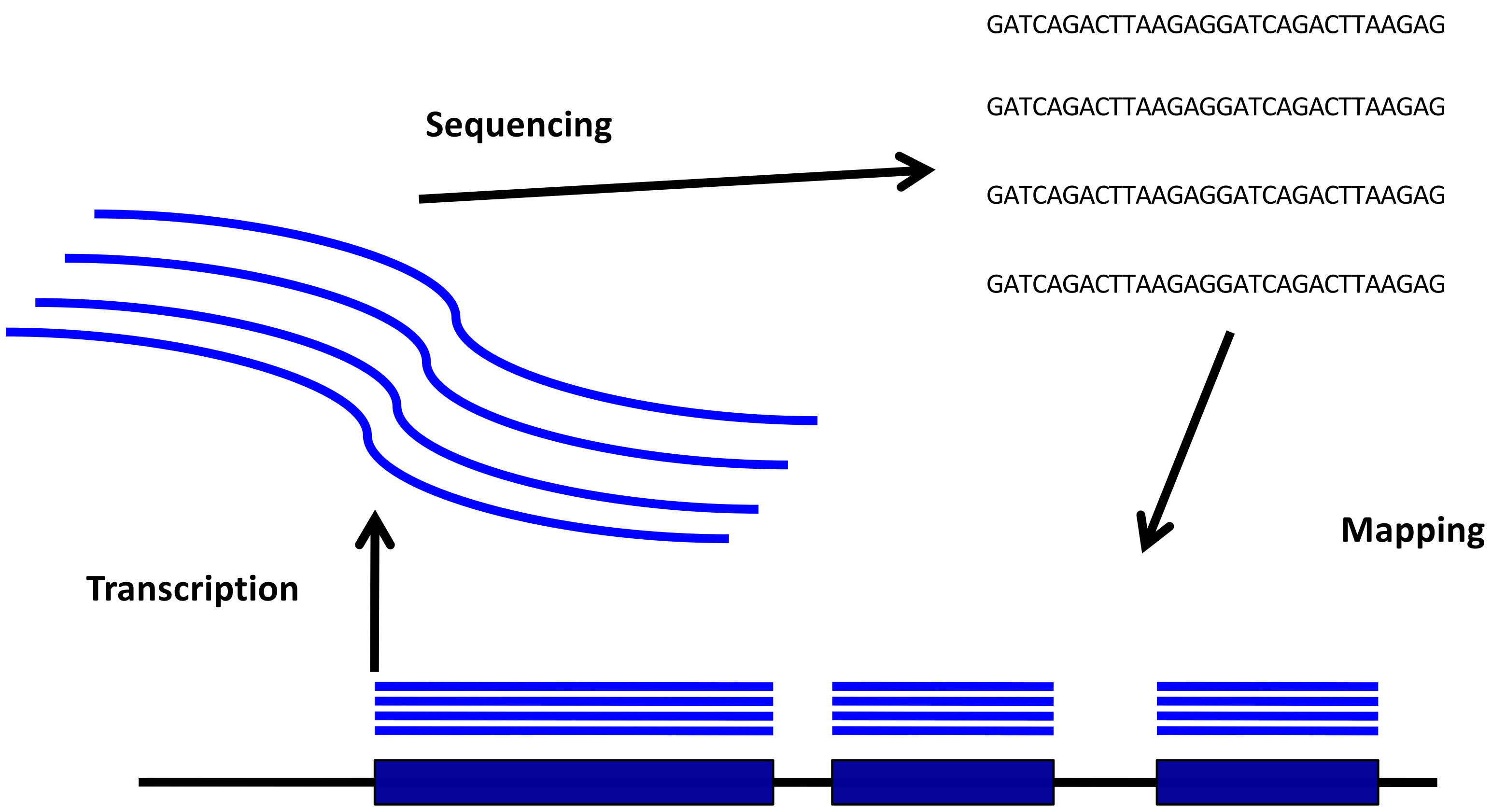
GATCAGACTTAAGAGGATCAGACTTAAGAG

GATCAGACTTAAGAGGATCAGACTTAAGAG

GATCAGACTTAAGAGGATCAGACTTAAGAG

GATCAGACTTAAGAGGATCAGACTTAAGAG

# Read mapping and counting (ideal)



GATCAGACTTAAGAGGATCAGACTTAAGAG

GATCAGACTTAAGAGGATCAGACTTAAGAG

GATCAGACTTAAGAGGATCAGACTTAAGAG

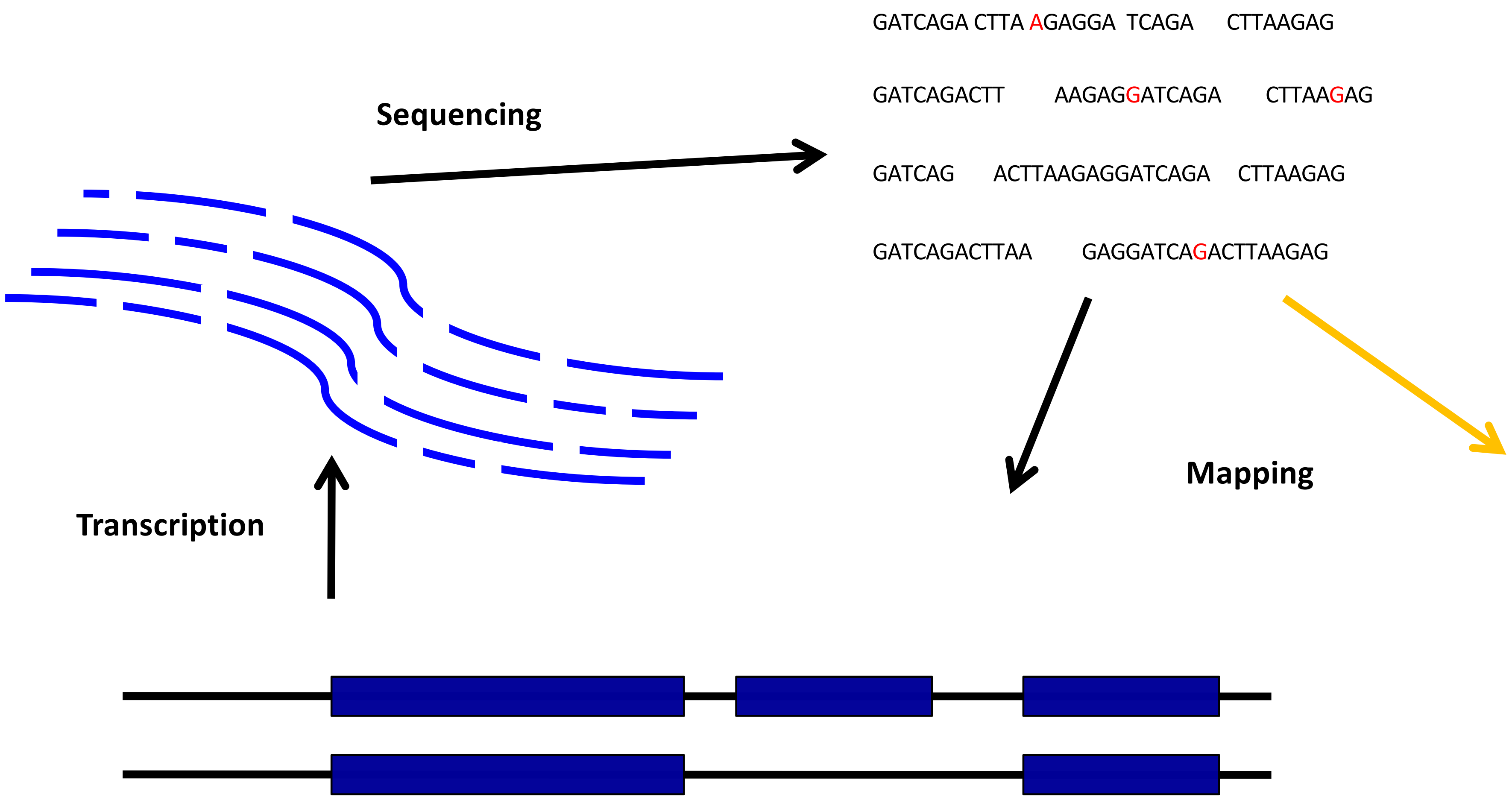
GATCAGACTTAAGAGGATCAGACTTAAGAG

Sequencing

Transcription

Mapping

# Read mapping and counting (today)



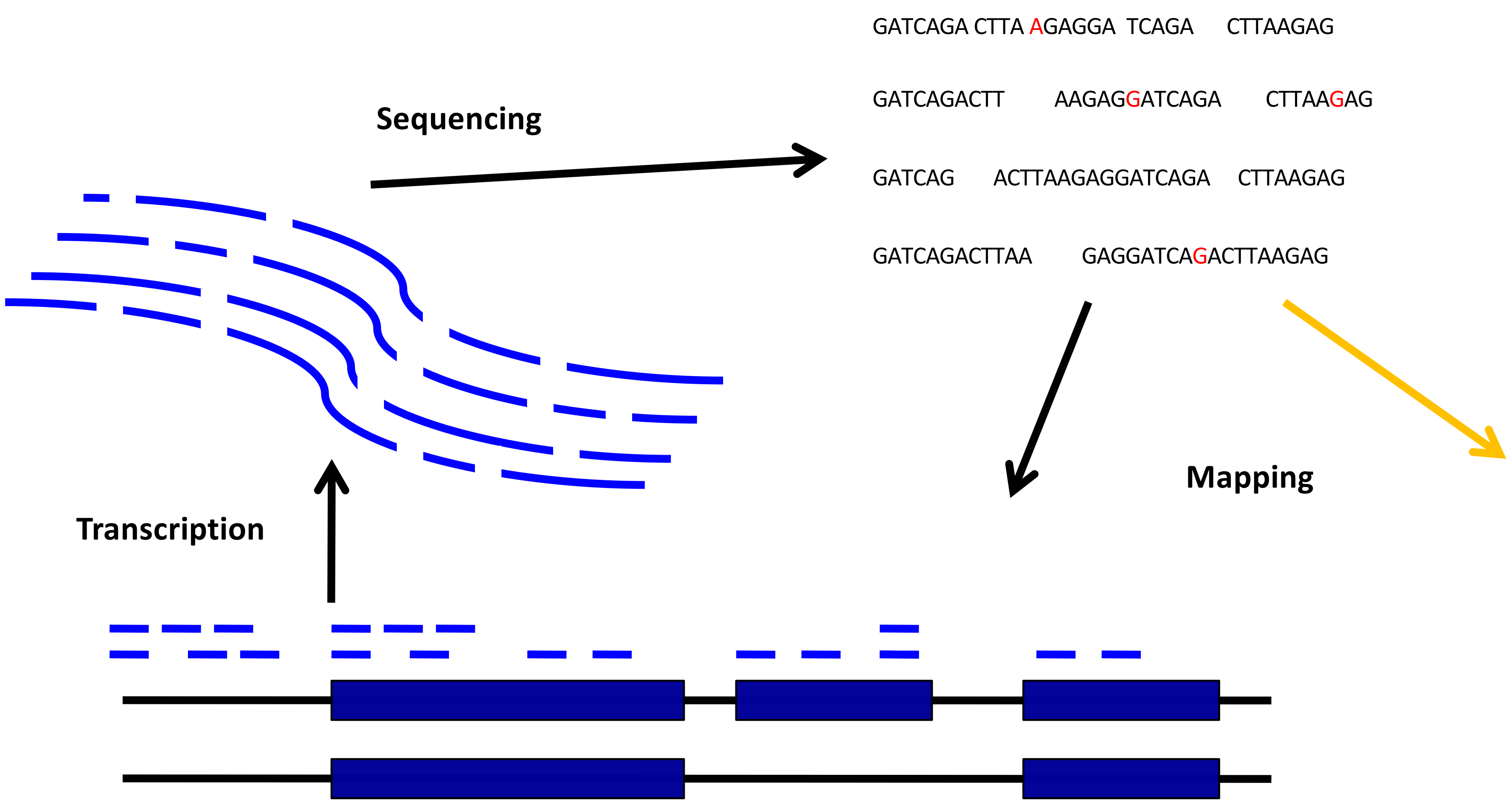
GATCAGA CTTA AGAGGA TCAGA CTTAAGAG

GATCAGACTT AAGAGGATCAGA CTTAAGAG

GATCAG ACTTAAGAGGATCAGA CTTAAGAG

GATCAGACTTAA GAGGATCAGACTTAAGAG

# Read mapping and counting (today)



GATCAGA CTTA AGAGGA TCAGA CTTAAGAG

GATCAGACTT AAGAGGATCAGA CTTAAGAG

GATCAG ACTTAAGAGGATCAGA CTTAAGAG

GATCAGACTTAA GAGGATCAGACTTAAGAG

## 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 that indicate the abundance relative to all other genes/isoforms**

## Gene-level Read Counts

- rather straightforward to compute:
  - **number of reads** that uniquely map to a gene locus
