Measuring transcriptomes with RNA-Seq

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2019

Colin Dewey

colin.dewey@wisc.edu

Overview

- RNA-Seq technology

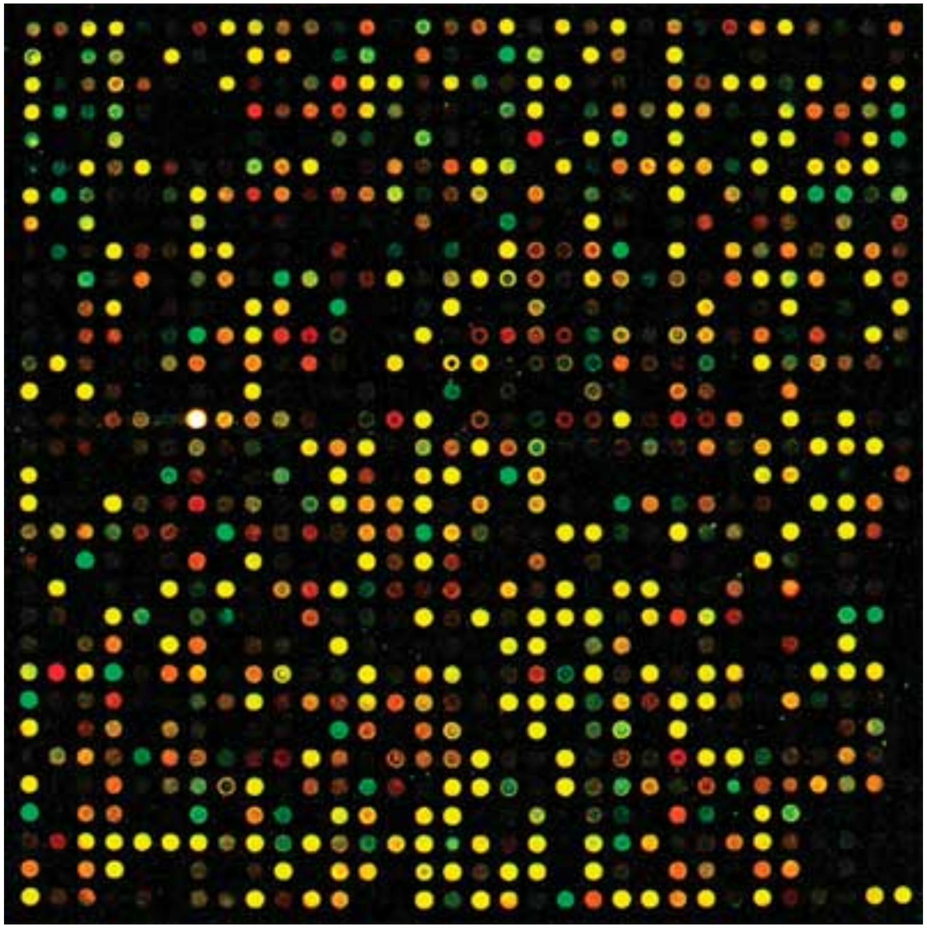
- The RNA-Seq quantification problem

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

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

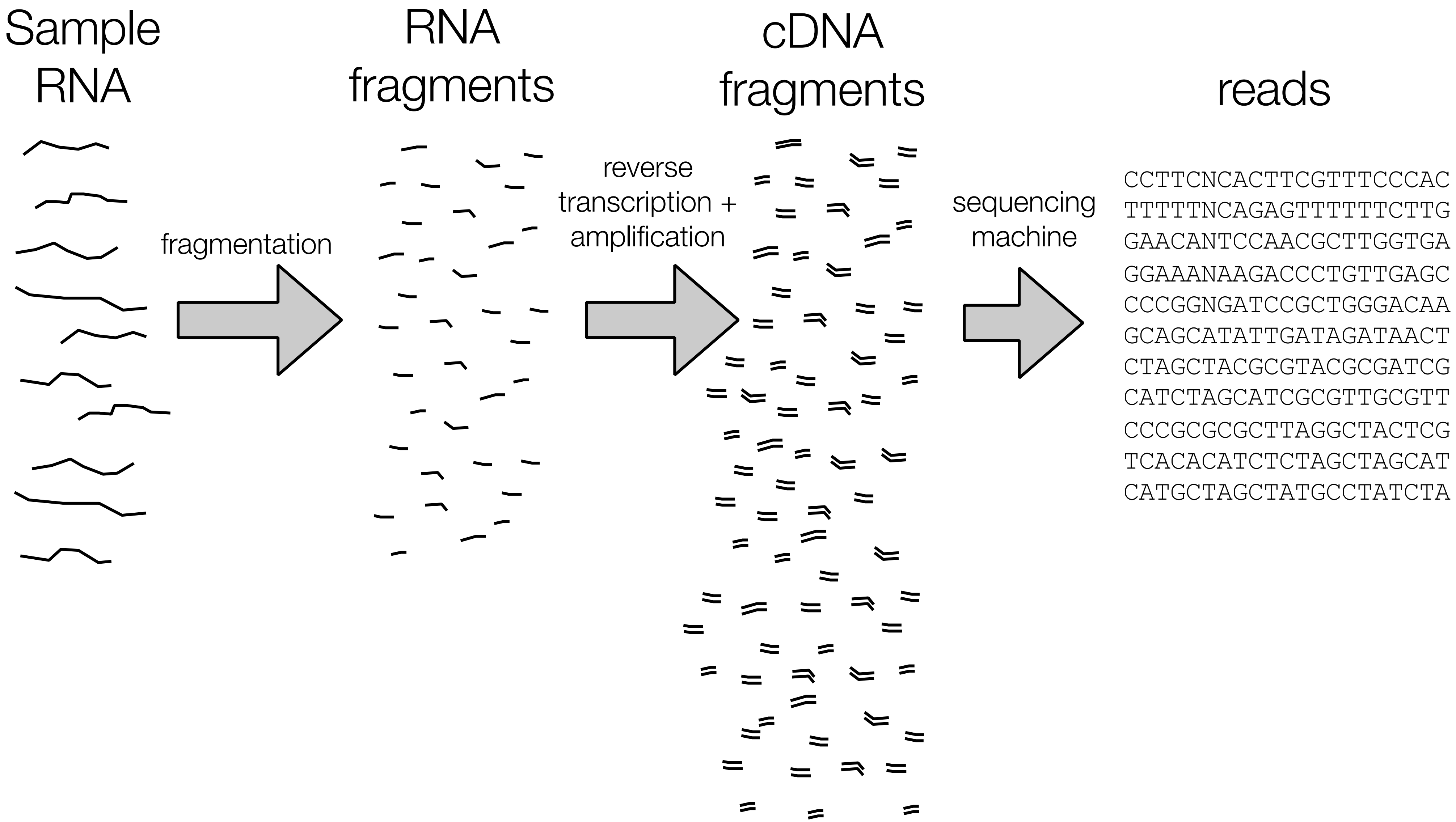
Advantages of RNA-Seq over microarrays

- 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

RNA fragments

cDNA fragments

reads

fragmentation

reverse transcription + amplification

sequencing machine

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

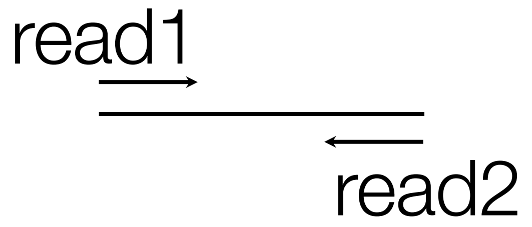
RNA-Seq data: FASTQ format

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

name
sequence
qualities

read

paired-end reads



1 Illumina HiSeq 2500 lane

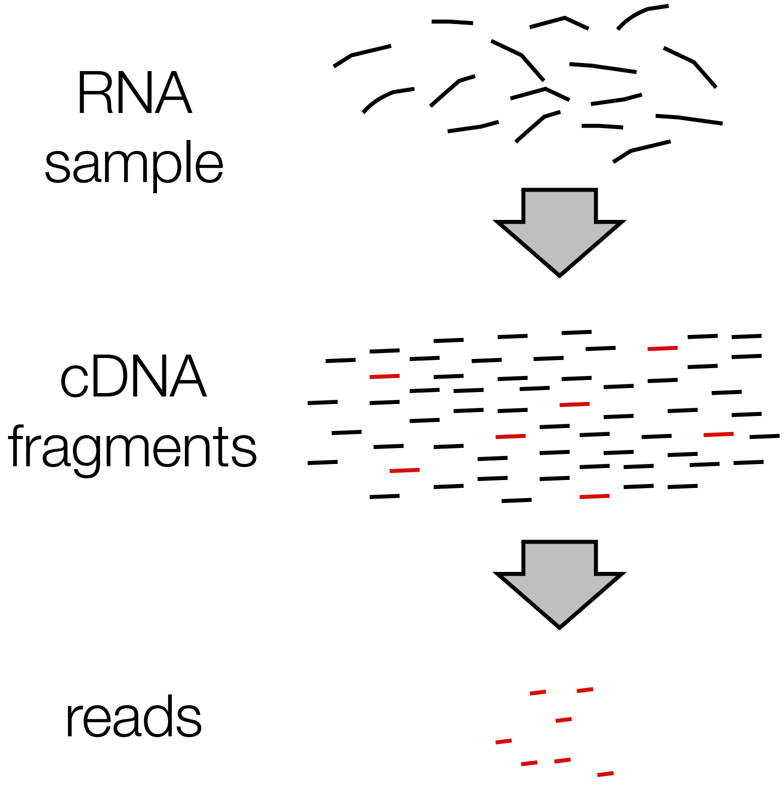
~150 million reads

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:**
  - 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

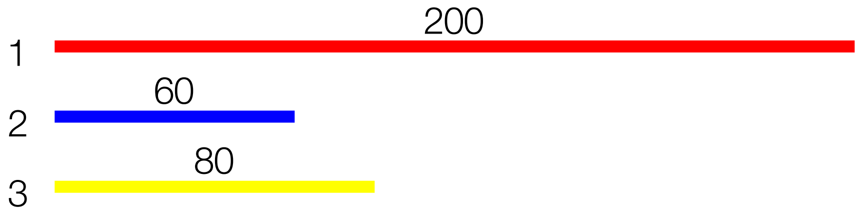
Gene	Sample 1 absolute abundance	Sample 1 relative abundance	Sample 2 absolute abundance	Sample 2 relative abundance
1	20	10%	20	5%
2	20	10%	20	5%
