Measuring transcriptomes with RNA-Seq

BMI/CS 776
www.biostat.wisc.edu/bmi776/
Spring 2012

Colin Dewey
cdewey@biostat.wisc.edu

Overview

- RNA-Seq technology

- The RNA-Seq quantification problem

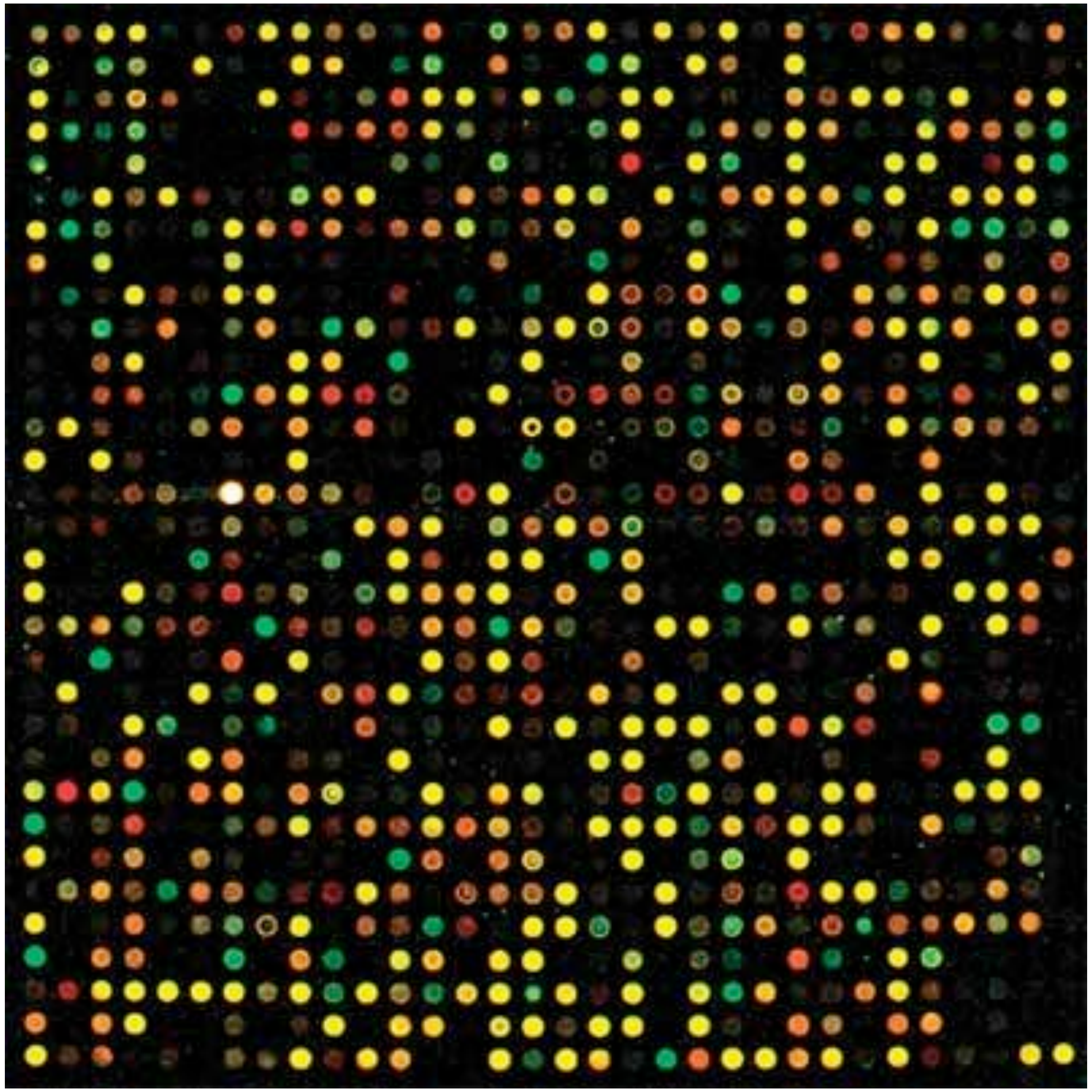
- Generative probabilistic models and Expectation-Maximization for the quantification task

- Probabilistic splice graph models for analysis of alternative splicing

What I want you to get from this 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

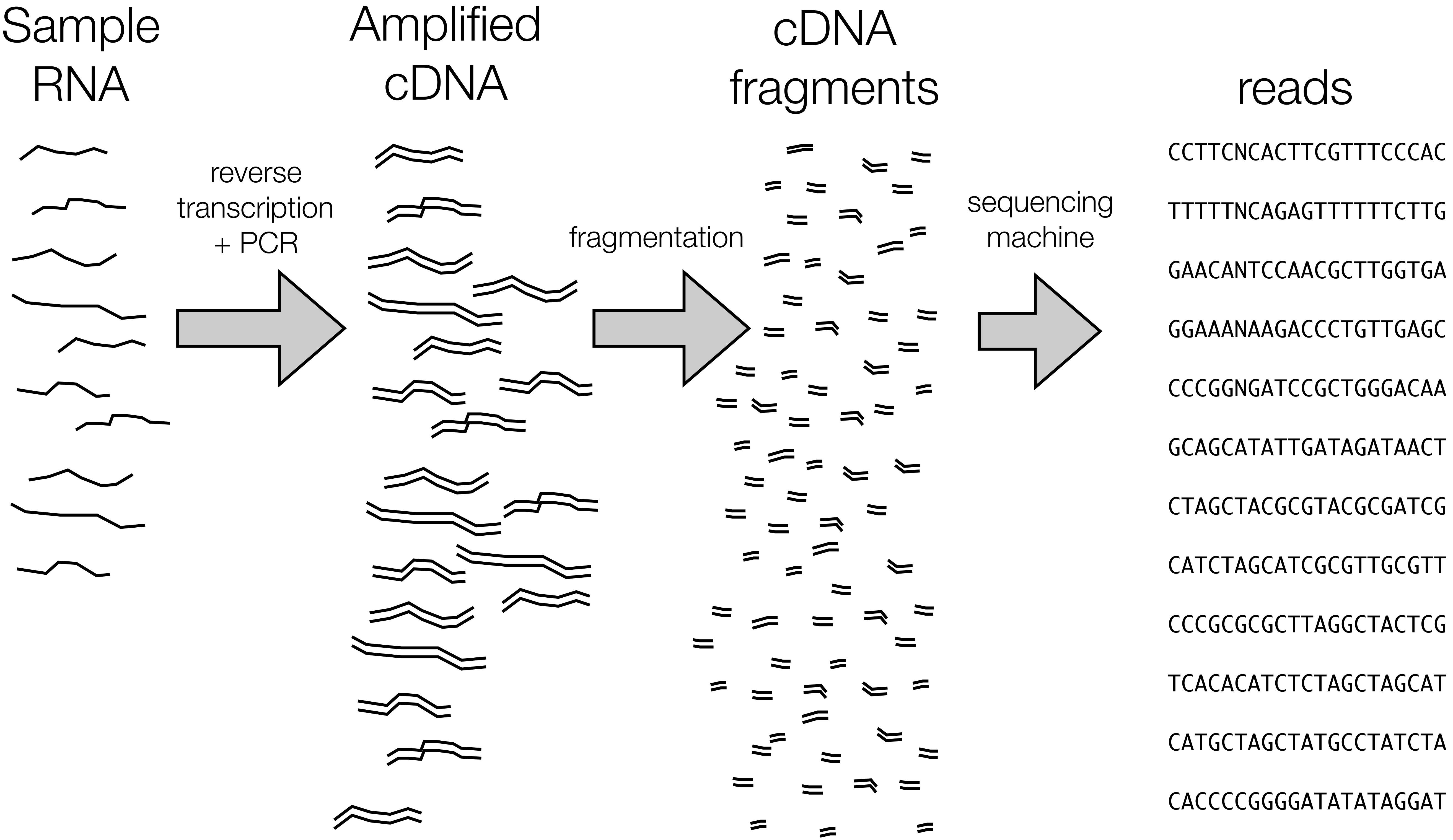


- Each spot has “probes” for a certain gene
- Probe: a DNA sequence complementary to a certain gene
- Relies on complementary hybridization
- Intensity/color of light from each spot is measurement of the number of transcripts for a certain gene in a sample
- Requires knowledge of gene sequences

RNA-Seq technology

- Leverages rapidly advancing sequencing technology (e.g., Illumina, SOLiD)
- 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

RNA-Seq protocol



Sample RNA

reverse transcription + PCR

Amplified cDNA

fragmentation

cDNA fragments

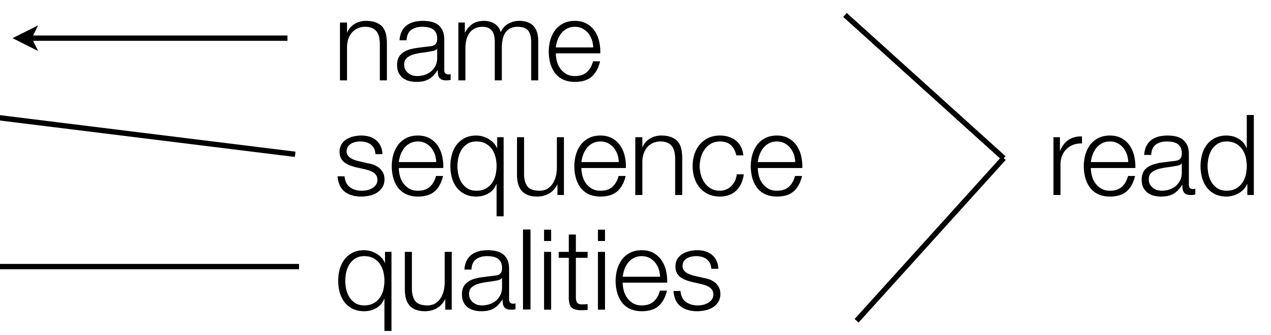
sequencing machine

reads

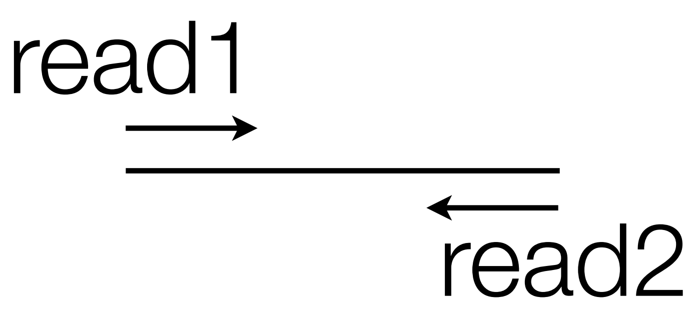
```
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA
CACCCCGGGGATATATAGGAT
```

