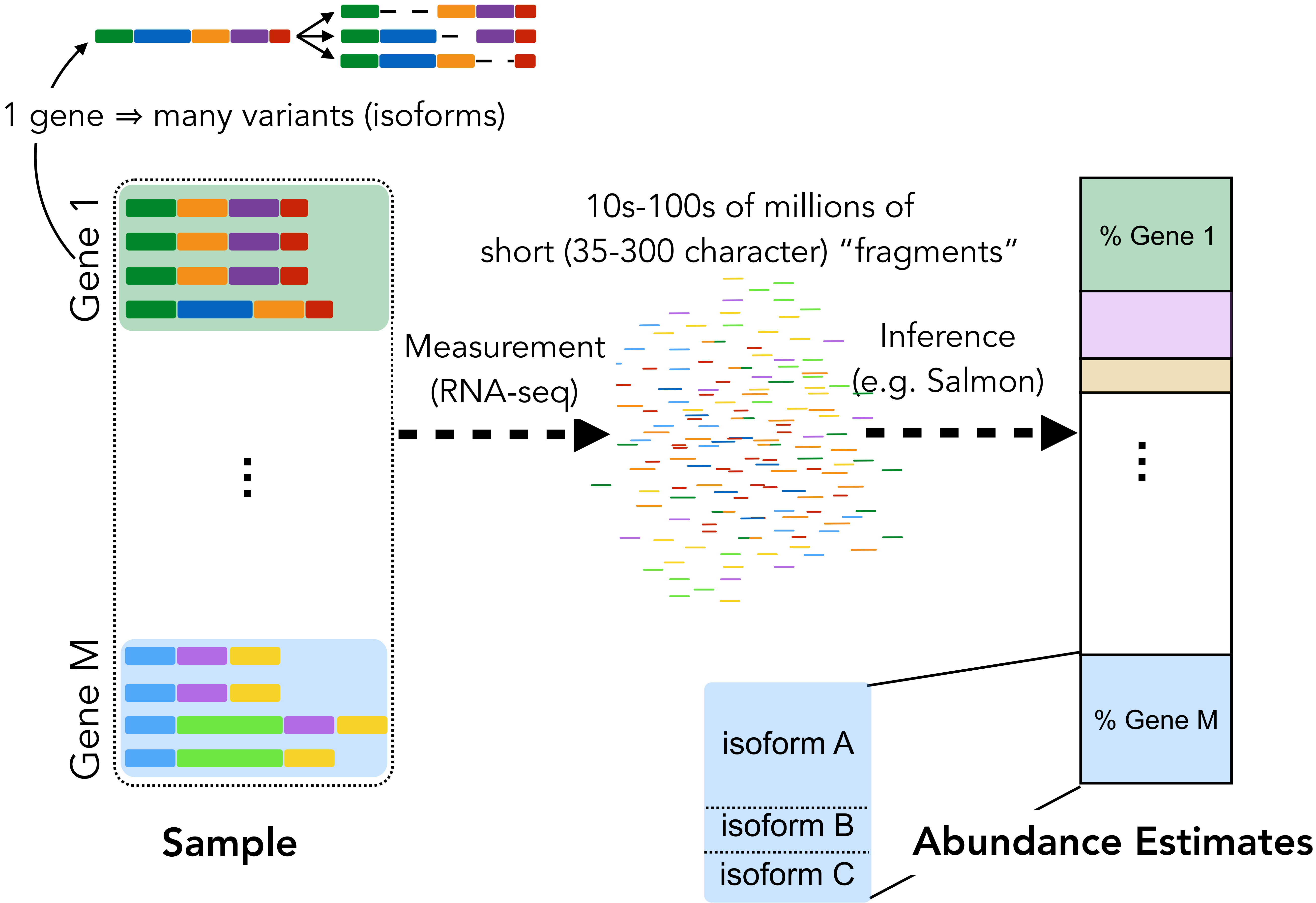
Analyzing gene and transcript expression using RNA-seq (II)

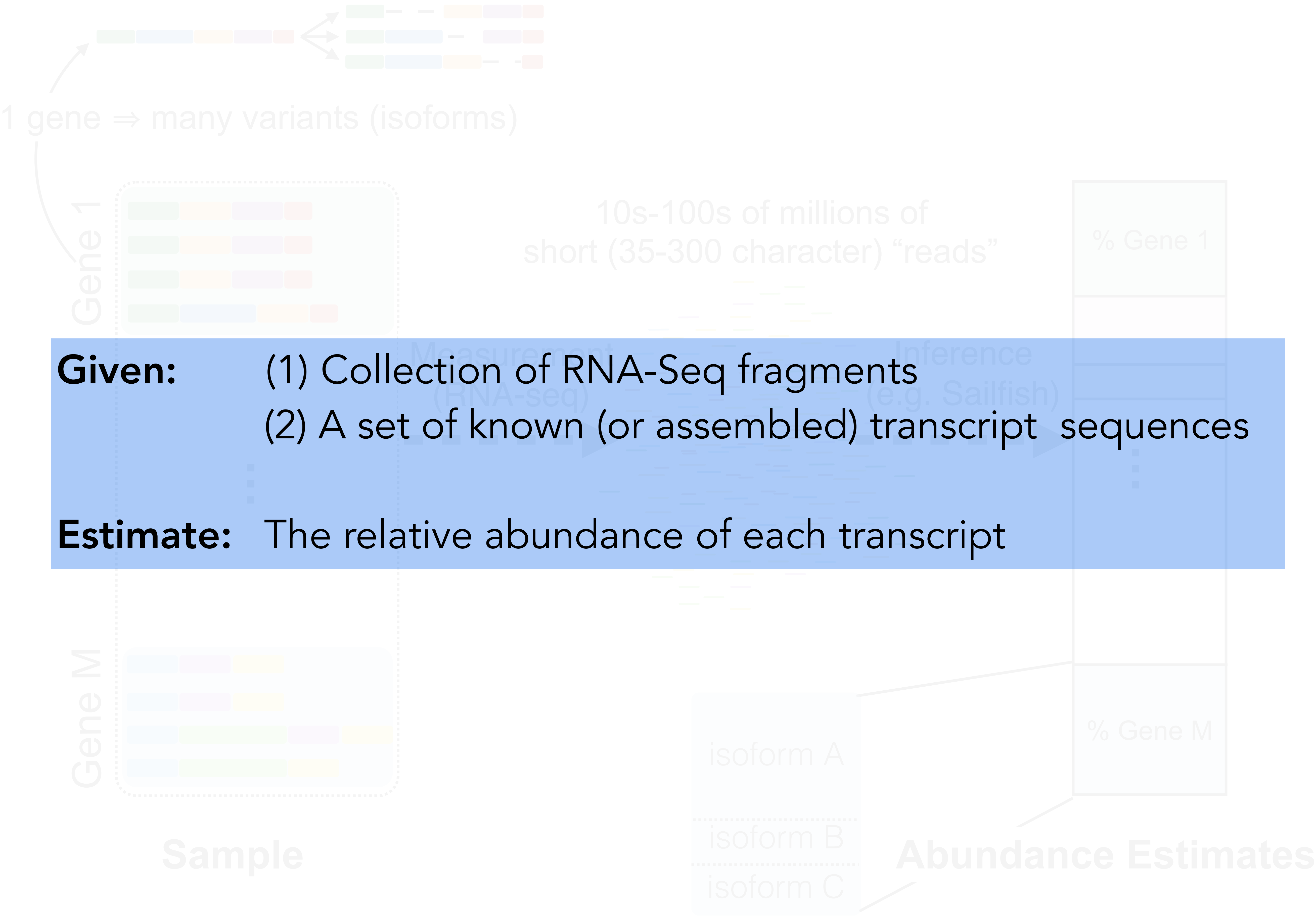
UNIVERSITY OF MARYLAND

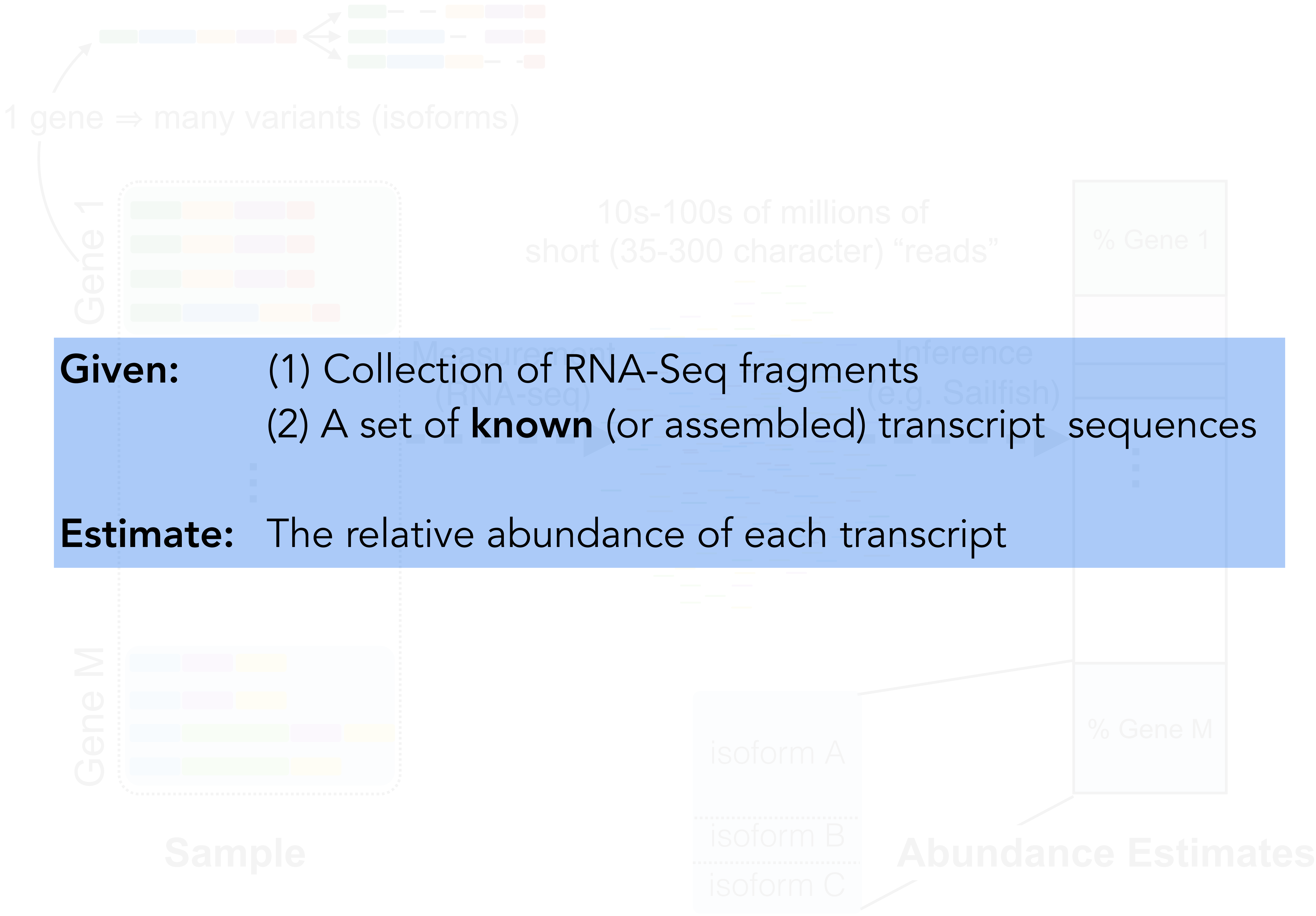
Transcript Quantification: An Overview



Given: (1) Collection of RNA-Seq fragments
(2) A set of known (or assembled) transcript sequences

Estimate: The relative abundance of each transcript





Given: (1) Collection of RNA-Seq fragments

(2) A set of **known** (or assembled) transcript sequences

Estimate: The relative abundance of each transcript

Why not simply “count” reads

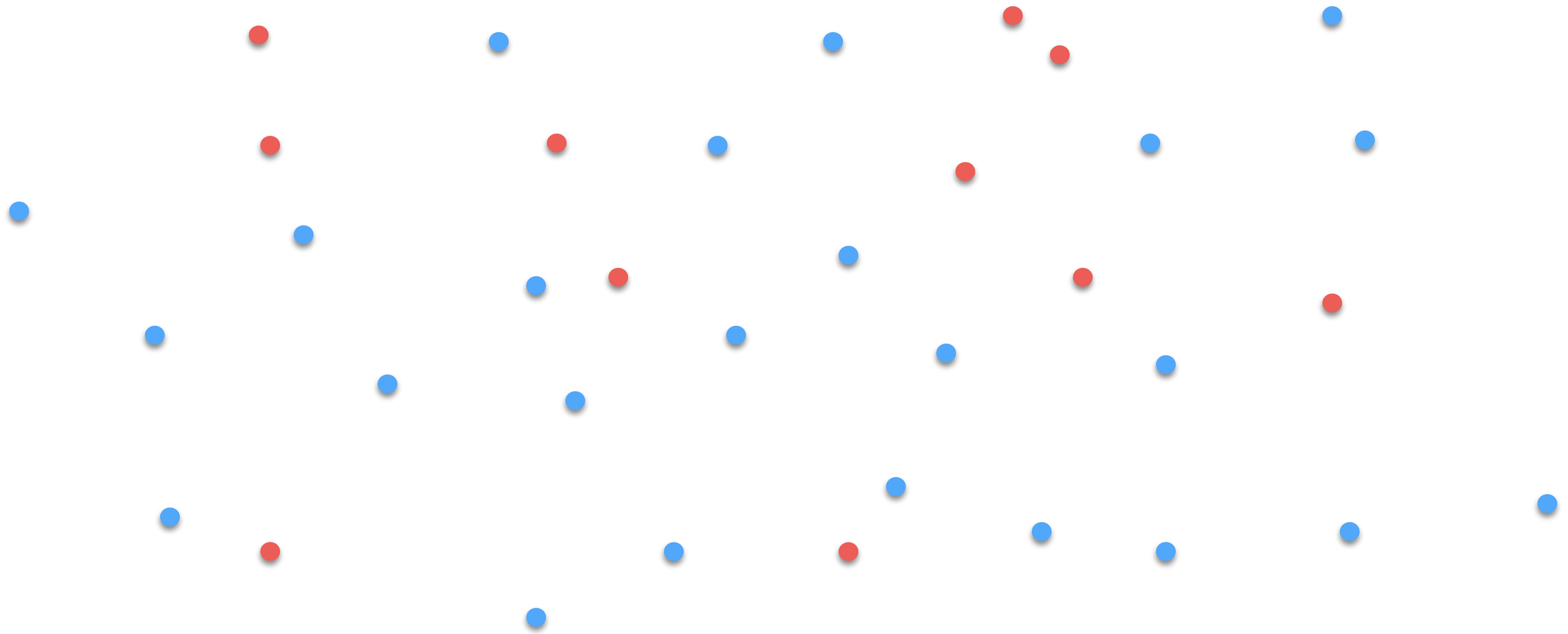
The RNA-seq reads are drawn from transcripts, and our spliced-aligners let us map them back to the transcripts on the genome from which they originate.

Problem: How do you handle reads that align equally-well to multiple isoforms / or multiple genes?

- Discarding multi-mapping reads leads to incorrect and biased quantification
- Even at the gene-level, the transcriptional output of a gene should depend on what isoforms it is expressing.

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



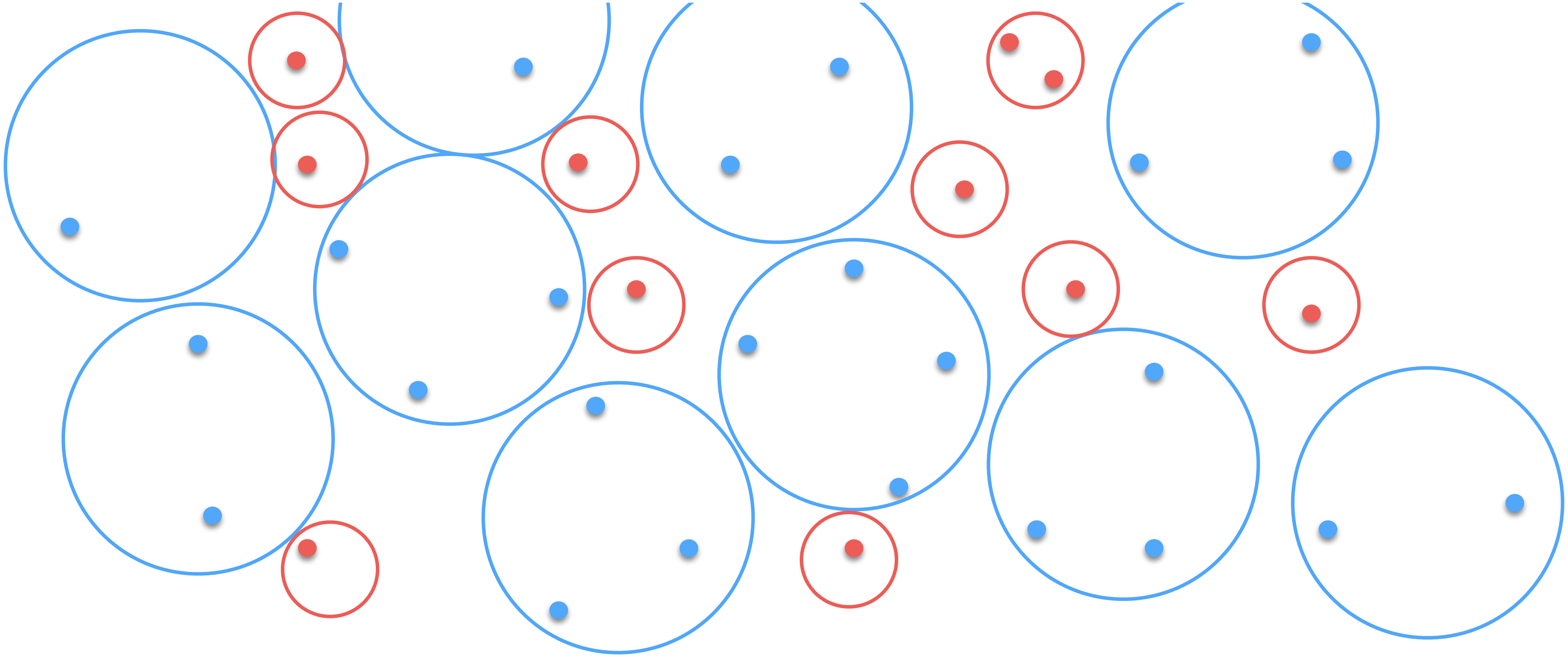
Here, a dot of a color means I hit a circle of that color.

What type of circle is more prevalent?

What is the fraction of red / blue circles?

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.

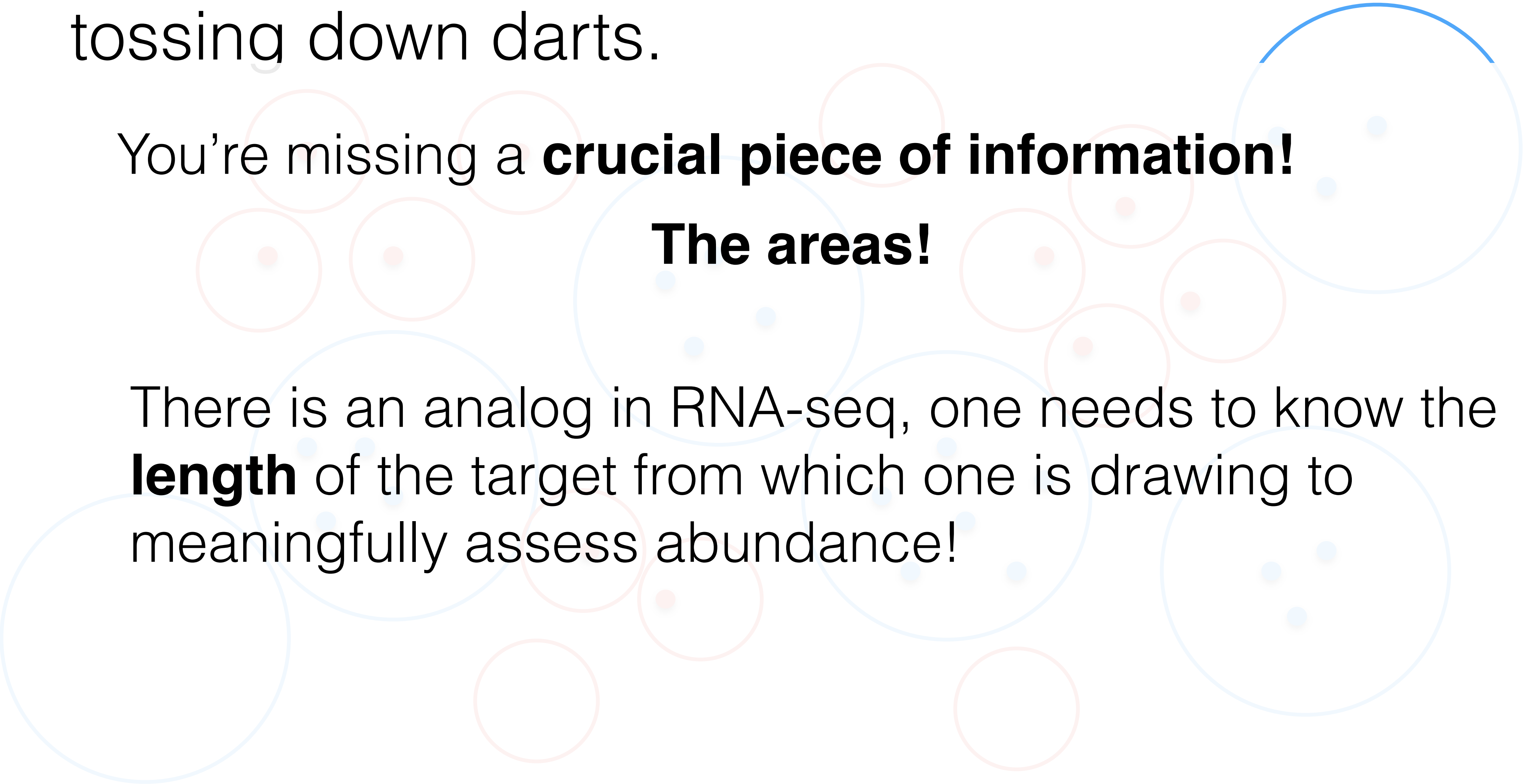


You're missing a **crucial piece of information!**

The areas!

First, consider this non-Biological example

Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



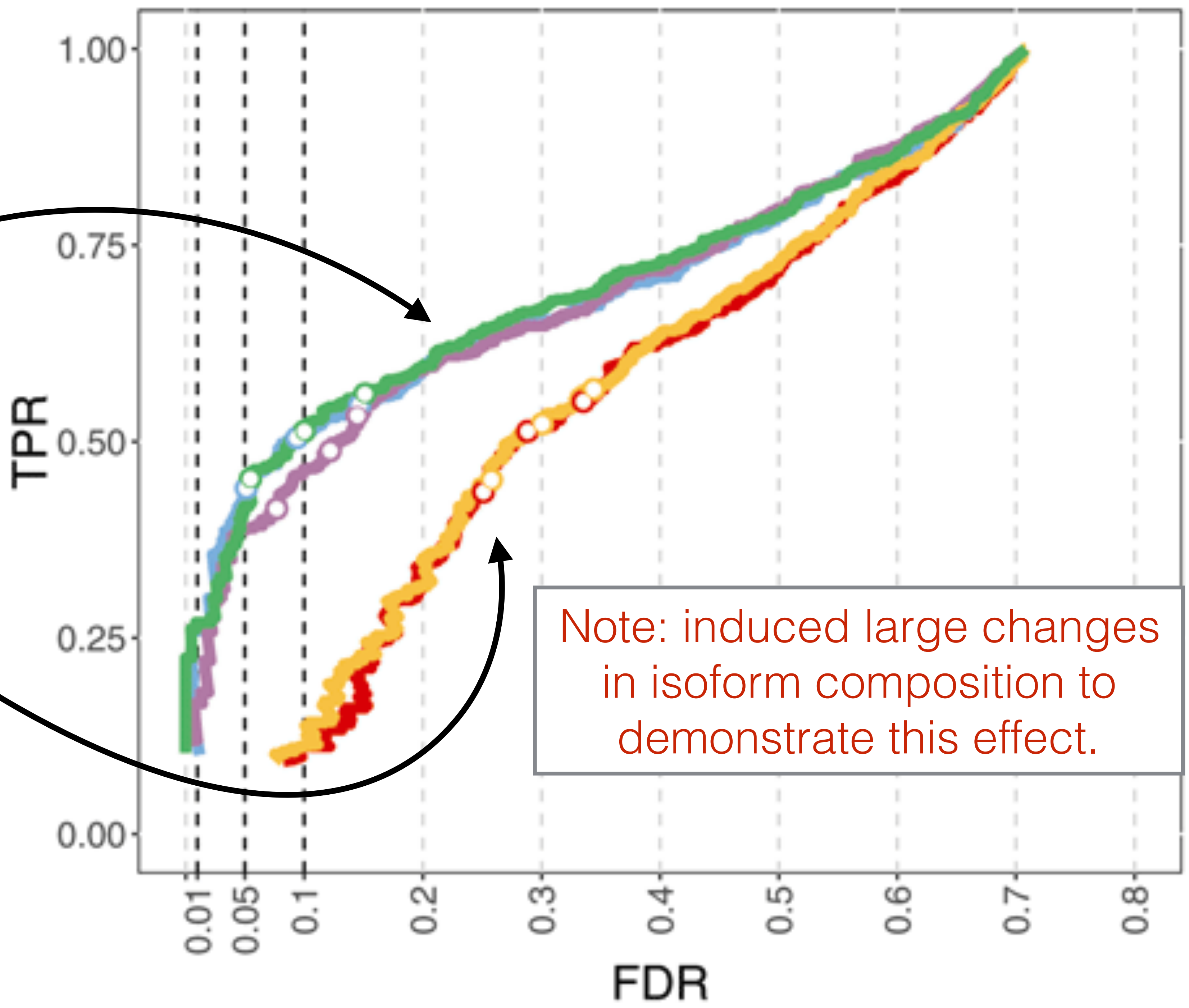
You're missing a **crucial piece of information!**

The areas!

There is an analog in RNA-seq, one needs to know the **length** of the target from which one is drawing to meaningfully assess abundance!

