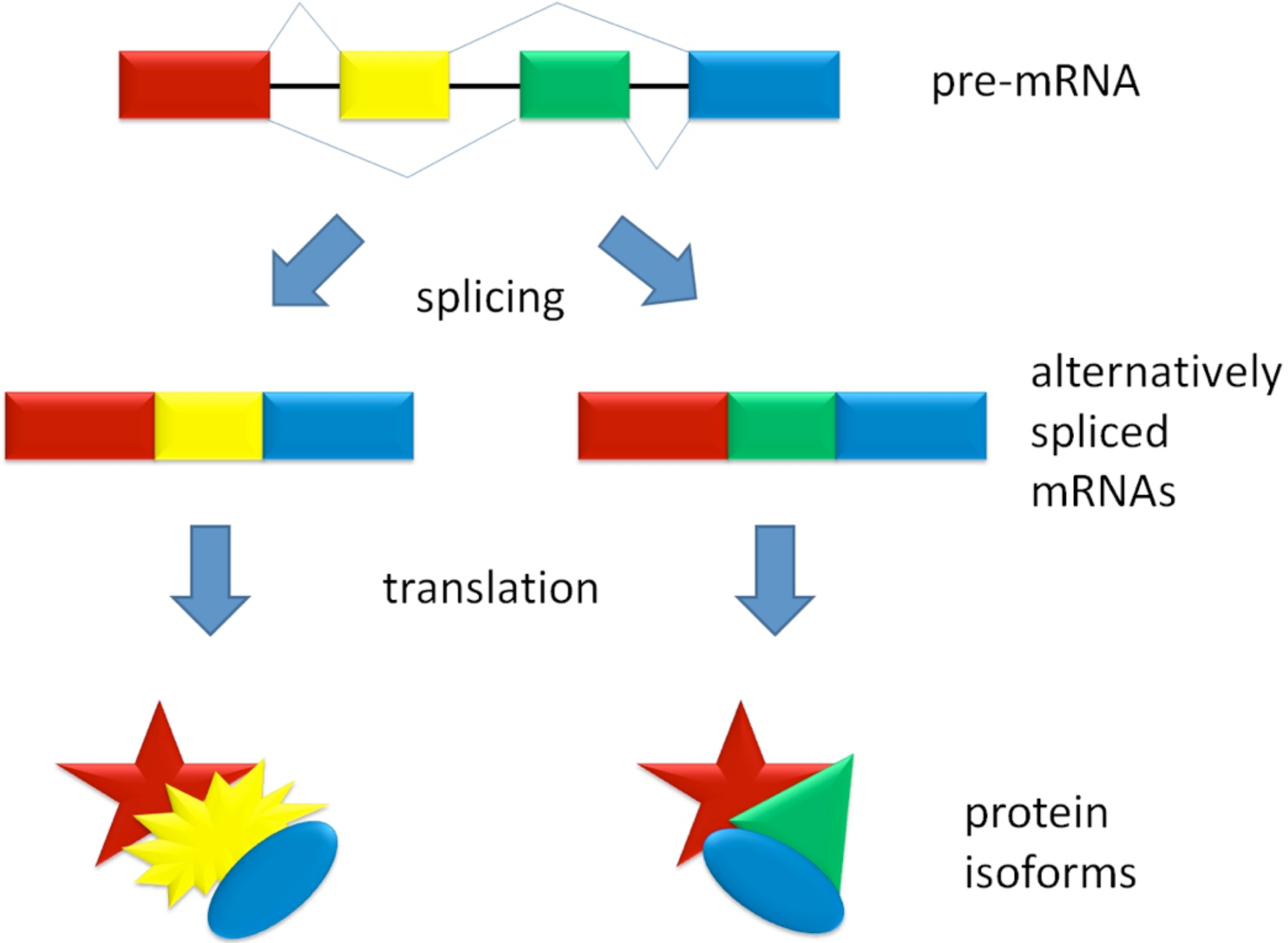
Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2019
Colin Dewey
cdewey@biostat.wisc.edu

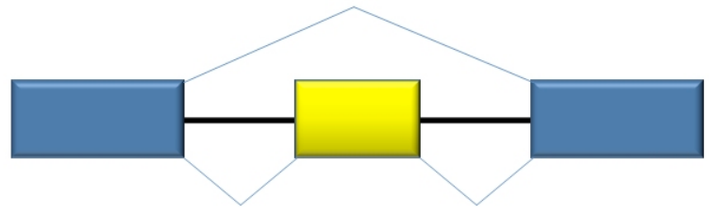
Overview

- Part I - Alternative splicing and the challenges it poses
- Part II - A solution: *Probabilistic Splice Graphs (PSGs)*
- Part III - Evaluating PSG methodology