RNA-Seq data

```
@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_]XdccdfbbL_\`a\aY_^]LZ^
@HWUSI-EAS1789_0001:3:2:5000:1304#0/1
CCCGGNGATCCGCTGGGACAAGCAGCATATTGATA
+HWUSI-EAS1789_0001:3:2:5000:1304#0/1
aaaaaBeeeeffffehhhhhhggdhhhhahhhadh
```



paired-end reads



1 Illumina (GAIIX) lane

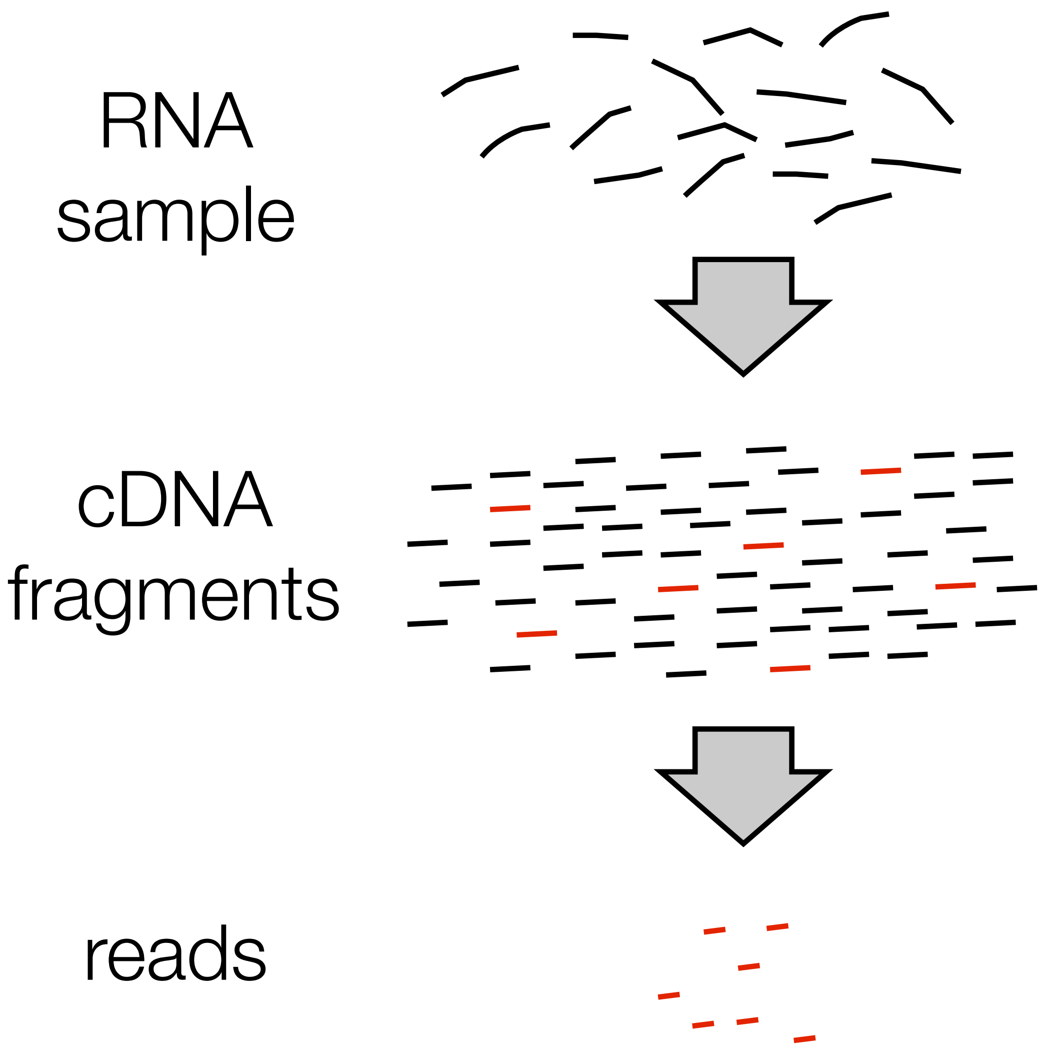
~20 million reads

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

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

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:
  - Given: RNA-Seq reads and transcript sequences
  - Do: Estimate the relative abundances of transcripts (“gene expression”)
- Differential expression:
  - Given: RNA-Seq reads from two different samples and transcript sequences
  - Do: Predict which transcripts have different abundances between the two samples

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 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

$$\hat{\theta}_i = \frac{c_i \longleftarrow \text{\# reads mapping to transcript i}}{N \longleftarrow \text{total \# of mappable reads}}$$

- 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

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

- TPM - Transcripts Per Million

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

- Prefer TPM to RPKM/FPKM 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

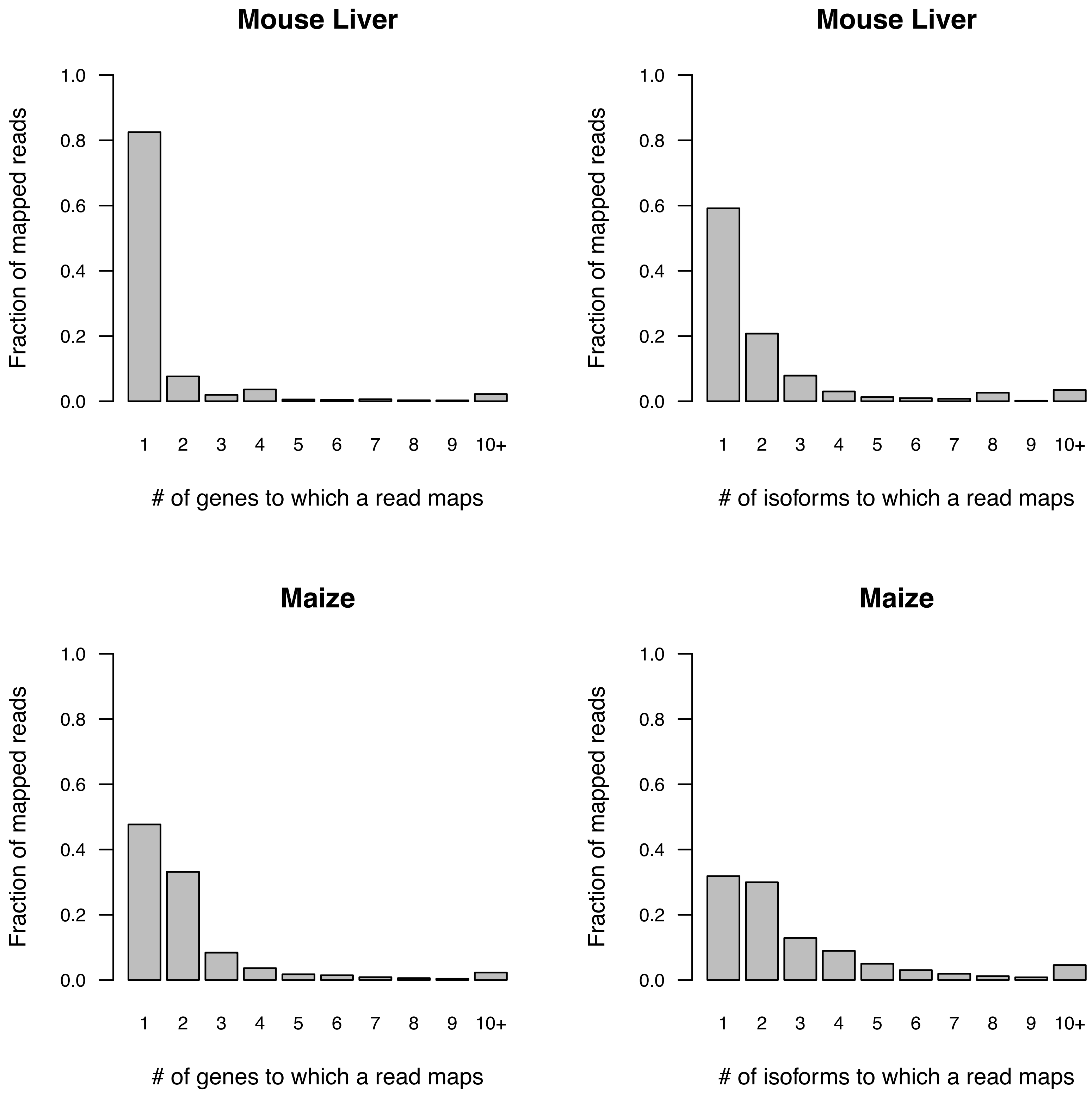
Are multireads really a problem?

Data set	% unmapped	% unique	% multireads	% filtered
Mouse liver (Mortazavi et al. 2008)	46.2	44.4	9.2	0.2
Maize simulation	47.5	25.0	27.1	0.4

25 base reads, 2 mismatches allowed

- Still an issue with longer and paired reads
    - mouse 75 base reads: 10% multireads (single-end), 8% (paired-end)
- Multireads arise due to homology, not chance similarity