Resolving multi-mapping is fundamental to quantification

These errors can affect DGE calls



Variants of Salmon

Variants of “counting”

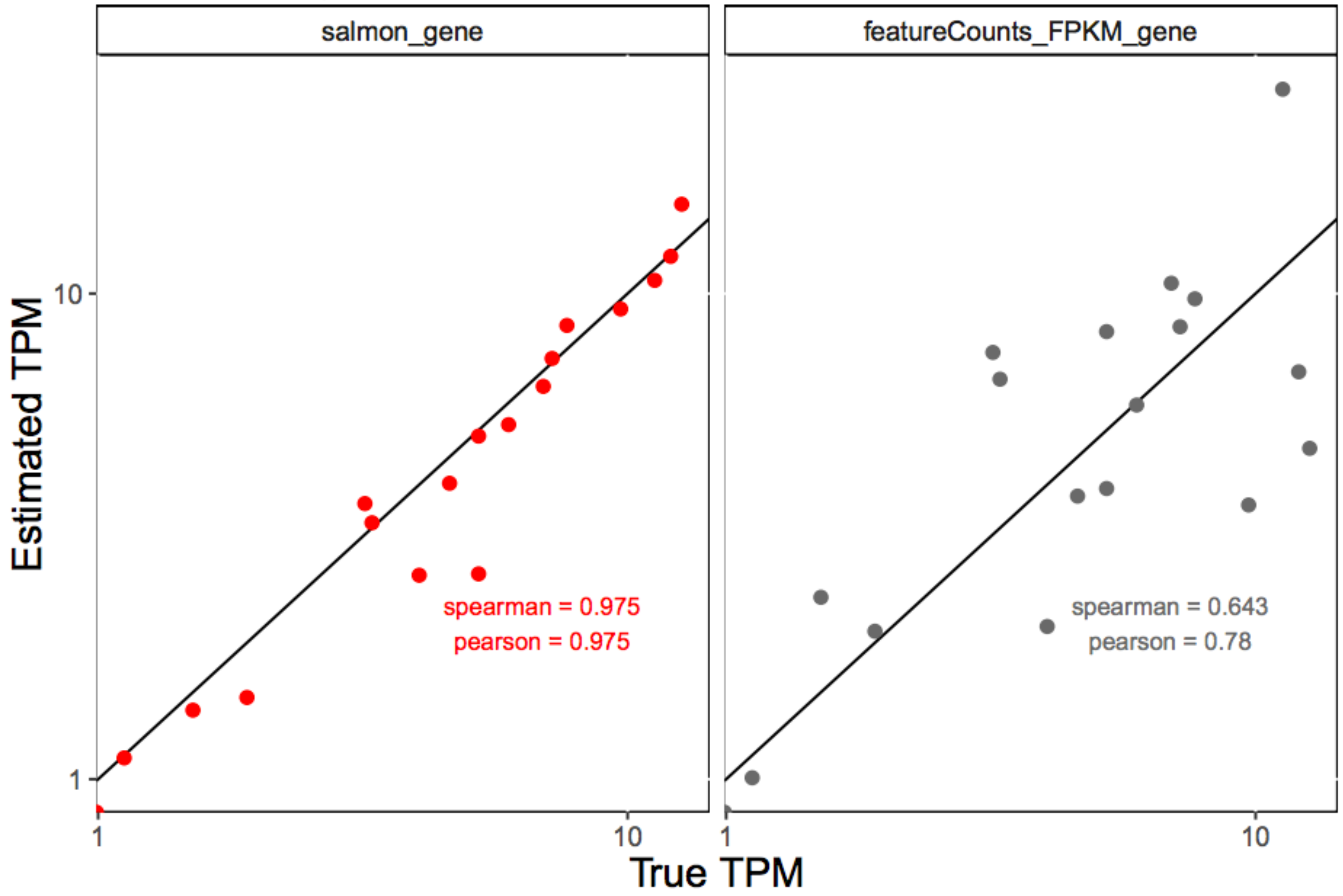
Note: induced large changes in isoform composition to demonstrate this effect.

From: Soneson C, Love MI and Robinson MD 2016 [version 2; referees: 2 approved] F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2)

Resolving multi-mapping is fundamental to quantification

Can even affect abundance estimation in **absence** of alternative-splicing (e.g. paralogous genes)

Paralogs of ENSG00000090612



From: Soneson C, Love MI and Robinson MD 2016 [version 2; referees: 2 approved] F1000Research 2016, 4:1521 (doi: 10.12688/f1000research.7563.2)

Main challenges of fast & accurate quantification

- finding locations of reads (alignment) is slower than necessary → simply aligning reads in a sample **can take hours**

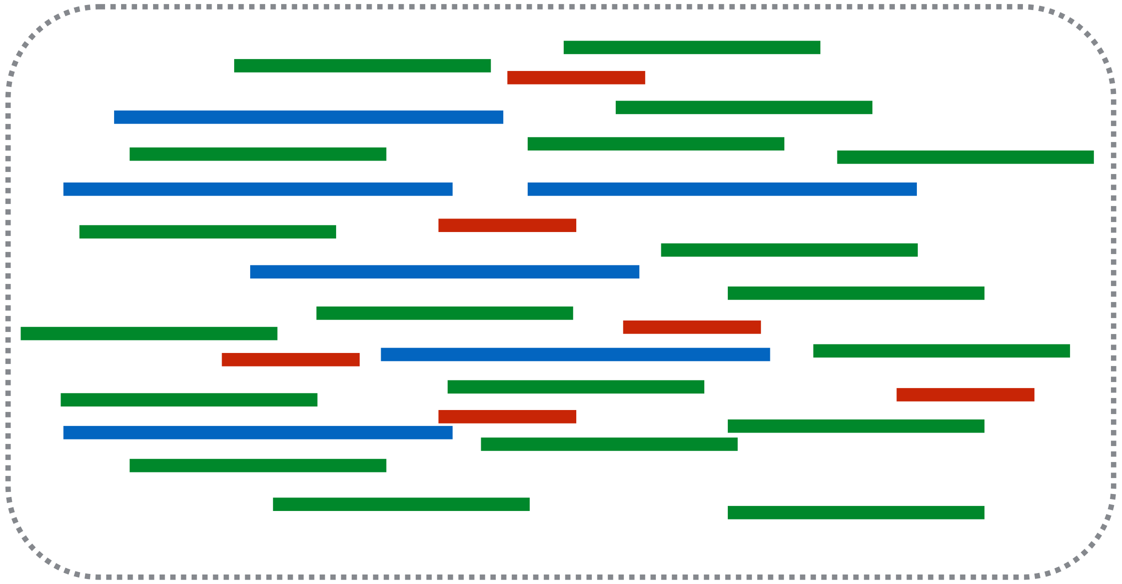
- alternative splicing and related sequences creates ambiguity about where reads came from → **multi-mapping reads** *cannot* be ignored / discarded or assigned naïvely

- sampling of reads is not uniform or idealized, exhibits multiple types of bias → RNA-seq can exhibit ***extensive*** **and** ***sample-specific bias***

- uncertainty in ML estimate of abundances → There is both technical (shot noise) and **inherent** ***inferential*** **uncertainty** in abundance estimates

How can we perform inference from sequenced fragments?

Experimental Mixture



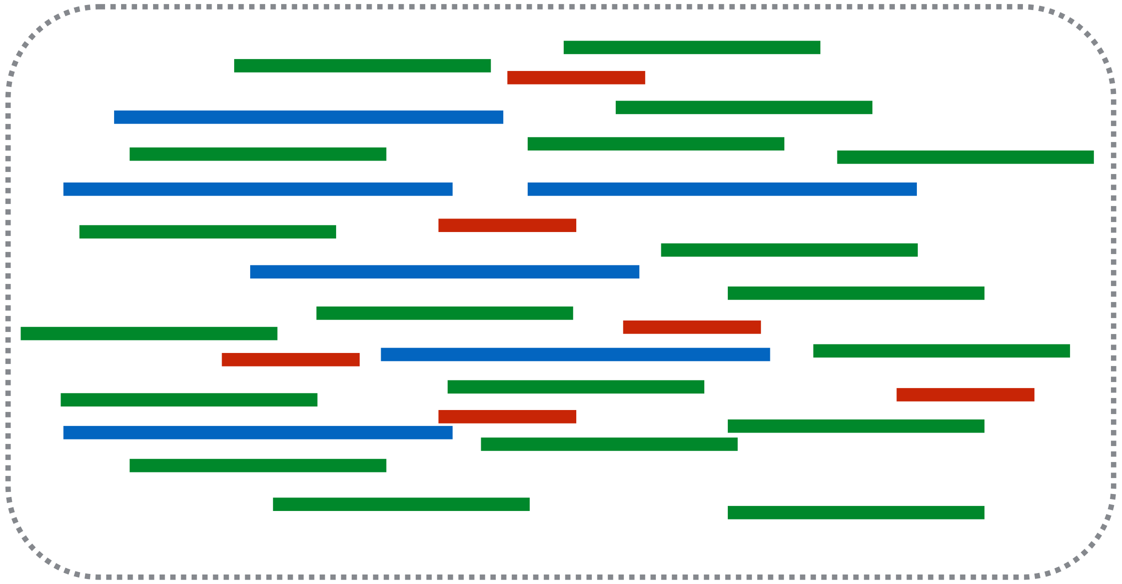
In an unbiased experiment, sampling fragments depends on:

- # of copies of each txp type
- length of each txp type

length(━━━━━━) = 100	x 6 copies	= 600 nt	~ 30% blue
length(━━━━) = 66	x 19 copies	= 1254 nt	~ 60% green
length(━━) = 33	x 6 copies	= 198 nt	~ 10% red

How can we perform inference from sequenced fragments?

Experimental Mixture



In an unbiased experiment, sampling fragments depends on:

- # of copies of each txp type
- length of each txp type

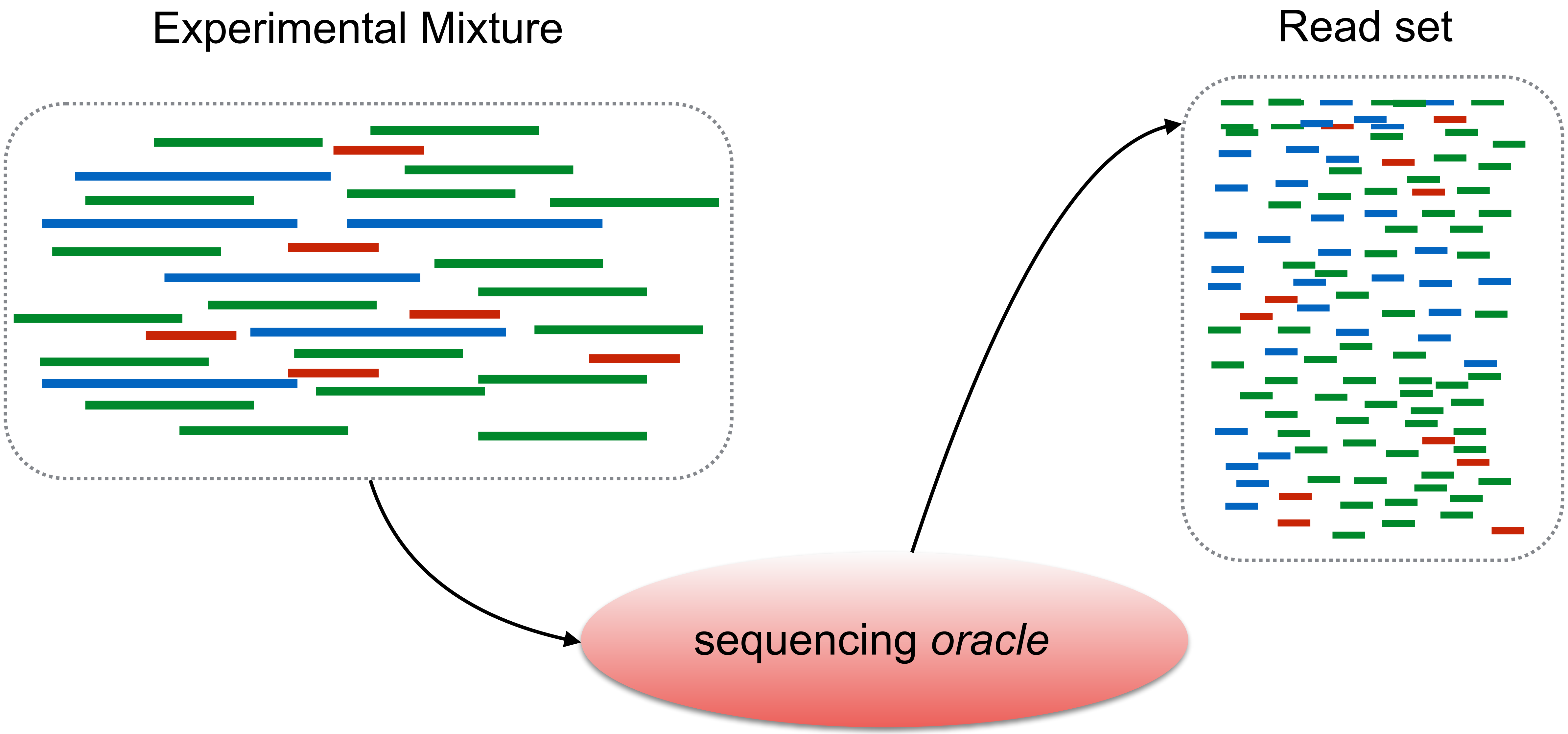
length(━━━━━━) = 100	x 6 copies	= 600 nt	~ 30% blue
length(━━━━) = 66	x 19 copies	= 1254 nt	~ 60% green
length(━━) = 33	x 6 copies	= 198 nt	~ 10% red

↑

We call these values η = [0.3, 0.6, 0.1] the nucleotide fractions, they become the primary quantity of interest

How can we perform inference from sequenced fragments?

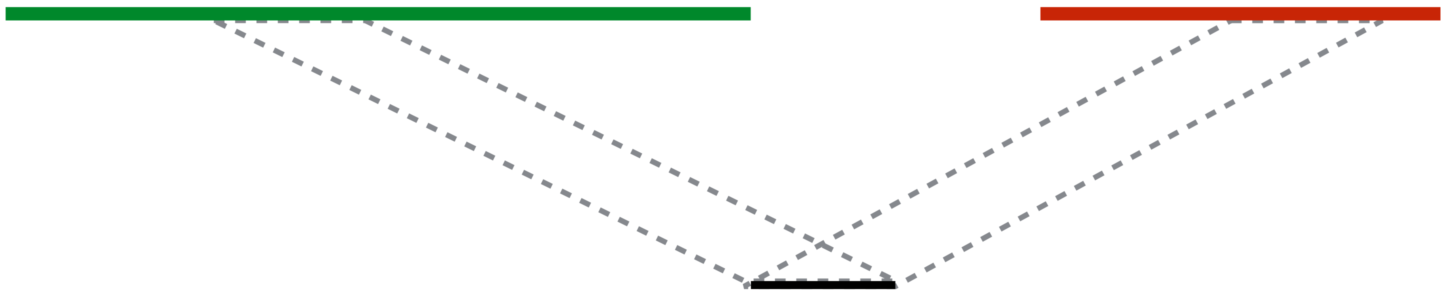
Think about the "ideal" RNA-seq experiment . . .



(1) Pick transcript **t** ∝ total available nucleotides = count * length

(2) Pick a position **p** on **t** "uniformly at random"

Resolving a single multi-mapping read



Say we *knew* the η, and observed a *single* read that mapped ambiguously, as shown above.

What is the probability that it truly originated from G or R?

$$\Pr\{r \text{ from } G\} = \frac{\frac{\eta_G}{\text{length}(G)}}{\frac{\eta_G}{\text{length}(G)} + \frac{\eta_R}{\text{length}(R)}} = \frac{\frac{0.6}{66}}{\frac{0.6}{66} + \frac{0.1}{33}} = 0.75$$

$$\Pr\{r \text{ from } R\} = \frac{\frac{\eta_R}{\text{length}(R)}}{\frac{\eta_G}{\text{length}(G)} + \frac{\eta_R}{\text{length}(R)}} = \frac{\frac{0.1}{33}}{\frac{0.6}{66} + \frac{0.1}{33}} = 0.25$$

normalization factor

length(🟦) = 100	x 6 copies	= 600 nt	~ 30% blue
length(🟩) = 66	x 19 copies	= 1254 nt	~ 60% green
length(🟥) = 33	x 6 copies	= 198 nt	~ 10% red

Units for Relative Abundance

TPM (Transcripts Per Million)

$$\mathrm{TPM}_i = \rho_i \times 10^6 \text{ where } 0 \leq \rho_i \leq 1 \text{ and } \sum_i \rho_i = 1$$

$$\rho_i = \frac{\underbrace{\frac{X_i}{\ell_i}}}{\sum_j \frac{X_j}{\ell_j}}$$

Reads coming from transcript i

Units for Relative Abundance

TPM (Transcripts Per Million)

$$\mathrm{TPM}_i = \rho_i \times 10^6 \text{ where } 0 \leq \rho_i \leq 1 \text{ and } \sum_i \rho_i = 1$$

$$\rho_i = \frac{\frac{X_i}{\ell_i}}{\sum_j \frac{X_j}{\ell_j}}$$

Reads coming from transcript i

Length of transcript i

Units for Relative Abundance

TPM (Transcripts Per Million)

$$\text{TPM}_i = \rho_i \times 10^6 \text{ where } 0 \le \rho_i \le 1 \text{ and } \sum_i \rho_i = 1$$

abundance of i as fraction of all measured transcripts

$$\rho_i = \frac{\frac{X_i}{\ell_i}}{\sum_j \frac{X_j}{\ell_j}}$$

Reads coming from transcript i

Length of transcript i

Units for Relative Abundance

TPM (Transcripts Per Million)

$$\text{TPM}_i = \rho_i \times 10^6 \text{ where } 0 \leq \rho_i \leq 1 \text{ and } \sum_i \rho_i = 1$$

abundance of i as fraction of all measured transcripts

$$\rho_i = \frac{\frac{X_i}{\ell_i}}{\sum_j \frac{X_j}{\ell_j}}$$

Reads coming from transcript i

Length of transcript i

A probabilistic view of RNA-Seq quantification

nucleotide fractions

known transcriptome

assumes independence of fragments

$$\Pr\{\mathcal{F} \mid \boldsymbol{\eta}, \mathcal{T}\} = \prod_{j=1}^{N} \Pr\{f_j \mid \boldsymbol{\eta}, \mathcal{T}\}$$

observed fragments (reads)

We can safely truncate $\Pr\{t_i \mid$ **η**$\}$ to 0 for transcripts where a fragment doesn't map/align.

$$= \prod_{j=1}^{N} \boxed{\sum_{i=1}^{M} \Pr\{t_i \mid \boldsymbol{\eta}\} \cdot \boxed{\Pr\{f_j \mid t_i, \boldsymbol{z}_{ji} = 1\}}}$$

Prob. of selecting t_i *given* **η**

Prob. of generating fragment f_j *given* that it originates from t_i

Depends on abundance estimate

Independent of abundance estimate

We want to find the values of **η** that ***maximize*** this probability. We can do this (at least locally) using the EM algorithm.

*Li, Bo, and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC bioinformatics 12.1 (2011): 1.

A probabilistic view of RNA-Seq quantification

E-step: (what is the “soft assignment” of each read to the transcripts where it aligns)

$$E_{Z|f,\eta^{(t)}} = P(Z_{nij} = 1 \mid f, \eta^{(t)}) = \frac{(\eta_i^{(t)}/\ell_i)P(f_n \mid Z_{nij} = 1)}{\sum_{i',j'} (\eta_{i'}^{(t)}/\ell'_i)P(f_n \mid Z_{ni'j'} = 1)}$$

M-step: Given these soft assignments, how abundant is each transcript?

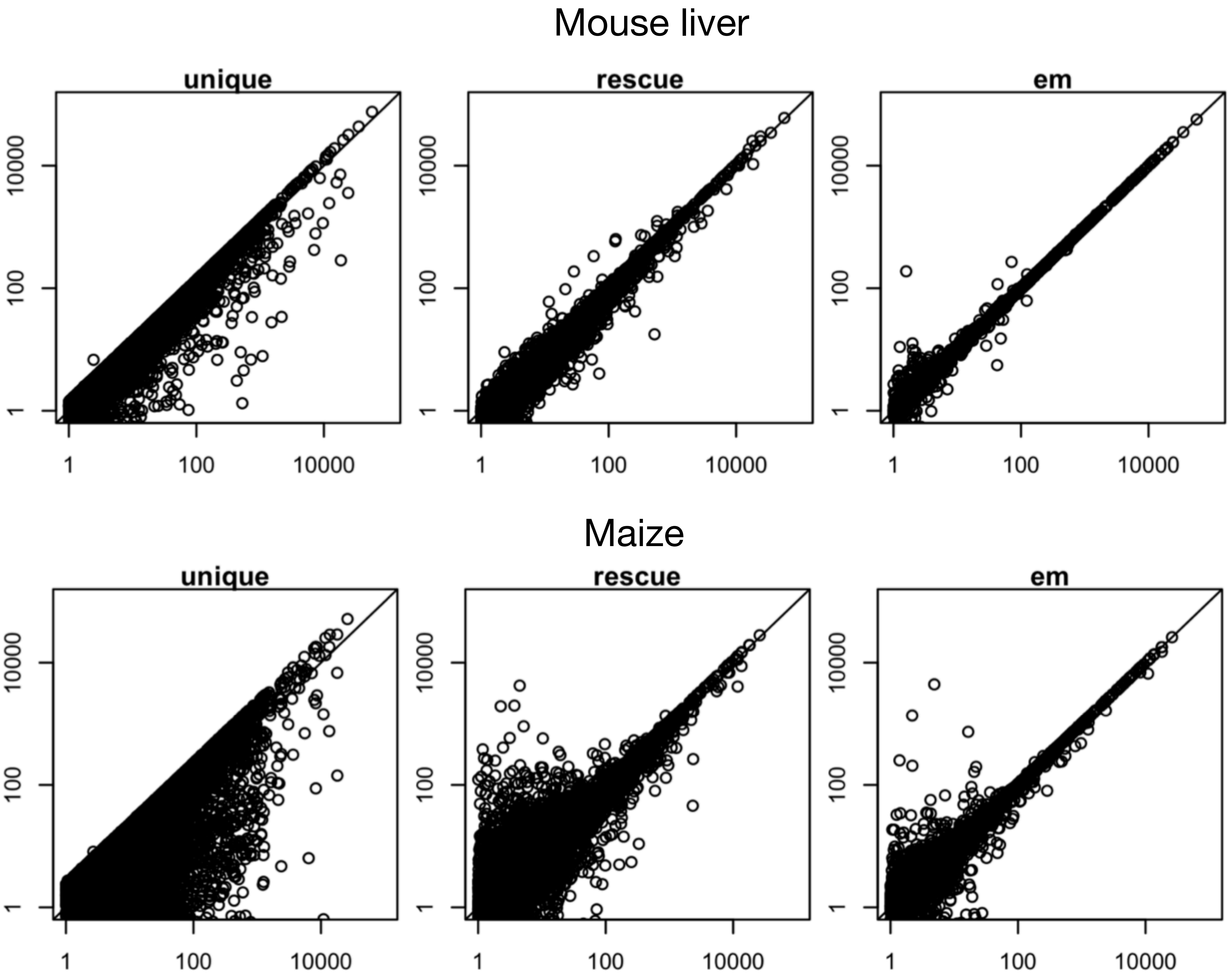
$$\eta_i^{(t+1)} = \frac{E_{Z|f,\eta^{(t)}}\left[C_i\right]}{N},$$

$$\text{where } C_i = \sum_{n,i,j} P(Z_{nij} = 1 \mid f, \eta^{(t)})$$

This approach is quite effective. Unfortunately, it’s also quite slow.

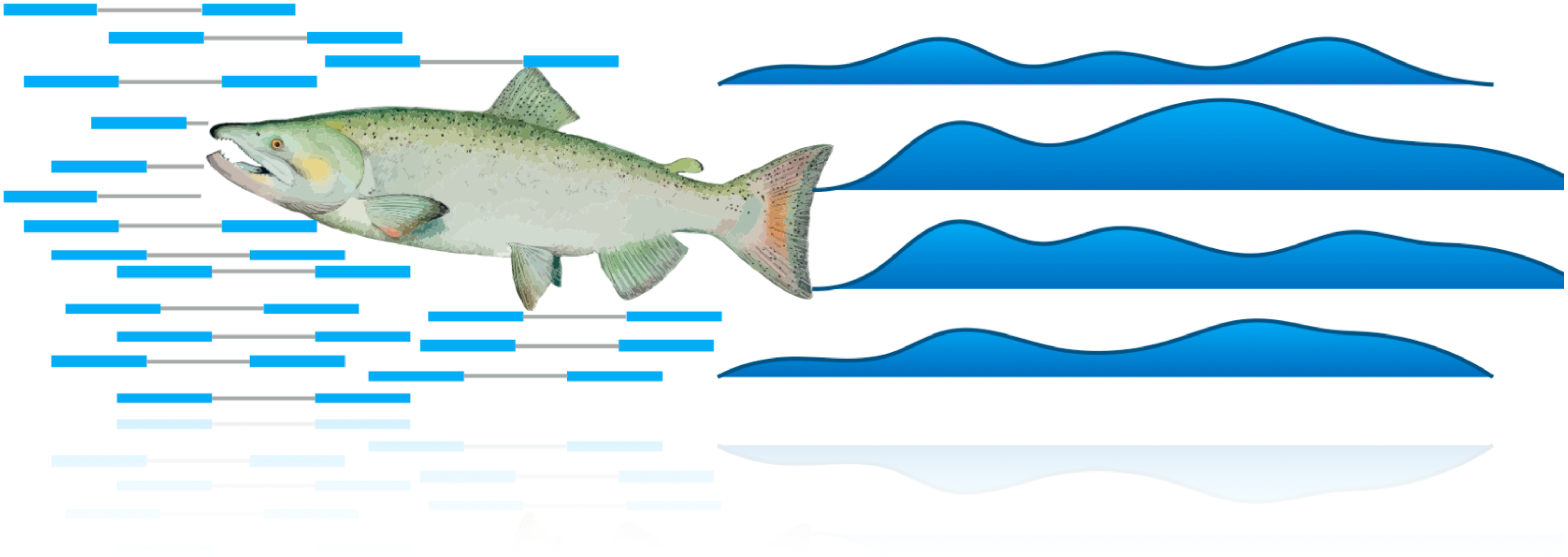
Equations adapted from: Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, Colin N. Dewey; RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics*, Volume 26, Issue 4, 15 February 2010, Pages 493–500, https://doi.org/10.1093/bioinformatics/btp692

Gene expression estimation accuracy in simulated data



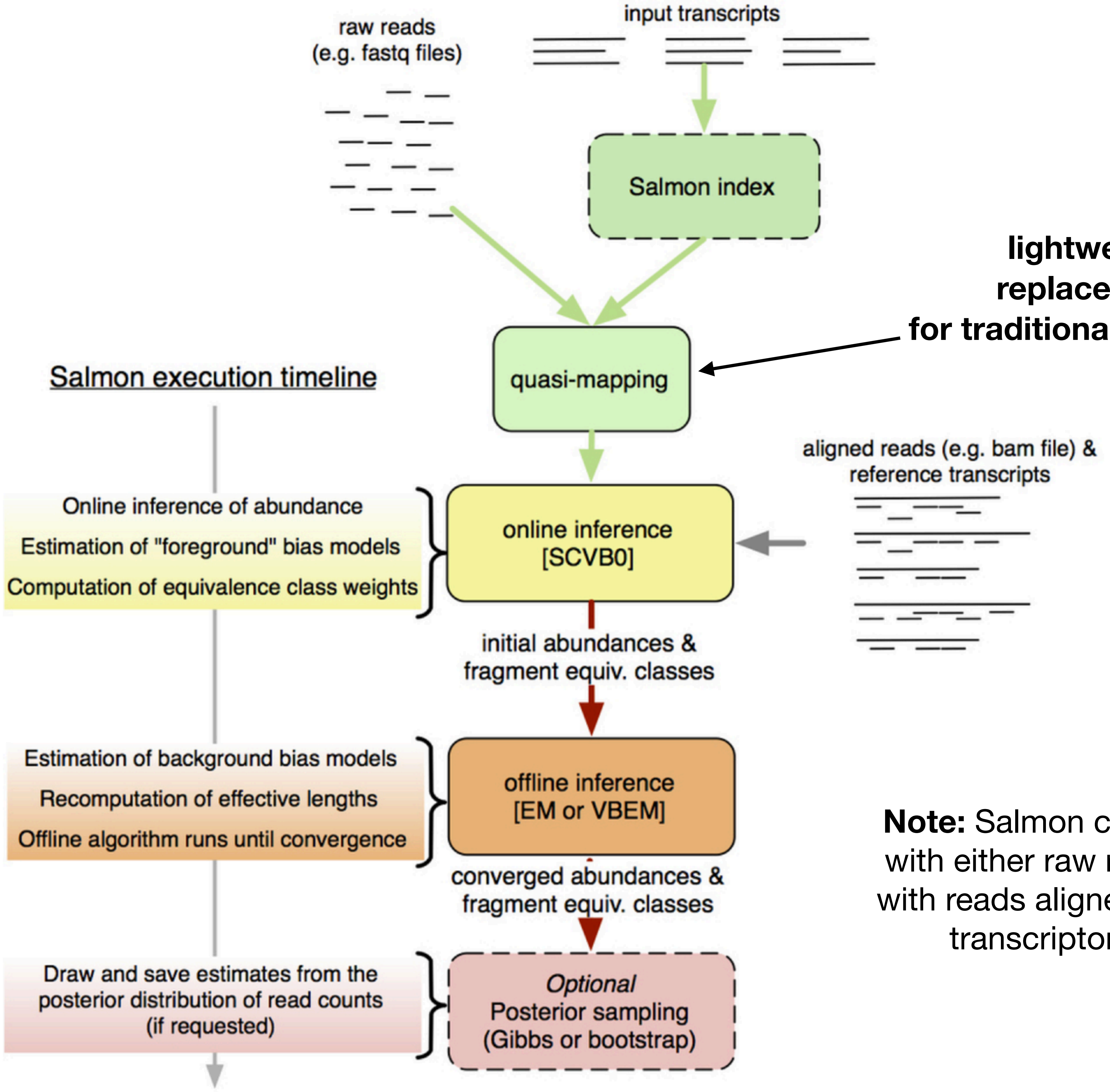
From supplementary material of : Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, Colin N. Dewey; RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics*, Volume 26, Issue 4, 15 February 2010, Pages 493–500, https://doi.org/10.1093/bioinformatics/btp692

Salmon provides fast and bias-aware quantification of transcript expression



Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017).
Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*.

