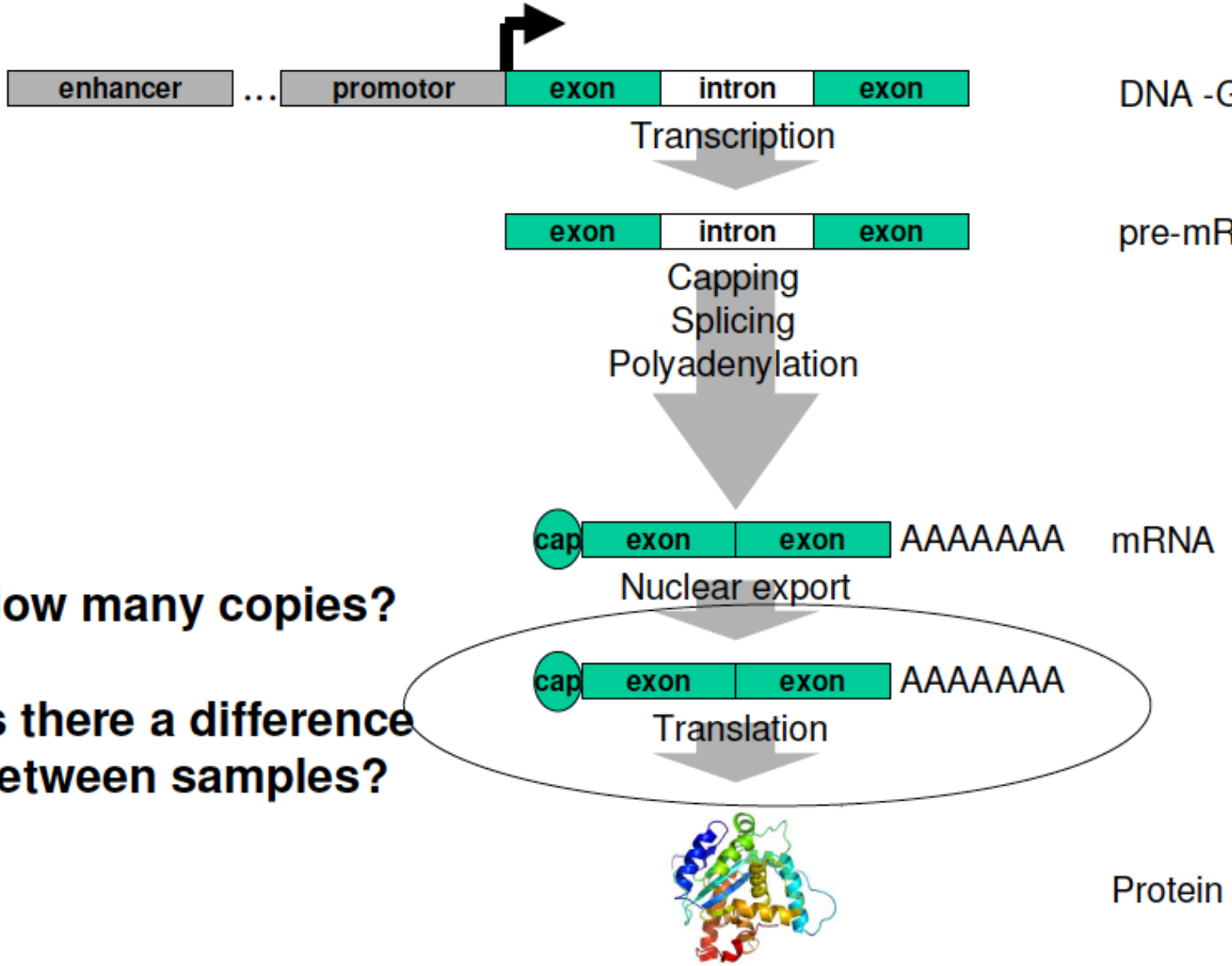
Genome-scale technologies 2/ Algorithmic and statistical aspects of DNA sequencing

RNA-Seq II: alternative splicing

Ewa Szczurek
University of Warsaw, MIMUW

szczurek@mimuw.edu.pl

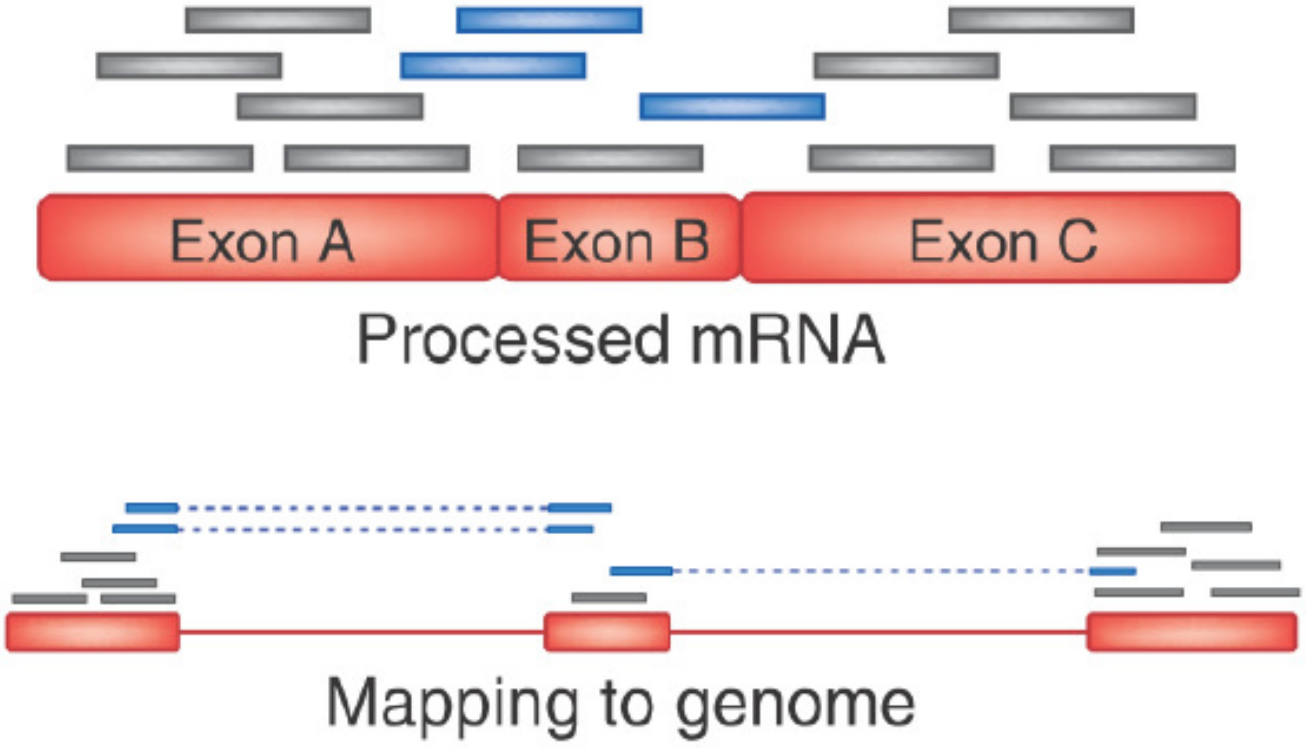
RNA-Seq



How many copies?