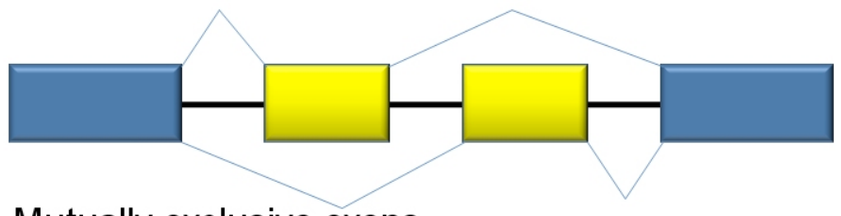
Alternative splicing



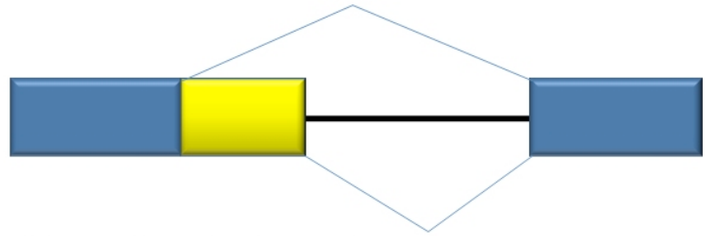
Classes of alternative splicing events



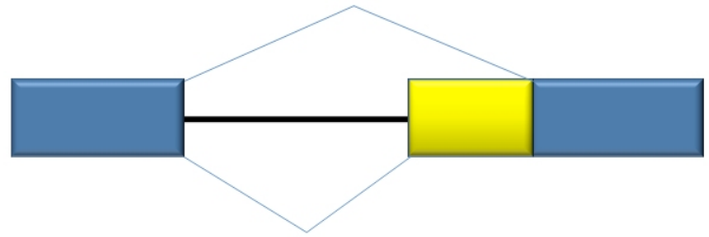
Exon skipping



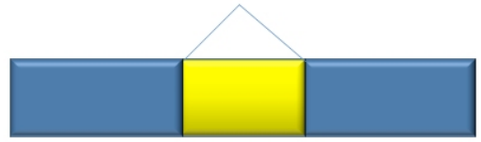
Mutually exclusive exons



Alternative 5' donor sites



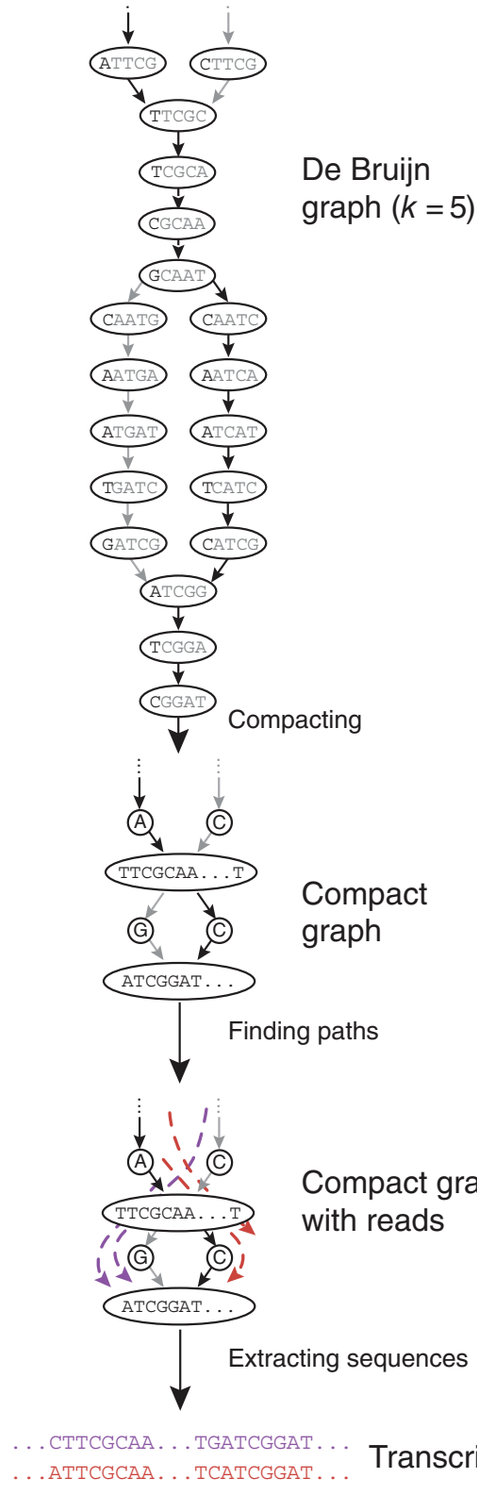
Alternative 3' acceptor sites



Intron retention

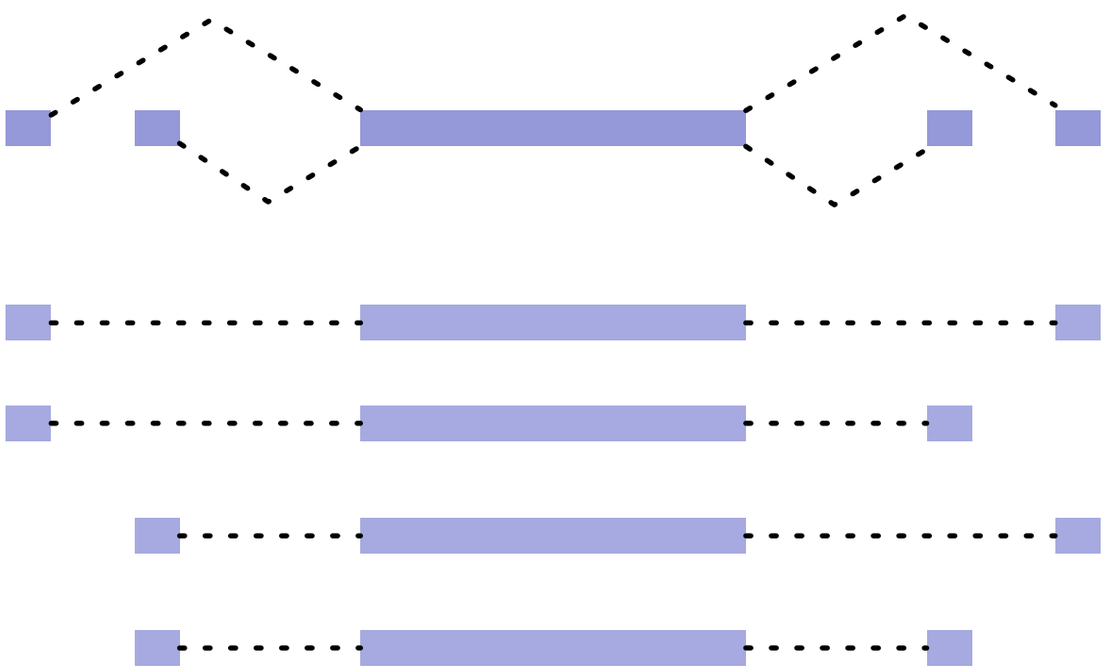
Complication 1: De novo transcriptome assembly

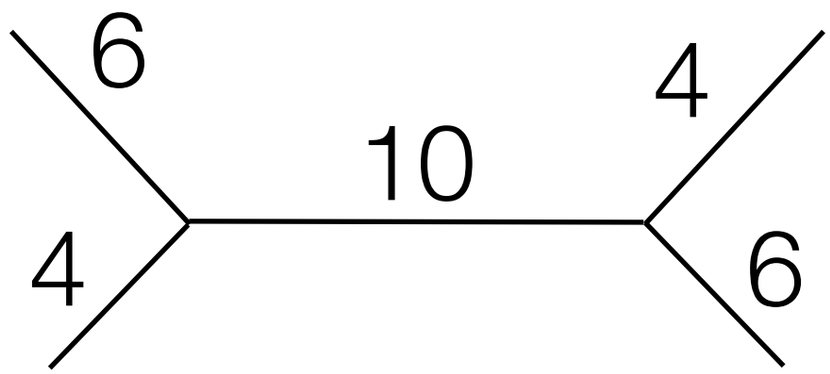
- RNA-Seq reads/fragments are relatively short
- Often insufficient to reconstruct full-length isoforms in the presence of alternative splicing
- Transcriptome assemblies perhaps best left in “graph” form
  - De Bruijn graph

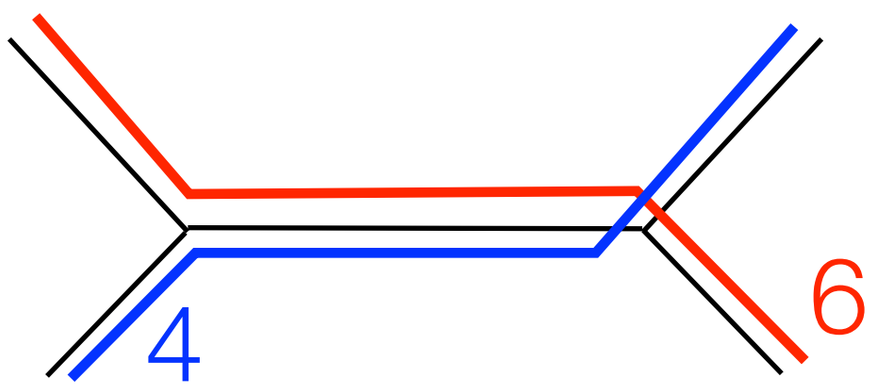


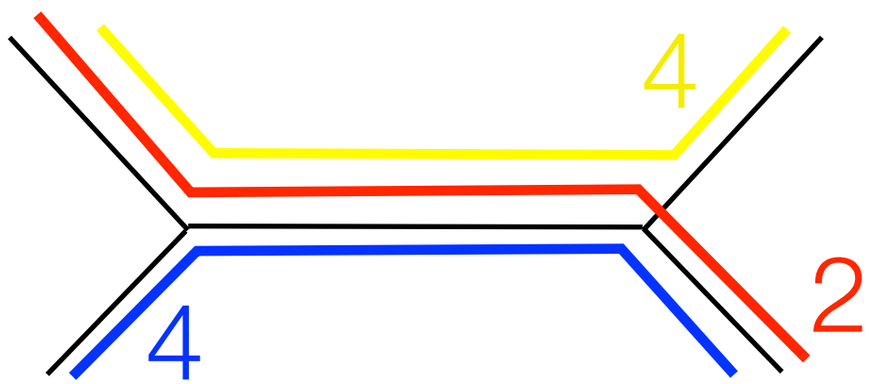
Graph constructed by the “Butterfly” module of Trinity (Grabherr et al. 2011)

Complication 2: Non-identifiability of full-length isoform models







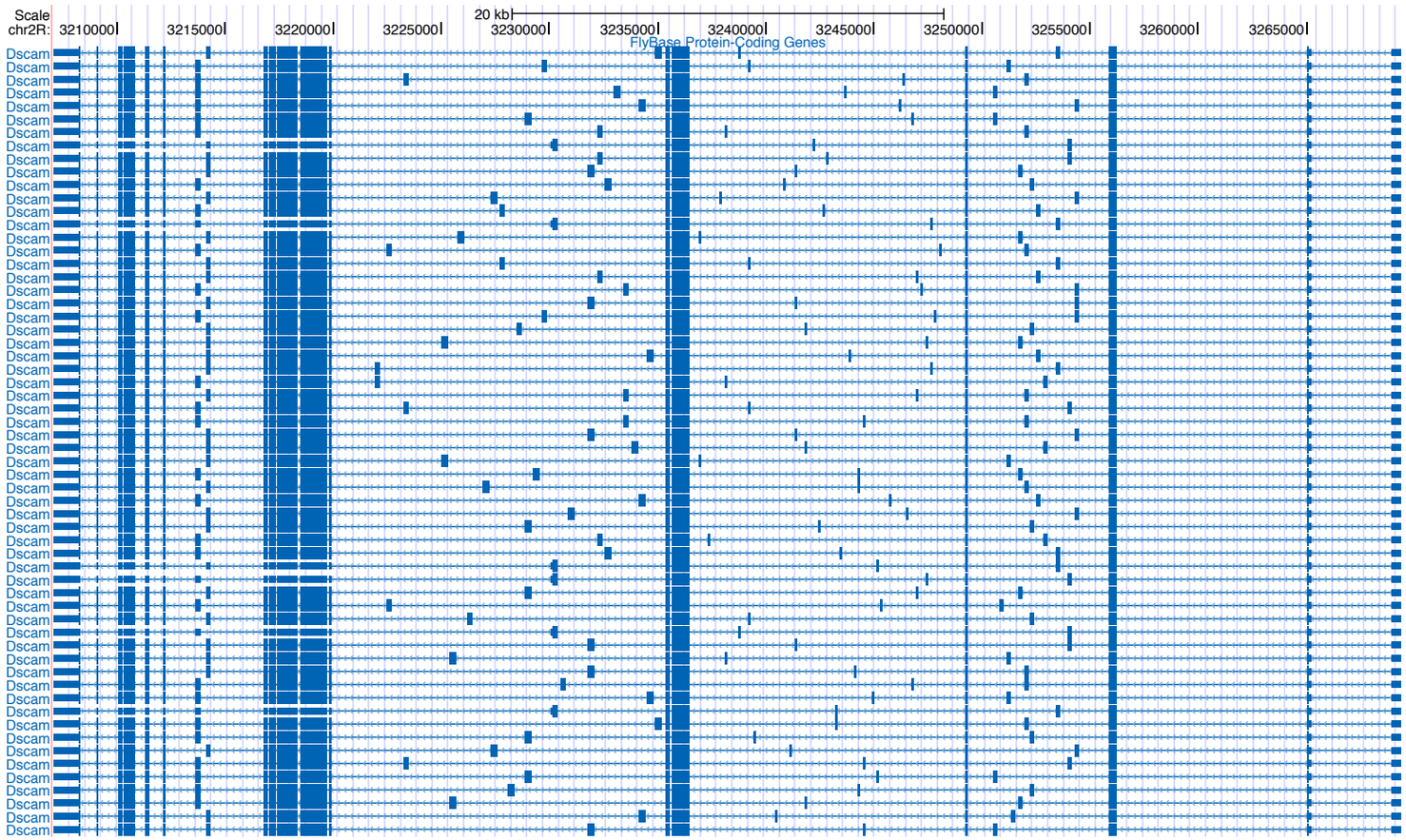


Lacroix et al. 2008; Hiller et al. 2009

Complication 3: Combinatorial explosion of distinct isoforms

- Combinatorial explosion of the number of possible isoforms for each gene

- Insufficient data to accurately estimate abundances of thousands of isoforms



Drosophila *Dscam*: more than 38,000 possible isoforms (Schmucker et al., 2000)

