Transcript quantification and Analysis of alternative splicing with RNA-Seq

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2020

Daifeng Wang

daifeng.wang@wisc.edu

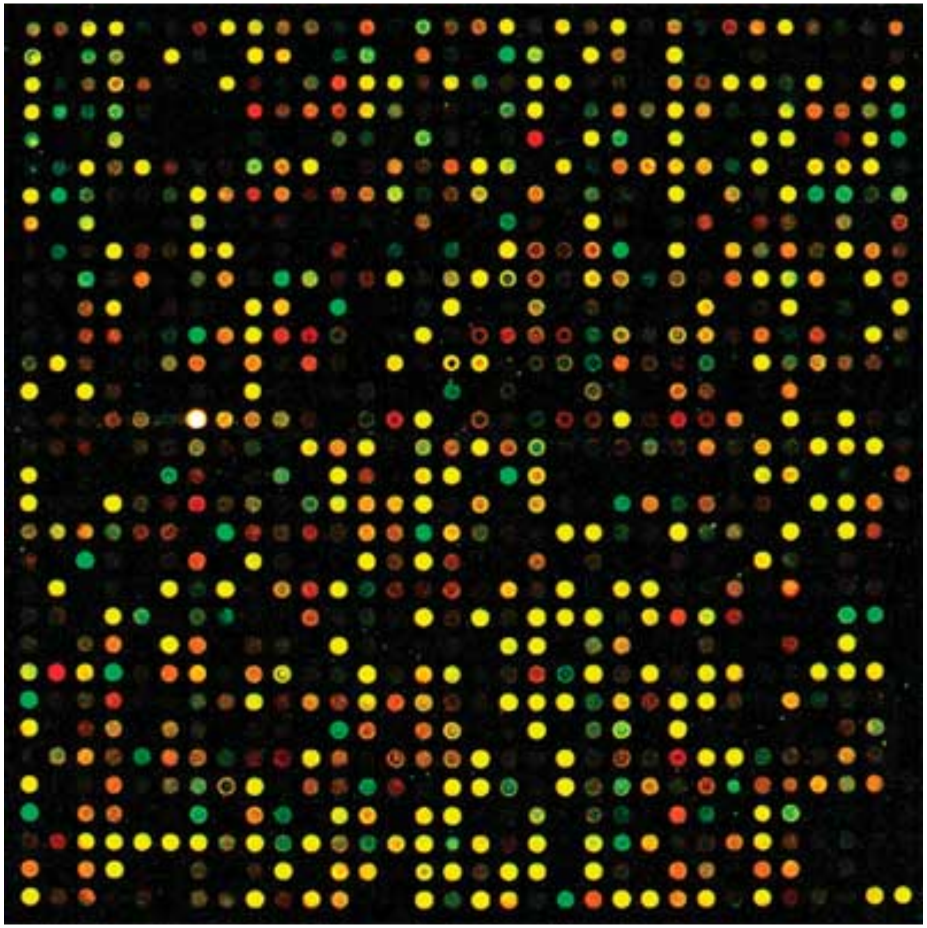
Overview

- RNA-Seq technology
- The RNA-Seq quantification problem
- Generative probabilistic models and Expectation-Maximization for the quantification task
- Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs

Goals for lecture

- What is RNA-Seq?
- How is RNA-Seq used to measure the abundances of RNAs within cells?
- What probabilistic models and algorithms are used for analyzing RNA-Seq?

Measuring transcription the old way: microarrays



- Each spot has “probes” for a certain gene
- Probe: a DNA sequence complementary to a certain gene
- Relies on complementary hybridization
- Intensity/color of light from each spot is measurement of the number of transcripts for a certain gene in a sample
- Requires knowledge of gene sequences

Advantages of RNA-Seq over microarrays

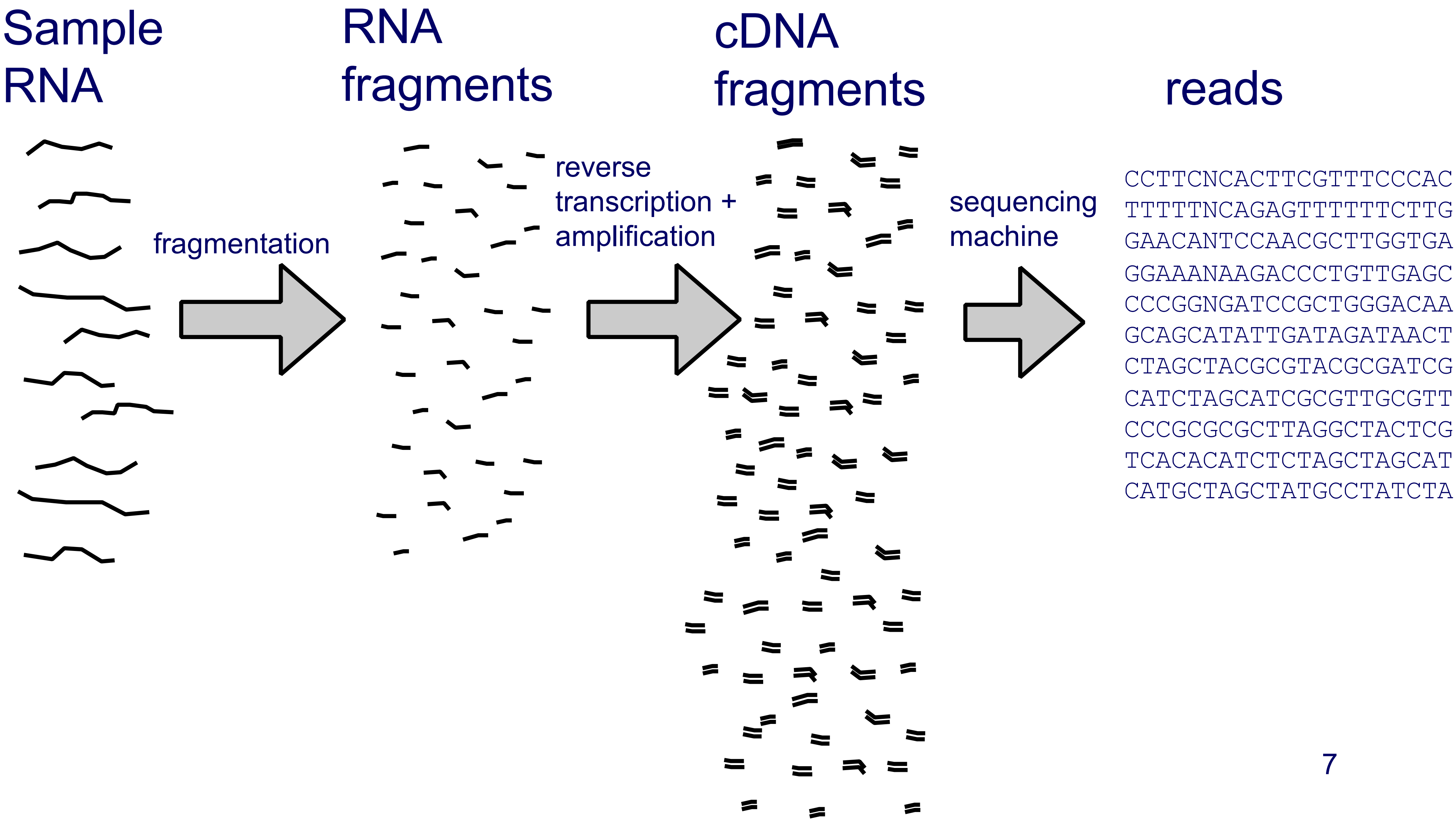
- No reference sequence needed
  - With microarrays, limited to the probes on the chip
- Low background noise
- Large dynamic range
  - 10^5 compared to 10^2 for microarrays
- High technical reproducibility
- Identify novel transcripts and splicing events

RNA-Seq technology

- Leverages rapidly advancing sequencing technology
- Transcriptome analog to whole genome shotgun sequencing
- Two key differences from genome sequencing:

1. Transcripts sequenced at different levels of coverage - expression levels
2. Sequences already known (in many cases) - coverage is measurement

A generic RNA-Seq protocol



Sample RNA

fragmentation

RNA fragments

reverse transcription + amplification

cDNA fragments

sequencing machine

reads

```
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA
```

RNA-Seq data: FASTQ format

```
@HWUSI-EAS1789_0001:3:2:1708:1305#0/1
CCTTCNCACTTCGTTTCCCACTTAGCGATAATTTG
+HWUSI-EAS1789_0001:3:2:1708:1305#0/1
VVULVBVYVYZZXZZ\ee[a^b`[a\a[\\a^^^\
@HWUSI-EAS1789_0001:3:2:2062:1304#0/1
TTTTTNCAGAGTTTTTTCTTGAACTGGAAATTTTT
+HWUSI-EAS1789_0001:3:2:2062:1304#0/1
a__[\Bbbb`edeeefd`cc`b]bffff`ffffff
@HWUSI-EAS1789_0001:3:2:3194:1303#0/1
GAACANTCCAACGCTTGGTGAATTCTGCTTCACAA
+HWUSI-EAS1789_0001:3:2:3194:1303#0/1
ZZ[[VBZZY][TWQQZ\ZS\[ZZXV__\OX`a[ZZ
@HWUSI-EAS1789_0001:3:2:3716:1304#0/1
GGAAANAAGACCCTGTTGAGCTTGACTCTAGTCTG
+HWUSI-EAS1789_0001:3:2:3716:1304#0/1
aaXWYBZVTXZX_]XdccdfbbB_\`a\aY_^]LZ^
@HWUSI-EAS1789_0001:3:2:5000:1304#0/1
CCCGGNGATCCGCTGGGACAAGCAGCATATTGATA
+HWUSI-EAS1789_0001:3:2:5000:1304#0/1
aaaaaBeeeeffffehhhhhhggdhhhahhhadh
```

name

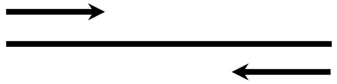
sequence

qualities

read

<u>paired-end reads</u>

read1



read2

1 Illumina HiSeq 2500 lane

~150 million reads

Tasks with RNA-Seq data

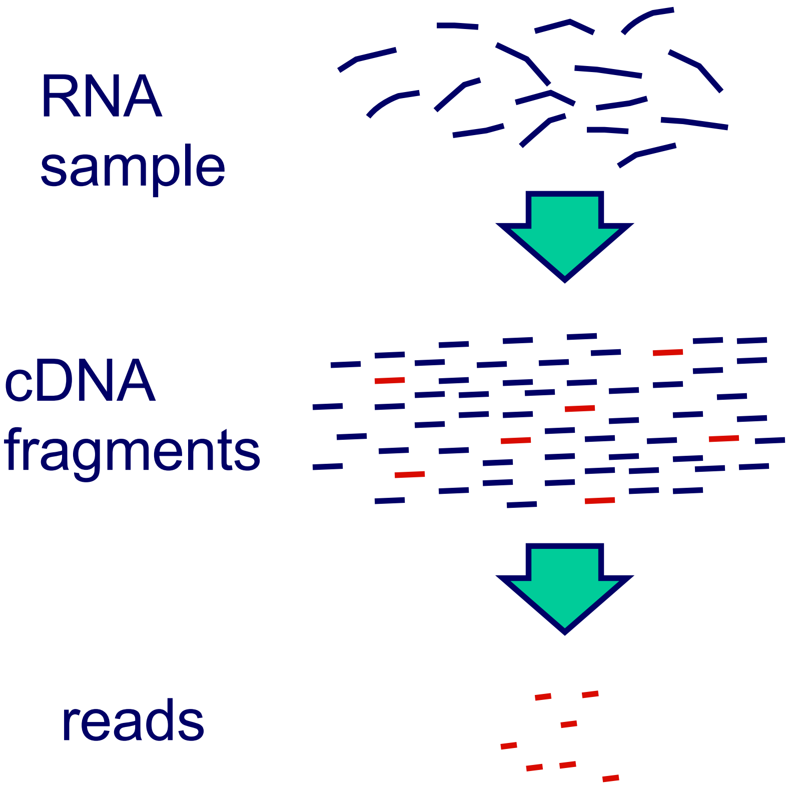
- **Assembly:**
  - Given: RNA-Seq reads (and possibly a genome sequence)
  - Do: Reconstruct full-length transcript sequences from the reads
- **Quantification (our focus):**
  - Given: RNA-Seq reads and transcript sequences
  - Do: Estimate the relative abundances of transcripts (“gene expression”)
- **Differential expression or additional downstream analyses:**
  - Given: RNA-Seq reads from two different samples and transcript sequences
  - Do: Predict which transcripts have different abundances between two samples

RNA-Seq is a *relative* abundance measurement technology

- RNA-Seq gives you reads from the ends of a random **sample** of fragments in your library

- Without additional data this only gives information about **relative** abundances

- Additional information, such as levels of "spike-in" transcripts, are needed for absolute measurements



RNA sample

cDNA fragments

reads

Issues with relative abundance measures

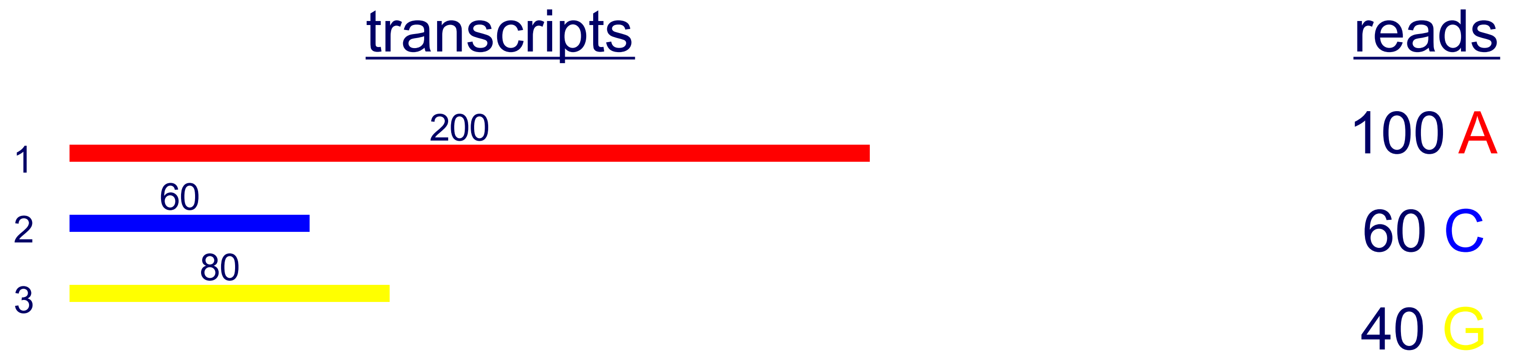
Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