Is there a difference between samples?

mRNAs are collections of exons



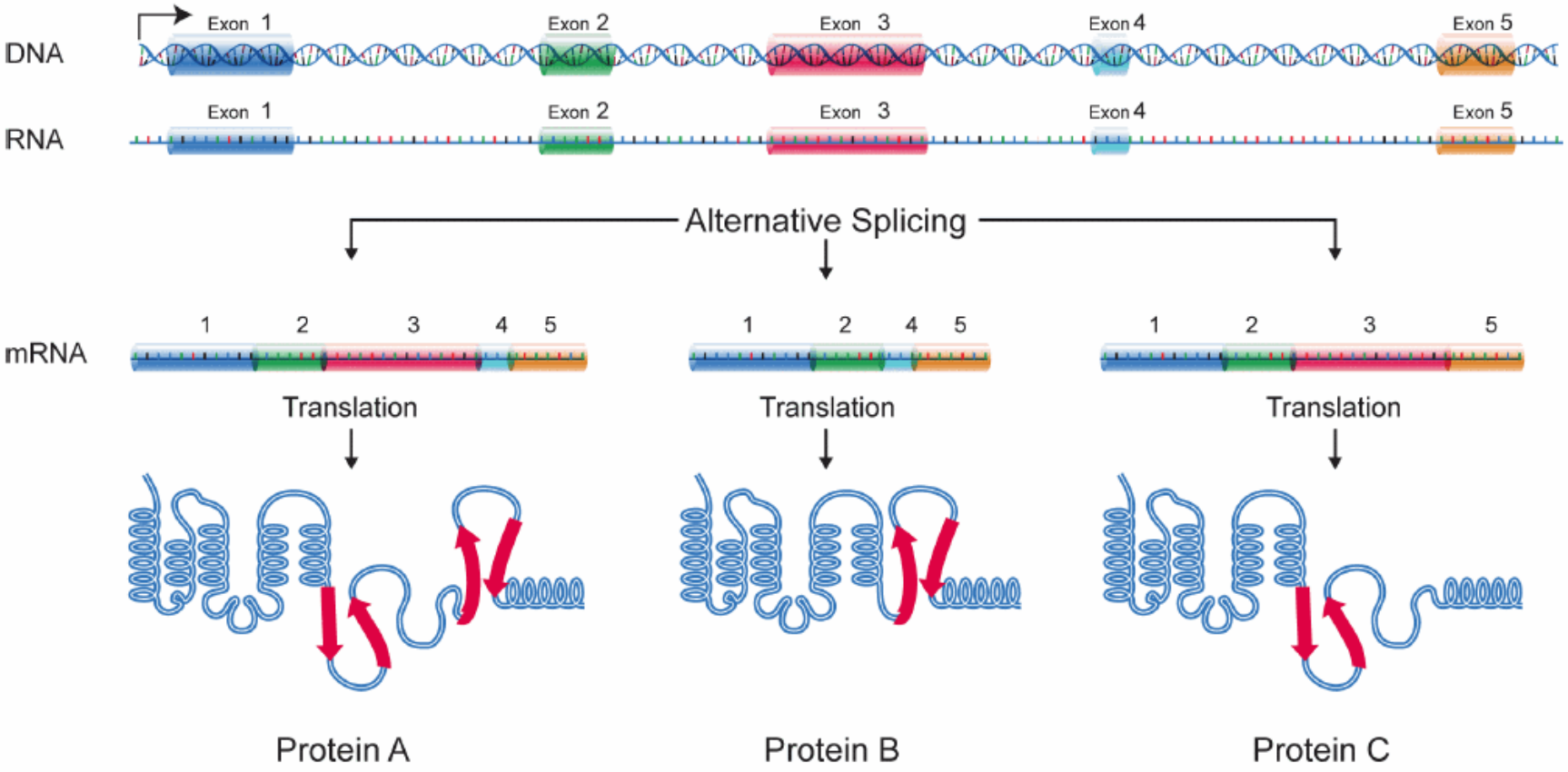
Trapnell (2009), Nature Biotech. 27, 455

Splicing

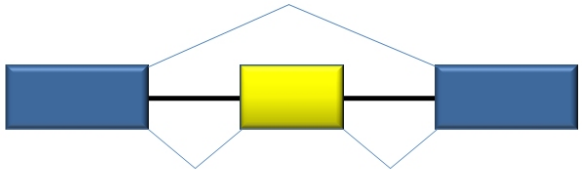
- Typical human gene:
  - 9-10 exons
  - spread out over ~2,227 bp.
  - the exons relatively short (~235 bp)

- The process by which the introns are removed is called **RNA splicing**.
- **Splice sites** are the sequences immediately surrounding the exon-intron boundaries.
- **GT-AG rule:** GT marks the first two and AG the last two positions of introns.

