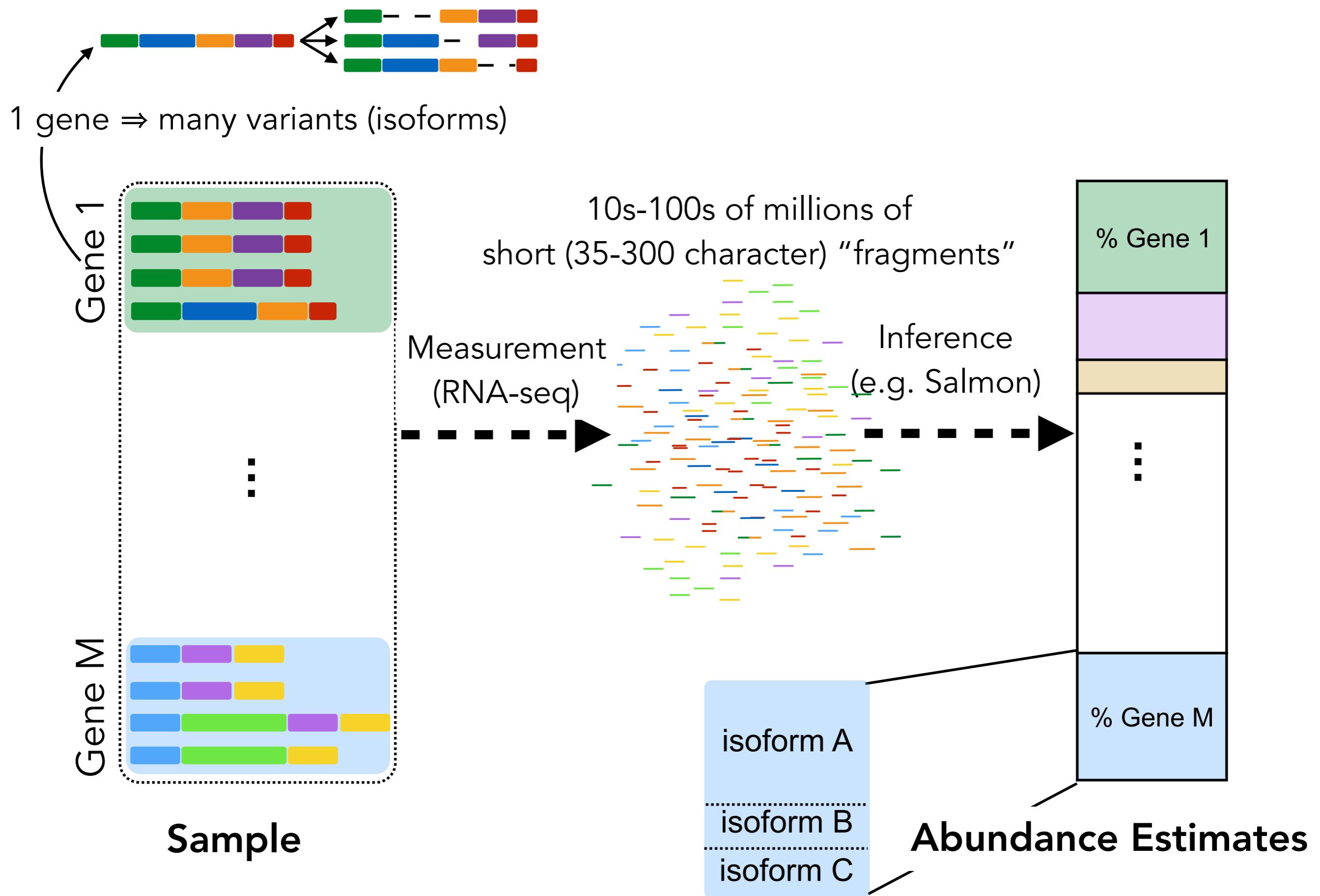
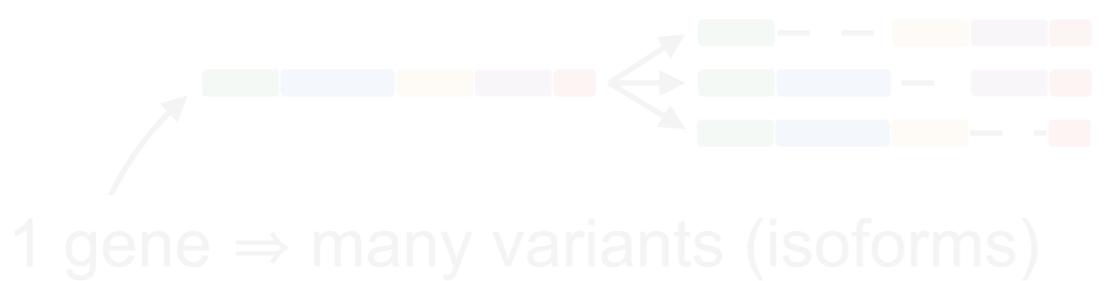


Analyzing gene and transcript expression using RNA-seq (II)

Transcript Quantification: An Overview

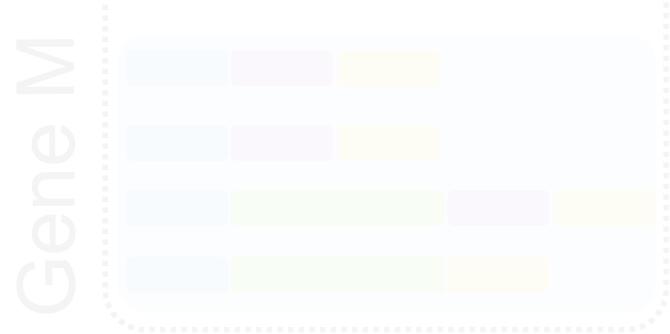




10s-100s of millions of
short (35-300 character) “reads”

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

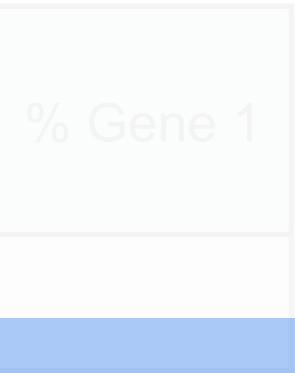
Estimate: The relative abundance of each transcript

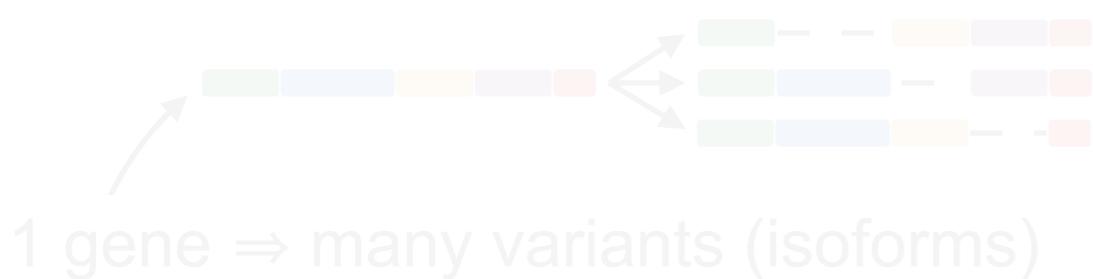


Sample

isoform A
isoform B
isoform C

Abundance Estimates

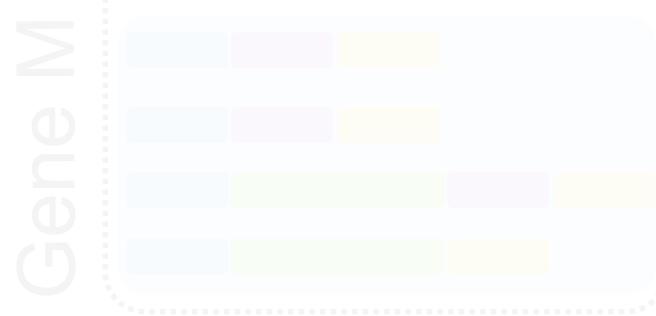




10s-100s of millions of
short (35-300 character) “reads”

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

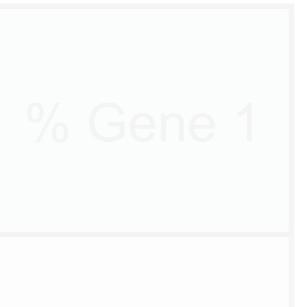
Estimate: The relative abundance of each transcript



Sample

isoform A
isoform B
isoform C

Abundance Estimates



Why not simply “count” reads

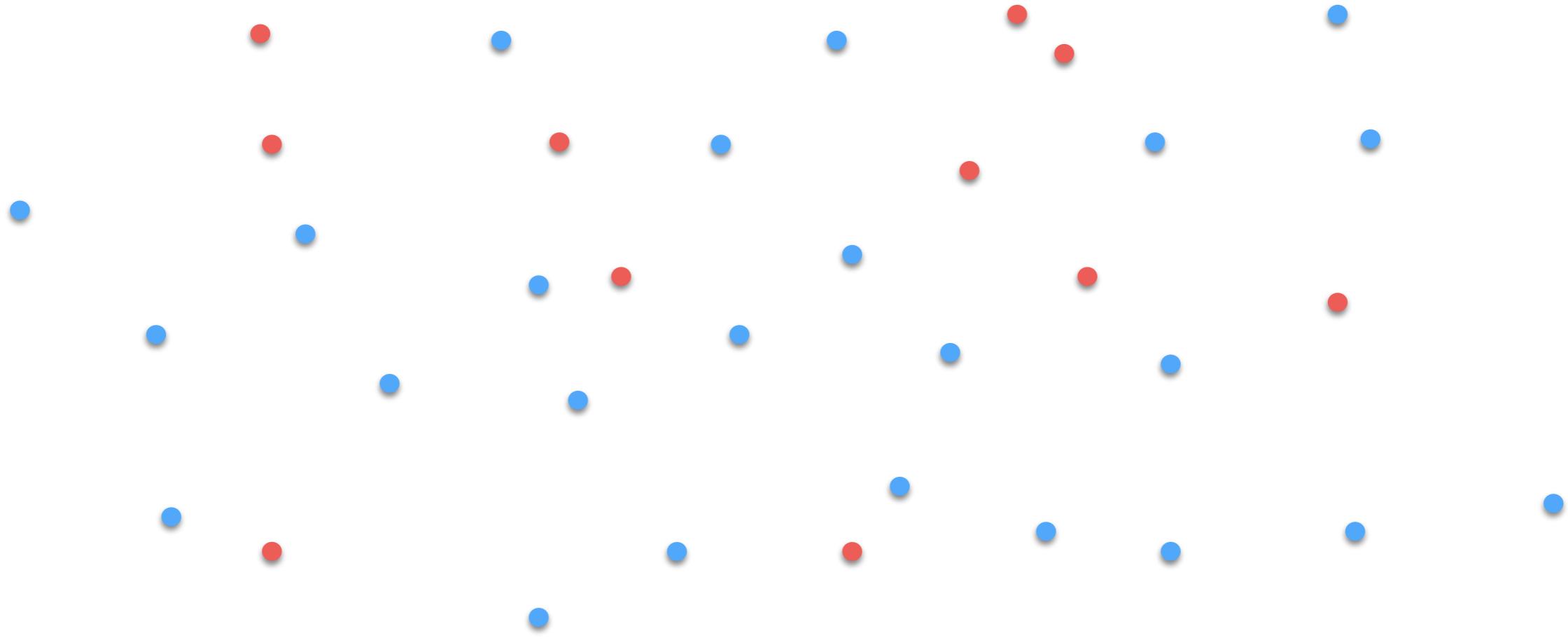
The RNA-seq reads are drawn from transcripts, and our spliced-aligners let us map them back to the transcripts on the genome from which they originate.