Distributions of alignment counts



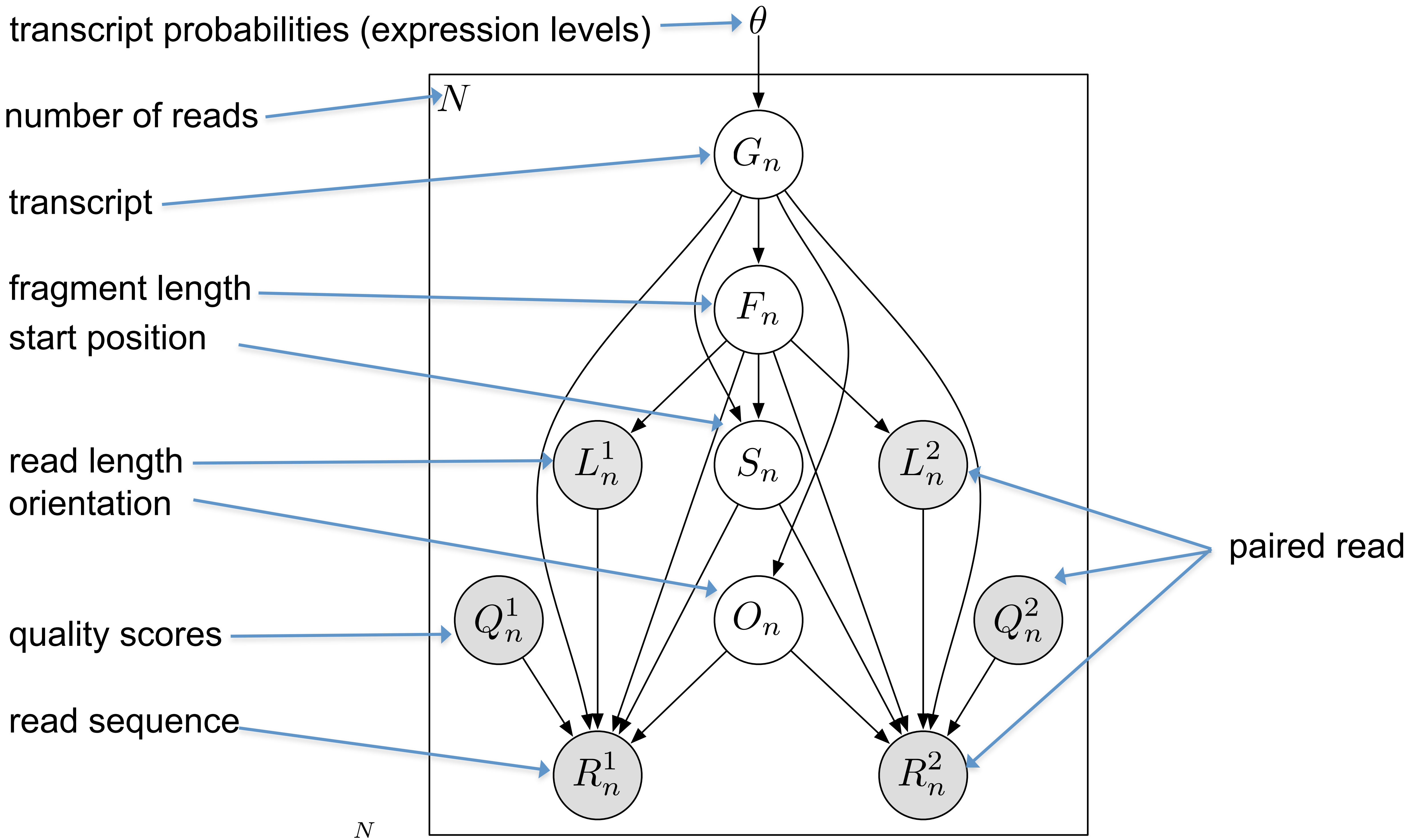
Some options for handling multireads

- Discard all 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

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

Our solution - 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

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

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 w.r.t. 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

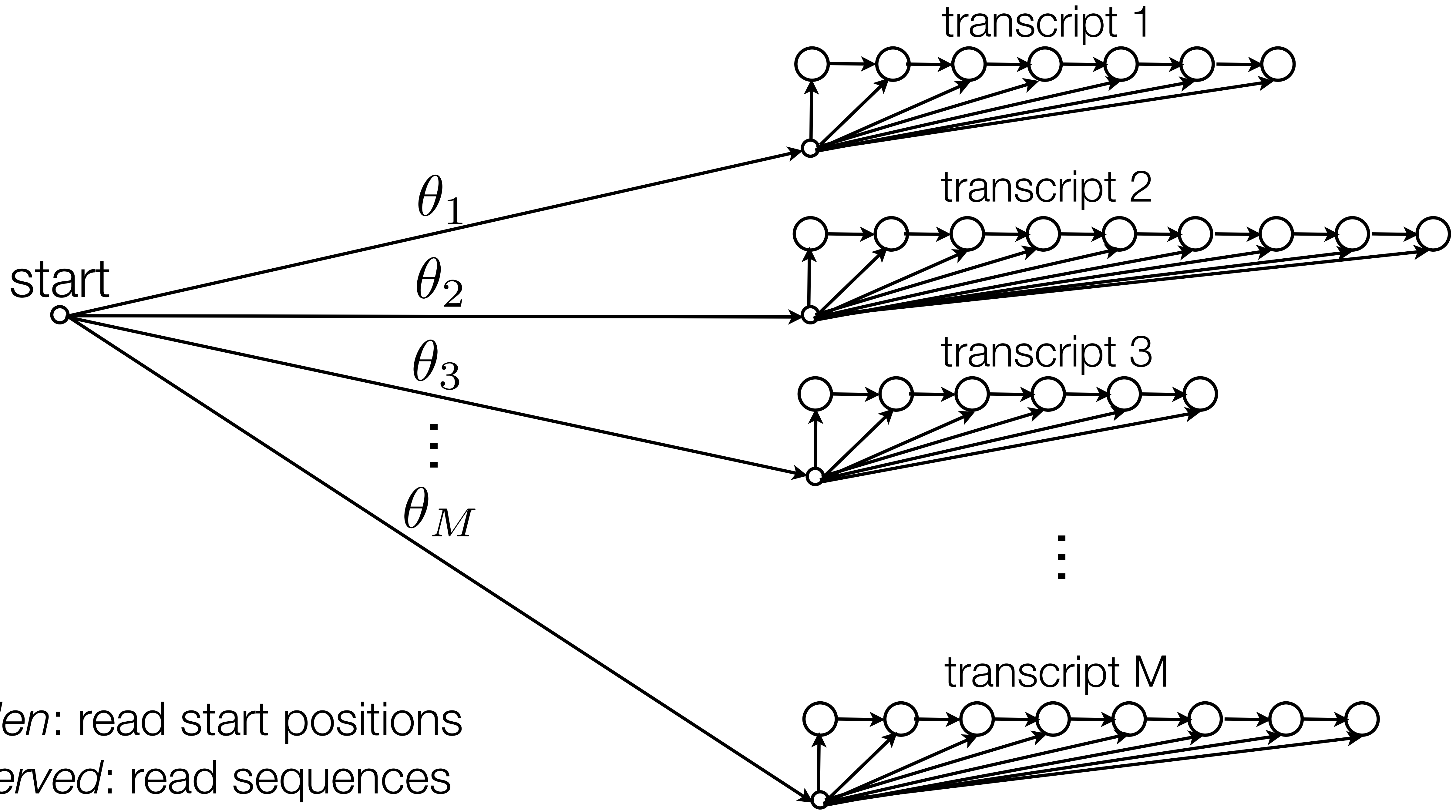
$$P(\mathbf{r}, \ell, \mathbf{q} | \theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \theta_i \sum_{j=0}^{L_i} \sum_{k=0}^{L_i} \sum_{o=0}^{1} P(R_n = r_n, L_n = \ell_n, Q_n = q_n, S_n = j, F_n = k, O_n = o | G_n = i)$$

- 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

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

all local alignments of read n with at most x mismatches

HMM Interpretation



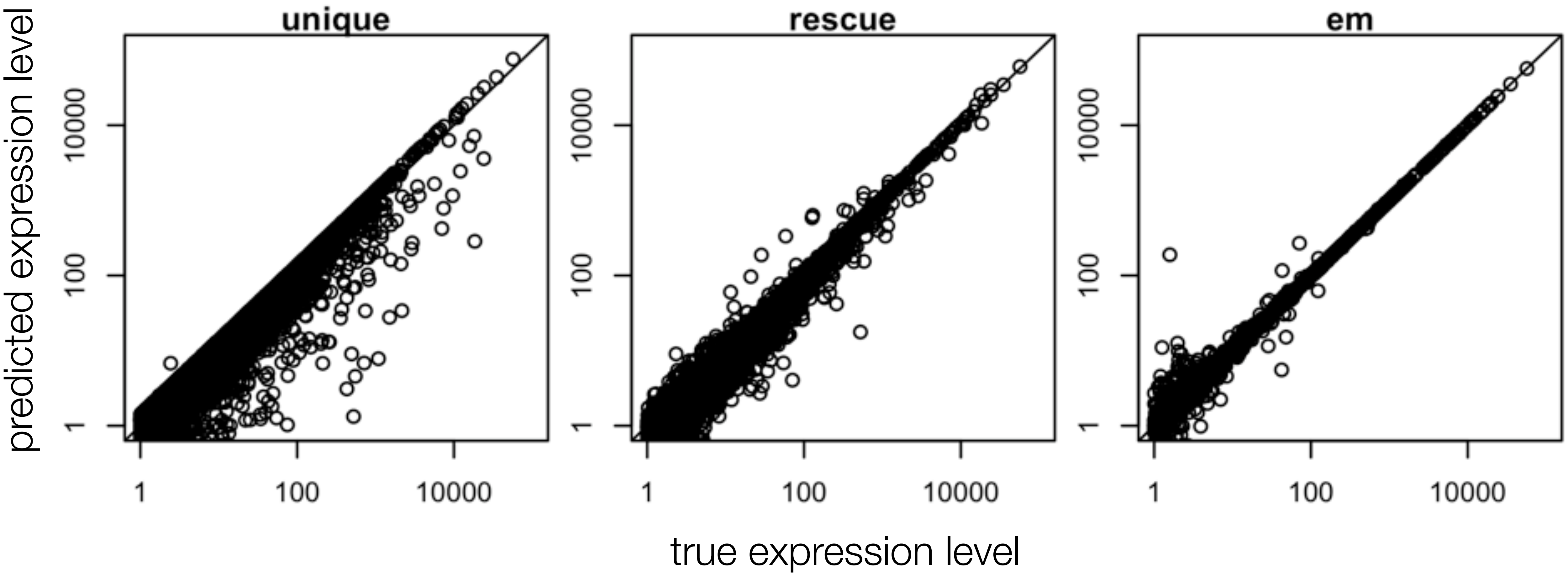
hidden: read start positions
observed: read sequences

Learning parameters: Baum-Welch Algorithm (EM for HMMs)
Approximation: Only consider a subset of paths for each read