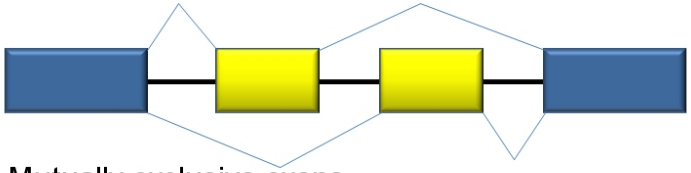
Alternative splicing



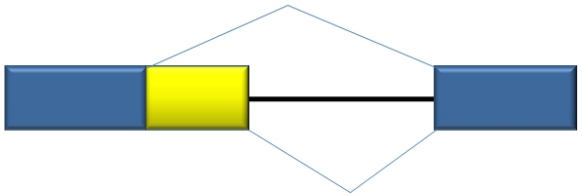
Alternative splicing modes



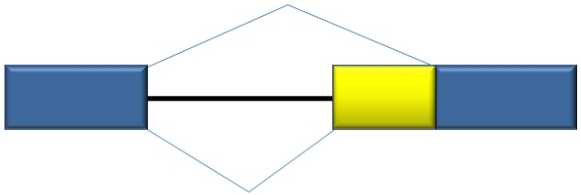
Exon skipping



Mutually exclusive exons



Alternative 5' donor sites

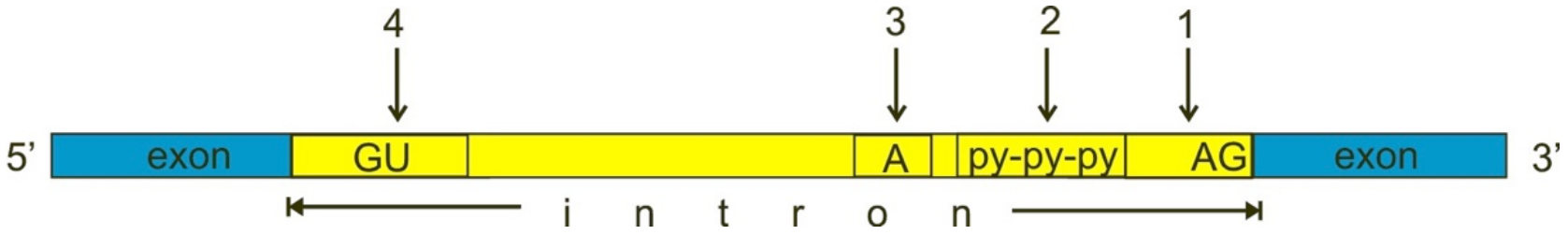


Alternative 3' acceptor sites

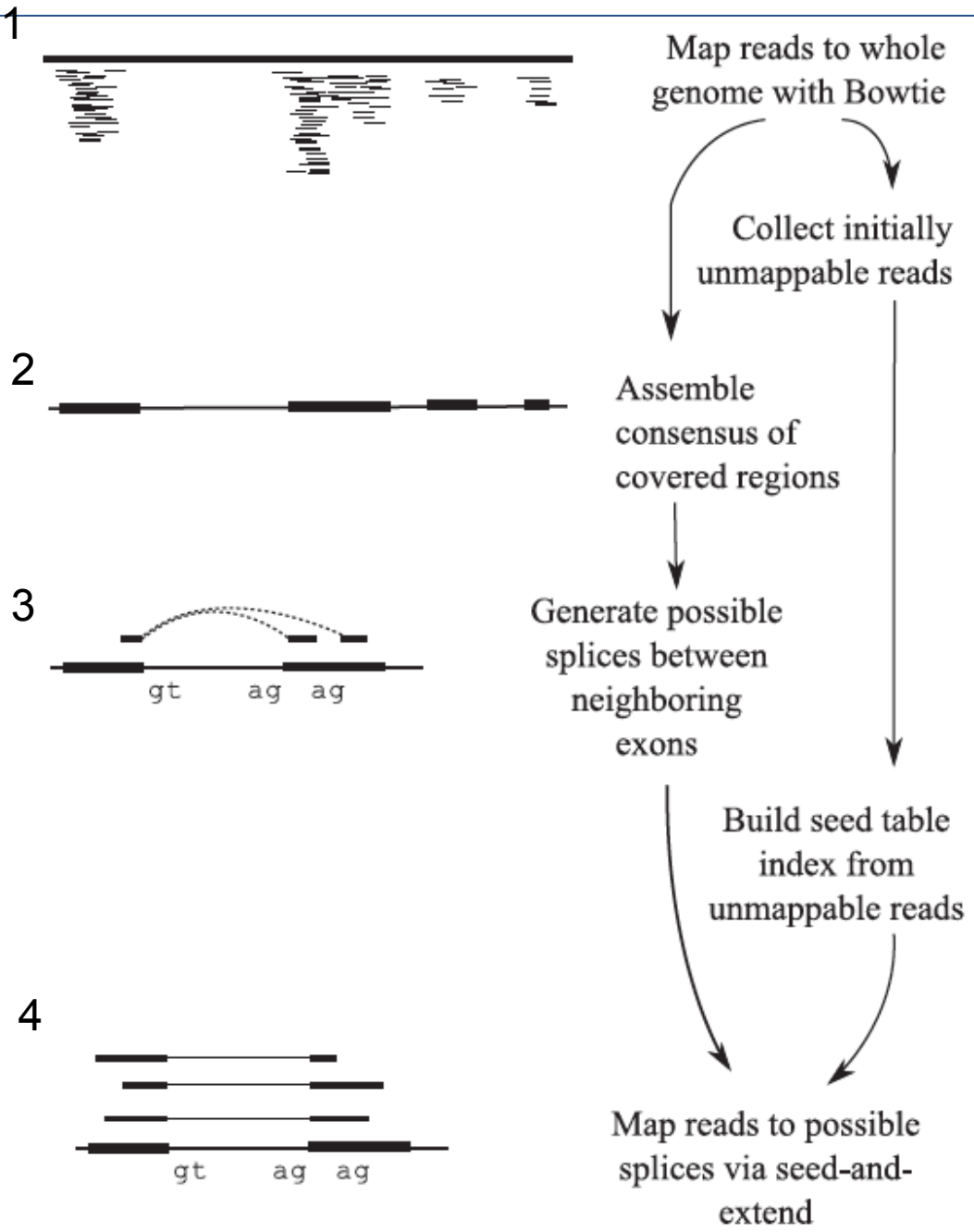
Intron retention

Introns (intragenic regions)

- Sites within introns required for splicing
  - **donor site**
    - 5' end of the intron,
    - GU sequence
  - **branch site**
    - near the 3' end of the intron
  - **acceptor site**
    - 3' end of the intron
    - AG sequence



Tophat



Trapnell et al. 2009

Tophat

Coverage islands: distinct regions of piled up reads

Island detection from *i* to *j*:

$$D_{ij} = \frac{\sum_{m=i}^{j} d_m}{j-i} \cdot \frac{1}{\sum_{m=0}^{n} d_m}$$

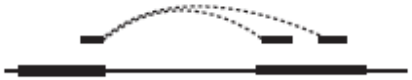
where

- d_m is the depth of coverage at coordinate *m*
- *n* is the genome length

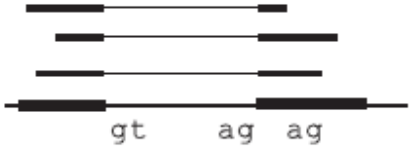
The D values scaled to [0,1000]
D>300 → island.