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

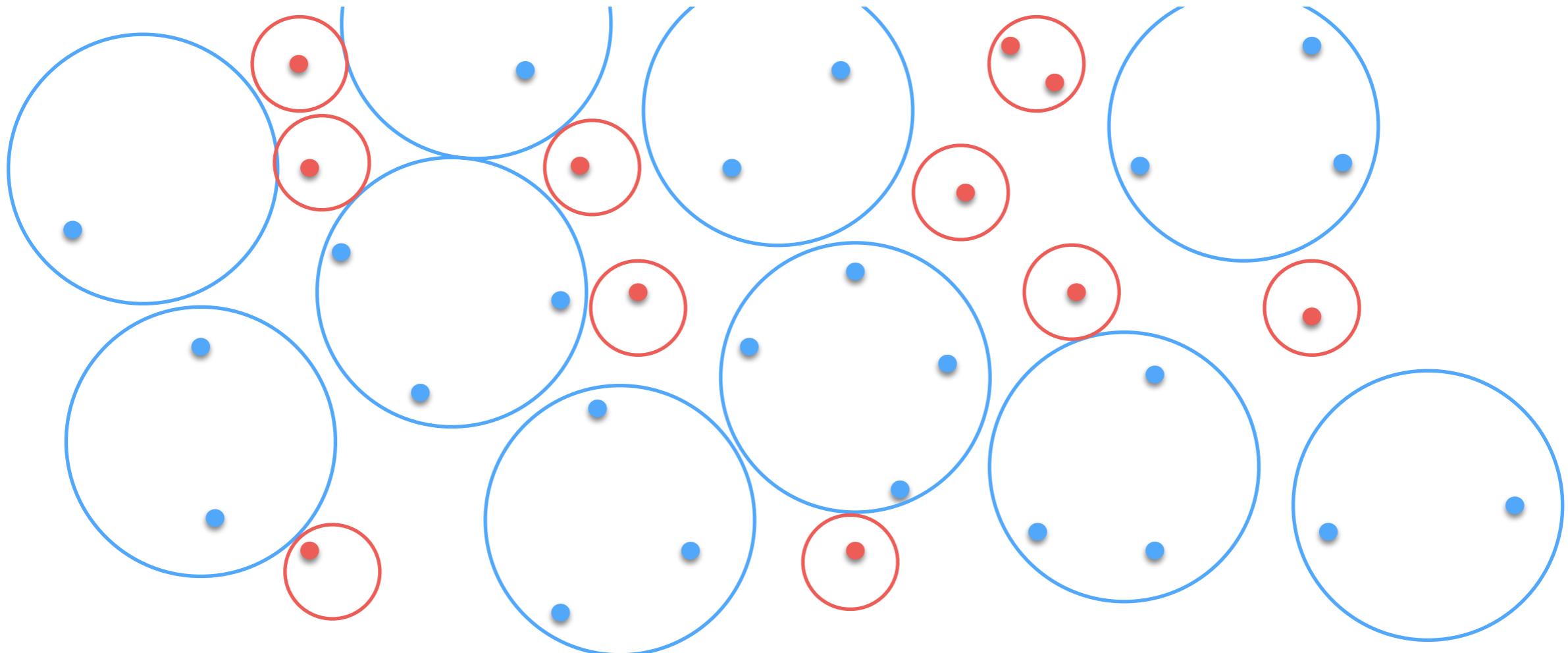
Imagine I have two colors of circle, **red** and **blue**. I want to estimate the **fraction of circles** that are **red** and **blue**. I'll *sample* from them by tossing down darts.



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

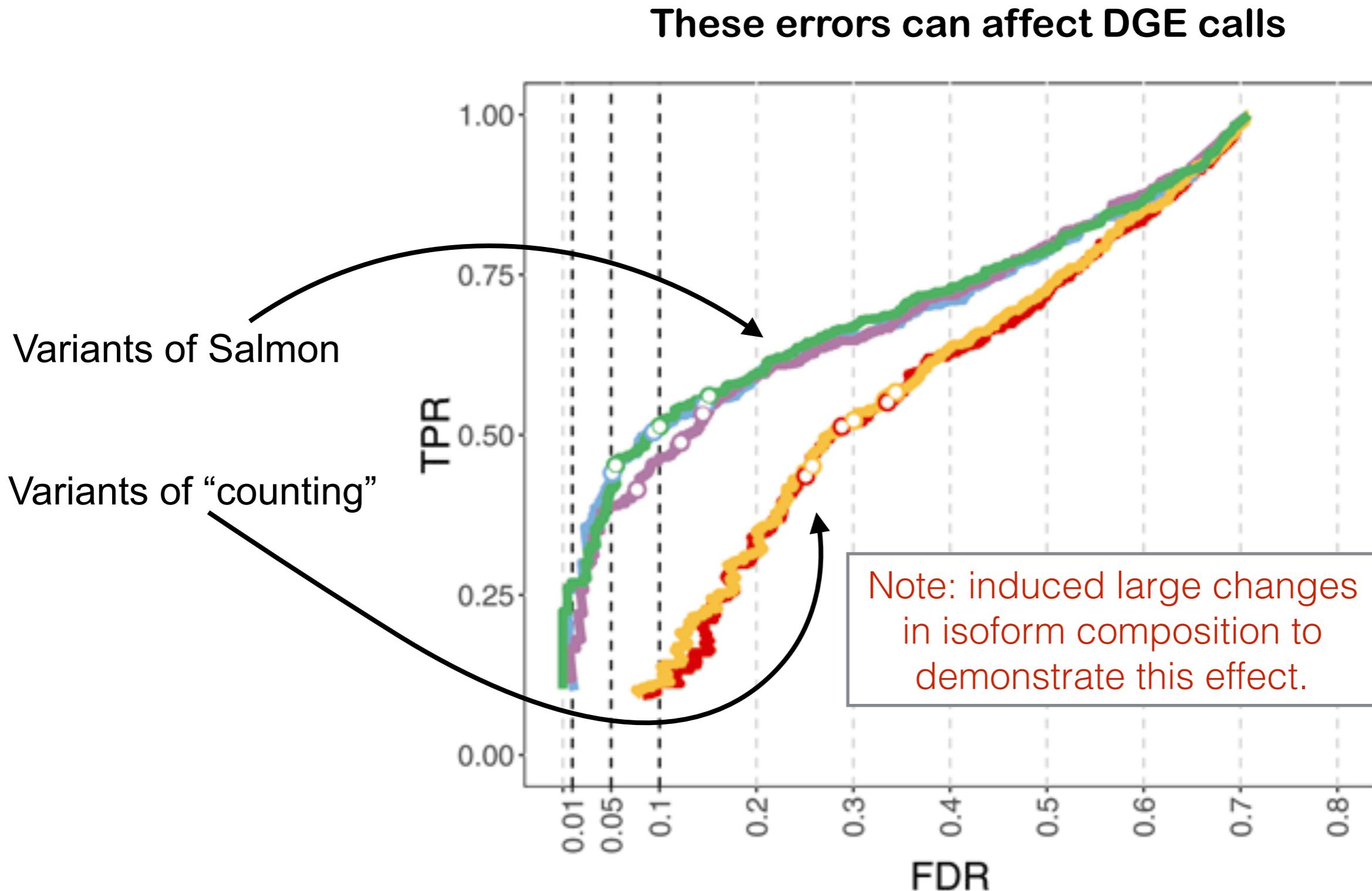
Imagine I have two colors of circle, **red** and **blue**. I want to estimate the **fraction of circles** that are **red** and **blue**. I'll *sample* from them by tossing down darts.

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

The areas!

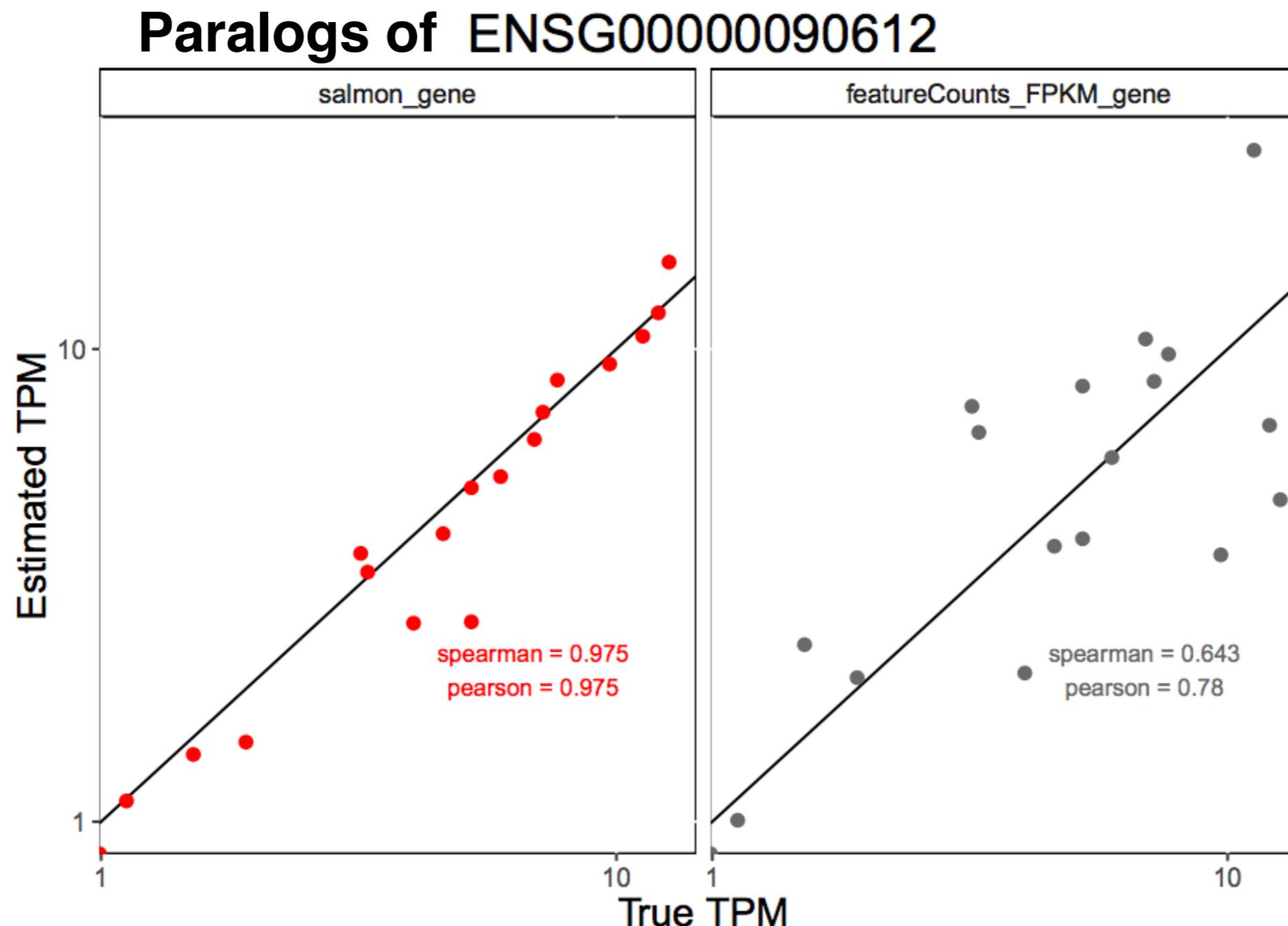
There is an analog in RNA-seq, one needs to know the **length** of the target from which one is drawing to meaningfully assess abundance!

Resolving multi-mapping is fundamental to quantification



Resolving multi-mapping is fundamental to quantification

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



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**

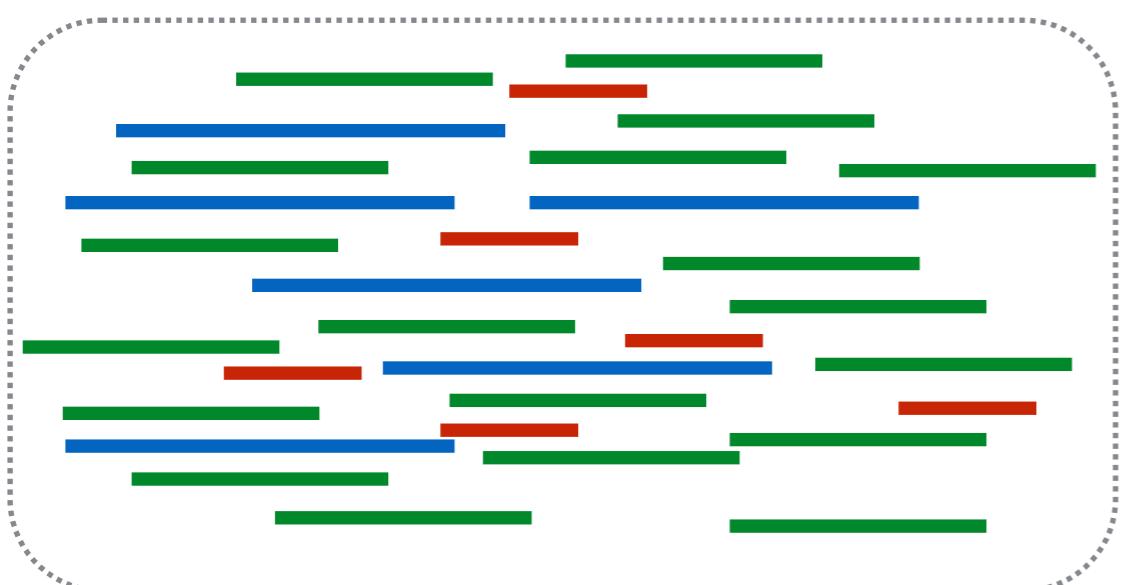
- uncertainty in ML estimate of abundances