EM Algorithm

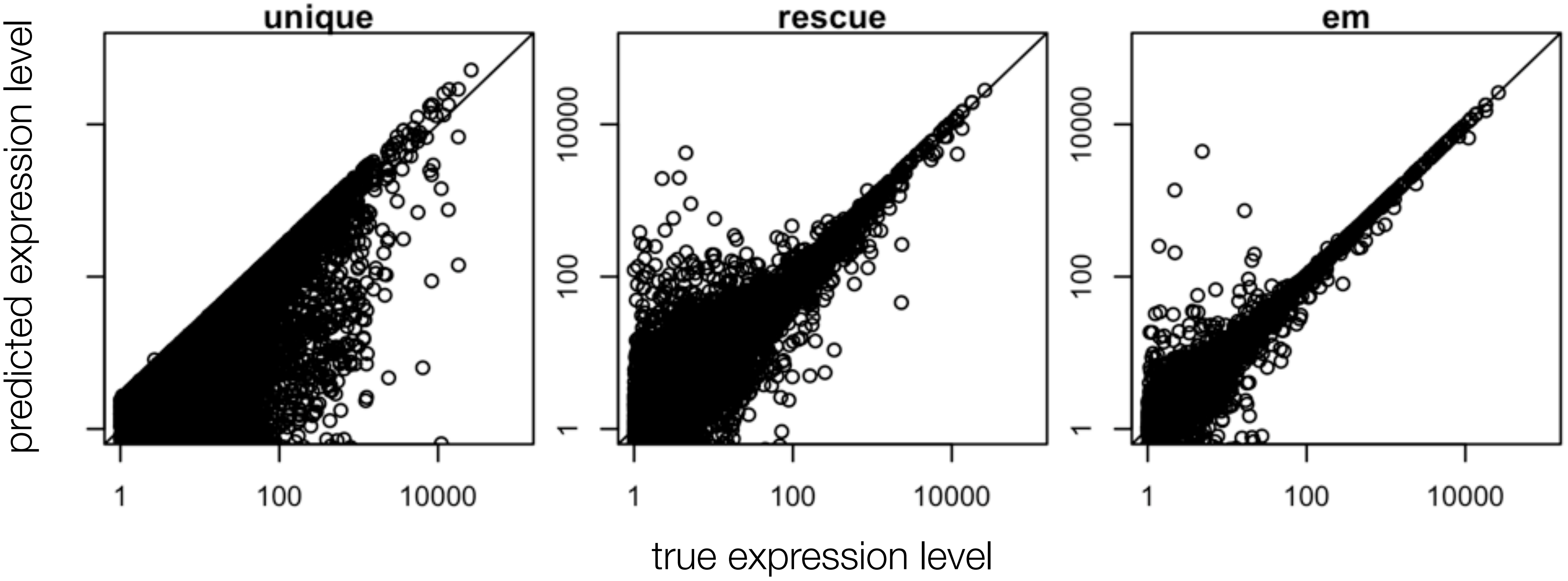
- Expectation-Maximization for RNA-Seq
  - E-step: Compute expected read counts given current expression levels
  - M-step: Compute expression values maximizing likelihood given expected read counts
- Rescue algorithm ≈ 1 iteration of EM