3	20	10%	20	5%
4	20	10%	20	5%
5	20	10%	20	5%
6	100	50%	300	75%

- 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

100 A

60 C

40 G

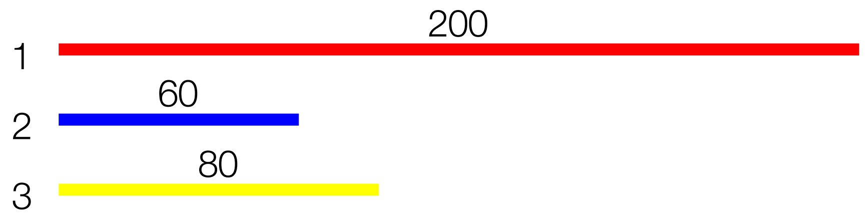
- What relative abundances would you estimate for these genes?

- 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

<u>transcripts</u>



<u>reads</u>

100 A

60 C

40 G

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

transcript 1 relative abundance

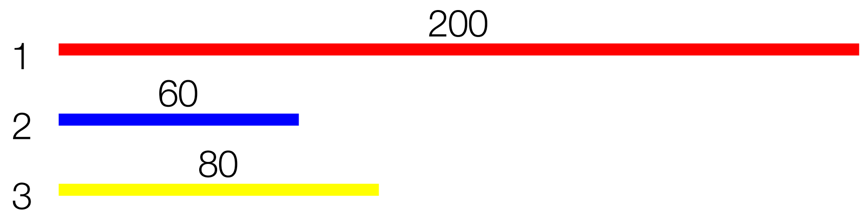
$$\hat{f}_1 \propto \frac{\frac{100}{200}}{200} = \frac{1}{400}$$

transcript 1 length

Length dependence

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

<u>transcripts</u>



<u>reads</u>

100 A

60 C

40 G

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

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

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

$$\xrightarrow{\text{normalize}}$$

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

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

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

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

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

- 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?