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

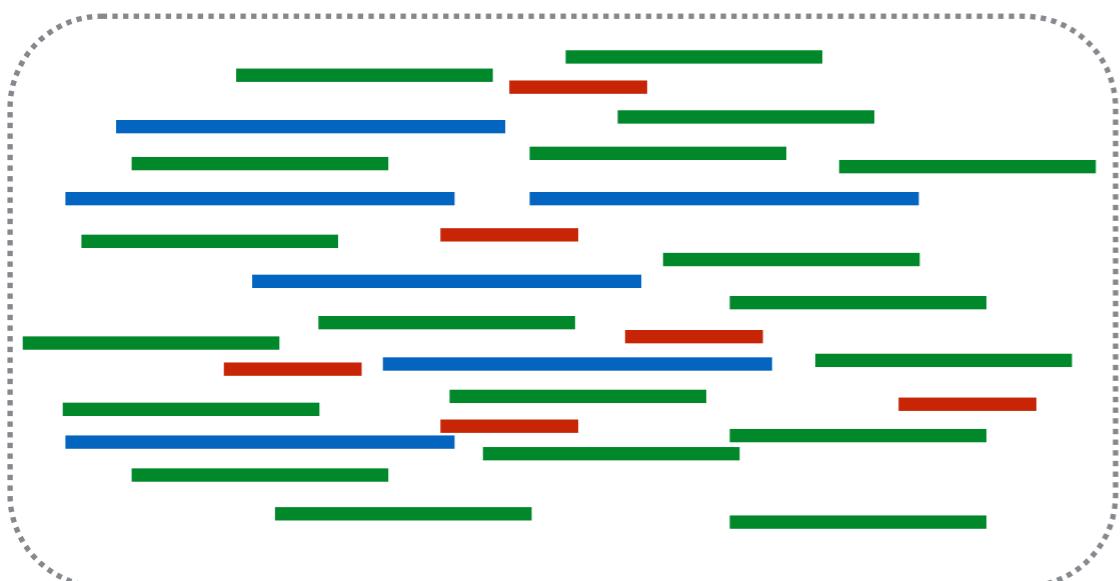
$$\text{length}(\text{blue bar}) = 100 \text{ nt} \times 6 \text{ copies} = 600 \text{ nt} \quad \sim 30\% \text{ blue}$$

$$\text{length}(\text{green bar}) = 66 \text{ nt} \times 19 \text{ copies} = 1254 \text{ nt} \quad \sim 60\% \text{ green}$$

$$\text{length}(\text{red bar}) = 33 \text{ nt} \times 6 \text{ copies} = 198 \text{ nt} \quad \sim 10\% \text{ red}$$

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

$$\text{length}(\text{--- blue ---}) = 100 \text{ nt} \times 6 \text{ copies} = 600 \text{ nt} \quad \sim 30\% \text{ blue}$$

$$\text{length}(\text{--- green ---}) = 66 \text{ nt} \times 19 \text{ copies} = 1254 \text{ nt} \quad \sim 60\% \text{ green}$$

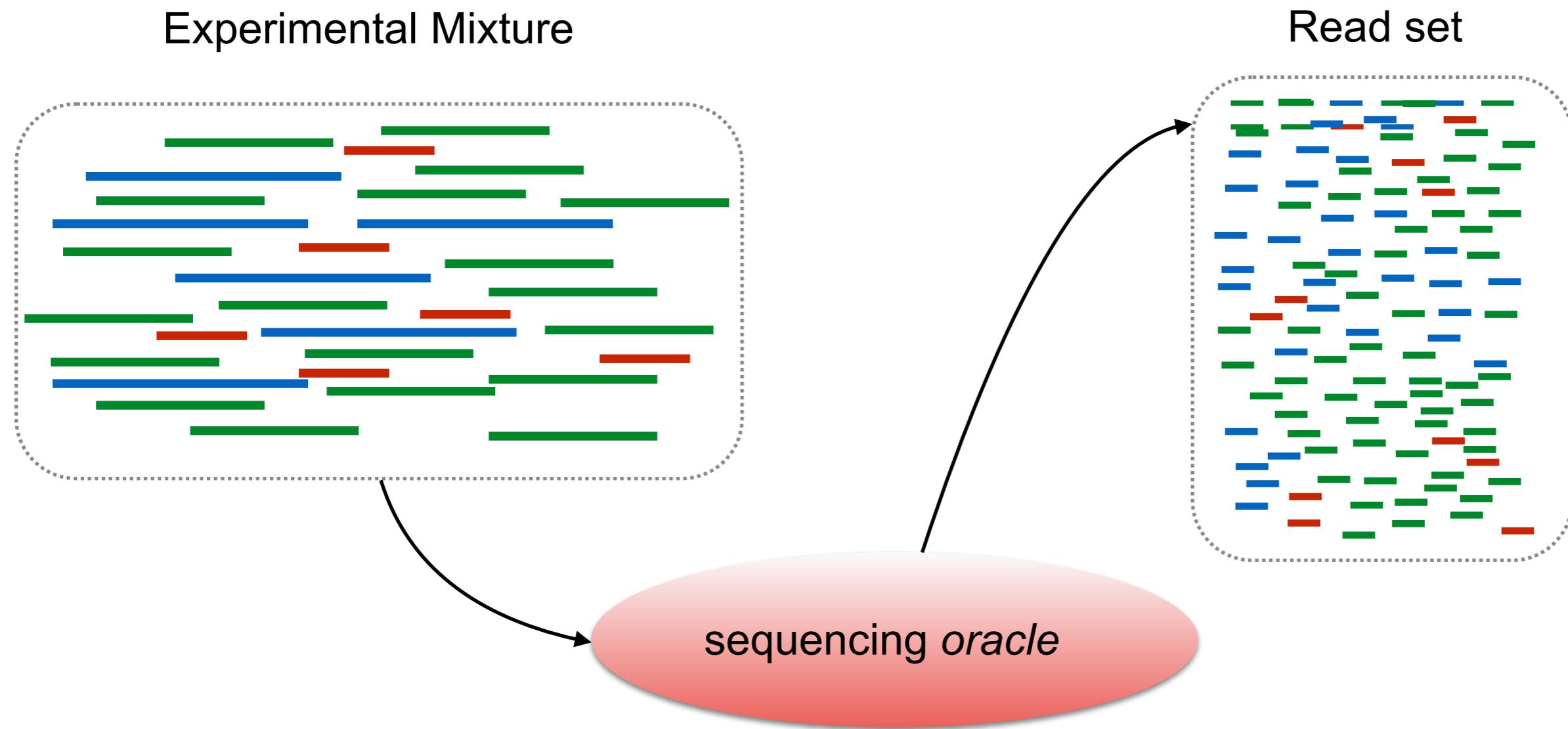
$$\text{length}(\text{--- red ---}) = 33 \text{ nt} \times 6 \text{ copies} = 198 \text{ nt} \quad \sim 10\% \text{ red}$$



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

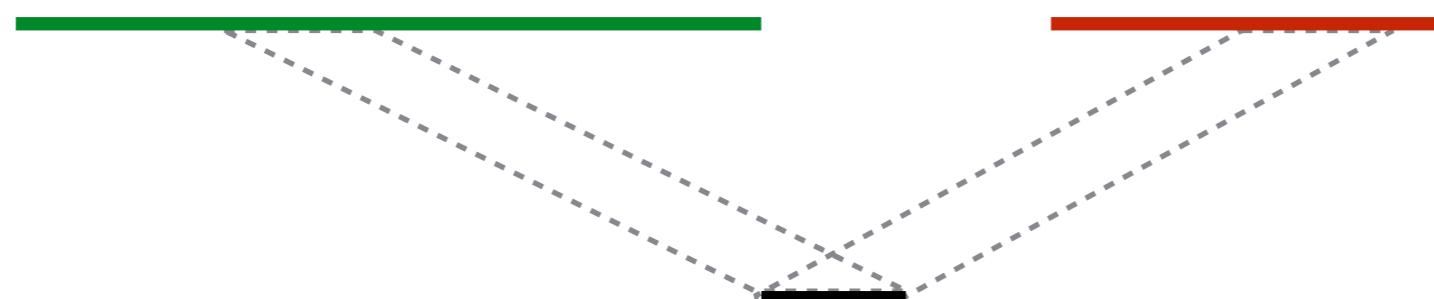
How can we perform inference from sequenced fragments?

Think about the “ideal” RNA-seq experiment . . .



- (1) Pick transcript $t \propto$ 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

$$\text{length}() = 100 \times 6 \text{ copies} = 600 \text{ nt} \sim 30\% \text{ blue}$$

$$\text{length}() = 66 \times 19 \text{ copies} = 1254 \text{ nt} \sim 60\% \text{ green}$$

$$\text{length}() = 33 \times 6 \text{ copies} = 198 \text{ nt} \sim 10\% \text{ red}$$

Units for Relative Abundance

TPM (Transcripts Per Million)

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

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

Reads coming from
transcript i

Units for Relative Abundance

TPM (Transcripts Per Million)

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

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

Reads coming from transcript i

Length of transcript i

Units for Relative Abundance

TPM (Transcripts Per Million)

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

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

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

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

nucleotide fractions known transcriptome assumes independence of fragments

observed fragments (reads)

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

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

Prob. of selecting t_i given $\boldsymbol{\eta}$

Depends on abundance estimate

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

Independent of abundance estimate

We want to find the values of $\boldsymbol{\eta}$ that **maximize** this probability.
We can do this (at least locally) using the EM algorithm.

A probabilistic view of RNA-Seq quantification

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

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

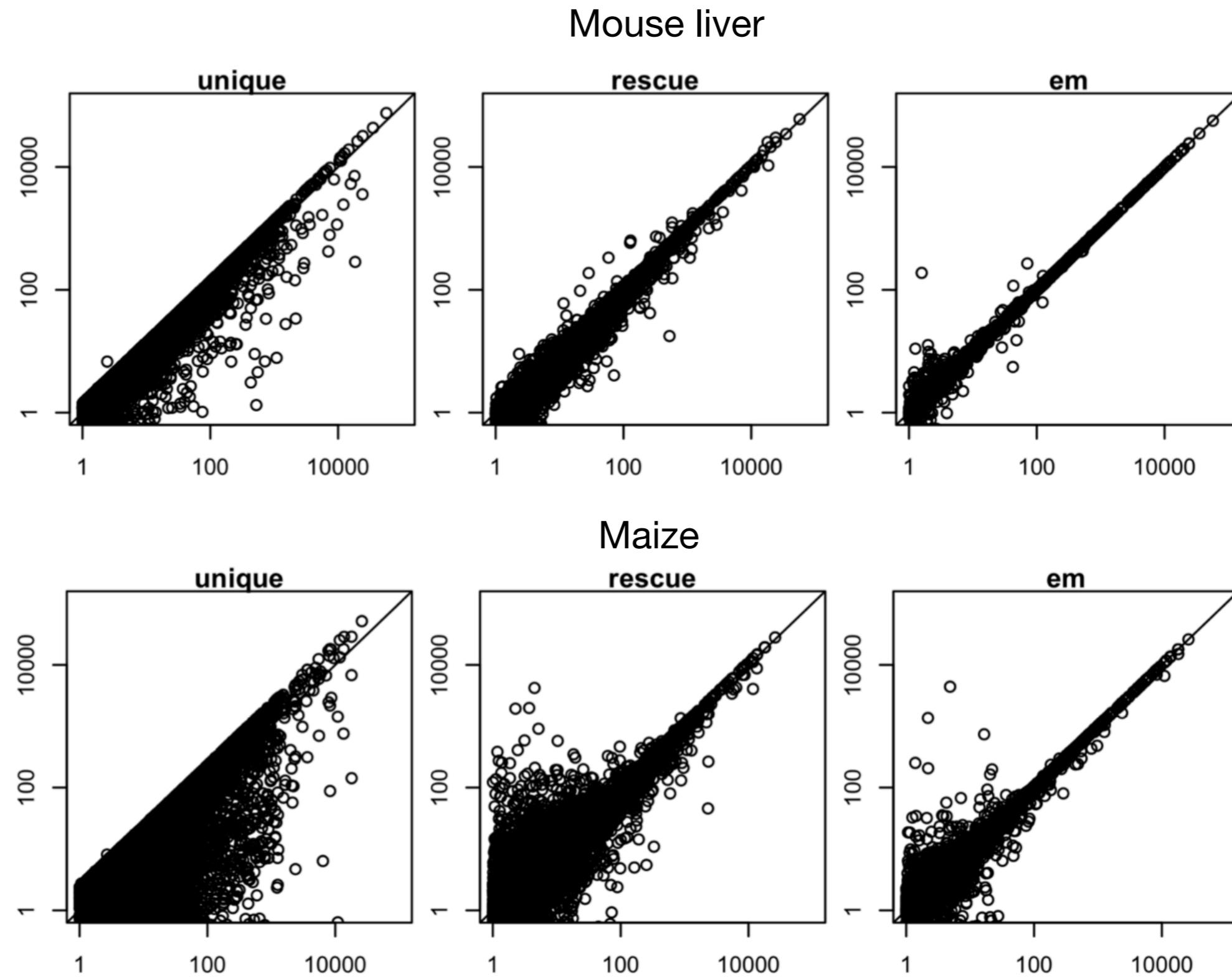
M-step: Given these soft assignments, how abundant is each transcript?