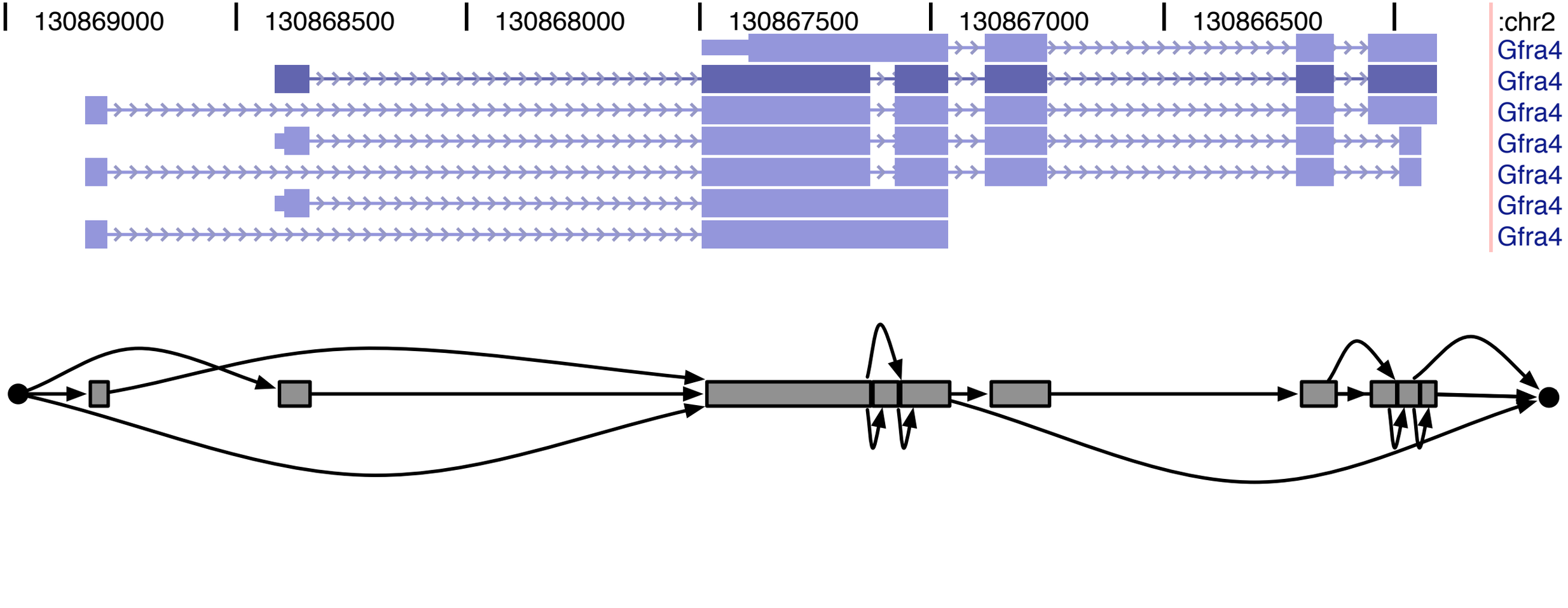
Overview

- Part I - Alternative splicing and the challenges it poses
- Part II - A solution: *Probabilistic Splice Graphs (PSGs)*
- Part III - Evaluating PSG methodology

Splice Graphs

- Heber et al. 2002

- Compact **data structure** for representing the possible isoforms of a gene



Splice Graphs with EST and RNA-Seq data

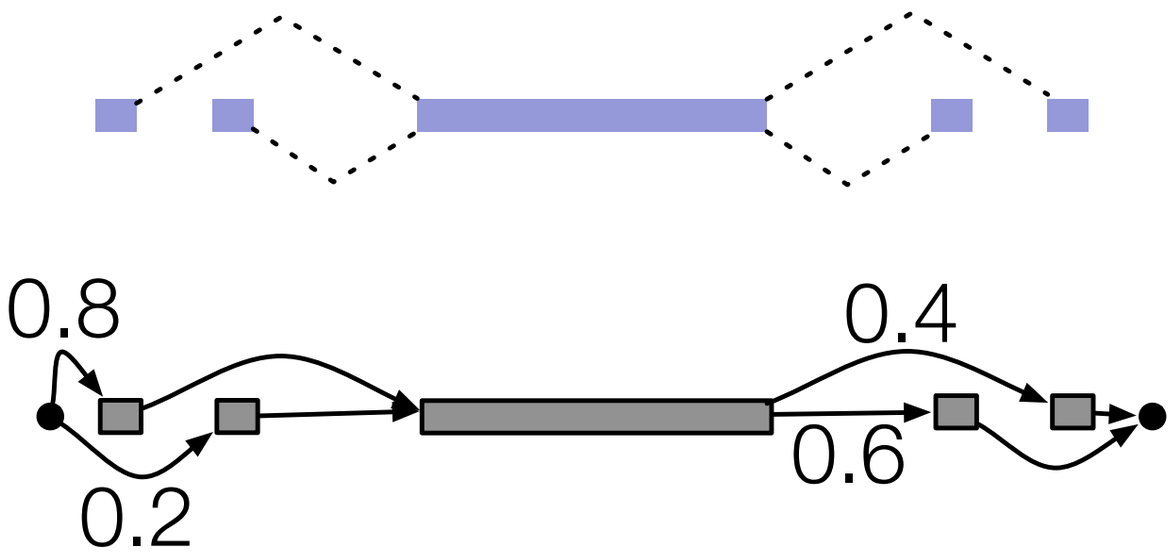
- Xing et al. 2006
    - EM algorithm for estimating abundances of all possible isoforms given splice graph and EST data
- Montgomery et al. 2010, Singh et al. 2011
    - Graph flow-based methods for quantification/differential splicing given RNA-Seq data
- Rogers et al. 2012
    - SpliceGrapher: construct splice graph structure given RNA-Seq data

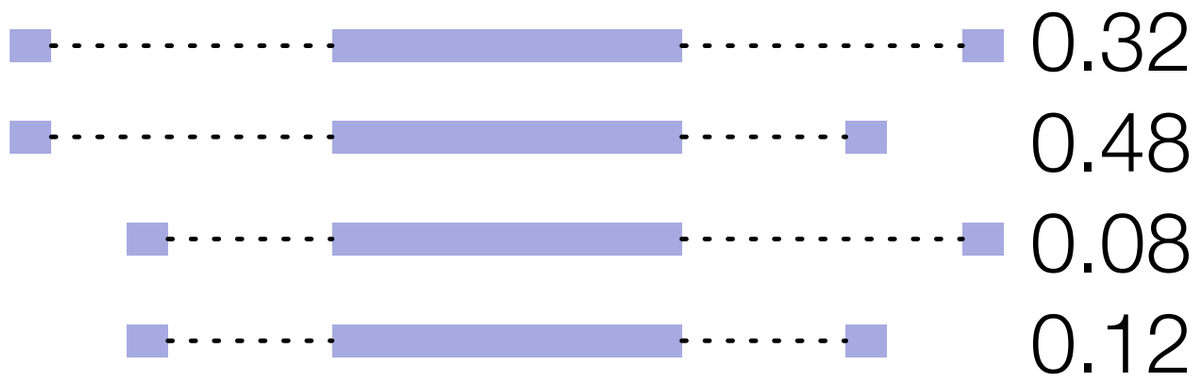
Probabilistic Splice Graphs

- Jenkins et al. 2006

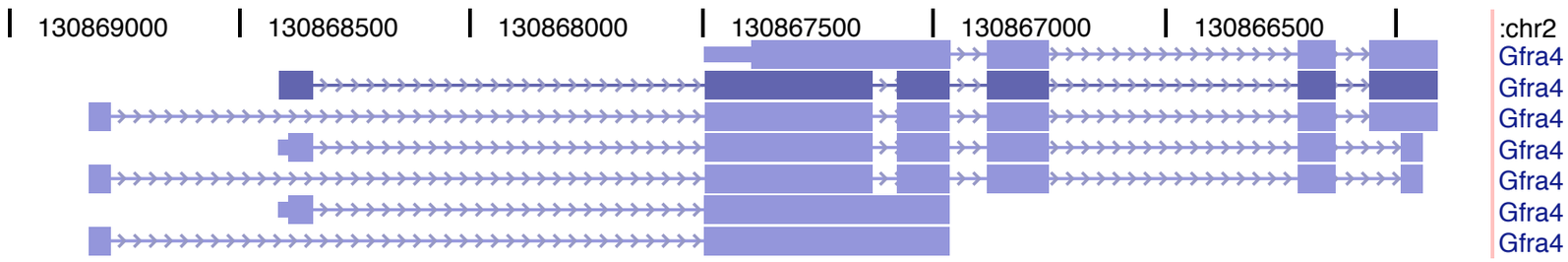
- Compact **probabilistic model** representing isoform frequencies in terms of frequencies of individual splice events

