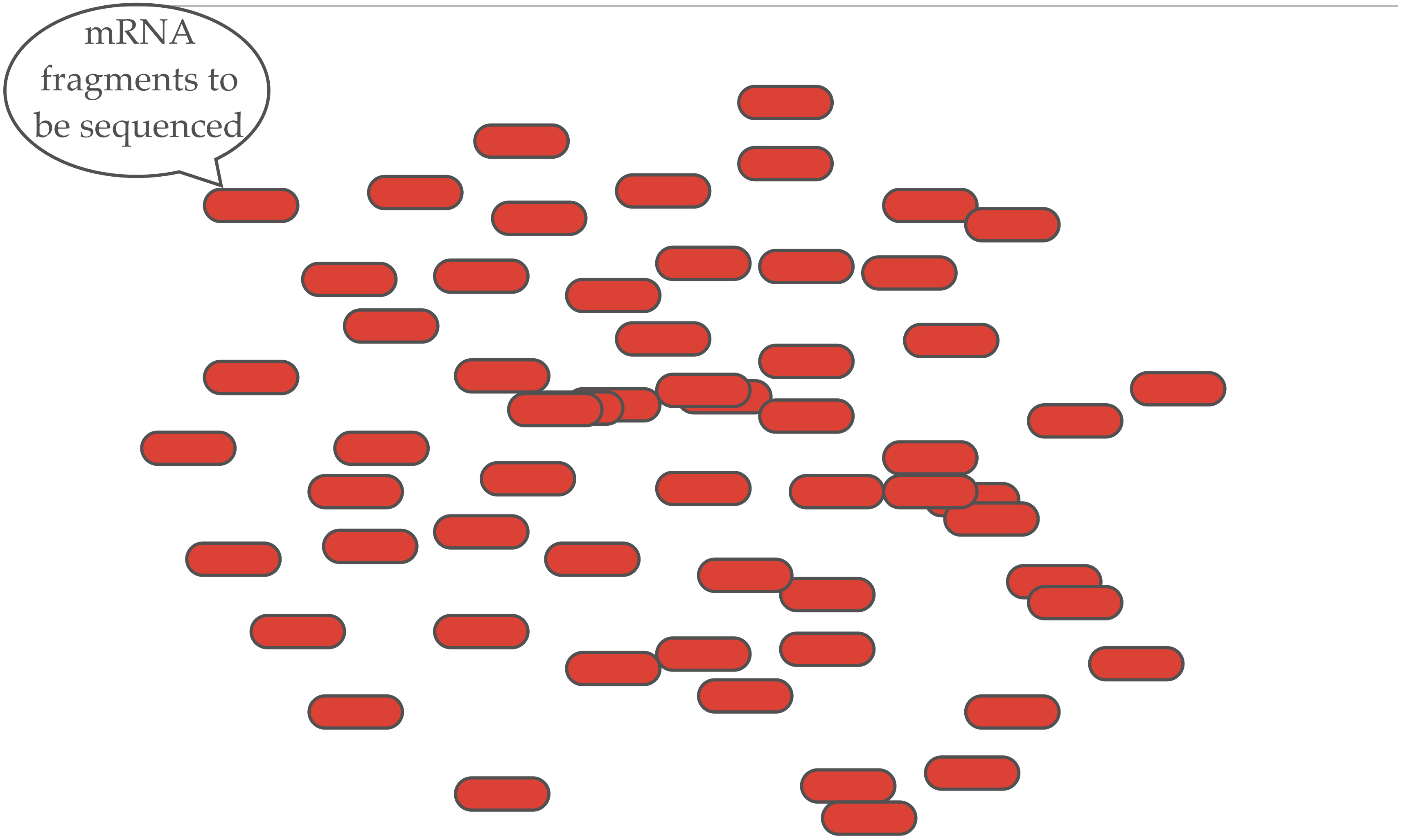
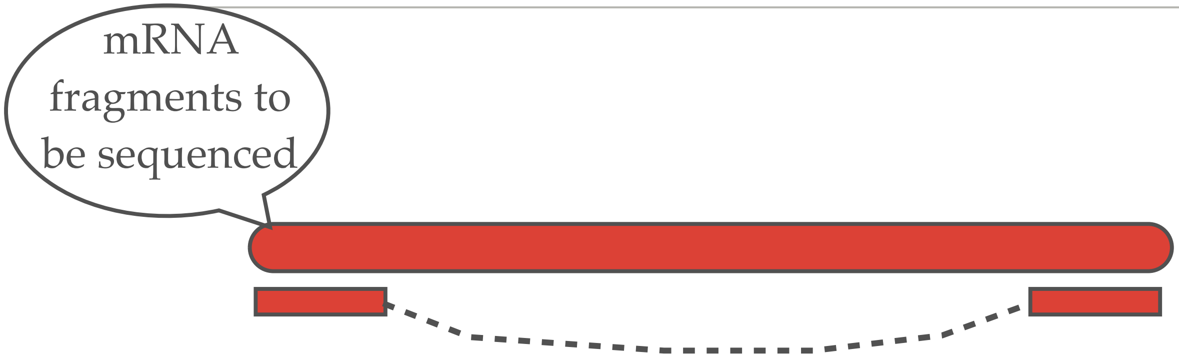
RNAseq: isoform expression quantification and transcript assembly