- Changes in absolute expression of high expressors is a major factor

- Normalization is required for comparing samples in these situations

The basics of quantification with RNA-Seq data

- For simplicity, suppose reads are of length **one** (typically they are > 35 bases)



transcripts

reads

1 200 — 100 A

2 60 — 60 C

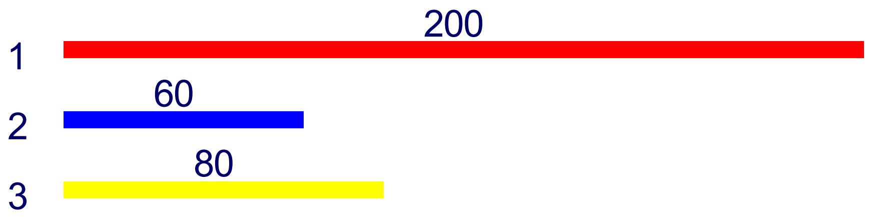
3 80 — 40 G

- What relative abundances would you estimate for these genes?
- Relative abundance is relative transcript levels in the cell, not proportion of observed reads

Length dependence

- Probability of a read coming from a transcript ∝ relative abundance × length

transcripts



reads

100 A

60 C

40 G

probability of read from transcript 1 = (transcript 1 reads) / (total reads)

transcript 1 relative abundance

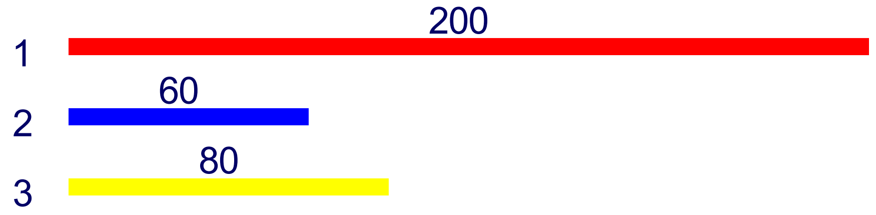
$$\hat{f}_1 \propto \frac{\frac{100}{200}}{200} = \frac{1}{400}$$

transcript 1 length

Length dependence

- Probability of a read coming from a transcript ∝ relative abundance × length

transcripts



reads

100 A

60 C

40 G

$$\hat{f}_1 \propto \frac{\frac{100}{200}}{200} = \frac{1}{400}$$

$$\hat{f}_2 \propto \frac{\frac{60}{200}}{60} = \frac{1}{200}$$

$$\hat{f}_3 \propto \frac{\frac{40}{200}}{80} = \frac{1}{400}$$

normalize

$$\hat{f}_1 = 0.25$$

$$\hat{f}_2 = 0.5$$

$$\hat{f}_3 = 0.25$$

The basics of quantification from RNA-Seq data

- Basic assumption:

$$\theta_i = P(\text{read from transcript } i) = Z^{-1}\tau_i\ell'_i$$

expression level
(relative abundance)

length

- Normalization factor is the mean length of expressed transcripts

$$Z = \sum_i \tau_i \ell'_i$$

The basics of quantification from RNA-Seq data

- Estimate the probability of reads being generated from a given transcript by counting the number of reads that align to that transcript

$$\hat{\theta}_i = \frac{c_i}{N}$$

# reads mapping to transcript i

total # of mappable reads

- Convert to expression levels by normalizing by transcript length

$$\hat{\tau}_i \propto \frac{\hat{\theta}_i}{\ell'_i}$$

The basics of quantification from RNA-Seq data

- Basic quantification algorithm
  - Align reads against a set of reference transcript sequences
  - Count the number of reads aligning to each transcript
  - Convert read counts into relative expression levels

Counts to expression levels

- RPKM - Reads Per Kilobase per Million mapped reads

$$\text{RPKM for gene i} = 10^9 \times \frac{c_i}{\ell'_i N}$$

- FPKM (fragments instead of reads, two reads per fragment, for paired end reads)
- TPM - Transcripts Per Million

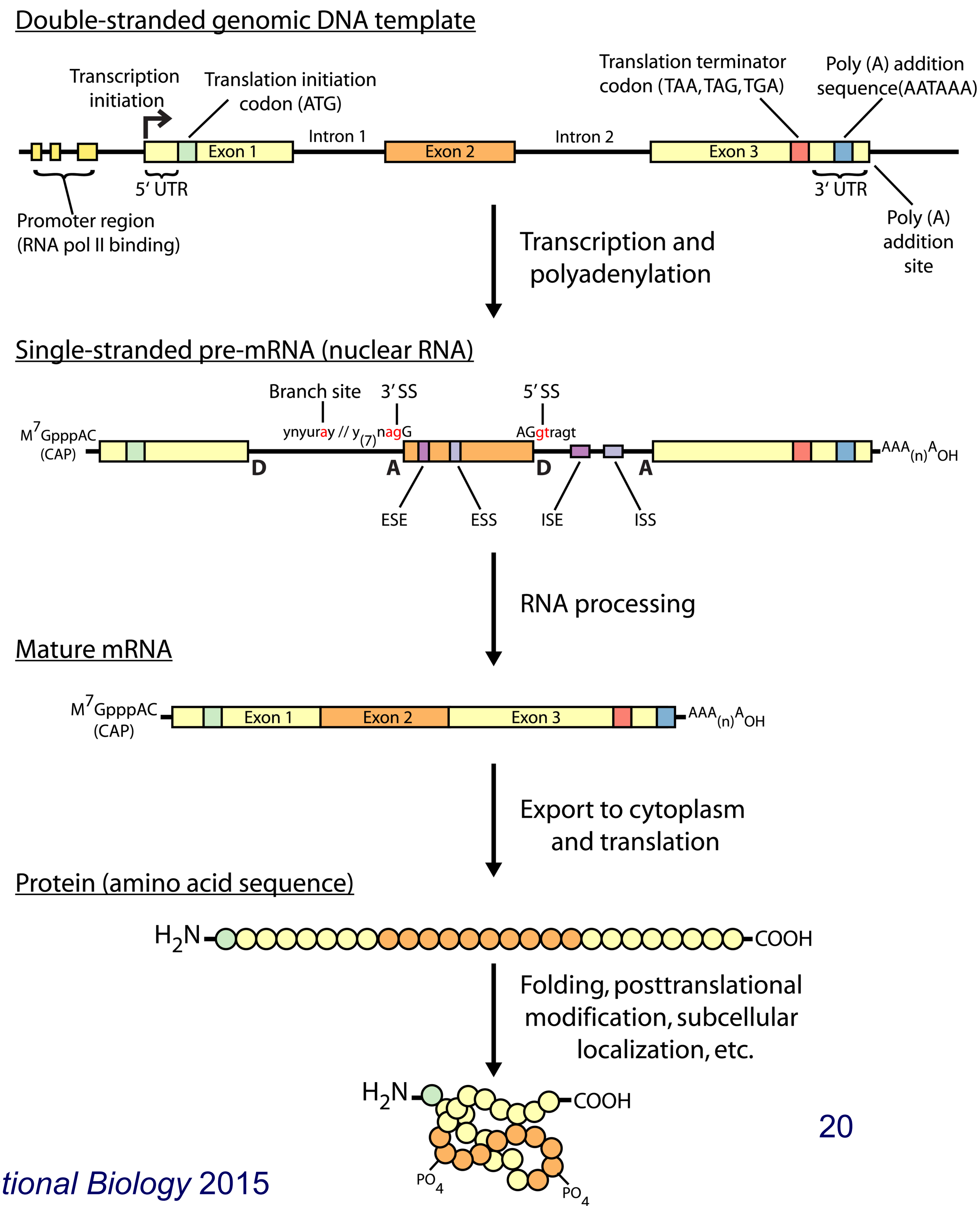
(estimate of) $$\text{TPM for isoform i} = 10^6 \times Z \times \frac{c_i}{\ell'_i N}$$

- Prefer TPM to RPKM because of normalization factor
  - TPM is a technology-independent measure (simply a fraction)

What if reads do not uniquely map to transcripts?

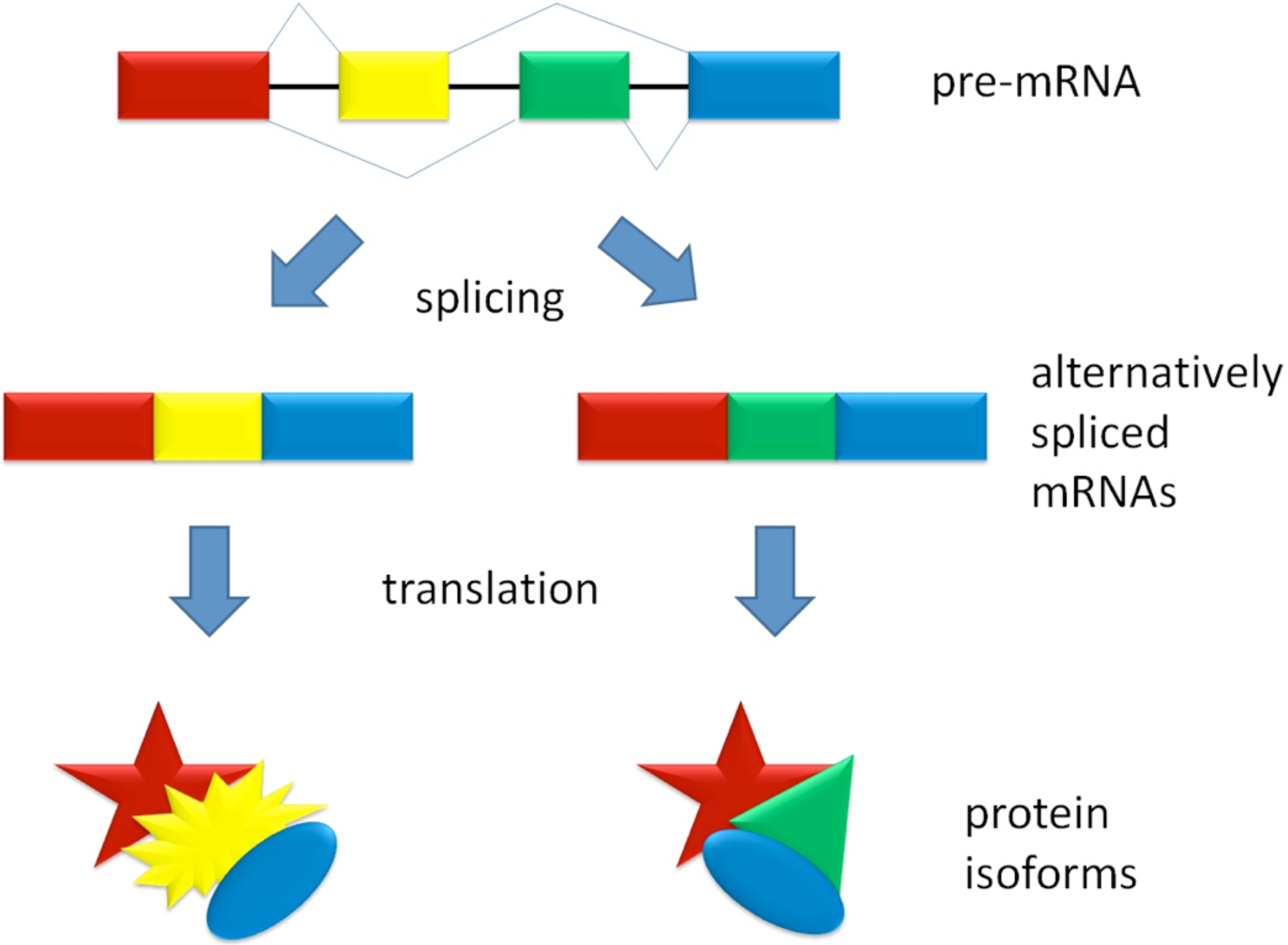
- The approach described assumes that every read can be uniquely aligned to a single transcript
- This is generally not the case
  - Some genes have similar sequences - gene families, repetitive sequences
  - Alternative splice forms of a gene share a significant fraction of sequence

Central dogma of molecular biology


Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
ynyuray // y(7)nagG
AGgtragt
M7GpppAC (CAP)
D
A
D
A
AAA(n)AOH
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
AAA(n)AOH
Export to cytoplasm and translation
Protein (amino acid sequence)
H2N
COOH
Folding, posttranslational modification, subcellular localization, etc.
H2N
COOH
PO4
PO4

Griffith et al. *PLoS Computational Biology* 2015

Alternative splicing

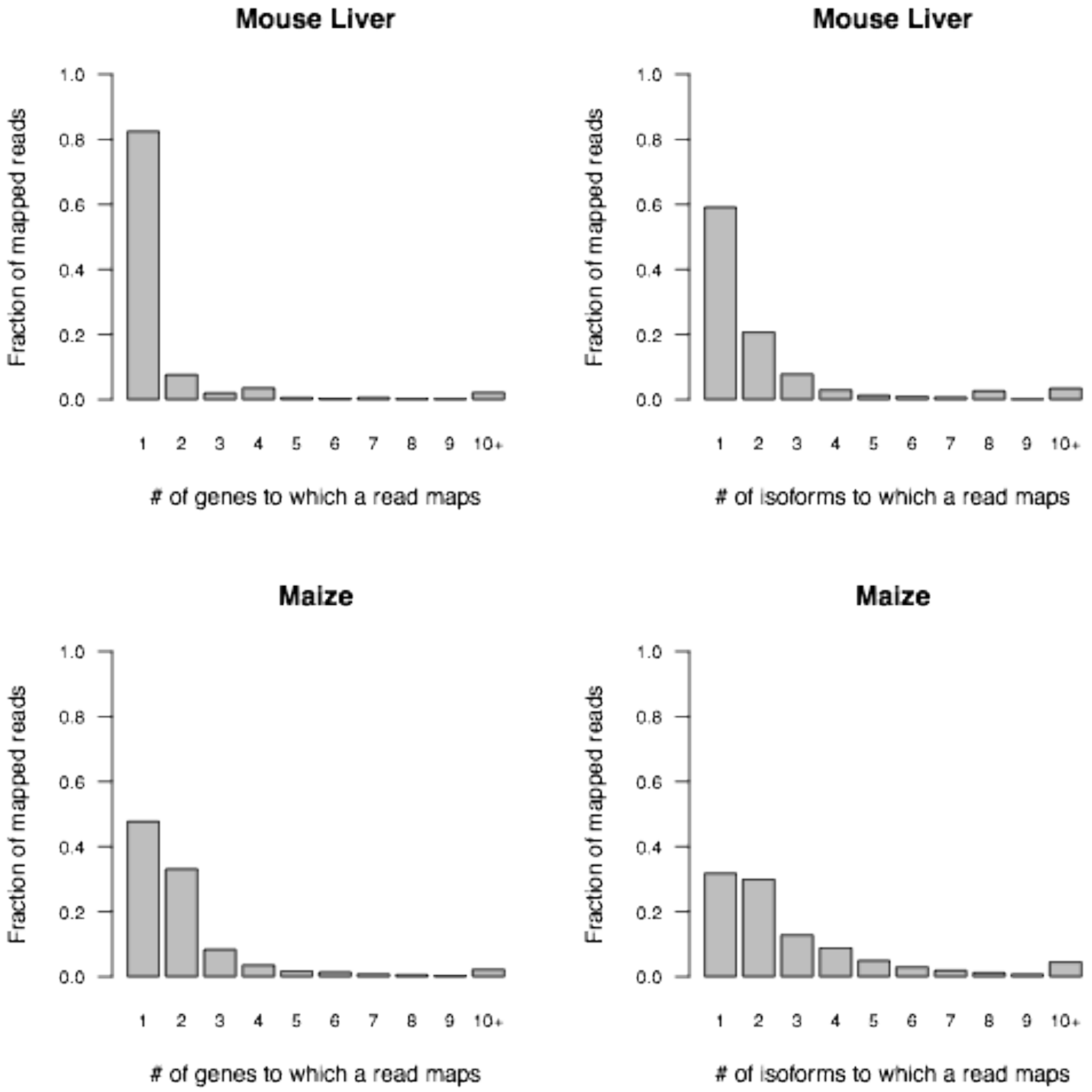


Multi-mapping reads in RNA-Seq

Species	Read length	% multi-mapping reads
Mouse	25	17%
Mouse	75	10%
Maize	25	52%
Axolotl	76	23%
Human	50	23%