- Originally used by Jenkins et al. for EST analysis

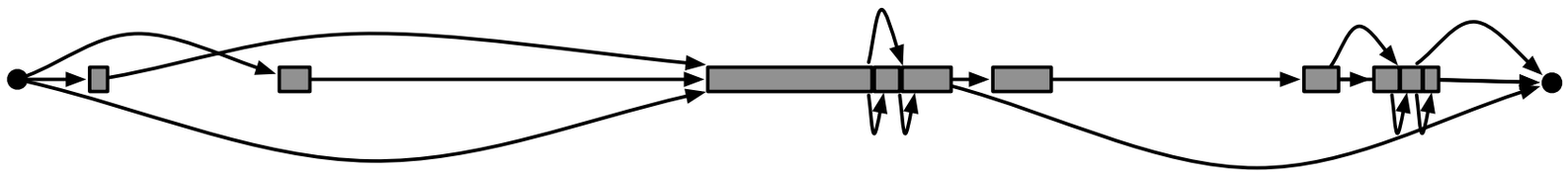




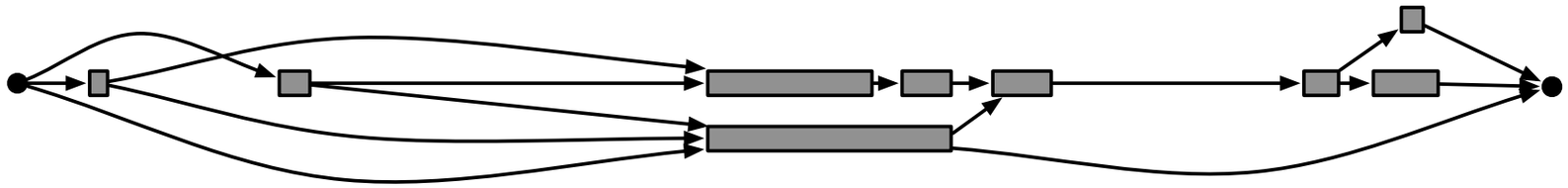
Probabilistic Splice Graph Complexity



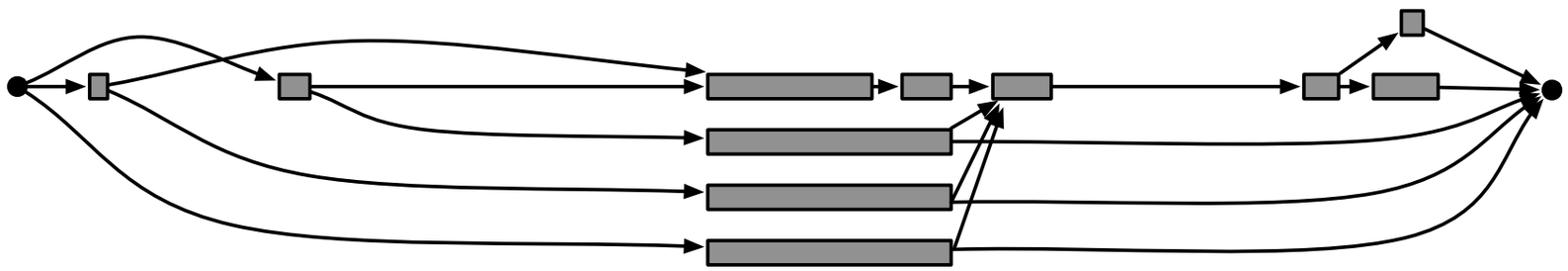
known isoforms



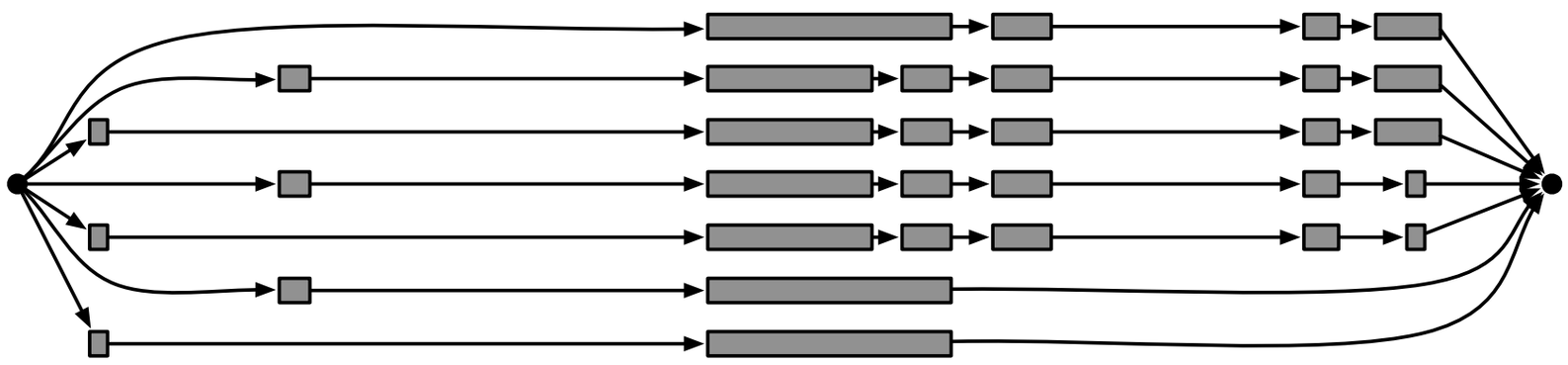
“line graph”



“exon graph”



“higher-order exon graph”



“unfactorized graph”

Advantages of PSGs

- Compact description of the possible isoforms of a gene
    - Models the frequencies of potentially exponentially many isoforms with a polynomial number of parameters
        - Models dependence or independence of splice events
- The parameters of a PSG are more often identifiable than a model that has a parameter for every possible isoform
- Splice graphs are naturally-produced structures from transcriptome assemblers

PSGs are alternative “parsimonious” models

- Other methods find smallest set of isoform structures that explain the data
  - Cufflinks (Trapnell et al., 2010)
  - IsoLasso (Li et al., 2011)
  - NSMAP (Xia et al., 2011)
  - SLIDE (Li et al., 2011)
- PSG models are another form of parsimonious model
  - Minimize the number of splice event parameters
  - Assumption of independence between splice events

Our contributions

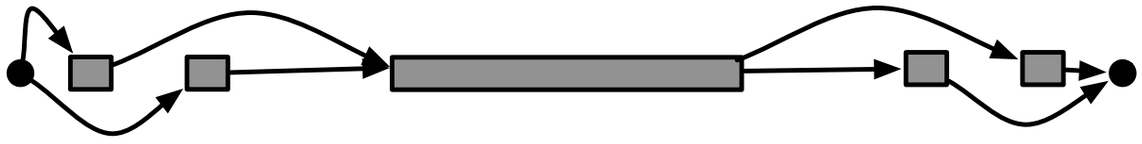
- Application of PSGs to RNA-Seq data
    - Combined model of PSG with RNA-Seq generative model
    - Efficient PSG parameter estimation with EM and dynamic programming
- Identifiability proofs for PSG with RNA-Seq data
- Differential processing (splicing) tests

L. Legault and C. Dewey. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. *Bioinformatics* 29(18):2300-2310.

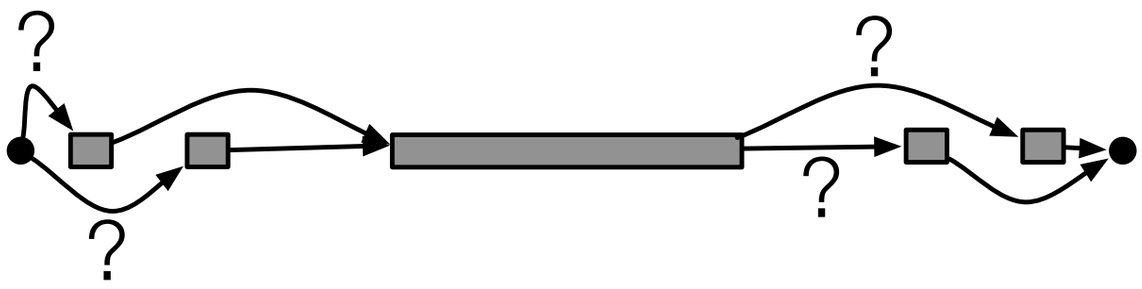
The PSG parameter inference task

- Given: RNA-Seq reads and a PSG structure

```
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
```



- Do: Estimate the (ML or MAP) parameters for the model



A model of RNA-Seq from PSGs

- RSEM model extended to probabilistic splice graphs
    - fragment length distribution, quality scores, read mapping ambiguity
- Dynamic programming algorithms → polynomial time inference for genes with an exponential number of isoforms

Probability of including vertex j given that vertex i was in transcript

$$f(i,j) = \sum_{s:s_1=i,s_{|s|}=j} w(s) = \begin{cases} 1 & i = j \\ \sum_k \alpha_{kj} f(i,k) & i \neq j \end{cases}$$