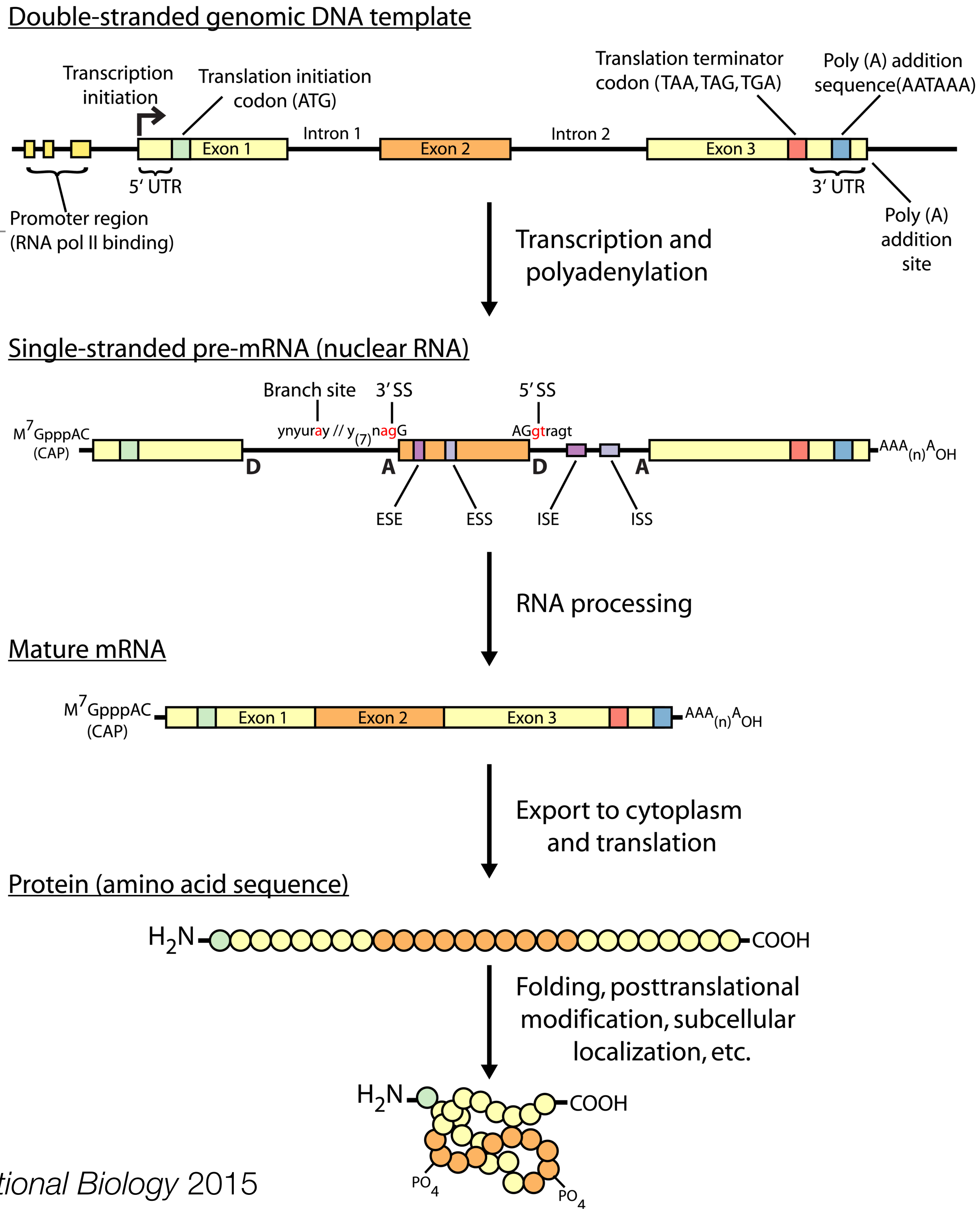
- The approach described assumes that every read can be uniquely aligned to a single transcript

- This is generally not the case

  - Some genes have similar sequences - gene families, repetitive sequences

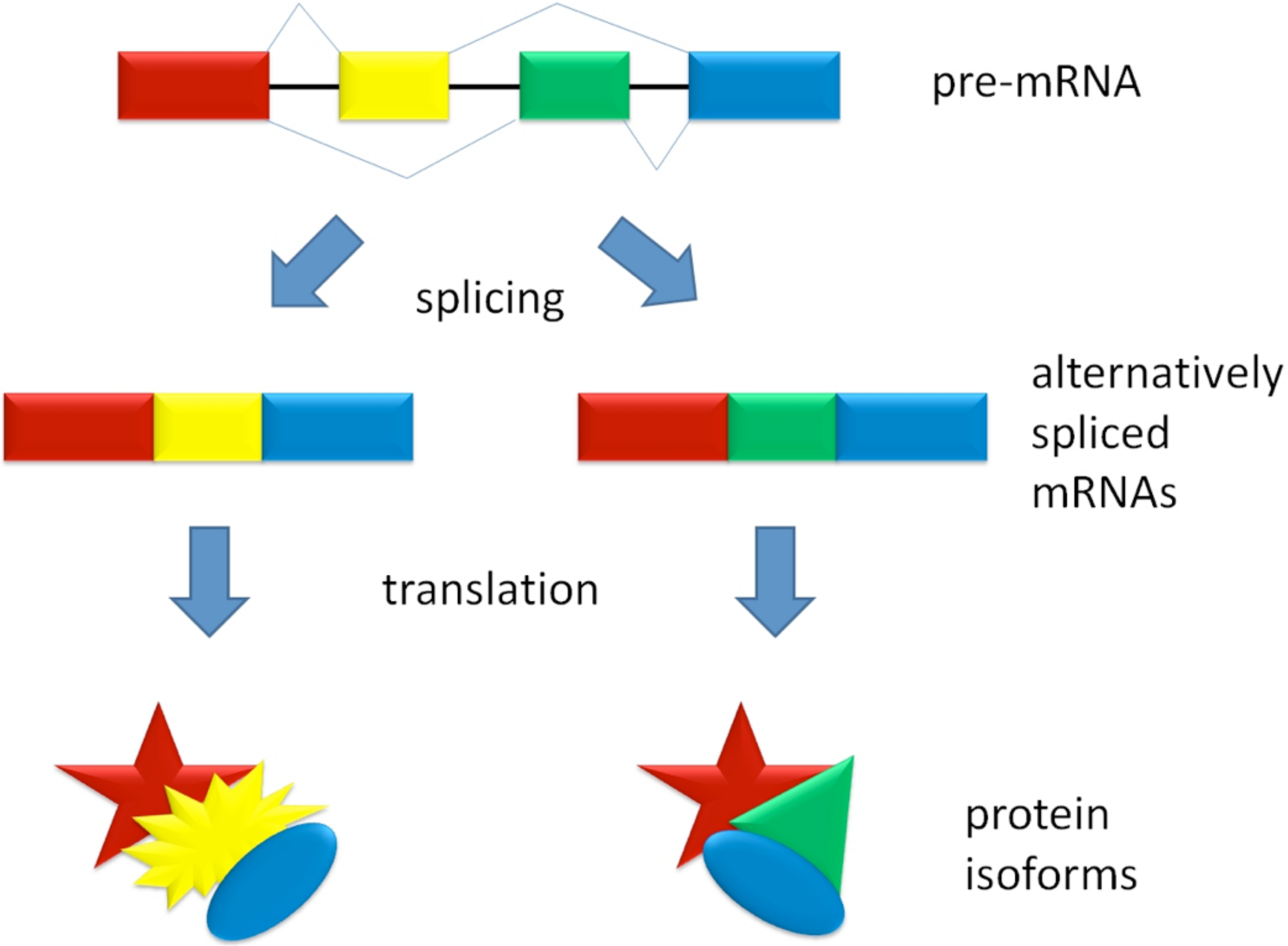
  - Alternative splice forms of a gene share a significant fraction of sequence