1

2



4



Map reads to whole genome with Bowtie

Collect initially unmappable reads

IUM reads

Assemble consensus of covered regions

Generate possible splices between neighboring exons

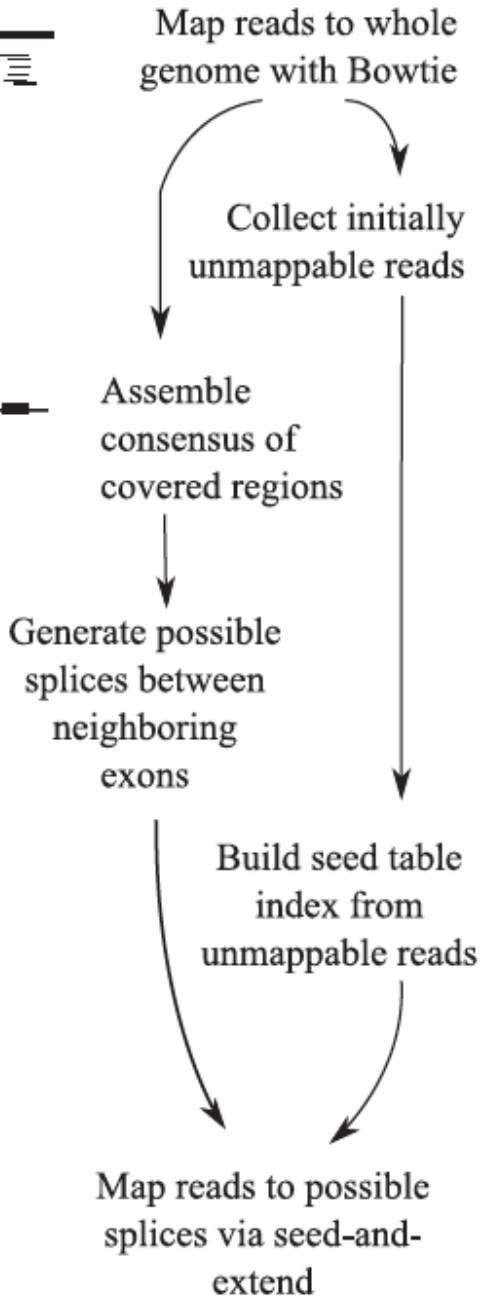
Build seed table index from unmappable reads

Map reads to possible splices via seed-and-extend

Trapnell et al. 2009

Tophat

2



Step 2:

Maq (Li et al, 2008)
→ "pseudoconsensus" assembly:

- Low quality bases replaced by the reference
- Splice junctions covered by few reads
  - Include a flanking sequence (45bp) from the reference on both sides of each island
- Coverage gaps < 6bp → exons merged

Trapnell et a

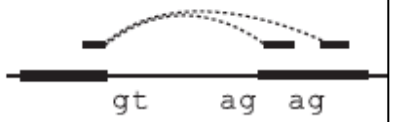
Tophat

Map reads to whole genome with Bowtie

Collect initially unmappable reads

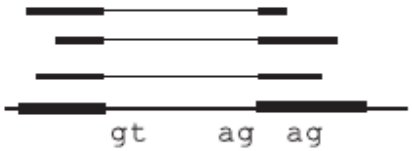
IUM reads

3



Step 3:

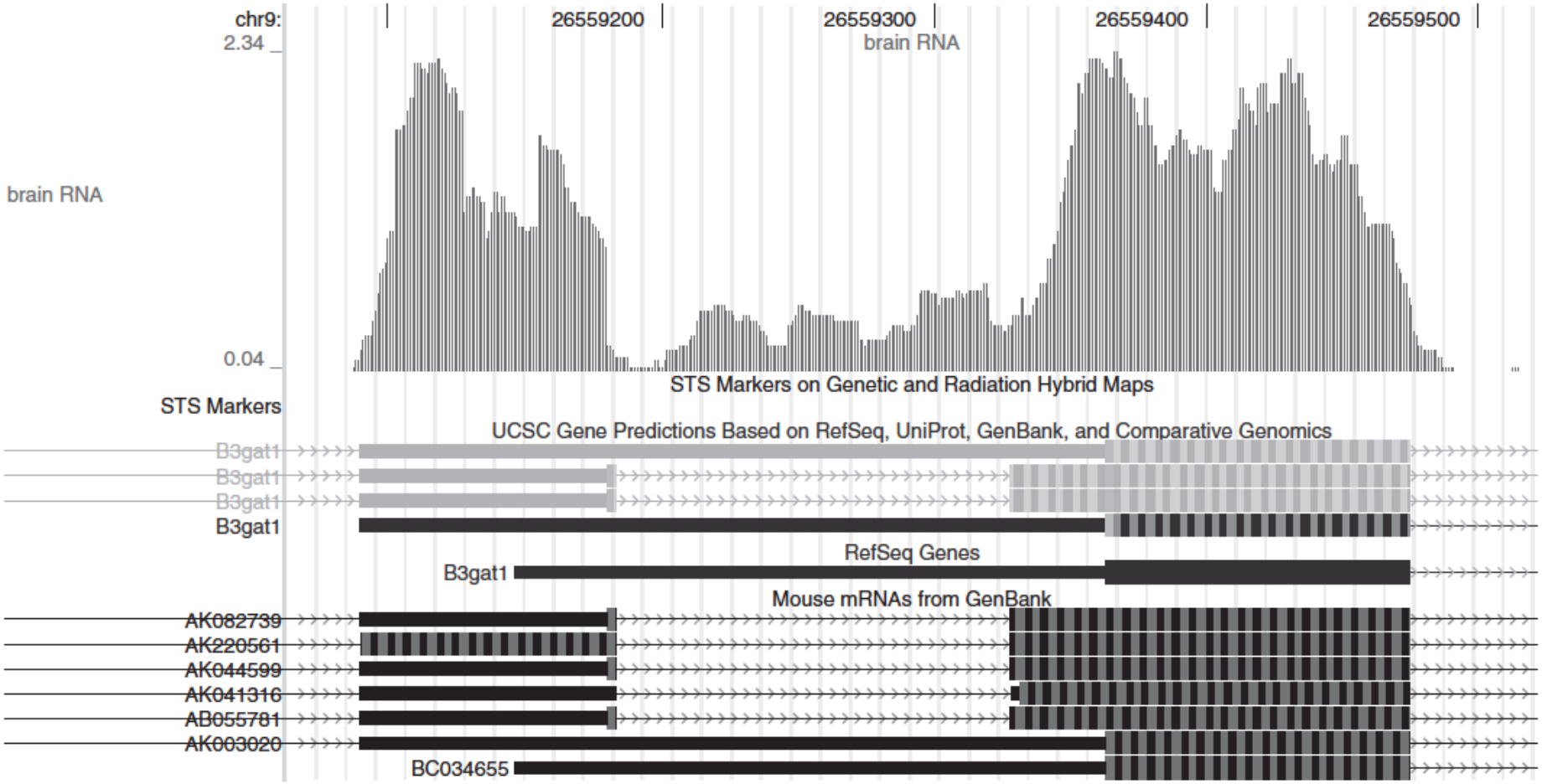
- Enumerate all donor-acceptor pairs within island sequences (and reverse complements)
- Only pairs between islands unless deep coverage
- Check each candidate intron (length from 70 to 20000 bp) with IUM reads