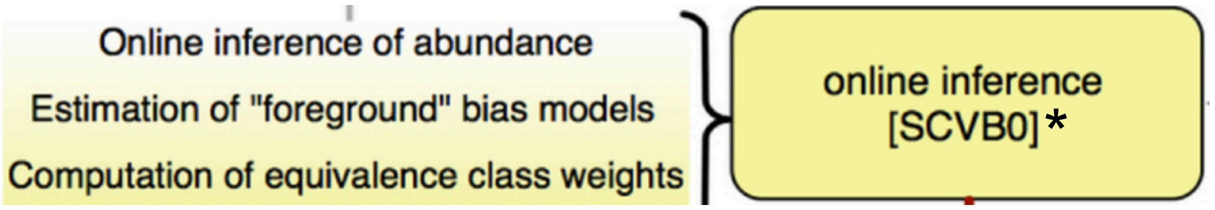
Salmon's "pipeline"


raw reads
(e.g. fastq files)
input transcripts
Salmon index
quasi-mapping
Salmon execution timeline
aligned reads (e.g. bam file) &
reference transcripts
Online inference of abundance
Estimation of "foreground" bias models
Computation of equivalence class weights
online inference
[SCVB0]
initial abundances &
fragment equiv. classes
Estimation of background bias models
Recomputation of effective lengths
Offline algorithm runs until convergence
offline inference
[EM or VBEM]
converged abundances &
fragment equiv. classes
Draw and save estimates from the
posterior distribution of read counts
(if requested)
Optional
Posterior sampling
(Gibbs or bootstrap)

lightweight replacement for traditional alignment

Note: Salmon can work with either raw reads *or* with reads aligned to the transcriptome.

Phase 1: Online Inference (asynchronous!)



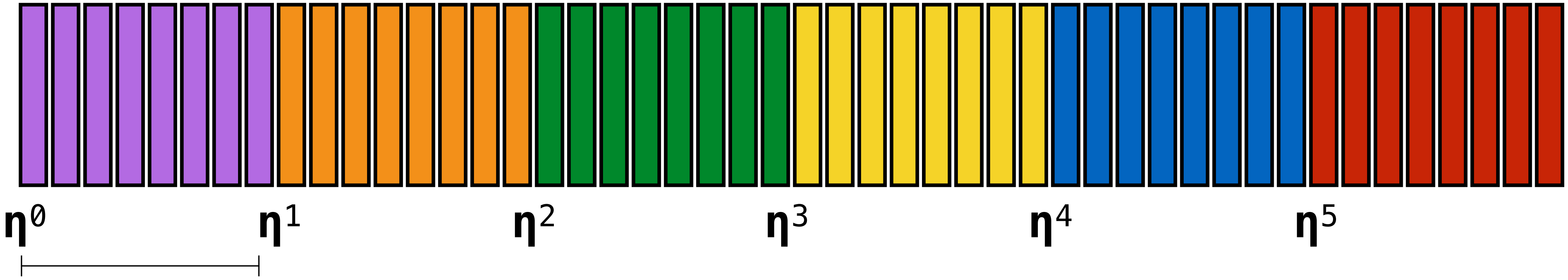
Online inference of abundance

Estimation of "foreground" bias models

Computation of equivalence class weights

online inference [SCVB0]*

Process fragments in batches:



Compute local **η**' using $\boldsymbol{\eta}^{t-1}$ & current “bias” model to allocate fragments

Update global nucleotide fractions: $\boldsymbol{\eta}^{t} = \boldsymbol{\eta}^{t-1} + a^{t}\,\boldsymbol{\eta}'$

Weighting factor that decays over time

Update “bias” model

Place mappings in **equivalence classes**

- Have access to ***all fragment-level information*** when making these updates
- Often converges very quickly.
- Compare-And-Swap (CAS) for synchronizing updates of different batches

*Based on: Foulds et al. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. ACM SIGKDD, 2013.

Broderick, Tamara, et al. "Streaming variational bayes." *Advances in Neural Information Processing Systems*. 2013.

Hsieh, Cho-Jui, Hsiang-Fu Yu, and Inderjit S. Dhillon. "PASSCoDe: Parallel ASynchronous Stochastic dual Co-ordinate Descent." *ICML*. Vol. 15. 2015.

Raman, Parameswaran, et al. "Extreme Stochastic Variational Inference: Distributed and Asynchronous." *arXiv preprint arXiv:1605.09499* (2016). (@*ICML* 2017)

```
Give each transcript appropriate prior mass $\eta^0$ (init.)
For each mini-batch $B^t$ of reads {

    For each read $r$ in $B^t$ {

        For each alignment $a$ of $r$ {
            compute (un-normalized) prob of $a$ using $\boldsymbol{\eta}^{t-1}$, and aux params
        }
        normalize alignment probs & update local transcript weights $\boldsymbol{\eta}'$
        add / update the equivalence class for read r
        sample a ∈ r to update auxiliary models
    }
    update global transcript weights given local transcript
    weights according to “update rule” $\Longrightarrow \boldsymbol{\eta}^t = \boldsymbol{\eta}^{t-1} + w^t \boldsymbol{\eta}'$
}
```

mini-batches processed in parallel by different threads

additive nature of updates mitigates effects of
no synchronization between mini-batches

Broderick, Tamara, et al. "Streaming variational bayes." *Advances in Neural Information Processing Systems*. 2013.

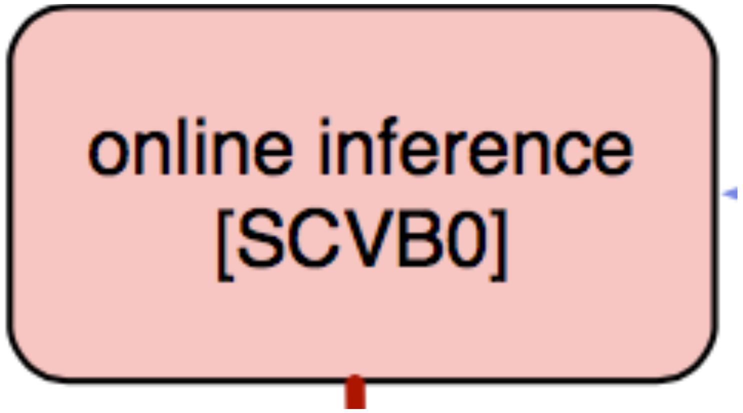
Hsieh, Cho-Jui, Hsiang-Fu Yu, and Inderjit S. Dhillon. "PASSCoDe: Parallel ASynchronous Stochastic dual Co-ordinate Descent." *ICML*. Vol. 15. 2015.

Raman, Parameswaran, et al. "Extreme Stochastic Variational Inference: Distributed and Asynchronous." *arXiv preprint arXiv:1605.09499* (2016). (@*ICML* 2017)

In this phase, we maintain *current* estimates of abundance.

Each group of fragments arrive (streaming), and we use their mapping locations & current estimates to:

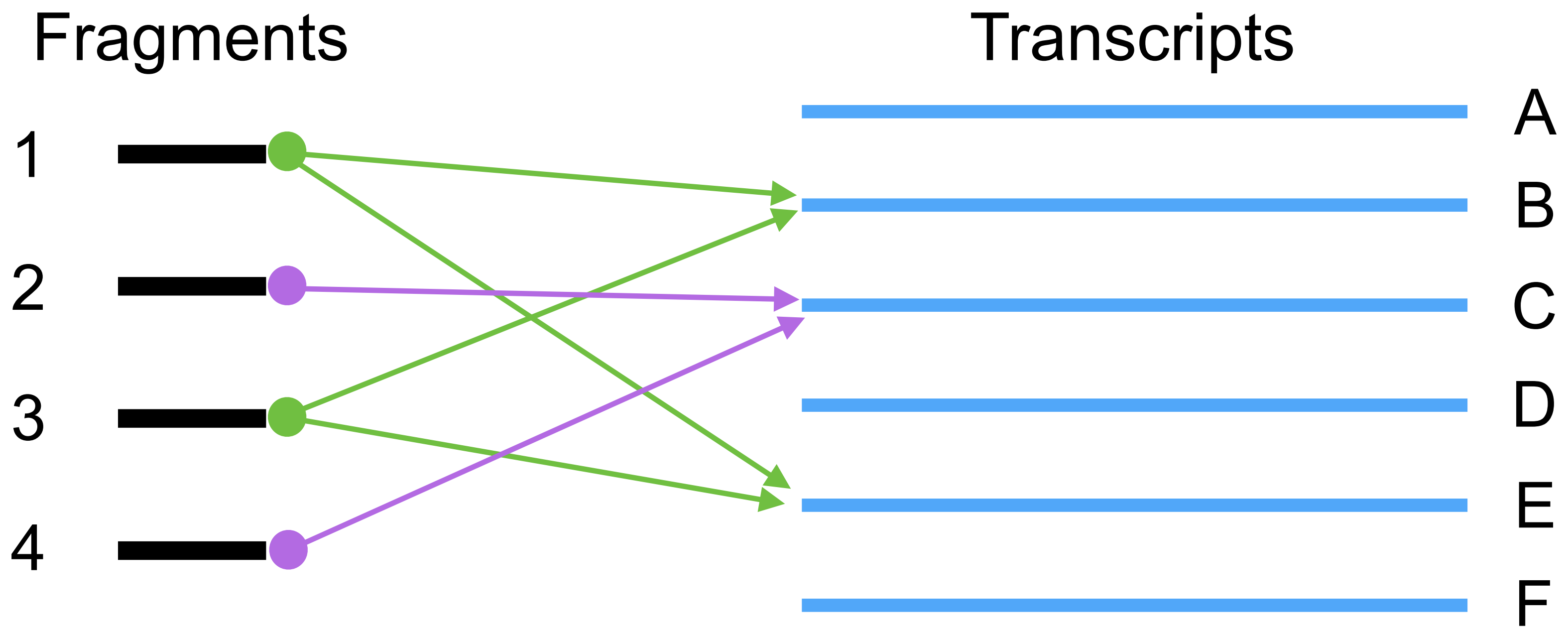
1. Allocate them to transcripts
2. Update auxiliary models
3. Place them in **equivalence classes**


online inference
[SCVB0]

We use a streaming, parallel, stochastic inference algorithm for Phase 1; a variant of **S**tochastic **C**ollapsed **V**ariational **B**ayesian Inference [SCVB0]*

* Foulds, James, et al. "Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

Fragment Equivalence Classes



Reads 1 & 3 both map to transcripts B & E
Reads 2 & 4 both map to transcript C

w^j_i encodes the "affinity" of class *j* to transcript *i* according to the model. This is $P\{f_j \mid t_i\}$, aggregated for all fragments in a class.

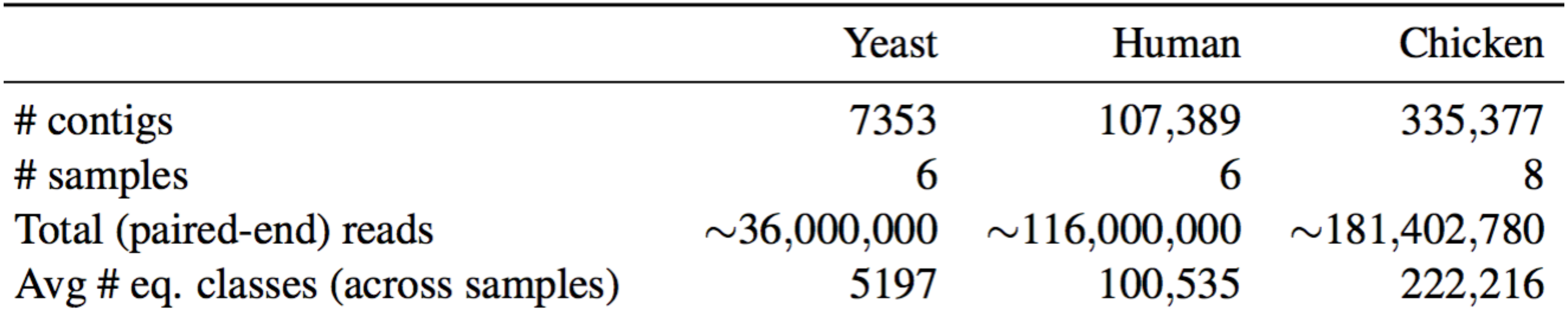
We have 4 reads, but only 2 eq. classes of reads

eq. Label	Count	Aux weights
{B,E}	2	$w^{\{B,E\}}_B, w^{\{B,E\}}_E$
{C}	2	$w^{\{C\}}_C$

This idea goes quite far back in the RNA-seq literature; at least to MMSeq (Turro et al. 2011)

Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome biology* 12.2 (2011): R13.

The number of equivalence classes is small



	Yeast	Human	Chicken
# contigs	7353	107,389	335,377
# samples	6	6	8
Total (paired-end) reads	∼36,000,000	∼116,000,000	∼181,402,780
Avg # eq. classes (across samples)	5197	100,535	222,216

The **# of equivalence classes grows with the complexity of the transcriptome** — independent of the # of sequence fragments.

Typically, ***two or more orders of magnitude*** fewer equivalence classes than sequenced fragments.

The offline **inference** algorithm **scales in # of fragment equivalence classes**.

This naturally handles different types of multi-mapping *without* having to rely on the annotation

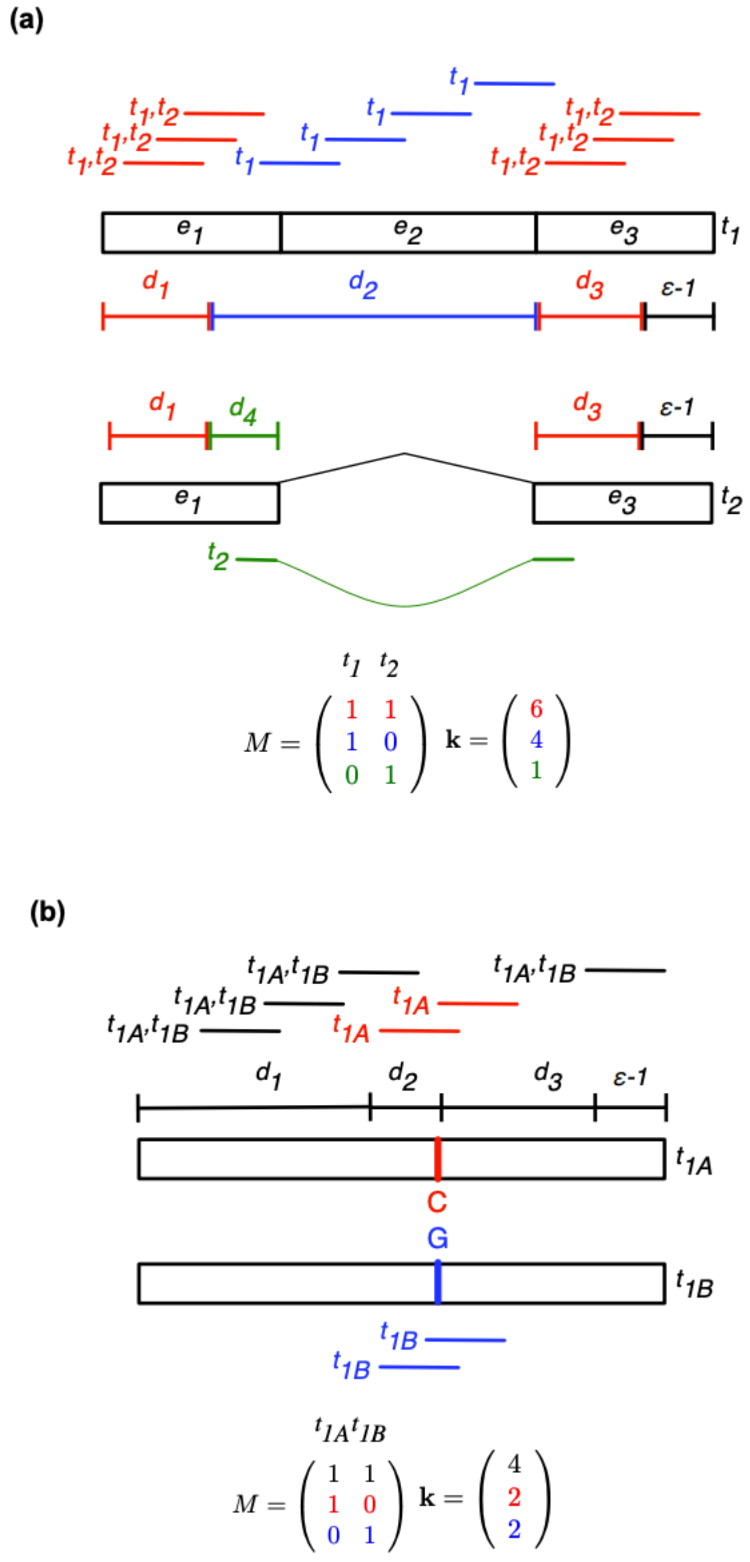

(a)
t_1,t_2
t_1
e_1
e_2
e_3
t_1
d_1
d_2
d_3
ε-1
d_1
d_4
d_3
ε-1
e_1
e_3
t_2
t_2
$M = \begin{array}{c} \begin{array}{cc} t_1 & t_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$ $\mathbf{k} = \begin{pmatrix} 6 \\ 4 \\ 1 \end{pmatrix}$
(b)
t_{1A},t_{1B}
t_{1A}
d_1
d_2
d_3
ε-1
t_{1A}
C
G
t_{1B}
t_{1B}
$M = \begin{array}{c} \begin{array}{cc} t_{1A} & t_{1B} \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$ $\mathbf{k} = \begin{pmatrix} 4 \\ 2 \\ 2 \end{pmatrix}$

Figure 2 from Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome biology* 12.2 (2011): R13.

This lets us approximate the likelihood efficiently

Approximate this:

sum over all alignments of fragment

$$\mathcal{L}\left(\boldsymbol{\eta};\mathcal{F}\right) = \prod_{f_j\in\mathcal{F}} \sum_{i=1}^{M} \Pr\left(t_i \mid \boldsymbol{\eta}\right) \Pr\left(f_j \mid t_i\right)$$

product over all fragments

with this:

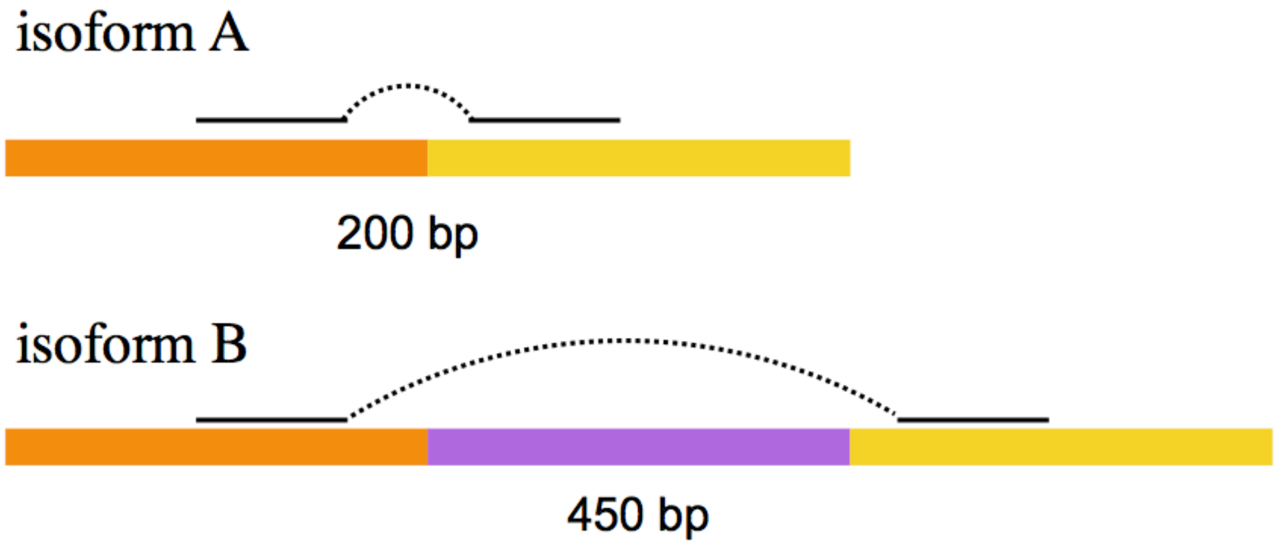
$$\mathcal{L}\left(\boldsymbol{\eta};\mathcal{F}\right) \approx \prod_{\mathcal{F}^q\in\boldsymbol{\mathcal{C}}} \left(\sum_{\langle i,t_i\rangle\in\Omega(\mathcal{F}^q)} \Pr\left(t_i \mid \boldsymbol{\eta}\right) \cdot \Pr\left(f \mid \mathcal{F}^q, t_i\right) \right)^{N^q}$$

sum over all transcripts labeling this eq. class

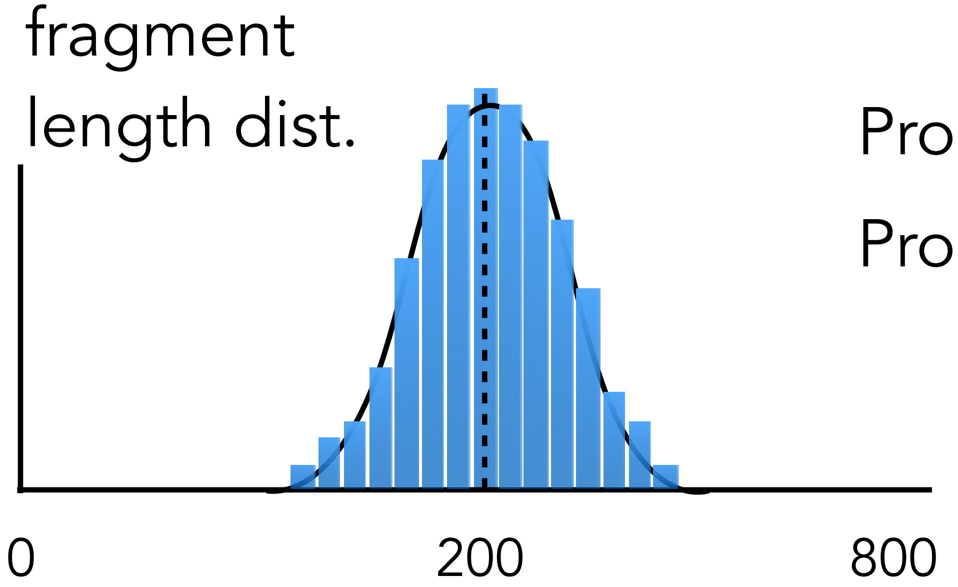
product over all equivalence classes

Why might $Pr(f_j \mid t_i)$ matter?

Consider the following scenario:



Conditional probabilities can provide valuable information about origin of a fragment! ***Potentially different for each transcript/fragment pair.***



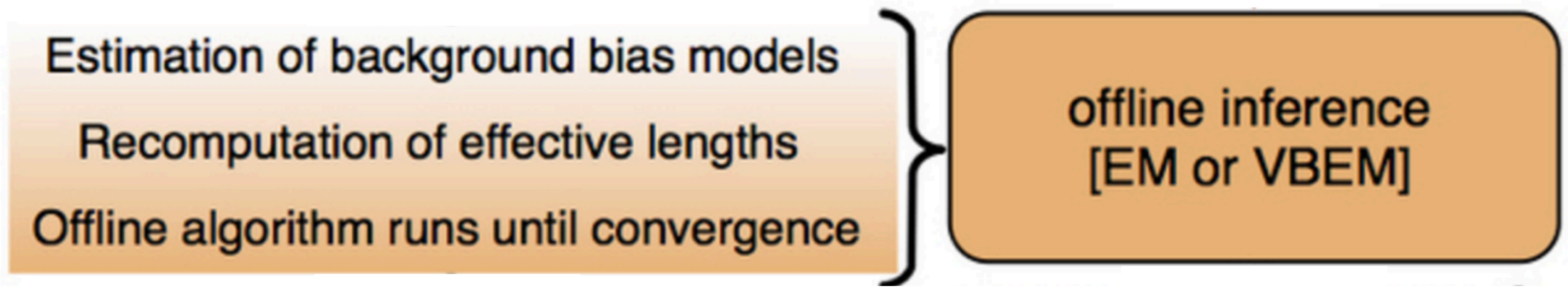
Prob of observing a fragment of size ~200 is **large**

Prob of observing a fragment of size ~450 is **small**

Many terms can be considered in a general “fragment-transcript agreement” model[1]. e.g. position, orientation, alignment path etc.