Slides courtesy from S. Salzberg, C. Trapnell, L. Pachter and K. Okrah

Sec-gen Sequencing

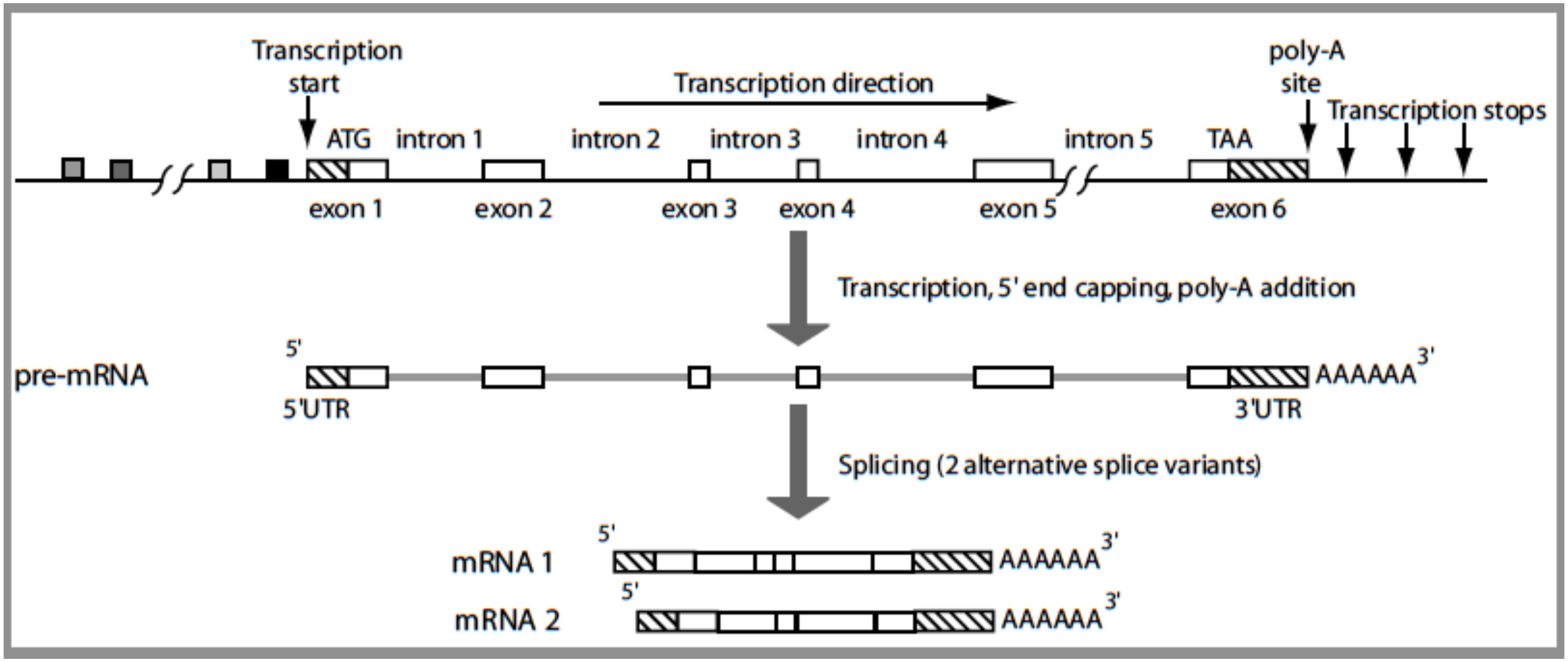


Sec-gen Sequencing
Paired-Ends



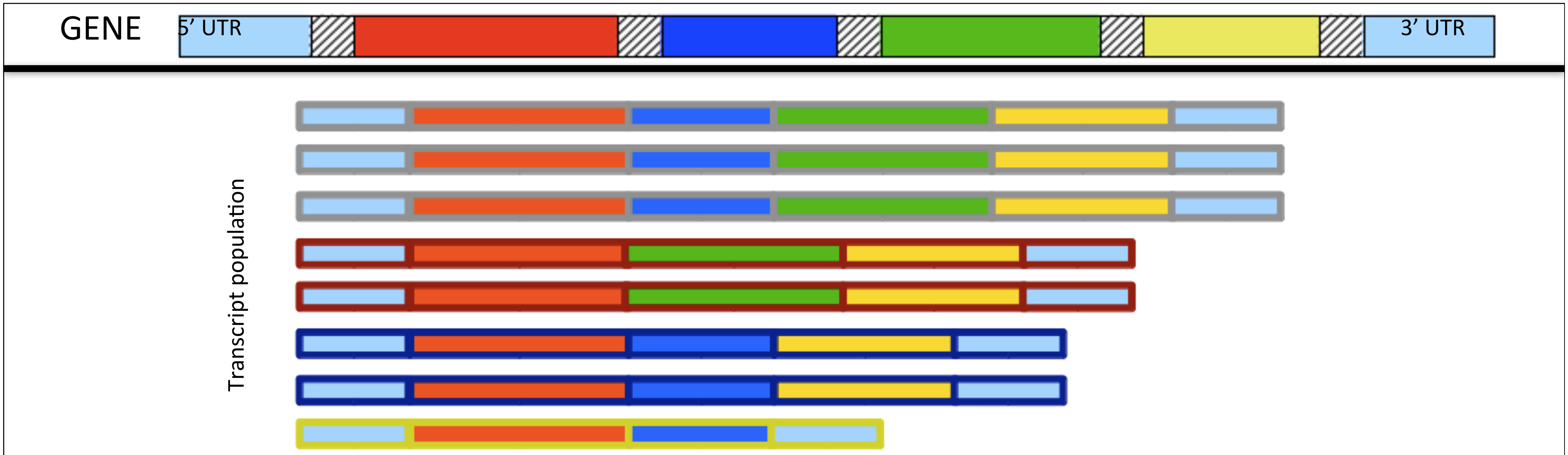
In paired-end sequencing reads are generated from both ends of a fragment

Recall:

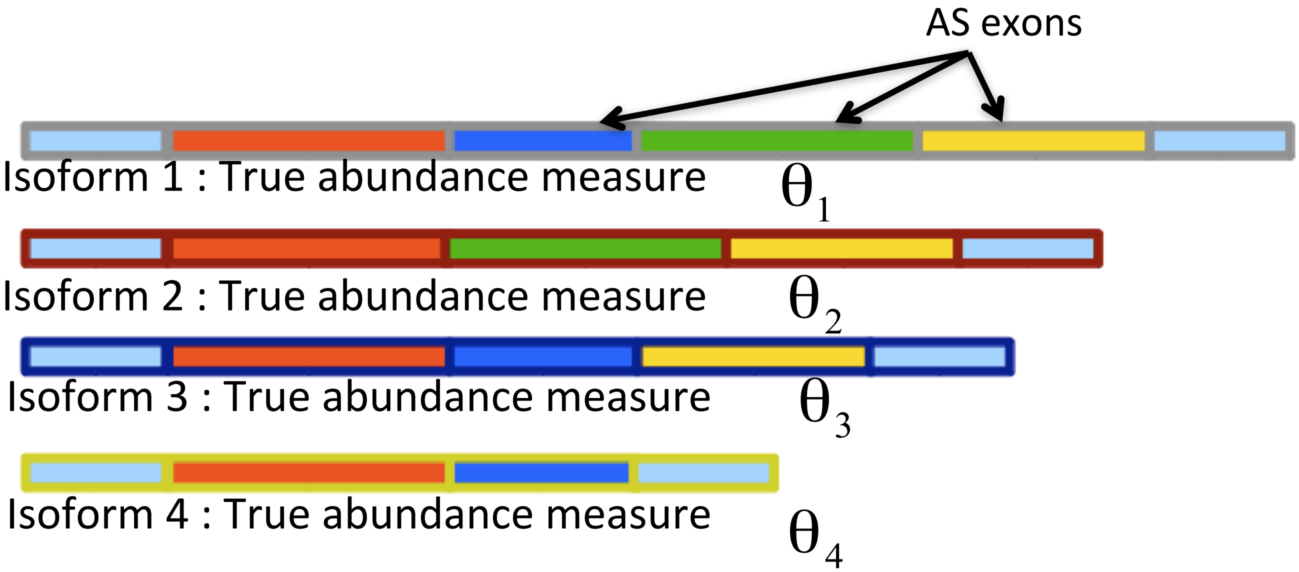


Source: Computational Genome Analysis

Goal: Develop and analyze a statistical model for measuring differential expression of **Isoforms** of the same gene using Rna-Seq.



Suppose we have a gene with 4 isoforms and 3 alternatively spliced (AS) exons as shown above.



Isoform 1 : True abundance measure θ_1

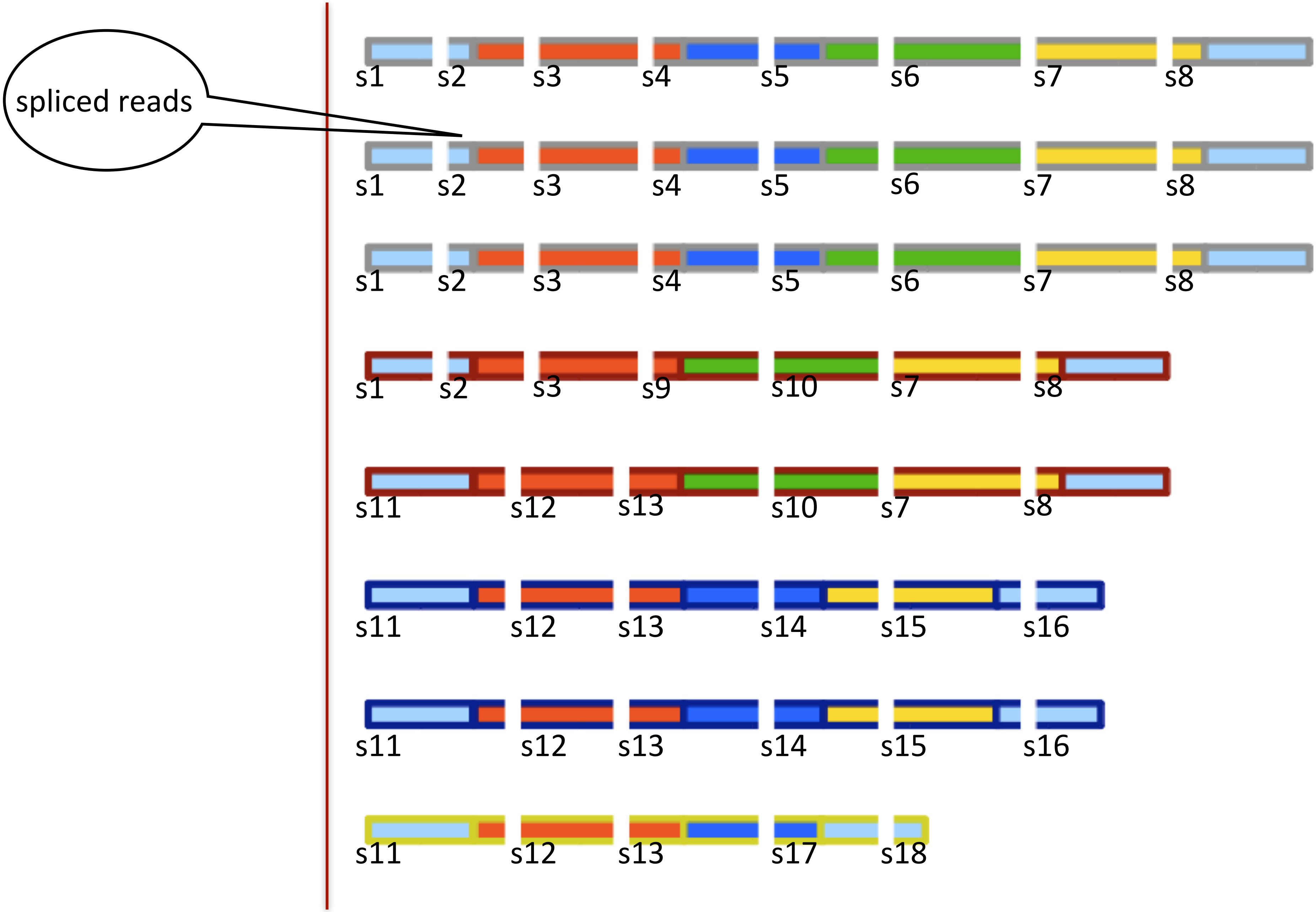
Isoform 2 : True abundance measure θ_2

Isoform 3 : True abundance measure θ_3

Isoform 4 : True abundance measure θ_4

The goal is to estimate the true abundance measure of the 4 isoforms.