- Throwing away multi-mapping reads leads to
  - – Loss of information
  - – Potentially biased estimates of abundance

Distributions of alignment counts


Mouse Liver

Mouse Liver

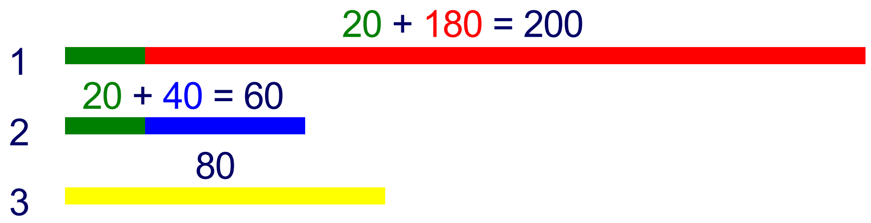
Maize

Maize

What if reads do not uniquely map to transcripts?

- Multiread: a read that could have been derived from multiple transcripts

transcripts



20 + 180 = 200

1

20 + 40 = 60

2

80

3

reads

90 A

40 C

40 G

30 T

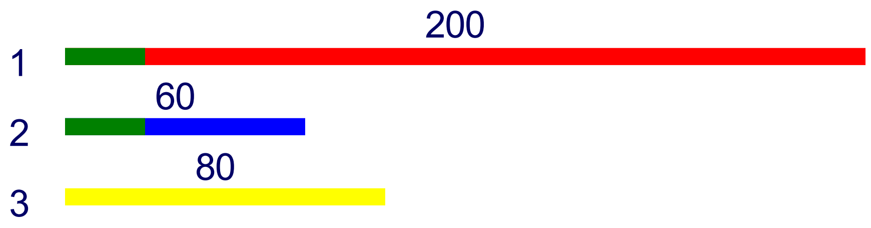
- How would you estimate the relative abundances for these transcripts?

Some options for handling multireads

- Discard multireads, estimate based on uniquely mapping reads only
- Discard multireads, but use “unique length” of each transcript in calculations
- “Rescue” multireads by allocating (fractions of) them to the transcripts
  - Three step algorithm
    1. Estimate abundances based on uniquely mapping reads only
    2. For each multiread, 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

Rescue method example - Step 1

transcripts



reads

90 A

40 C

40 G

30 T

Step 1

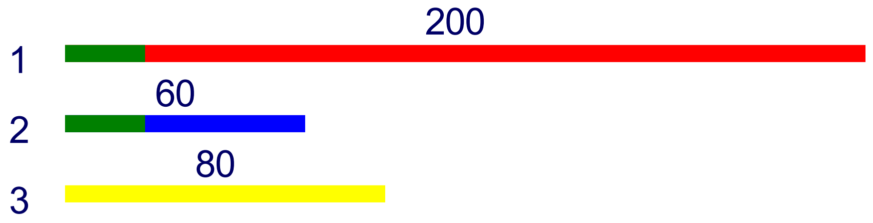
$$\hat{f}_1^{unique} = \frac{\frac{90}{200}}{\frac{90}{200} + \frac{40}{60} + \frac{40}{80}} = 0.278$$

$$\hat{f}_2^{unique} = 0.412$$

$$\hat{f}_3^{unique} = 0.309$$

Rescue method example - Step 2

transcripts



reads

90 A

40 C

40 G

30 T

Step 2

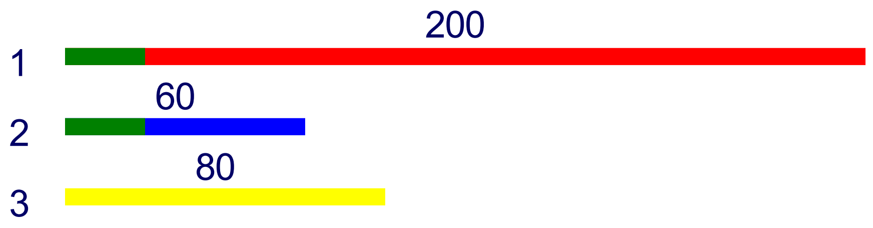
$$c_1^{rescue} = 90 + 30 \times \frac{0.278}{0.278 + 0.412} = 102.1$$

$$c_2^{rescue} = 40 + 30 \times \frac{0.412}{0.278 + 0.412} = 57.9$$

$$c_3^{rescue} = 40 + 0 = 40$$

Rescue method example - Step 3

transcripts



reads

90 A

40 C

40 G

30 T

Step 3

$$\hat{f}_1^{rescue} = \frac{\frac{102.1}{200}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.258$$

$$\hat{f}_2^{rescue} = \frac{\frac{57.9}{60}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.488$$

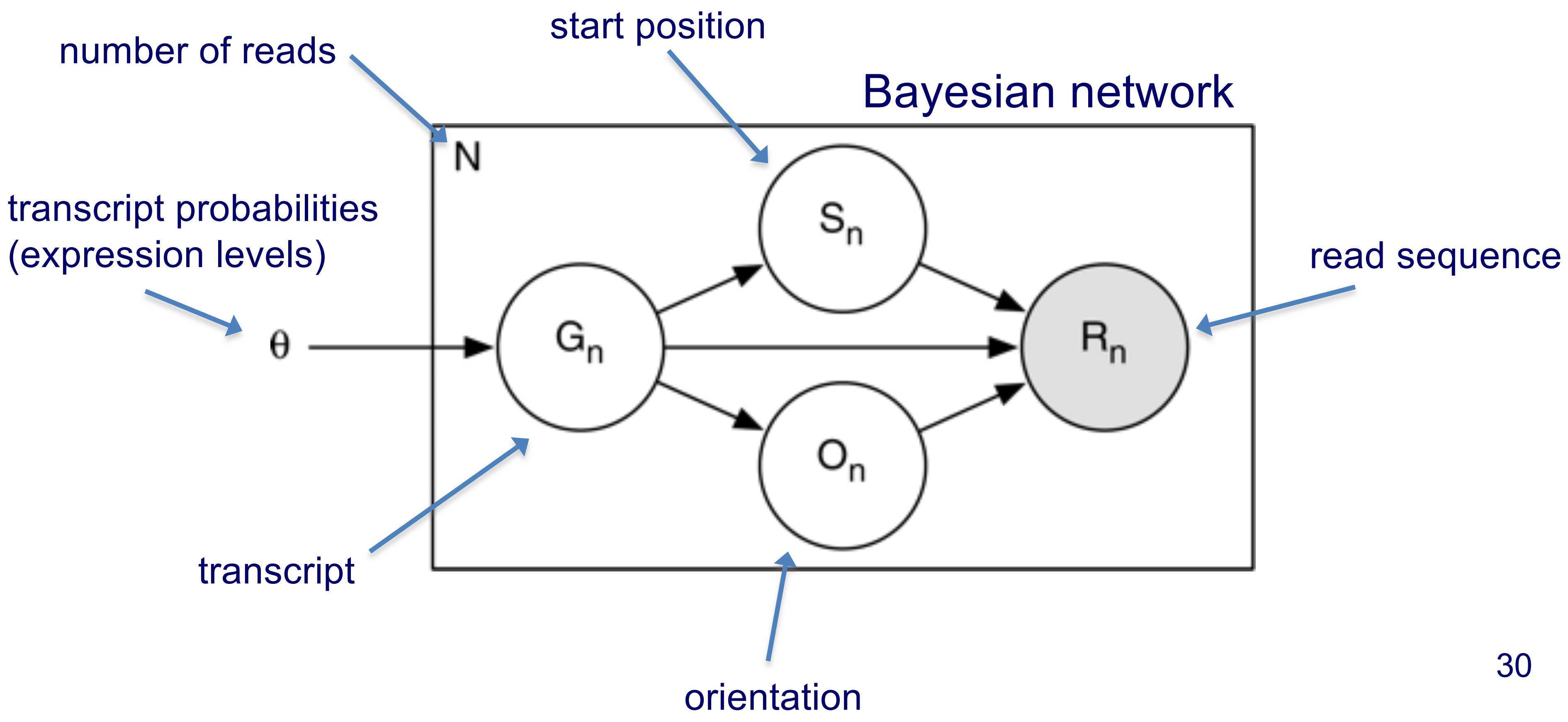
$$\hat{f}_3^{rescue} = \frac{\frac{40}{80}}{\frac{102.1}{200} + \frac{57.9}{60} + \frac{40}{80}} = 0.253$$

An observation about the rescue method

- Note that at the end of the rescue algorithm, we have an updated set of abundance estimates
- These new estimates could be used to reallocate the multireads
- And then we could update our abundance estimates once again
- And repeat!
- This is the intuition behind the statistical approach to this problem

RSEM (**R**NA-**S**eq by **E**xpectation-**M**aximization) - a generative probabilistic model

- Simplified view of the model (plate notation)
  - Grey – observed variable
  - White – latent (unobserved) variables



start position

number of reads

Bayesian network

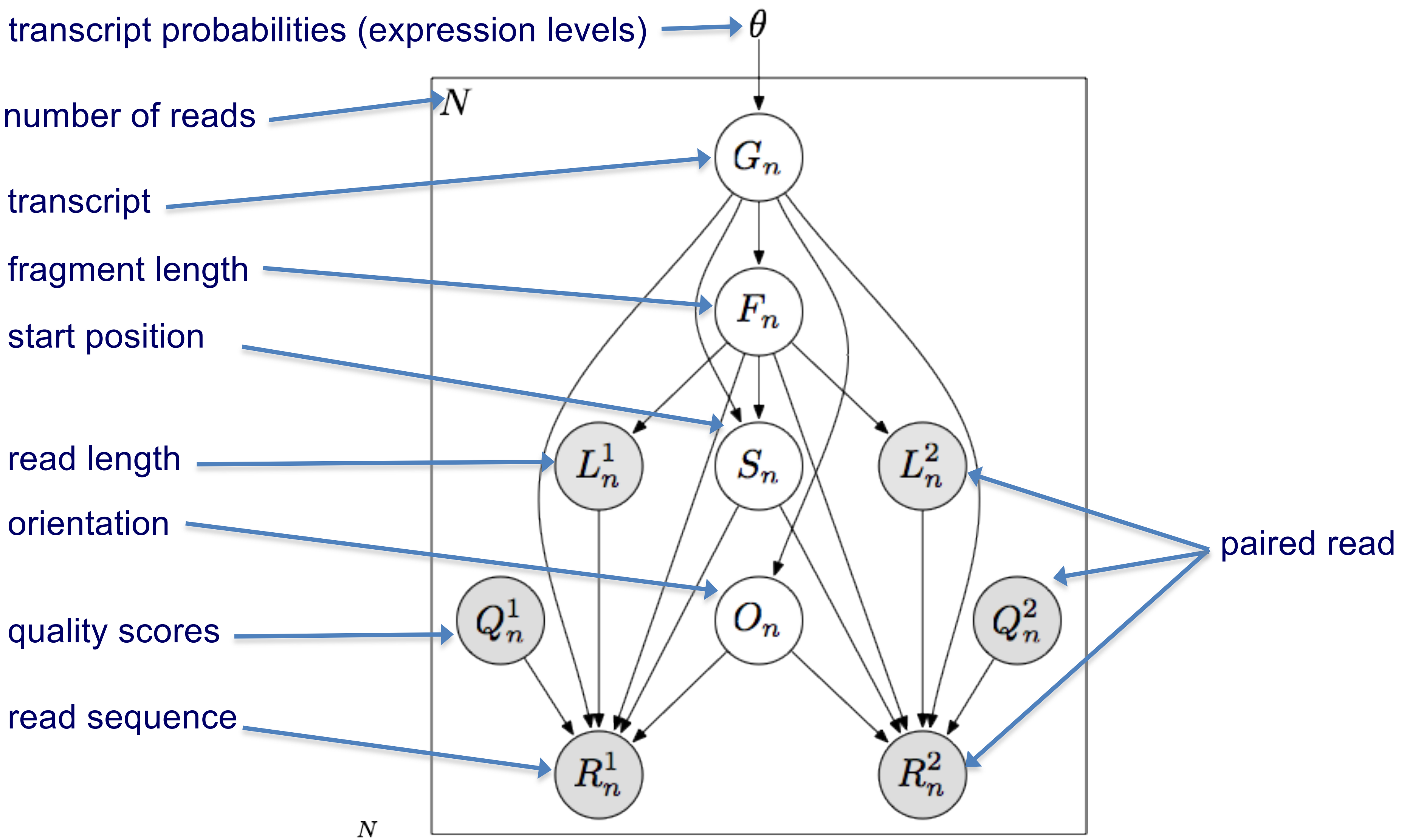
transcript probabilities (expression levels)

read sequence

transcript

orientation

RSEM - a generative probabilistic model



transcript probabilities (expression levels)

number of reads

transcript

fragment length

start position

read length

orientation

quality scores

read sequence

paired read

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

Quantification as maximum likelihood inference

- Observed data likelihood

$$P(\mathbf{r}, \ell, \mathbf{q} | \theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \theta_i \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{o=0}^{1} P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = o | G_n = i)$$

- Likelihood function is concave with respect to θ
  - Has a global maximum (or global maxima)
- Expectation-Maximization for optimization