    → biased by length, discards information in multi-mappers
  - **number of 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

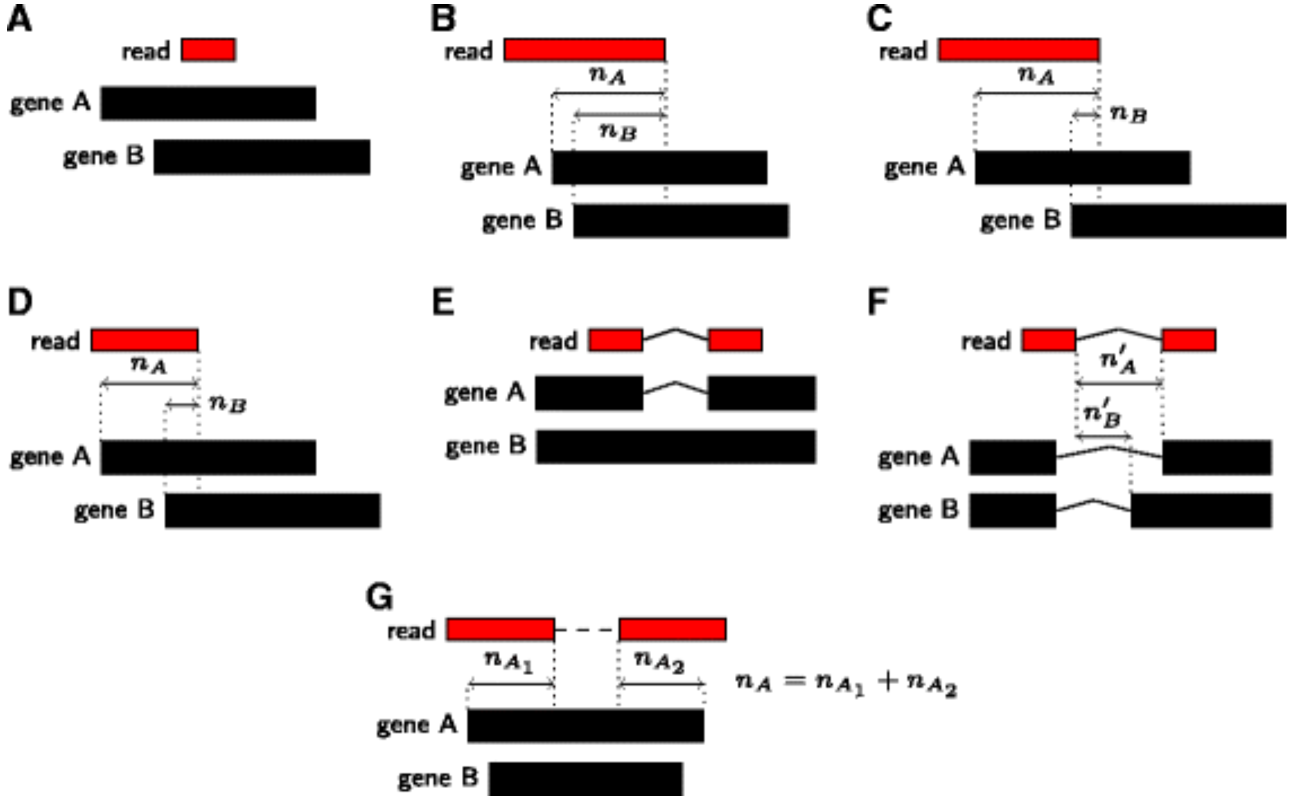
## *Rsubread::featureCounts* – assigning reads to genes

- versatile function to count reads towards exons, transcripts, 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

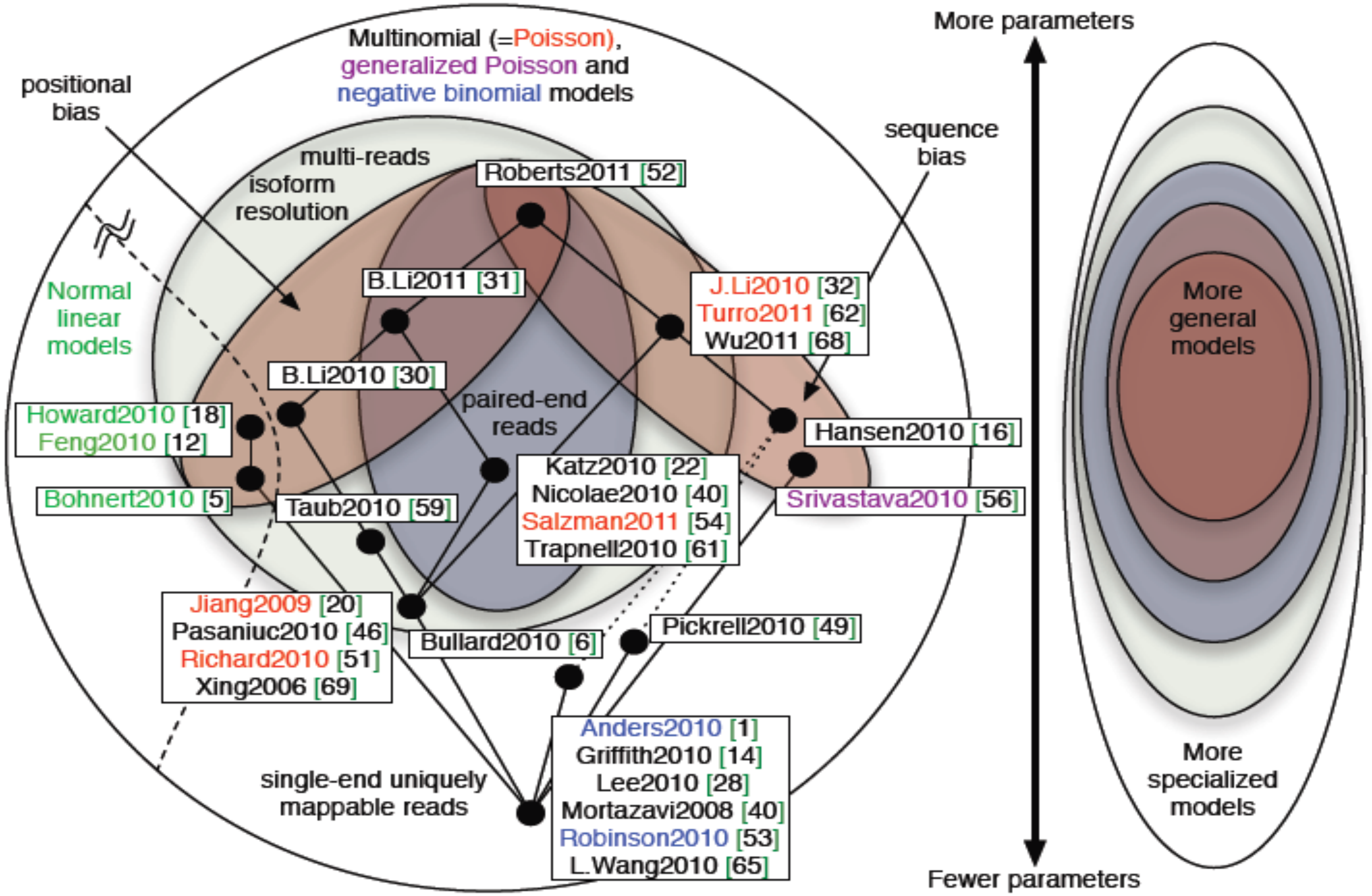
# Model-free Counting of Overlapping reads – Count Modes

mmquant: resolve multi-mapping reads based on heuristics



https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1816-4

# Generative Models for RNA-seq quantification



Pachter 2011, https://doi.org/10.48550/arXiv.1104.3889

## RNA-seq model

Likelihood that a read/fragment from transcript *t* is generated:

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

with:

$\rho_t$ expression level / abundance / fraction

$l_t$ transcript length

$\mathrm{Z} = \sum_t \rho_t l_t$ normalization factor

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

## RNA-seq model

Estimate of the probability that a read from a specific transcript is generated:

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

Abundance estimates:

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

## Maximum Likelihood Estimation

- The estimated abundances represent unique MLE estimates

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

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

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

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

## 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 should 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:

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

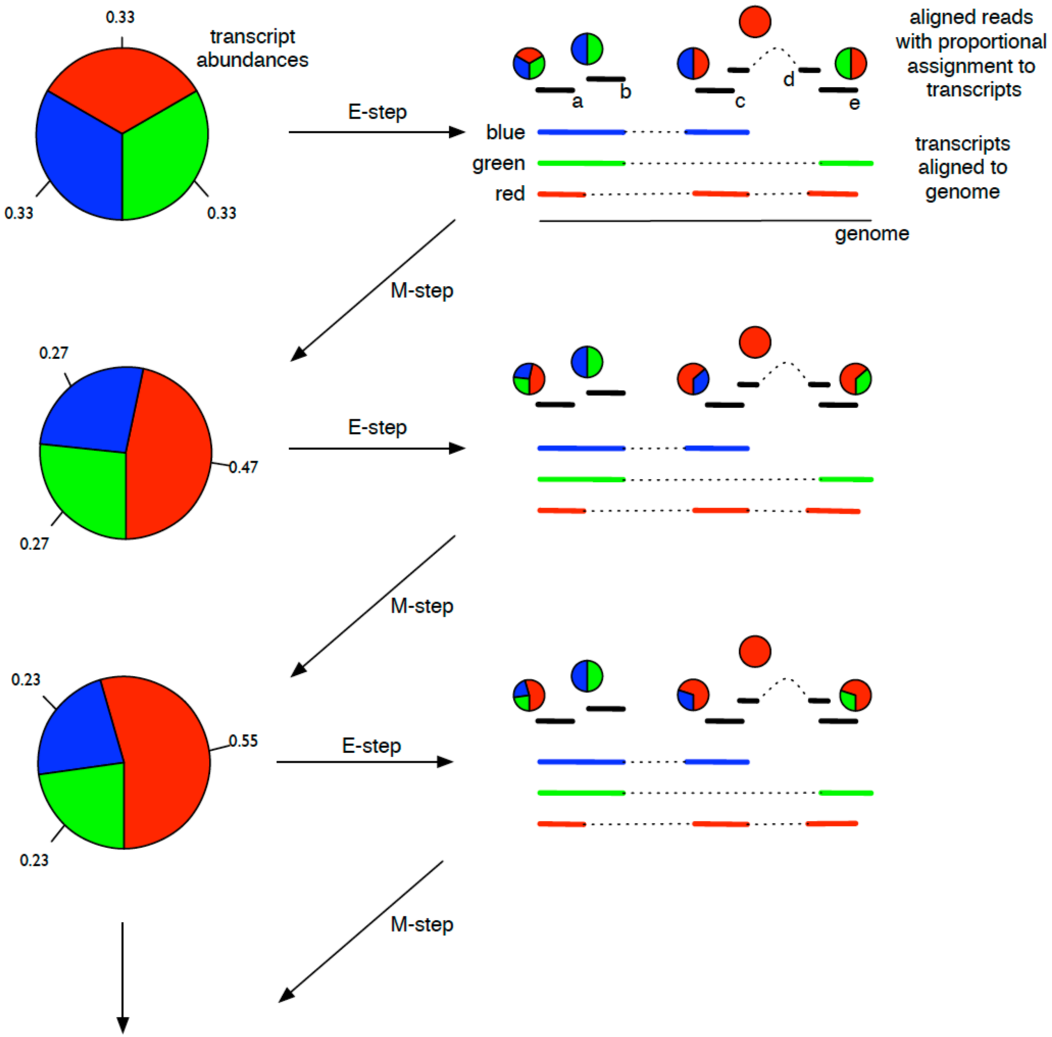
- but now abundances must be estimated iteratively

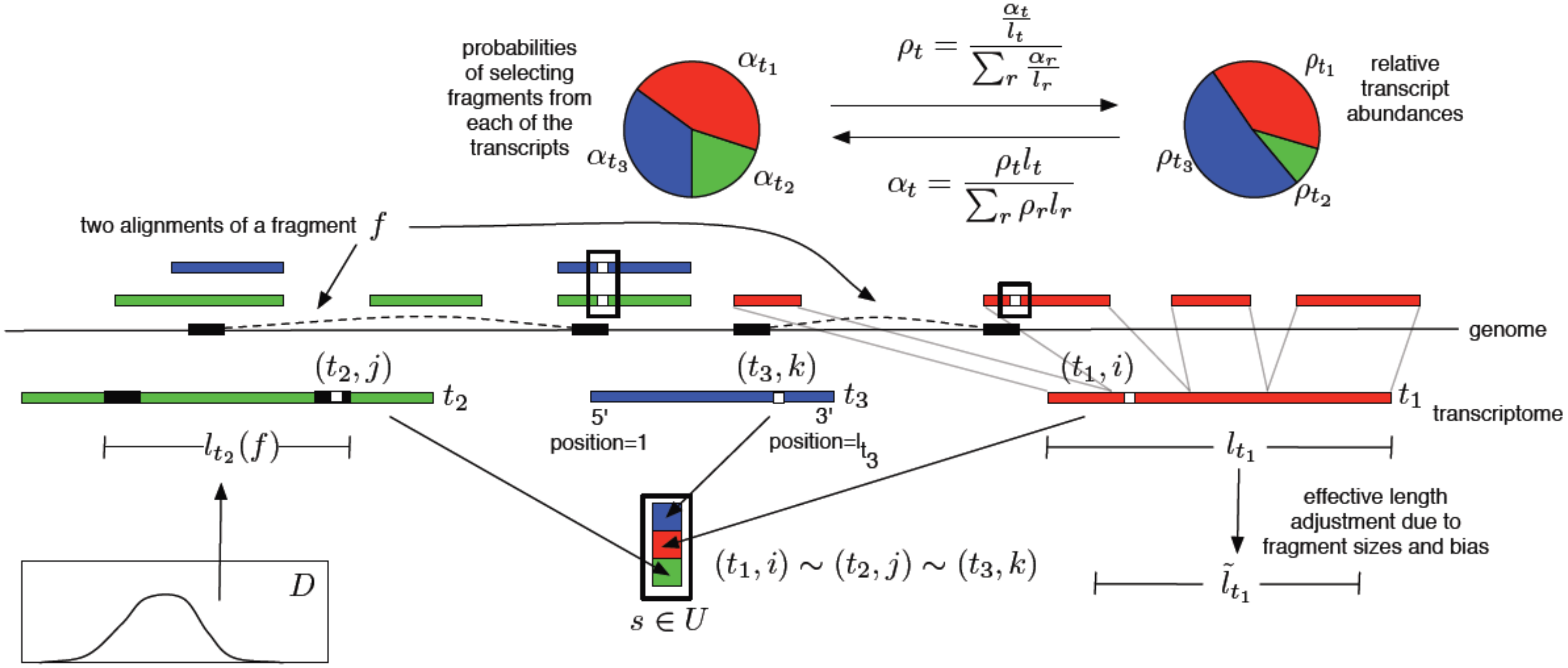
### Iterative Estimation

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

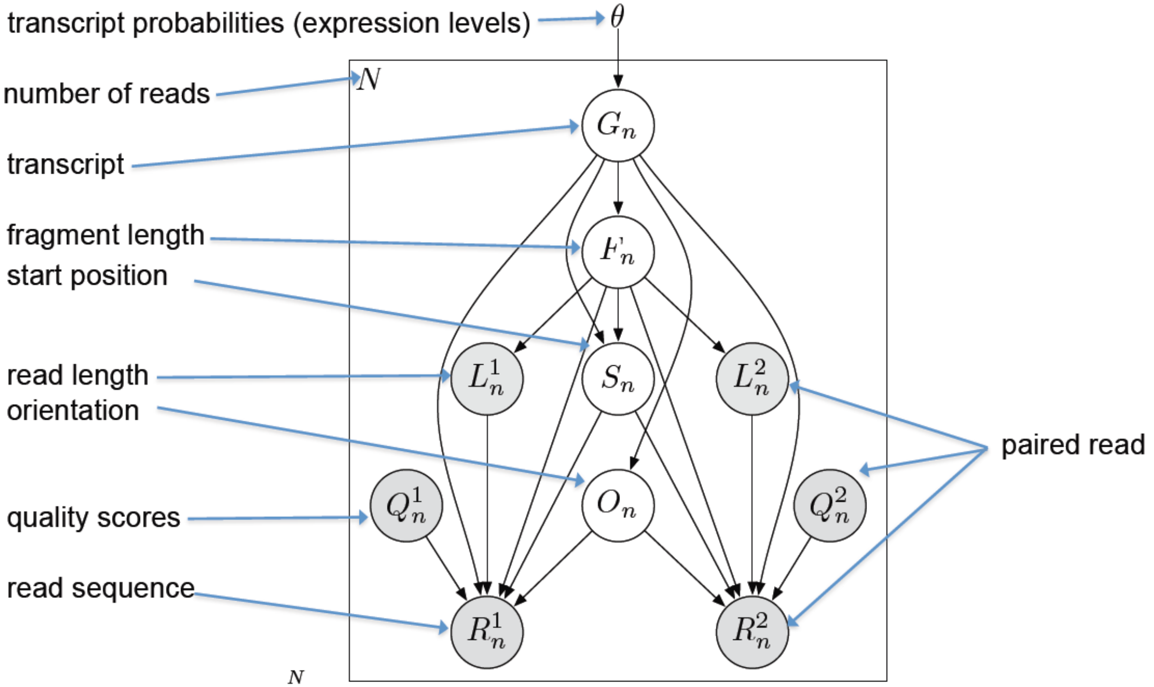
## General Formulation of Abundance Estimation

A full model for the abundance estimation would ideally consider:

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

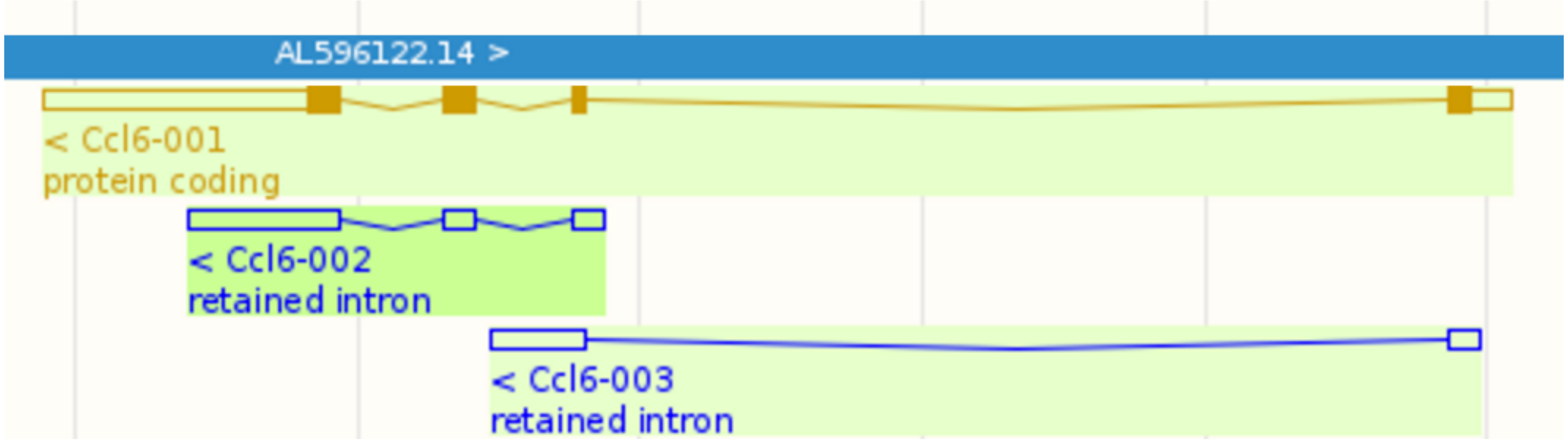
# RSEM: Model specification

Dewey: RSEM



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

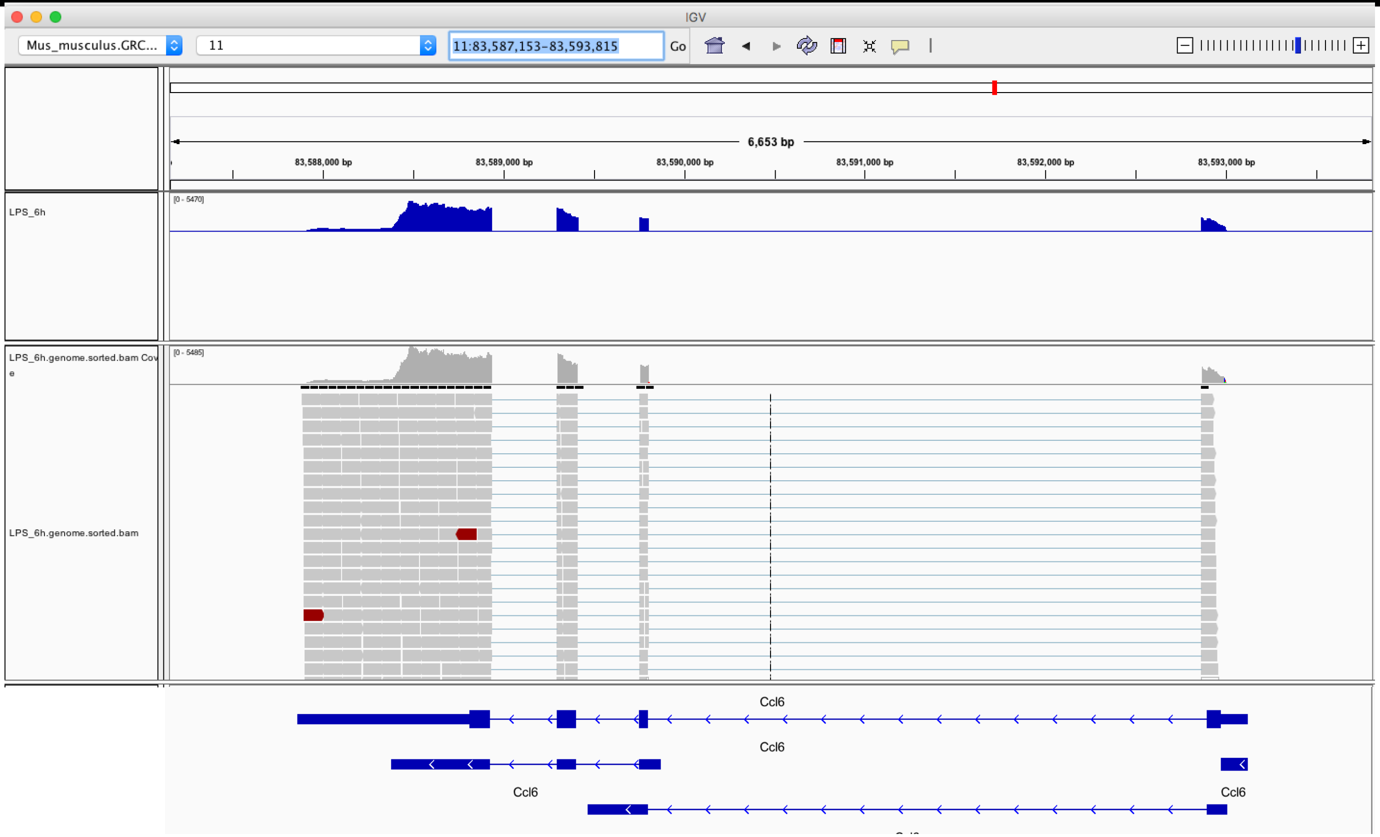
## Example: RSEM



Ccl6 gene locus with 3 isoforms

follows the example:

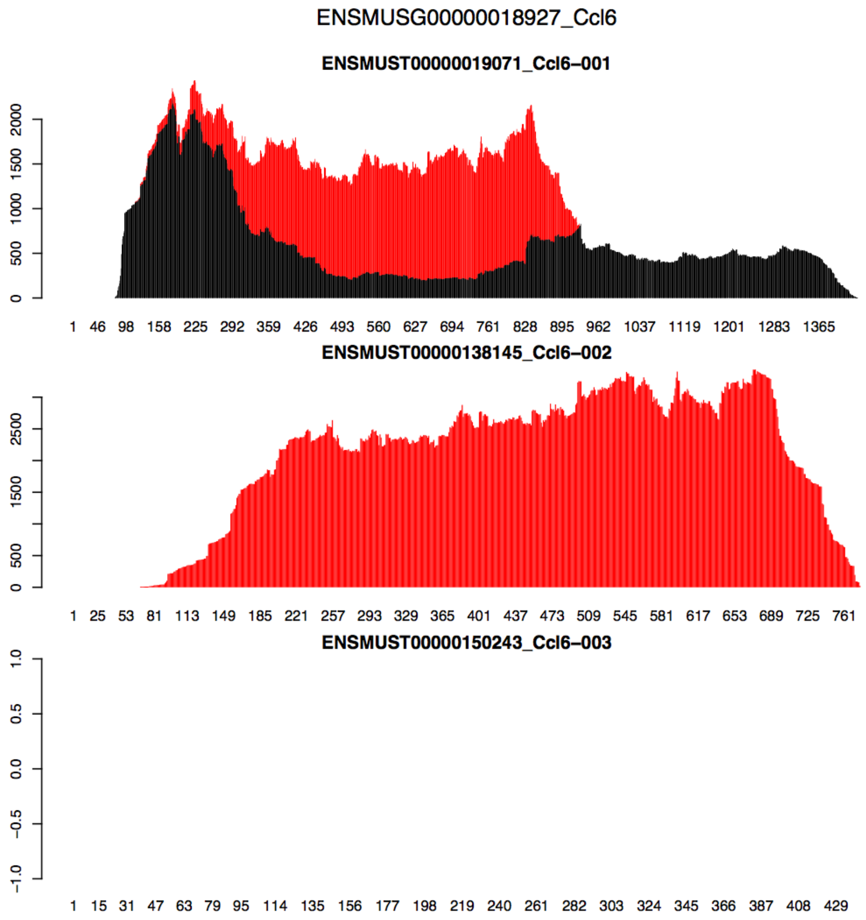
https://github.com/bli25broad/RSEM_tutorial



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 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).

- **Review:**
  Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889* (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).