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



pre-mRNA

splicing

alternatively spliced mRNAs

translation

protein isoforms

Multi-mapping reads in RNA-Seq

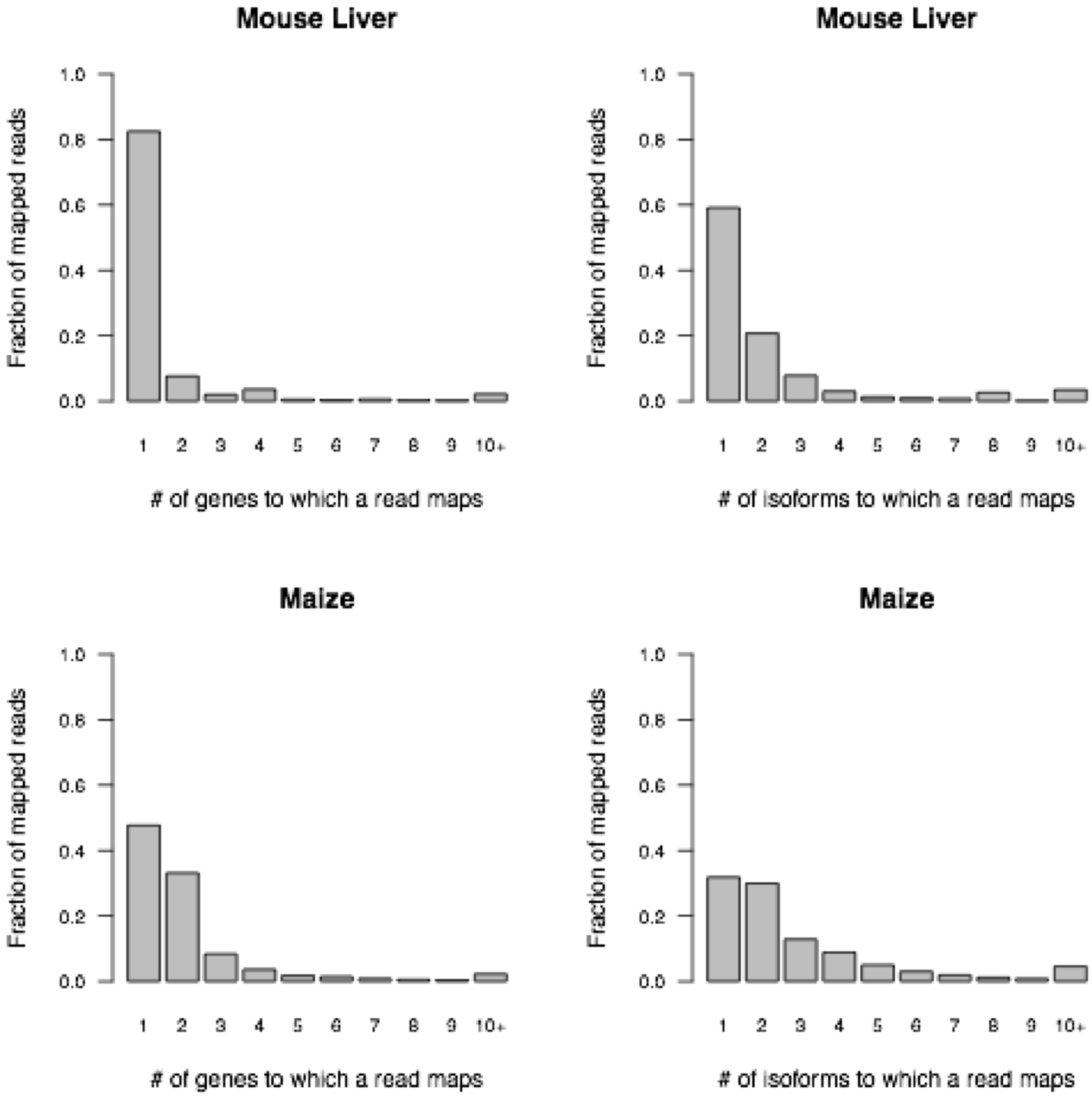
Species	Read length	% multi-mapping reads
Mouse	25	17%
Mouse	75	10%
Maize	25	52%
Axolotl	76	23%
Human	50	23%

- Throwing away multi-mapping reads leads to

  1. Loss of information

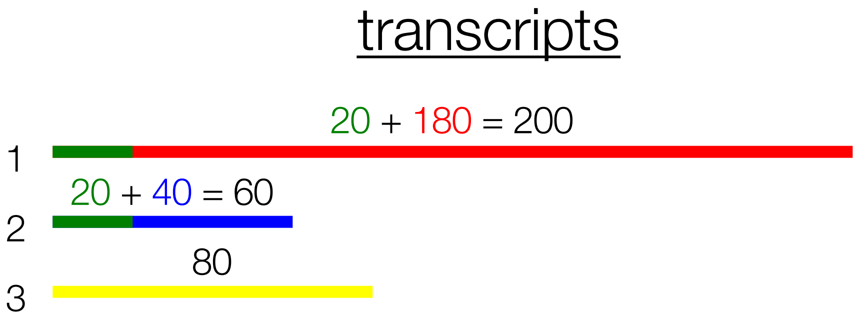
  2. Potentially biased estimates of abundance

Distributions of alignment counts



What if reads do not uniquely map to transcripts?

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



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

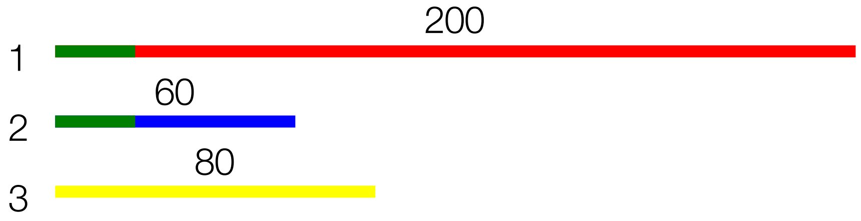
        1. Estimate abundances based on uniquely mapping reads only

        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

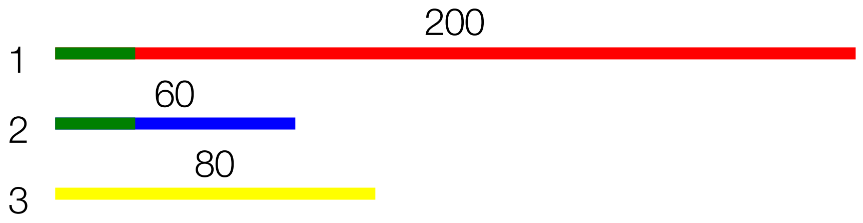
$$\hat{f}_1^{unique} = \frac{\frac{90}{200}}{\frac{90}{200} + \frac{40}{60} + \frac{40}{80}} = 0.278$$