"RNA-Seq gene expression estimation with read mapping uncertainty"
Li, B., Ruotti, V., Stewart, R., Thomson, J., Dewey, C.
Bioinformatics, 2010

Approximate inference with read alignments

$$P(\mathbf{r},\ell,\mathbf{q}|\theta) = \prod_{n=1}^{N}\sum_{i=0}^{M}\theta_i\sum_{j=0}^{L_i}\sum_{k=0}^{L_i}\sum_{o=0}^{1} P(R_n=r_n, L_n=\ell_n, Q_n=q_n, S_n=j, F_n=k, O_n=o|G_n=i)$$

- Full likelihood computation requires O(NML^2) time
  - N (number of reads) ~ 10^7
  - M (number of transcripts) ~ 10^4
  - L (average transcript length) ~ 10^3
- Approximate by alignment

$$P(\mathbf{r},\ell,\mathbf{q}|\theta) = \prod_{n=1}^{N}\sum_{(i,j,k,o)\in\pi_n^x}\theta_i P(R_n=r_n, L_n=\ell_n, Q_n=q_n, Z_{nijko}=1|G_n=i)$$

$\Uparrow$

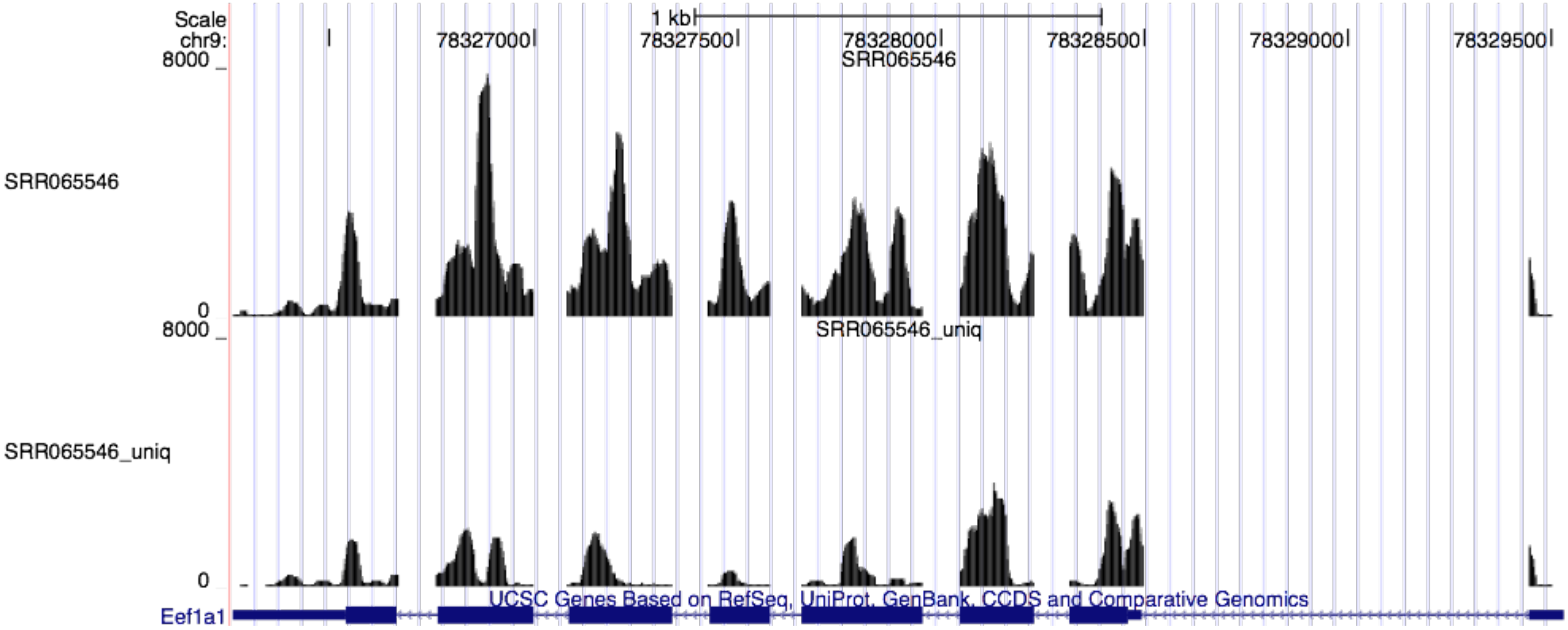
all local alignments of read n with at most x mismatches

EM Algorithm

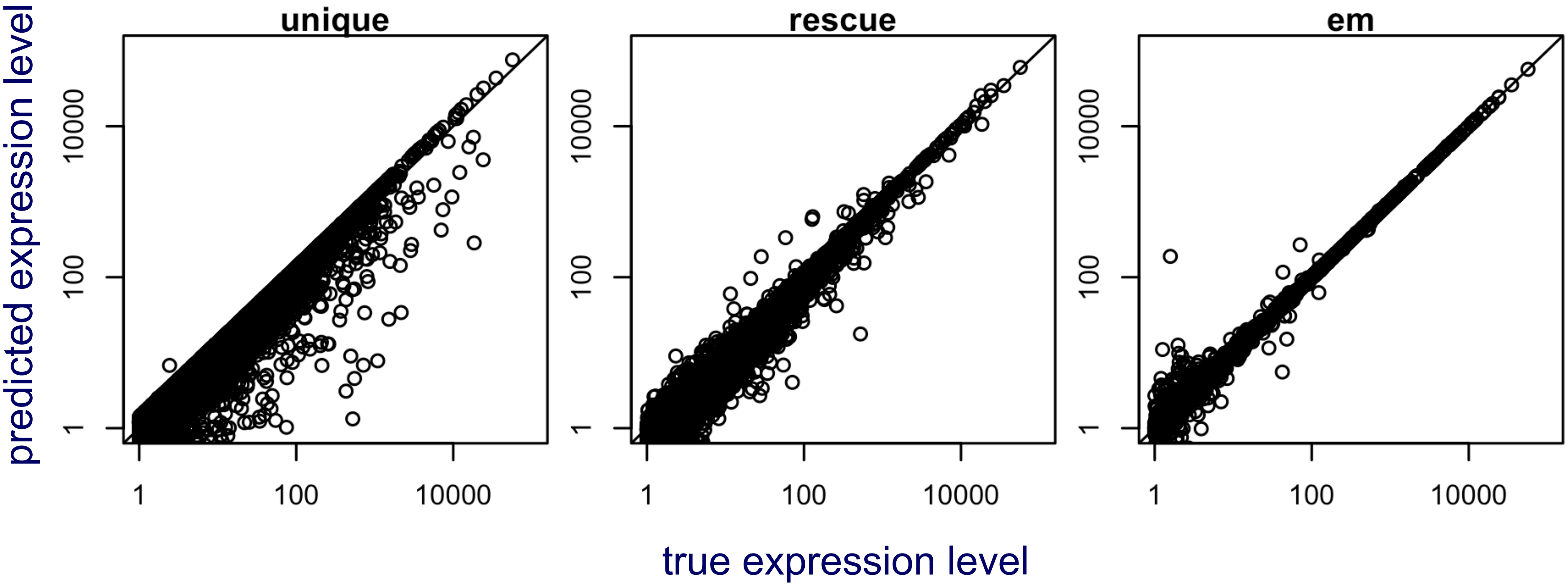
- Expectation-Maximization for RNA-Seq
  - E-step: Compute expected read counts given current expression levels
  - M-step: Compute expression values maximizing likelihood given expected read counts

- Rescue algorithm ≈ 1 iteration of EM

Expected read count visualization

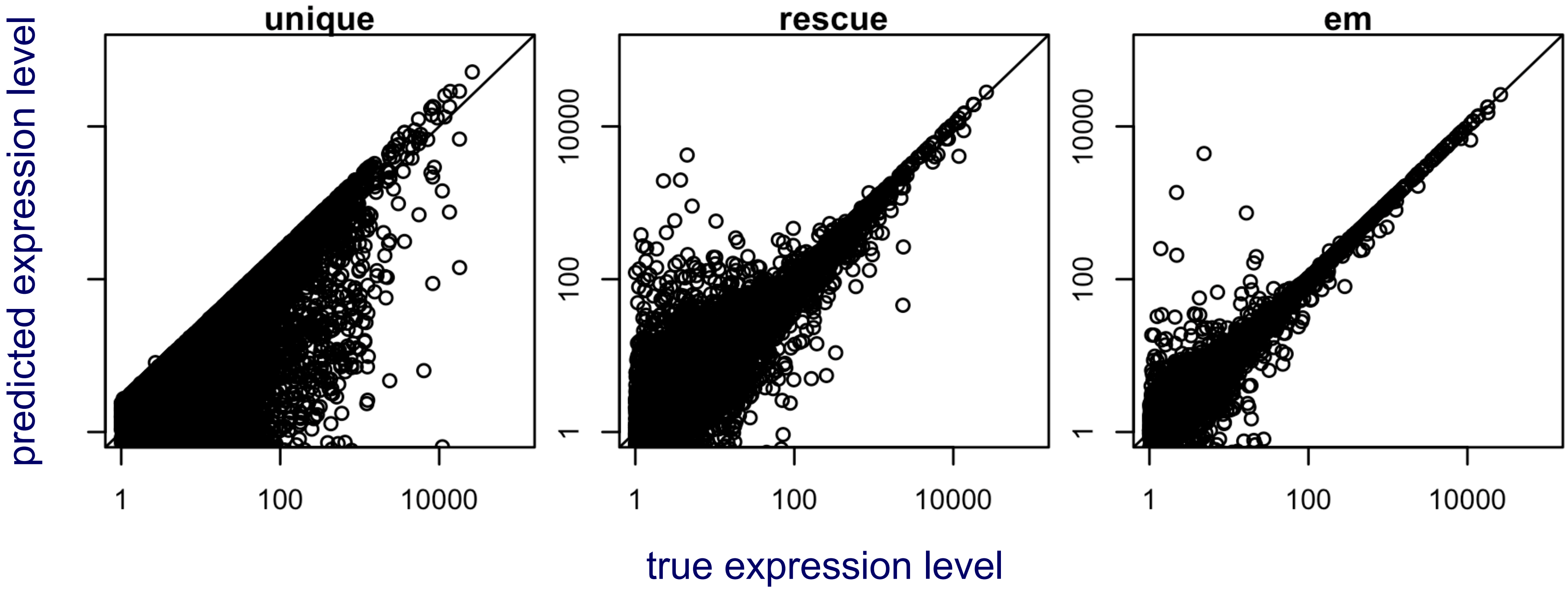


Improved accuracy over unique and rescue



Mouse gene-level expression estimation

Improving accuracy on repetitive genomes: maize



Maize gene-level expression estimation

RNA-Seq and RSEM summary

- RNA-Seq is the preferred technology for transcriptome analysis in most settings
- The major challenge in analyzing RNA-Seq data: the reads are much shorter than the transcripts from which they are derived
- Tasks with RNA-Seq data thus require handling hidden information: which gene/isoform gave rise to a given read
- The Expectation-Maximization algorithm is extremely powerful in these situations

Recent developments in RNA-Seq

- Long read sequences: PacBio and Oxford Nanopore

- Single-cell RNA-Seq: review
  - Observe heterogeneity of cell populations
  - Model technical artifacts (e.g. artificial 0 counts)
  - Detect sub-populations
  - Predict pseudotime through dynamic processes
  - Detect gene-gene and cell-cell relationships

- Alignment-free quantification:
  - Kallisto
  - Salmon

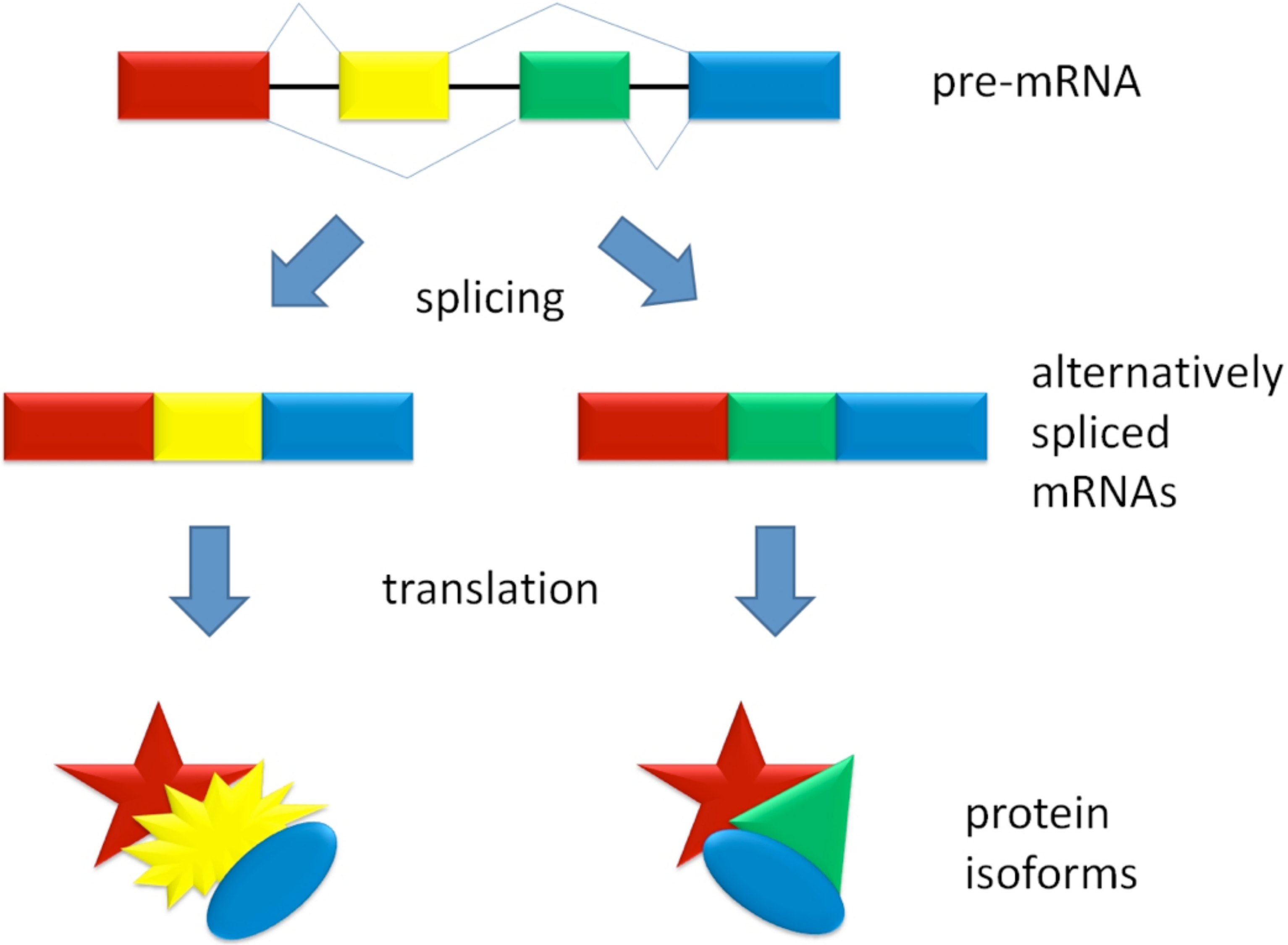
Public sources of RNA-Seq data

- Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo/
  - Both microarray and sequencing data
- Sequence Read Archive (SRA): http://www.ncbi.nlm.nih.gov/sra
  - All sequencing data (not necessarily RNA-Seq)
- ArrayExpress: https://www.ebi.ac.uk/arrayexpress/
  - European version of GEO
- Homogenized data: MetaSRA, Toil, recount2, ARCHS^4