[1] “Salmon provides fast and bias-aware quantification of transcript expression”, Nature Methods 2017

Optimizing the objective



Estimation of background bias models

Recomputation of effective lengths

Offline algorithm runs until convergence

offline inference
[EM or VBEM]

our ML objective has a simple, **closed-form update rule** in terms of our eq. classes

count of eq.
class j

$$\alpha_i^{u+1} = \sum_{\mathcal{F}^q \in \boldsymbol{\mathcal{C}}} N^q \left(\frac{\alpha_i^u w_i^q}{\sum_{\langle k, t_k \rangle \in \Omega(\mathcal{F}^q)} \alpha_k^u w_k^q} \right)$$

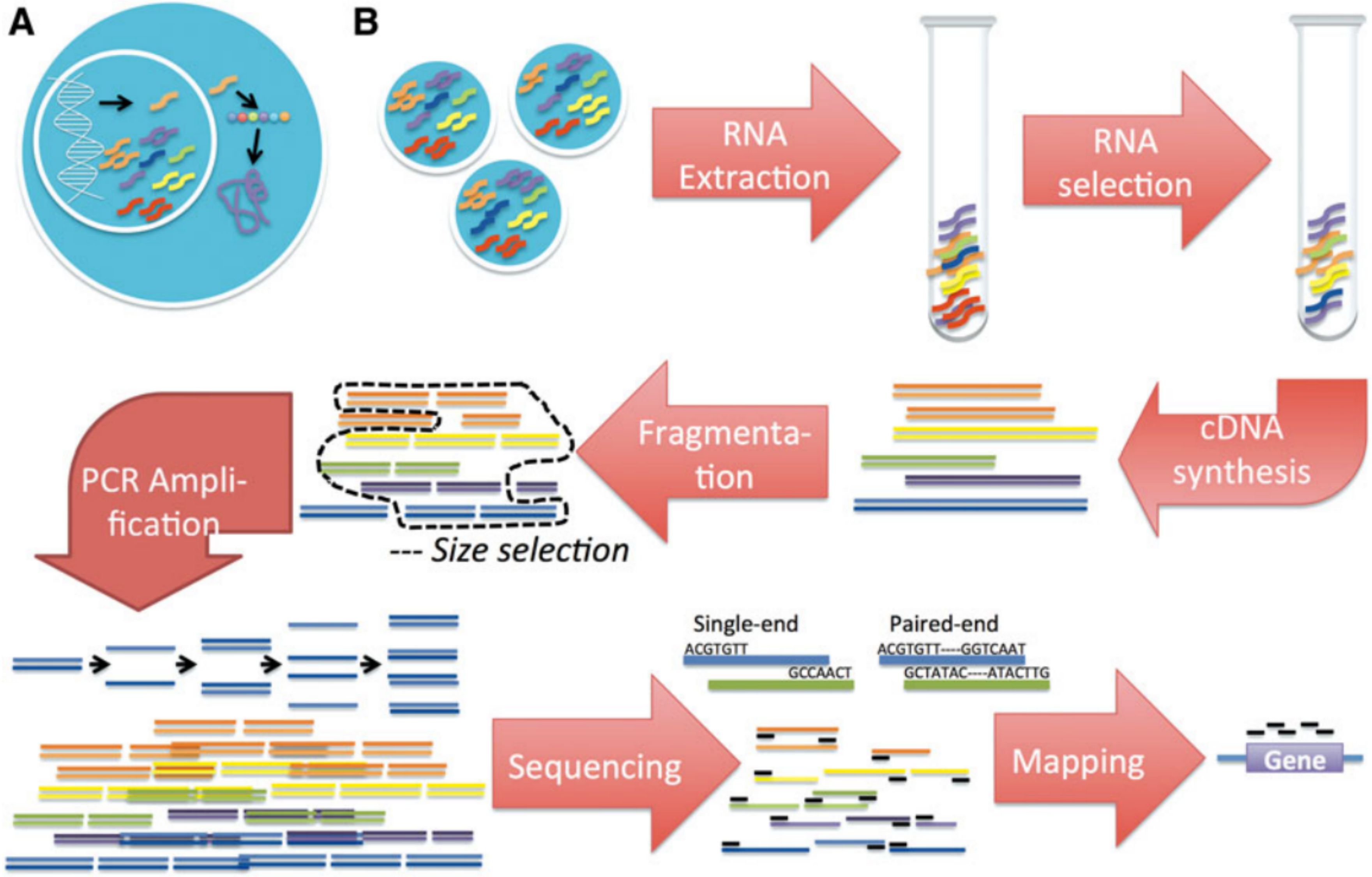
weight of t_i in eq.
class q

estimated read count from transcript i
at iteration u+1

$$\hat{\eta}_i = \frac{\alpha_i}{\sum_j \alpha_j}$$

we also provide the *option* to use a **variational Bayesian** objective instead

Actual RNA-seq protocols are a bit more “involved”



There is **substantial** potential for biases and deviations from the *basic* model — indeed, we see quite a few.

Prakash, Celine, and Arndt Von Haeseler. "An Enumerative Combinatorics Model for Fragmentation Patterns in RNA Sequencing Provides Insights into Nonuniformity of the Expected Fragment Starting-Point and Coverage Profile." *Journal of Computational Biology* 24.3 (2017): 200-212.

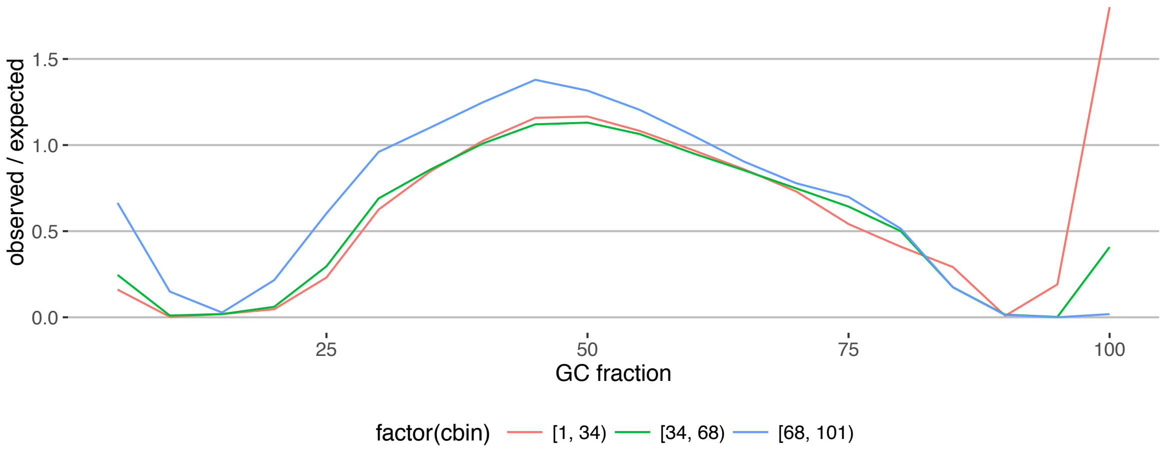
Biases abound in RNA-seq data

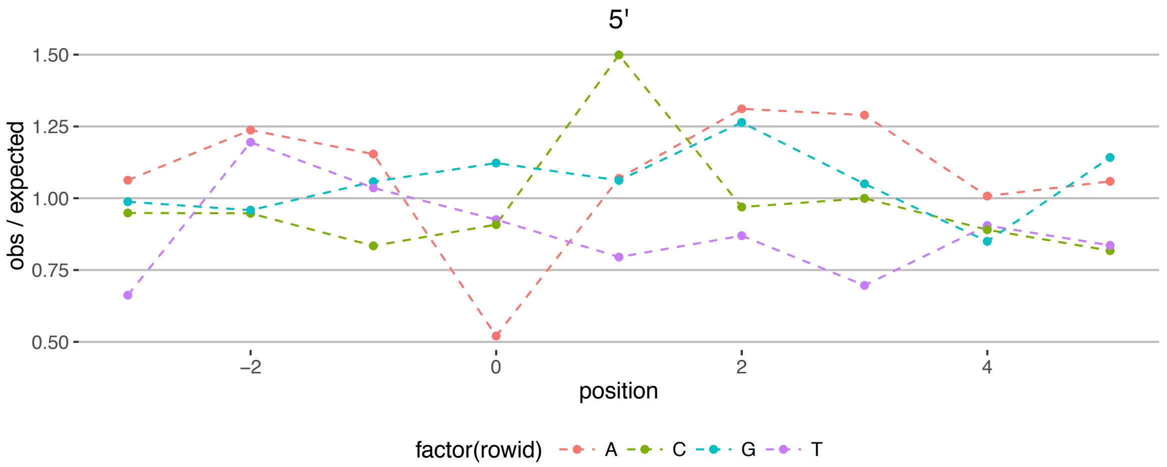
Biases in prep & sequencing can have a significant effect on the fragments we see:

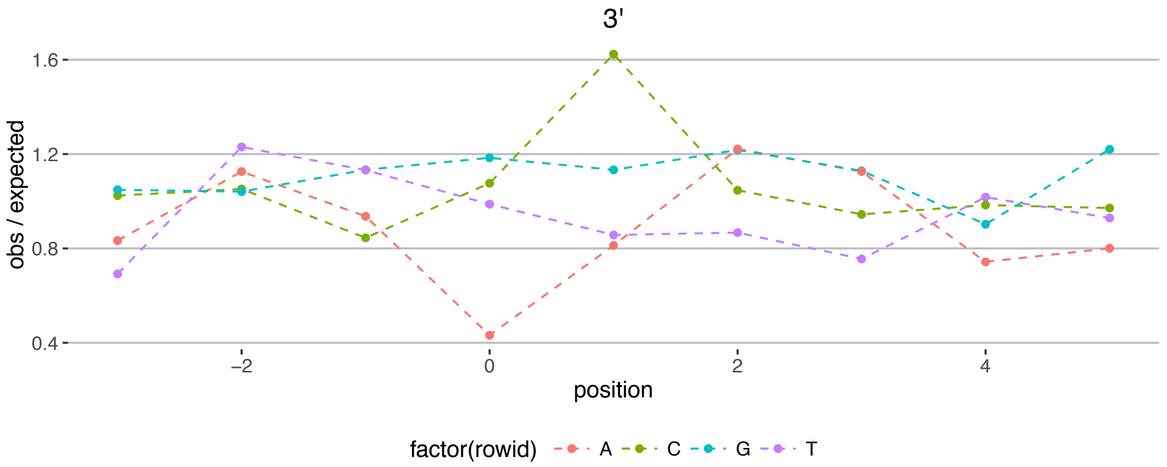
Fragment gc-bias[1]— The GC-content of the fragment affects the likelihood of sequencing

Sequence-specific bias[2]— sequences surrounding fragment affect the likelihood of sequencing

Positional bias[2]— fragments sequenced non-uniformly across the body of a transcript







[1] Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.

[2] Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Basic idea (1): Modify the "effective length" of a transcript to account for changes in the sampling probability. This leads to changes in soft-assignment in EM -> changes in TPM.

Basic idea (2):The effective length of a transcript is the sum of the bias terms at each position across a transcript. The bias term at a given position is simply the (observed / expected) sampling probability.

The trick is how to define "expected" given only biased data.

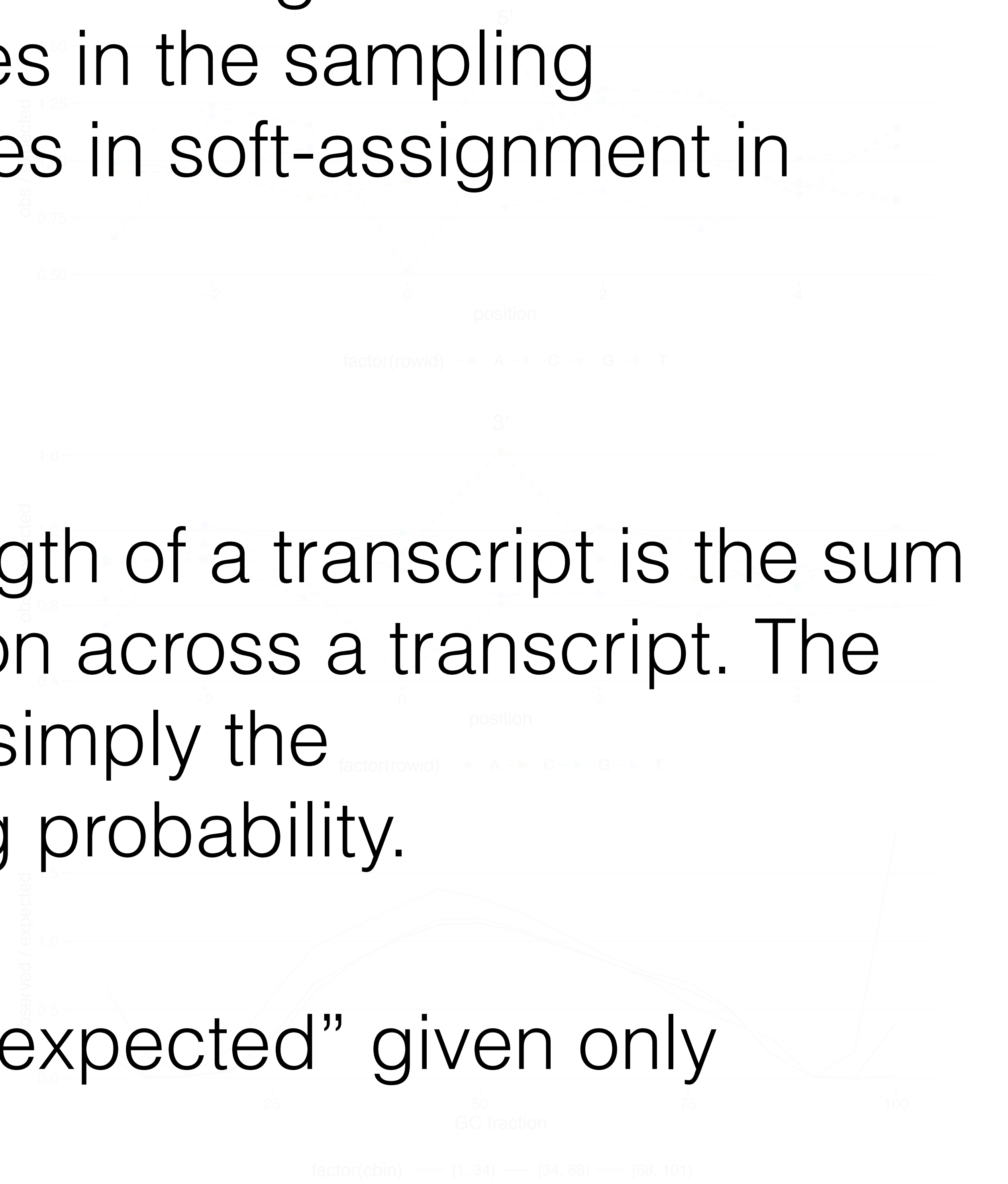
Biases abound in RNA-seq data

Biases in cDNA & sequencing can have a significant effect on the fragments we see:

Fragment gc-bias[1]— The GC-content of the fragment affects the likelihood of sequencing

sequences surrounding fragment affect the likelihood of sequencing

Positional bias[2]— fragments sequenced non-uniformly across the body of a transcript



[1] Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.

[2] Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Bias Modeling

Bias correction works by adjusting the effective lengths of the transcripts: The effective length becomes the sum of the per-base biases

$$\tilde{\ell}'_i = \sum_{j=1}^{j \le \ell_i} \sum_{k=1}^{k \le f_i(j,L)} \frac{b_{gc^+}\left(t_i, j, j+k\right)}{b_{gc^-}\left(t_i, j, j+k\right)} \cdot \frac{b^{5'}_{s+}\left(t_i, j\right)}{b^{5'}_{s-}\left(t_i, j\right)} \cdot \frac{b^{3'}_{s+}\left(t_i, j+k\right)}{b^{3'}_{s-}\left(t_i, j+k\right)} \cdot \frac{b^{5'}_{p+}\left(t_i, j+k\right)}{b^{5'}_{p-}\left(t_i, j+k\right)} \cdot \frac{b^{3'}_{p+}\left(t_i, j+k\right)}{b^{3'}_{p-}\left(t_i, j+k\right)} \cdot \Pr\left\{X = j\right\}$$

Fragment GC bias model:

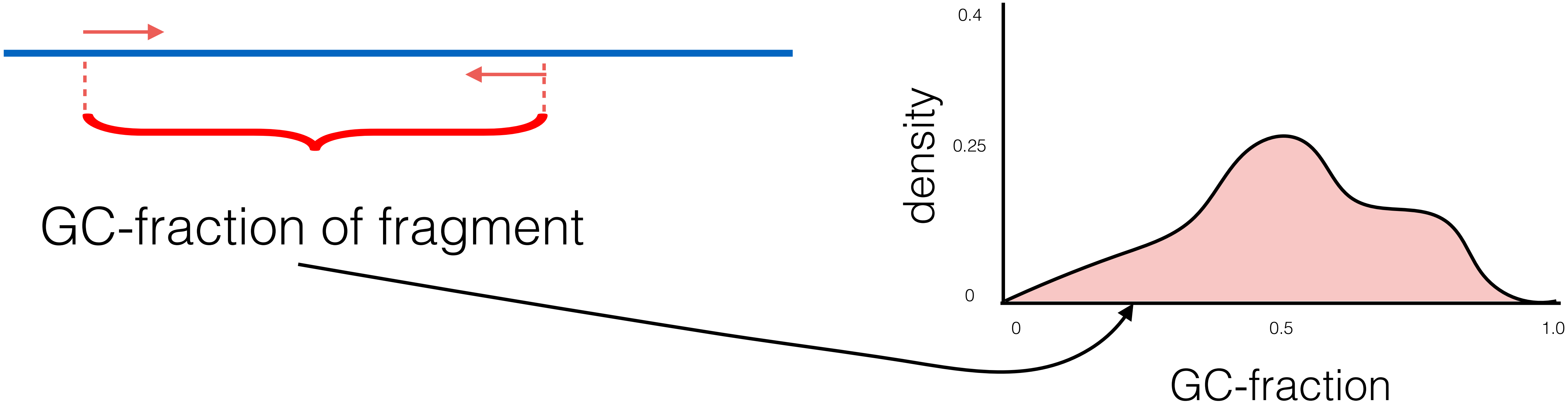
Density of fragments with specific GC content, **conditioned** on GC fraction at read start/end

Foreground:

Observed

Background:

Expected given est. abundances



First explored in Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." *Nature biotechnology* 34.12 (2016): 1287.

Bias Modeling

Bias correction works by adjusting the effective lengths of the transcripts: The effective length becomes the sum of the per-base biases

$$\tilde{\ell}'_i = \sum_{j=1}^{j\le \ell_i} \sum_{k=1}^{k\le f_i(j,L)} \frac{b_{gc^+}(t_i,j,j+k)}{b_{gc^-}(t_i,j,j+k)} \cdot \frac{b^{5'}_{s+}(t_i,j)}{b^{5'}_{s-}(t_i,j)} \cdot \frac{b^{3'}_{s+}(t_i,j+k)}{b^{3'}_{s-}(t_i,j+k)} \cdot \frac{b^{5'}_{p+}(t_i,j+k)}{b^{5'}_{p-}(t_i,j+k)} \cdot \frac{b^{3'}_{p+}(t_i,j+k)}{b^{3'}_{p-}(t_i,j+k)} \cdot \Pr\{X=j\}$$

Seq-specific bias model*:

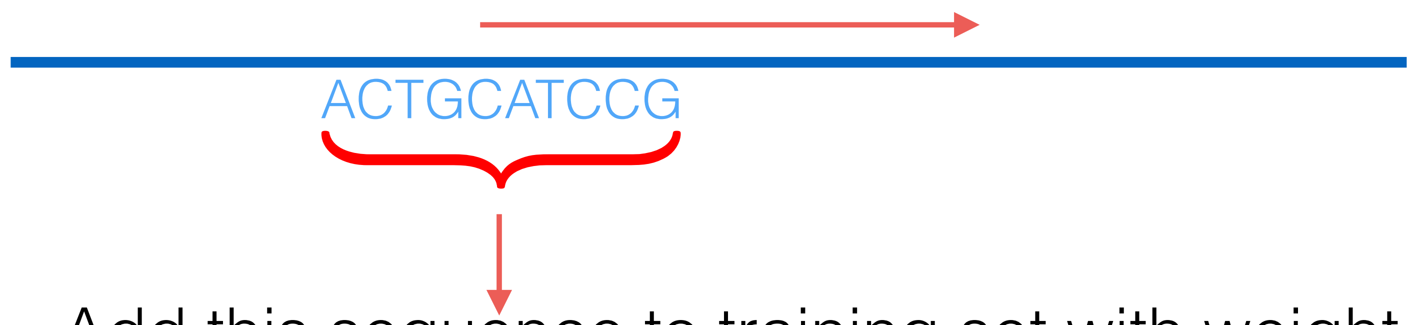
VLMM for the 10bp window surrounding the 5' read start site and the 3' read start site

Foreground:

Observed

Background:

Expected given est. abundances

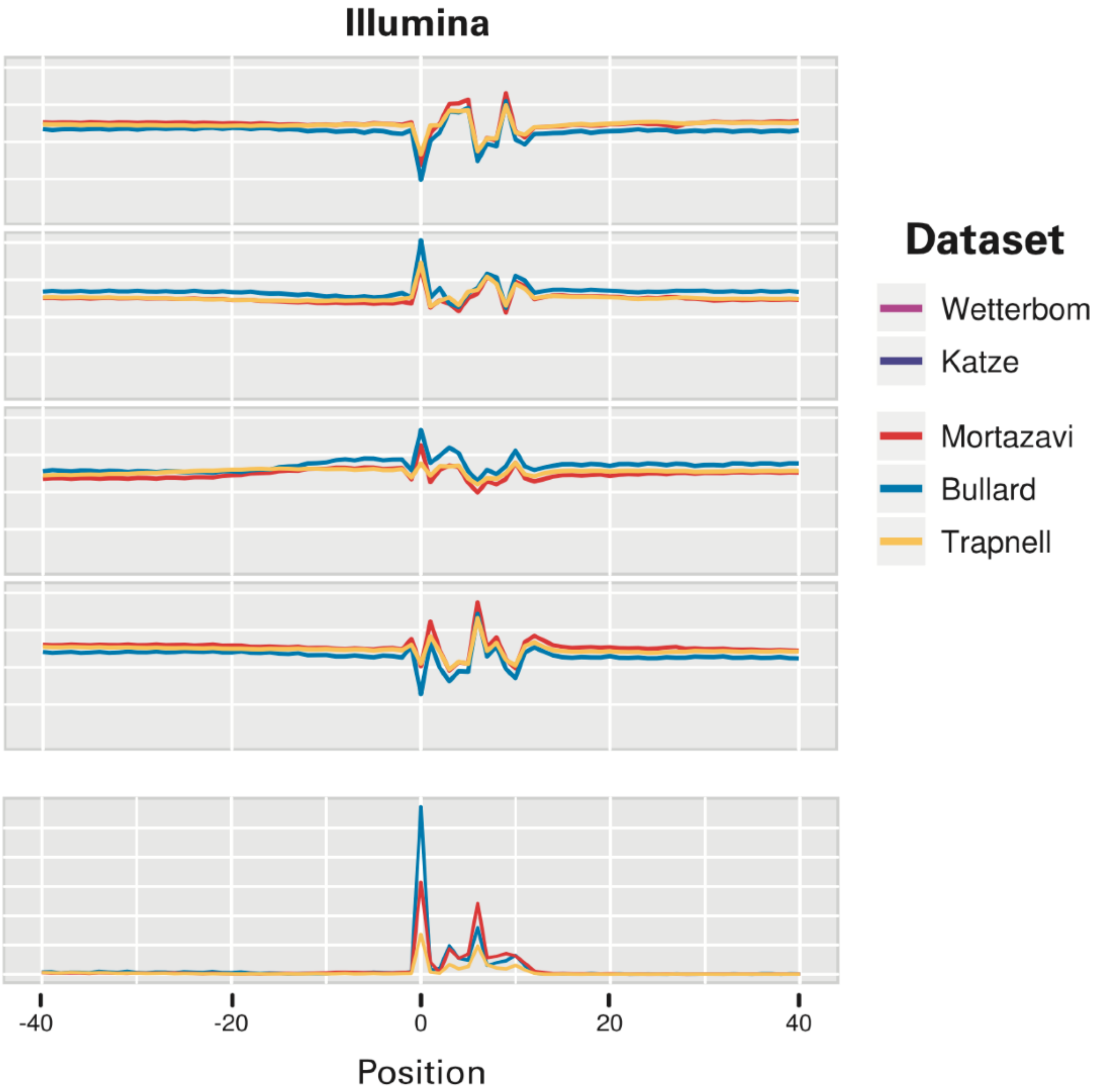


Add this sequence to training set with weight = P{f | t_i}

Same, but independent model for 3' end

*Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Priming bias is sample & sequence-specific



Jones, Daniel C., et al. "A new approach to bias correction in RNA-Seq." *Bioinformatics* 28.7 (2012): 921-928.

Basic idea

The sequencing is unbiased w.r.t. sequence context if

$$E[x_i \mid s_i] = N\,\Pr[m_i \mid s_i] = N\,\Pr[m_i] = E[x_i]$$

Expected read count at pos i *conditioned* on sequence = Unconditional expected read count at pos i

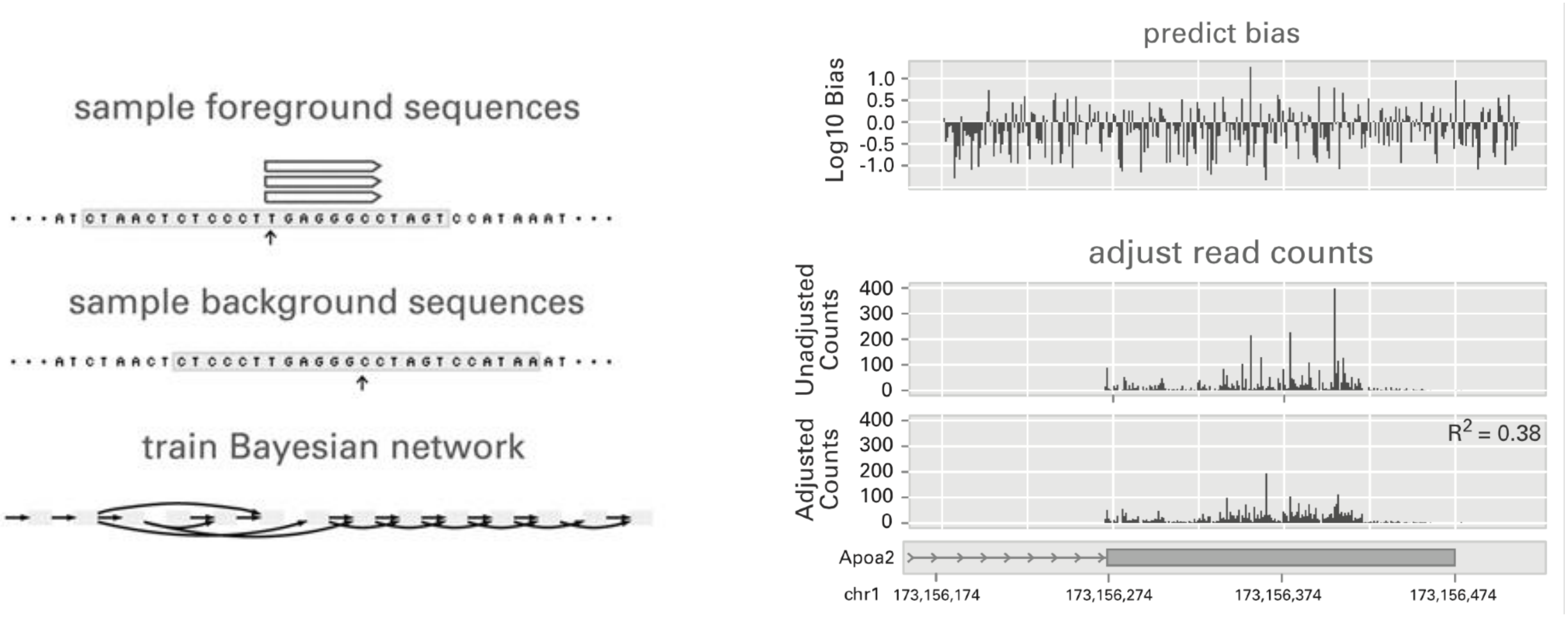
Define the sequence bias as: $b_i = \Pr[s_i] / \Pr[s_i \mid m_i]$

So that:

$$E[b_i x_i \mid s_i] = b_i E[x_i \mid s_i] = N b_i \Pr[m_i | si] \quad = N \frac{\Pr[m_i \mid s_i]\Pr[s_i]}{\Pr[s_i \mid s_i]} = N\,\Pr[m_i] = E[x_i]$$

Jones, Daniel C., et al. "A new approach to bias correction in RNA-Seq." *Bioinformatics* 28.7 (2012): 921-928.