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

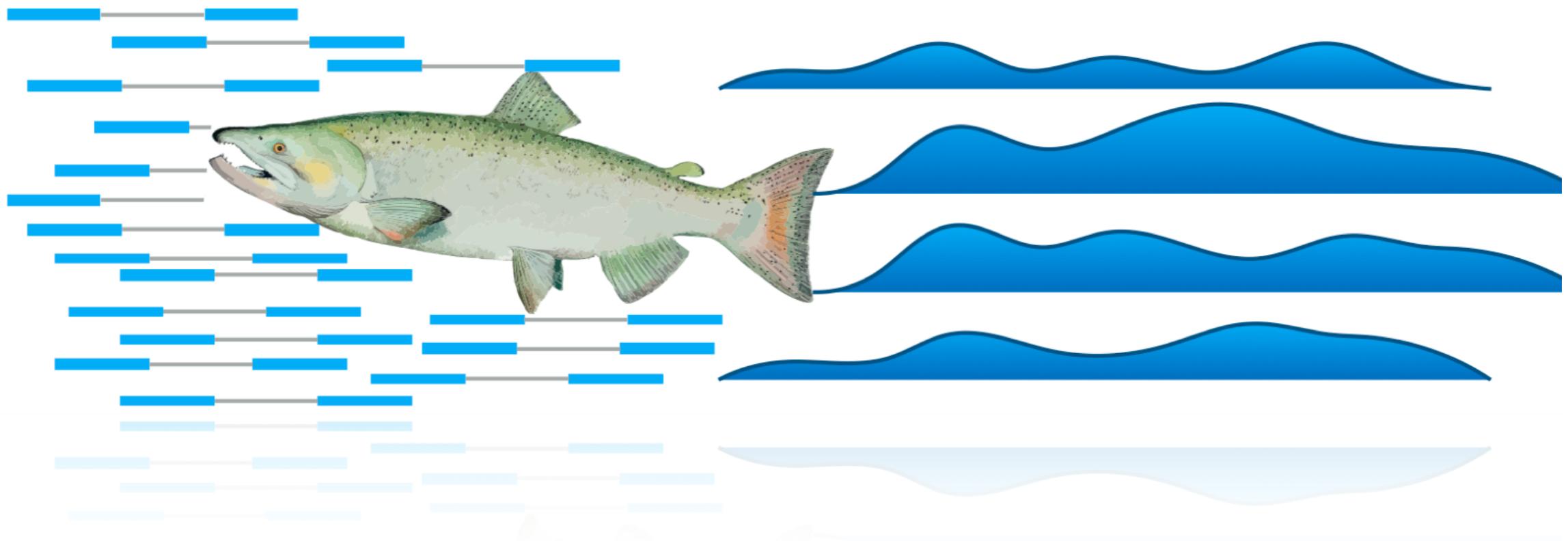
$$\text{where } C_i = \sum_{n,i,j} P(Z_{nij} = 1 | f, \eta^{(t)})$$

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

Gene expression estimation accuracy in simulated data

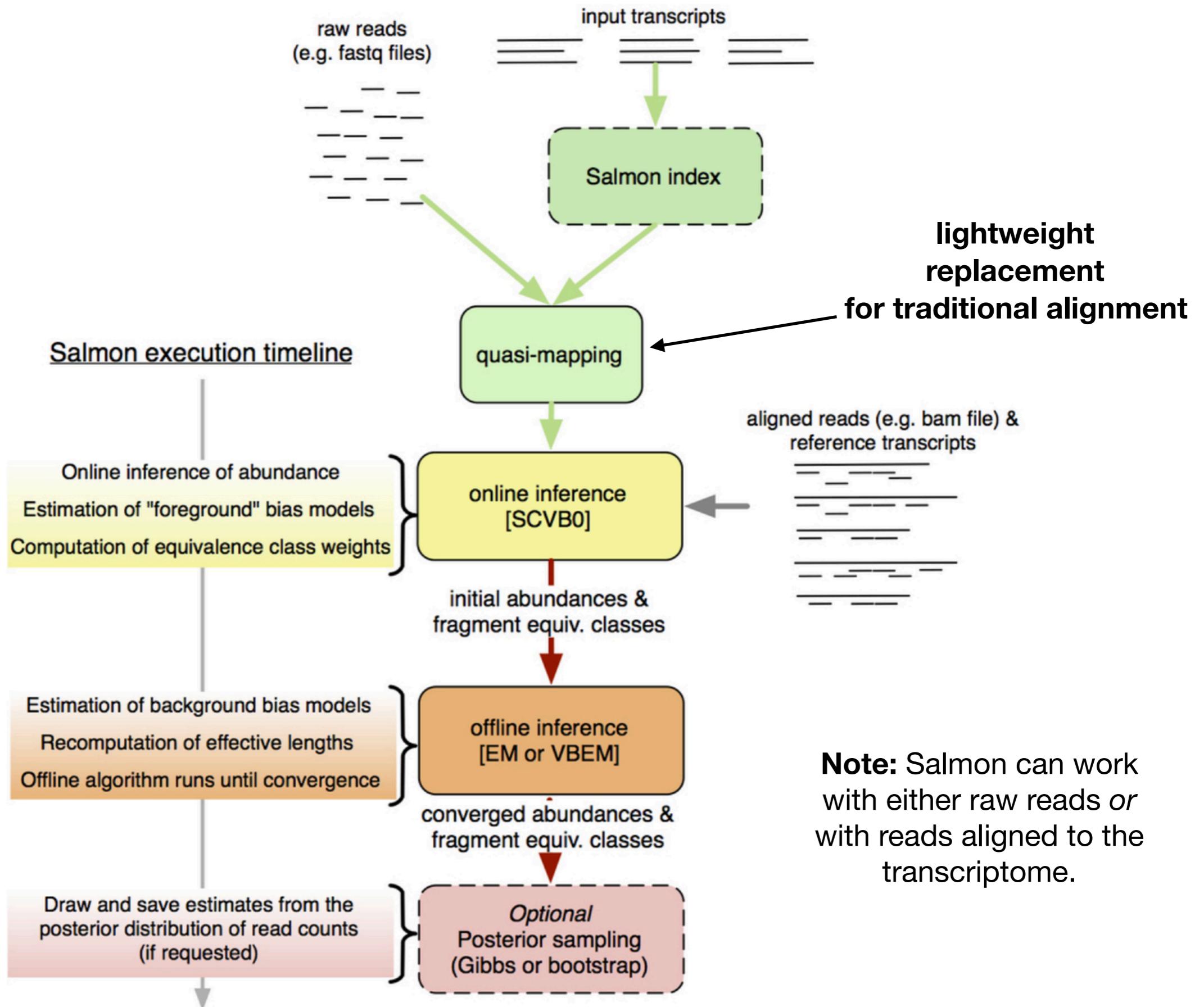


Salmon provides fast and bias-aware quantification of transcript expression



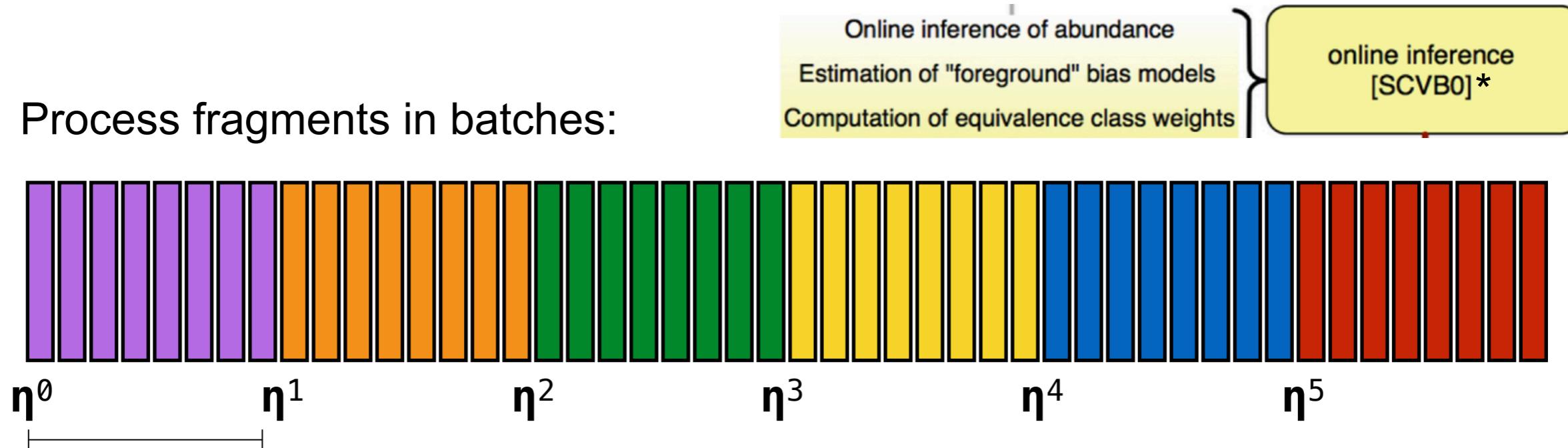
Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017).
Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*.

Salmon's “pipeline”



Phase 1: Online Inference (asynchronous!)

Process fragments in batches:



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

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

Update “bias” model

Place mappings in **equivalence classes**

Weighting factor that decays over time

- 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 η^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 η^{t-1} , and aux params

}

normalize alignment probs & update local transcript weights η'

add / update the equivalence class for read r

sample $a \in r$ to update auxiliary models

}

update global transcript weights given local transcript

weights according to “update rule” $\Rightarrow \eta^t = \eta^{t-1} + w^t \eta'$

}

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:

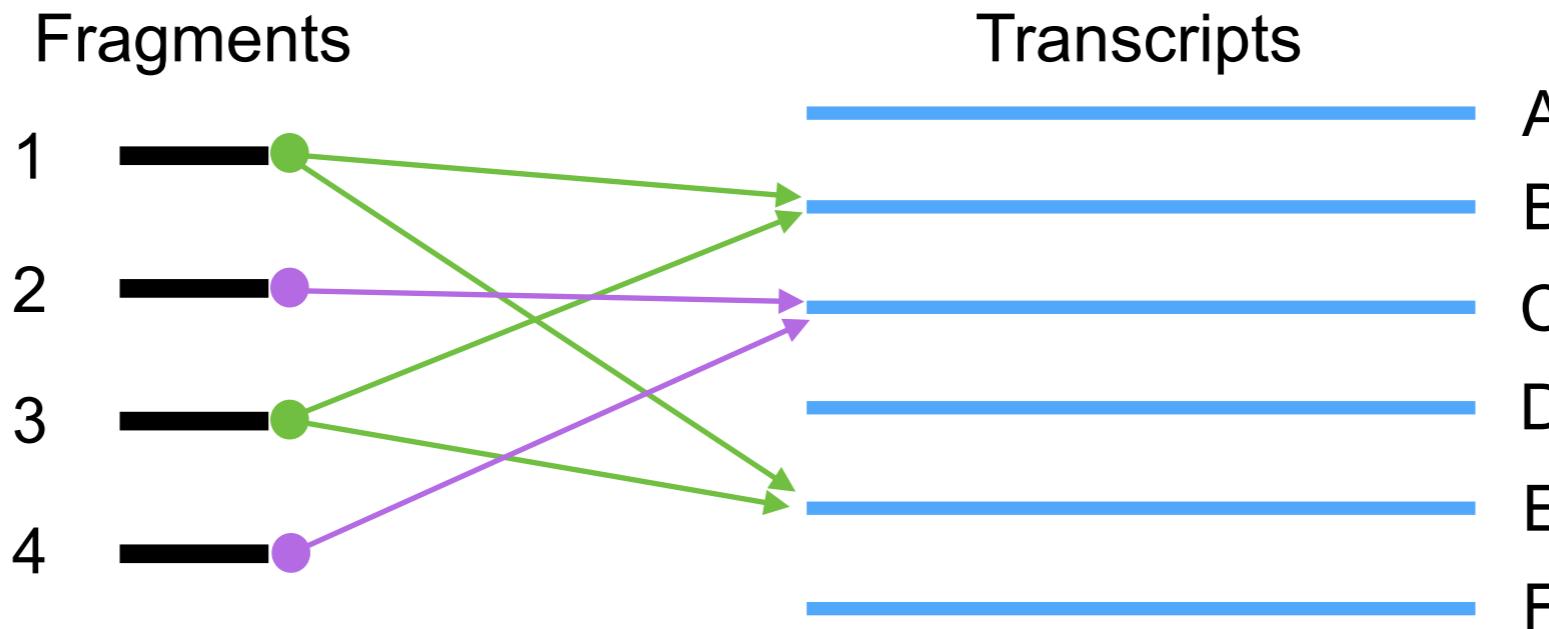
1. Allocate them to transcripts
2. Update auxiliary models
3. Place them in **equivalence classes**

online inference
[SCVB0]