Map reads to possible splices via seed-and-extend

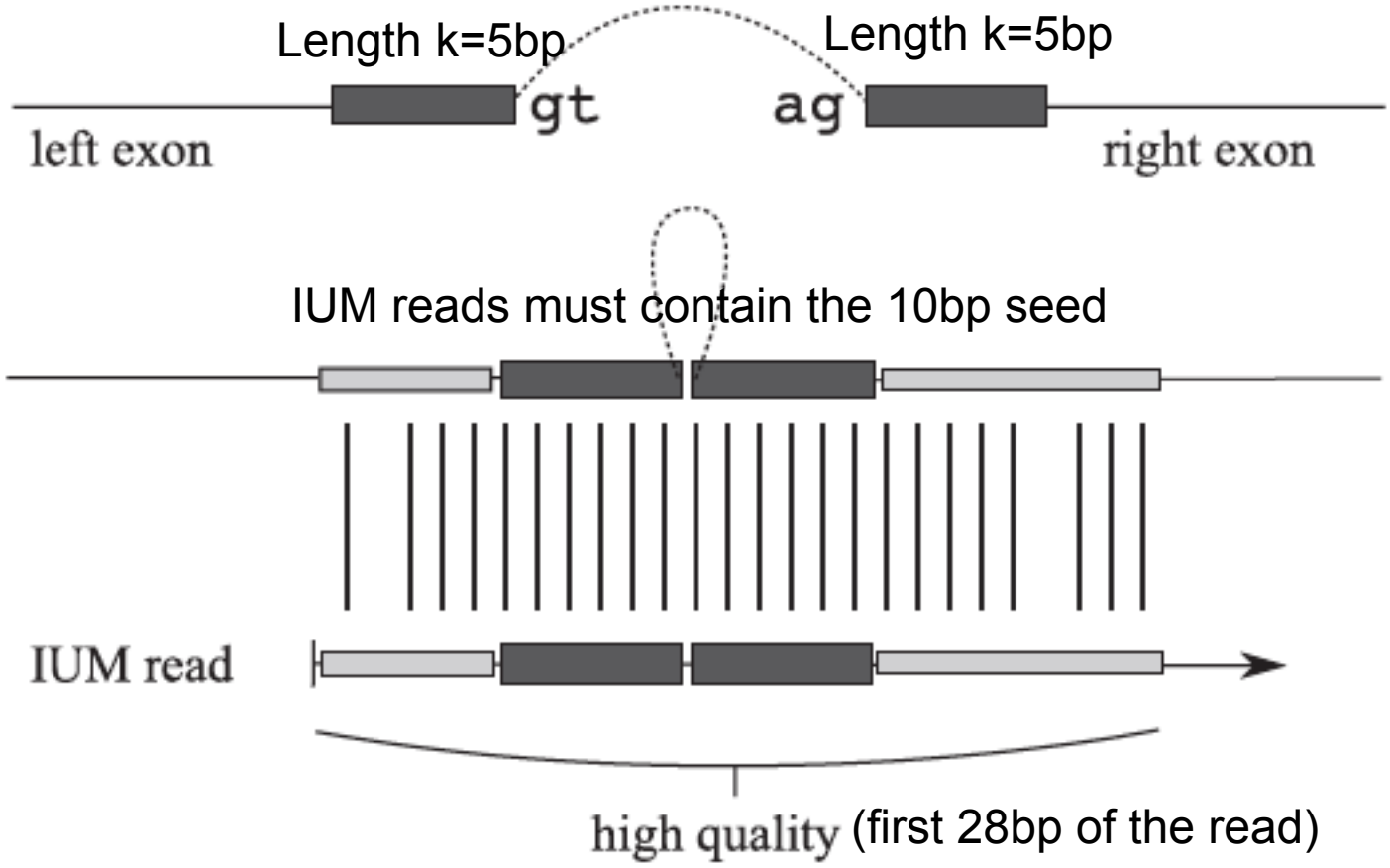
Trapnell et al. 2009

Single island intron



An intron entirely overlapped by 5’ end of another transcript

Seed and extend strategy to map reads to splice variants



Filtering the identified splice junctions

- 86% of the minor isoforms expressed at at least 15% of the level of the major isoform (Wang et al 2008).
- For each junction compute the depth of coverage for the left and right flanking regions
- Filter if the splice junction covered <15% of its flanking exons