Priming bias is sample & sequence-specific



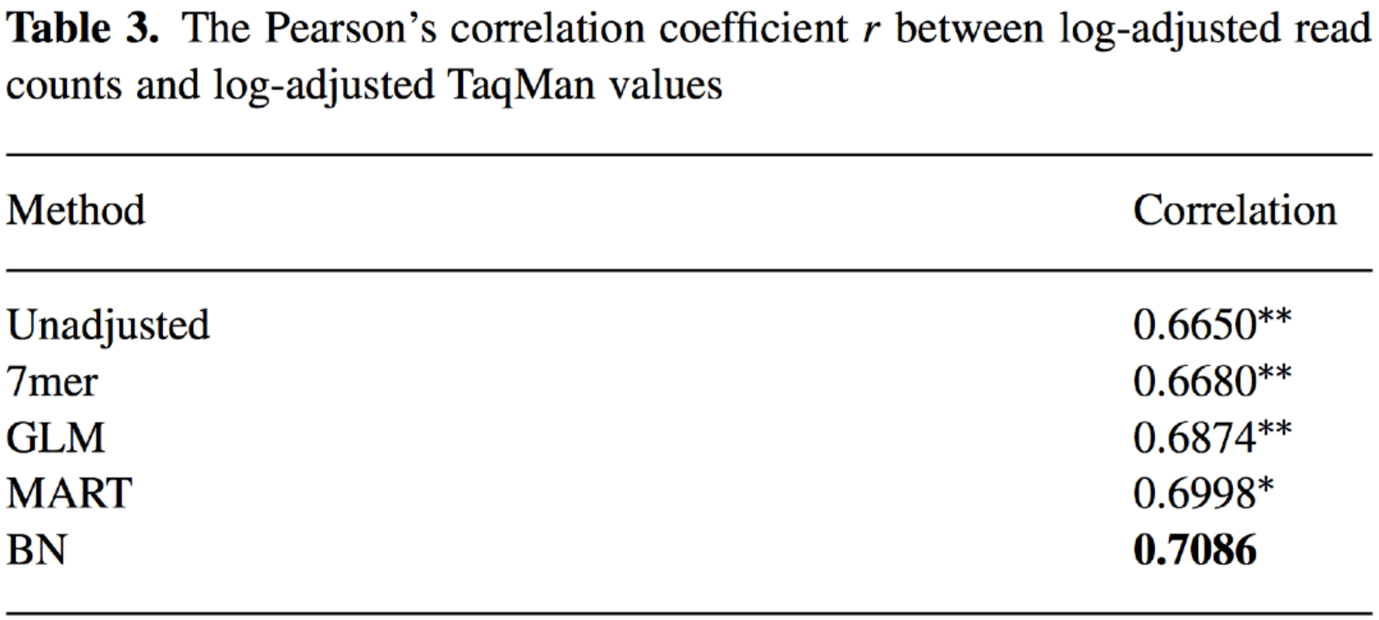
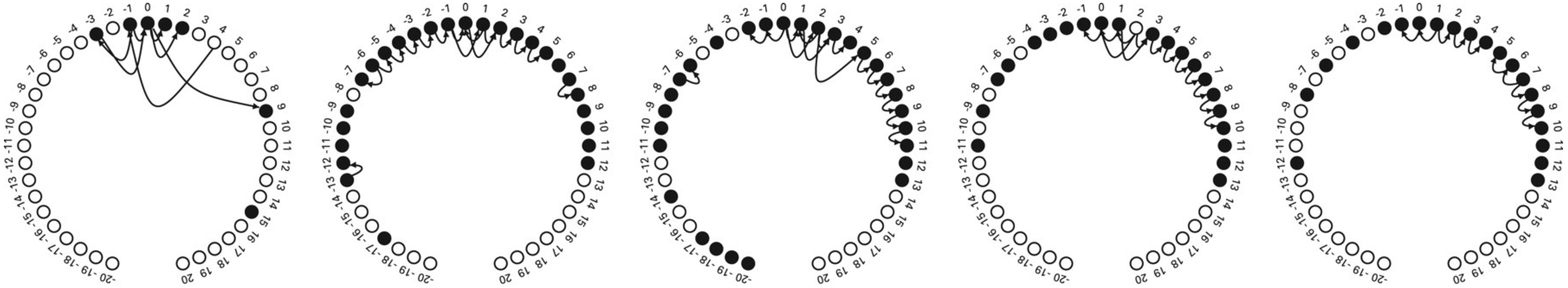


Table 3. The Pearson's correlation coefficient *r* between log-adjusted read counts and log-adjusted TaqMan values

Method	Correlation
Unadjusted	0.6650**
7mer	0.6680**
GLM	0.6874**
MART	0.6998*
BN	**0.7086**

Jones, Daniel C., et al. "A new approach to bias correction in RNA-Seq." *Bioinformatics* 28.7 (2012): 921-928.

The best *model* may also be sample-specific



Wetterbom (282 parameters) **Katze (684 parameters)** **Bullard (696 parameters)** **Mortazavi (582 parameters)** **Trapnell (360 parameters)**

Contrast with Roberts et al. which uses a fixed-structure

VLMM to model the sample-specific bias.

Jones, Daniel C., et al. "A new approach to bias correction in RNA-Seq." *Bioinformatics* 28.7 (2012): 921-928.

Bias Modeling

Bias correction works by adjusting the effective lengths of the transcripts:
The effective length becomes the sum of the per-base biases

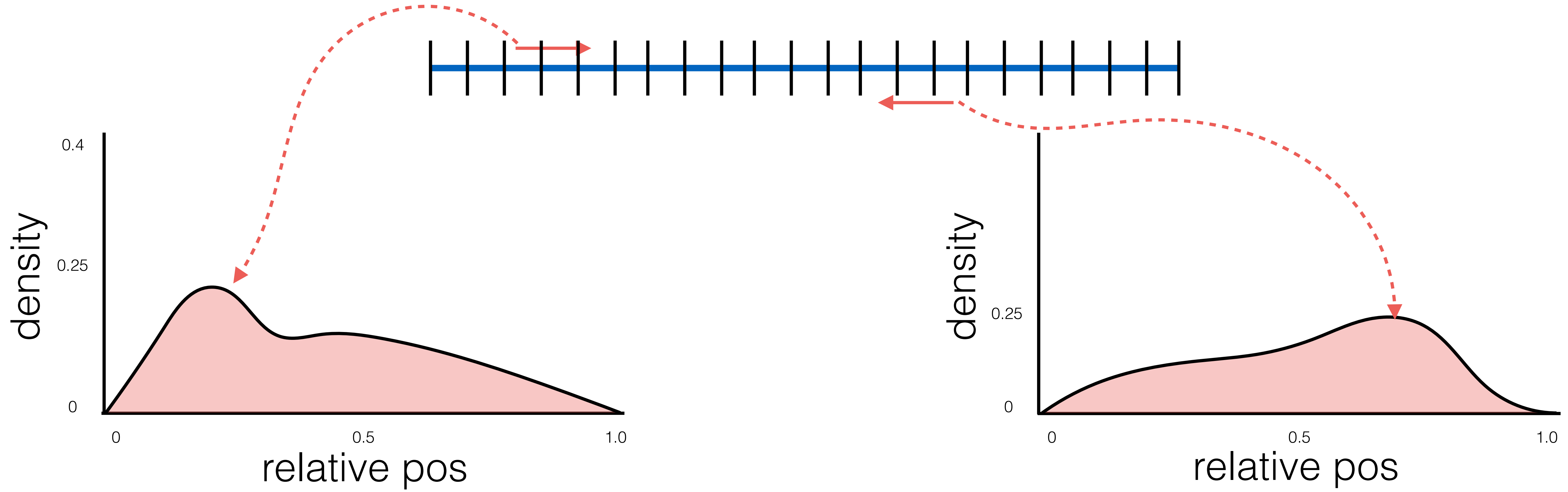
$$\tilde{\ell}'_i = \sum_{j=1}^{j \leq \ell_i} \sum_{k=1}^{k \leq f_i(j,L)} \frac{b_{gc^+}(t_i, j, j+k)}{b_{gc^-}(t_i, j, j+k)} \cdot \frac{b^{5'}_{s+}(t_i, j)}{b^{5'}_{s-}(t_i, j)} \cdot \frac{b^{3'}_{s+}(t_i, j+k)}{b^{3'}_{s-}(t_i, j+k)} \cdot \frac{b^{5'}_{p+}(t_i, j+k)}{b^{5'}_{p-}(t_i, j+k)} \cdot \frac{b^{3'}_{p+}(t_i, j+k)}{b^{3'}_{p-}(t_i, j+k)} \cdot \Pr\{X = j\}$$

Position bias model*:

Foreground:
Observed

Density of 5’ and 3’ read start positions —
different models for transcripts of different length

Background:
Expected given est. abundances



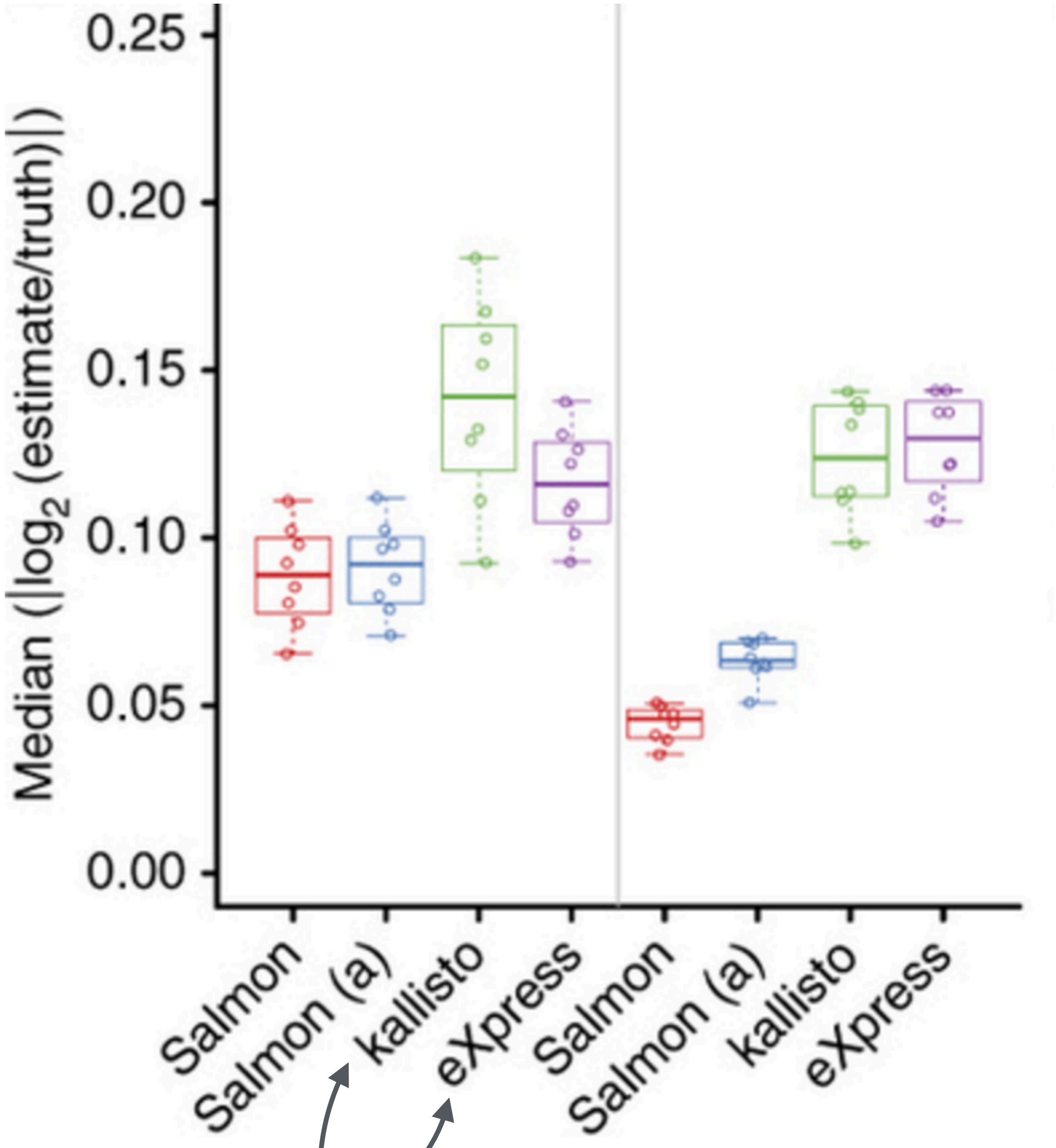
*Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Accuracy difference can be larger with biased data

Simulated data:
2 conditions; 8 samples each

- Simulated transcripts across entire genome with known abundance using Polyester (modified to account for GC bias)
- How well do we recover the underlying relative abundances?
- How does accuracy vary with level of bias?

Lower is better



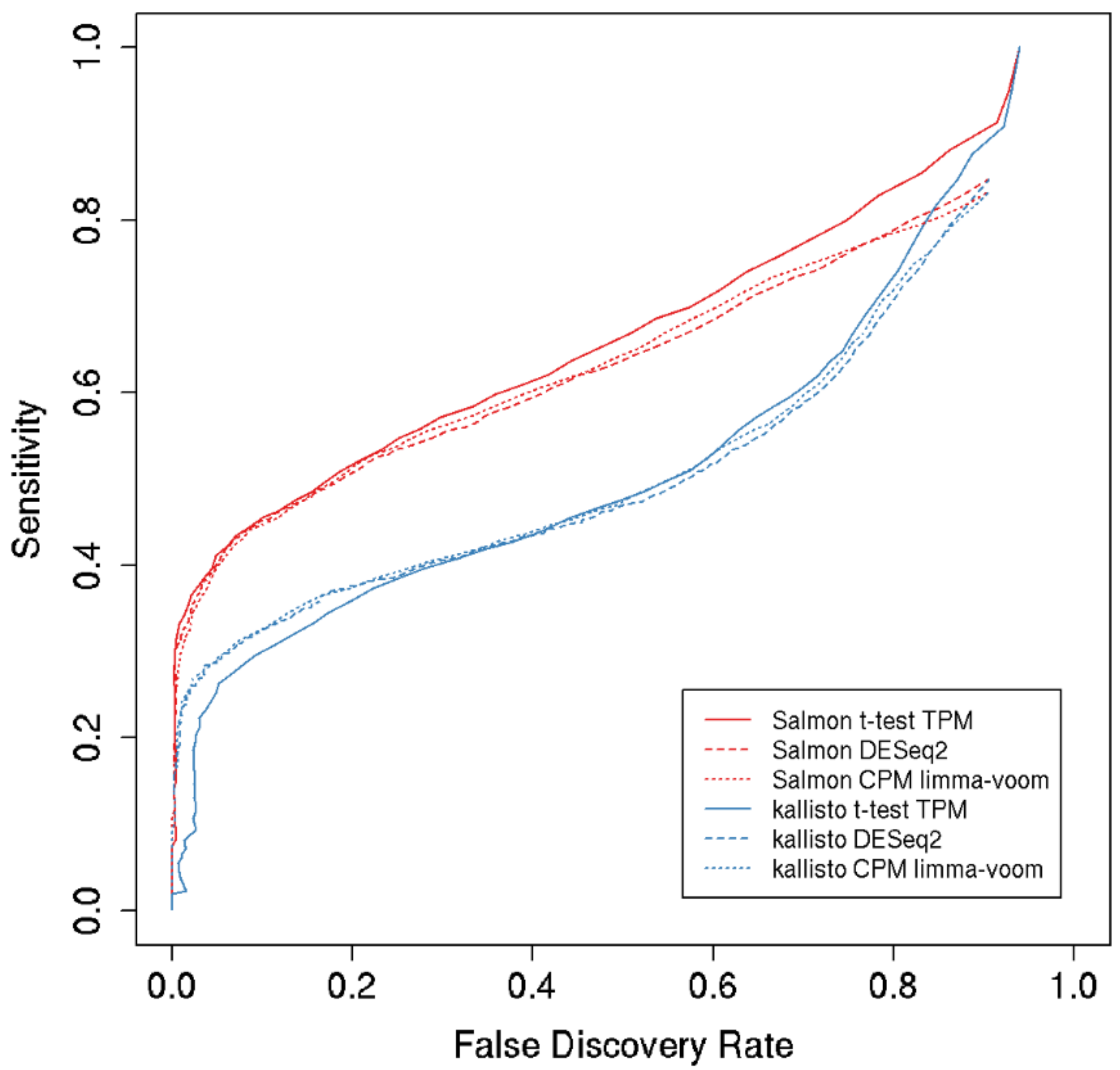
Sequence-bias models don’t account for fragment-level GC bias

Mis-estimates confound downstream analysis

Simulated data:
2 conditions; 8 replicates each

- set 10% of txps to have fold change of 1/2 or 2 — rest unchanged.
- How well do we recover true DE?
- Since bias is systematic, effect may be even worse than accuracy difference suggests.

Recovery of DE transcripts



Importance with experimental data

30 samples from the GEUVADIS study:
15 samples from UNIGE sequencing center
15 samples from CNAG_CRG sequencing center

Same human population, expect few-to-no *real* DE

Randomized condition assignments result it << 1 DE txp

DE of data between centers (FDR < 1%) (TPM > 0.1)

	Salmon	Kallisto	eXpress
All transcripts	**1,183**	2,620	2,472
Transcripts of 2 isoform genes	**228**	545	531

Bias and **batch effects** are ***substantial***, and must be accounted for.

Importance with experimental data

30 samples from the GEUVADIS study:
15 samples from UNIGE sequencing center
15 samples from CNAG_CRG sequencing center

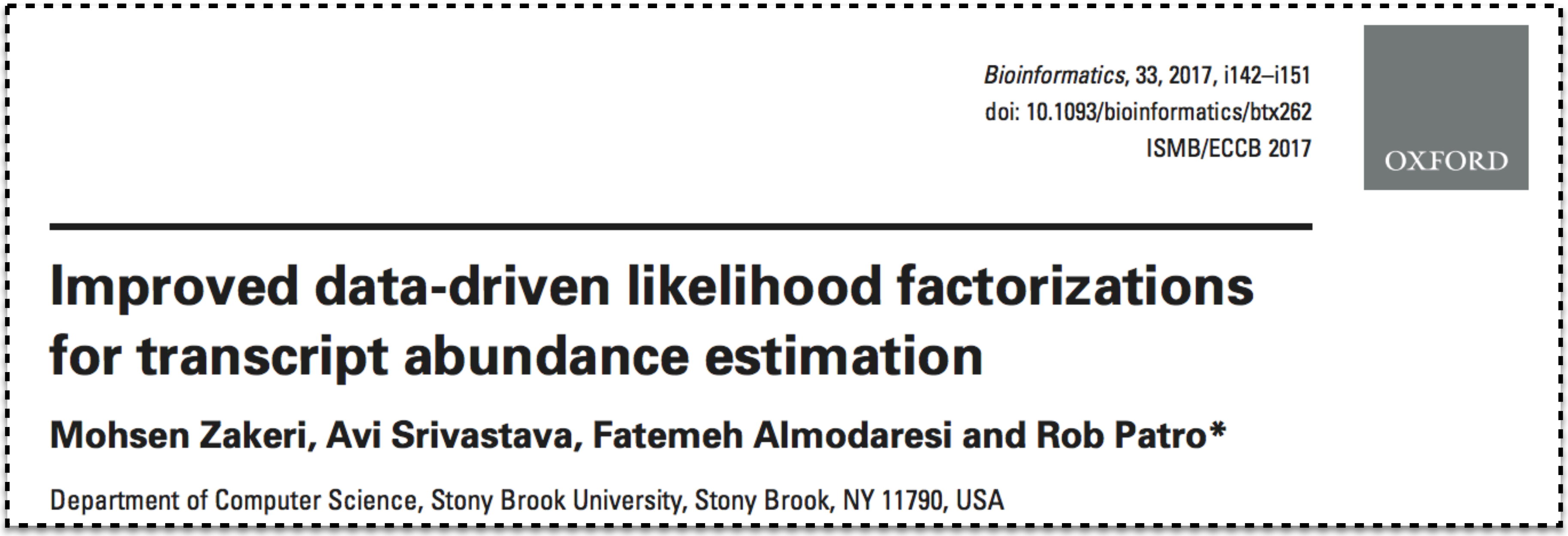
Effects seem **at least as extreme** at the gene level

DE of data between centers (FDR < 1%) (TPM > 0.1)

	Salmon	Kallisto	eXpress
All genes	**455**	1,200	1,582
Transcripts of 2 isoform genes	**224**	545	531

Bias and **batch effects** are ***substantial***, and must be accounted for.

Further improving the factorization (at low computational cost)



Bioinformatics, 33, 2017, i142–i151
doi: 10.1093/bioinformatics/btx262
ISMB/ECCB 2017

Improved data-driven likelihood factorizations for transcript abundance estimation

Mohsen Zakeri, Avi Srivastava, Fatemeh Almodaresi and Rob Patro*

Department of Computer Science, Stony Brook University, Stony Brook, NY 11790, USA

A probabilistic view of RNA-Seq quantification

We want to find the values of **η** that ***maximize*** this probability. We can do this (at least locally) using the EM algorithm.

but

This leads to an iterative EM algorithm where *each iteration* scales in the total number of **alignments** in the sample (typically on the order of 10^7 — 10^8), and typically 10^2—10^3 **iterations**

$$\mathcal{L}(\boldsymbol{\eta}; \mathcal{F}, \mathcal{T}) = \boxed{\prod_{f \in \mathcal{F}} \sum_{t_i \in \Omega(f)}} \Pr(t_i \mid \boldsymbol{\eta}) \Pr(f \mid t_i)$$

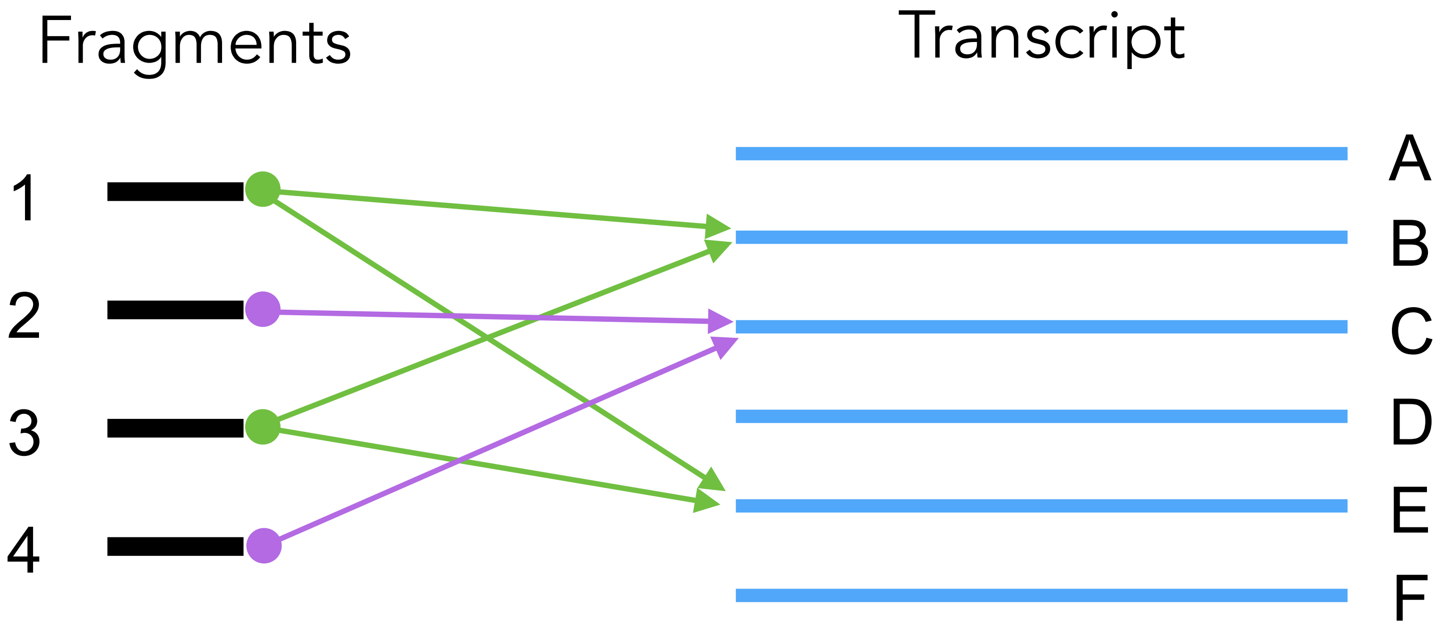
Set of transcripts where f maps/aligns

*Li, Bo, and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC bioinformatics 12.1 (2011): 1.