### 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

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

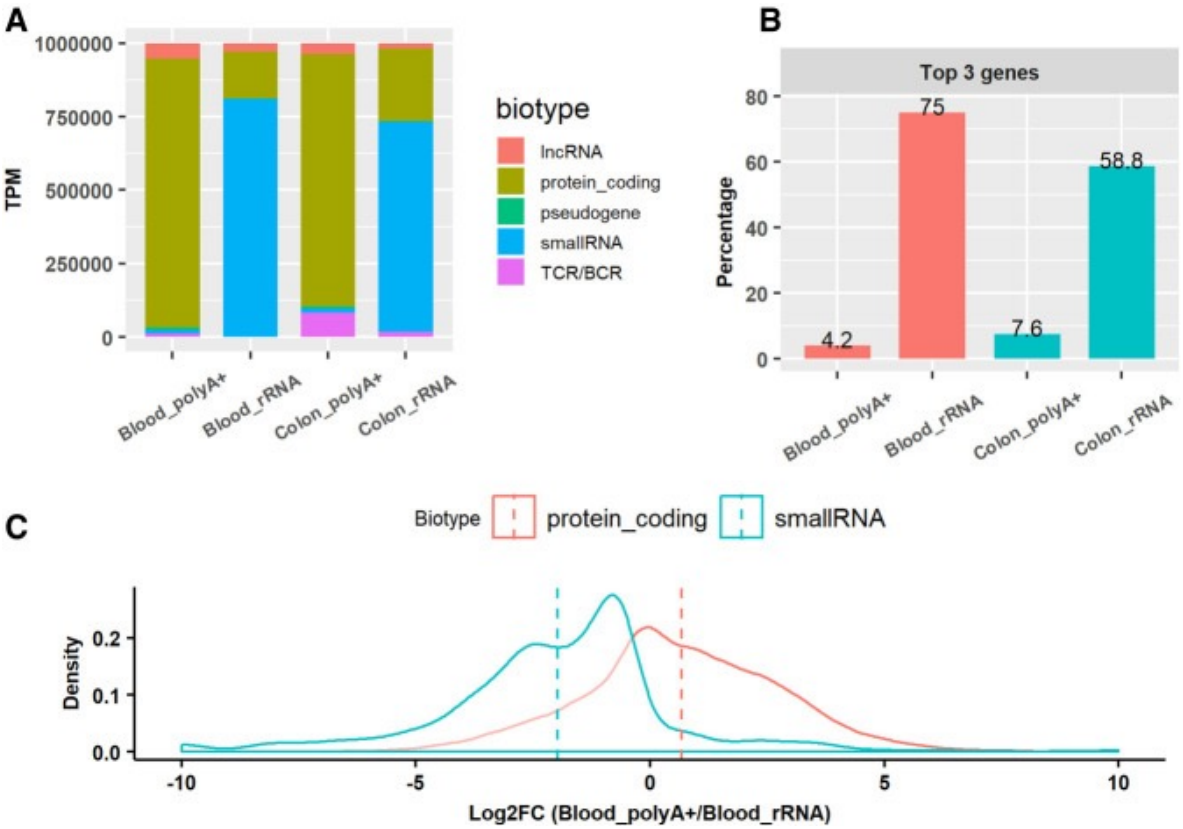
## Shortcomings of RPKM and TPM

- 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

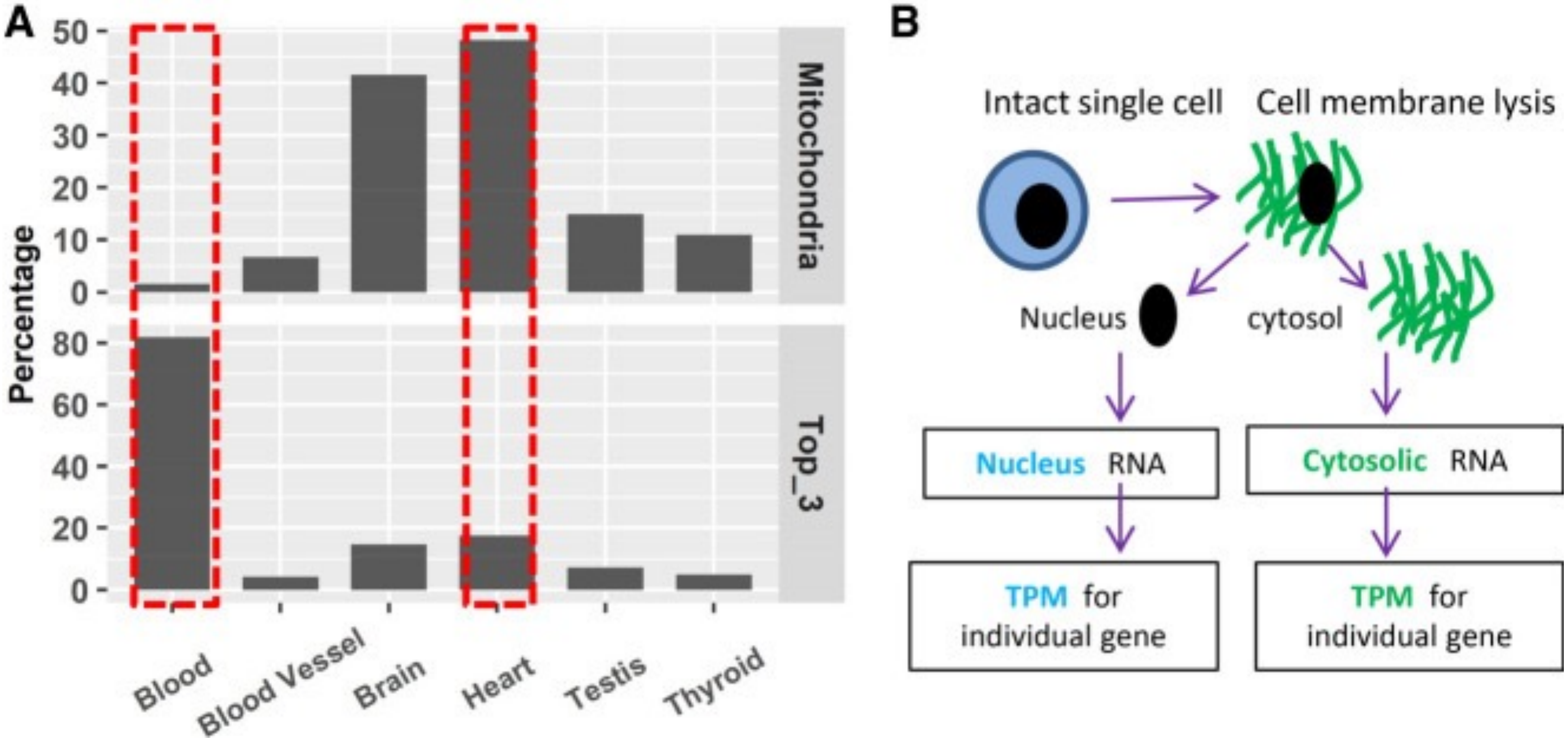


issues:

- 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)

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

## Comparing samples across tissues

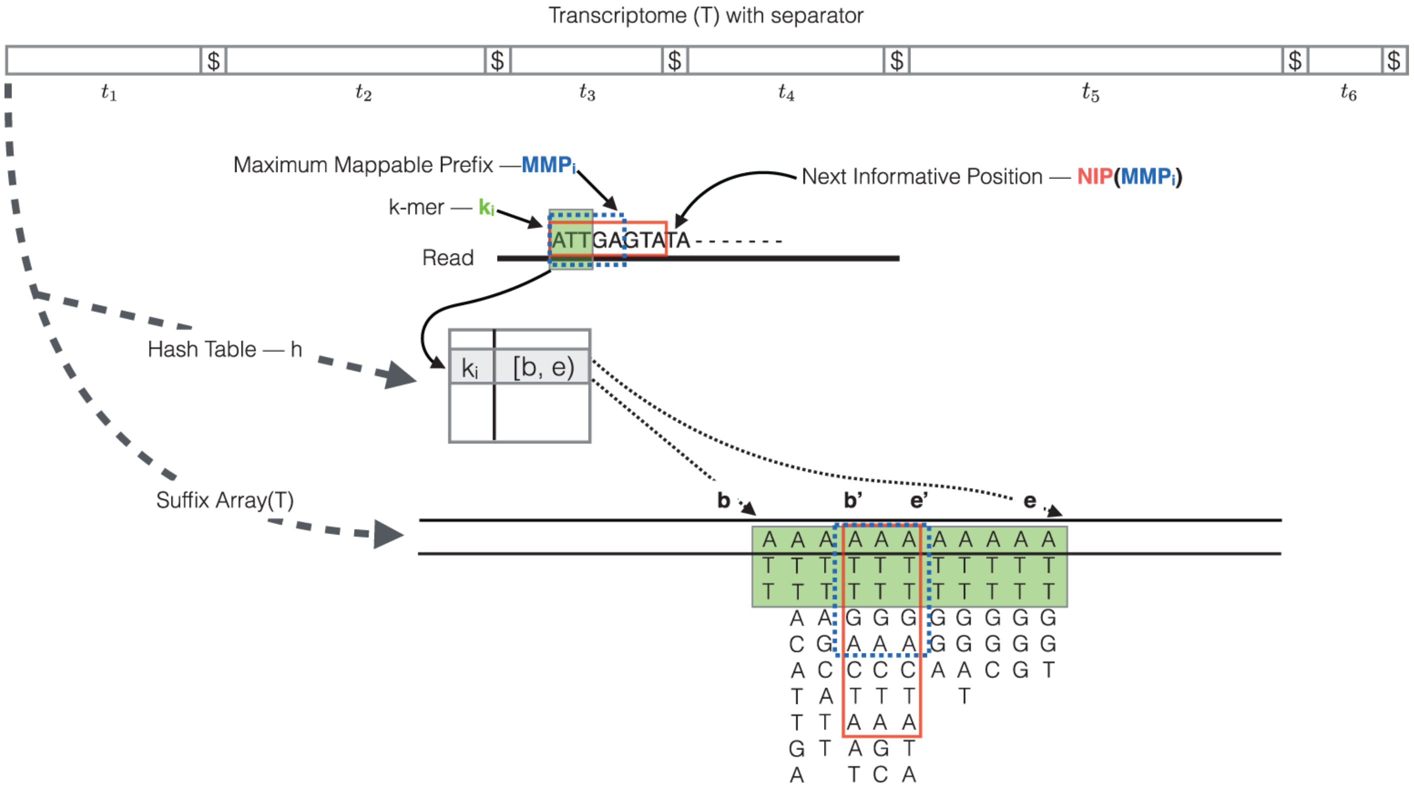


Different tissues may have different populations of genes expressed

# Fast approaches to get the Read-Transcript Compatibility Matrix

- Salmon: quasi-mapping
- kallisto: pseudo-alignments

# Quasi-mapping



Default k-mer size: 31

https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/04_quasi_alignment_salmon.html

# 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

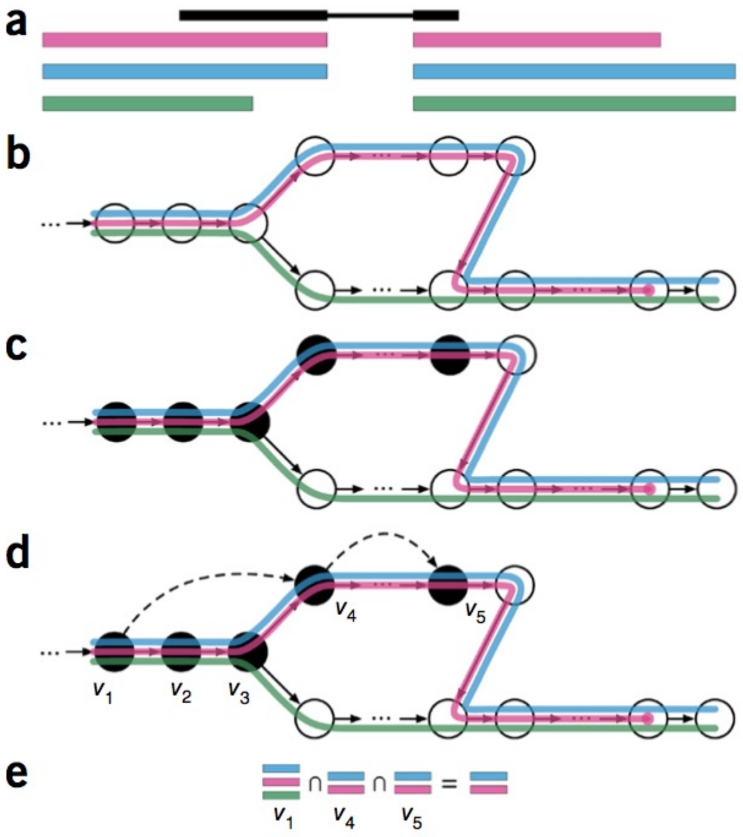
- Result: Read-Transcript compatibility matrix