Expected prefix length

$$d_p(i) = \ell_i + \frac{1}{f(0,i)} \sum_j f(0,j)\alpha_{ji} d_p(j)$$

Expected suffix length

$$d_q(i) = \ell_i + \sum_j \alpha_{ij} d_q(j)$$

EM for PSG parameter estimation

- E-step: compute the expectation of the number of times edge (i,j) is used

$$E[Z_{nij}] = \frac{\sum_{(b,s)\in\pi(r)} g(s,i,j)}{\sum_{(b,s)\in\pi(r)} g(s)}$$

$$g(s) = f(0,s_1)w(s)$$

$$g(s,i,j) = \begin{cases} f(0,s_1)w(s) & (i,j)\in s \\ f(0,i)\alpha_{ij}f(j,s_1)w(s) & \text{if } \exists \text{ path from } v_j \text{ to } s_1 \\ f(0,s_1)w(s)f(s_{|s|},i)\alpha_{ij} & \text{if } \exists \text{ path from } s_{|s|} \text{ to } v_i \\ 0 & \text{otherwise} \end{cases}$$

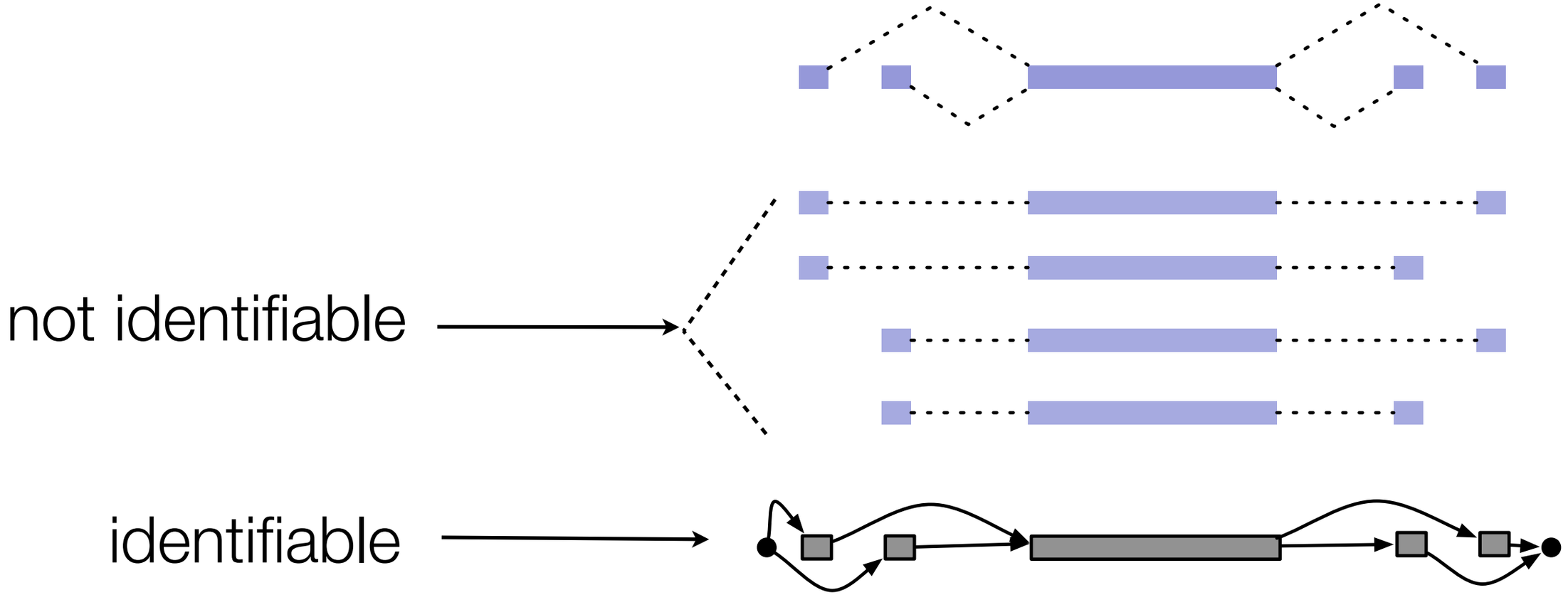
- M-step: maximize the completely-observed likelihood given the edge counts

$$\alpha_{ij} = \frac{\frac{z_{ij}}{(d_p(i)+d_q(j))}}{\sum_k \frac{z_{ik}}{(d_p(i)+d_q(k))}}$$

Identifiability of PSGs with RNA-Seq data

- Identifiability: $P(D|M,\theta) = P(D|M,\theta'), \forall D \Leftrightarrow \theta = \theta'$

- *Proposition: If for all edges (u, v), there exists a read that is uniquely derived from that edge, or v has indegree 1 and there exists a read that is uniquely derived from v, then the PSG is identifiable.*



The differential processing (DP) task

- Given: RNA-Seq reads from two conditions and a PSG structure

condition 1

```
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
```

condition 2

```
CATATCGTCGTAGCTAGTACG
CCACACTAGGCTACGTGCGCA
TCGACGCTACCGGCATCGCGC
ACTAGTACGTACGTAGTAGCT
GGATGCTCAGATGGCTATCGG
CGCATTACGGAAGCTCATCGA
AACCATCGGAAGGCCGTTTAA
CAGCTAGGCGCTAGGCGCTTT
CATGCTAGCGCGATCGCGTAG
GCATCGACTCGCGACCGATCC
ACGCATCGACTCGCGCATCGC
```

- Do: Determine if the processing frequencies are different

$$\alpha_1 = \alpha'_1 \text{ and } \alpha_2 = \alpha'_2 \; ? \qquad \alpha_1 = \alpha'_1 \text{ or } \alpha_2 = \alpha'_2 \; ?$$

Our approach to the differential processing (DP) task

- Simple likelihood ratio tests with PSG model

- Test for null hypothesis that all frequencies are the same

$$\text{LR} = \frac{P(R^1|\hat{\alpha}^1)P(R^2|\hat{\alpha}^2)}{P(R^1 \cup R^2|\hat{\alpha}^{12})}$$

- Test for null hypothesis that frequencies of edges out of one vertex (*i)* are the same

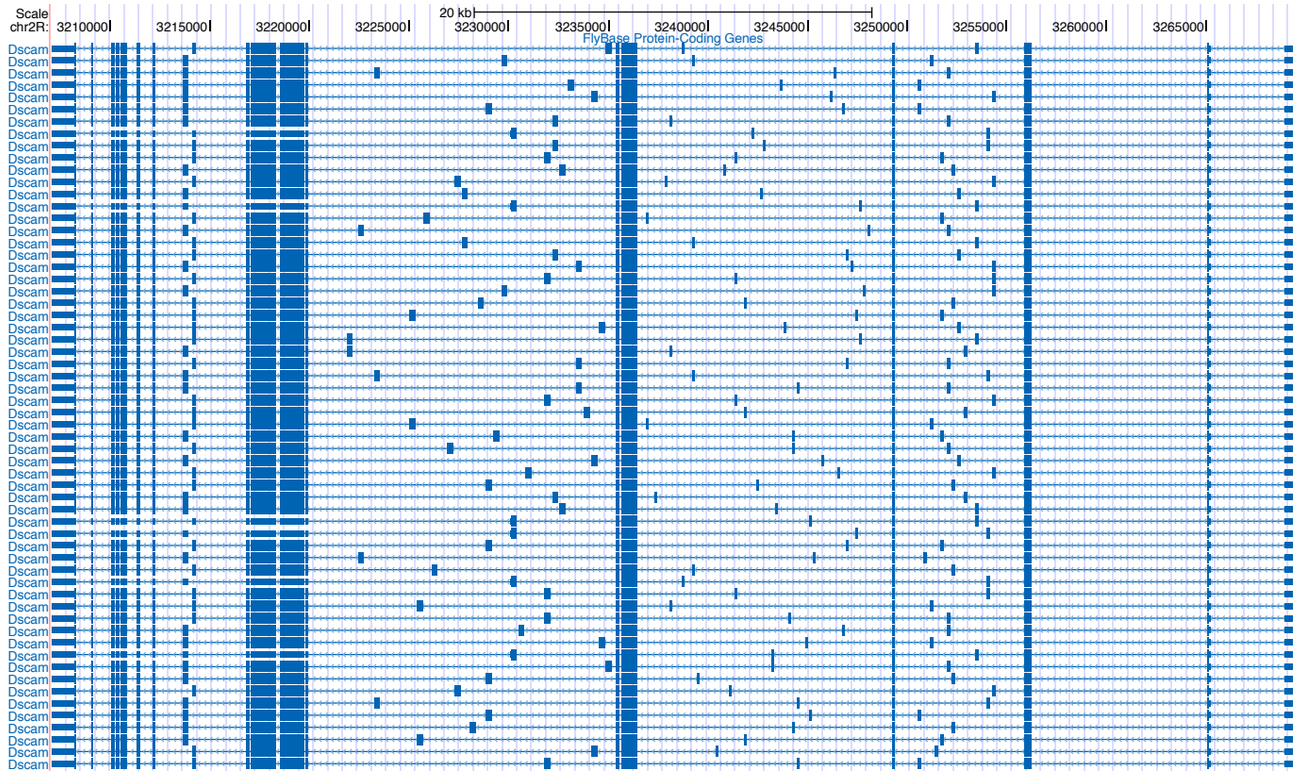
$$\text{LR} = \frac{P(R^1|\hat{\alpha}^1)P(R^2|\hat{\alpha}^2)}{P(R^1, R^2|\hat{\alpha}^1_{\setminus i}, \hat{\alpha}^2_{\setminus i}, \hat{\alpha}^{12}_i))}$$

Overview

- Part I - The problem
- Part II - A solution: *Probabilistic Splice Graphs (PSGs)*
- Part III - Evaluating PSG methodology