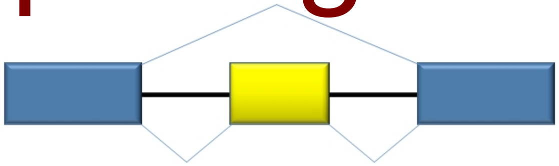
Inference of alternative splicing from RNA-Seq data

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

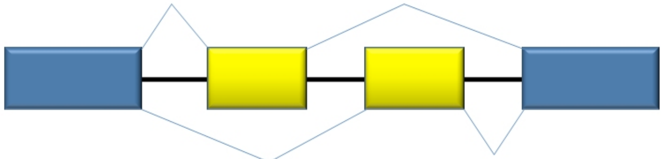
Alternative splicing



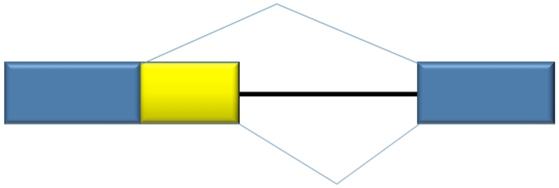
Classes of alternative splicing events



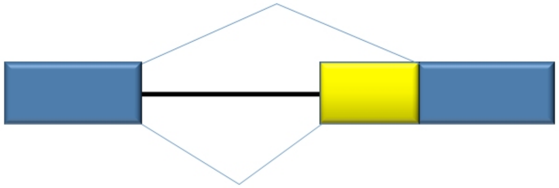
Exon skipping



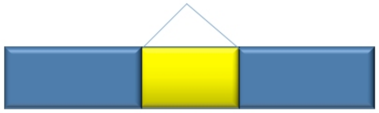
Mutually exclusive exons



Alternative 5' donor sites



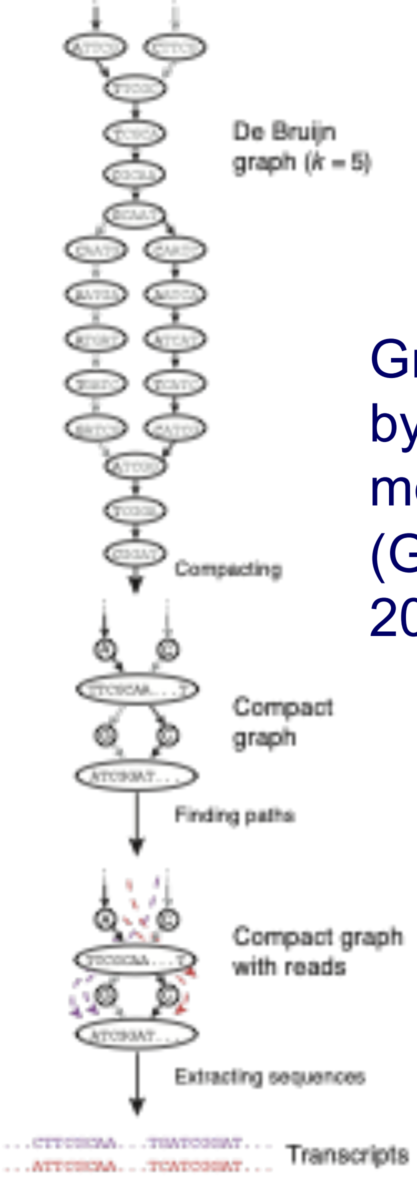
Alternative 3' acceptor sites



Intron retention

Complication 1: De novo transcriptome assembly

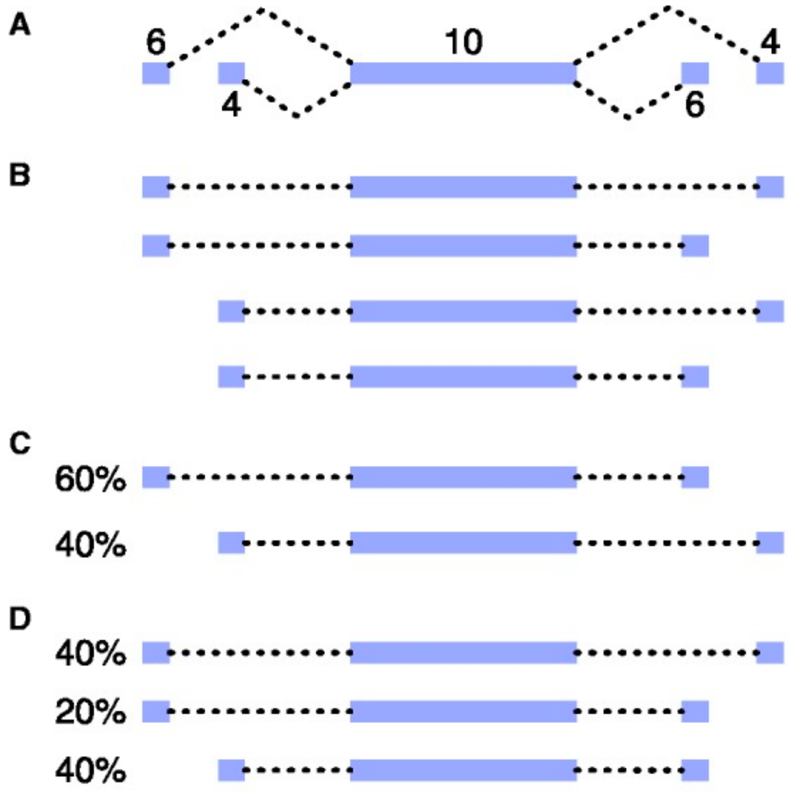
- RNA-Seq reads/fragments are relatively short
- Often insufficient to reconstruct full-length isoforms in the presence of alternative splicing
- Transcriptome assemblies perhaps best left in “graph” form
  - –De Bruijn graph
  - –String graphs



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

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

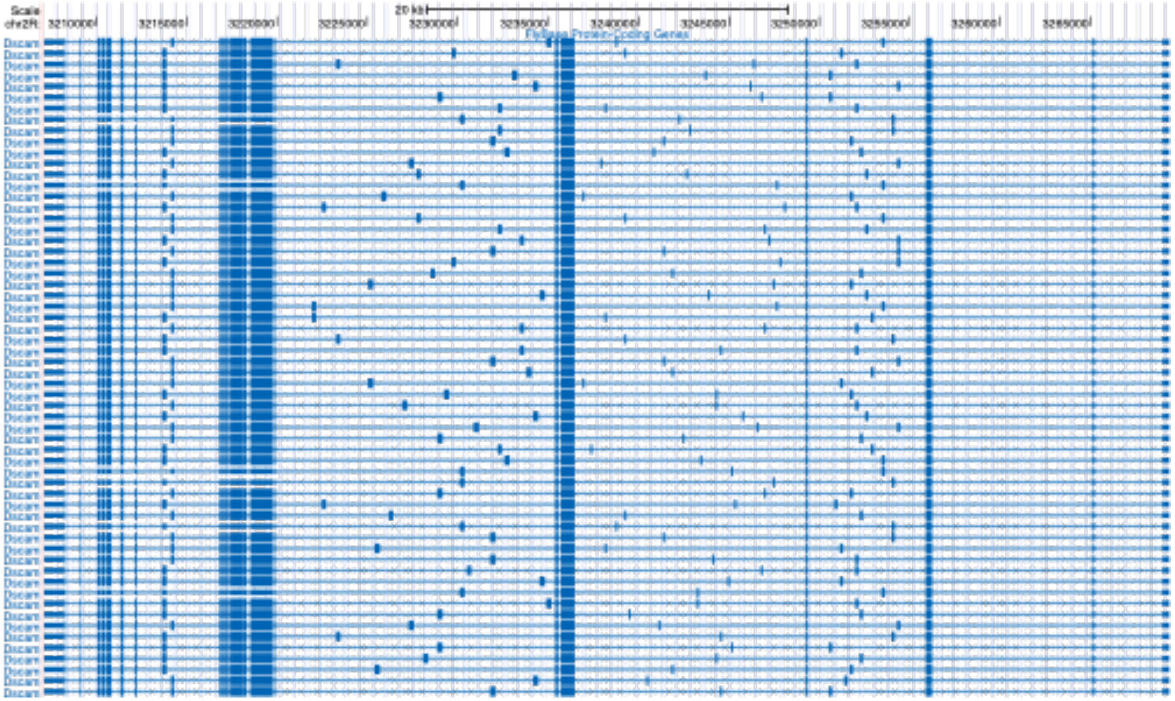
two (of infinitely many) possible isoform abundances



LeGault et al. 2013

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



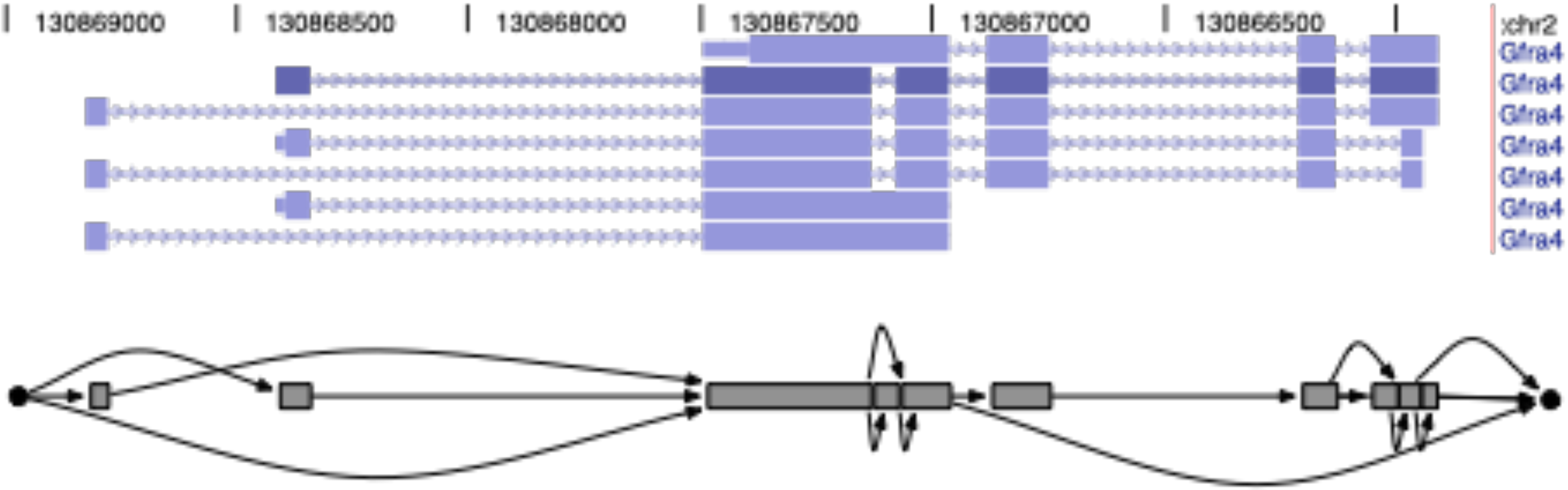
Drosophila *Dscam*: more than 38,000 possible isoforms (Schmucker et al., 2000)