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

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

Rescue method example - Step 2

transcripts



reads

90 A

40 C

40 G

30 T

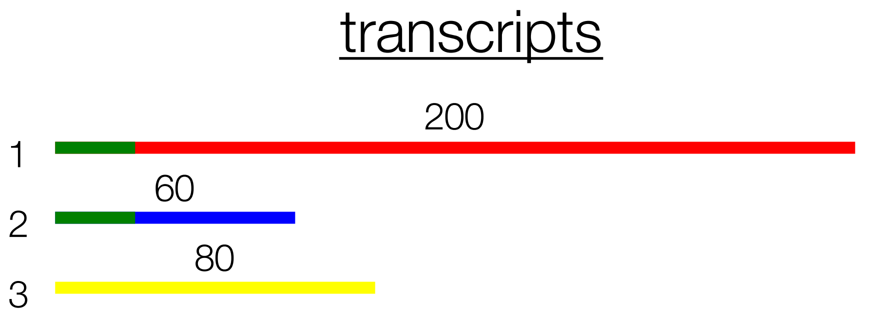
Step 2

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

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

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

Rescue method example - Step 3



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

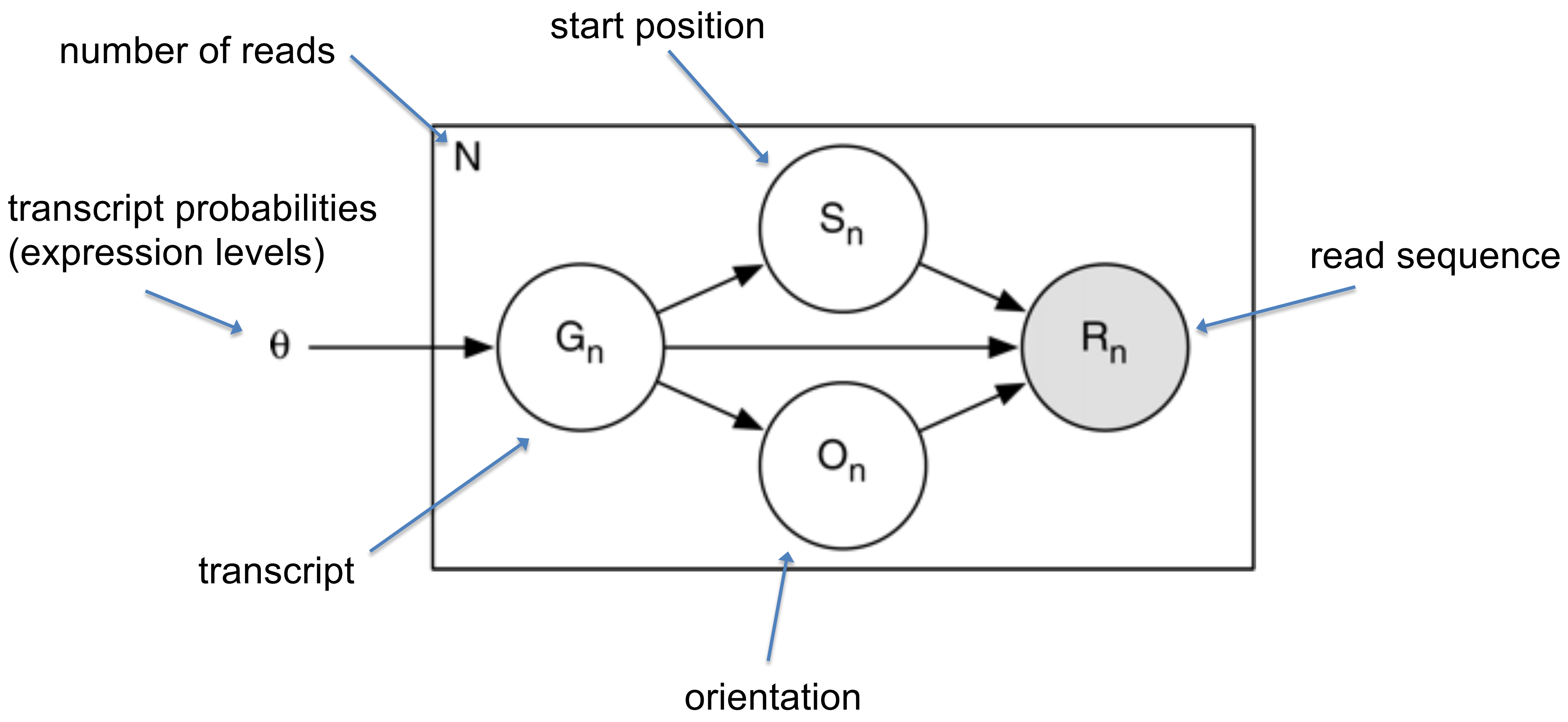
- And then we could update our abundance estimates once again

- 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



number of reads

start position

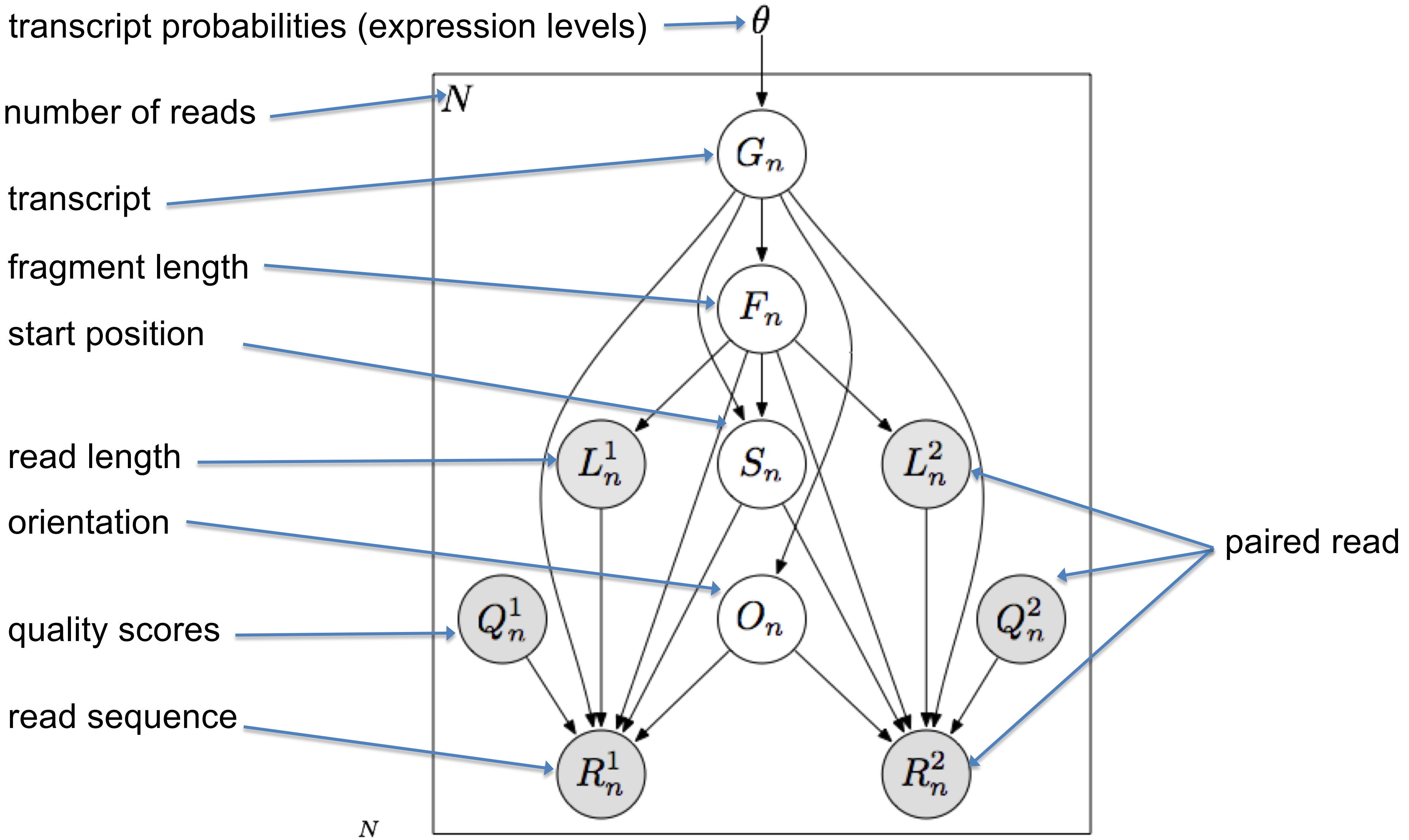
transcript probabilities (expression levels)