Improved accuracy over unique and rescue



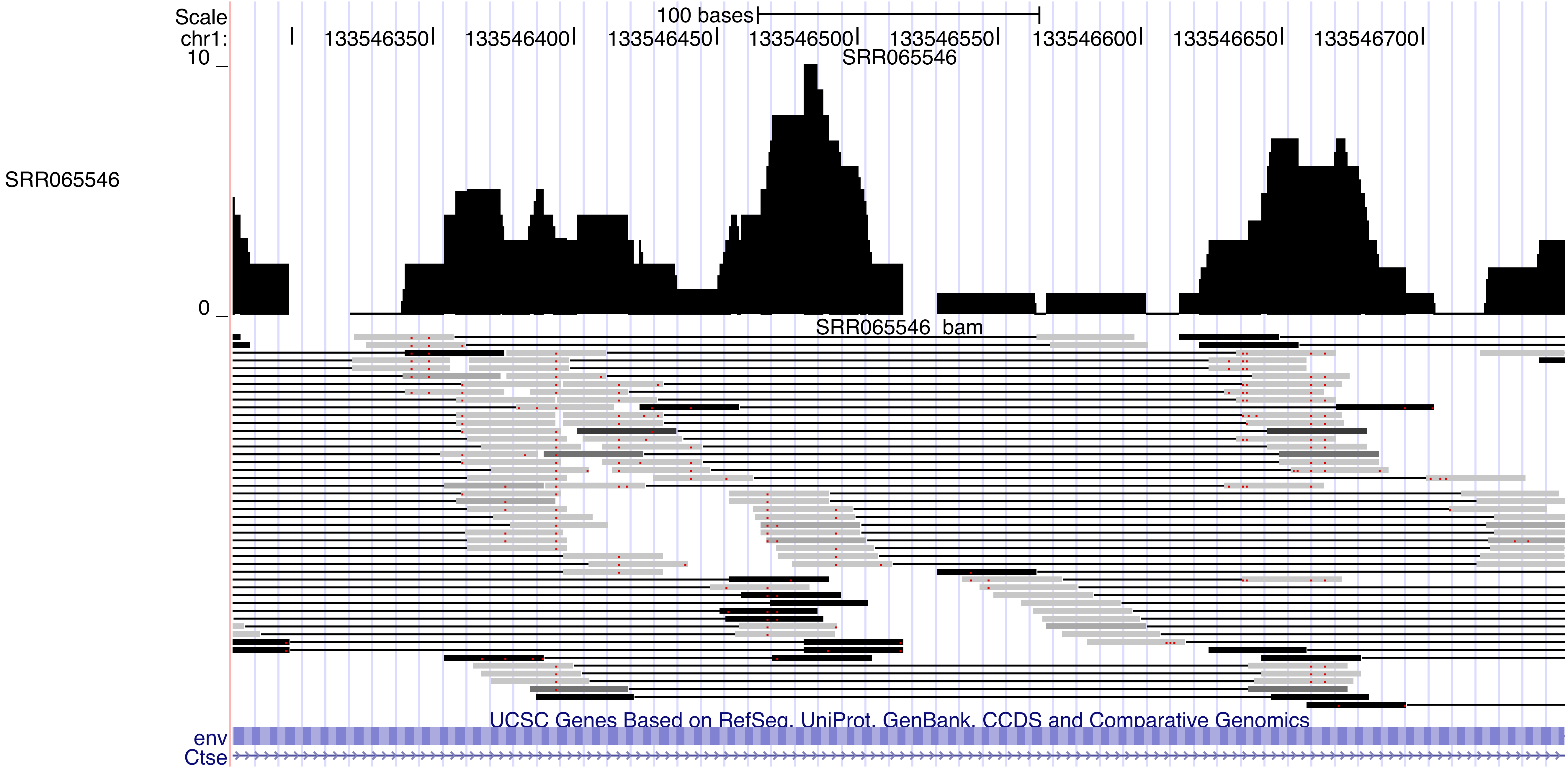
Gene-level expression estimation

Improving accuracy on repetitive genomes: maize

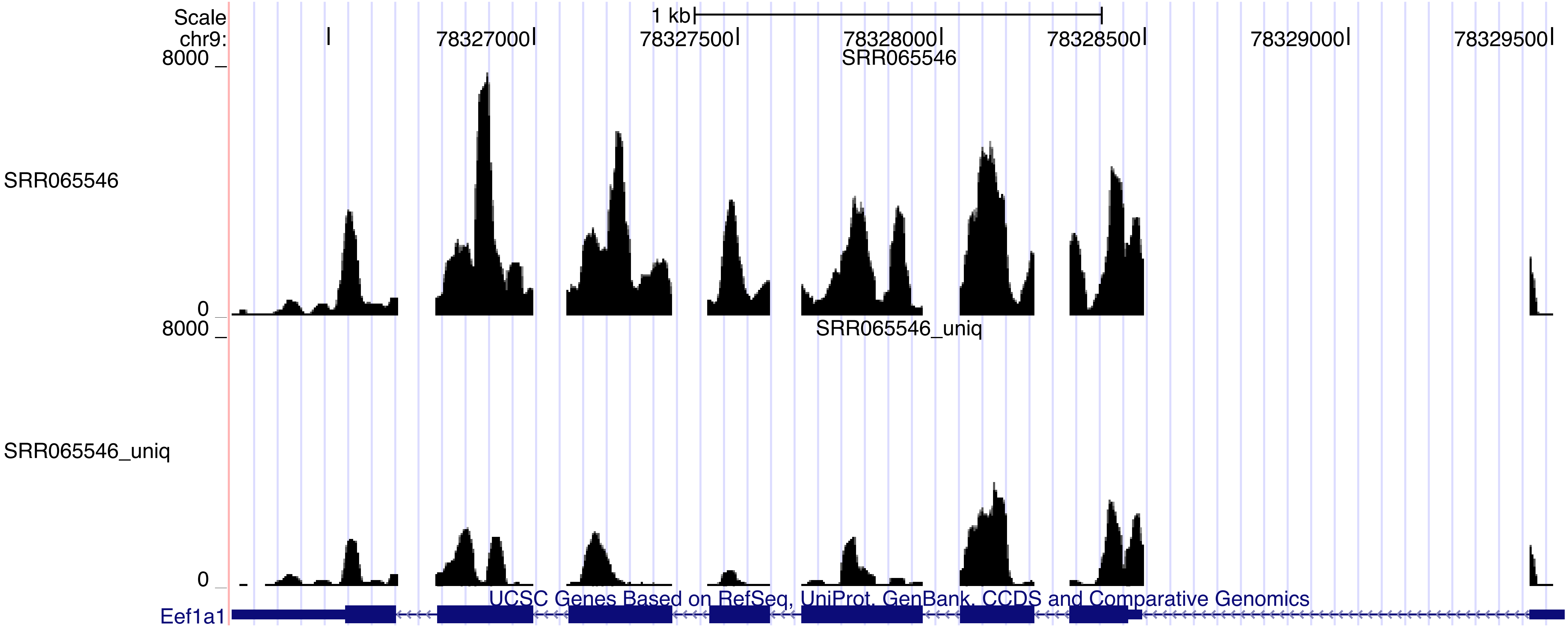


Gene-level expression estimation

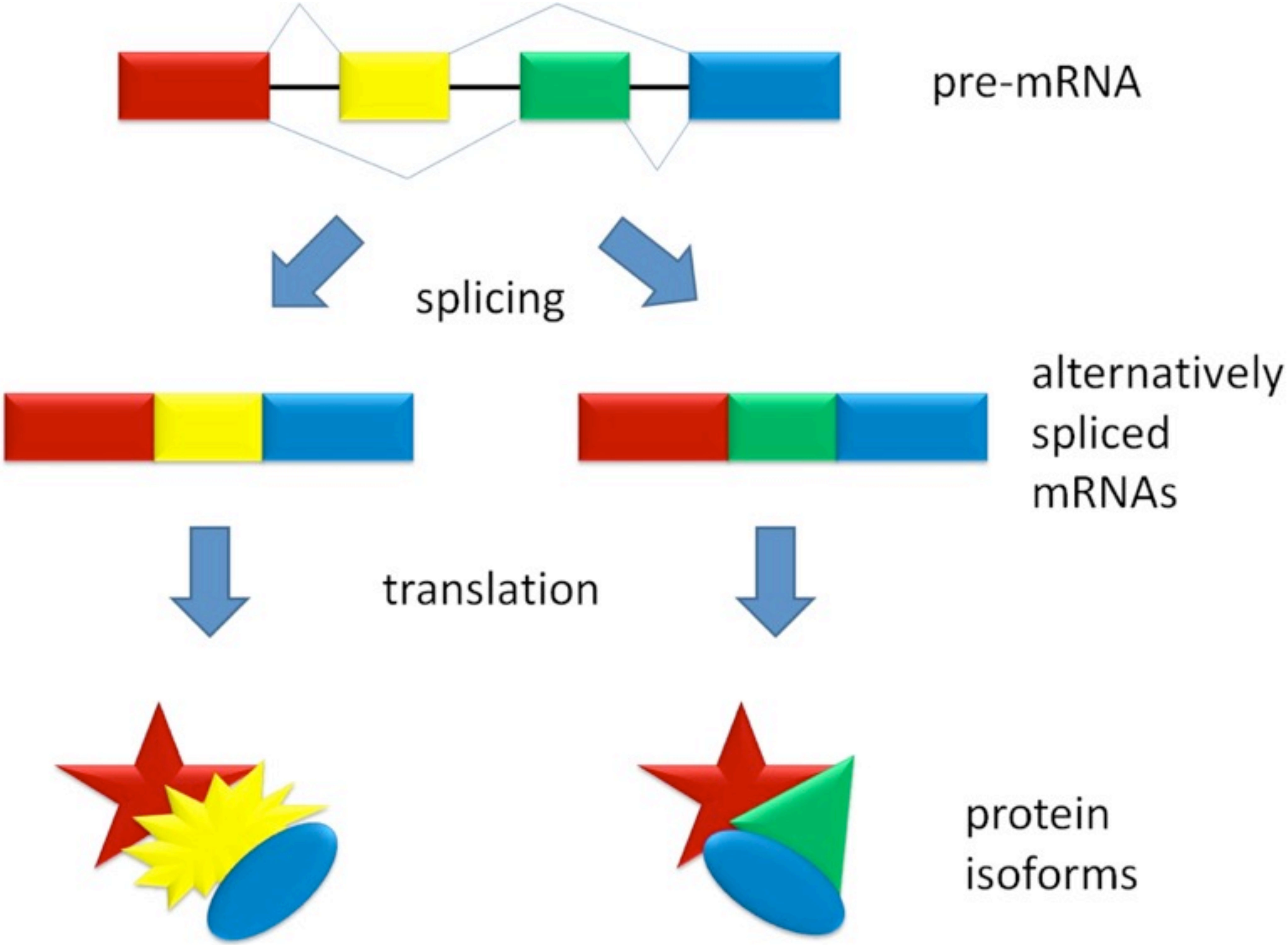
Probabilistically-weighted alignments



Expected read count visualization



Alternative splicing



pre-mRNA

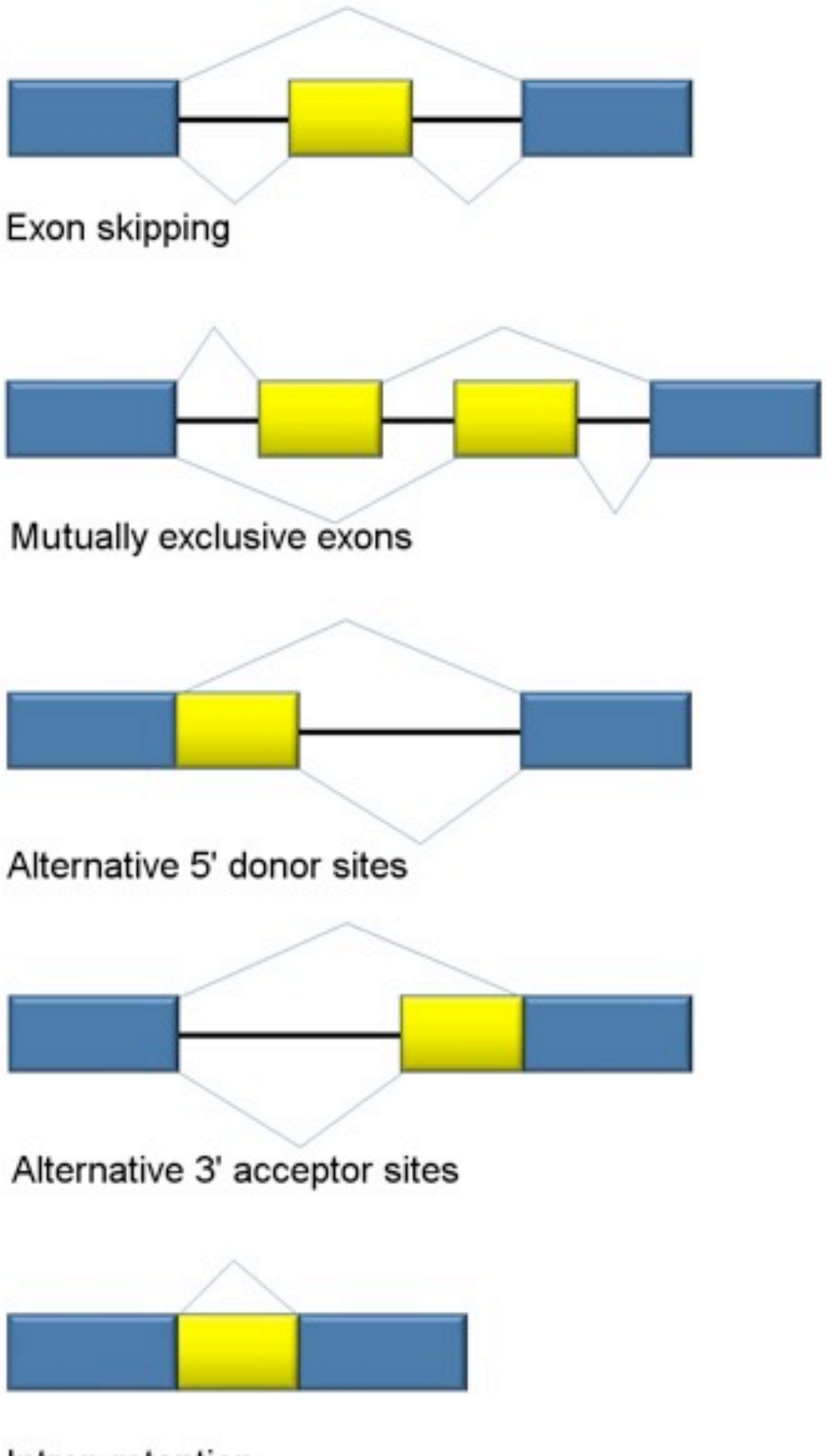
splicing

alternatively spliced mRNAs

translation

protein isoforms

Forms of alternative splicing



Exon skipping

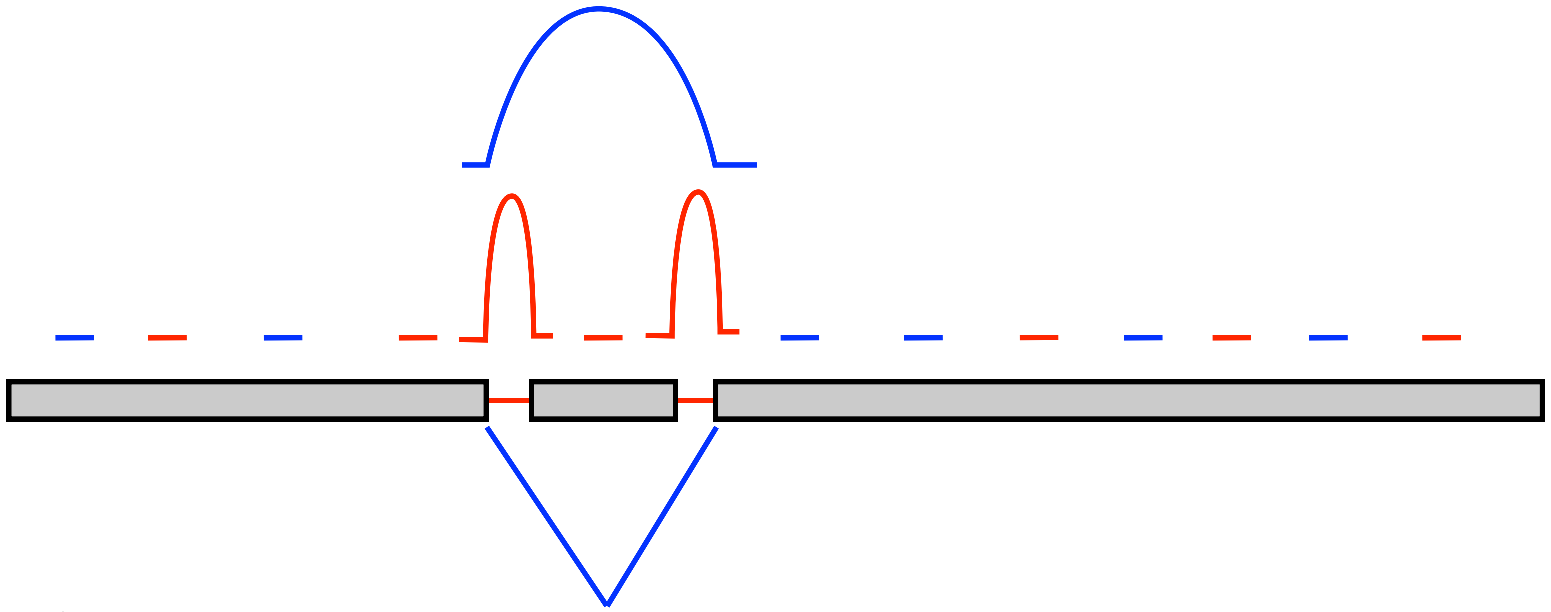
Mutually exclusive exons

Alternative 5' donor sites

Alternative 3' acceptor sites

Intron retention

Alternative splicing analysis with RNA-Seq