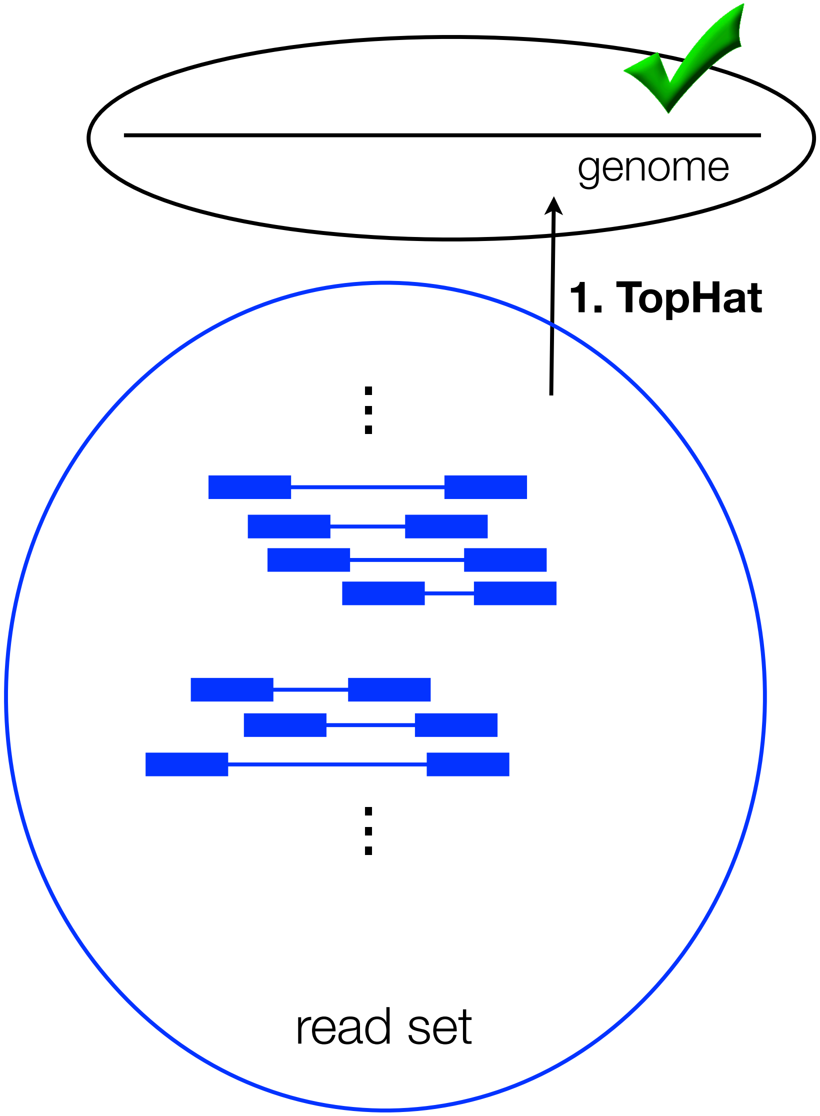
Fragmented mRNAs: 54 total reads with 18 unique types.



spliced reads

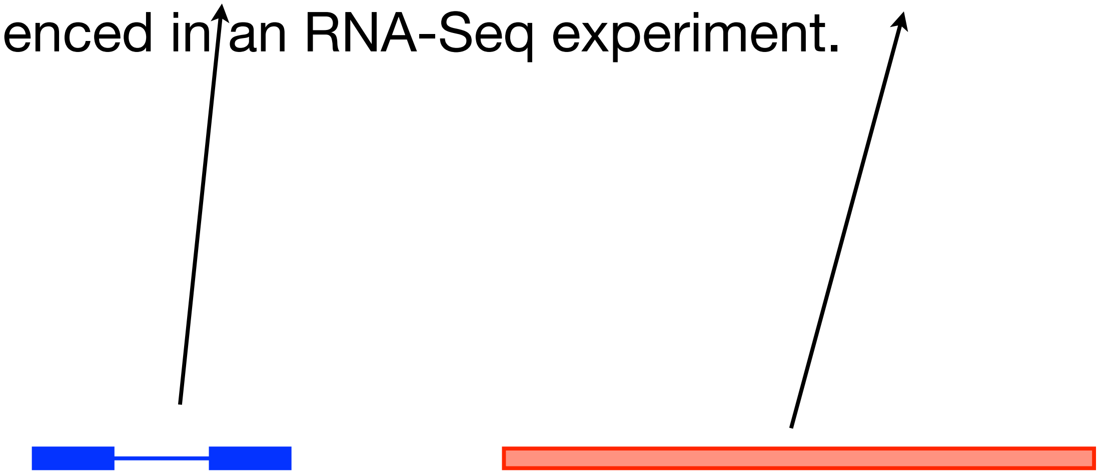
TopHat for second generation RNA-Seq: spliced read alignment

- Suitable for
  - short reads (25-50bp)
  - long reads (100+ bp)
  - paired end reads
- New features since 0.8x (Trapnell et al., *Bioinformatics* 2009)
  - Much faster, almost fully threaded
  - Semi-canonical introns (GC-AG and AT-AC) and some support for microexons



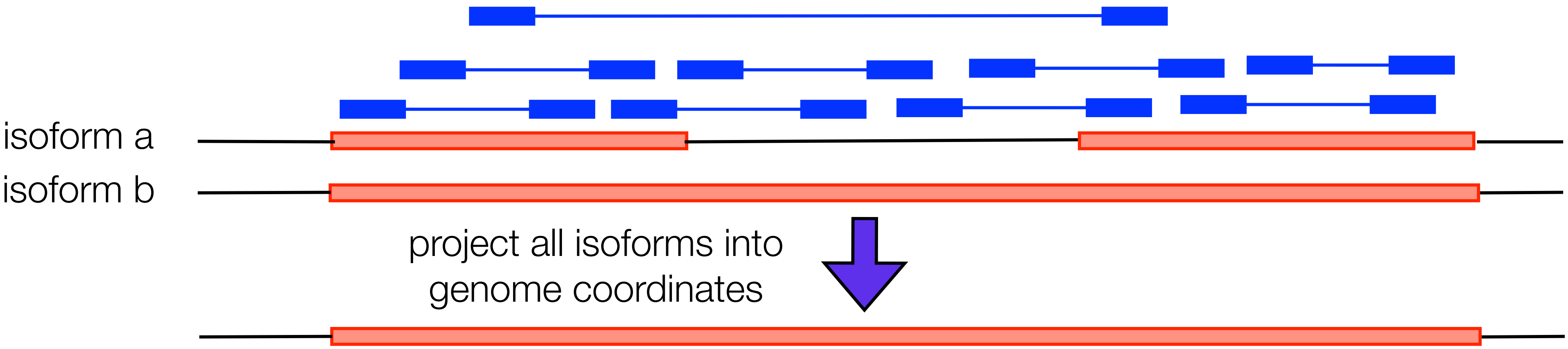
FPKM

- Expected number of **F**ragments **P**er **K**ilobase (of transcript) per **M**illion fragments sequenced in an RNA-Seq experiment.



- These units are proportional to the θ_i.

Projective normalization underestimates expression



isoform a

isoform b

project all isoforms into genome coordinates

R reads total, *r* reads for the gene:

- r_a for isoform *a*
- r_b for isoform *b*

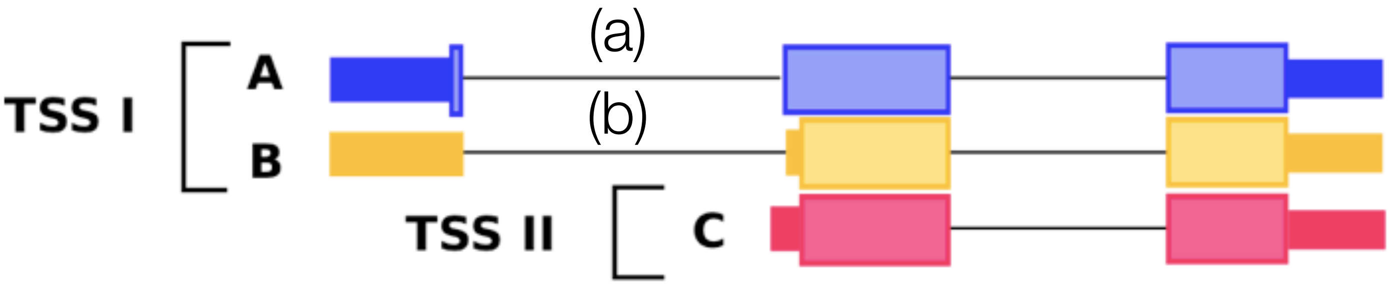
$$FPKM_g = \frac{1}{R}\left(\frac{r_a}{length_a}\right) + \frac{1}{R}\left(\frac{r_b}{length_b}\right)$$

$$FPKM_{proj(g)} = \frac{1}{R}\left(\frac{r_a + r_b}{length_{proj(g)}}\right)$$

but $\frac{r_a}{length_a} \geq \frac{r_a}{length_{proj(g)}}, \frac{r_b}{length_b} \geq \frac{r_b}{length_{proj(g)}}$ so