Inference of alternative splicing from RNA-Seq data

- 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

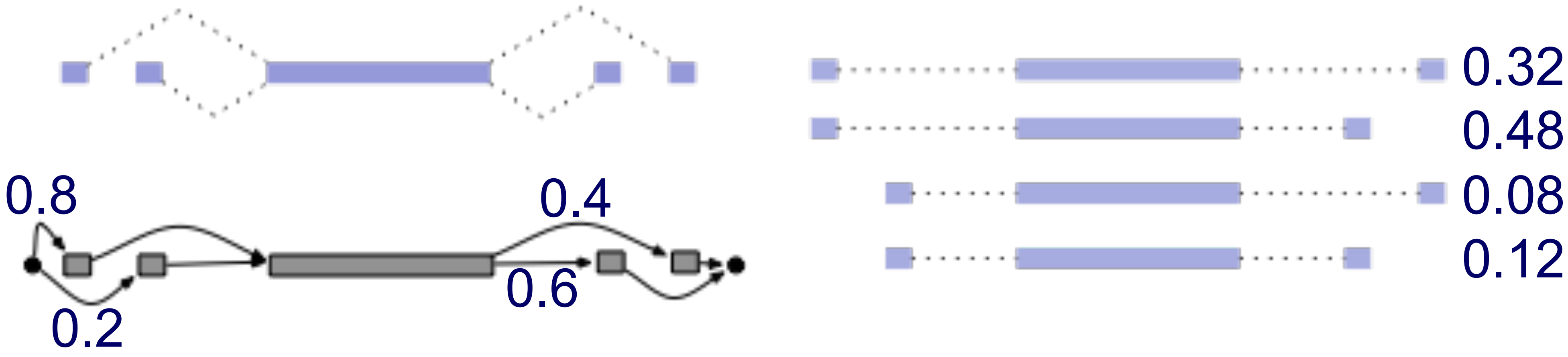


Splice Graphs with EST and RNA-Seq data

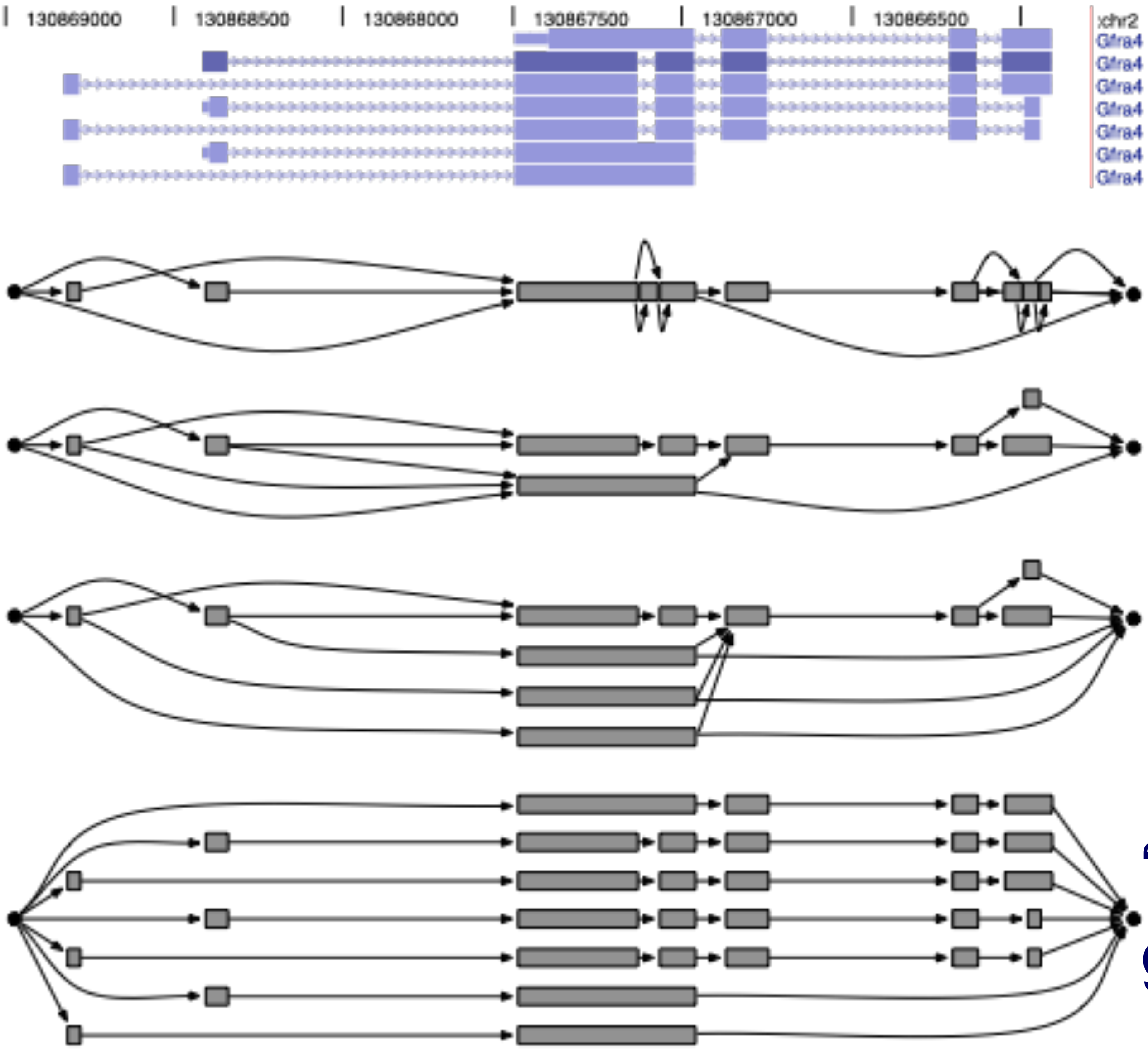
- Xing et al. 2006
  - EM algorithm for estimating abundances of all possible isoforms given splice graph and EST data
  - Expressed Sequence Tag (EST), 74.2 million in 2013
- Montgomery et al. 2010, Singh et al. 2011
  - Graph flow-based methods for quantification/differential splicing given RNA-Seq 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

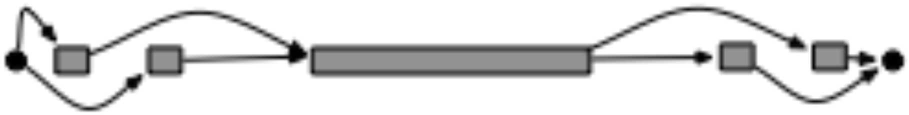
Application of PSGs to RNA-Seq data

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

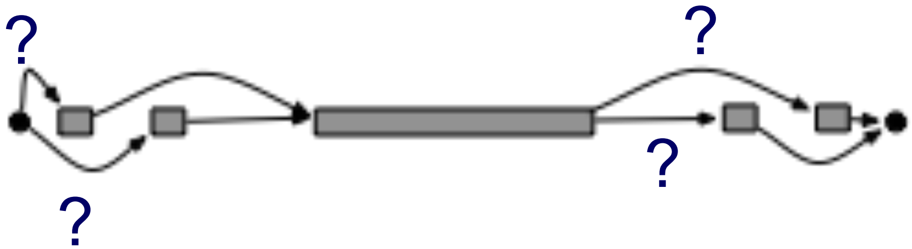
The PSG parameter inference task

- Given: RNA-Seq reads and a PSG structure

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

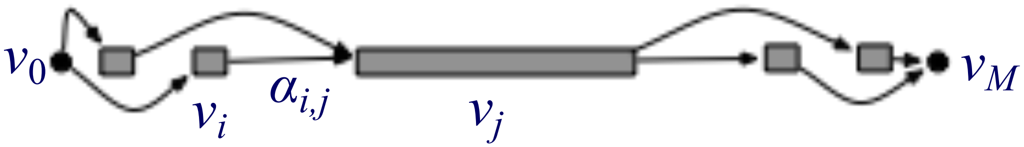


- Do: Estimate the (ML or MAP) parameters for the model



PSG notations

- A directed acyclic graph (DAC)
- Vertex v_i is a sequence with length l_i
- Edge (v_i, v_j) with weight 0<= $\alpha_{i,j}$ <=1
- An isoform is a path s with weight $w(s) = \prod_{i=1}^{|s|-1} \alpha_{s_i,\, s_{i+1}}$



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 from v_0 to v_i

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

Expected suffix length from v_i to v_M

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

EM for PSG parameter estimation

- E-step: compute the expectation of the number of times edge (i,j) is used $E[Z_{nij}] = \frac{\sum_{(b,s)\in\pi(r_n)} g(s,i,j)}{\sum_{(b,s)\in\pi(r_n)} g(s)}$

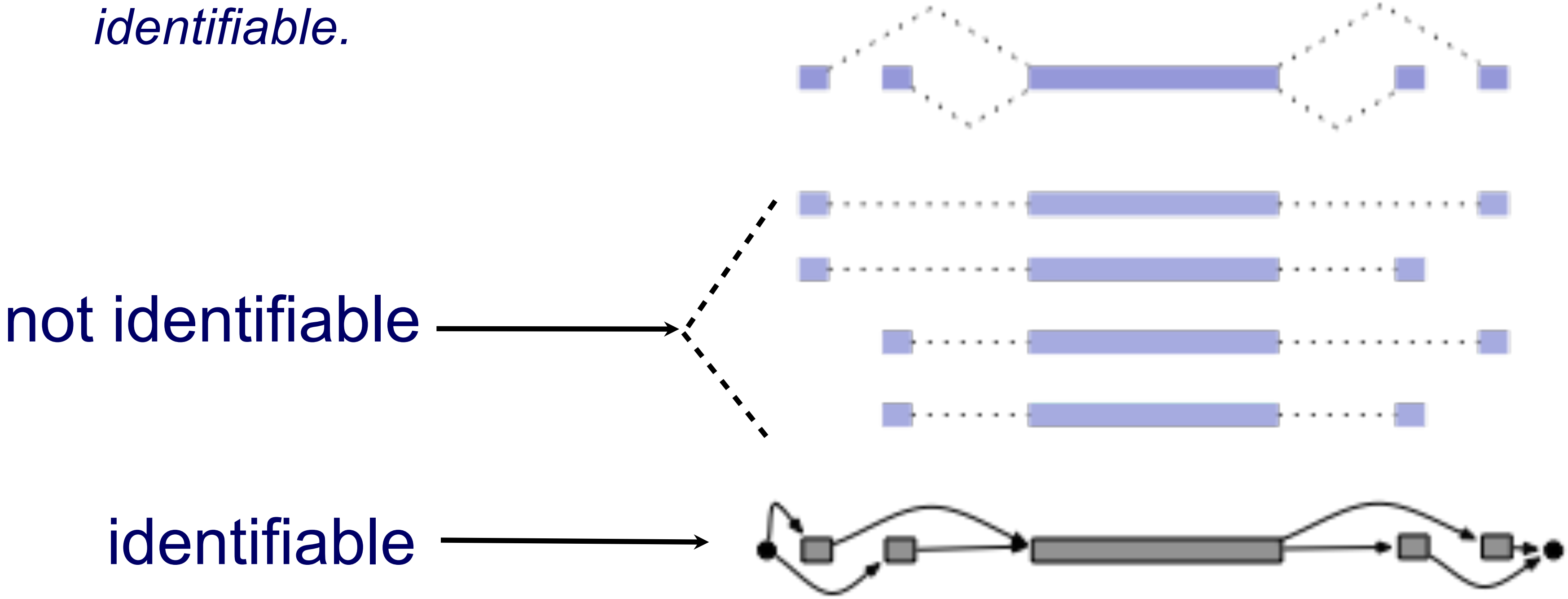
$$
\begin{aligned}
g(s) &= f(0,s_1)w(s) \\
g(s,i,j) &= \begin{cases} f(0,s_1)w(s) & (i,j) \in s \\ f(0,i)\alpha_{ij} f(j,s_1)w(s) & \text{if } \exists \text{ path from } v_j \text{ to } s_1 \\ f(0,s_1)w(s)f(s_{|s|},i)\alpha_{ij} & \text{if } \exists \text{ path from } s_{|s|} \text{ to } v_i \\ 0 & \text{otherwise}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

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

$$
\alpha_{ij} = \frac{\frac{c_{ij}}{(d_p(i)+d_q(j))}}{\sum_k \frac{c_{ik}}{(d_p(i)+d_q(k))}} \qquad c_{ij} = E_{\alpha^{(t)}}[Z_{ij}]
$$

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.*



not identifiable

identifiable

The differential processing (DP) task

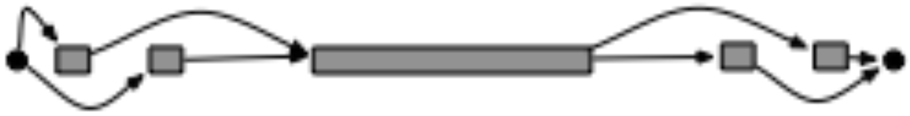
- Given: RNA-Seq reads from two conditions and a PSG structure

condition 1

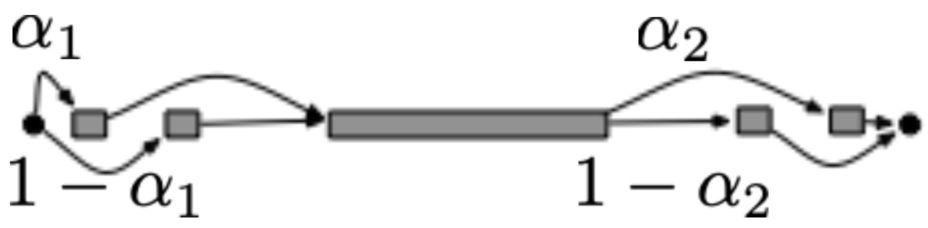
```
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
```

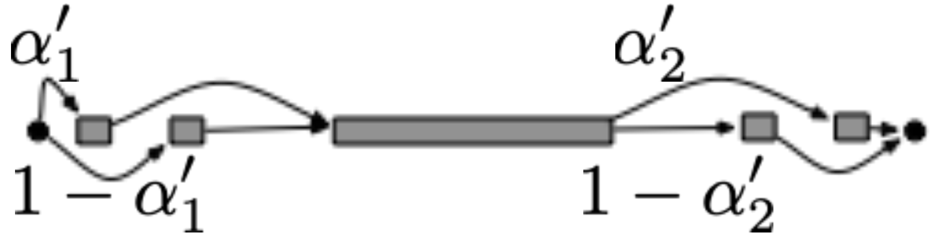
condition 2

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



- Do: Determine if the processing frequencies are different





$\alpha_1 = \alpha_1'$ and $\alpha_2 = \alpha_2'$?

$\alpha_1 = \alpha_1'$ 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

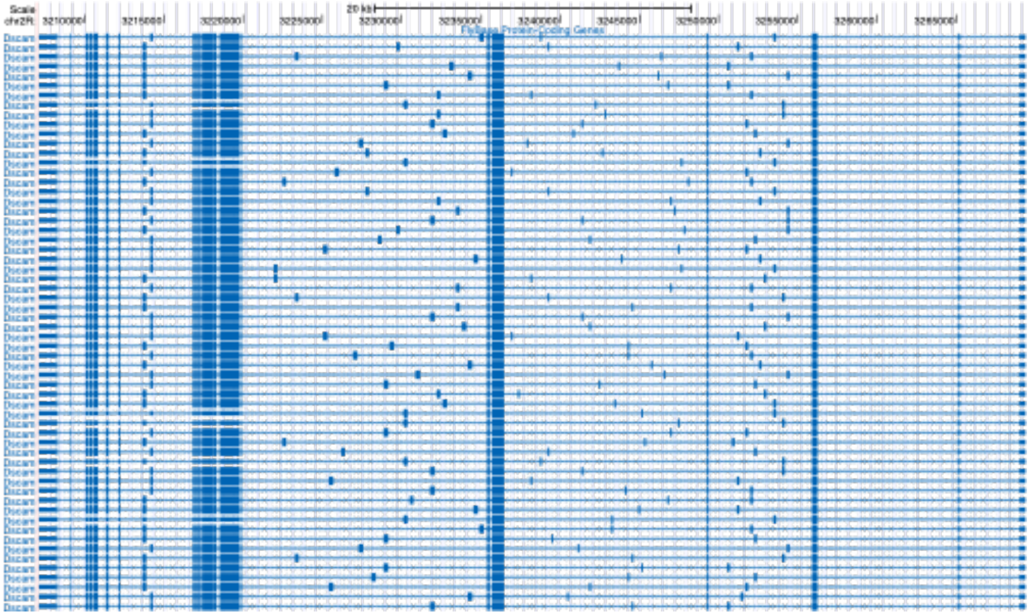
$$\text{LR} = \frac{P(R^1|\hat{\alpha}^1)P(R^2|\hat{\alpha}^2)}{P(R^1, R^2|\hat{\alpha}^1_{\setminus i}, \hat{\alpha}^2_{\setminus i}, \hat{\alpha}^{12}_i))}$$