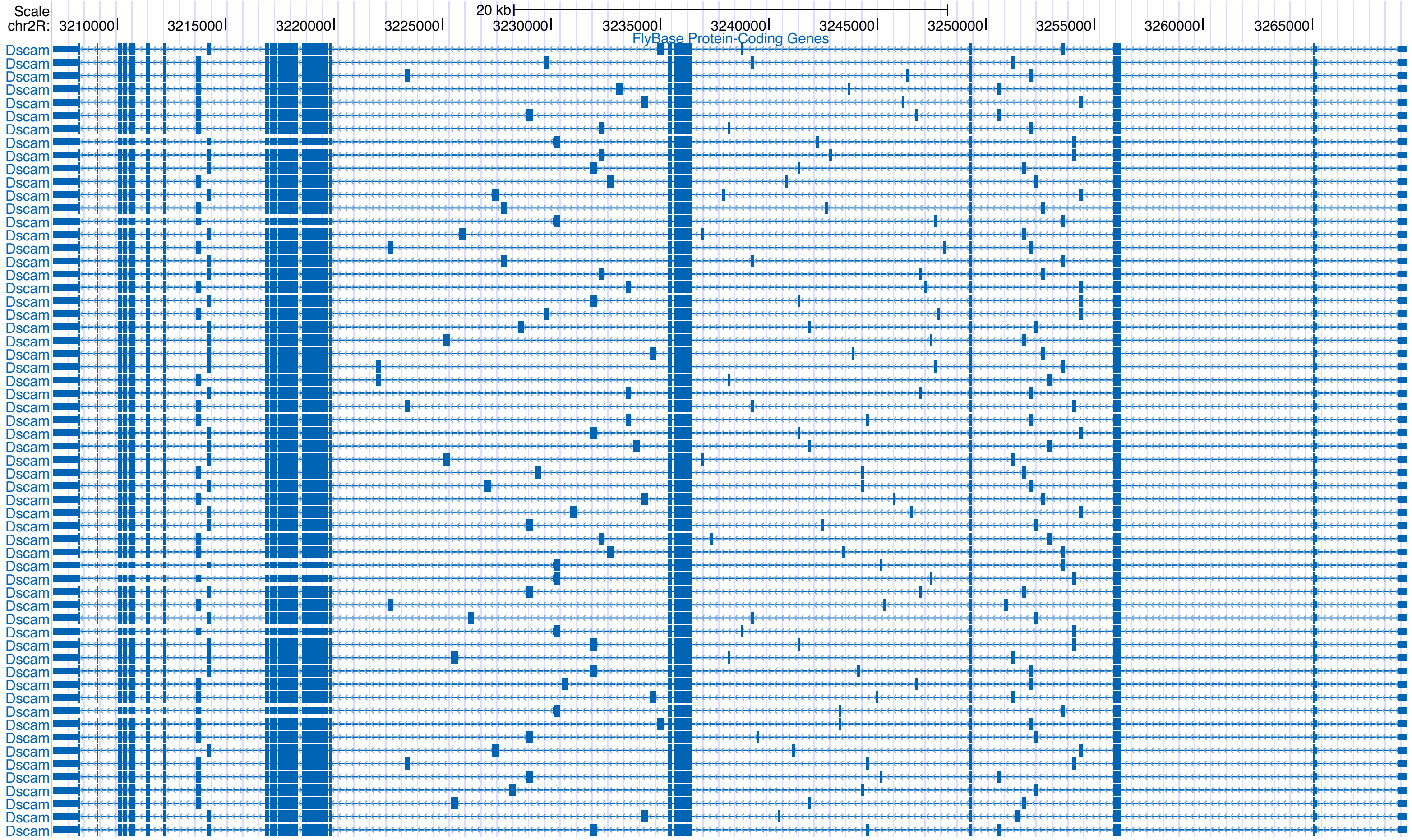
- RNA-Seq: powerful for analyzing alternative splicing
  - Discovery of novel splice junctions
  - Precise quantification of splice events: low background, large dynamic range

- Analysis challenges
  - Genes with many isoforms
  - Non-identifiability of abundances
  - Difficulty in *de novo* assembly of full-length isoforms

Combinatorial explosion of distinct isoforms

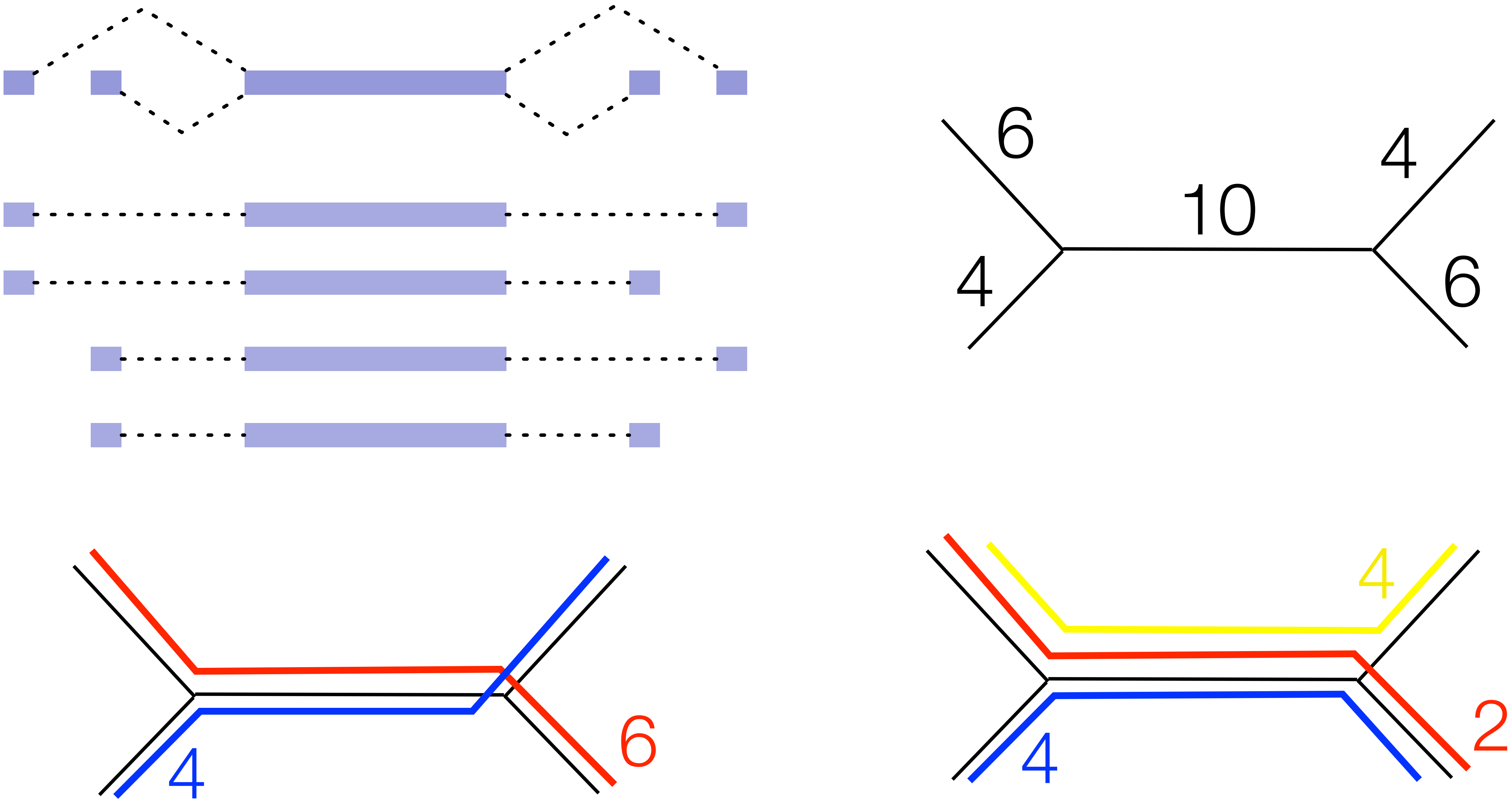
- Combinatorial explosion of the number of possible isoforms for each gene

- Insufficient data to accurately estimate abundances of thousands of isoforms



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

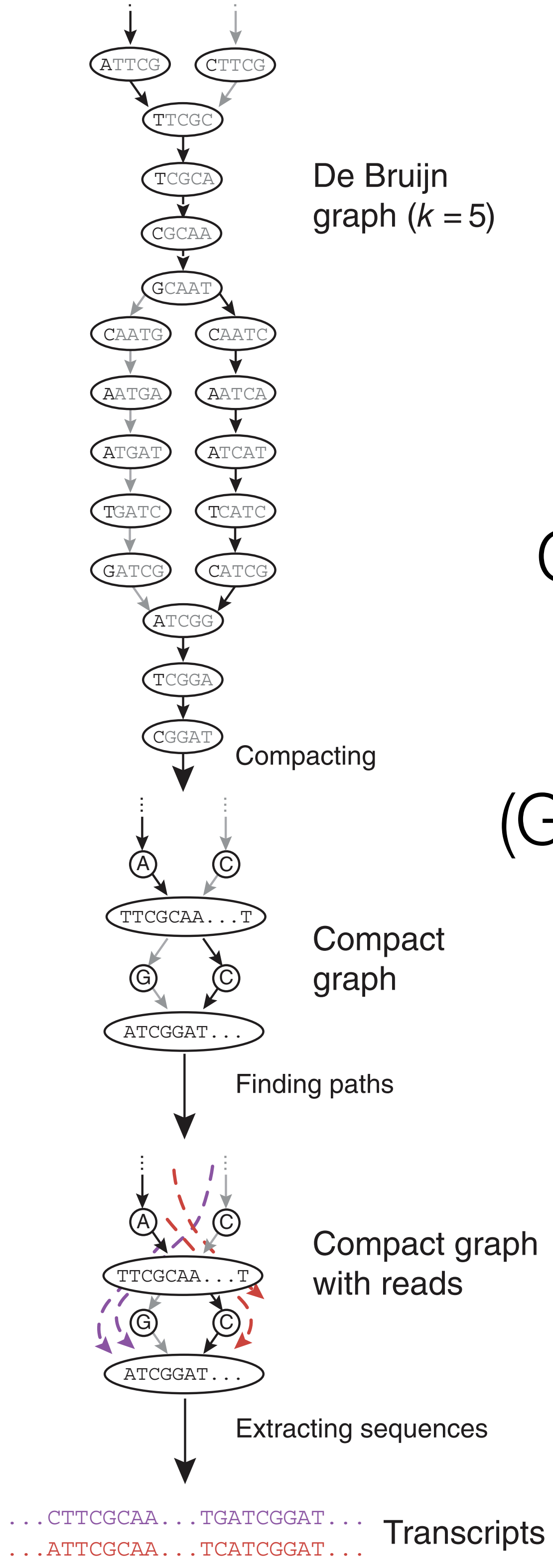
Non-identifiability of full-length isoform models