Recall : Fragment Equivalence Classes

$$f \sim f' \iff \Omega(f) = \Omega(f')$$

$$\Omega(f) = \{t \mid f \text{ maps to } t\}$$



Reads 1 & 3 both map to transcripts B & E

Reads 2 & 4 both map to transcript C

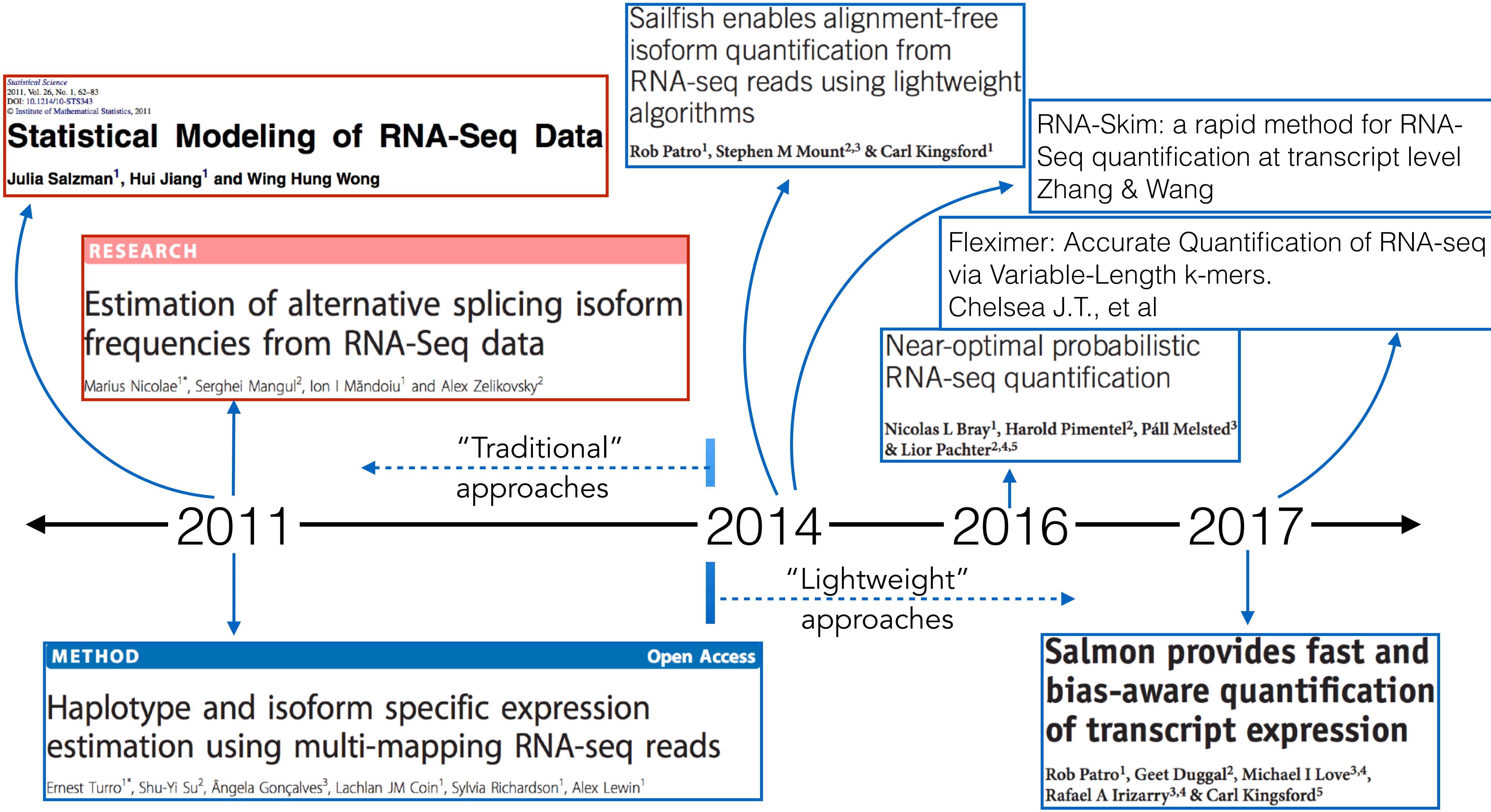
We have 4 reads, but only 2 eq. classes/types of reads

eq. Label	Count
{B,E}	2
{C}	2

Equivalence classes in RNA-Seq quantification

Long history of this idea — collapsing “redundant” reads

This list is not-complete (just illustrative)



Statistical Science
2011, Vol. 26, No. 1, 62–83
DOI: 10.1214/10-STS343
© Institute of Mathematical Statistics, 2011

Statistical Modeling of RNA-Seq Data

Julia Salzman[1], Hui Jiang[1] and Wing Hung Wong

RESEARCH

Estimation of alternative splicing isoform frequencies from RNA-Seq data

Marius Nicolae[1*], Serghei Mangul[2], Ion I Măndoiu[1] and Alex Zelikovsky[2]

METHOD Open Access

Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads

Ernest Turro[1*], Shu-Yi Su[2], Ângela Gonçalves[3], Lachlan JM Coin[1], Sylvia Richardson[1], Alex Lewin[1]

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms

Rob Patro[1], Stephen M Mount[2,3] & Carl Kingsford[1]

RNA-Skim: a rapid method for RNA-Seq quantification at transcript level
Zhang & Wang

Fleximer: Accurate Quantification of RNA-seq via Variable-Length k-mers.
Chelsea J.T., et al

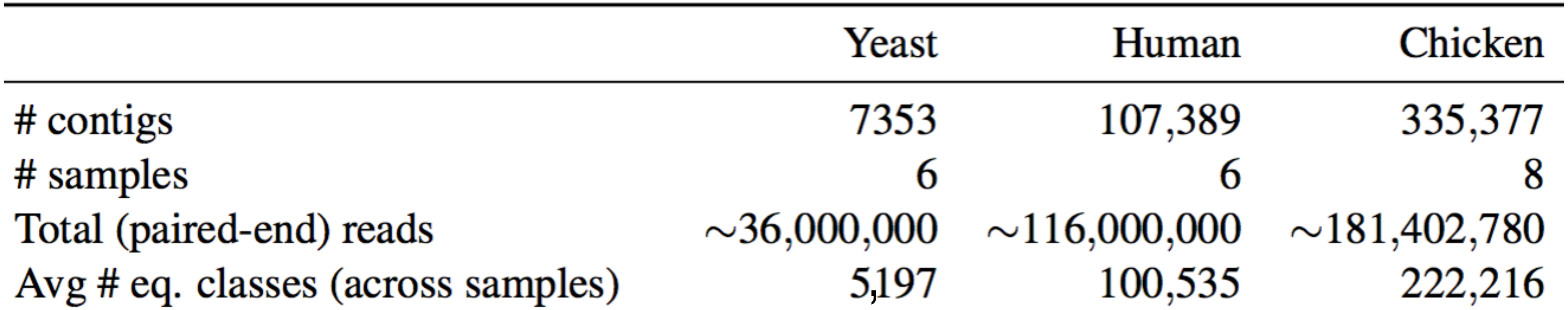
Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray[1], Harold Pimentel[2], Páll Melsted[3] & Lior Pachter[2,4,5]

Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro[1], Geet Duggal[2], Michael I Love[3,4], Rafael A Irizarry[3,4] & Carl Kingsford[5]

The number of equivalence classes is small



	Yeast	Human	Chicken
# contigs	7353	107,389	335,377
# samples	6	6	8
Total (paired-end) reads	∼36,000,000	∼116,000,000	∼181,402,780
Avg # eq. classes (across samples)	5,197	100,535	222,216

# of equivalence classes grows with the complexity of the transcriptome — not (asymptotically) with the # of sequence fragments.

Typically, ***two or more orders of magnitude*** fewer equivalence classes than sequenced fragments.

The **inference** algorithm **scales in # of fragment equivalence classes**.

This lets us approximate the likelihood efficiently

Approximate this:

sum over all alignments of fragment

$$\mathcal{L}(\boldsymbol{\eta}; \mathcal{F}, \mathcal{T}) = \prod_{f \in \mathcal{F}} \sum_{t_i \in \Omega(f)} \Pr(t_i \mid \boldsymbol{\eta}) \Pr(f \mid t_i)$$

product over all fragments

with this:

$$\mathcal{L}\left(\boldsymbol{\eta}; \mathcal{F}, \mathcal{T}\right) \approx \prod_{\mathcal{F}^q \in \boldsymbol{\mathcal{C}}} \left(\sum_{t_i \in \Omega(\mathcal{F}^q)} \Pr\left(t_i \mid \boldsymbol{\eta}\right) \cdot \Pr\left(f \mid \mathcal{F}^q, t_i\right) \right)^{N^q}$$

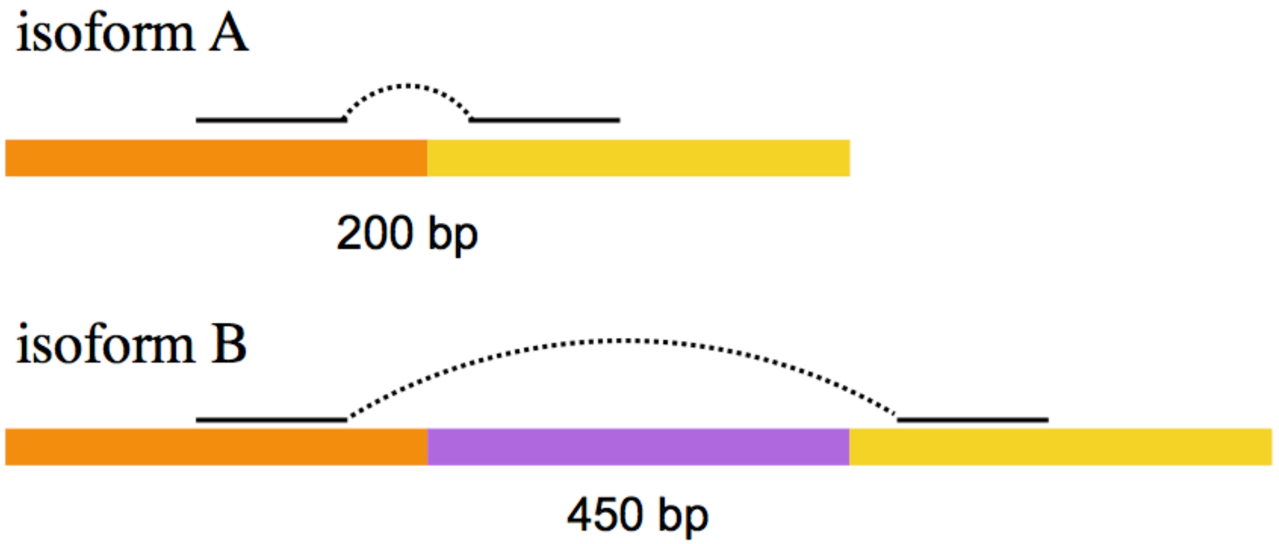
sum over all transcripts labeling this eq. class

product over all equivalence classes

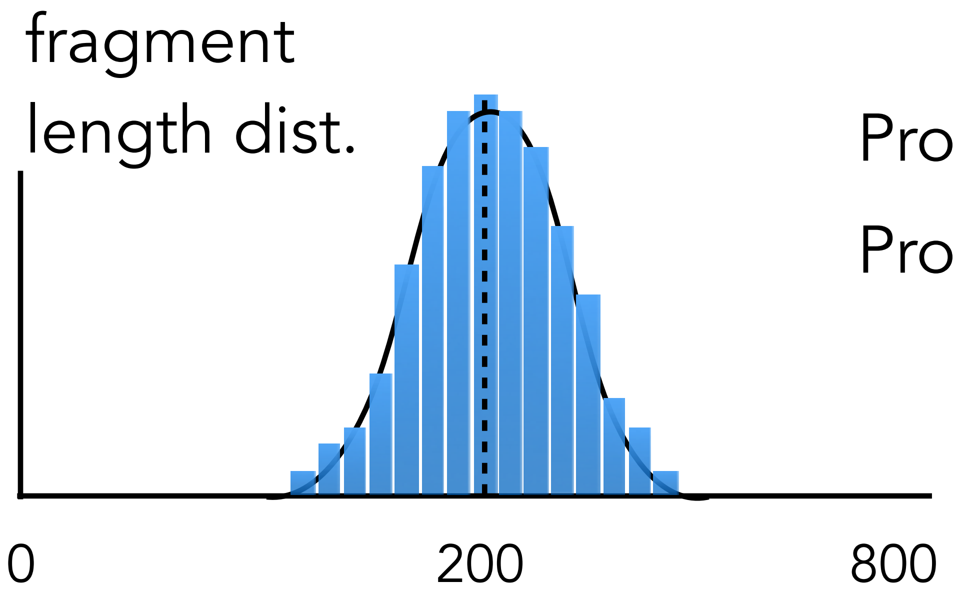
The approximation applies because **all** f in F^q have **the same** conditional probability given t$_i$ —- i.e. $\Pr\left(f \mid \mathcal{F}^q, t_i\right)$

Why might $\Pr(f_j \mid t_i)$ matter?

Consider the following scenario:



Conditional probabilities can provide valuable information about origin of a fragment! ***Potentially different for each transcript/fragment pair.***



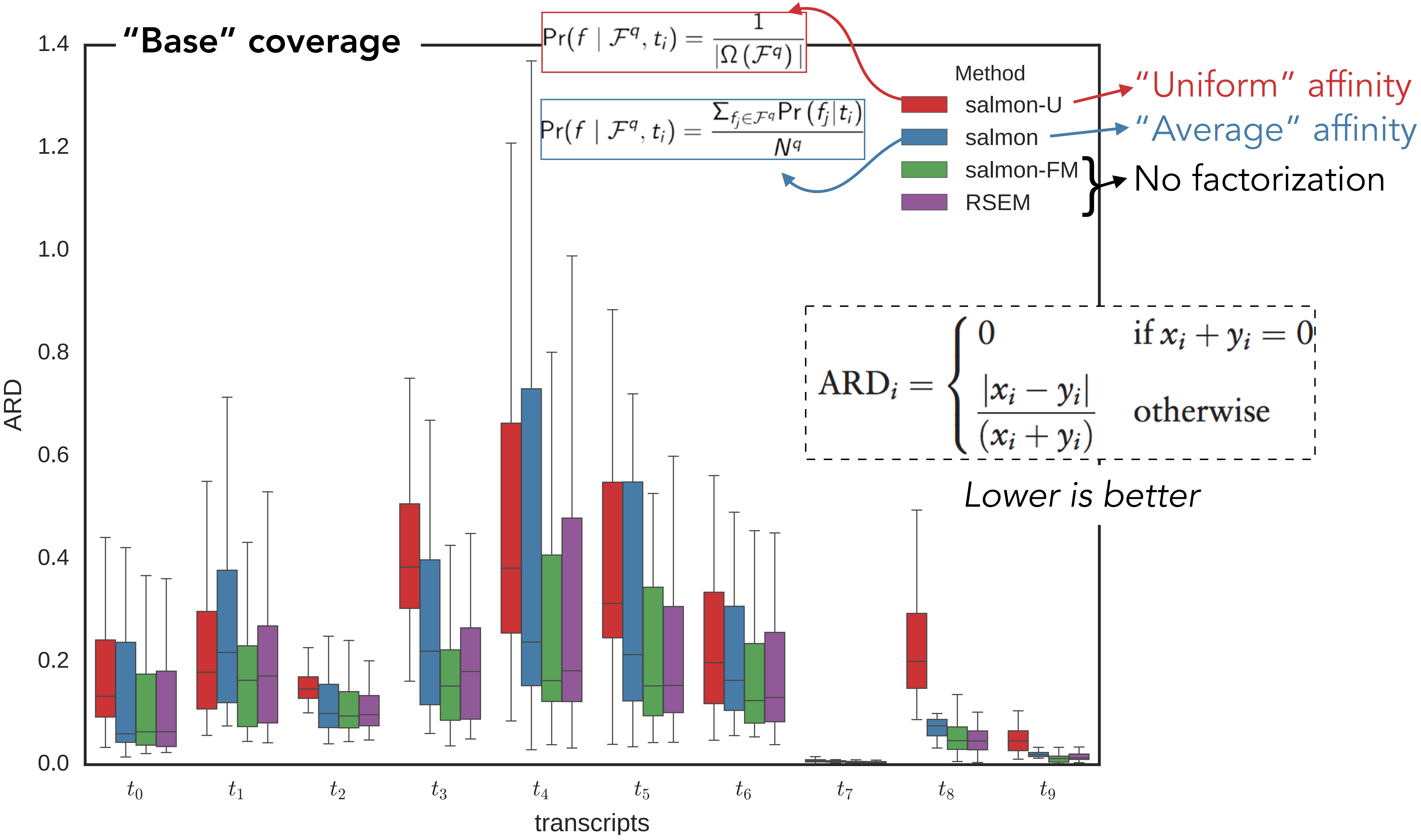
Prob of observing a fragment of size ~200 is **large**

Prob of observing a fragment of size ~450 is **small**

Many terms can be considered in a general "fragment-transcript agreement" model[1]. e.g. position, orientation, alignment path etc.

[1] "Salmon provides fast and bias-aware quantification of transcript expression", Nature Methods 2017

Does this term matter?



$$\Pr(f \mid \mathcal{F}^q, t_i) = \frac{1}{|\Omega(\mathcal{F}^q)|}$$

"Uniform" affinity

$$\Pr(f \mid \mathcal{F}^q, t_i) = \frac{\Sigma_{f_j \in \mathcal{F}^q} \Pr(f_j | t_i)}{N^q}$$

"Average" affinity

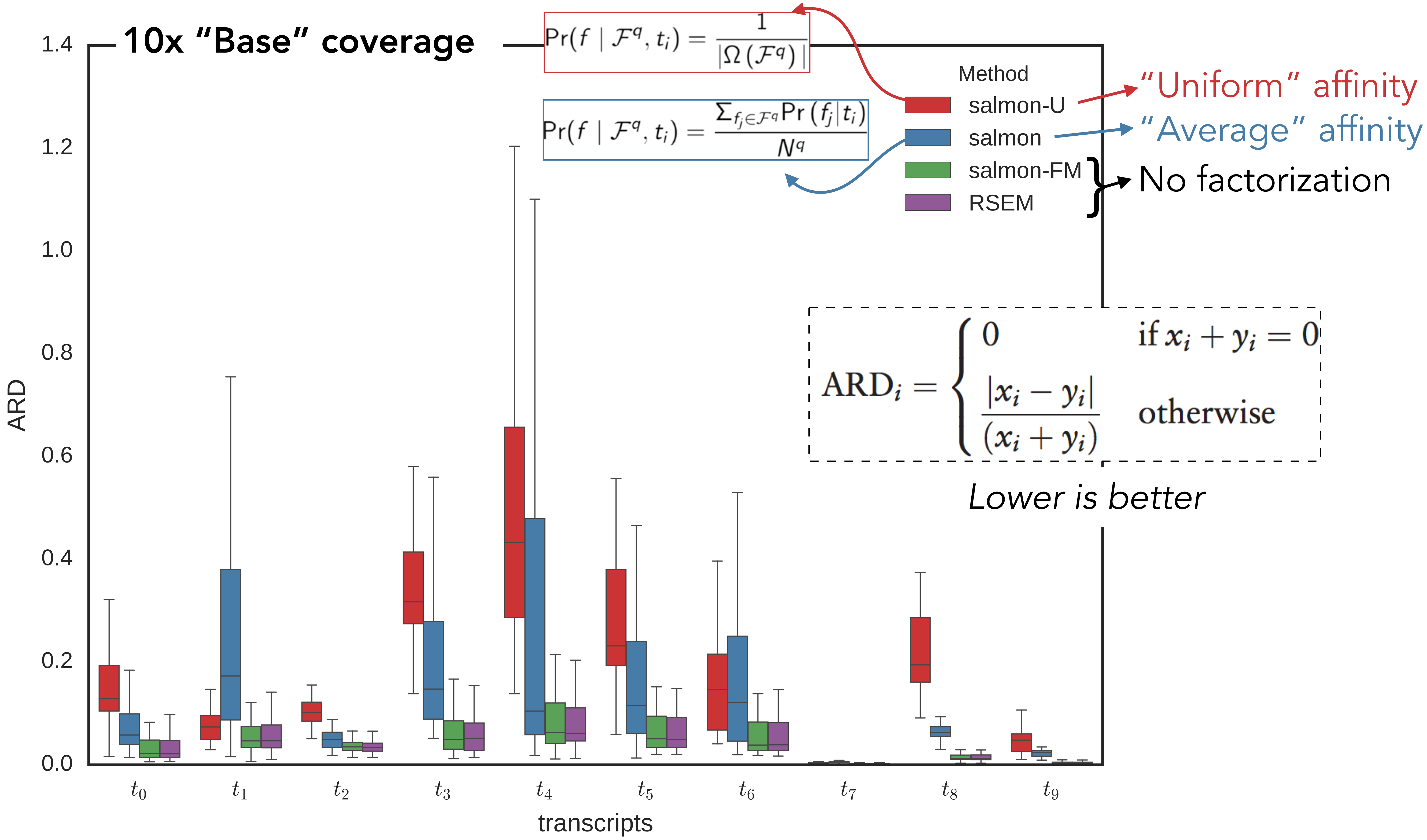
No factorization

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i + y_i = 0 \\ \dfrac{|x_i - y_i|}{(x_i + y_i)} & \text{otherwise} \end{cases}$$

Lower is better

- Transcripts of RAD51 gene — txp coverage drawn randomly in [1,200]
- Distribution over 30 random replicates of this distribution

Does this term matter?



$$\Pr(f \mid \mathcal{F}^q, t_i) = \frac{1}{|\Omega(\mathcal{F}^q)|}$$

$$\Pr(f \mid \mathcal{F}^q, t_i) = \frac{\sum_{f_j \in \mathcal{F}^q} \Pr(f_j | t_i)}{N^q}$$

"Uniform" affinity

"Average" affinity

No factorization

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i + y_i = 0 \\ \dfrac{|x_i - y_i|}{(x_i + y_i)} & \text{otherwise} \end{cases}$$

Lower is better

- Transcripts of RAD51 gene — txp coverage drawn randomly in [1,200]
- Distribution over 30 random replicates of this distribution

Range-factorized equivalence relation

Recall:

$$f \sim f' \iff \Omega(f) = \Omega(f')$$

$$\Omega(f) = \{t \mid f \text{ maps to } t\}$$

Now:

$$b_k(f, \langle t_{i_1}, \ldots, t_{i_j} \rangle)$$

Given a fragment and vector of transcripts, returns a vector of bin indices — each in [0,k) — that encode the conditional bin into which f falls with respect to each transcript.

# of conditional bins. Default = $4 + \lceil \sqrt{(}\,|\Omega(\mathcal{F}^q))|\rceil$

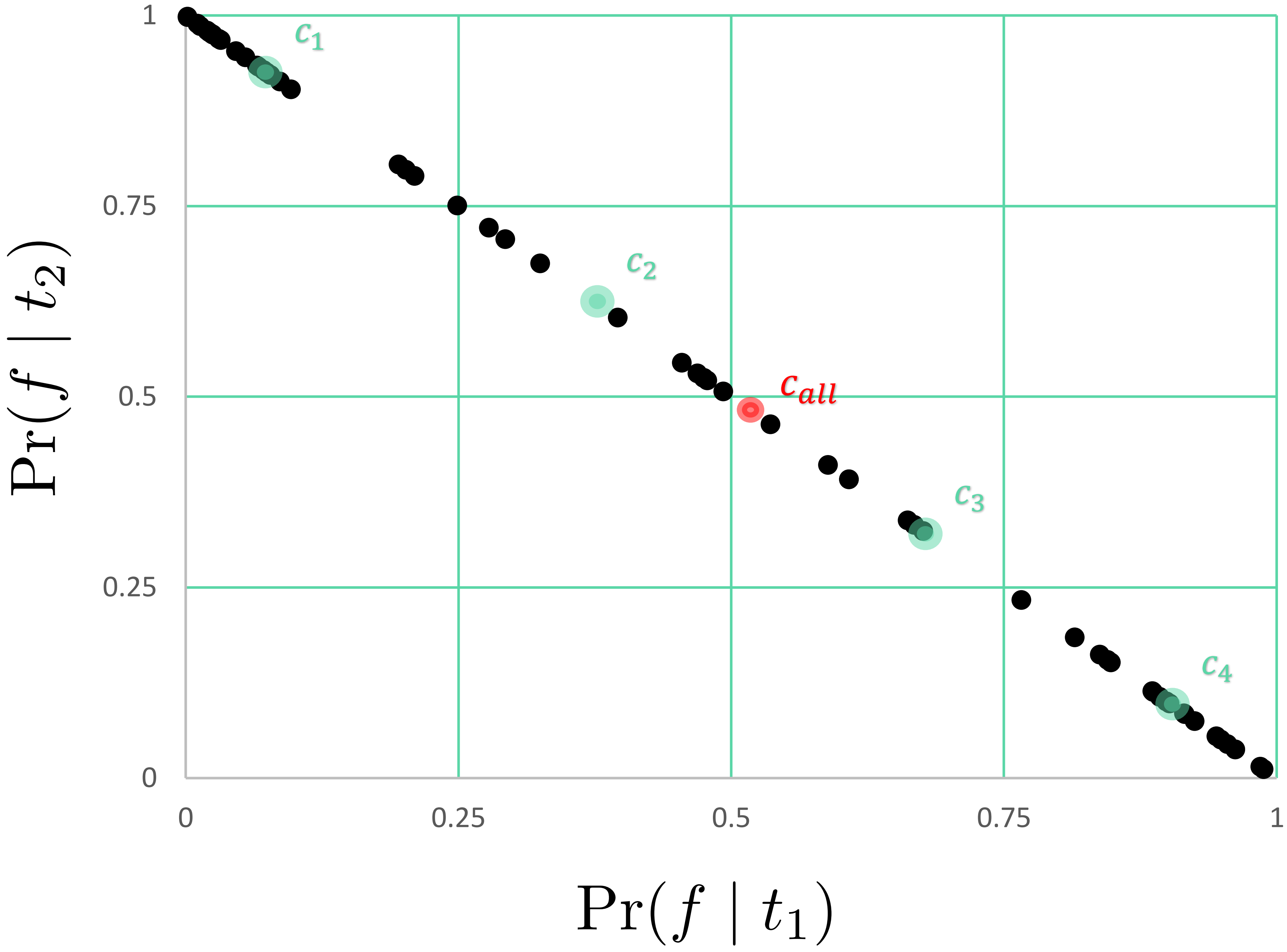
$$f \sim_r f' \iff \Omega(f) = \Omega(f') \wedge b_k(f, \Omega(f)) = b_k(f', \Omega(f'))$$

Maps to the same set of transcripts

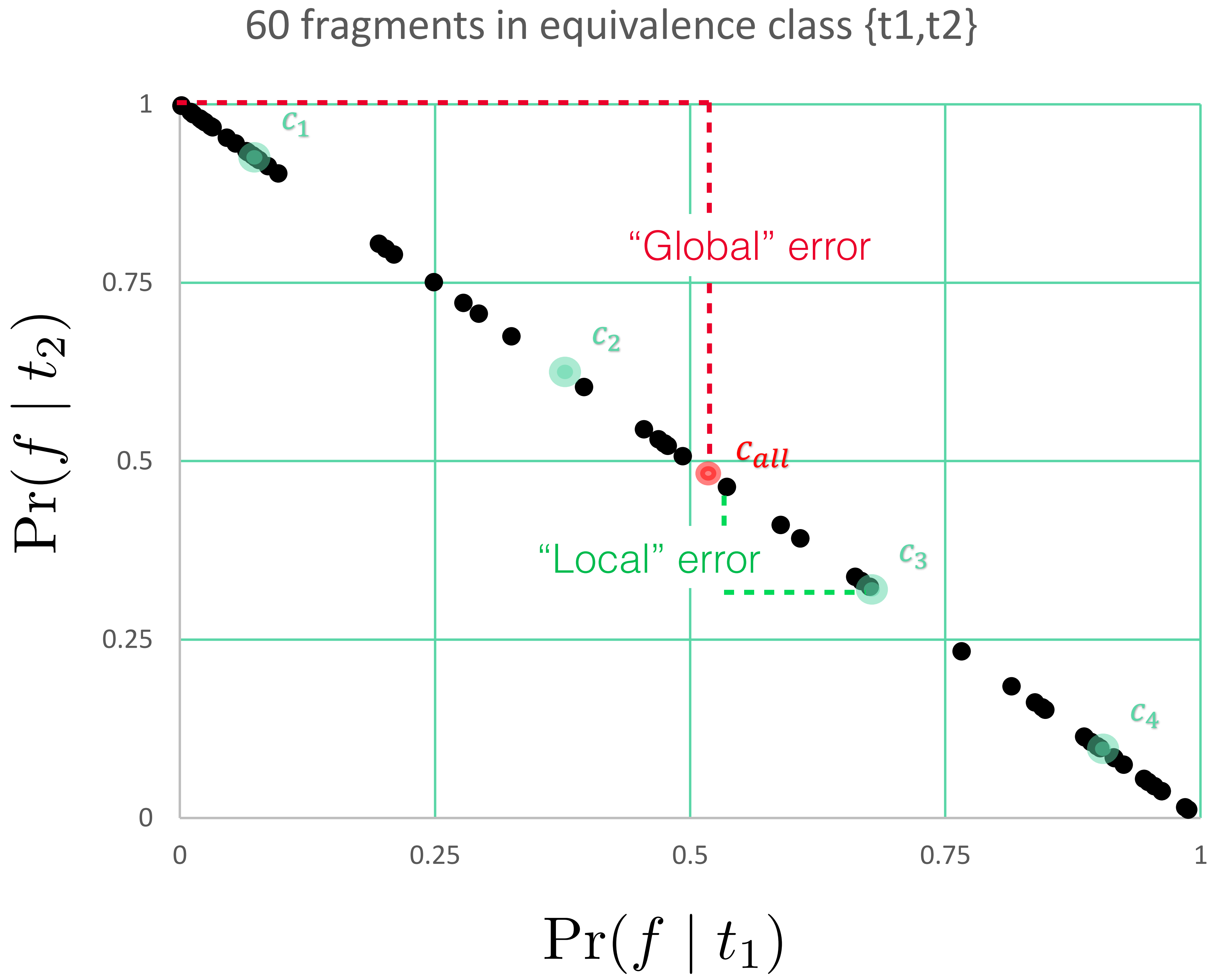
Has the same *binned* cond. prob vector

Range-based factorization

60 fragments in equivalence class {t1,t2}

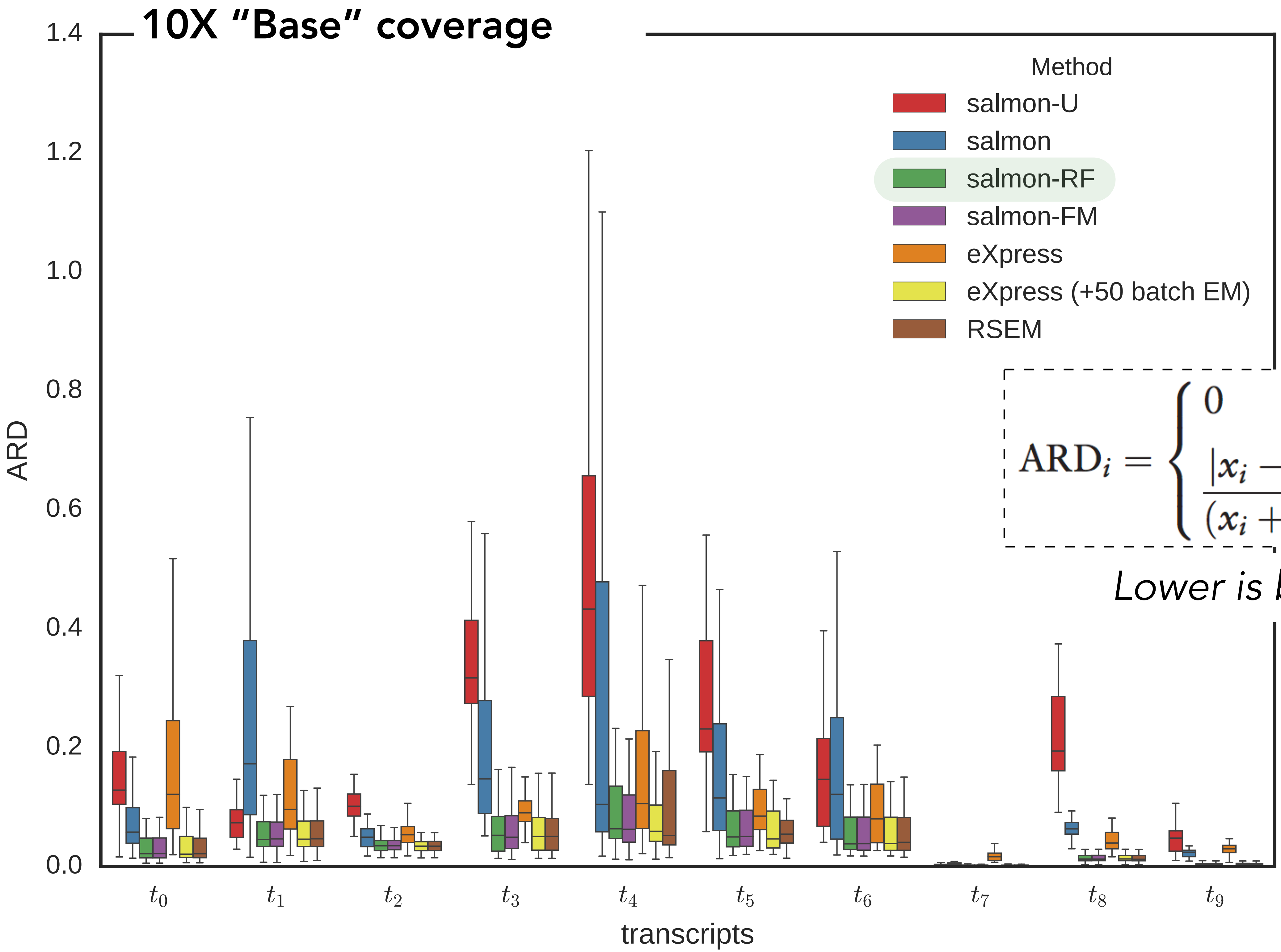


Range-based factorization improves approximation



- Provides a way to control the divergence between the full and factorized conditional likelihood distributions of an equivalence class

How well does this work?