- 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

# kallisto: 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

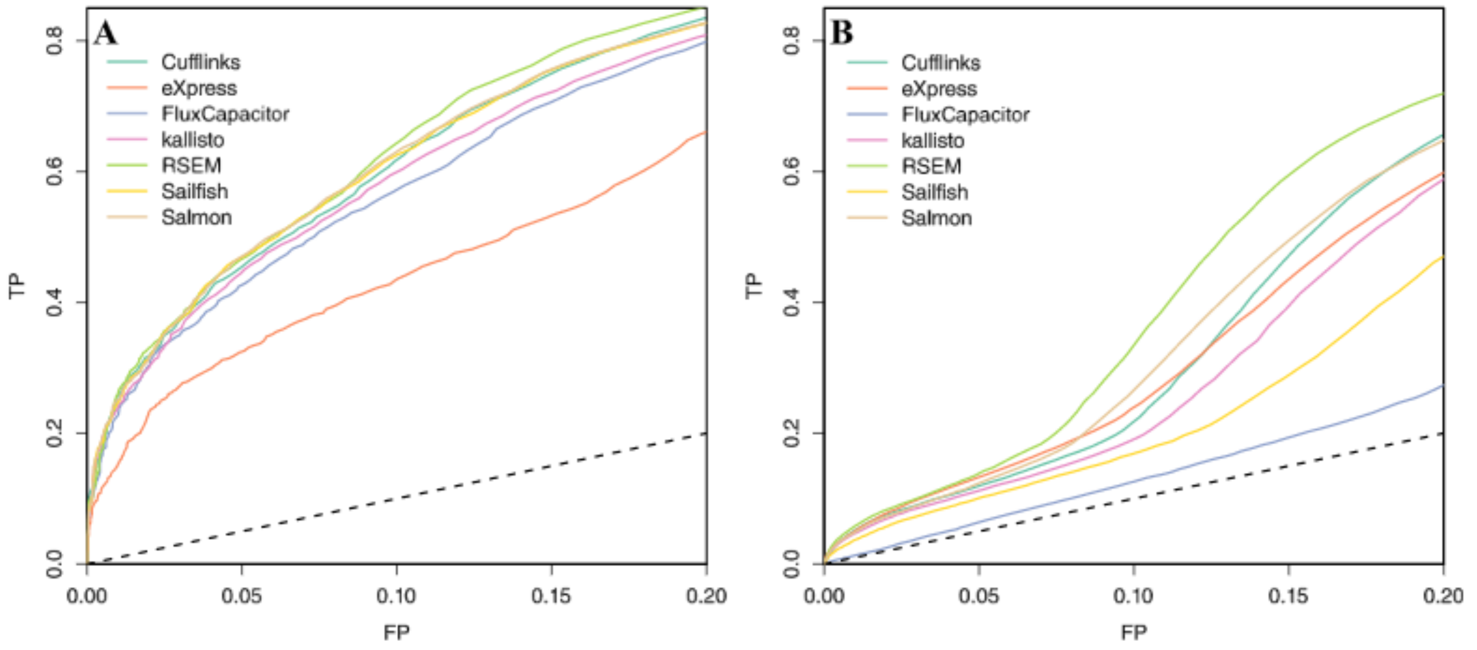
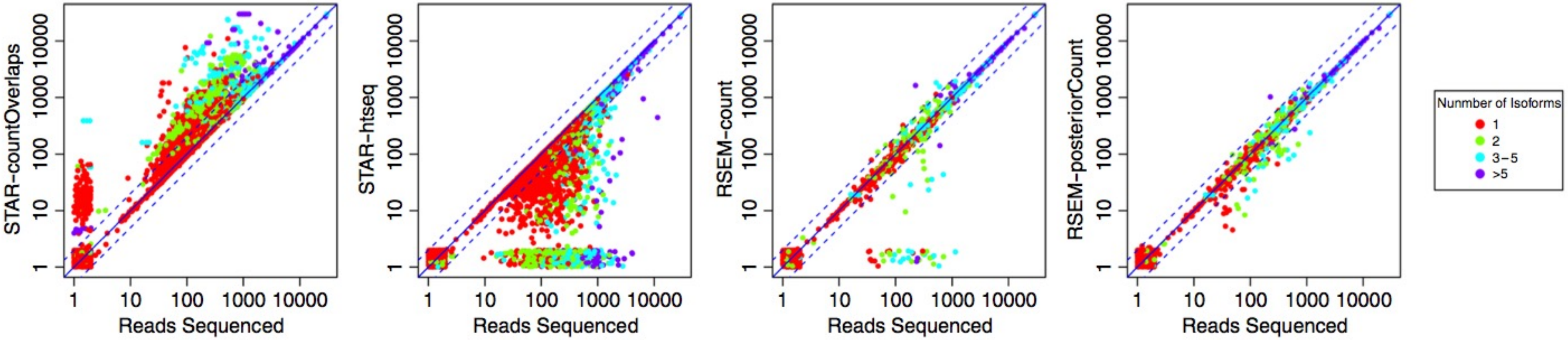


**Fig. 6** ROC curves indicating performance of quantification methods based on differential expression analysis of **a** an experimental dataset and **b** a simulation dataset. Seven quantification methods are shown. *FP* false positive, *TP* true positive

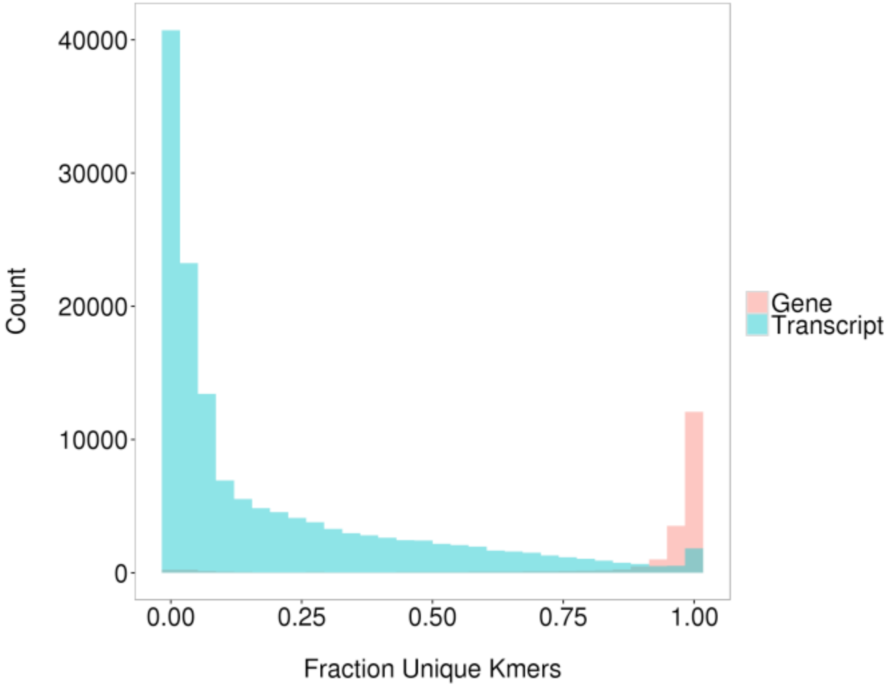
Teng et al. Genome Biology, 2016

## Read Counting Accuracy



Rehrauer et al.BMC Bioinformatics, 2014

# Uniqueness: Isoform-level vs gene-level

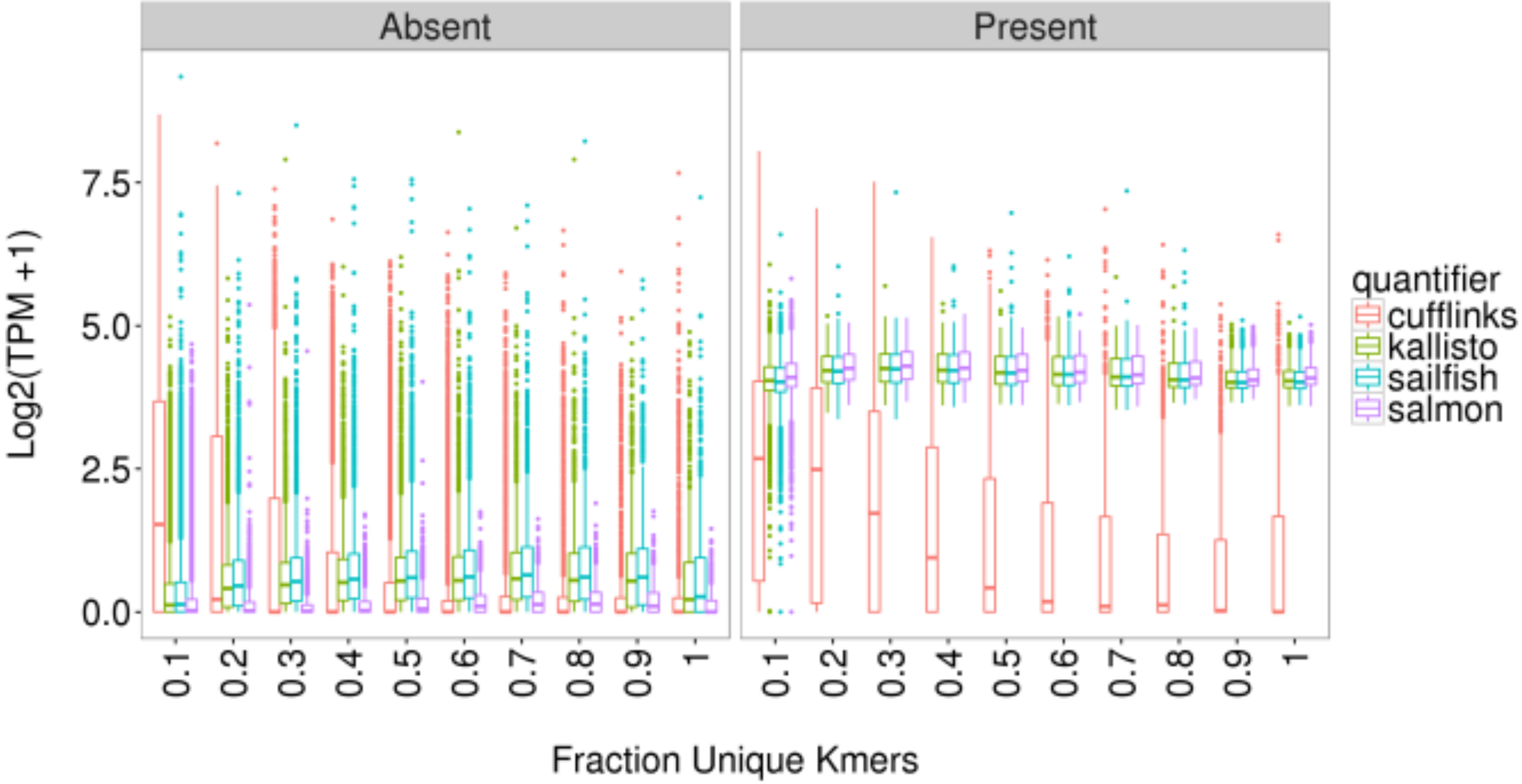


- Fraction of unique k-mer sequences for genes and transcripts
- Ambiguity is mainly between alternative transcripts from the same gene locus