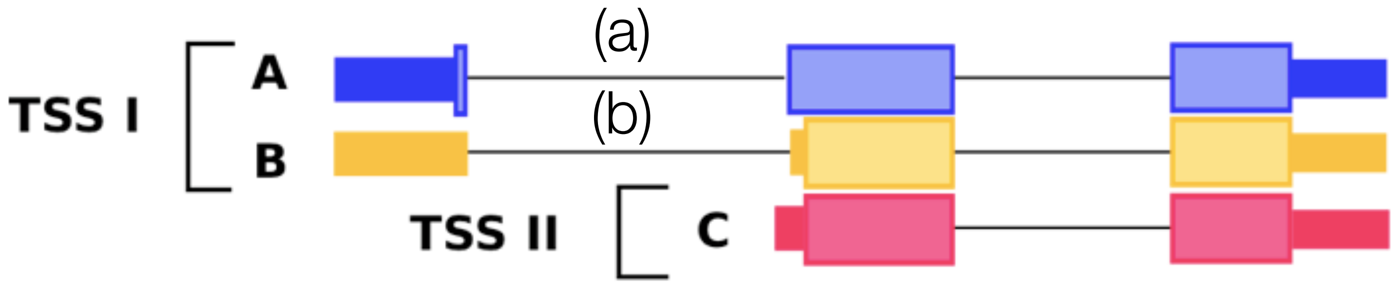
$$FPKM_g \geq FPKM_{proj(g)}$$

How should expression levels be estimated?



- A-B are distinguished by the presence of splice junction (a) or (b).
- A-C are distinguished by the presence of splice junction (a) and change in UTR
- B-C are distinguished by the presence of splice junction (b) and change in UTR

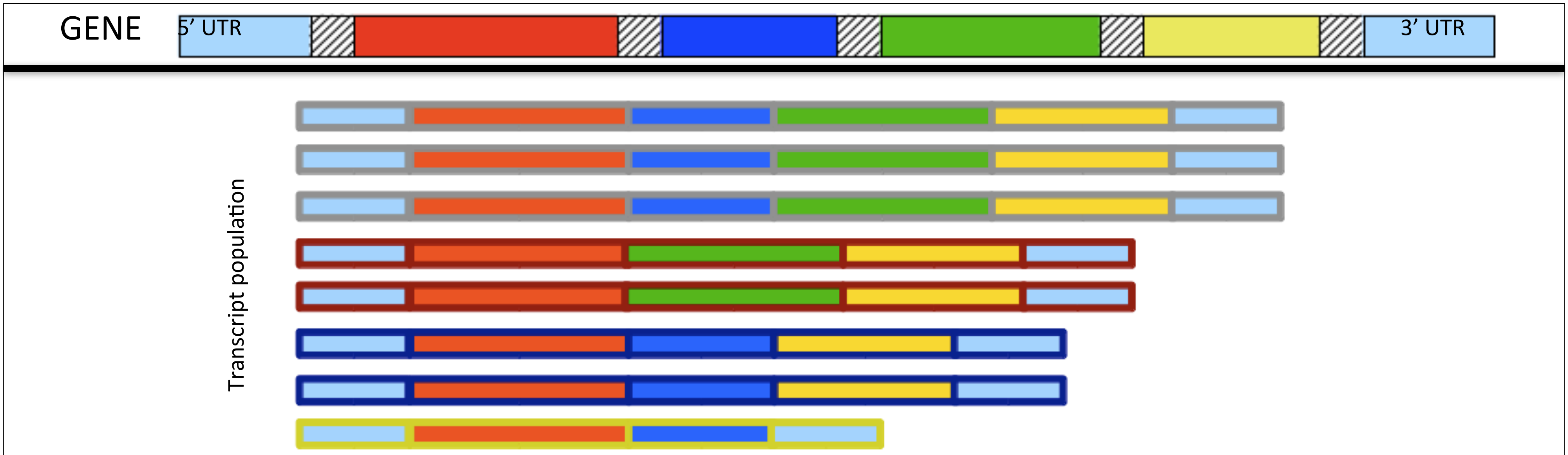
How should expression levels be estimated?



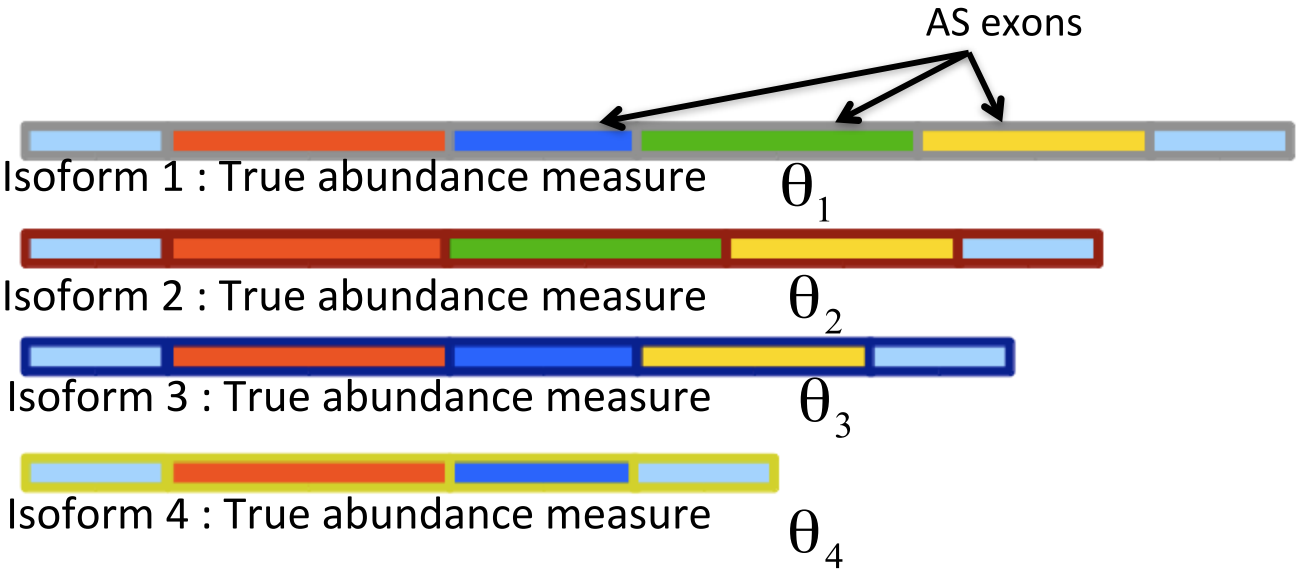
- Longer transcripts contain more reads.
- Reads that could have originated from multiple transcripts are informative.
- Relative abundance estimation requires “discriminatory reads”.

Isoform-level expression quantification

Jiang and Wong. Bioinformatics, 2009.
Salzman, Jiang and Wong. Statistical Science, 2011.



Suppose we have a gene with 4 isoforms and 3 alternatively spliced (AS) exons as shown above.