Efficient inference for highly-spliced genes

- DSCAM running time test
  - 23,976 isoforms
  - 184 read pairs from a modENCODE sample

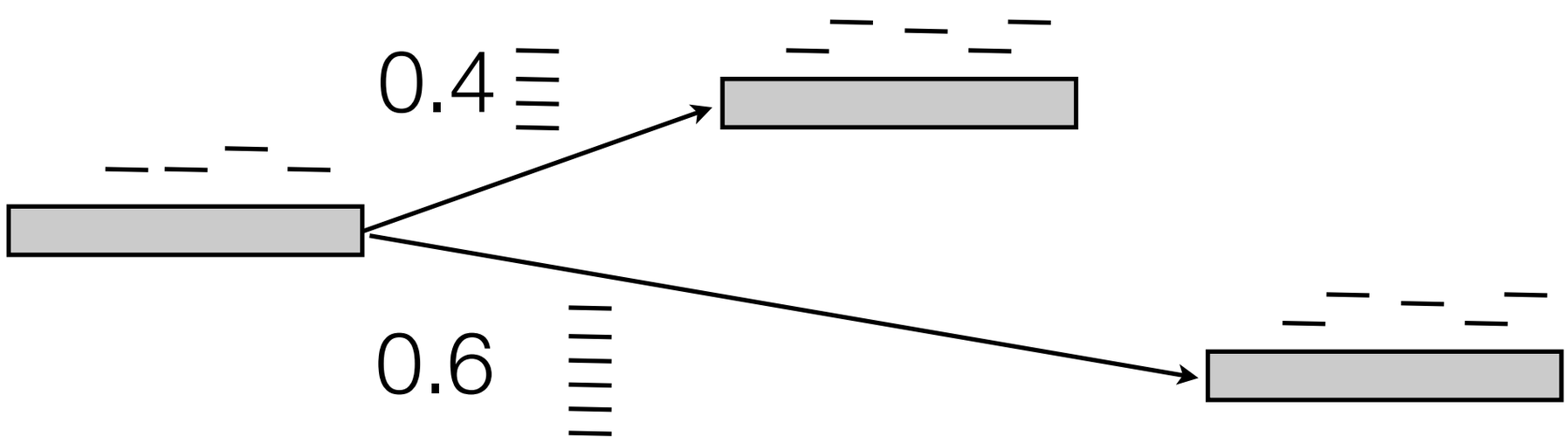


Method	RSEM	Cufflinks	PSG EM
Running time	Not possible	> 6 hours (> 90 GB RAM)	< 3 seconds

A simple method for comparison

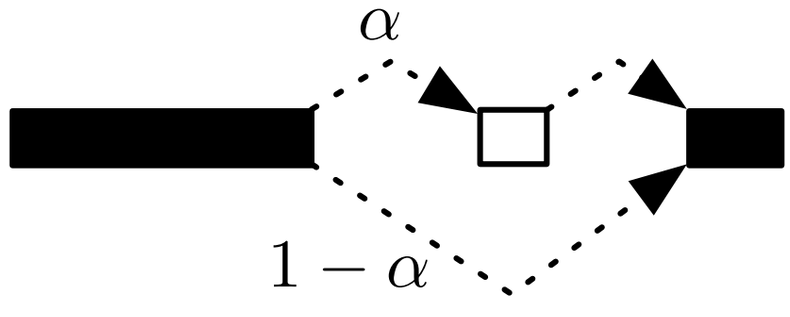
- The **Junction-Read** (JR) method

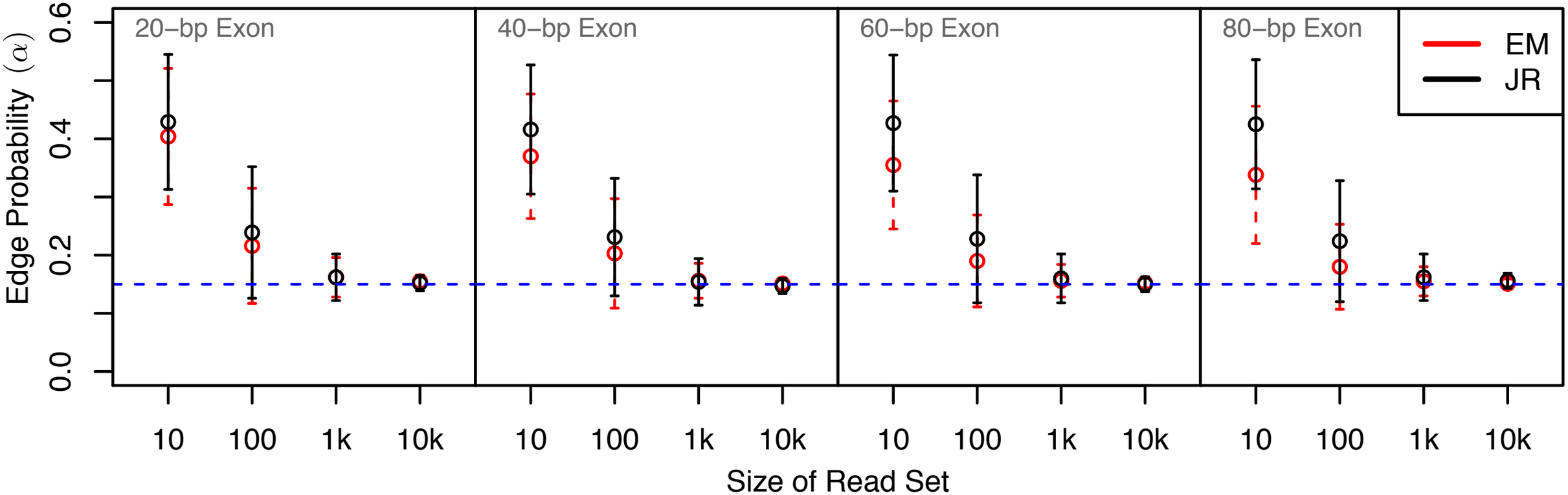
- Keep only reads that align to the splice junctions (edges in the PSG)