https://cgatoxford.wordpress.com/tag/kallisto/

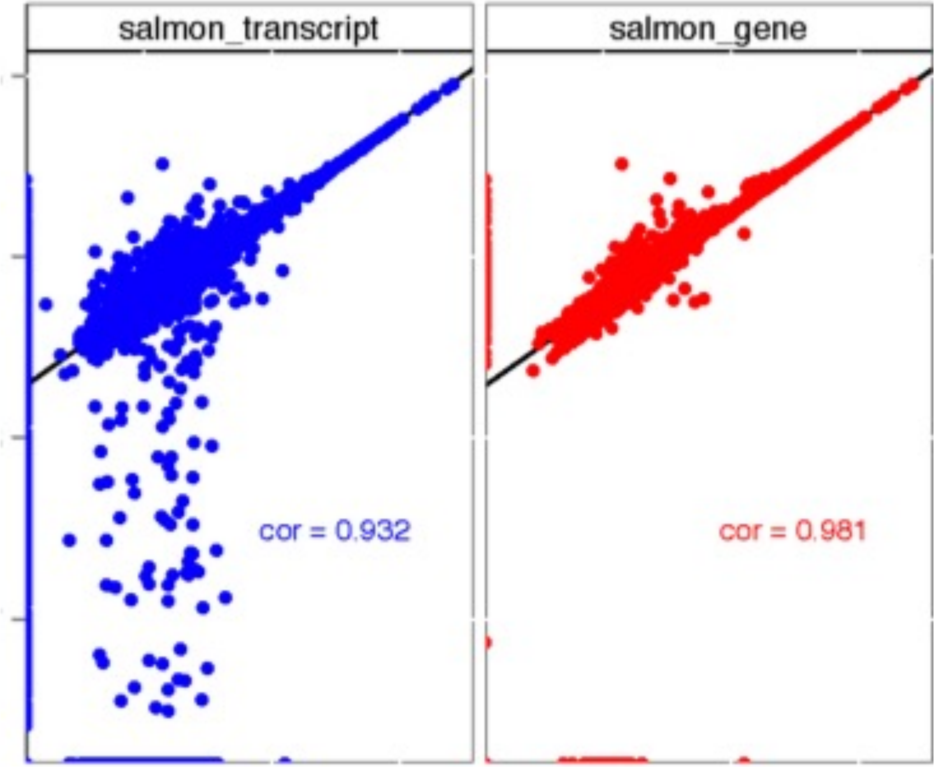
# Misassignments to unexpressed transcripts