We use a streaming, parallel, stochastic inference algorithm for Phase 1; a variant of **Stochastic Collapsed Variational Bayesian 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 $P\{f_j | t_i\}$, 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)

The number of equivalence classes is small

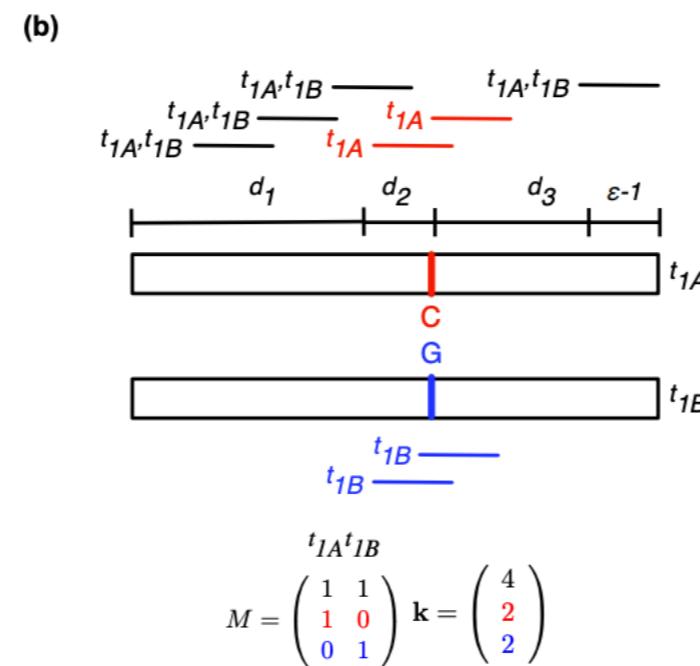
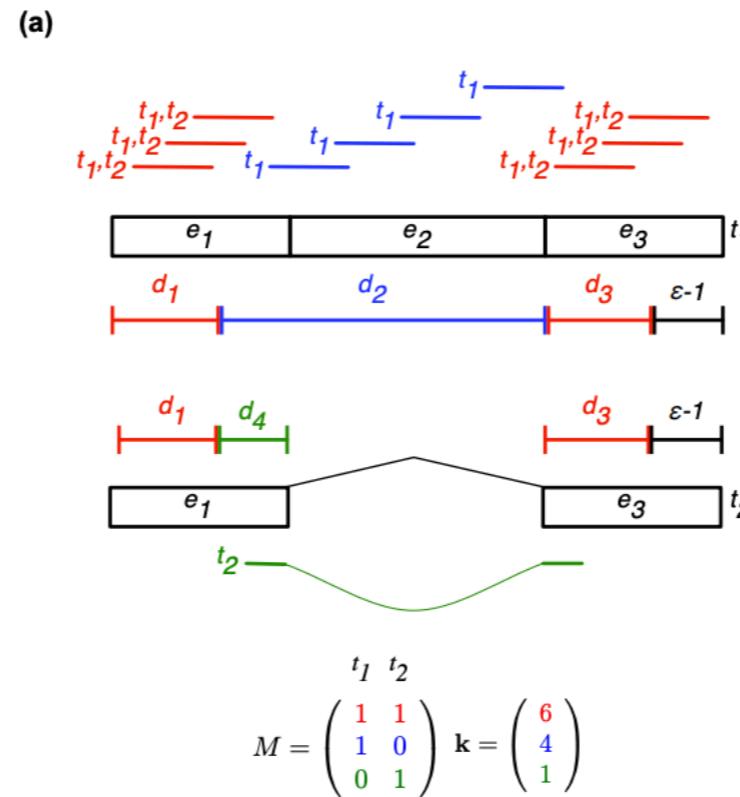
	Yeast	Human	Chicken
# contigs	7353	107,389	335,377
# samples	6	6	8
Total (paired-end) reads	~36,000,000	~116,000,000	~181,402,780
Avg # eq. classes (across samples)	5197	100,535	222,216

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



This lets us approximate the likelihood efficiently

Approximate this:

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

product over all fragments

sum over all alignments of fragment

with this:

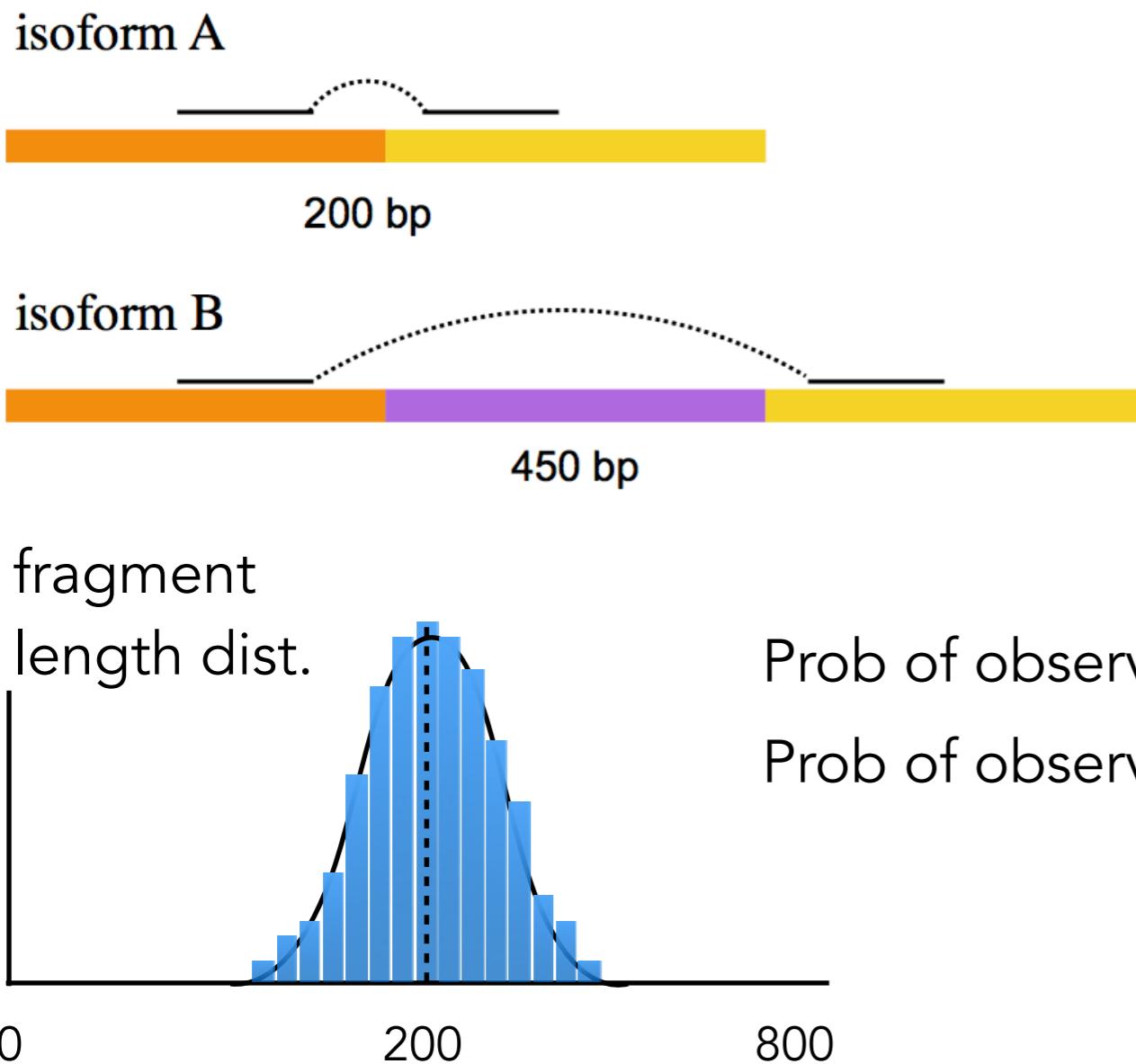
$$\mathcal{L}(\boldsymbol{\eta}; \mathcal{F}) \approx \prod_{\mathcal{F}^q \in \mathcal{C}} \left(\sum_{\langle i, t_i \rangle \in \Omega(\mathcal{F}^q)} \Pr(t_i \mid \boldsymbol{\eta}) \cdot \Pr(f \mid \mathcal{F}^q, t_i) \right)^{N^q}$$

product over all equivalence classes

sum over all transcripts labeling this eq. class

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

Consider the following scenario:



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

Many terms can be considered in a general “fragment-transcript agreement” model¹. 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

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

estimated read count from transcript i
at iteration u+1

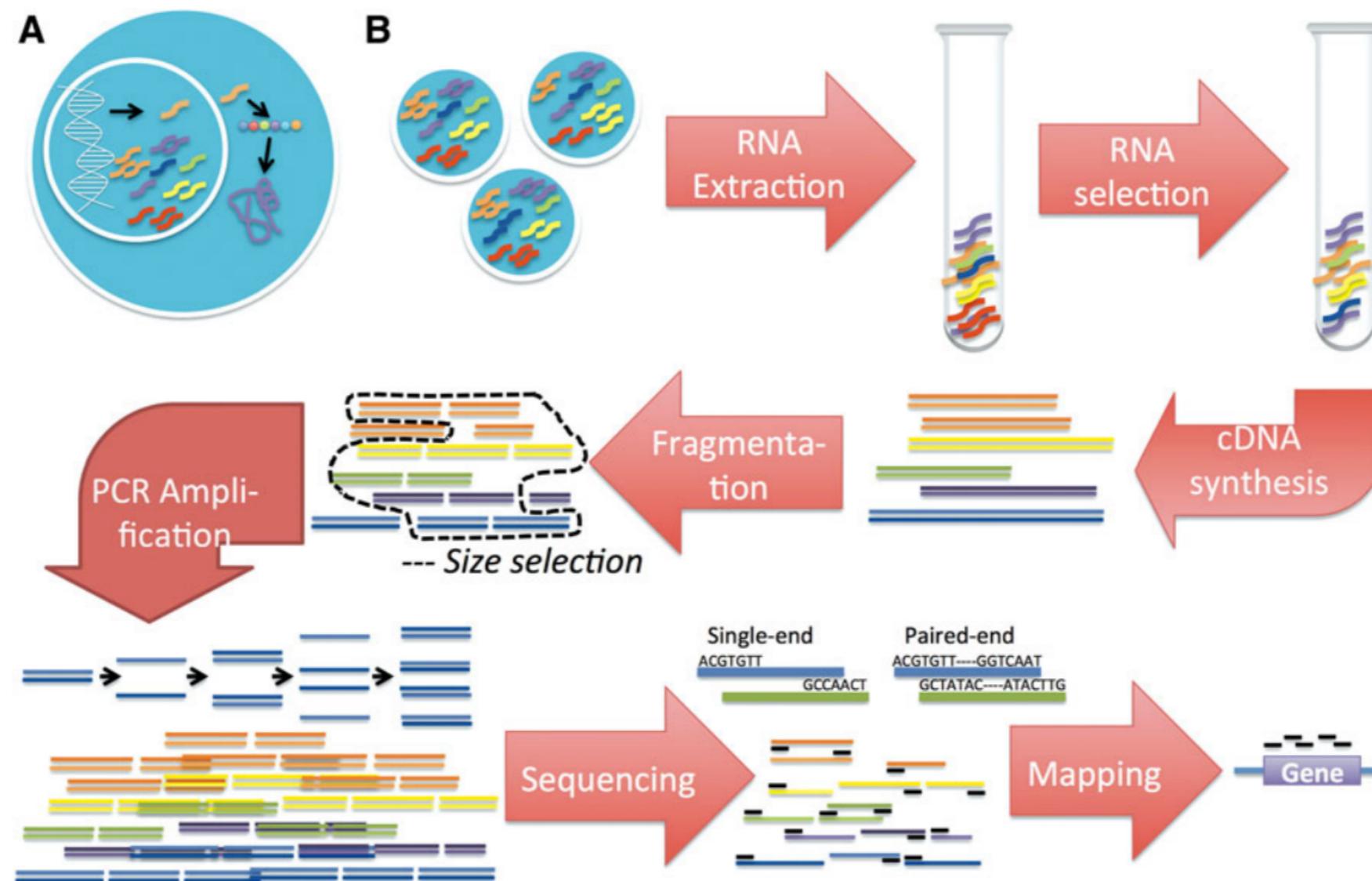
count of eq.
class j

weight of t_i in eq.
class q

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

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.