Isoform 1 : True abundance measure θ_1

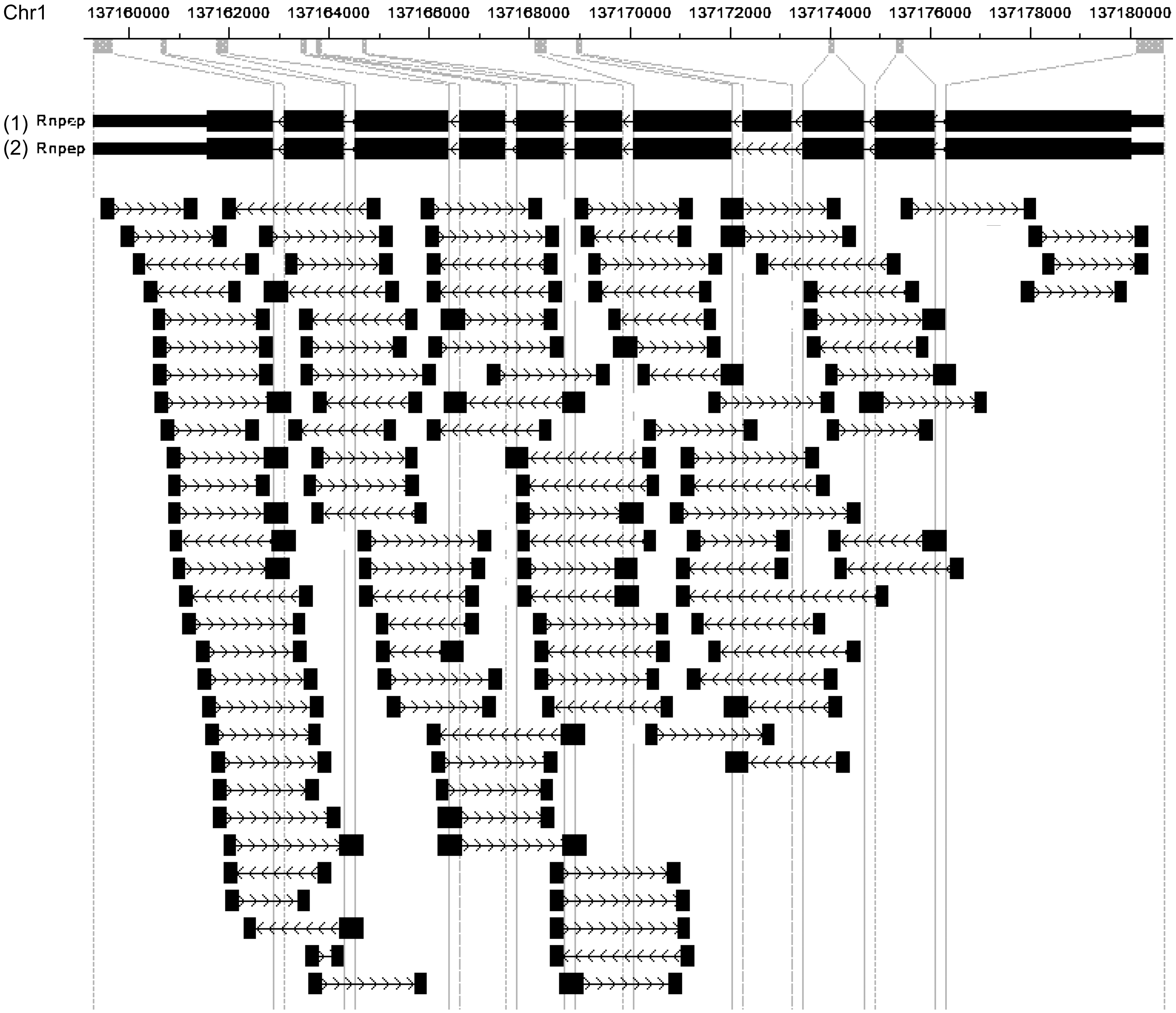
Isoform 2 : True abundance measure θ_2

Isoform 3 : True abundance measure θ_3

Isoform 4 : True abundance measure θ_4

The goal is to estimate the true abundance measure of the 4 isoforms.

Example: mouse RNAseq data

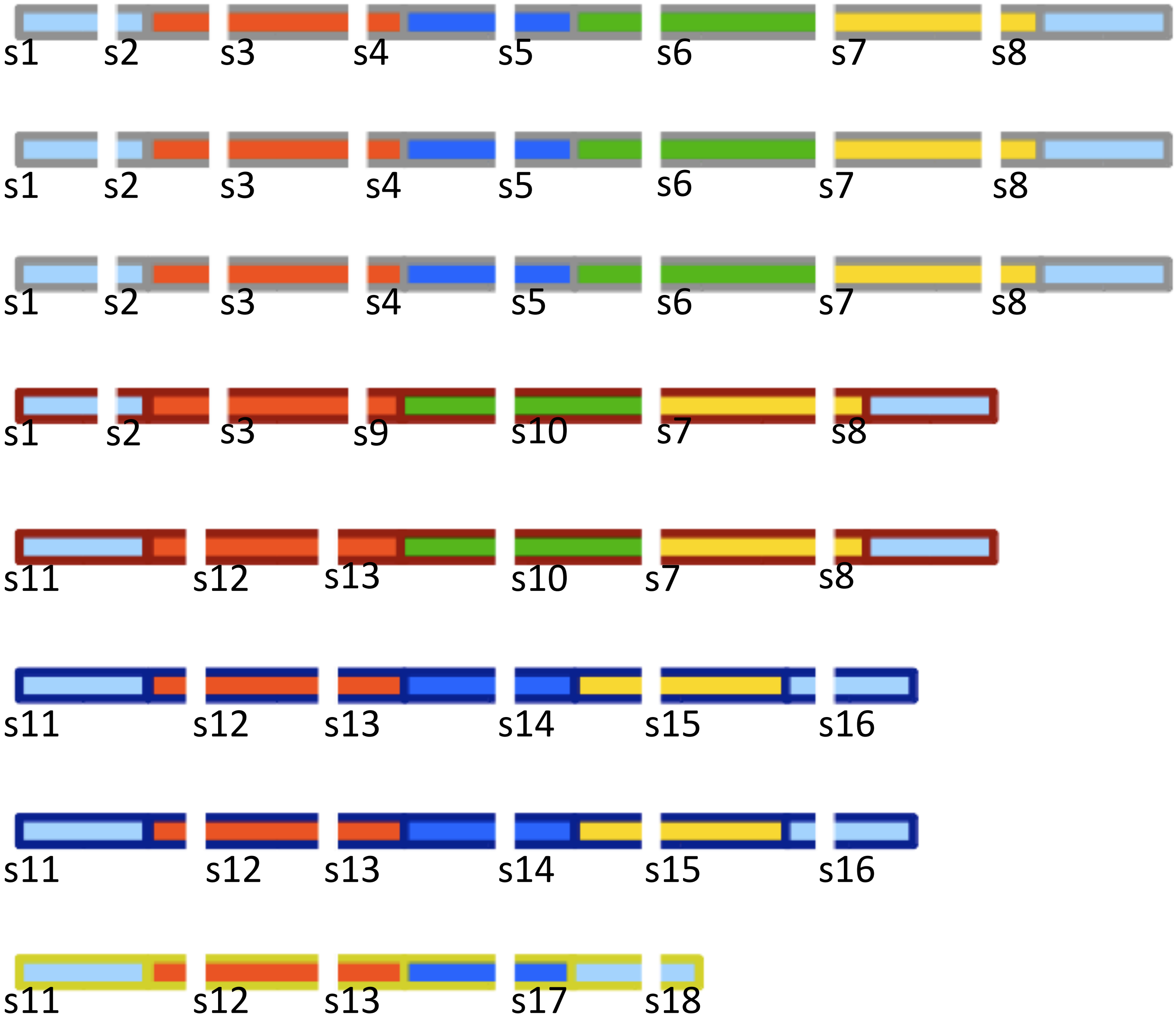


Fragmented mRNAs: 54 total reads with 18 unique types.

Sampling rate:

The ability for each of the 54 reads to be sequenced depends on:

1.Transcript fragmentation.
2. Size selection.
3. Sequence specific amplification of selection.



3.3 Likelihood Function

n_{ij} matrix = the number of reads type s_j generated by transcript θ_i.

	s1	**s2**	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s18	
θ1	3	3	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	24
θ2	1	1	1	0	0	0	2	2	1	2	1	1	1	0	0	0	0	0	13
θ3	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	0	0	12
θ4	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	5
n_j	4	4	4	3	3	3	5	5	1	2	4	4	4	2	2	2	1	1	54

For each read type, we only observe n_j.

We want to estimate last column (transcript abundance).

Last lecture concentrated on using the sum over the entire table (54) for positions that overlap *every* transcript

3.3 Likelihood Function

n_{ij} matrix = the number of reads type s_j generated by transcript θ_i.

	s1	**s2**	s3	s4	s5	s6	s7	s8	s9	s10	s11	s12	s13	s14	s15	s16	s17	s18	
θ1	3	3	3	3	3	3	3	3	0	0	0	0	0	0	0	0	0	0	24
θ2	1	1	1	0	0	0	2	2	1	2	1	1	1	0	0	0	0	0	13
θ3	0	0	0	0	0	0	0	0	0	0	2	2	2	2	2	2	0	0	12
θ4	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	5
n_j	4	4	4	3	3	3	5	5	1	2	4	4	4	2	2	2	1	1	54

- In reality we only observe $n_j = \sum_{i=1}^{I} n_{ij}$.
- $n_j \sim Poisson(\sum_{i=1}^{I} \theta_i a_{ij} = \theta^T a_j)$, where $\theta = \begin{bmatrix} \theta_1 \\ \dots \\ \theta_I \end{bmatrix}$, $a_j = \begin{bmatrix} a_{1j} \\ \dots \\ a_{Ij} \end{bmatrix}$.
- Likelihood: $f_\theta(n_1, n_2, \dots, n_J) = \prod_{j=1}^{J} \frac{(\theta^T a_j)^{n_j} e^{-\theta^T a_j}}{n_j!}$.