RPKM

arbitrary scaling to 1 Million per 1 KB

per base average expression

$$RPKM(G) = \frac{R}{L} \cdot \frac{10^6 \cdot 10^3}{M}$$

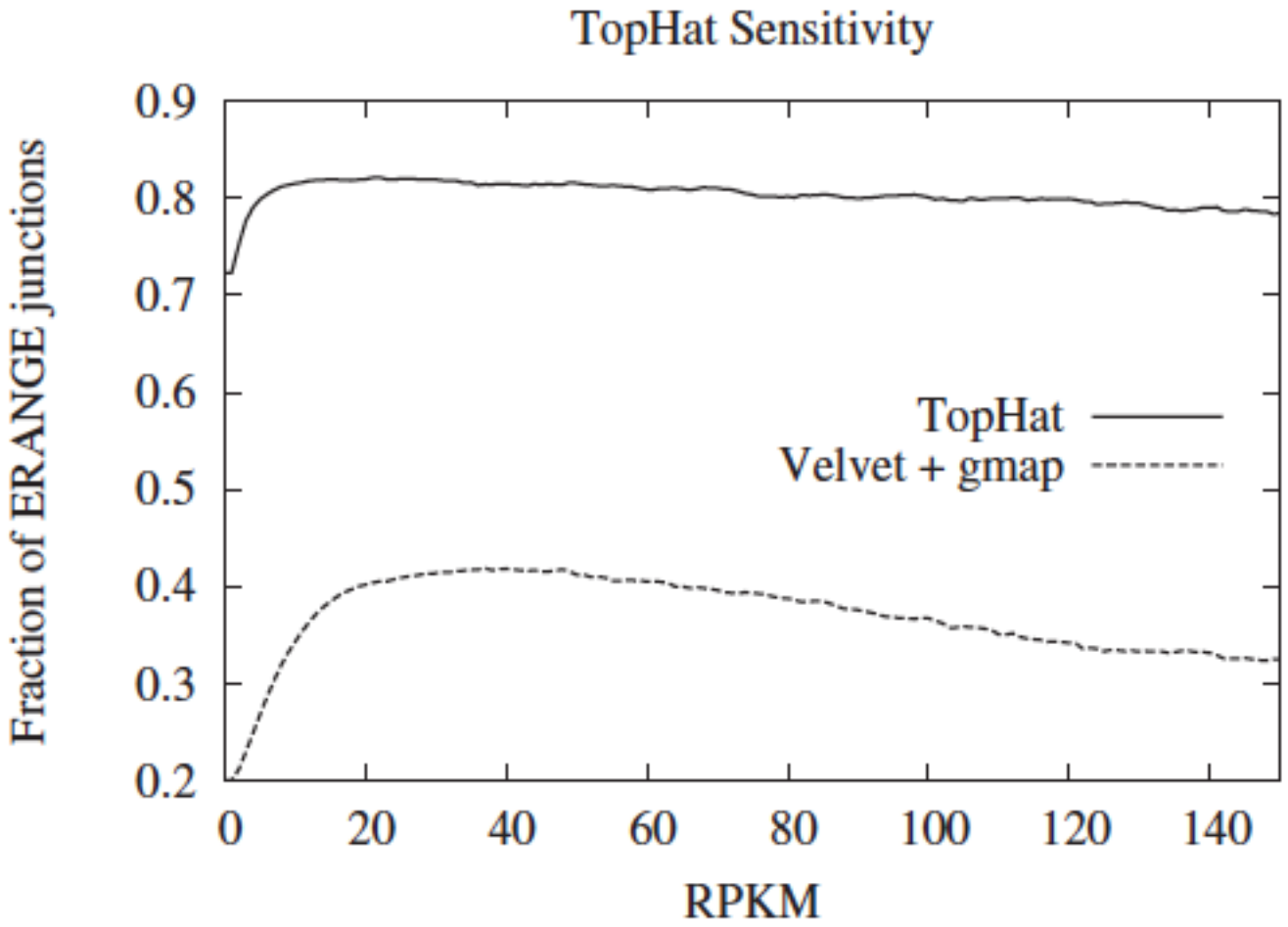
R = reads mapped to gene G
L =length of gene G
M=number of reads mapped in Experiment

experiment correction factor

Example: $$RPKM(G) = 0.5 \cdot \frac{10^6 \cdot 10^3}{10^7} = 50$$

In gene G, 50 reads are found every 1 KB for 1 million mapped reads

Tophat sensitivity



TopHat On simulated data

Table 1. TopHat junction finding under simulated sequencing of transcripts

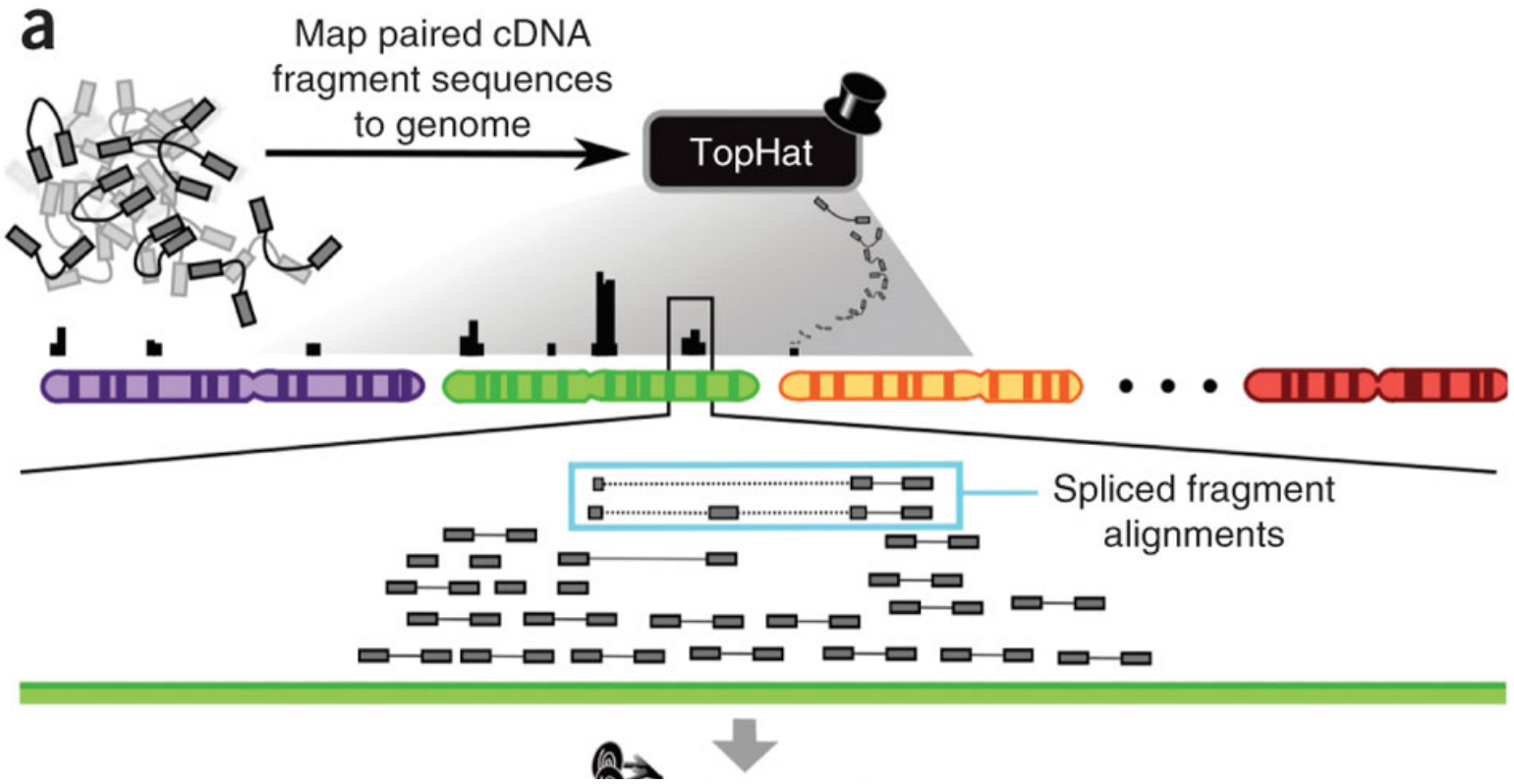
Depth of sequence coverage	True positives	Total (%)	False positives	Reported (%)
1	1744	17	114	6
5	7666	77	585	7
10	8737	88	428	4
25	9275	93	267	2
50	9351	94	235	2