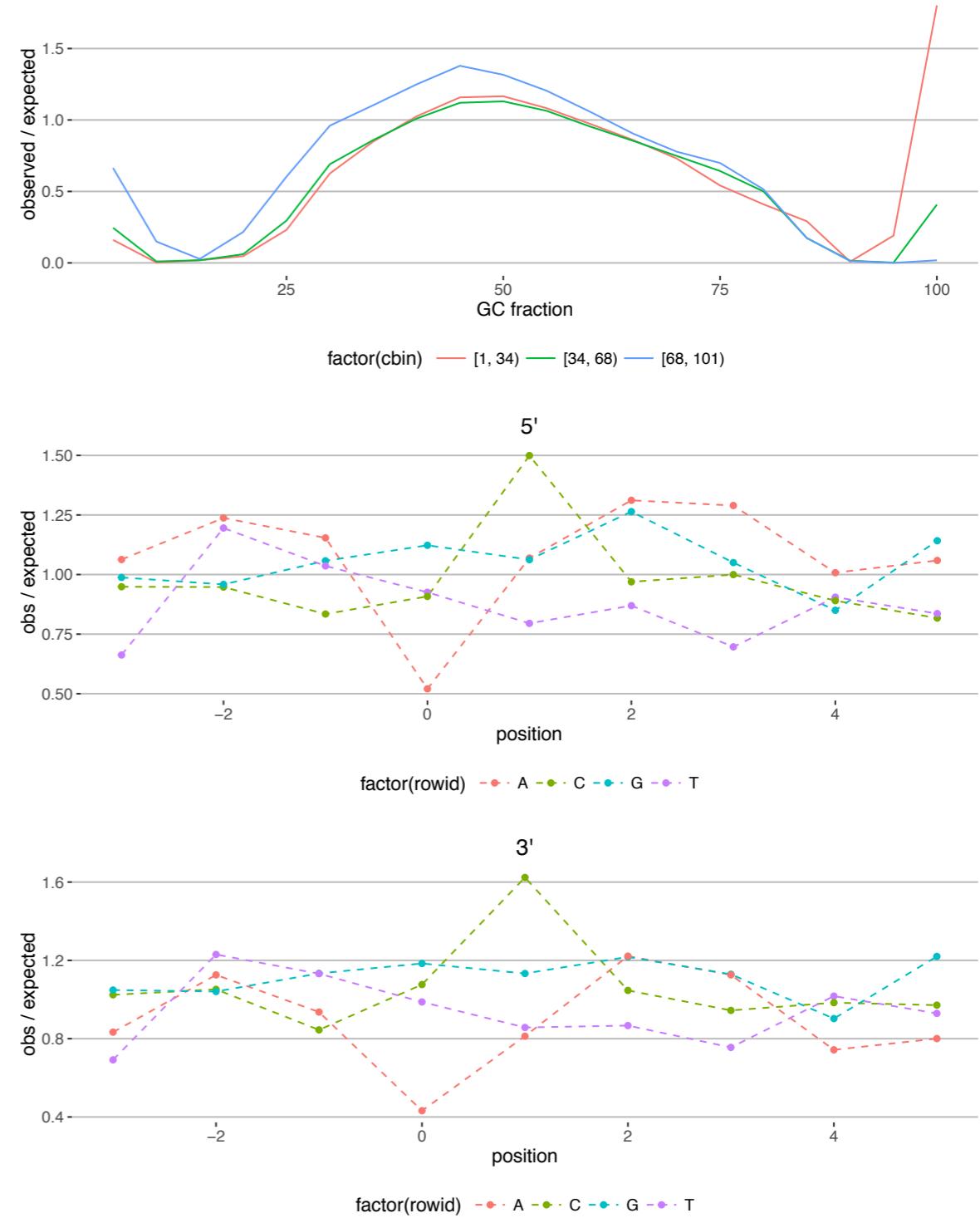
Biases abound in RNA-seq data

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

Fragment gc-bias¹—
The GC-content of the fragment affects the likelihood of sequencing

Sequence-specific bias²—
sequences surrounding fragment affect the likelihood of sequencing

Positional bias²—
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.

Biases abound in RNA-seq data

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.

Fragment gc-bias¹—

The GC-content of the fragment

affects the likelihood of sequencing

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.

Positional bias²—

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

1:Love, Michael L., 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.

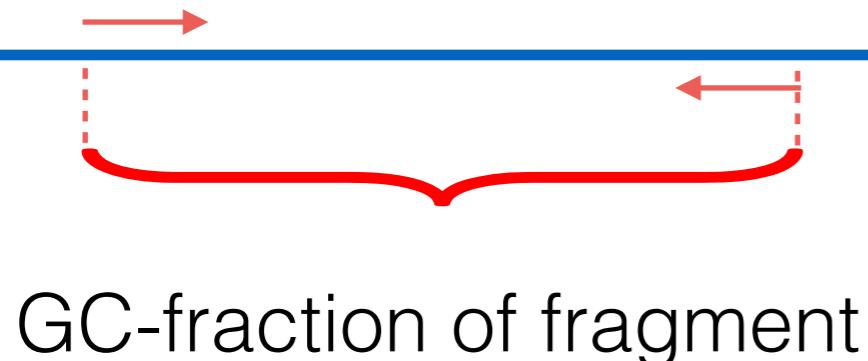
Bias Modeling

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

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

Fragment GC bias model:

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

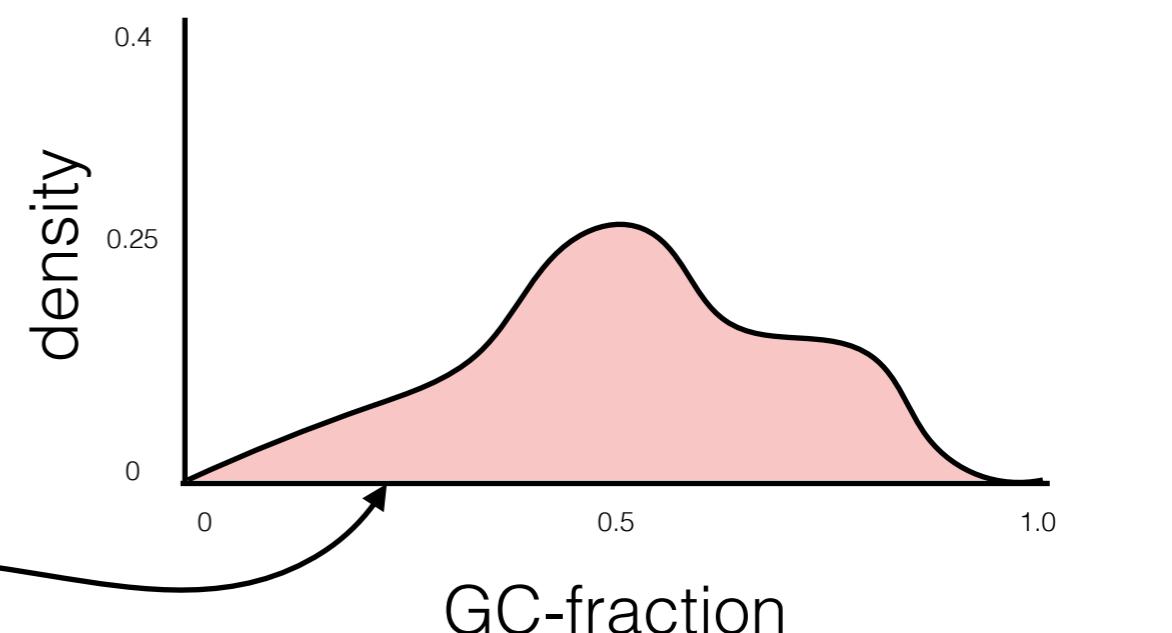


Foreground:

Observed

Background:

Expected given est. abundances



Bias Modeling

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

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

Seq-specific bias model*:

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

Foreground:

Observed

Background:

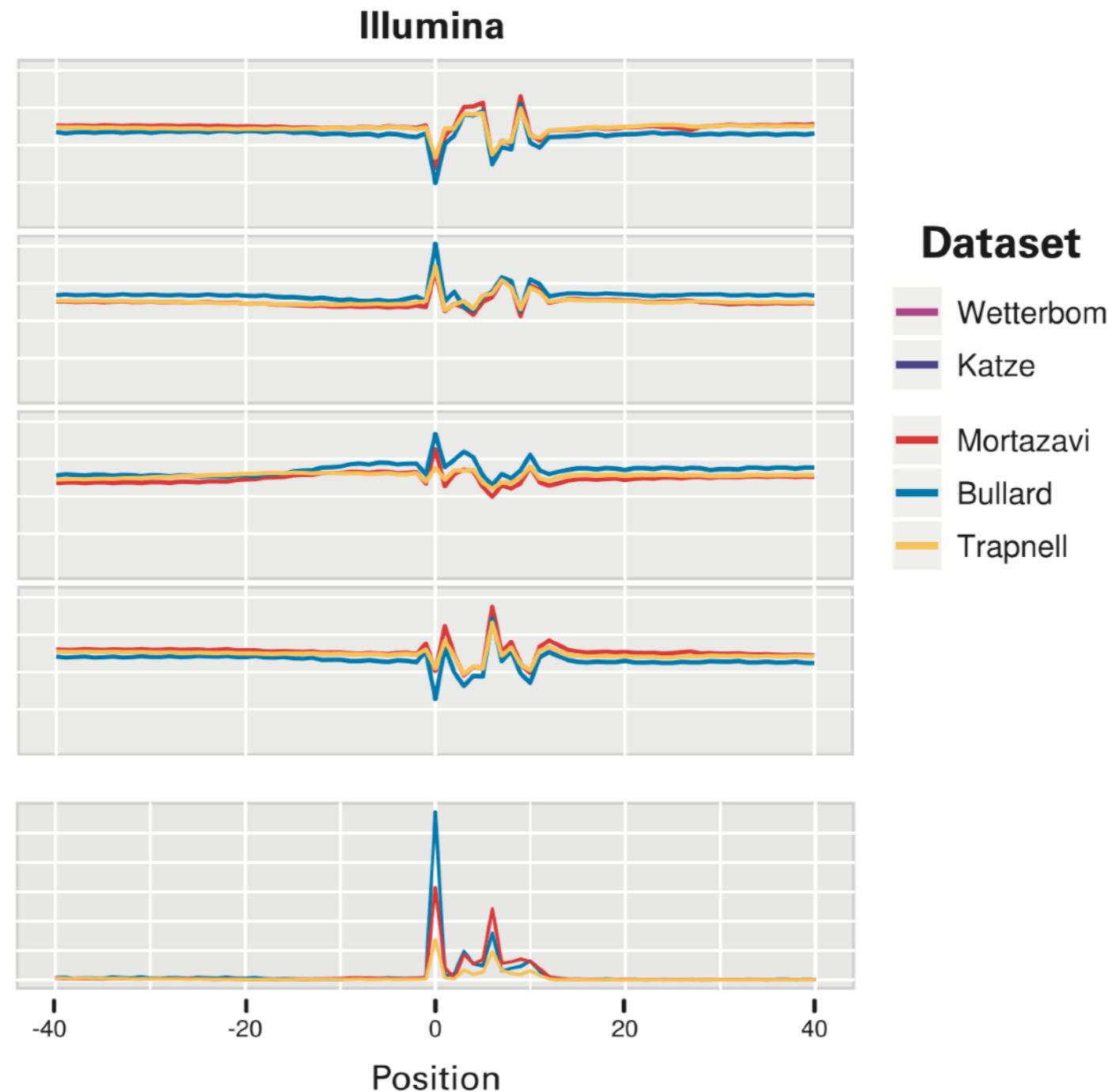
Expected given est. abundances



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

Same, but independent
model for 3' end

Priming bias is sample & sequence-specific



Basic idea

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

$$E[x_i | s_i] = N \Pr[m_i | s_i] = N \Pr[m_i] = E[x_i]$$

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

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

So that:

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

Priming bias is sample & sequence-specific

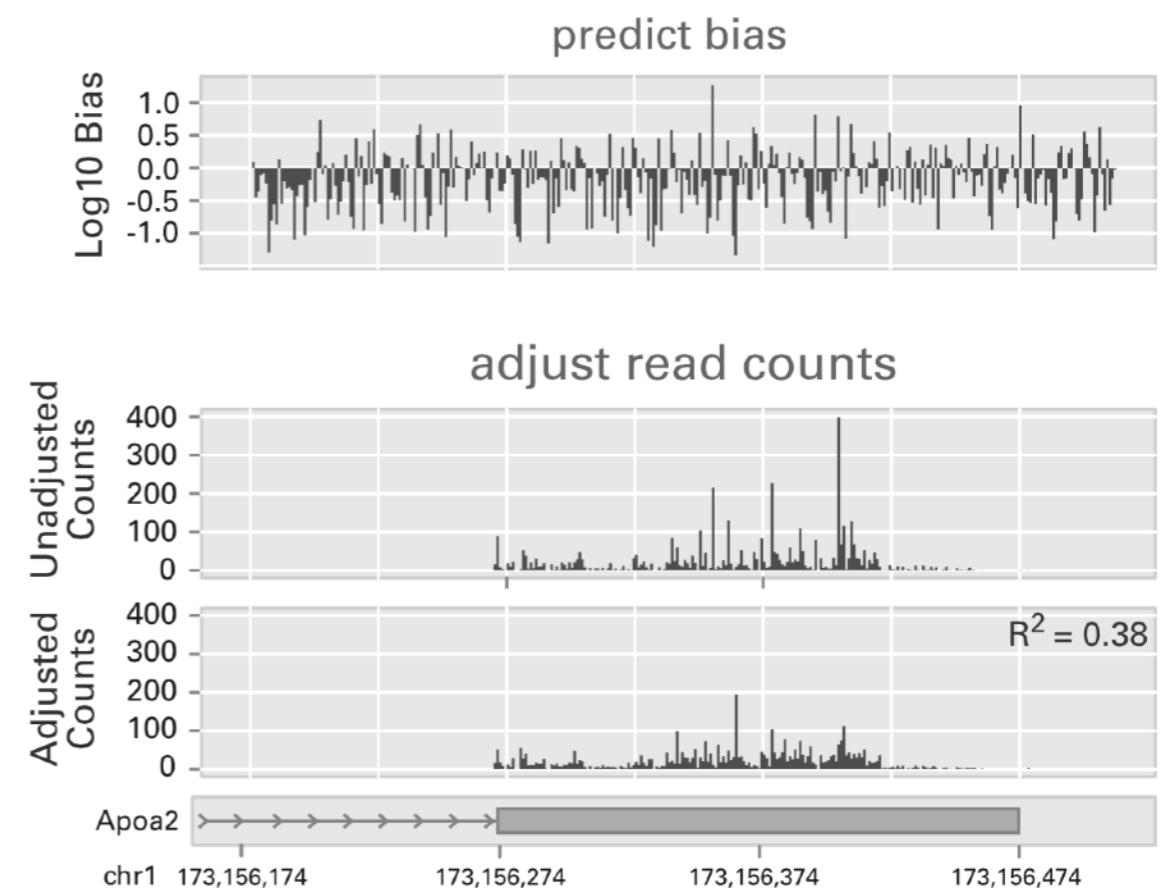
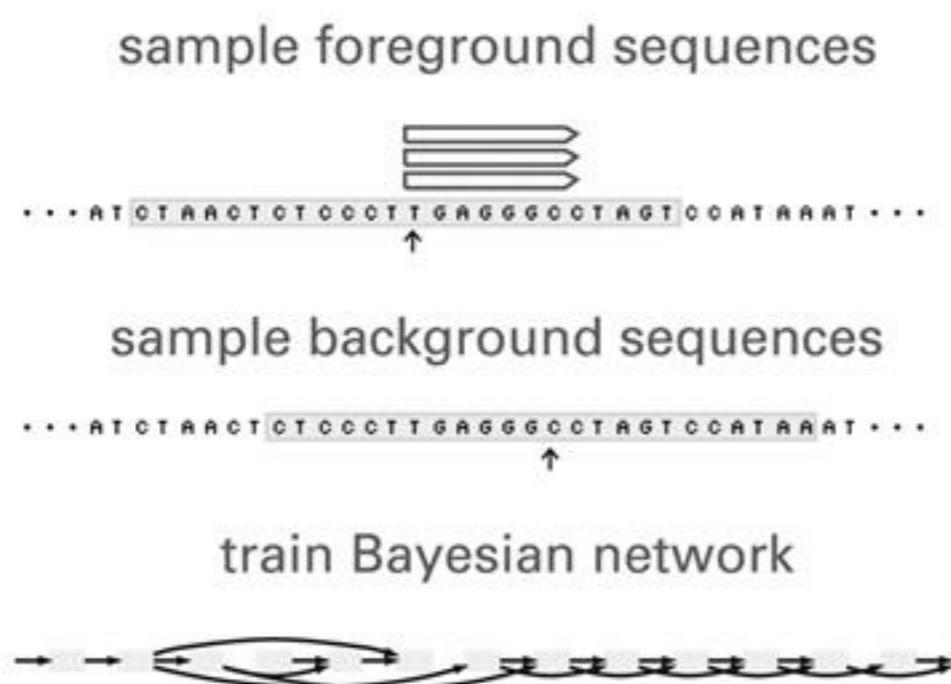
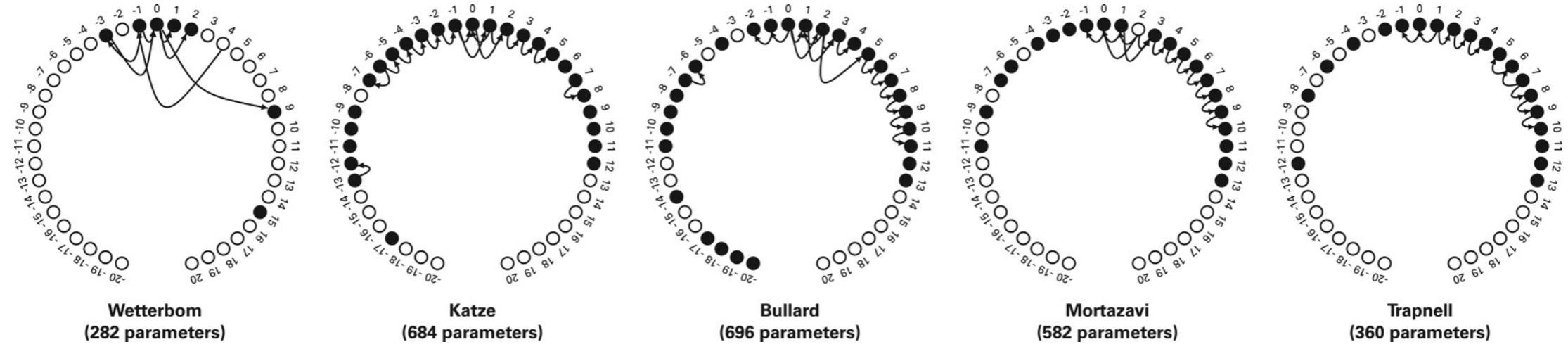


Table 3. The Pearson's correlation coefficient r between log-adjusted read counts and log-adjusted TaqMan values

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

The best *model* may also be sample-specific



Contrast with Roberts et al. which uses a fixed-structure VLMM to model the sample-specific bias.

Bias Modeling

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

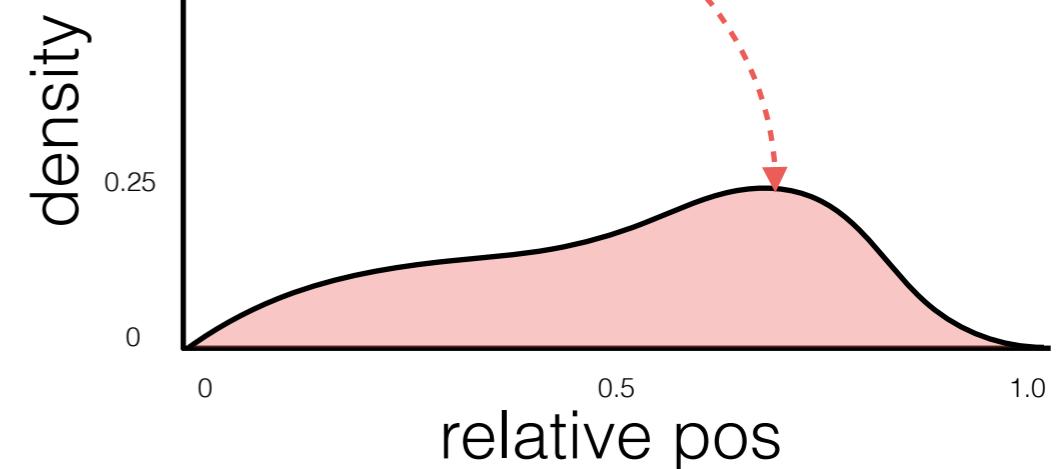
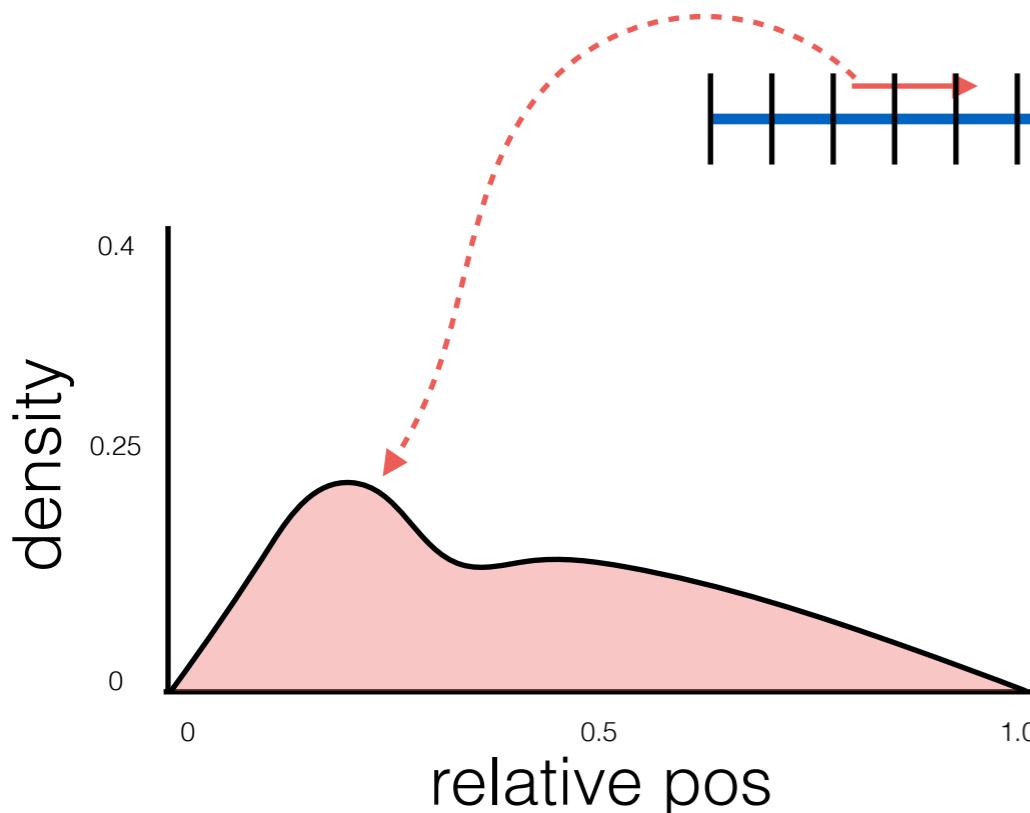
$$\tilde{\ell}'_i = \sum_{j=1}^{j \leq \ell_i} \sum_{k=1}^{k \leq f_i(j, L)} \frac{b_{gc+}(t_i, j, j+k)}{b_{gc-}(t_i, j, j+k)} \cdot \frac{b_{s+}^{5'}(t_i, j)}{b_{s-}^{5'}(t_i, j)} \cdot \frac{b_{s+}^{3'}(t_i, j+k)}{b_{s-}^{3'}(t_i, j+k)} \cdot \frac{b_{p+}^{5'}(t_i, j+k)}{b_{p-}^{5'}(t_i, j+k)} \cdot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

Position bias model*:

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

Foreground:
Observed

Background:
Expected given est. abundances



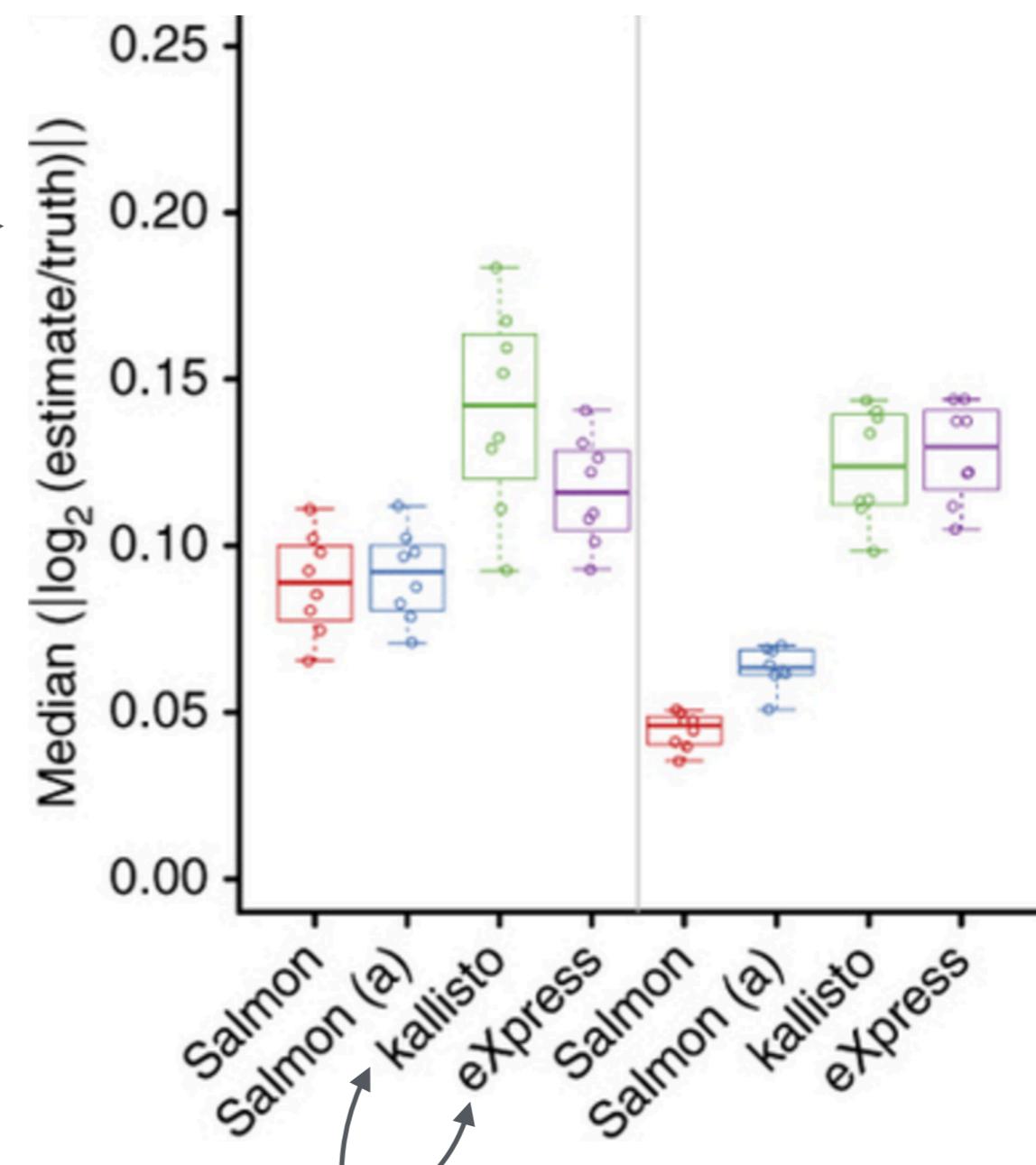
*Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