Lacroix et al. 2008; Hiller et al. 2009

De novo transcriptome assembly

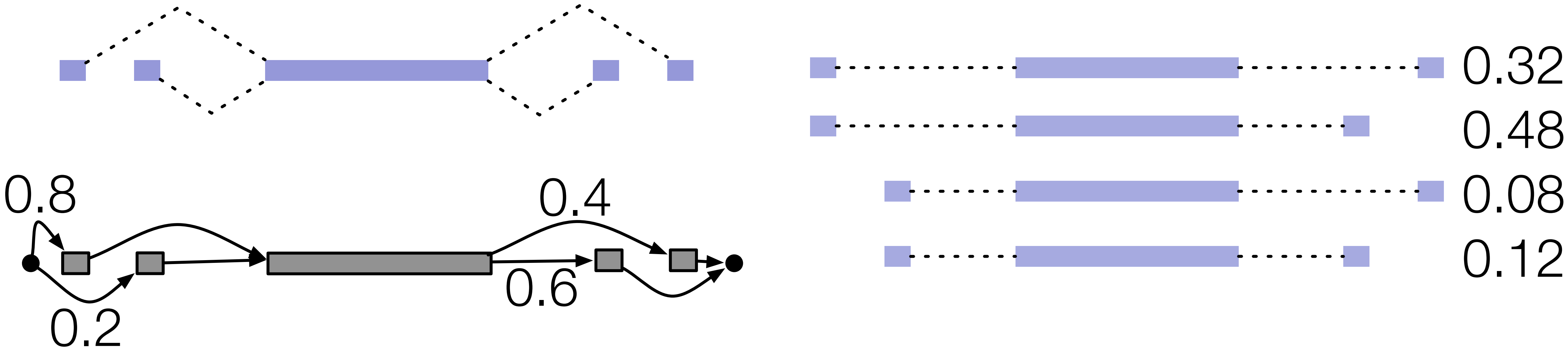
- RNA-Seq reads/fragments are relatively short
- Often insufficient to reconstruct full-length isoforms in the presence of alternative splicing
- Transcriptome assemblies perhaps best left in “graph” form
  - De Bruijn graph
  - String graphs



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

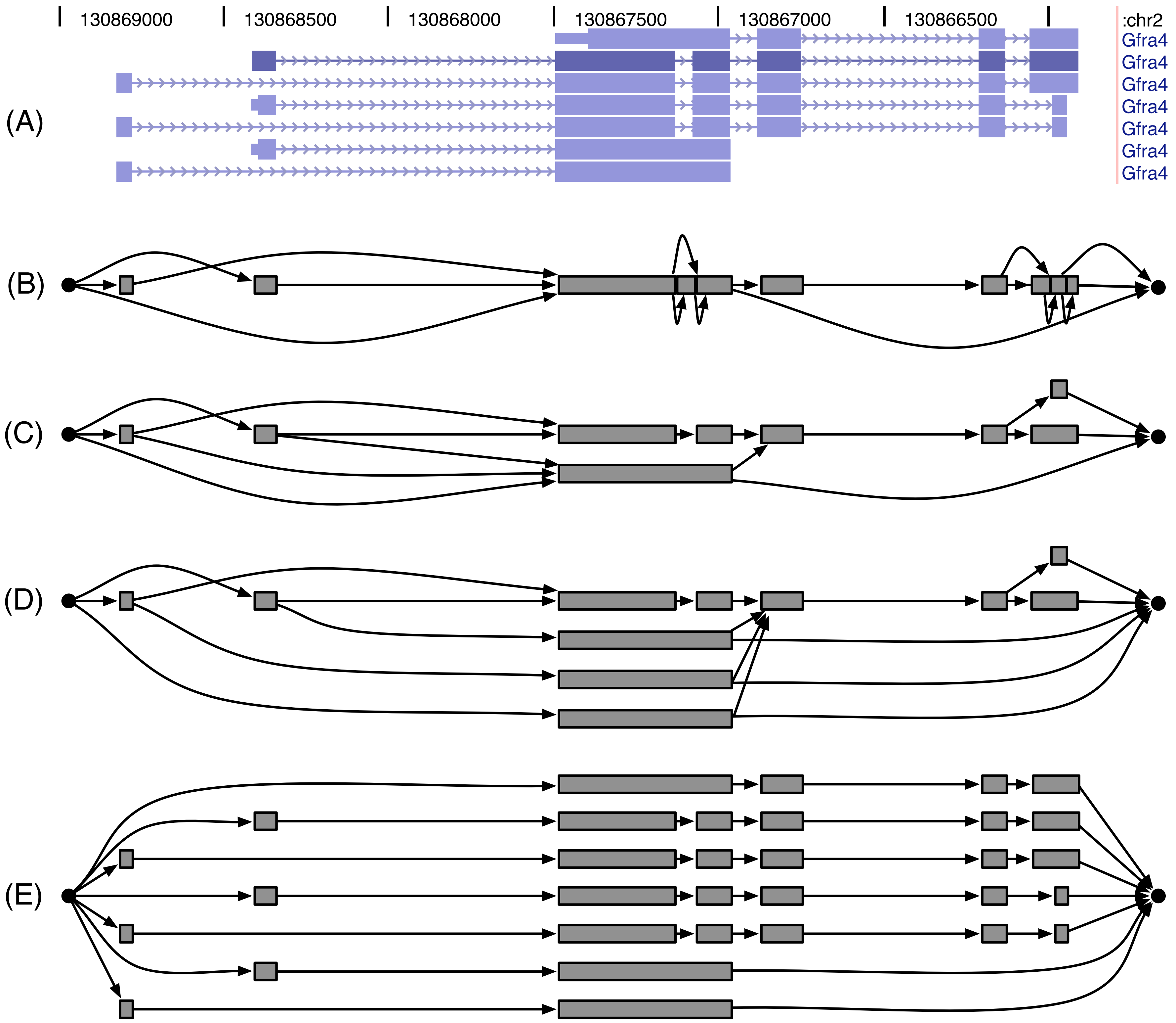
Our solution: Probabilistic Splice Graphs

- Splice Graphs (Heber et al. 2002)
  - Compact representation of possible isoforms for a gene
- Statistical models with splice graphs (Jenkins et al. 2006)
  - Modeling of EST data



L. Legault and C. Dewey. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. *Submitted*.

Probabilistic Splice Graph Complexity



known isoforms

“line graph”

“exon graph”

“higher-order exon graph”

“unfactorized graph”

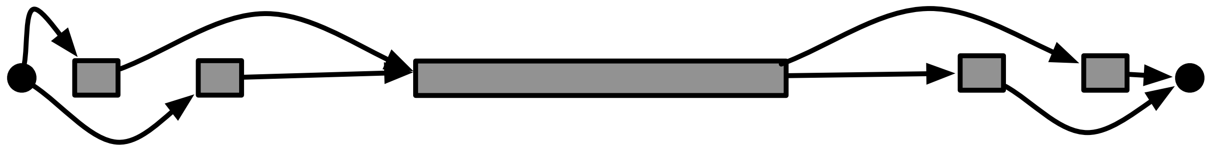
Advantages of PSGs

- 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

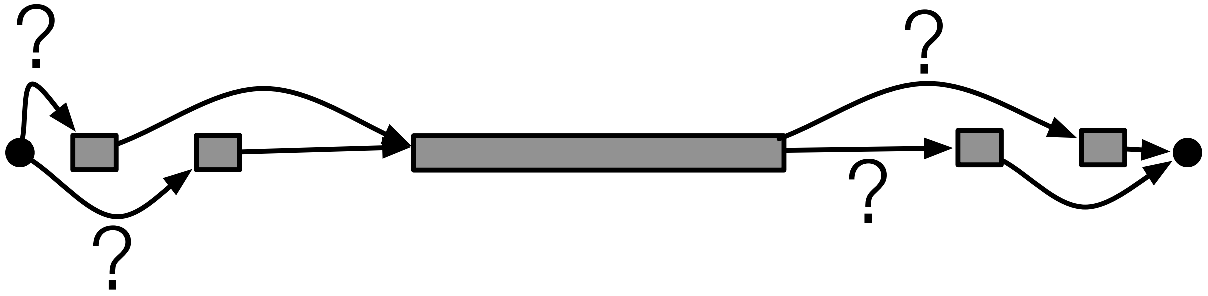
The PSG parameter inference problem

- Given: RNA-Seq reads and a PSG structure

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



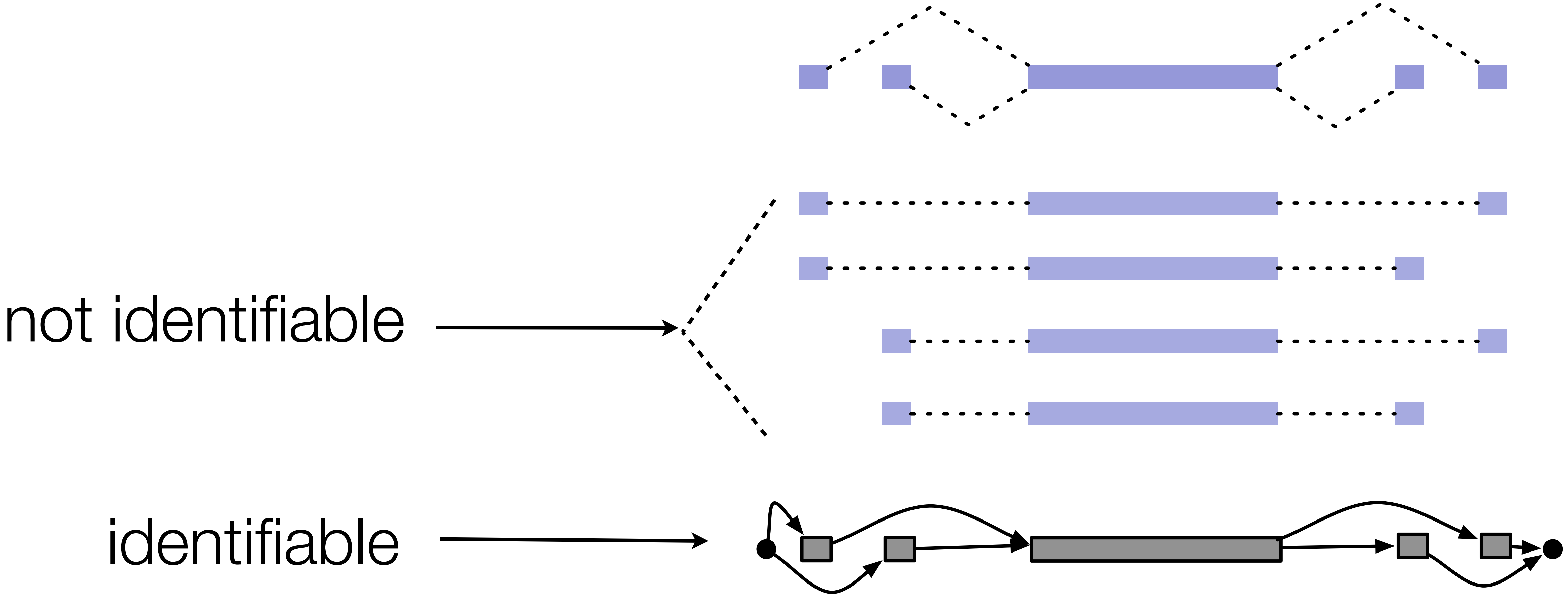
- Do: Estimate the (maximum likelihood) parameters for the model