The simulation sampled a set of transcripts with 9879 true splice junctions.

TopHat2

- accurate alignment of transcriptomes in the presence of
  - insertions,
  - deletions and
  - gene fusions

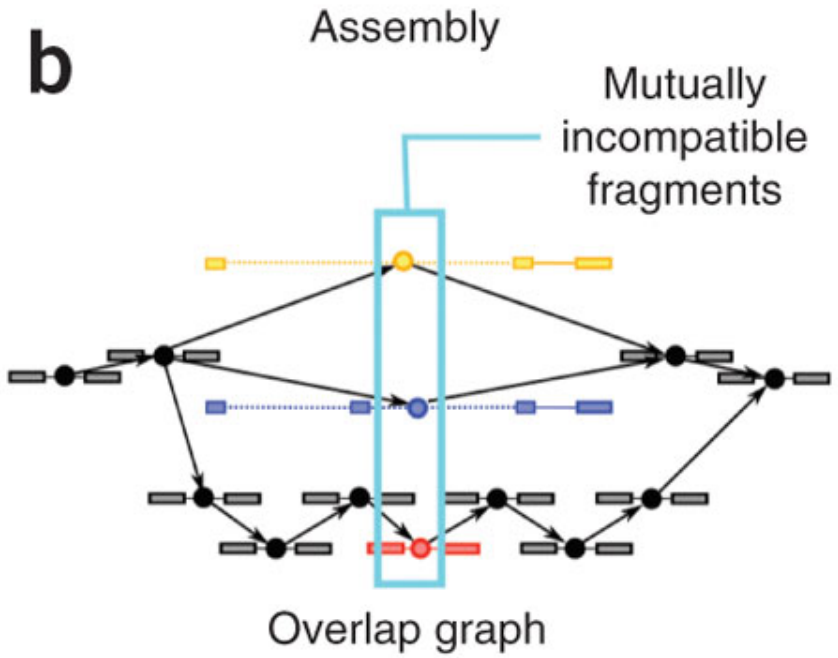
Cufflinks: input



- Input: fragment sequences that have been aligned to the genome by software capable of producing spliced alignments, such as TopHat

Trapnell et al., 2010

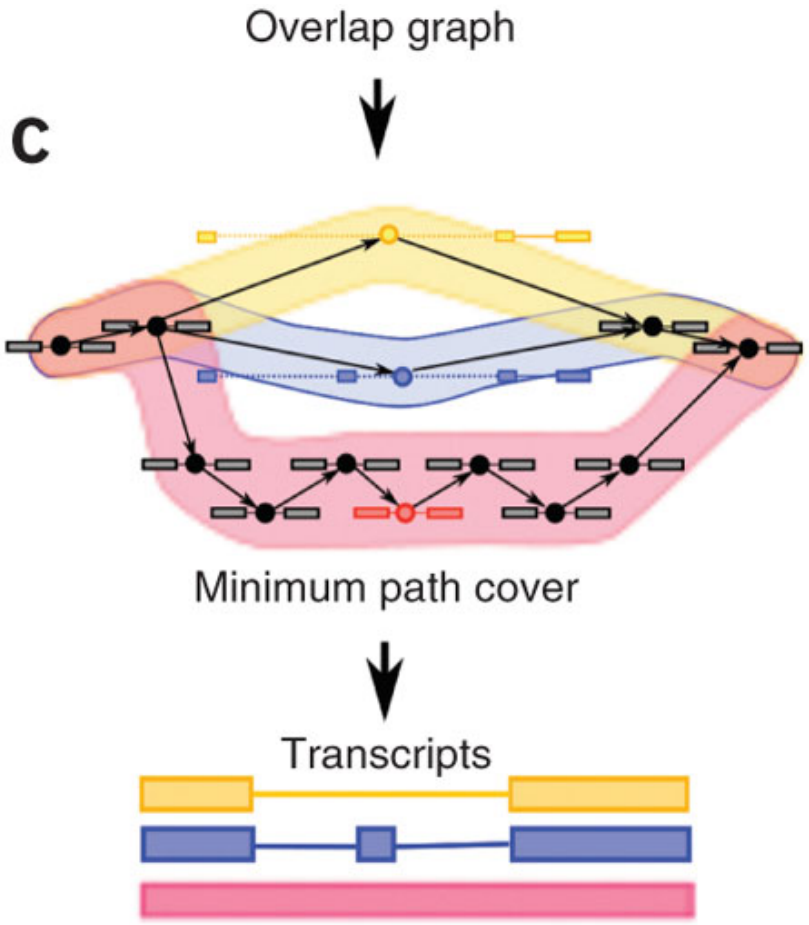
Cufflinks: assembly



- Assemble overlapping 'bundles' of fragment alignments separately, to reduce running time and memory use (each bundle typically contains the fragments from no more than a few genes)
- Identify pairs of 'incompatible' fragments that must have originated from distinct spliced mRNA isoforms