- Simulated data show that misassignments do happen



https://cgatoxford.wordpress.com/tag/kallisto/

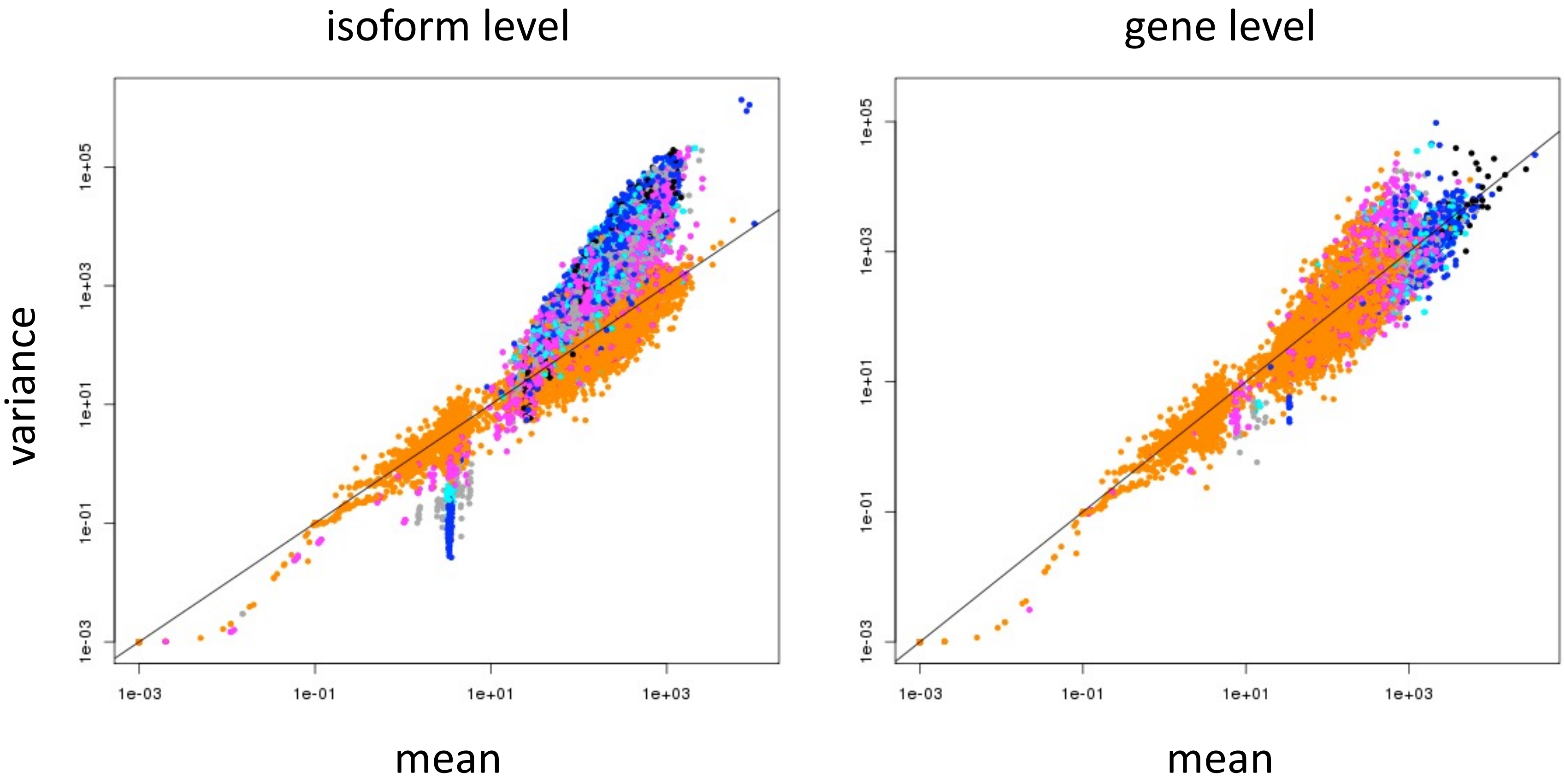
# Accuracy: Isoform-level vs gene-level



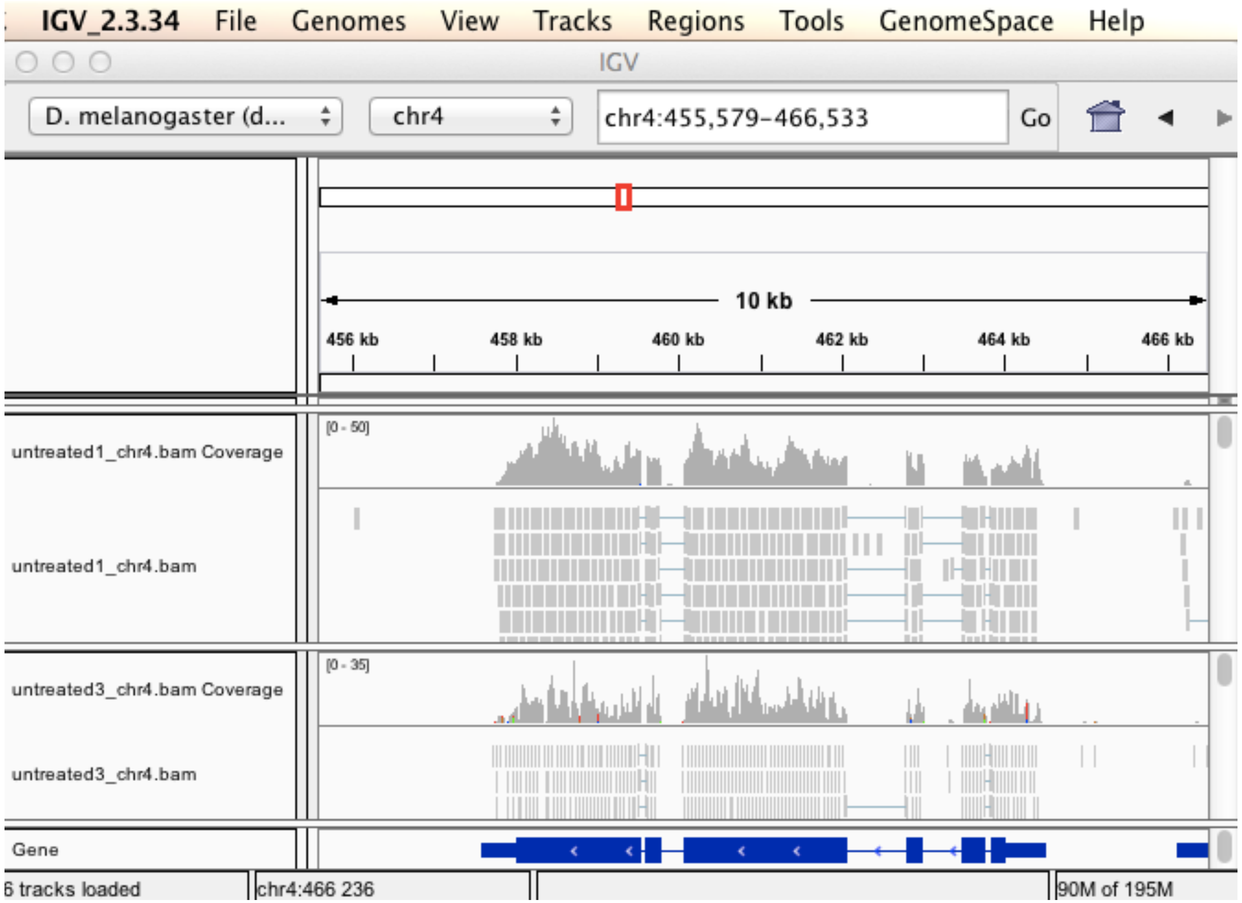
remove

Soneson et al. F1000, 2015

# Isoform level has higher variability



# Positional bias of read distributions

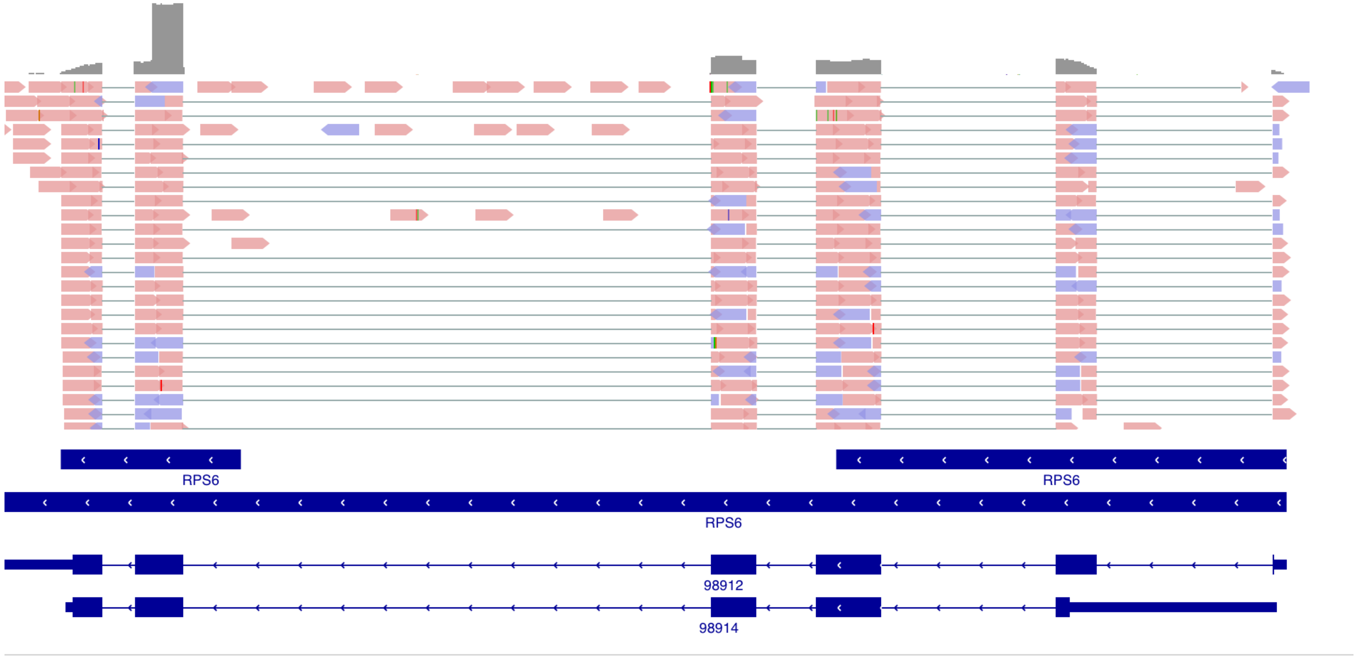

IGV_2.3.34 File Genomes View Tracks Regions Tools GenomeSpace Help
IGV
Go
10 kb
456 kb
458 kb
460 kb
462 kb
464 kb
466 kb
[0 - 50]
untreated1_chr4.bam Coverage
untreated1_chr4.bam
[0 - 35]
untreated3_chr4.bam Coverage
untreated3_chr4.bam
Gene

D. melanogaster (d... | chr4 | chr4:455,579-466,533

6 tracks loaded | chr4:466 236 | 90M of 195M

## 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)

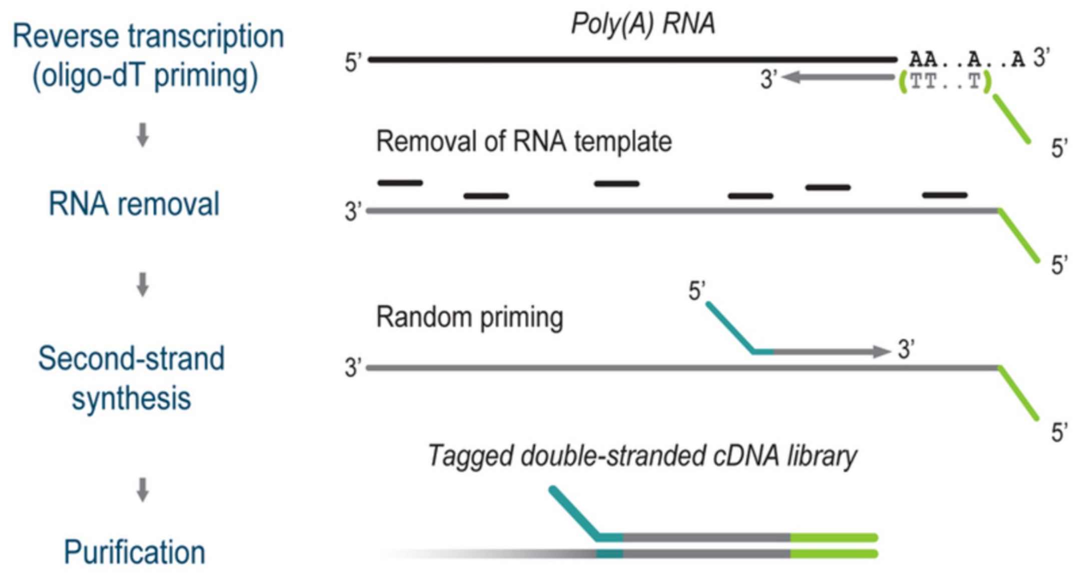
# 3’-Tagging

Reads are only generated near the 3’-end

Isoforms can not be resolved

Allows counting of the reads at the 3’-end

→ Assignment is to genes



# Summary

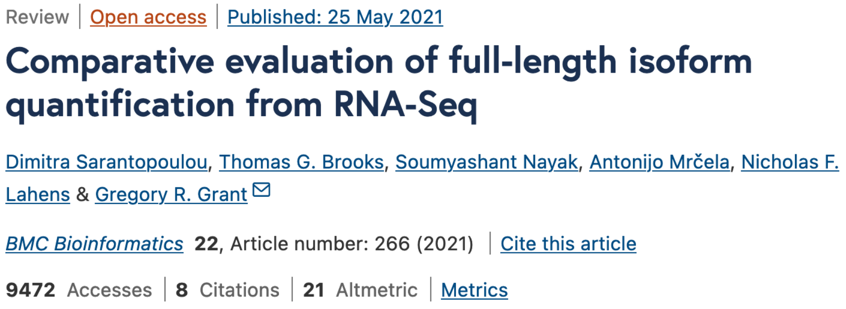
- Alignment + expression estimation using generative models give good results, let you inspect the aligned reads, and can be used to discover new genes and new isoforms
- Pseudo-alignment is reliable and fast but needs as input the accurate and complete set of transcripts
- 3'-end tag sequencing provides only gene-level estimates without isoform resolution
- Full-length transcript sequencing detects isoforms accurately, sequencing depth is typically lower than for short-read sequencing

# References

- Compares kallisto, salmon, featureCounts, … https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-021-04198-1

- from the salmon people, performance evaluation and tuning options (especially genomic “decoy”) https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02151-8

- from the salmon people, confirm that decoys help https://www.biorxiv.org/content/10.1101/2021.01.17.426996v1



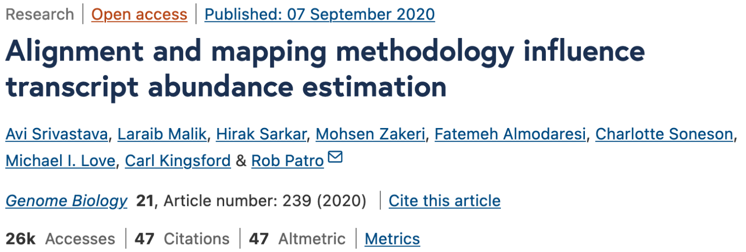
Review | Open access | Published: 25 May 2021

**Comparative evaluation of full-length isoform quantification from RNA-Seq**

Dimitra Sarantopoulou, Thomas G. Brooks, Soumyashant Nayak, Antonijo Mrčela, Nicholas F. Lahens & Gregory R. Grant

*BMC Bioinformatics* **22**, Article number: 266 (2021) | Cite this article

**9472** Accesses | **8** Citations | **21** Altmetric | Metrics



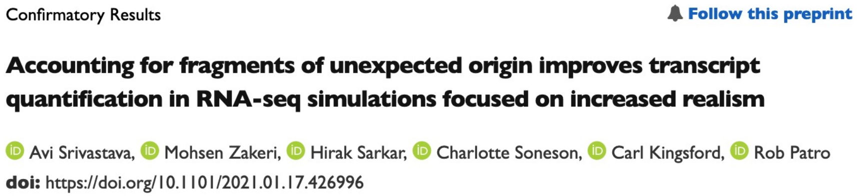
Research | Open access | Published: 07 September 2020

**Alignment and mapping methodology influence transcript abundance estimation**

Avi Srivastava, Laraib Malik, Hirak Sarkar, Mohsen Zakeri, Fatemeh Almodaresi, Charlotte Soneson, Michael I. Love, Carl Kingsford & Rob Patro

*Genome Biology* **21**, Article number: 239 (2020) | Cite this article

**26k** Accesses | **47** Citations | **47** Altmetric | Metrics



Confirmatory Results

**Follow this preprint**

**Accounting for fragments of unexpected origin improves transcript quantification in RNA-seq simulations focused on increased realism**

Avi Srivastava, Mohsen Zakeri, Hirak Sarkar, Charlotte Soneson, Carl Kingsford, Rob Patro

**doi:** https://doi.org/10.1101/2021.01.17.426996