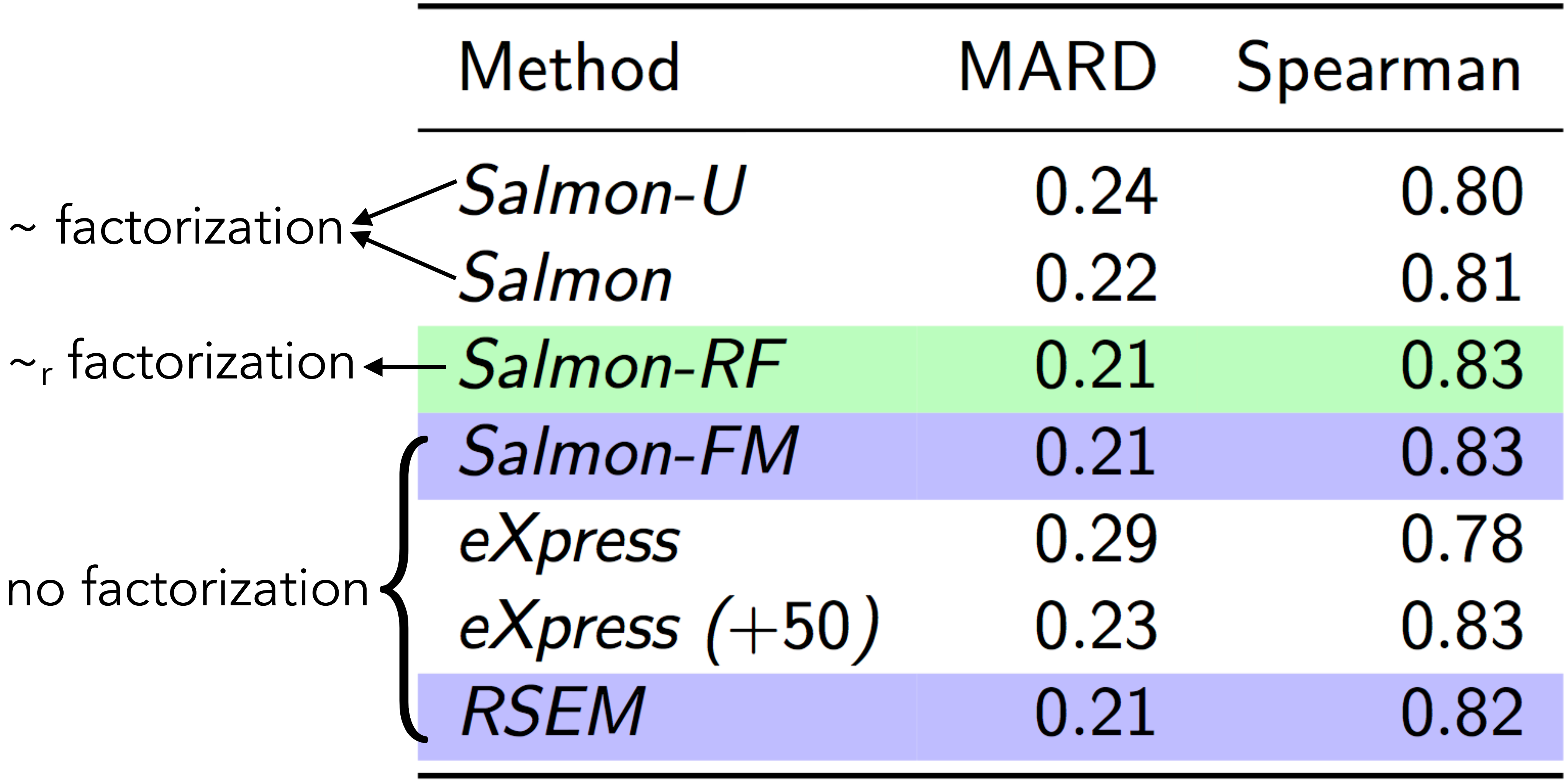
$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i + y_i = 0 \\ \dfrac{|x_i - y_i|}{(x_i + y_i)} & \text{otherwise} \end{cases}$$

Lower is better

- Transcripts of RAD51 gene — txp coverage drawn randomly in [1,200]
- Distribution over 30 random replicates of this distribution

Transcriptime-wide assessment can mask important differences

- Over tens of thousands of transcripts — overall differences are small
- But, we know this; factorized approaches are known to work well generally[1,2,3,4]



	Method	MARD	Spearman
~ factorization	*Salmon-U*	0.24	0.80
	Salmon	0.22	0.81
$\sim_r$ factorization	*Salmon-RF*	0.21	0.83
no factorization	*Salmon-FM*	0.21	0.83
	eXpress	0.29	0.78
	eXpress (+50)	0.23	0.83
	RSEM	0.21	0.82

- 30M paired-end reads, simulated with RSEM-Sim

1) Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome biology* 12.2 (2011): R13.

2) Srivastava, Avi, et al. "RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes." *Bioinformatics* 32.12 (2016): i192-i200.

3) Bray, N. L., et al. "Near-optimal probabilistic RNA-seq quantification." *Nature biotechnology* 34.5 (2016): 525.

4) Patro, Rob, et al. "Salmon provides fast and bias-aware quantification of transcript expression." *Nature Methods* 14.4 (2017): 417-419.

Transcriptime-wide assessment can mask important differences

- Focus on a subset of “critical” transcripts (not too easy, not intractable)
- Transcripts where RSEM yields an ARD in [0.25,0.75]

	Method	MARD	Spearman
$\sim$ factorization	*Salmon-U*	0.46	0.56
	Salmon	0.43	0.58
$\sim_r$ factorization	*Salmon-RF*	0.41	0.64
no factorization	*Salmon-FM*	0.41	0.65
	eXpress	0.53	0.54
	eXpress (+50)	0.48	0.59
	RSEM	0.41	0.65

- 30M paired-end reads, simulated with RSEM-Sim

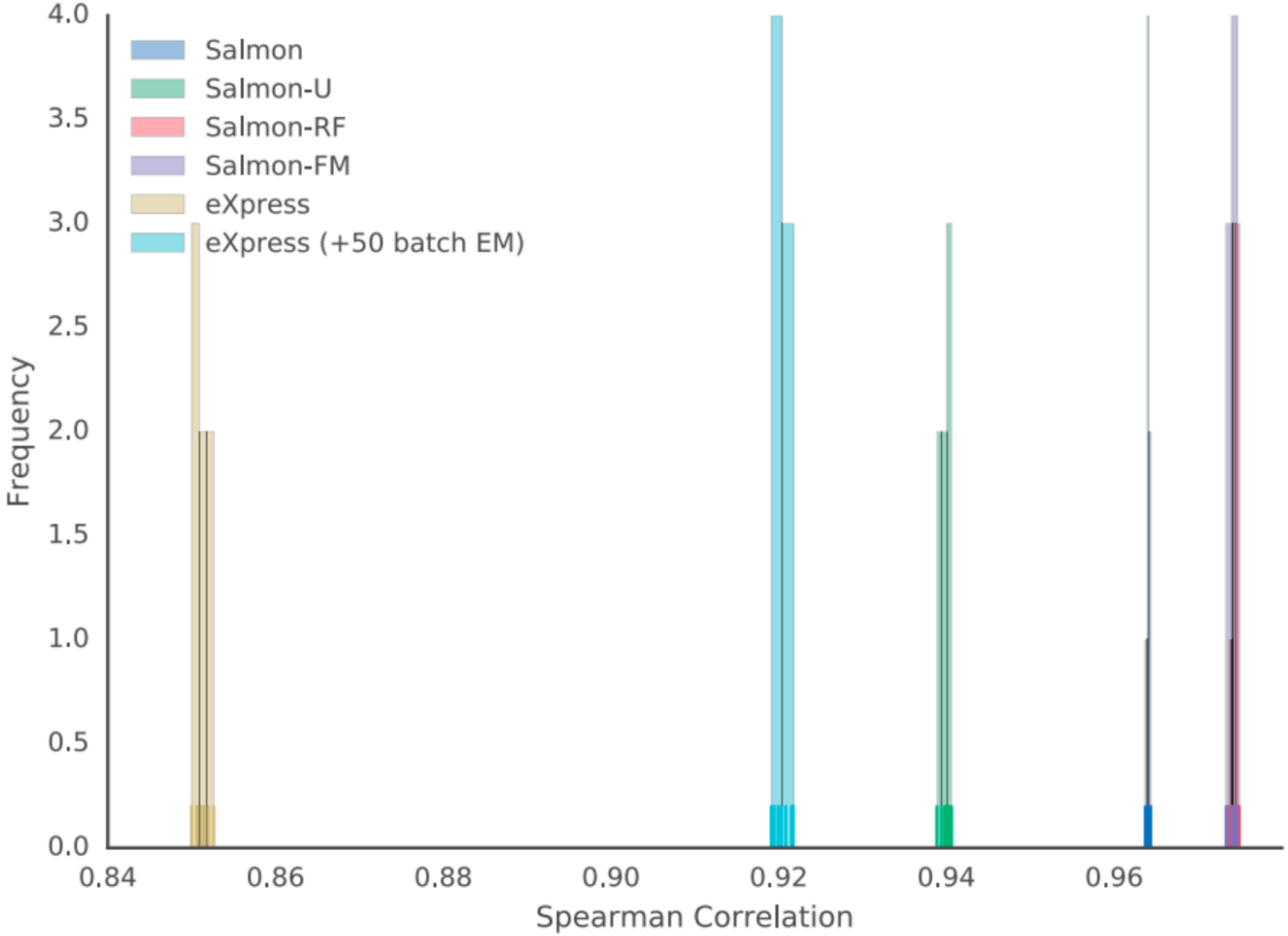
1) Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome biology* 12.2 (2011): R13.

2) Srivastava, Avi, et al. "RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes." *Bioinformatics* 32.12 (2016): i192-i200.

3) Bray, N. L., et al. "Near-optimal probabilistic RNA-seq quantification." *Nature biotechnology* 34.5 (2016): 525.

4) Patro, Rob, et al. "Salmon provides fast and bias-aware quantification of transcript expression." *Nature Methods* 14.4 (2017): 417-419.

Range-factorization improves correlation with full-model on experimental data



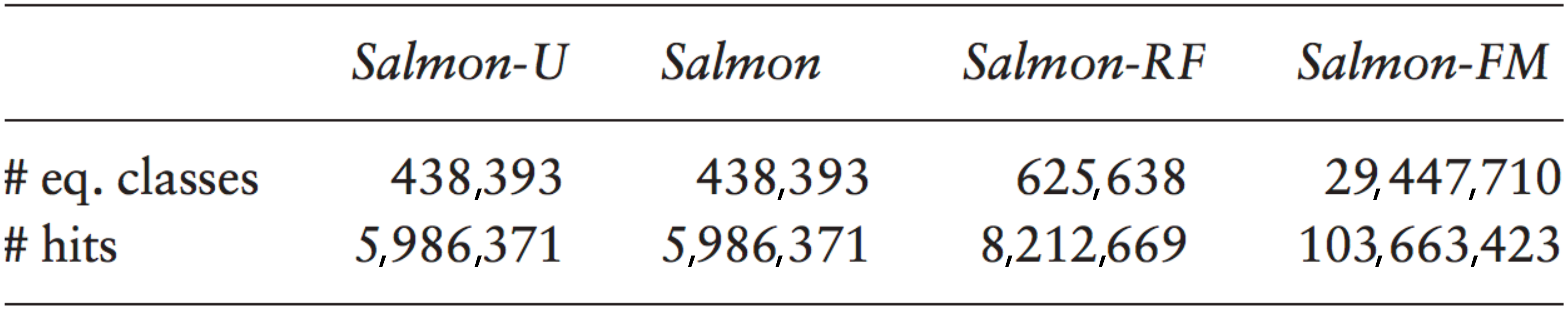
SEQC samples from UHRR (SRR1215996 - SRR1217002)

7 technical replicates to define distribution

Treat RSEM results as **ground truth** (though clearly, it's not perfect)

Range-factorization is still very (computationally) efficient

Factorization “size” on simulated data



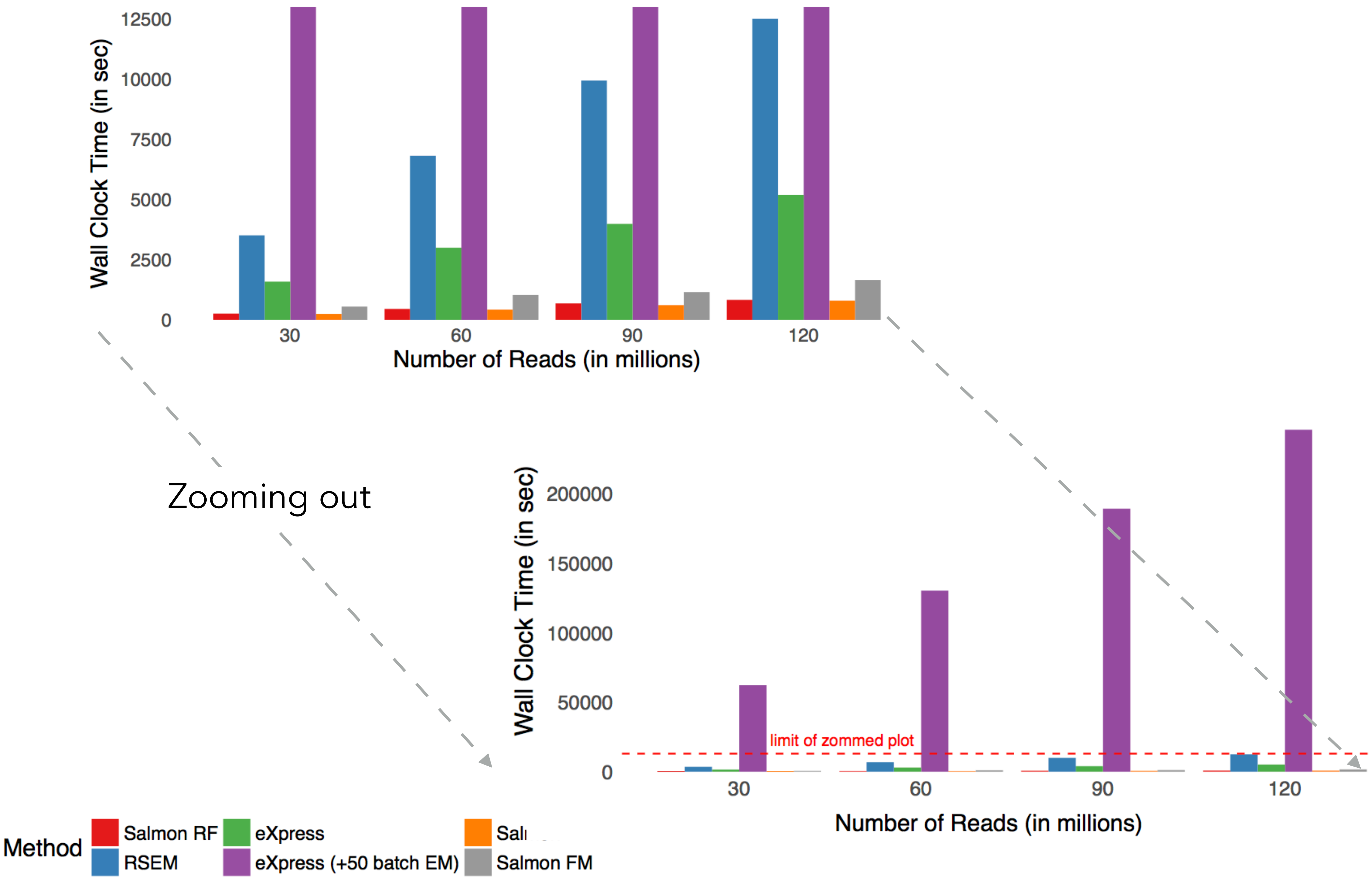
	Salmon-U	*Salmon*	*Salmon-RF*	*Salmon-FM*
# eq. classes	438,393	438,393	625,638	29,447,710
# hits	5,986,371	5,986,371	8,212,669	103,663,423

# eq. classes : The number of different “types” of read — i.e. $\sum_{\mathcal{F}^q \in \mathcal{C}} 1$

# hits : The number of hits is the sum, over each equivalence class, of the number of transcripts in this equivalence class — i.e. $\sum_{\mathcal{F}^q \in \mathcal{C}} |\Omega(\mathcal{F}^q)|$

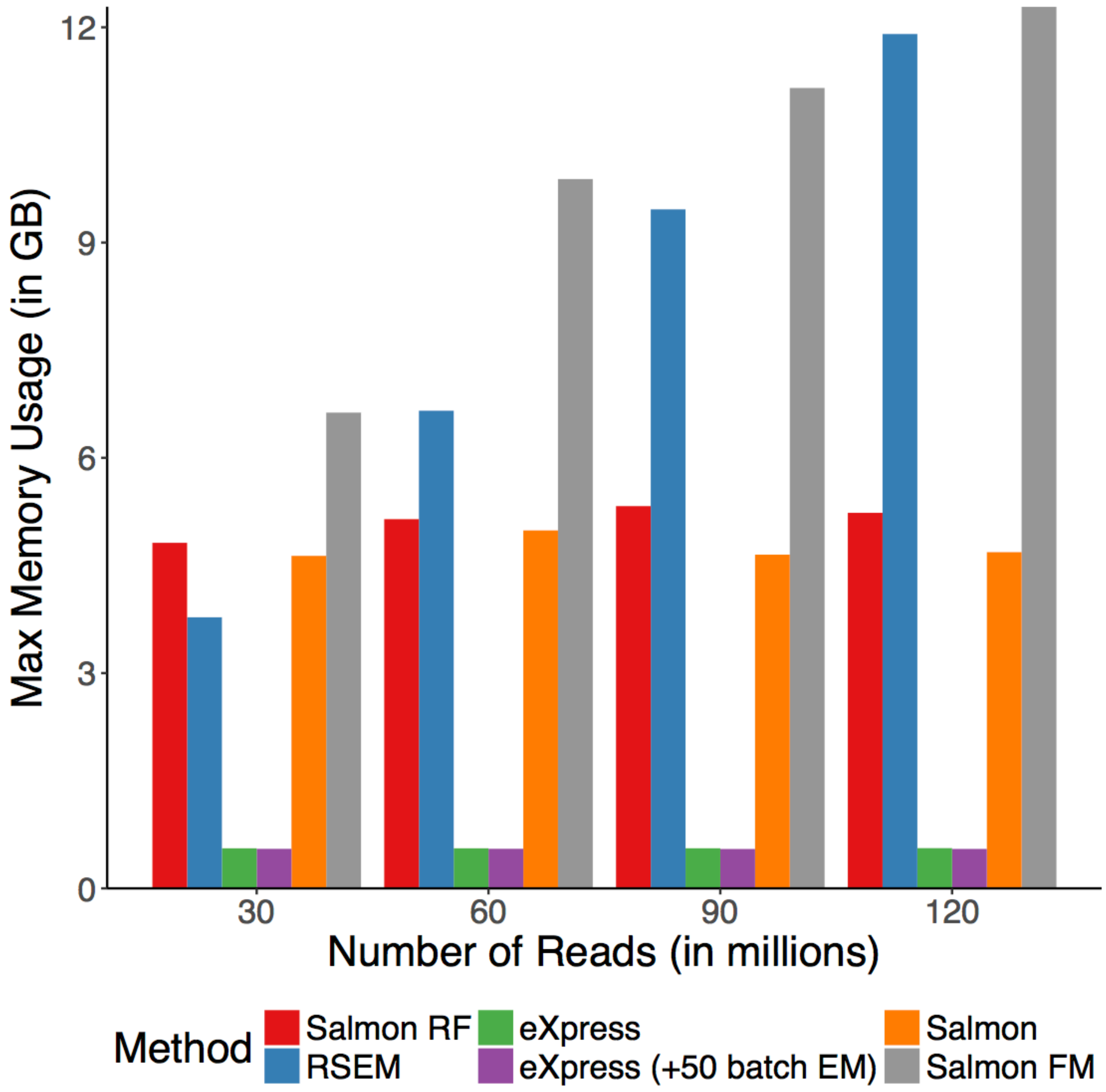
Difference is *marginal* with respect to # of reads / alignments

Range-factorization is still very (computationally) efficient



Zooming out

Range-factorization controls memory requirements



Estimating Posterior Uncertainty

One “issue” with maximum likelihood (ML)

The generative statistical model is a principled and elegant way to represent the RNA-seq process.

It can be optimized efficiently using e.g. the EM / VBEM algorithm.

but, these efficient optimization algorithms return “point estimates” of the abundances. That is, there is no notion of how *certain* we are in the computed abundance of transcript.

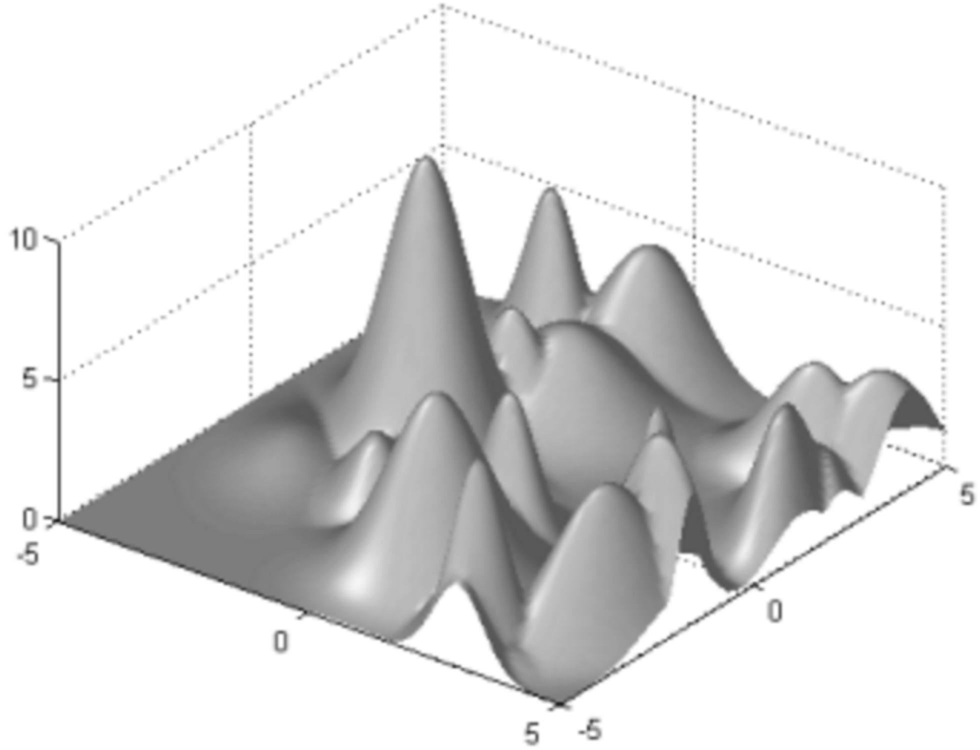
One “issue” with maximum likelihood (ML)

There are multiple sources of uncertainty e.g.