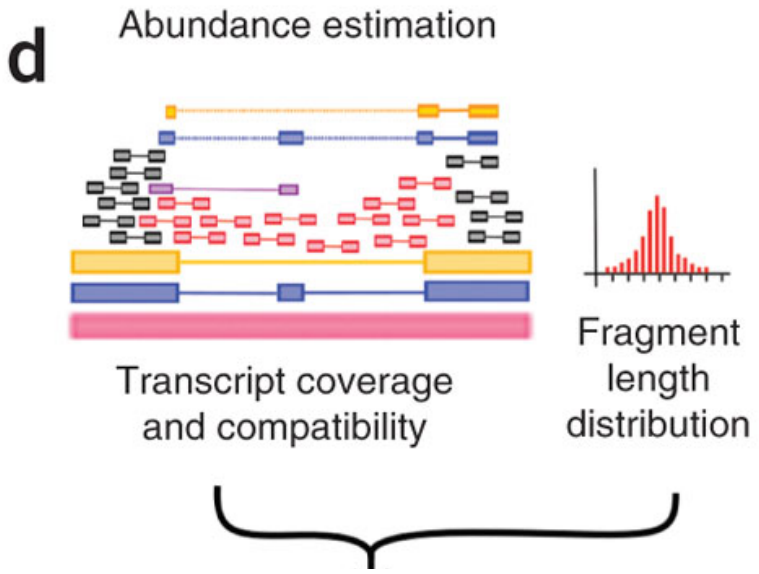
Cufflinks: overlap graph

- Fragments connected in an 'overlap graph' when
  - they are compatible
  - their alignments overlap in the genome.
- Each fragment has one node in the graph
- An edge, directed from left to right along the genome, is placed between each pair of compatible fragments.
- Isoforms are then assembled from the overlap graph



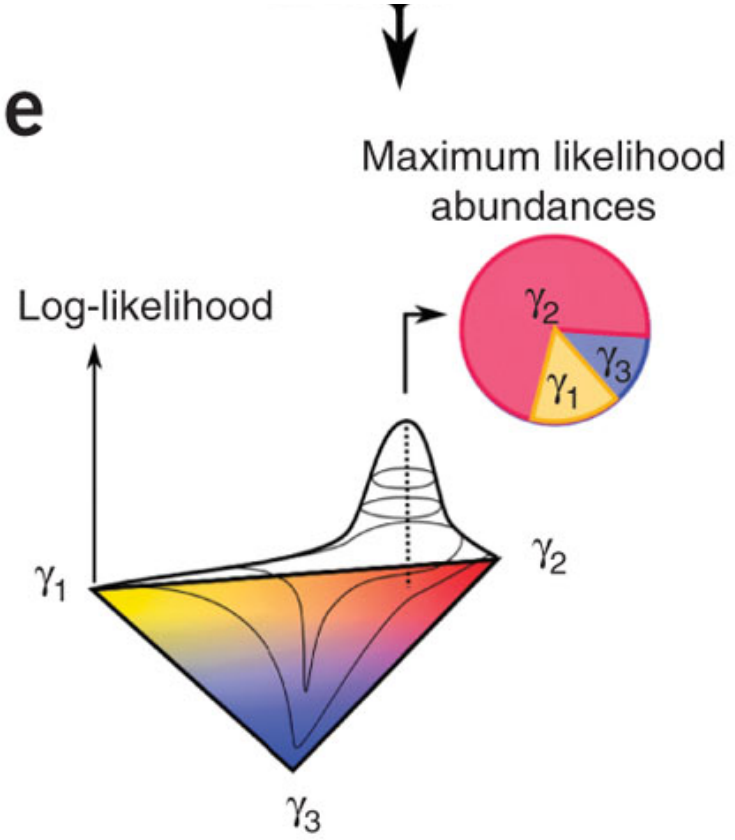
Cufflinks: abundance estimation

- Fragments are matched to transcripts from which they could have originated (color)
- The violet fragment: from the blue or red isoform. Gray fragments: any of the three.
- Statistical model
  - the probability of observing each fragment is a linear function of the abundances of the transcripts from which it could have originated.
  - Because only the ends of each fragment are sequenced, the length of each may be unknown.
  - Assigning a fragment to different isoforms →a different length.



- Length distribution incorporated to help assign fragments to isoforms.
- For example, the violet fragment would be much longer, and very improbable if from the red isoform instead of the blue isoform.

Cufflinks: Maximum likelihood abundancies



- Maximize a function that assigns a likelihood to all possible sets of relative abundances of the isoforms (yellow, red and blue: γ1,γ2,γ3)
- output: the abundances that best explain the observed fragments

Cufflinks pipeline

- Cufflinks: assembles transcriptomes from RNA-Seq data and quantifies their expression.

- Cuffcompare: helps perform comparisons of assembled transcriptiome to a reference and assess the quality of assembly.

- Cuffmerge: For multiple RNA-Seq libraries and assembled transcriptomes from each of them, merges these assemblies into a master transcriptome. Required for a differential expression analysis of the new assembled transcripts.

- Cuffquant (Cufflinks version >2.2.0): allows to compute the gene and transcript expression profiles and save these profiles to files that can be analyzed later with Cuffdiff or Cuffnorm. This can help to distribute the computational load over a cluster and is recommended for analyses involving more than a handful of libraries.

Cufflinks pipeline

- Cuffdiff: Comparing expression levels of genes and transcripts in RNA-Seq. Can tell not only which genes are up- or down-regulated between two or more conditions, but also which genes are differentially spliced or are undergoing other types of isoform-level regulation.

- Cuffnorm: normalizes the expression levels from a set of RNA-Seq libraries so that they're all on the same scale, facilitating downstream analyses such as clustering. Expression levels reported by Cufflinks in FPKM units are usually comparable between samples, but in certain situations, applying an extra level of normalization can remove sources of bias in the data. Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results obtained with other downstream tools.

Trapnell et al. 2010 (cufflinks) results

- sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series
- 13,692 known transcripts detected
- 3,724 previously unannotated ones
  - 62% are supported by independent expression data or by homologous genes in other species.

Bibliography

- Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi,Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold & Lior Pachter, *Nature Biotechnology* **28**, 511–515 (2010)

- Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics doi:10.1093/bioinformatics/btp120