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

A model of RNA-Seq from PSGs

- RSEM model extended to probabilistic splice graphs
- Efficient inference of parameters (splice event frequencies) with EM
  - 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

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

Expected suffix length

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

EM for PSG parameter estimation

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

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

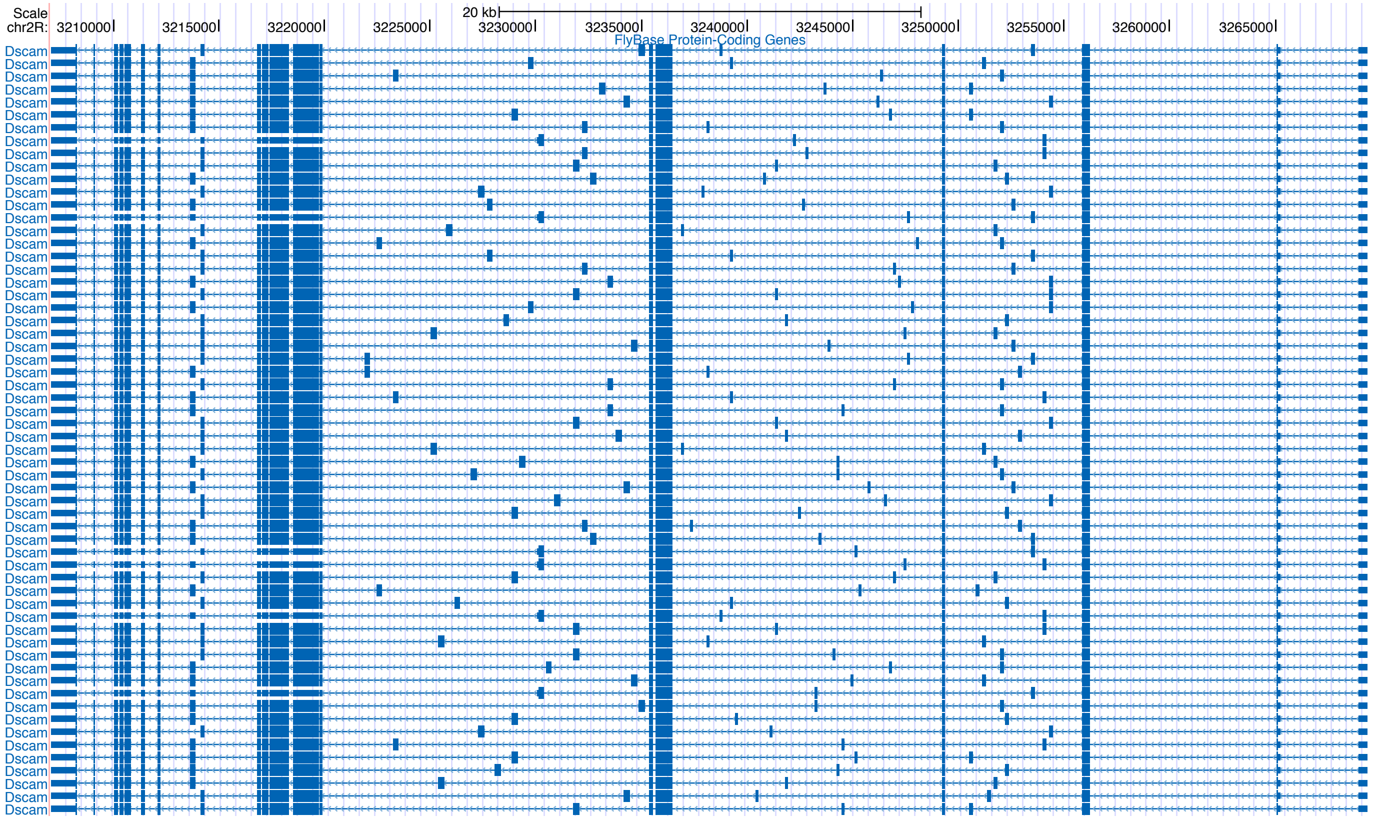
$$\begin{aligned} g(s) &= f(0,s_1)w(s) \\ g(s,i,j) &= \begin{cases} f(0,s_1)w(s) & (i,j)\in s \\ f(0,i)\alpha_{ij}f(j,s_1)w(s) & \text{if } \exists \text{ path from } v_j \text{ to } s_1 \\ f(0,s_1)w(s)f(s_{|s|},i)\alpha_{ij} & \text{if } \exists \text{ path from } s_{|s|} \text{ to } v_i \\ 0 & \text{otherwise}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

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

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

Efficient inference for highly-spliced genes

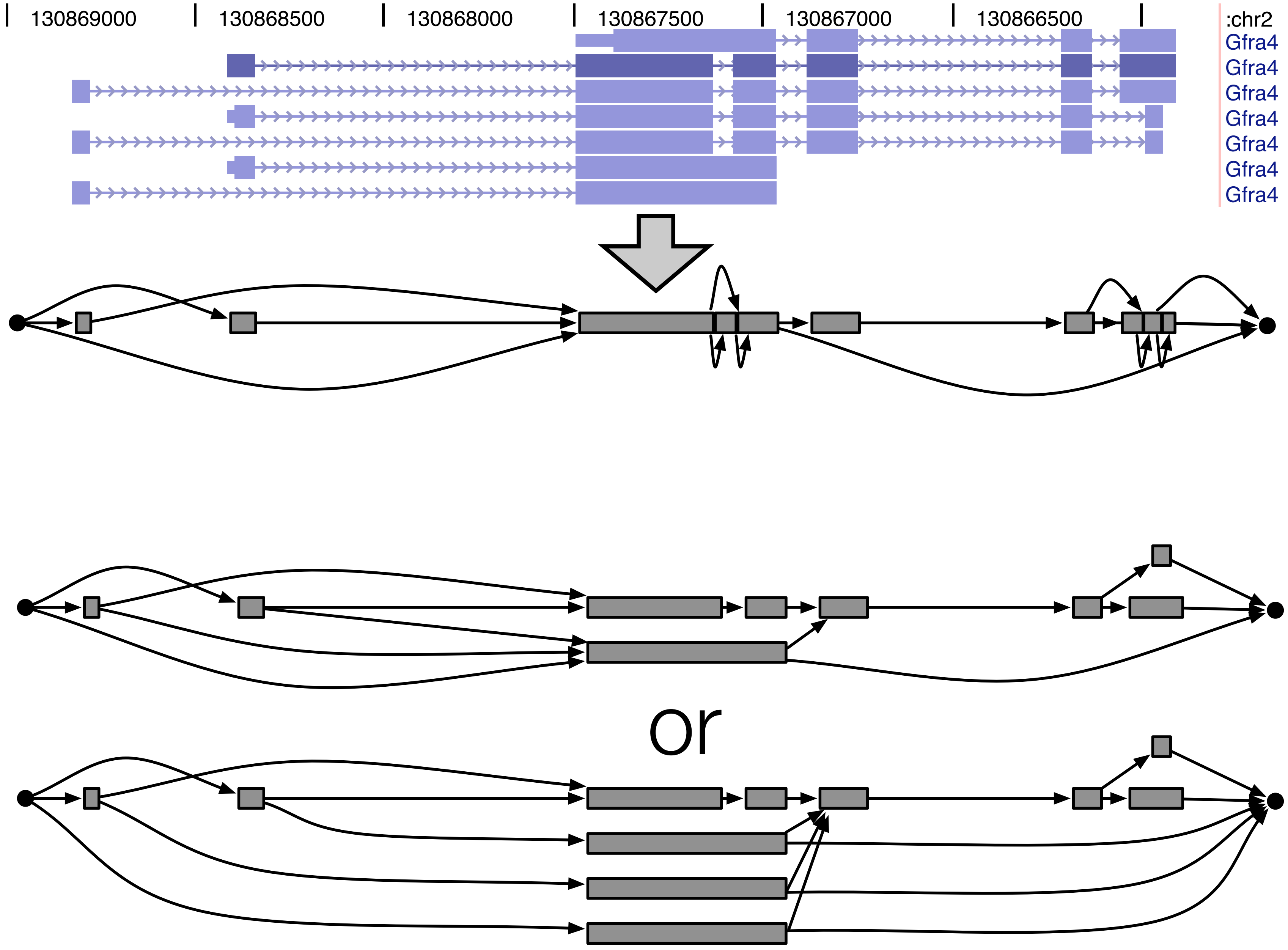
- DSCAM running time test
    - 23,976 isoforms
    - Simulated 10 reads



Method	RSEM	Cufflinks	Line graph PSG
Running time	Not possible	> 15 hours (> 50 GB RAM)	< 1 second

Next steps for modeling RNA-Seq with PSGs

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



Summary

- RNA-Seq is likely the future of transcriptome analysis

- 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

- Alternative splicing complicates matters further

- Probabilistic splice graphs are compact and efficient models for RNA-Seq data with alternatively spliced genes (dynamic programming!)