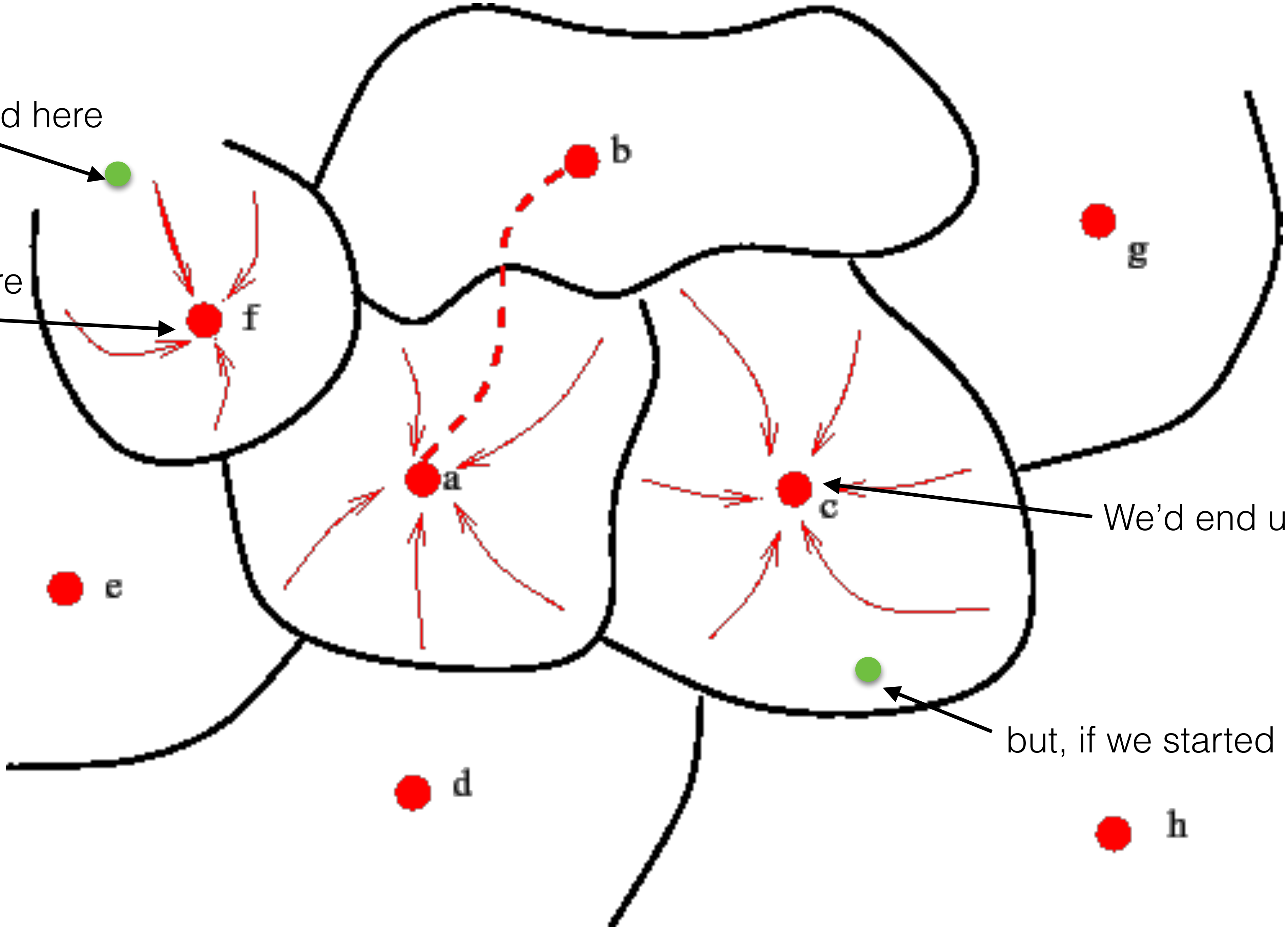
- Technical variance : If we sequenced the *exact* same sample again, we’d get a different set of fragments, and, potentially a different solution.

- Uncertainty in inference: We are almost never guaranteed to find a unique, globally optimal result. If we started our algorithm with different initialization parameters, we might get a different result.

We’re trying to find the *best* parameters in a space with 10s to 100s of thousands of dimensions!



One “issue” with maximum likelihood (ML)



If we started here

We’d end up here

We’d end up here

but, if we started here

https://commons.wikimedia.org/wiki/File:Local_search_attraction_basins.png (CC BY-SA 3.0)

Assessing Uncertainty

There are a few ways to address this “issue”

Do a fully Bayesian inference[1]:

Infer the entire posterior distribution of parameters, not just a ML estimate (e.g. using MCMC) — too slow!

✓ Posterior Gibbs Sampling[2,3]:

Starting from our ML estimate, do MCMC sampling to explore how parameters vary — if our ML estimate is good, this can be made *quite fast.*

✓ Bootstrap Sampling[4]:

Resample (from range-factorized equivalence class counts) with replacement, and re-run the ML estimate for each sample. This can be made reasonably fast.

Happy to discuss details / implications of this further.

[1] BitSeq (with MCMC) actually does this. It’s very accurate, but very slow. [Glaus, Peter, Antti Honkela, and Magnus Rattray. "Identifying differentially expressed transcripts from RNA-seq data with biological variation." Bioinformatics 28.13 (2012): 1721-1728.]

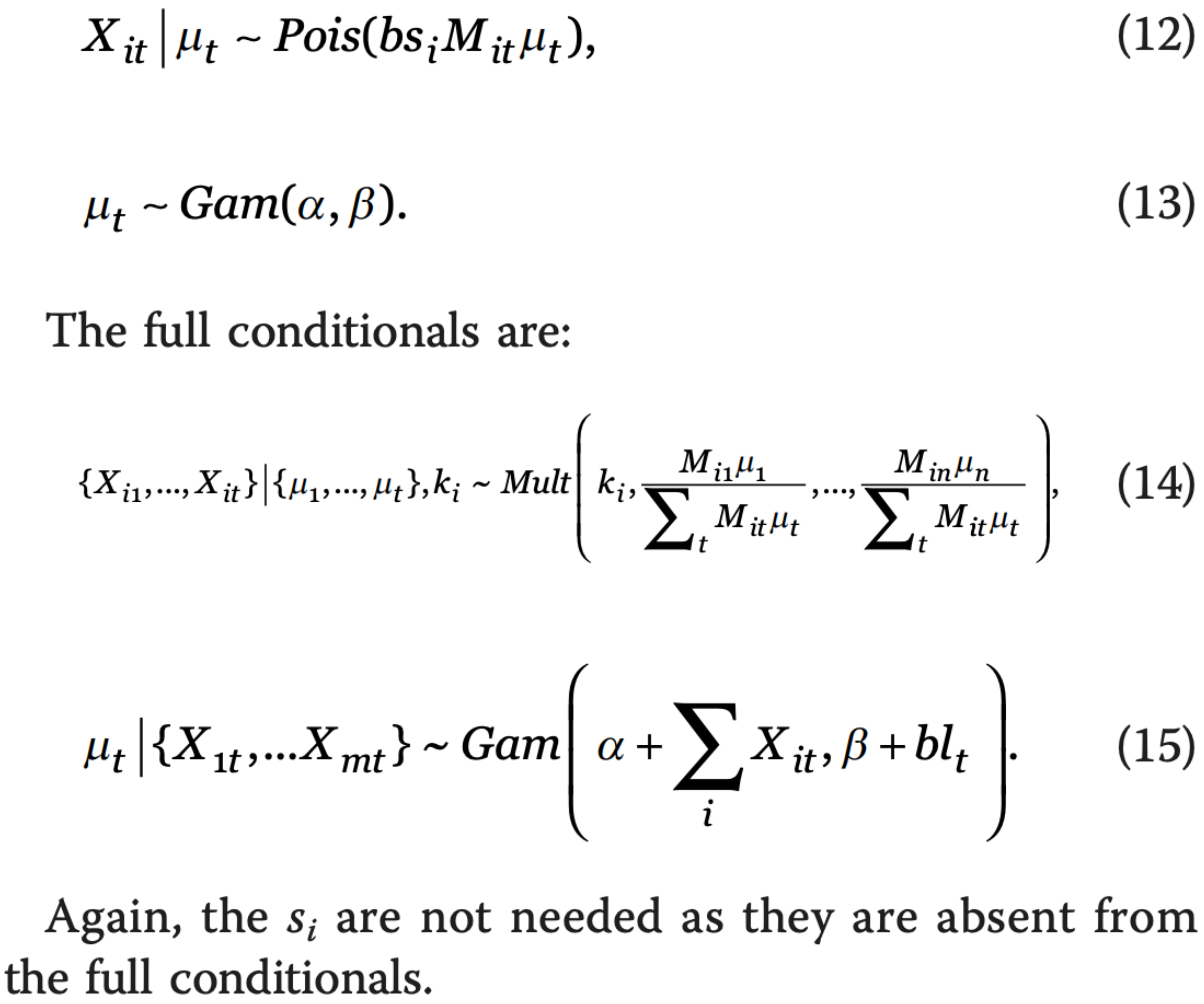
[2] RSEM has the ability to do this, and it seems to work well, but each sample scales in the # of reads. [Li, Bo, and Colin N. Dewey. "RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome." BMC bioinformatics 12.1 (2011): 1.]

[3] MMSEQ can perform Gibbs sampling over shared variables (i.e. equiv classes), producing estimates from the mean of the posterior dist.Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." Genome biology 12.2 (2011): 1.

[4] IsoDE introduced the idea of bootstrapping counts to assess quantification uncertainty. [Al Seesi, Sahar, et al. "Bootstrap-based differential gene expression analysis for RNA-Seq data with and without replicates." BMC genomics 15.8 (2014): 1.], but it was first made practical / fast in kallisto by doing the bootstrapping over equivalence classes.

A few ways to implement Gibbs Sampling for this problem

The model of MMSeq



$$X_{it} \,|\, \mu_t \sim Pois(bs_i M_{it} \mu_t), \tag{12}$$

$$\mu_t \sim Gam(\alpha, \beta). \tag{13}$$

The full conditionals are:

$$\{X_{i1},...,X_{it}\} \,|\, \{\mu_1,...,\mu_t\}, k_i \sim Mult\left(k_i, \frac{M_{i1}\mu_1}{\sum_t M_{it}\mu_t}, ..., \frac{M_{in}\mu_n}{\sum_t M_{it}\mu_t} \right), \tag{14}$$

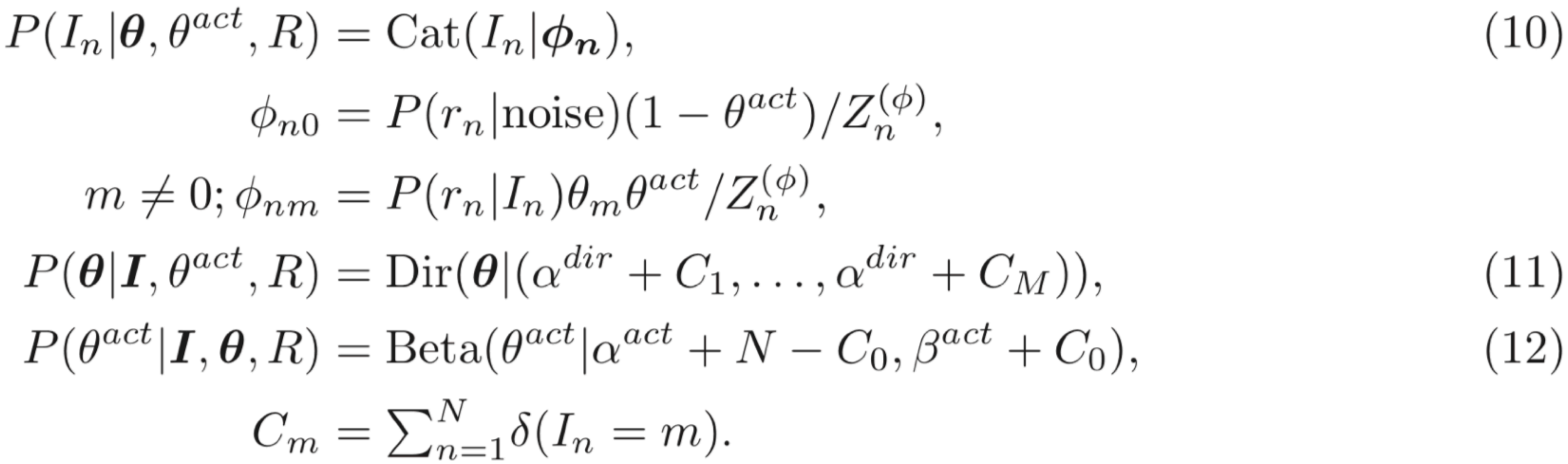
$$\mu_t \,|\, \{X_{1t},...X_{mt}\} \sim Gam\left(\alpha + \sum_i X_{it}, \beta + bl_t \right). \tag{15}$$

Again, the s_i are not needed as they are absent from the full conditionals.

Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." Genome biology 12.2 (2011): 1.

A few ways to implement Gibbs Sampling for this problem

The model of BitSeq



$$
\begin{align}
P(I_n|\boldsymbol{\theta}, \theta^{act}, R) &= \mathrm{Cat}(I_n|\boldsymbol{\phi_n}), \tag{10}\\
\phi_{n0} &= P(r_n|\mathrm{noise})(1-\theta^{act})/Z_n^{(\phi)}, \notag\\
m \neq 0; \phi_{nm} &= P(r_n|I_n)\theta_m\theta^{act}/Z_{\boldsymbol{n}}^{(\phi)}, \notag\\
P(\boldsymbol{\theta}|\boldsymbol{I}, \theta^{act}, R) &= \mathrm{Dir}(\boldsymbol{\theta}|(\alpha^{dir}+C_1, \ldots, \alpha^{dir}+C_M)), \tag{11}\\
P(\theta^{act}|\boldsymbol{I}, \boldsymbol{\theta}, R) &= \mathrm{Beta}(\theta^{act}|\alpha^{act}+N-C_0, \beta^{act}+C_0), \tag{12}\\
C_m &= \textstyle\sum_{n=1}^{N}\delta(I_n = m). \notag
\end{align}
$$

[Glaus, Peter, Antti Honkela, and Magnus Rattray. "Identifying differentially expressed transcripts from RNA-seq data with biological variation." Bioinformatics 28.13 (2012): 1721-1728.]

A few ways to implement Gibbs Sampling for this problem

The model of BitSeq (collapsed sampler)

$$
\begin{aligned}
P(I_n|I^{(-n)}, R) &= \mathrm{Cat}(I_n|\boldsymbol{\phi_n^*}), \qquad (9)\\
\phi^*_{n0} &= P(r_n|\mathrm{noise})(\beta^{act} + C_0^{(-n)})/Z_n^{(\phi^*)},\\
m \neq 0; \phi^*_{nm} &= P(r_n|I_n)(\alpha^{act} + C_+^{(-n)})\tfrac{(\alpha^{dir}+C_m^{(-n)})}{(M\alpha^{dir}+C_+^{(-n)})}/Z_n^{(\phi^*)},\\
C_m^{(-n)} &= \textstyle\sum_{i\neq n}\delta(I_i = m),\\
C_+^{(-n)} &= \textstyle\sum_{i\neq n}\delta(I_i > 0)\ ,
\end{aligned}
$$

with $Z_n^{(\phi^*)}$ being a constant normalising $\boldsymbol{\phi_n}^*$ to sum up to 1, and $\alpha^{dir} = 1, \alpha^{act} = 2, \beta^{act} = 2.$

[Glaus, Peter, Antti Honkela, and Magnus Rattray. "Identifying differentially expressed transcripts from RNA-seq data with biological variation." Bioinformatics 28.13 (2012): 1721-1728.]

This uncertainty matters

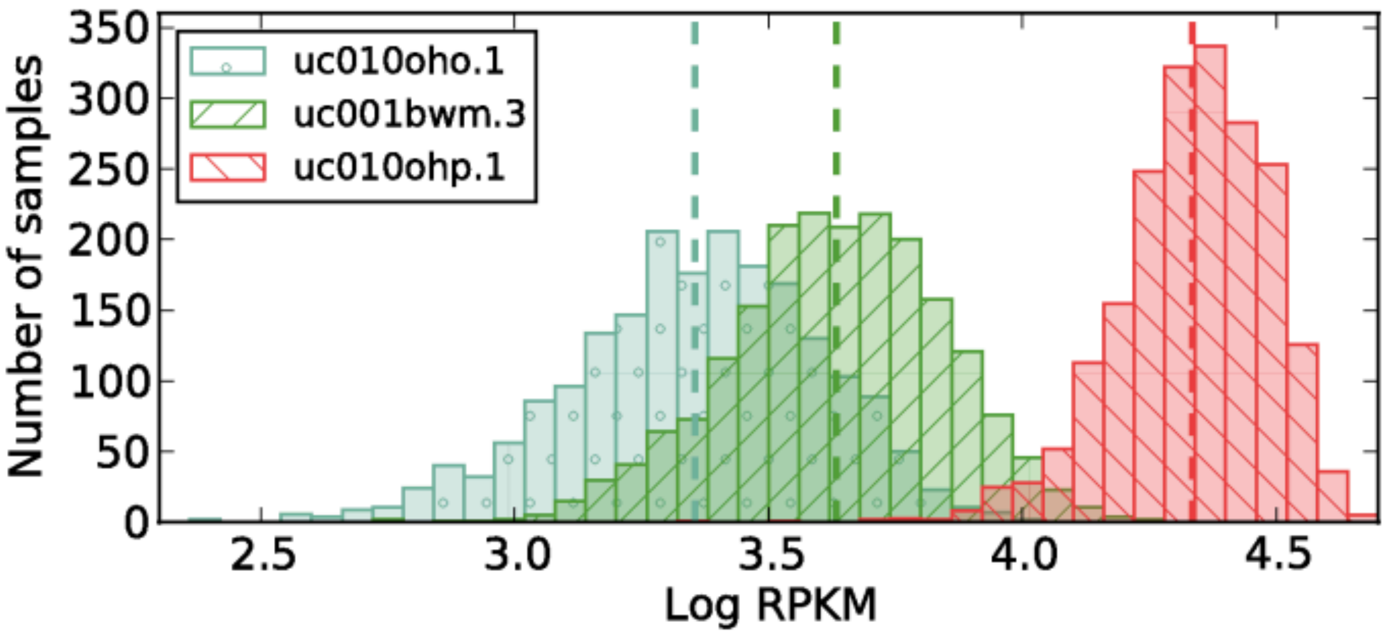
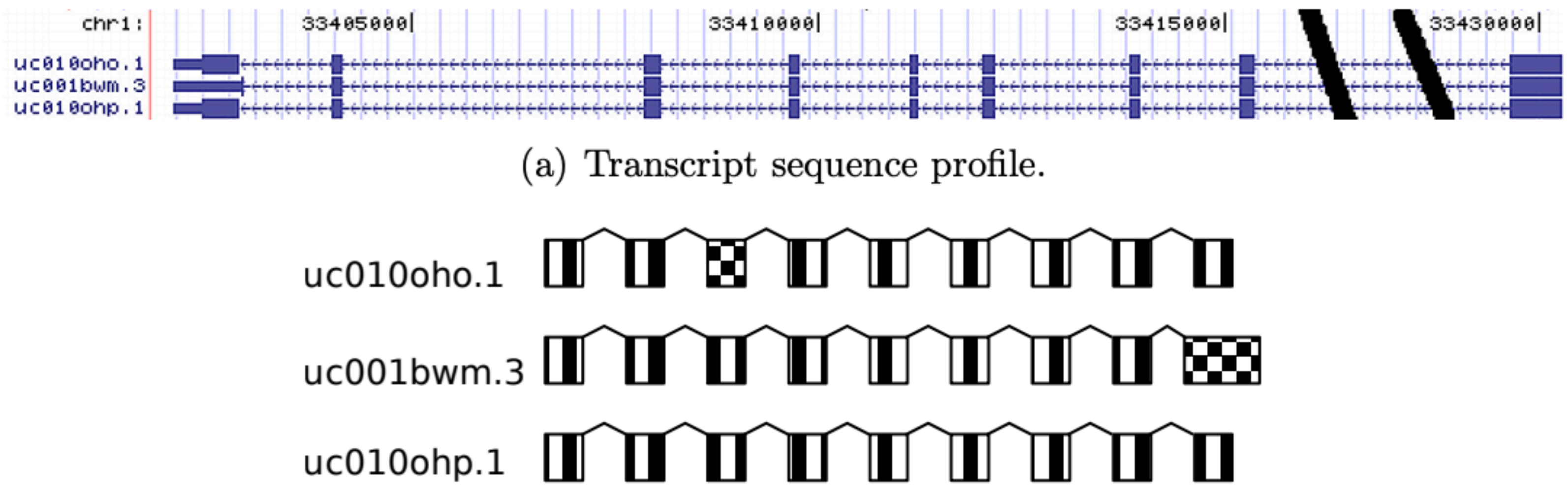


Figure 2.10: **Posterior distribution of expression levels of three transcripts of gene Q6ZMZ0.** The posterior distribution is represented in form of a histogram of expression samples converted into Log RPKM expression measure. The dashed lines mark the mean expression for each transcript.

*Glaus, Peter. *Bayesian Methods for Gene Expression Analysis from High-throughput Sequencing Data*. Diss. University of Manchester, 2014.

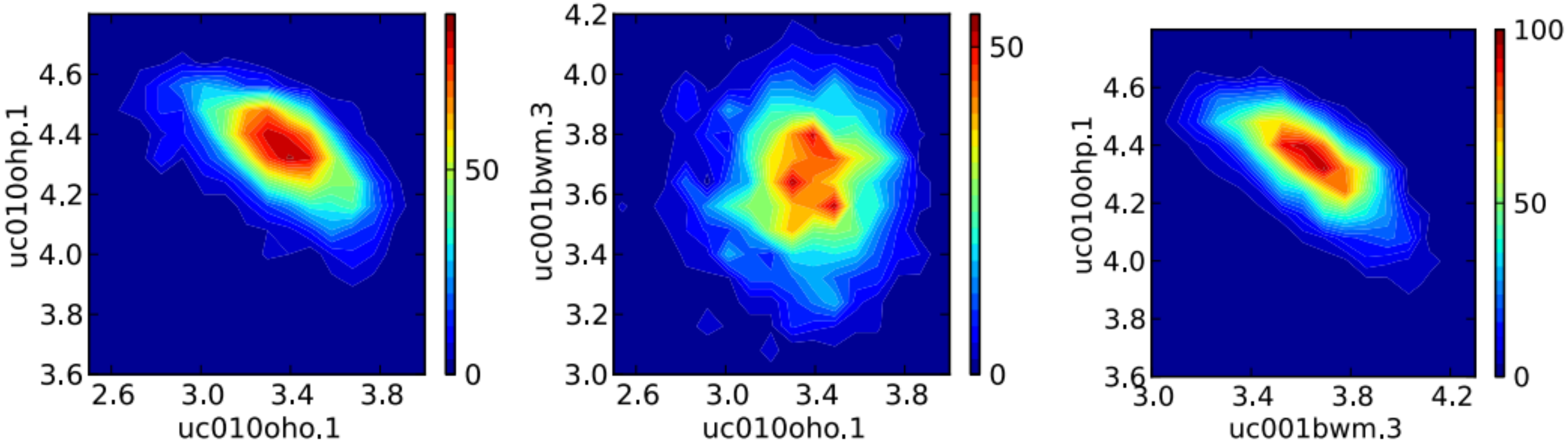
This uncertainty matters



(a) Transcript sequence profile.

(b) Splice variant model.

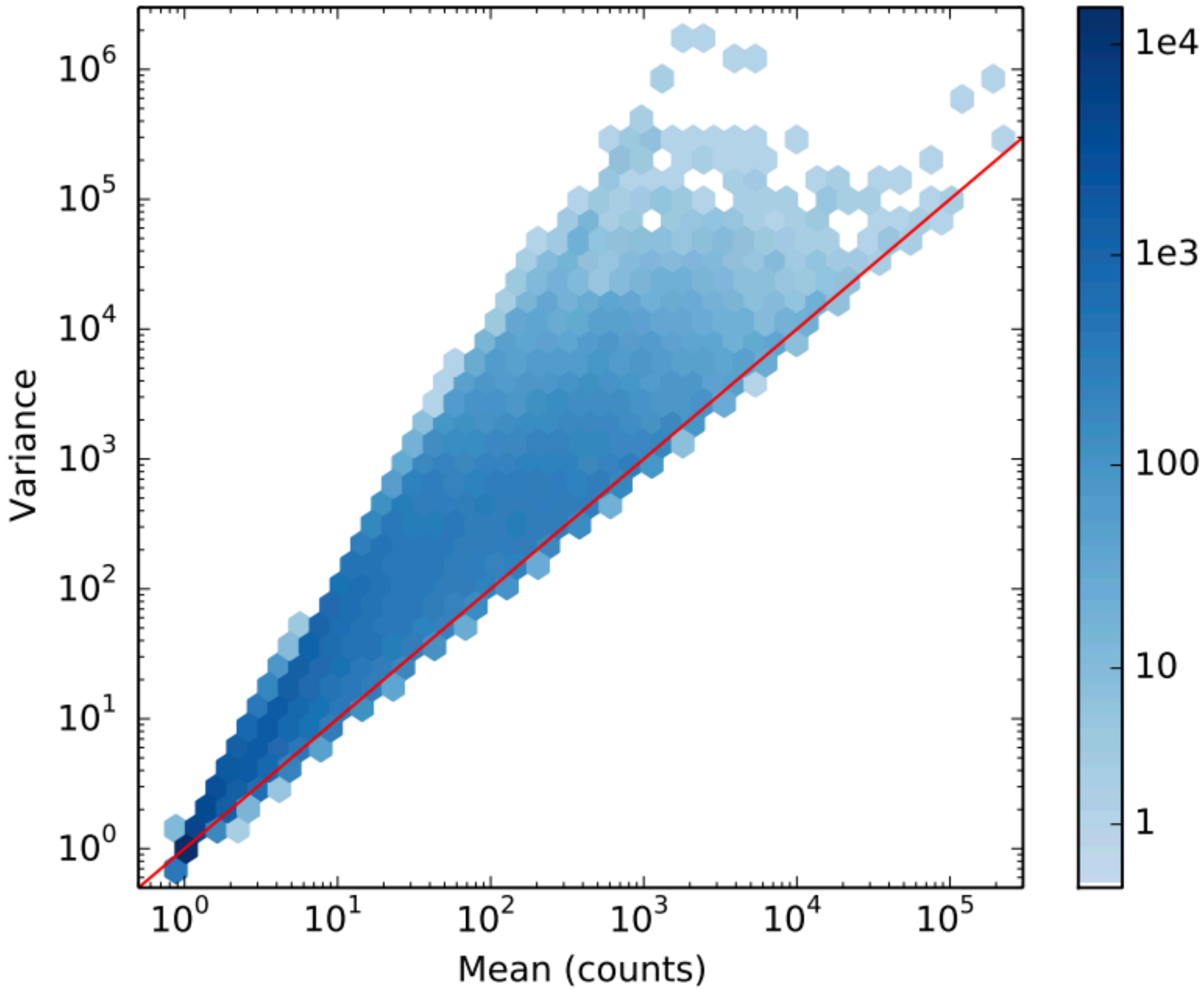
Figure 2.12: **Exon model of transcripts of gene Q6ZMZ0.** (a) transcript sequence profile obtained from the UCSC genome browser (Kuhn et al., 2013). In this annotation, transcript uc001bwm.3 has different 3' untranslated region and transcript uc010oho.1 has extra nucleotides at the end of second exon. As the second change cannot be distinguished in the UCSC genome browser diagram, we provide schematic splice variant model highlighting the differences (b).



*Glaus, Peter. *Bayesian Methods for Gene Expression Analysis from High-throughput Sequencing Data*. Diss. University of Manchester, 2014.

This uncertainty matters

We observe considerably increased variance due to read mapping ambiguity



If we know this increased uncertainty, we can propagate it & use it in downstream analysis (differential expression)!

*Glaus, Peter. *Bayesian Methods for Gene Expression Analysis from High-throughput Sequencing Data*. Diss. University of Manchester, 2014.