- Throws away data, but is very robust to model assumption violations

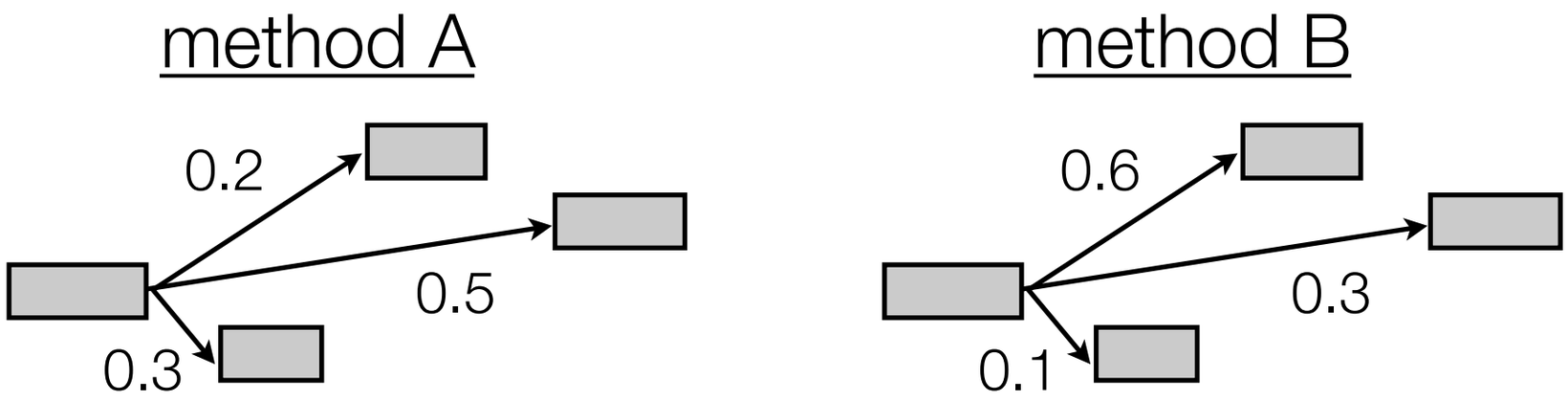
Convergence with simulated data





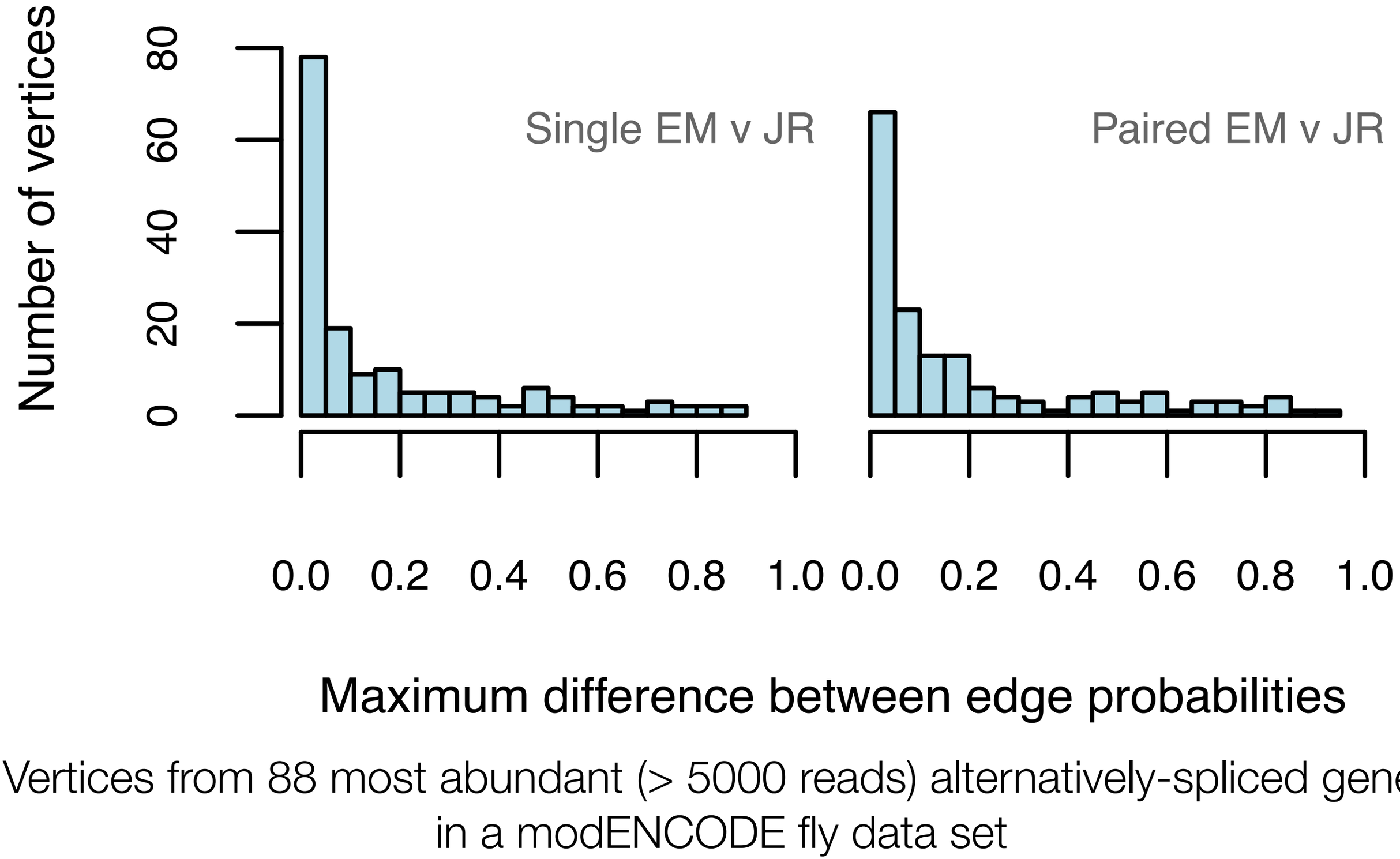
Comparisons on real data

- Require notion of “distance” between estimates from different methods
- Our distance measure:
  - per vertex
  - maximum difference between probability estimates on out-edges of vertex (L-∞ norm)



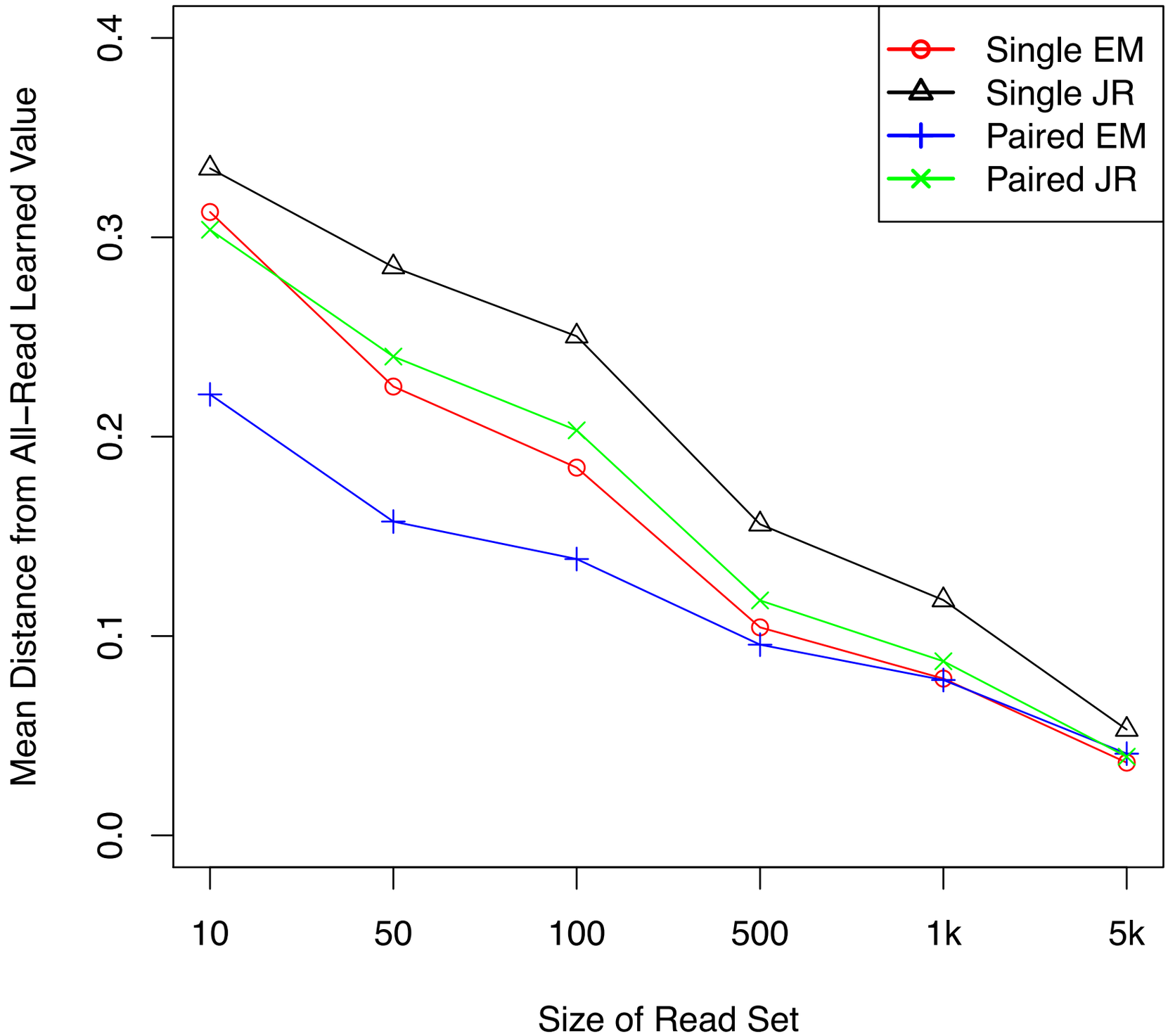
$$distance_v(A, B) = max(|0.6 - 0.2|, |0.5 - 0.3|, |0.3 - 0.1|) = 0.4$$

How close are the estimates from JR and EM on real data?

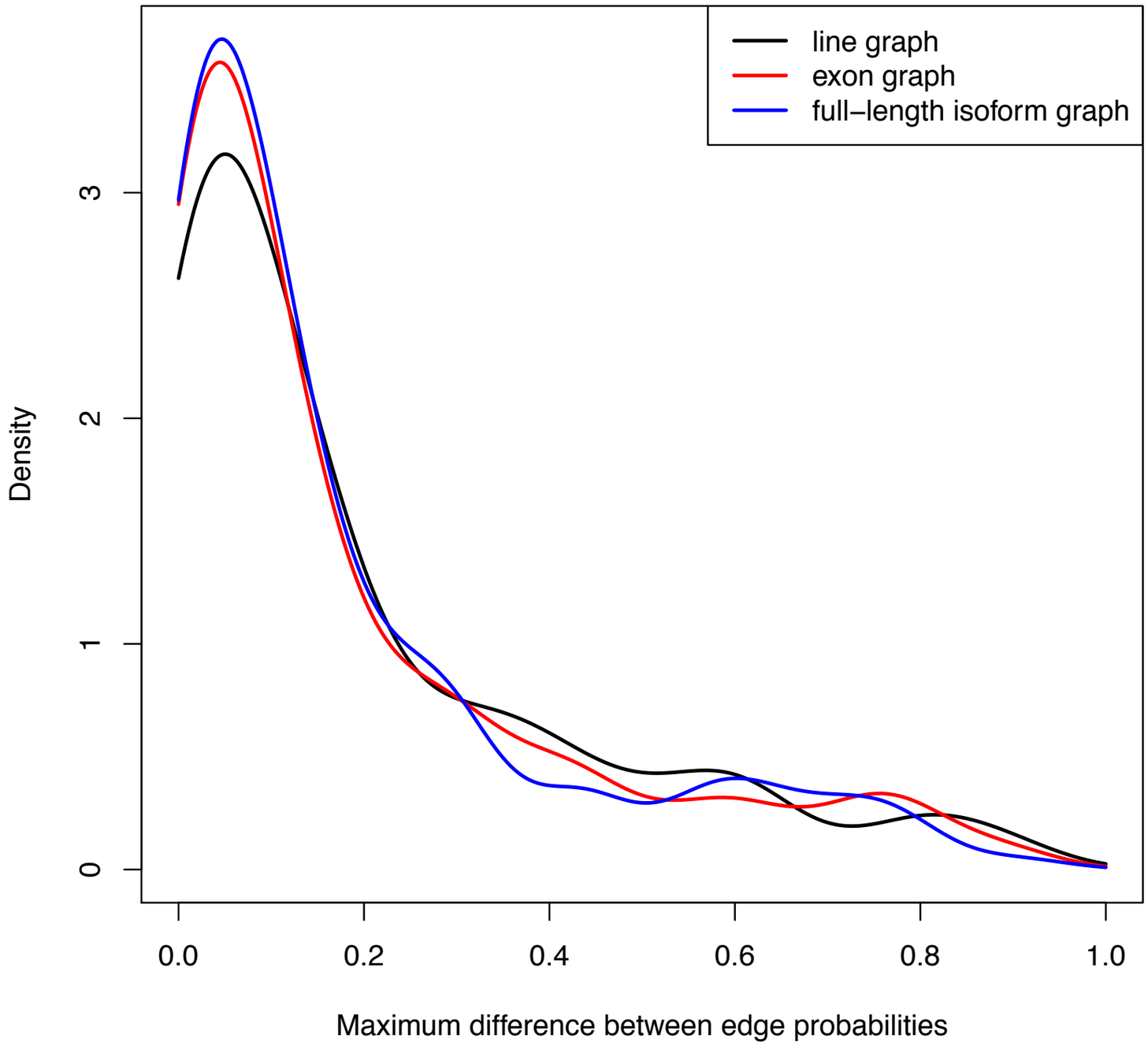


Vertices from 88 most abundant (> 5000 reads) alternatively-spliced genes in a modENCODE fly data set