Uniform sampling model

- Appropriate for single read data. (transcript length is not considered)

Model for A:

$a_{ij} = 0$ if transcript i cannot generate read s_j,

otherwise,

$a_{ij} = n$, where n is the total number of reads.

Interpretation of abundance:

This choice of this A means that $\theta_i = \frac{c_i}{\sum_i l_i c_i}$,

Remember FPKM!

where l_i is the length of transcript i and c_i isthe number of copies in the ith transcript in the sample.

How do you fit it?

- The use of Poisson model makes things very easy

- The idea is to use *Maximum Likelihood Estimation:* find estimates that maximize the probability of observed data under Poisson model!

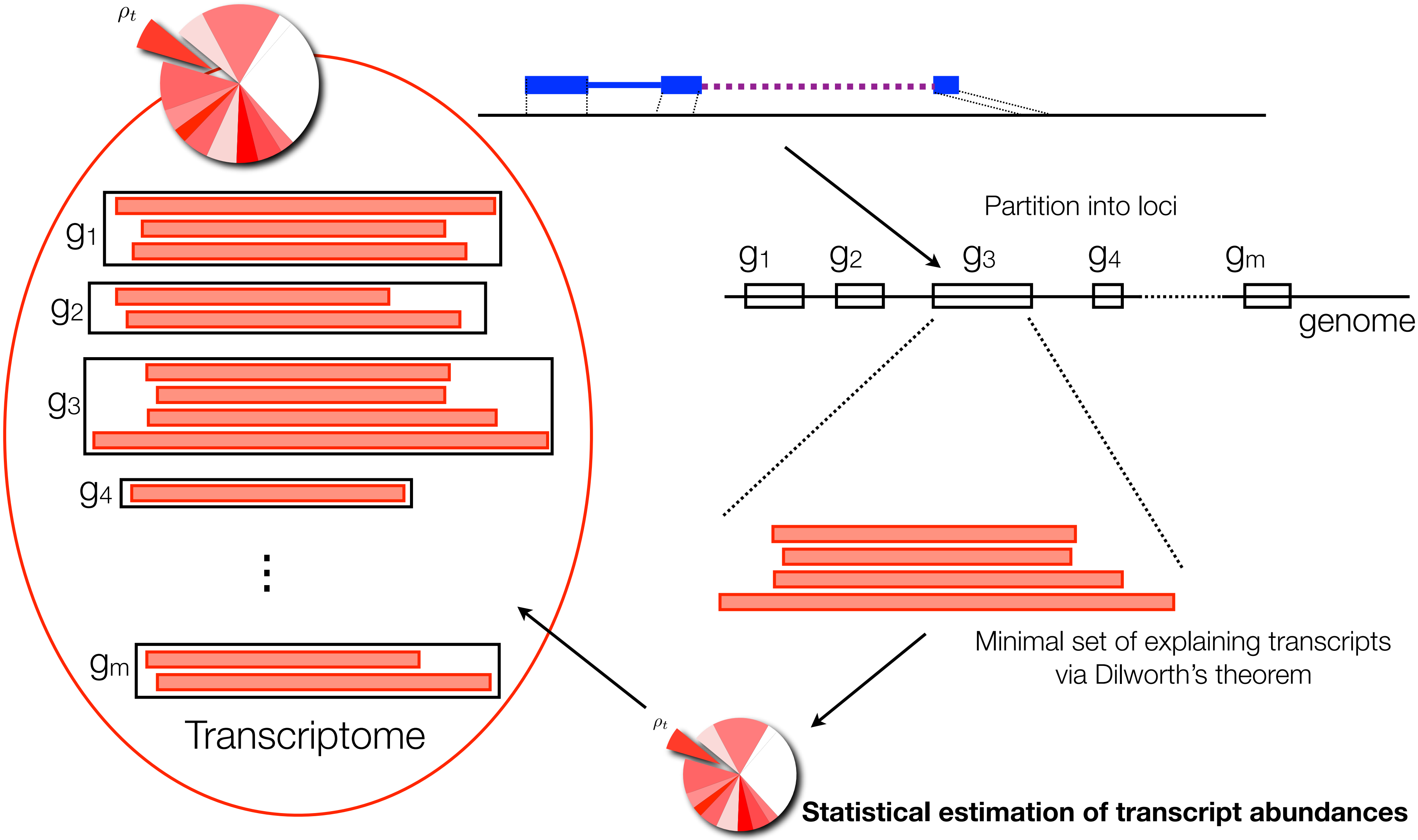
- Equivalent to a convex optimization problem:

$$\begin{array}{rc} \text{maximize} & n^T \log(A^T\theta) - \text{sum}(A^T\theta) \\ \text{s.t.} & \theta \geq 0 \end{array}$$

RNAseq: transcript assembly and quantification

All Slides courtesy from S. Salzberg, C. Trapnell and L. Pachter
Trapnell, et al. Nature Biotechnology, 2010.

Overview of cufflinks



Transcriptome

Partition into loci

Minimal set of explaining transcripts via Dilworth’s theorem

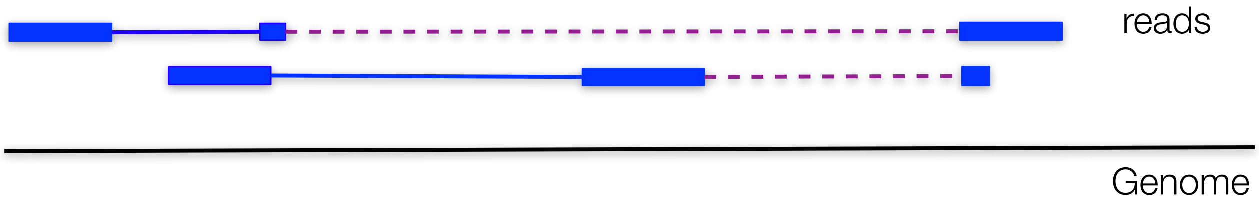
Statistical estimation of transcript abundances

Comparative transcript assembly

- Desirable properties of an assembly: consistency, parsimony and identifiability.
- Dilworth's theorem and its application to transcript assembly.
- The Cufflinks assembler.
- Promoter discovery and novel isoforms.
- Lessons learned.

Transcriptome assembly with a reference genome

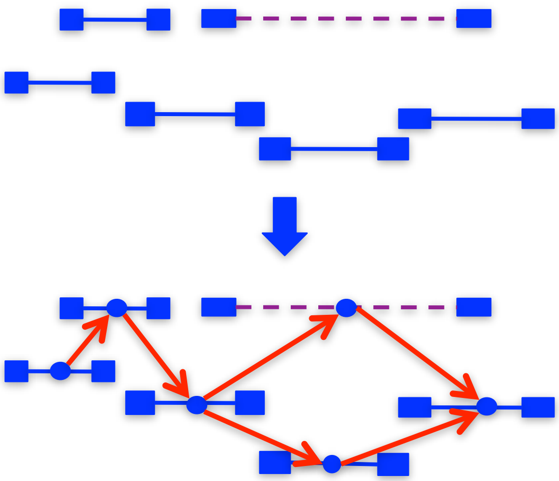
Don’t know that two reads came from the same transcripts, but sometimes know that they came from **different** transcripts



How many transcripts?

A partial order on paired end read alignments

- Alignment x $\prec$ y when
  - x starts to the left of y in the reference
  - x and y overlap consistently
  - y is not contained in x

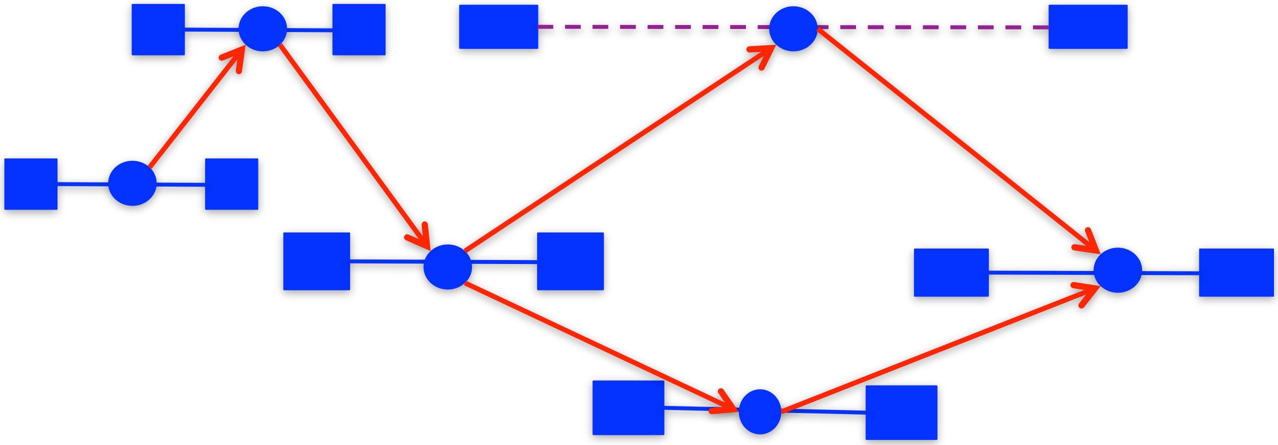


- That is, x $\prec$ y when they could have come from the same transcript

Dilworth’s theorem applied to the read partial order

- **Definition:** an *antichain* in the read partial order is a set of alignments with the property that no two are compatible (i.e. could arise from the same transcript).