Inference of alternative splicing from RNA-Seq data

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

Efficient inference for highly-spliced genes

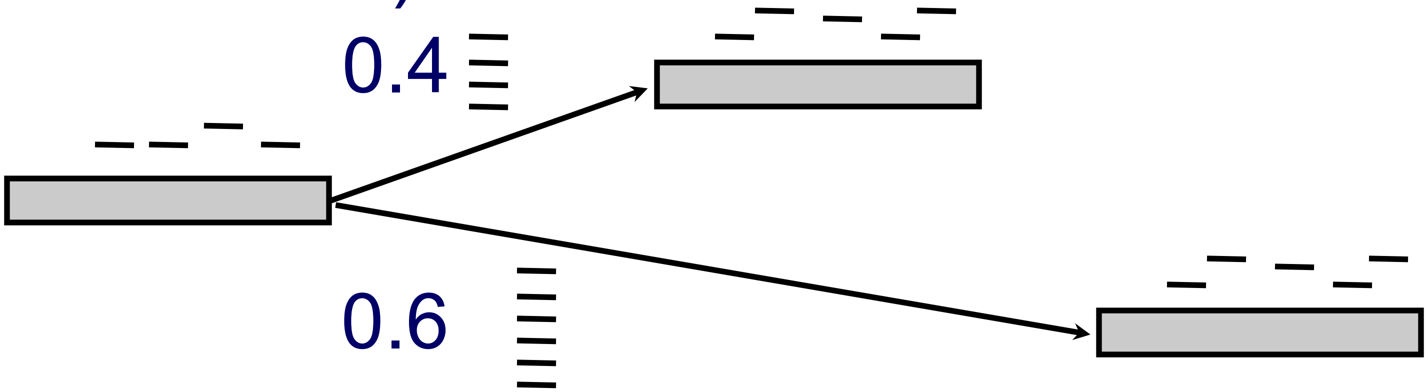
- DSCAM running time test
  - 23,976 isoforms
  - 184 read pairs from a modENCODE sample



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

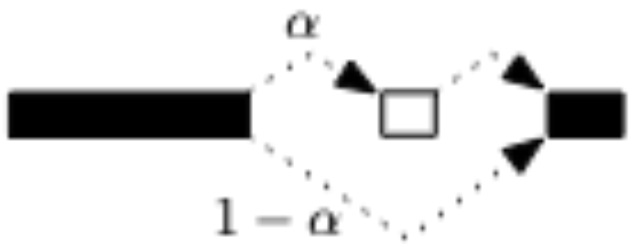
A simple method for comparison

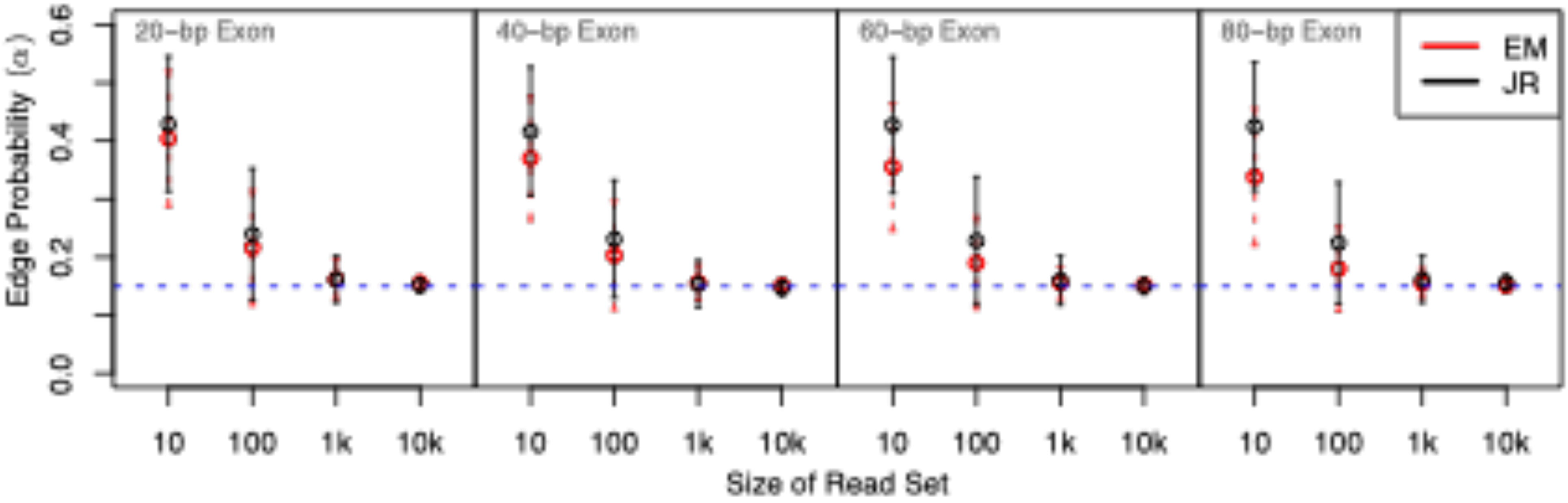
- The **Junction-Read** (JR) method
- Keep only reads that align to the splice junctions (edges in the PSG)



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

Convergence with simulated data

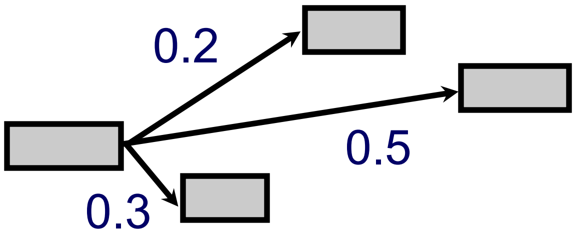




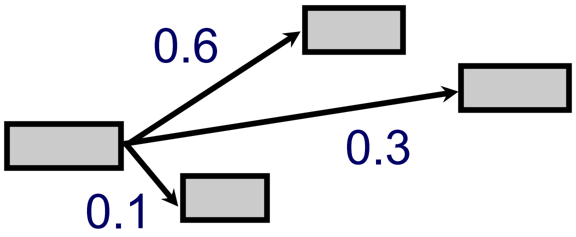
Comparisons on real data

- 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)

method A

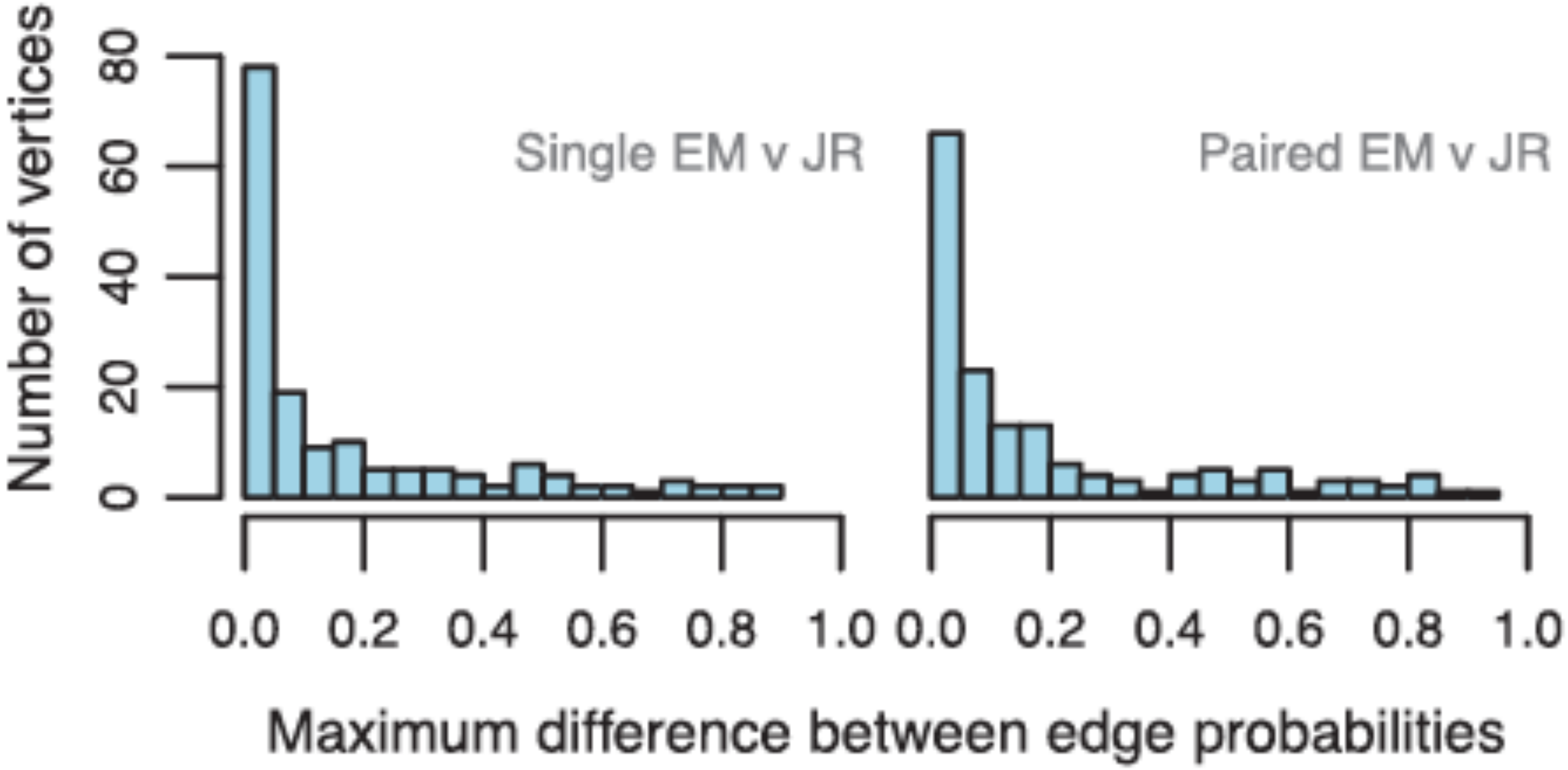


method B



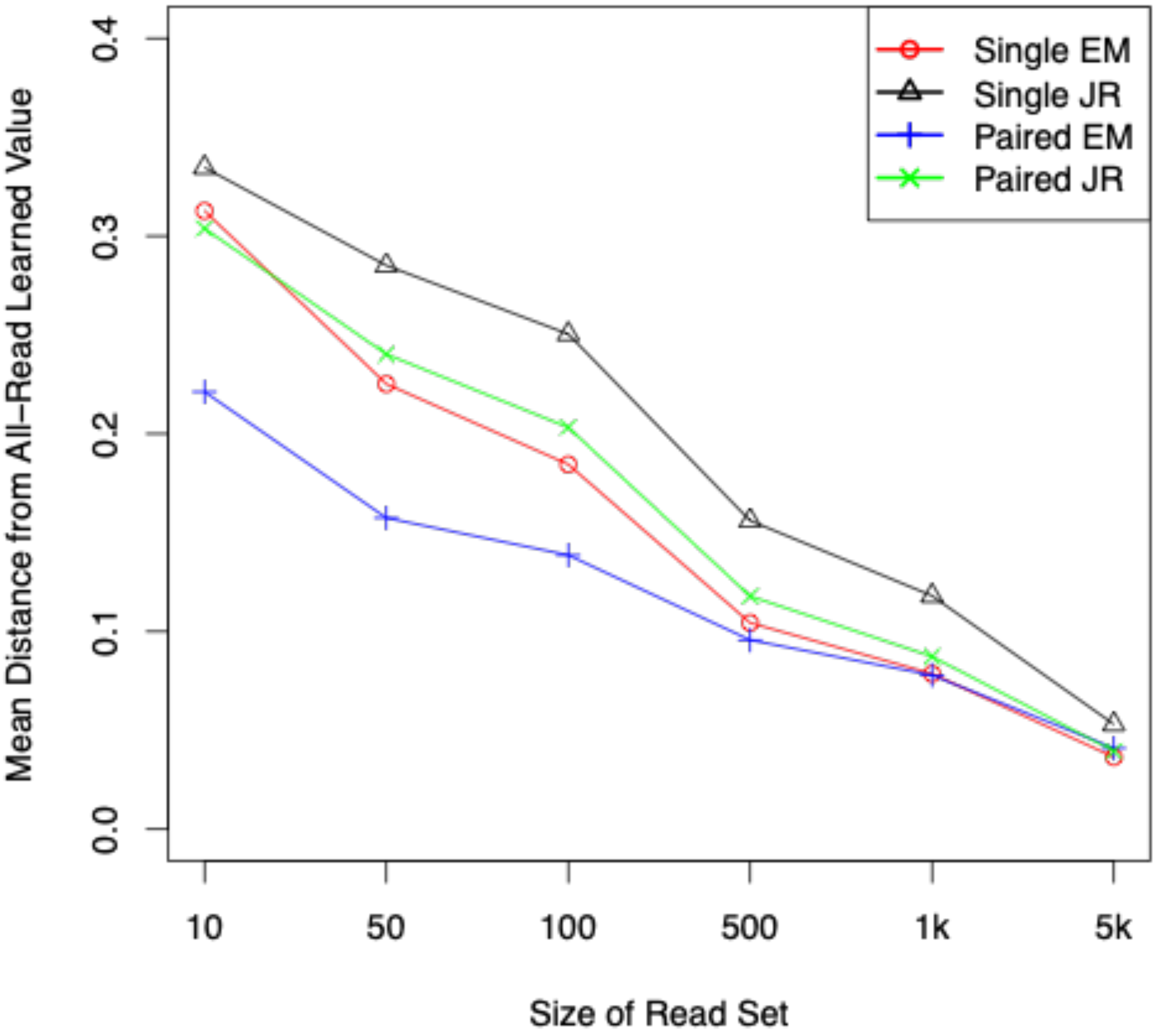
$$distance_v(A, B) = max(|0.6 - 0.2|, |0.5 - 0.3|, |0.3 - 0.1|) = 0.4$$