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

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

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

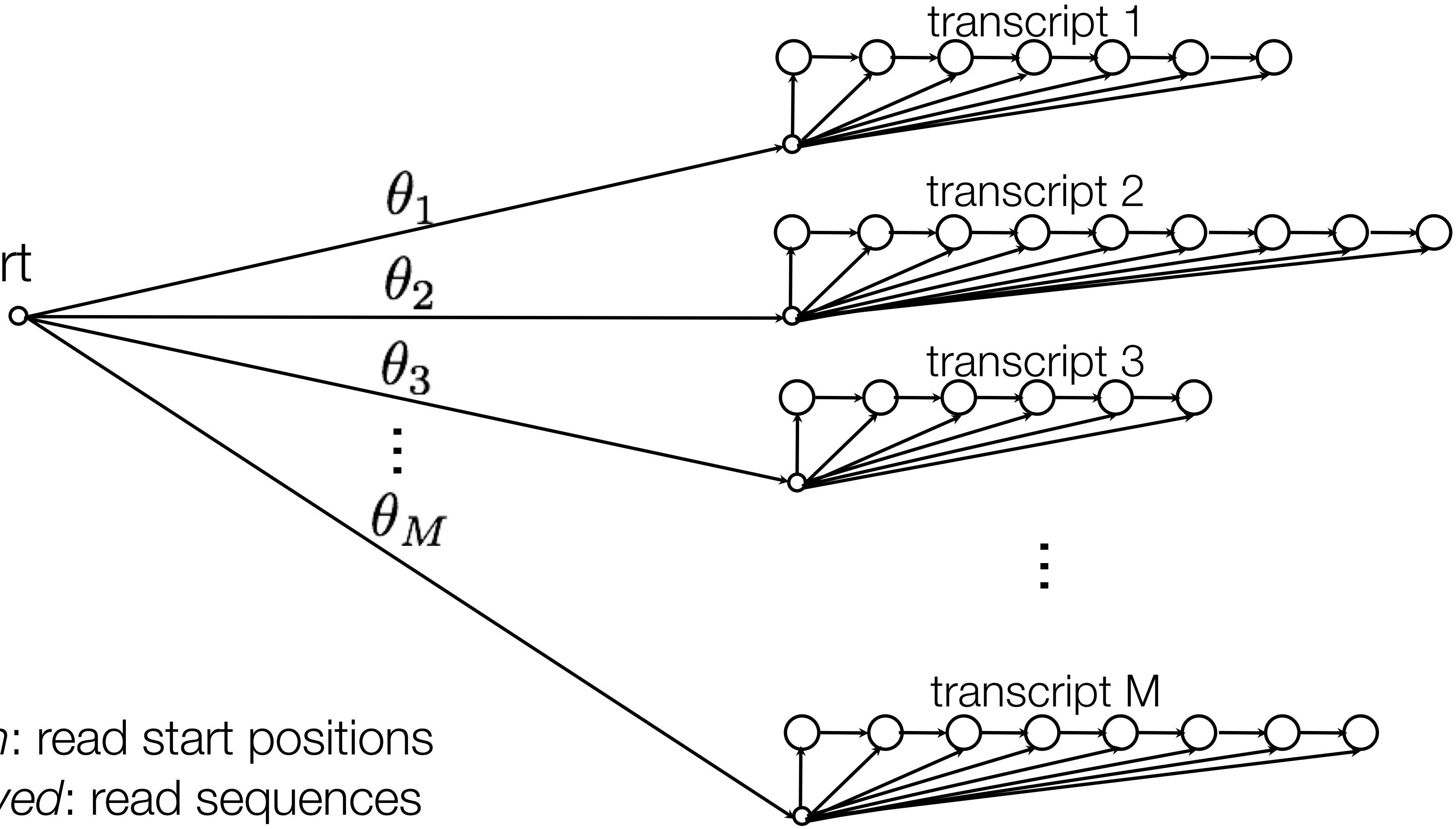
$$P(\mathbf{r}, \ell, \mathbf{q}|\theta) = \prod_{n=1}^{N} \sum_{(i,j,k,o) \in \pi_n^x} \theta_i P(R_n = r_n, L_n = \ell_n, Q_n = q_n, Z_{nijko} = 1 | G_n = i)$$