- **Theorem [R.P. Dilworth, “A decomposition theorem for partially ordered sets”, Annals of Mathematics, 1950]:** The size of the largest antichain is equal to the minimum size of a chain partition.



Dilworth’s theorem applied to the read partial order

- **Definition:** an *antichain* in the read partial order is a set of alignments with the property that no two are compatible (i.e. could arise from the same transcript).

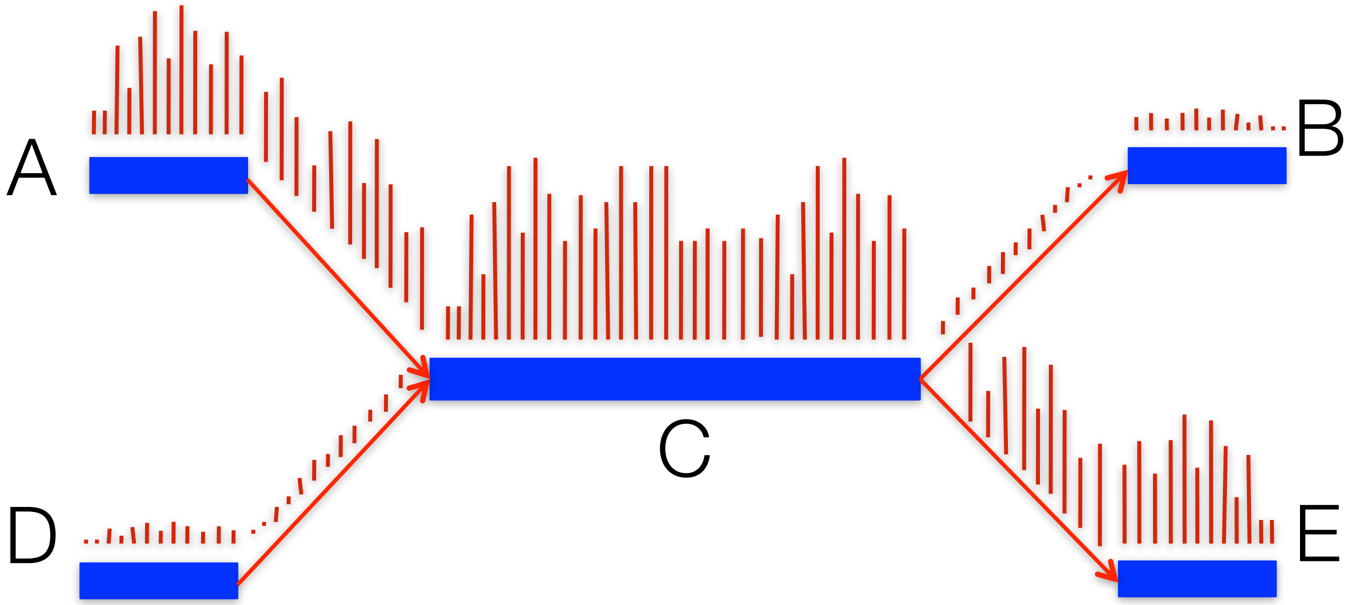
- **Theorem [R.P. Dilworth, “A decomposition theorem for partially ordered sets”, Annals of Mathematics, 1950]:** The size of the largest antichain is the minimum number of transcripts needed to explain the alignments.

- There is a constructive proof of the theorem, which reduces the problem to finding a maximum matching in a bipartite graph. The Hopcroft-Karp algorithm solves this problem in $O(\sqrt{V}E)$ time where we have *V*=*M*, the number of fragments sequenced.

- We rely instead on a maximum weighted matching algorithm; the best running time for weighted maximum matching is $O(V^2 logV + VE)$.

- This approach builds on ideas from N. Eriksson et al. (*PLoS Computational Biology* 2008) where a similar parsimony approach is used for viral population estimation.

Phasing splicing events using weighted matching



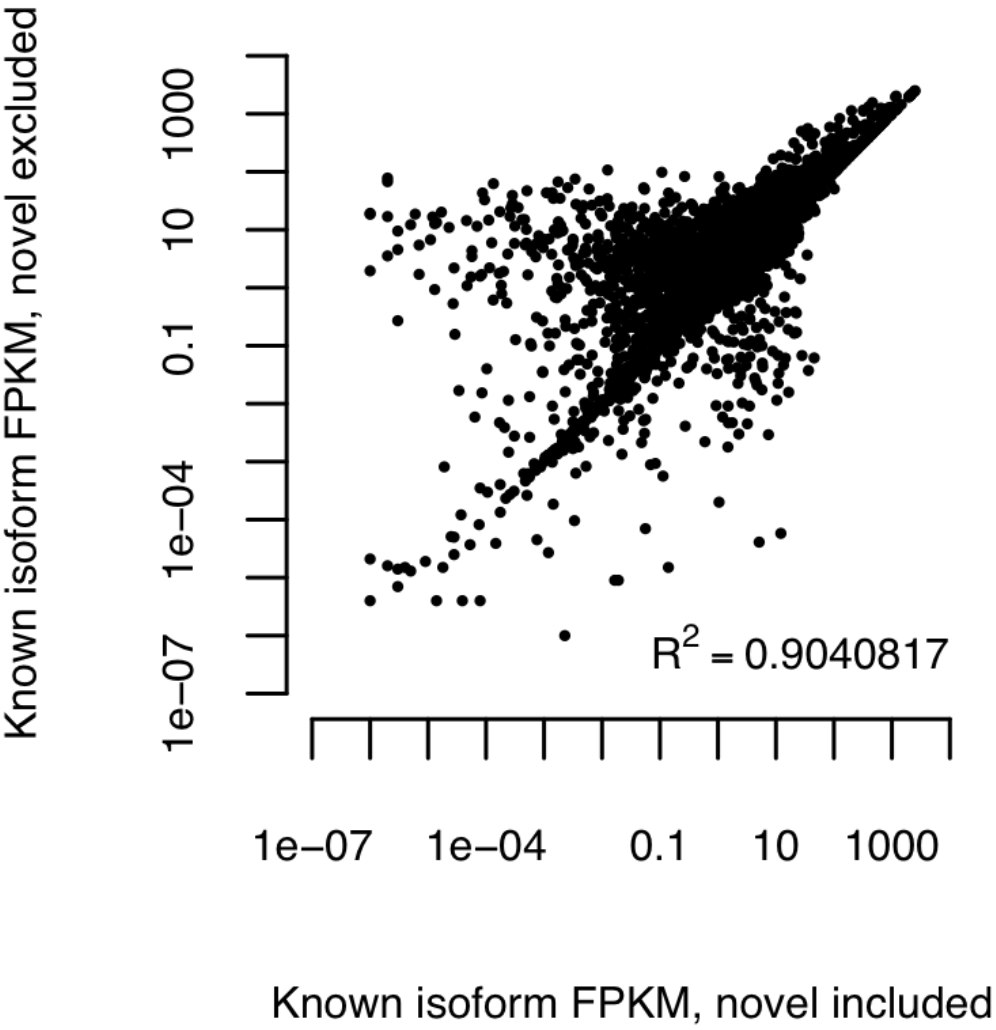
Properties of Cufflinks assemblies

- The assemblies are parsimonious- guarantee that the number of assembled transcripts is minimal.

- In the case of multiple minimal assemblies, likelihoods are compared in order to pick the best phasing.

- Identifiability of the resulting models is a corollary of Dilworth's theorem (the maximum antichain is a permutation submatrix of the read-transcript matrix, hence the latter is full rank).