Convergence of estimates on real data



Comparing PSGs of different complexity



- Same set of fly data
- Estimated with three classes of PSG: line, exon, full-length
- Compared estimates to those from JR (gold-standard)
- No statistically-significant difference between exon and full-length graph estimates

Summary of Junction-Read comparison results

- Estimates using PSG models are generally close to those from the simplistic JR-method
    - ⇒PSG model assumptions appear to be reasonable
- PSG estimates converge more quickly as the data set increases in size
    - ⇒Our EM estimation procedure uses information from all reads, not just those that span splice junctions
- Exon-graph estimates as good as those using traditional full-length isoform models
    - ⇒Independence assumptions of exon graphs appear to be reasonable

Differential processing detection

DP Accuracy on real data

		# of DP genes		
Sample 1	Sample 2	PSG	FDM	Cuffdiff
CEU Rep 1	**CEU Rep 2**	**0**	**0**	**1187**
CEU Rep 1	Yoruban Rep 1	39	24	269
CEU Rep 1	Yoruban Rep 2	46	24	282
CEU Rep 2	Yoruban Rep 1	45	22	253
CEU Rep 2	Yoruban Rep 2	38	29	260
Yoruban Rep 1	**Yoruban Rep 2**	**0**	**0**	**1253**
CME_W1_Cl.8+ Rep 1	**CME_W1_Cl.8+ Rep 2**	**16**	**32**	**204**
CME_W1_Cl.8+ Rep 1	Kc167	365	207	7
CME_W1_Cl.8+ Rep 1	ML-DmBG3-c2	232	164	6
CME_W1_Cl.8+ Rep 1	S2-DRSC	406	228	12
CME_W1_Cl.8+ Rep 2	Kc167	319	211	16
CME_W1_Cl.8+ Rep 2	ML-DmBG3-c2	260	126	16
CME_W1_Cl.8+ Rep 2	S2-DRSC	353	220	17
Kc167	ML-DmBG3-c2	384	321	12
Kc167	S2-DRSC	419	209	12
ML-DmBG3-c2	S2-DRSC	431	287	4
HUVEC Rep 1	**HUVEC Rep 2**	**35**	**43**	**440**
HUVEC Rep 1	K562 Rep 1	376	344	8
HUVEC Rep 1	K562 Rep 2	379	302	12
HUVEC Rep 2	K562 Rep 1	442	382	8
HUVEC Rep 2	K562 Rep 2	355	285	10
K562 Rep 1	**K562 Rep 2**	**224**	**308**	**168**

Differential processing detection

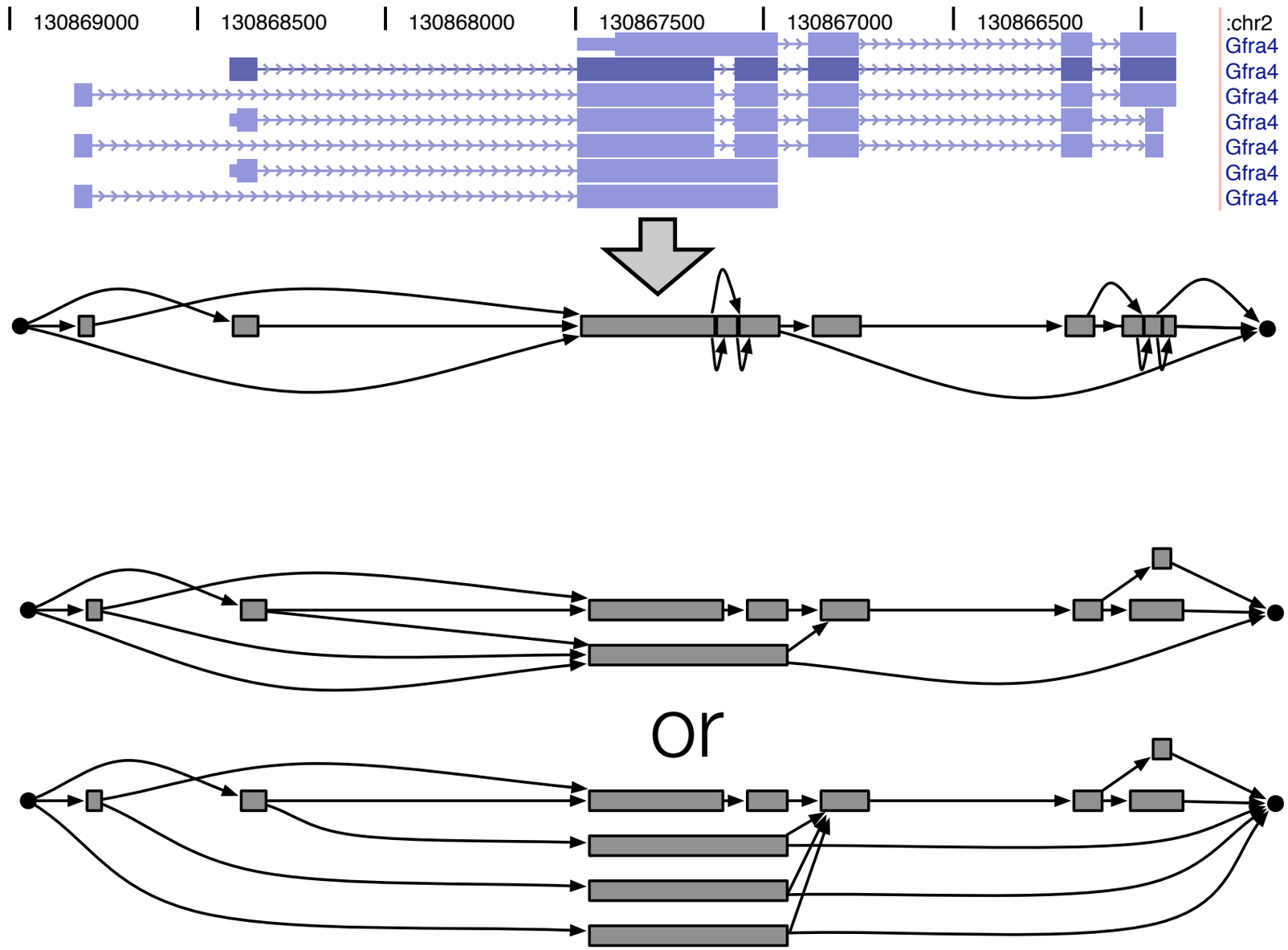
DP accuracy on simulated data

Method	Sample 1	Sample 2	Predicted DP	Recall	Precision
PSG	**A Rep 1**	**A Rep 2**	**4**		
	A Rep 1	B Rep 1	257	0.60	0.95
	A Rep 1	B Rep 2	230	0.54	0.95
	A Rep 2	B Rep 1	251	0.59	0.94
	A Rep 2	B Rep 2	235	0.54	0.93
	B Rep 1	**B Rep 2**	**0**		
Cuffdiff	**A Rep 1**	**A Rep 2**	**379**		
	A Rep 1	B Rep 1	49	0.11	0.92
	A Rep 1	B Rep 2	58	0.13	0.88
	A Rep 2	B Rep 1	48	0.12	0.98
	A Rep 2	B Rep 2	51	0.11	0.88
	B Rep 1	**B Rep 2**	**148**		
FDM	**A Rep 1**	**A Rep 2**	**11**		
	A Rep 1	B Rep 1	311	0.39	0.51
	A Rep 1	B Rep 2	255	0.28	0.44
	A Rep 2	B Rep 1	320	0.37	0.47
	A Rep 2	B Rep 2	242	0.24	0.40
	B Rep 1	**B Rep 2**	**148**		

Simulations based on two ENCODE cell lines, 10% of genes selected to be DP

Next steps for modeling RNA-Seq with PSGs

- Graph construction
  - Exon discovery
  - Splice junction discovery
- Model selection
  - Learning dependencies between splice events



Summary

- Alternative splicing is a significant complication in RNA-Seq analysis
- Probabilistic Splice Graphs enable identifiable models for alternatively spliced genes with efficient inference algorithms
- Differential processing (splicing) tests with PSG models look promising