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

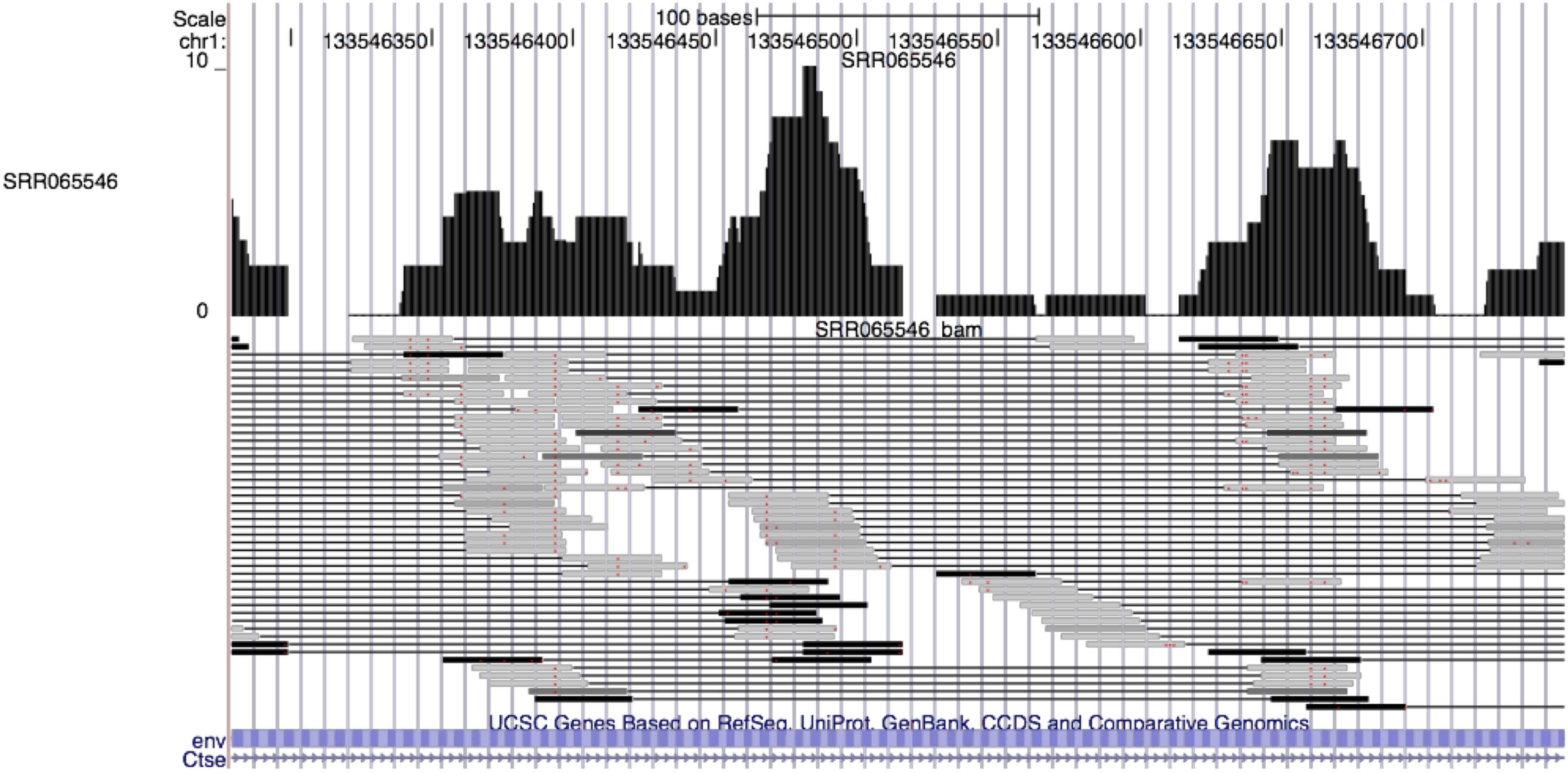
HMM Interpretation



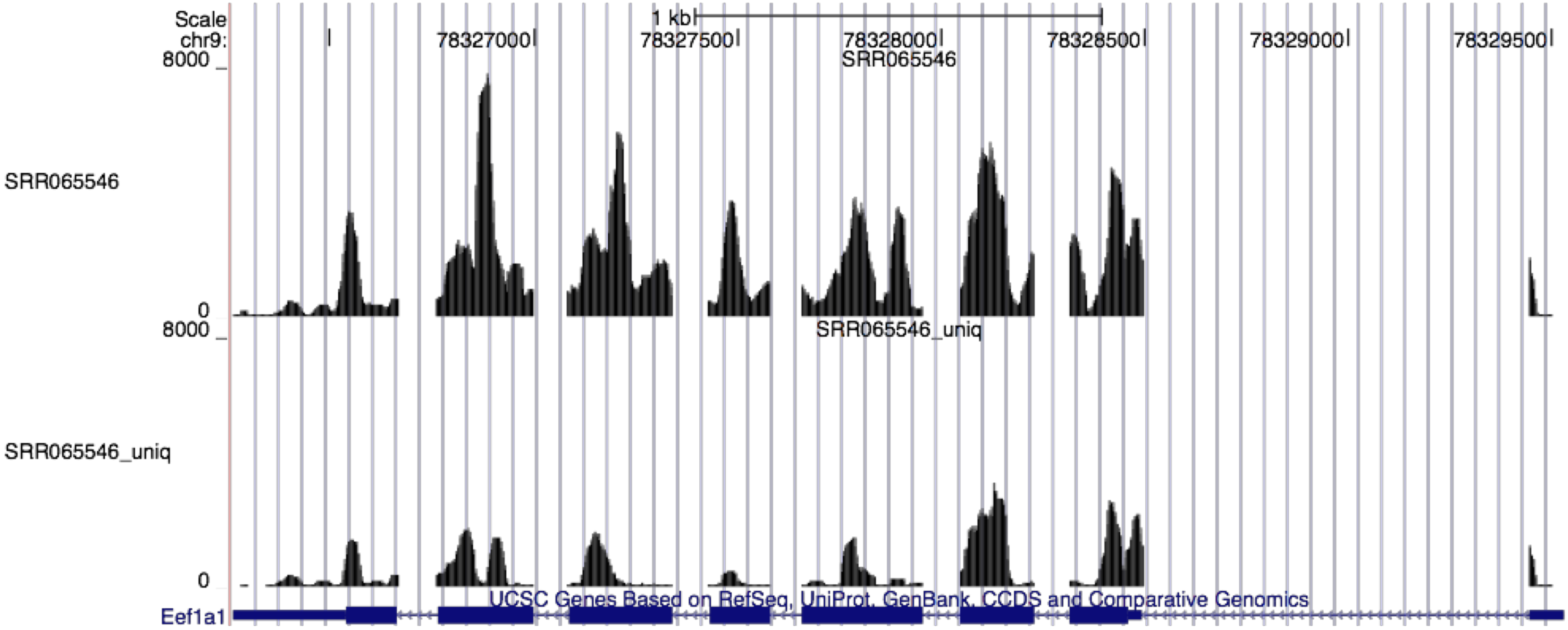
hidden: read start positions
observed: read sequences