Discovery is necessary for accurate abundance estimates

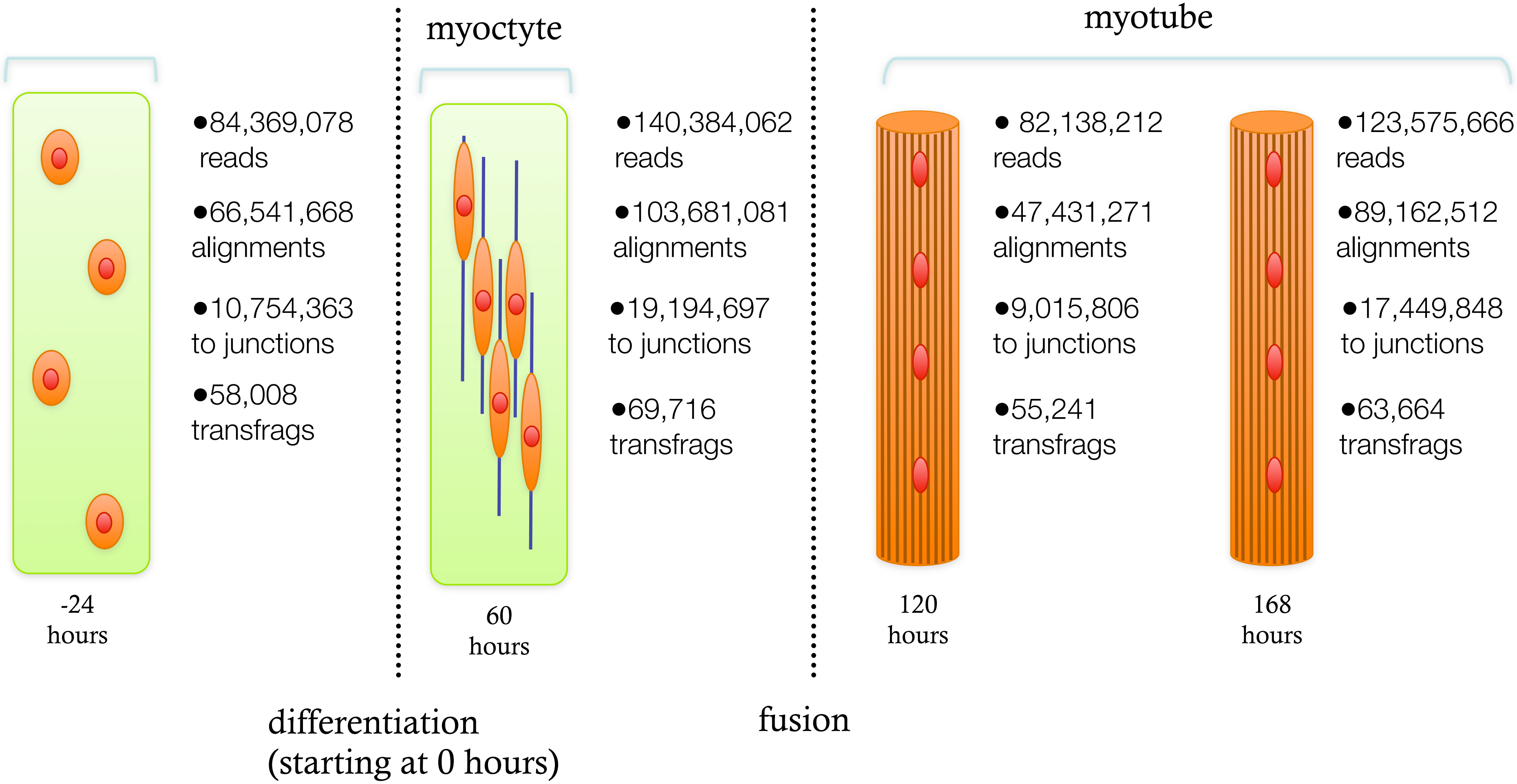


RNA-Seq time course analysis

- Measuring changes in relative abundances over time.
- Iosoform switching and generalizations.
- Inference of transcriptional versus post-transcriptional regulation.

The skeletal myogenesis transcriptome

RNA-Seq (2x75bp GAIIx) along time course of mouse C2C12 differentiation



myoctyte

myotube

- 84,369,078 reads
- 66,541,668 alignments
- 10,754,363 to junctions
- 58,008 transfrags

-24 hours

- 140,384,062 reads
- 103,681,081 alignments
- 19,194,697 to junctions
- 69,716 transfrags

60 hours

- 82,138,212 reads
- 47,431,271 alignments
- 9,015,806 to junctions
- 55,241 transfrags

120 hours

- 123,575,666 reads
- 89,162,512 alignments
- 17,449,848 to junctions
- 63,664 transfrags

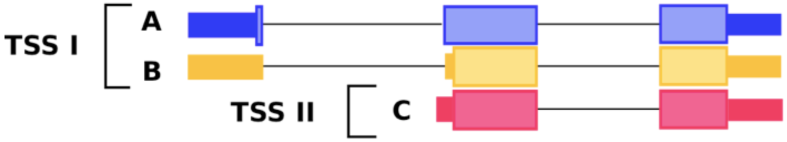
168 hours

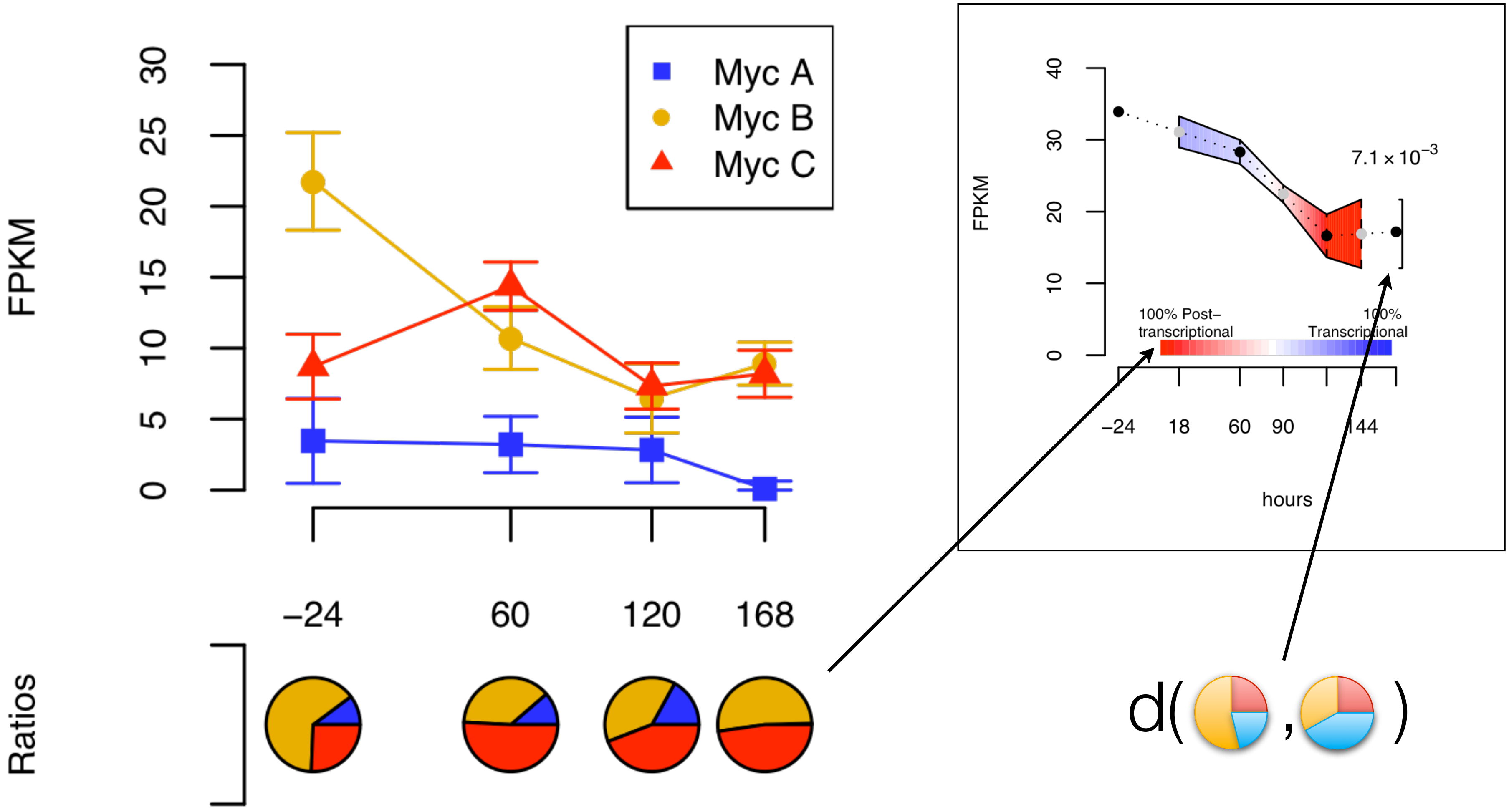
differentiation (starting at 0 hours)

fusion

Illustration based on: Ohtake et al, *J. Cell Sci., 2006; 119:3822-3832*

Dynamics of Myc expression


TSS I
A
B
TSS II
C


FPKM
Myc A
Myc B
Myc C
Ratios
FPKM
7.1 × 10^{-3}
100% Post–
transcriptional
Transcriptional
hours
d(,)