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

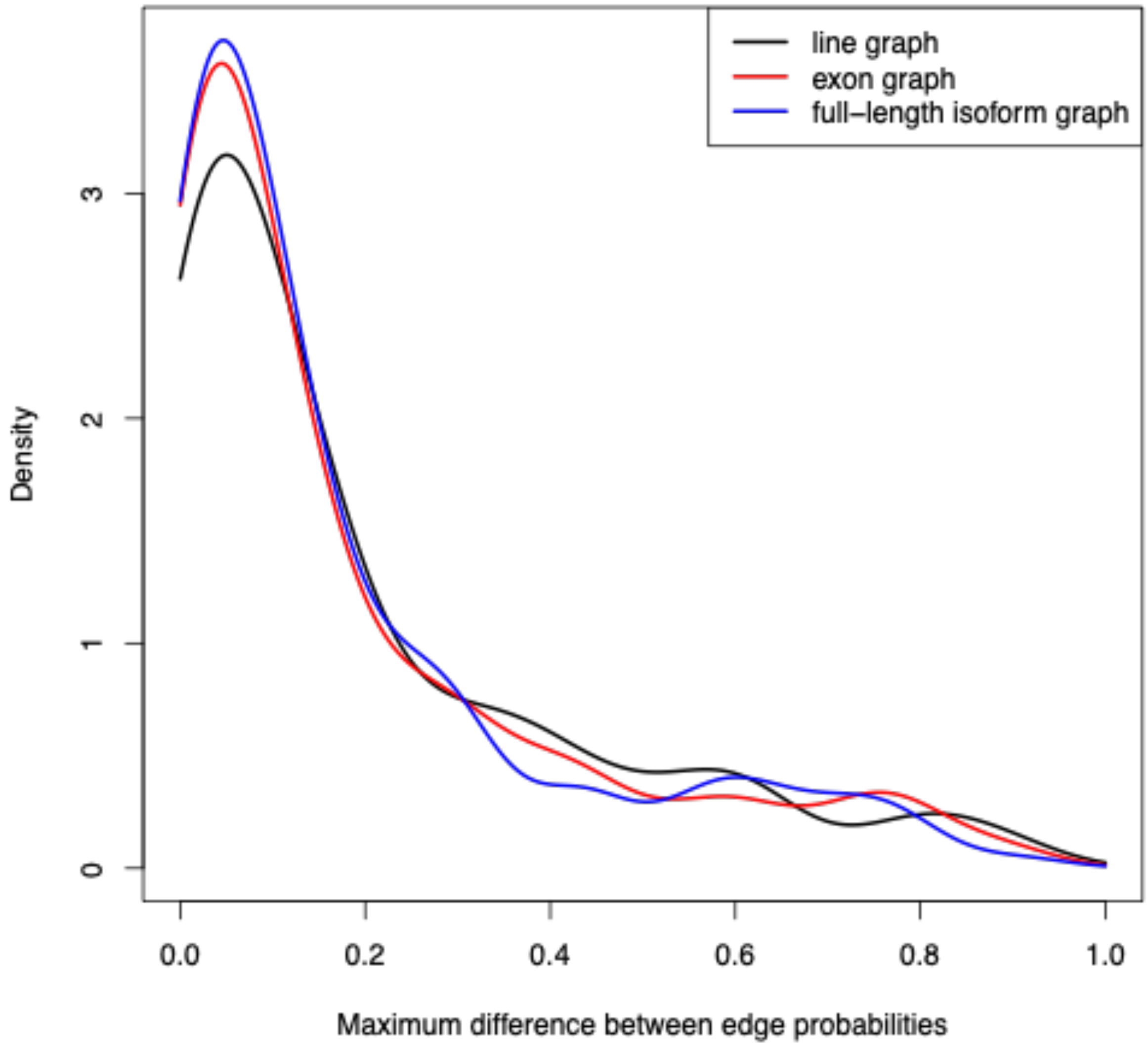


Vertices from 88 most abundant (> 5000 reads) alternatively-spliced genes in a modENCODE fly data set

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

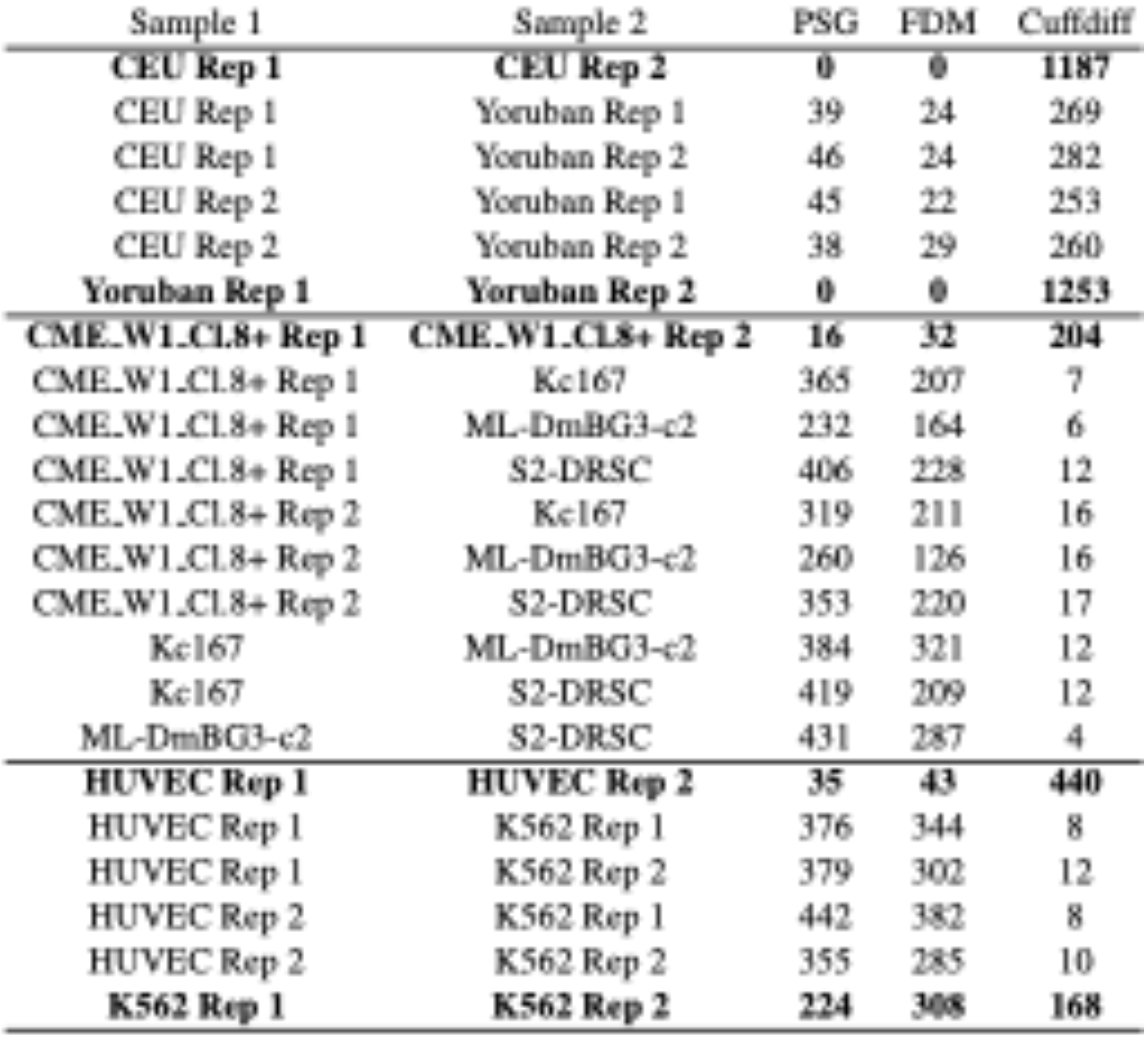
- PSG estimates converge more quickly as the data set increases in size
  - ⇒Our EM estimation procedure uses information from all reads, not just those that span splice junctions

- 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

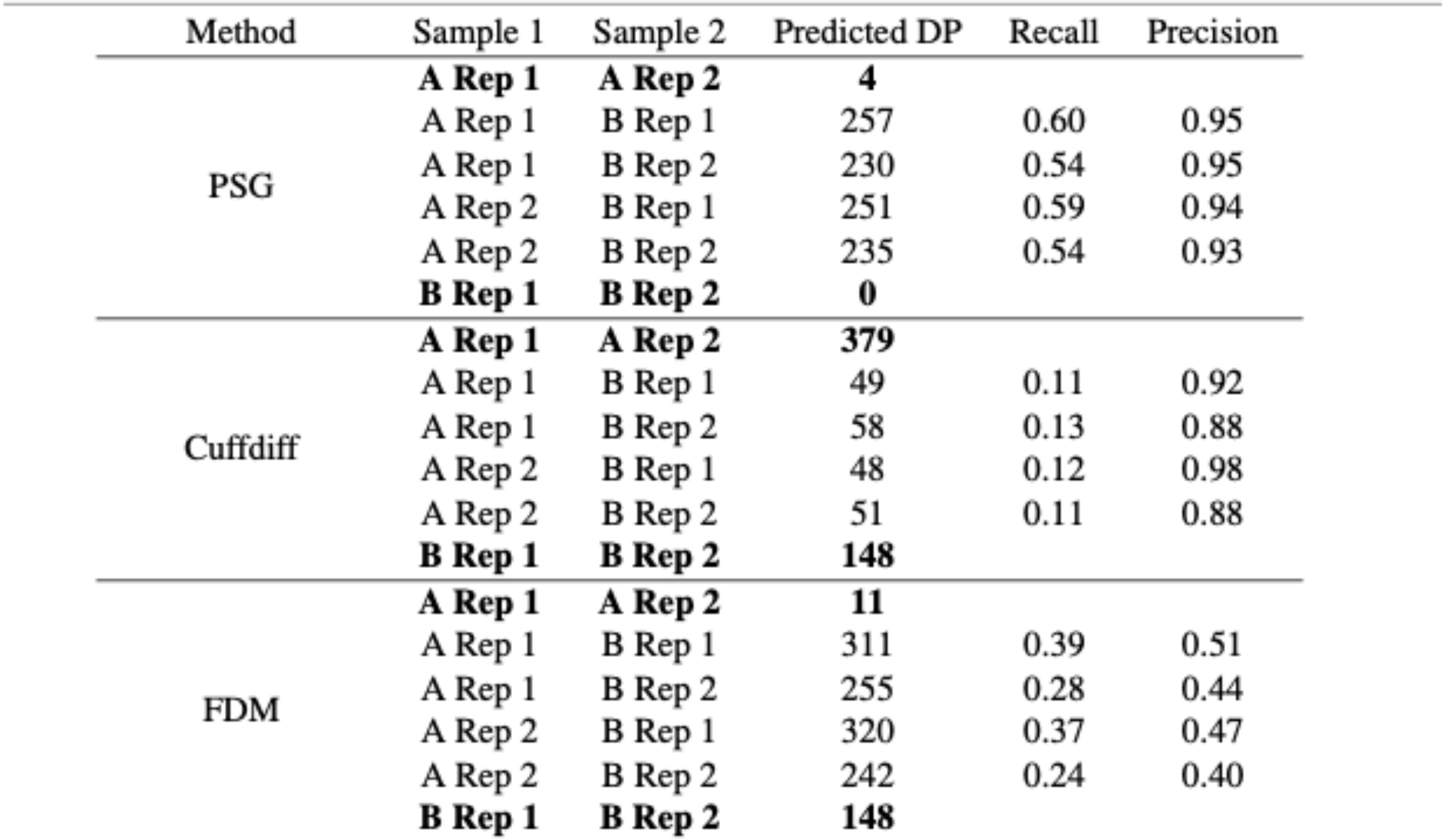
# of DP genes



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

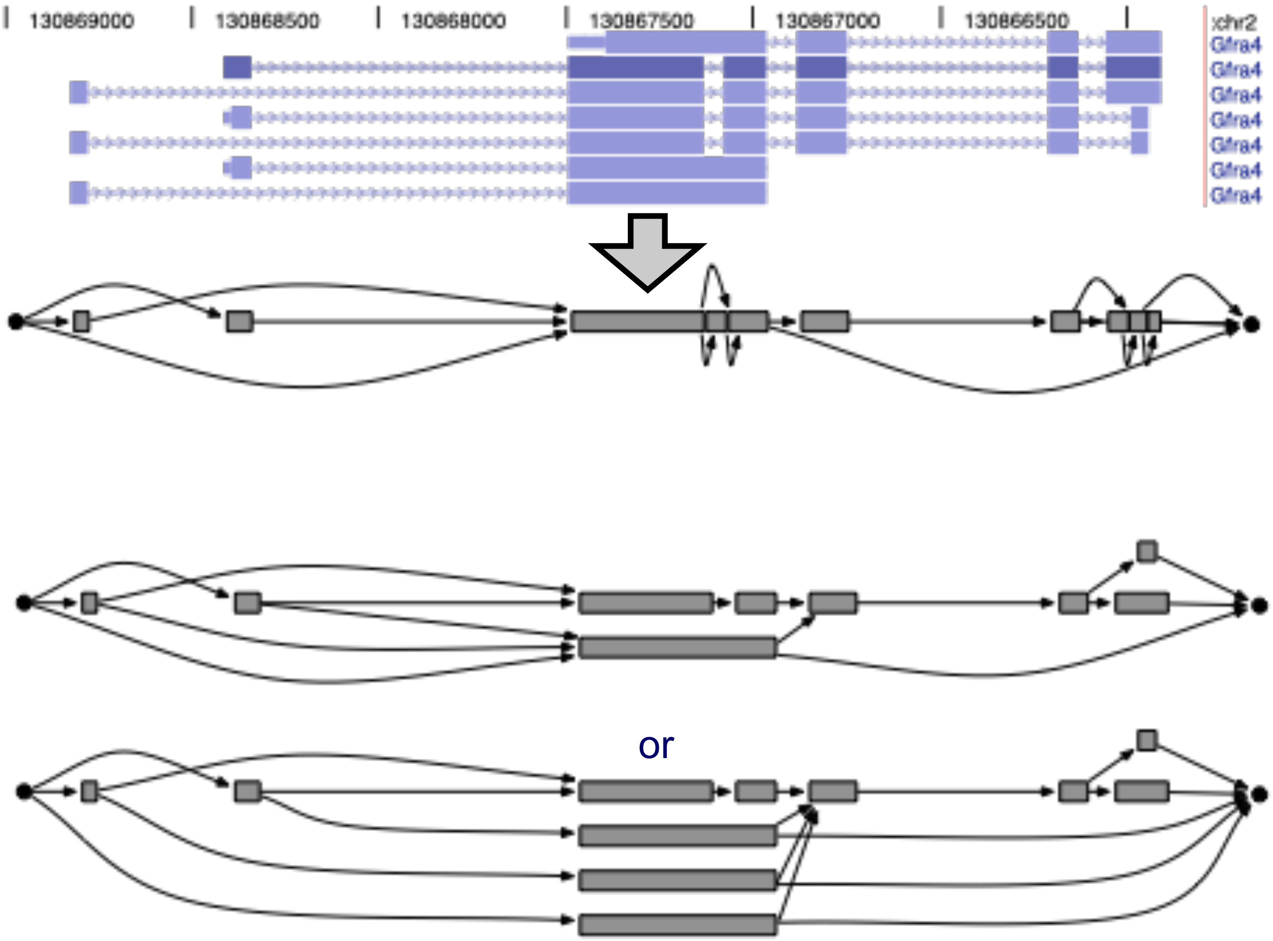


Method	Sample 1	Sample 2	Predicted DP	Recall	Precision
PSG	**A Rep 1**	**A Rep 2**	**4**		
	A Rep 1	B Rep 1	257	0.60	0.95
	A Rep 1	B Rep 2	230	0.54	0.95
	A Rep 2	B Rep 1	251	0.59	0.94
	A Rep 2	B Rep 2	235	0.54	0.93
	B Rep 1	**B Rep 2**	**0**		
Cuffdiff	**A Rep 1**	**A Rep 2**	**379**		
	A Rep 1	B Rep 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