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

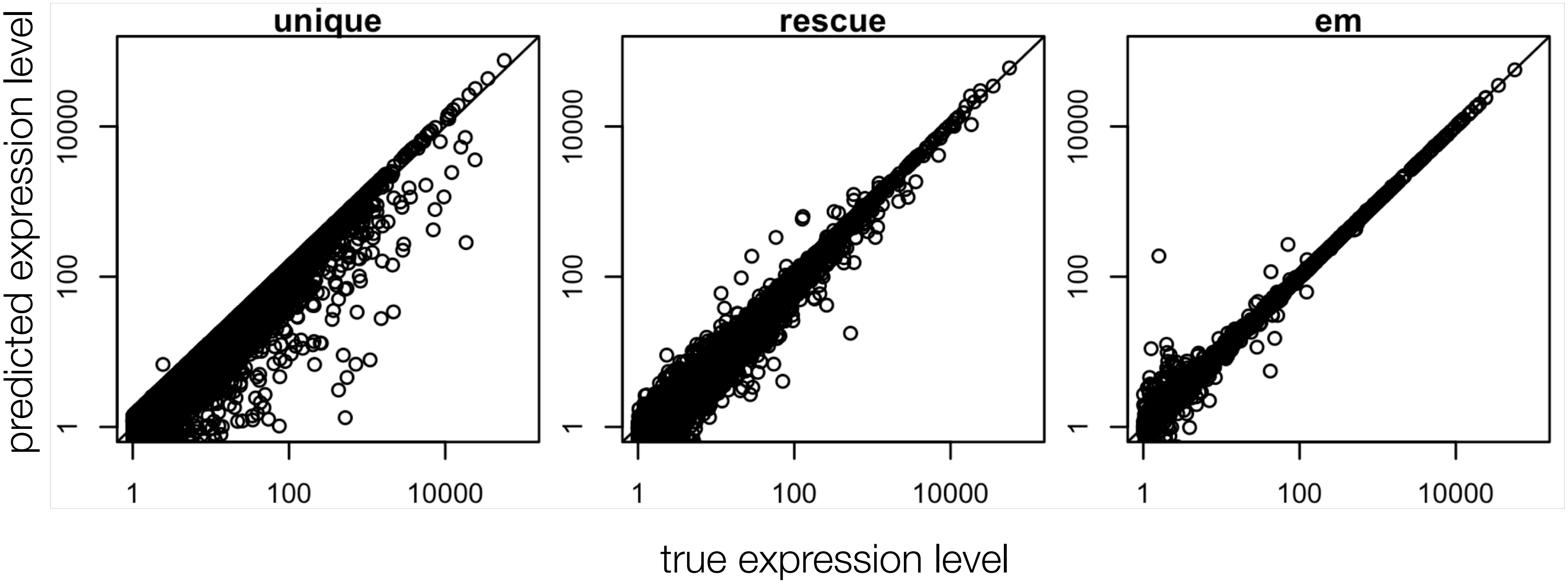
Probabilistically-weighted alignments



Expected read count visualization

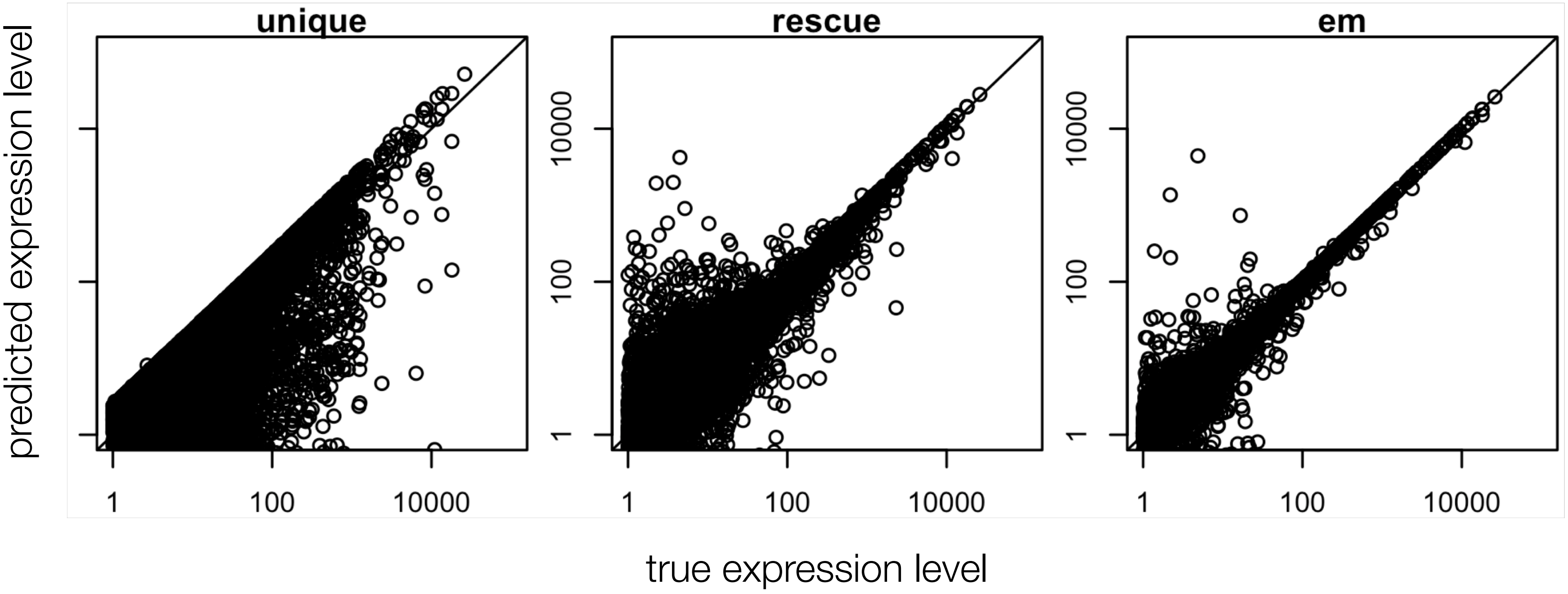


Improved accuracy over unique and rescue



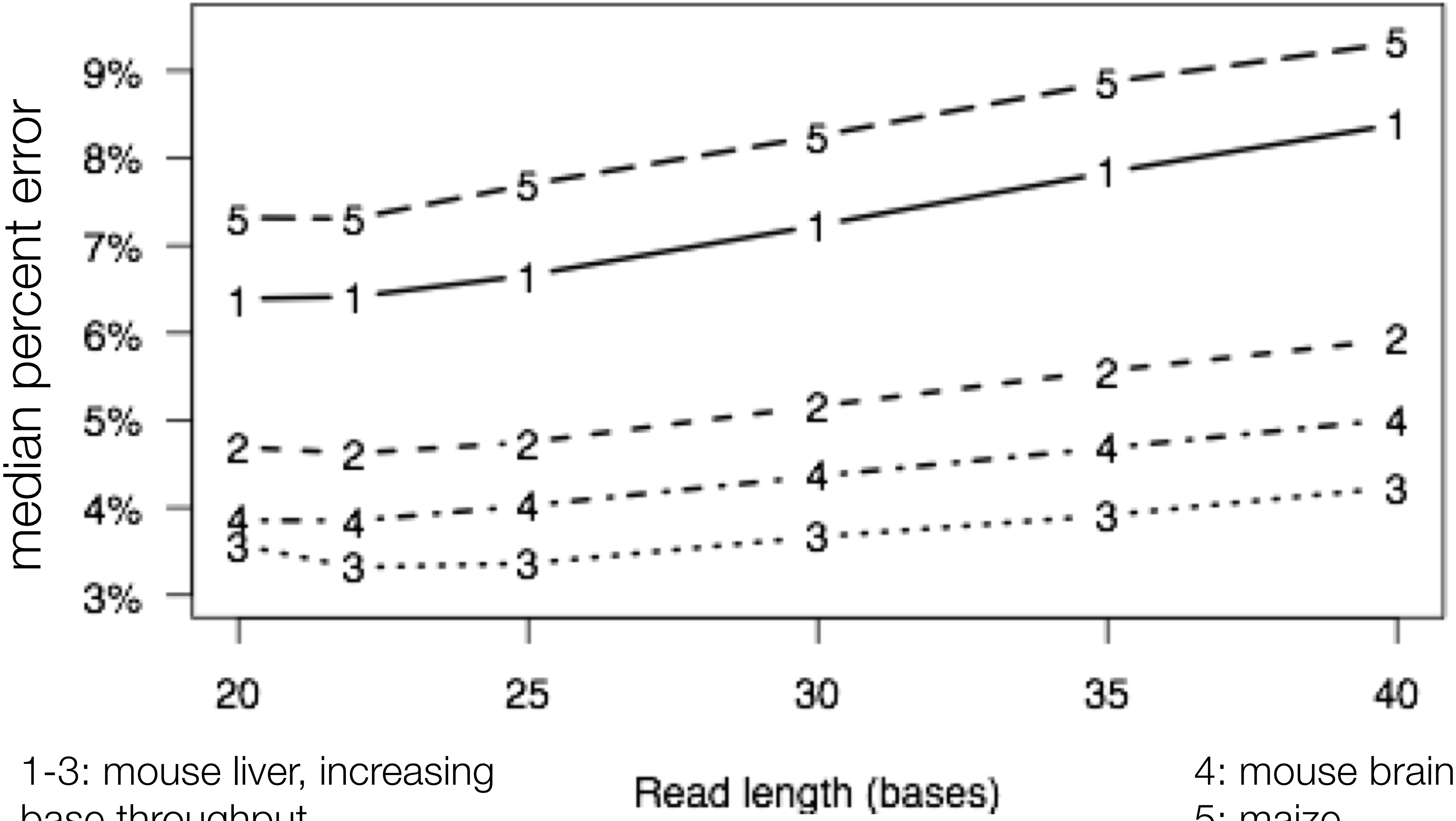
Mouse gene-level expression estimation

Improving accuracy on repetitive genomes: maize



Maize gene-level expression estimation

Finding the optimal read length



1-3: mouse liver, increasing base throughput

4: mouse brain
5: maize

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