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



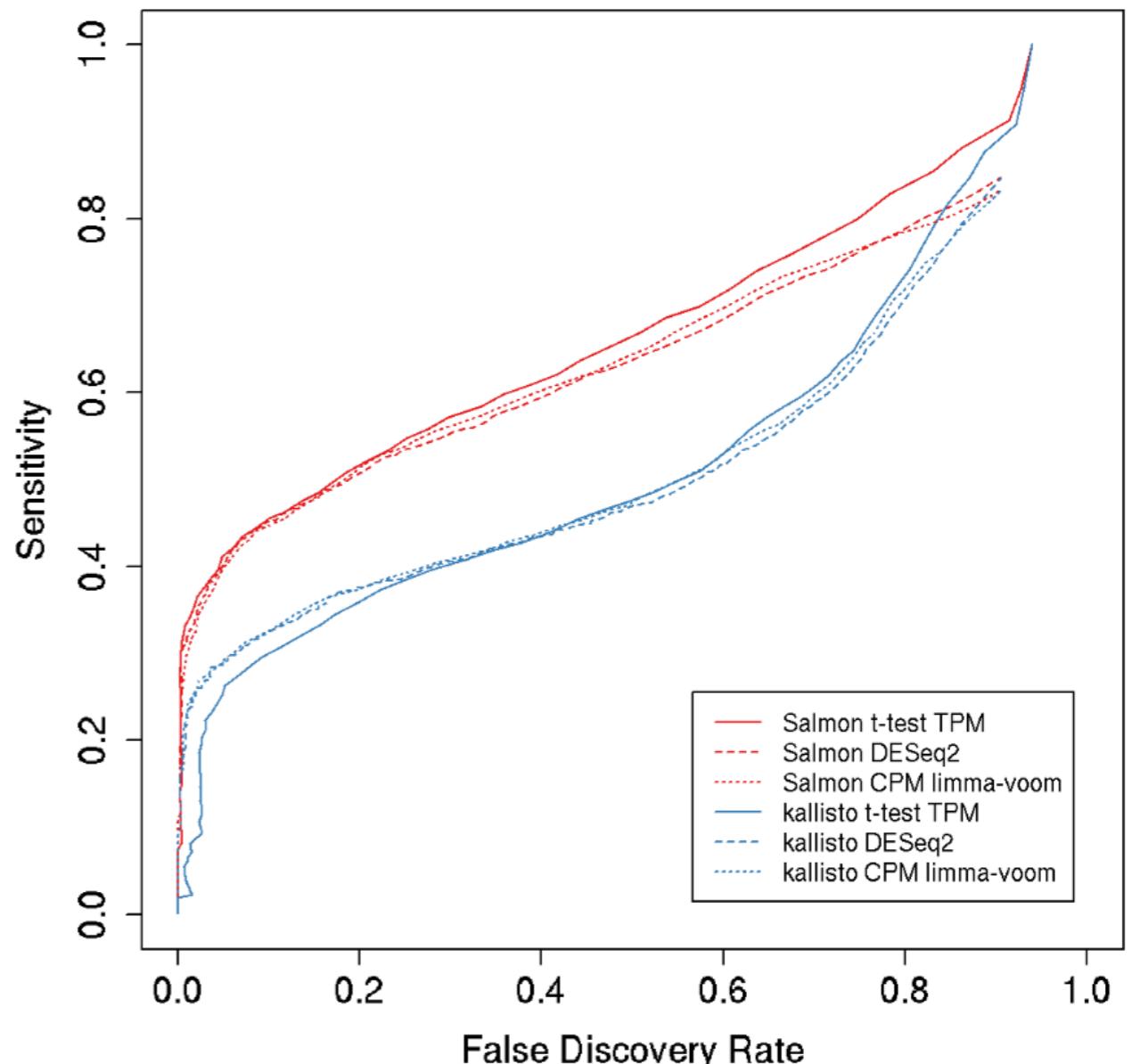
Sequence-bias models don't account for fragment-level GC bias

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 in << 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**

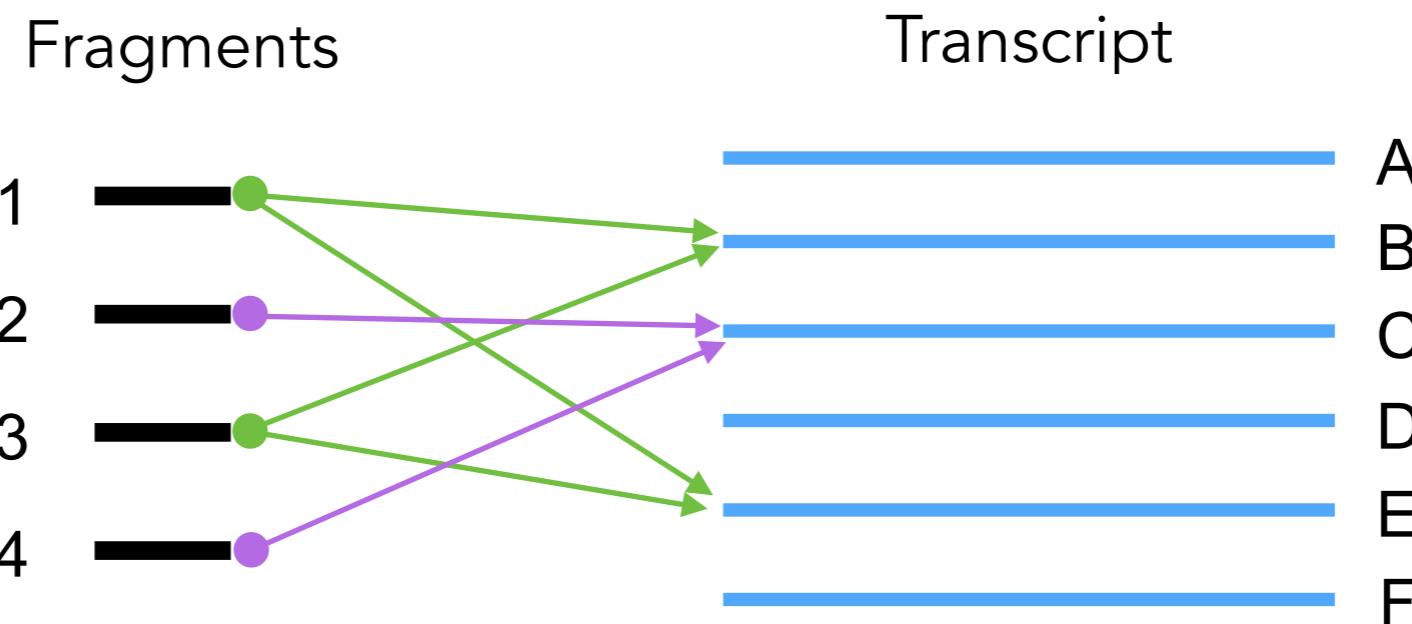
$$\mathcal{L}(\boldsymbol{\eta}; \mathcal{F}, \mathcal{T}) = \prod_{f \in \mathcal{F}} \sum_{t_i \in \Omega(f)} \Pr(t_i \mid \boldsymbol{\eta}) \Pr(f \mid t_i)$$

Set of transcripts where f maps/aligns

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 Salzman¹, Hui Jiang¹ and Wing Hung Wong

RESEARCH

Estimation of alternative splicing isoform frequencies from RNA-Seq data

Marius Nicolae^{1*}, Serghei Mangul², Ion I Măndoiu¹ and Alex Zelikovsky²

“Traditional”
approaches



METHOD

Open Access

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

Ernest Turro^{1*}, Shu-Yi Su², Ângela Gonçalves³, Lachlan JM Coin¹, Sylvia Richardson¹, Alex Lewin¹

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

Rob Patro¹, Stephen M Mount^{2,3} & Carl Kingsford¹

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 Bray¹, Harold Pimentel², Pál Melsted³
& Lior Pachter^{2,4,5}

Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro¹, Geet Duggal², Michael I Love^{3,4},
Rafael A Irizarry^{3,4} & Carl Kingsford⁵

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:

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

sum over all alignments of fragment

product over all fragments

with this:

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

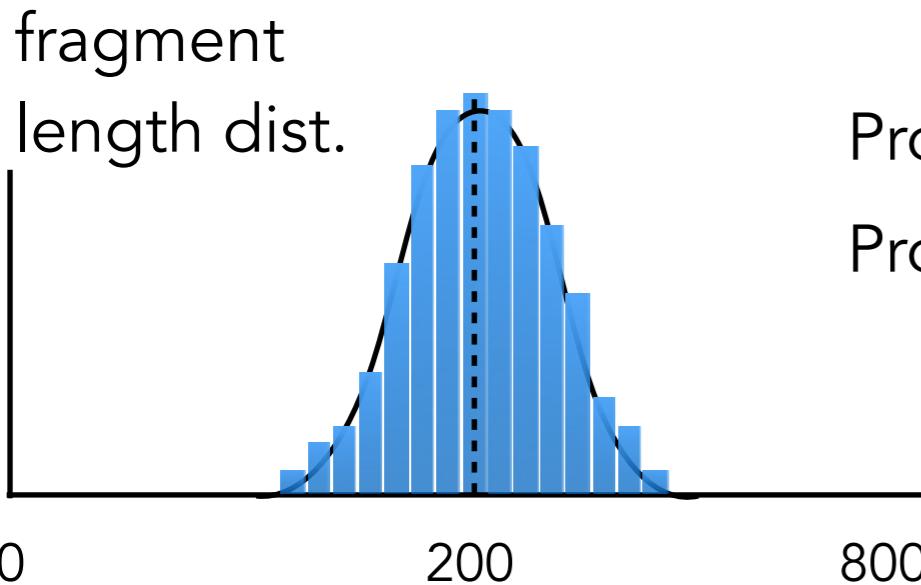
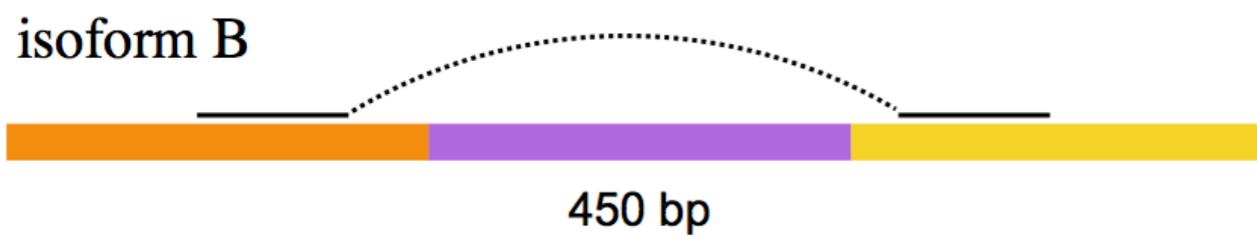
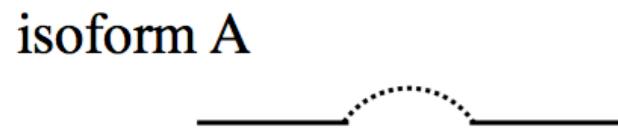
sum over all transcripts labeling this eq. class

product over all equivalence classes

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

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

Consider the following scenario:



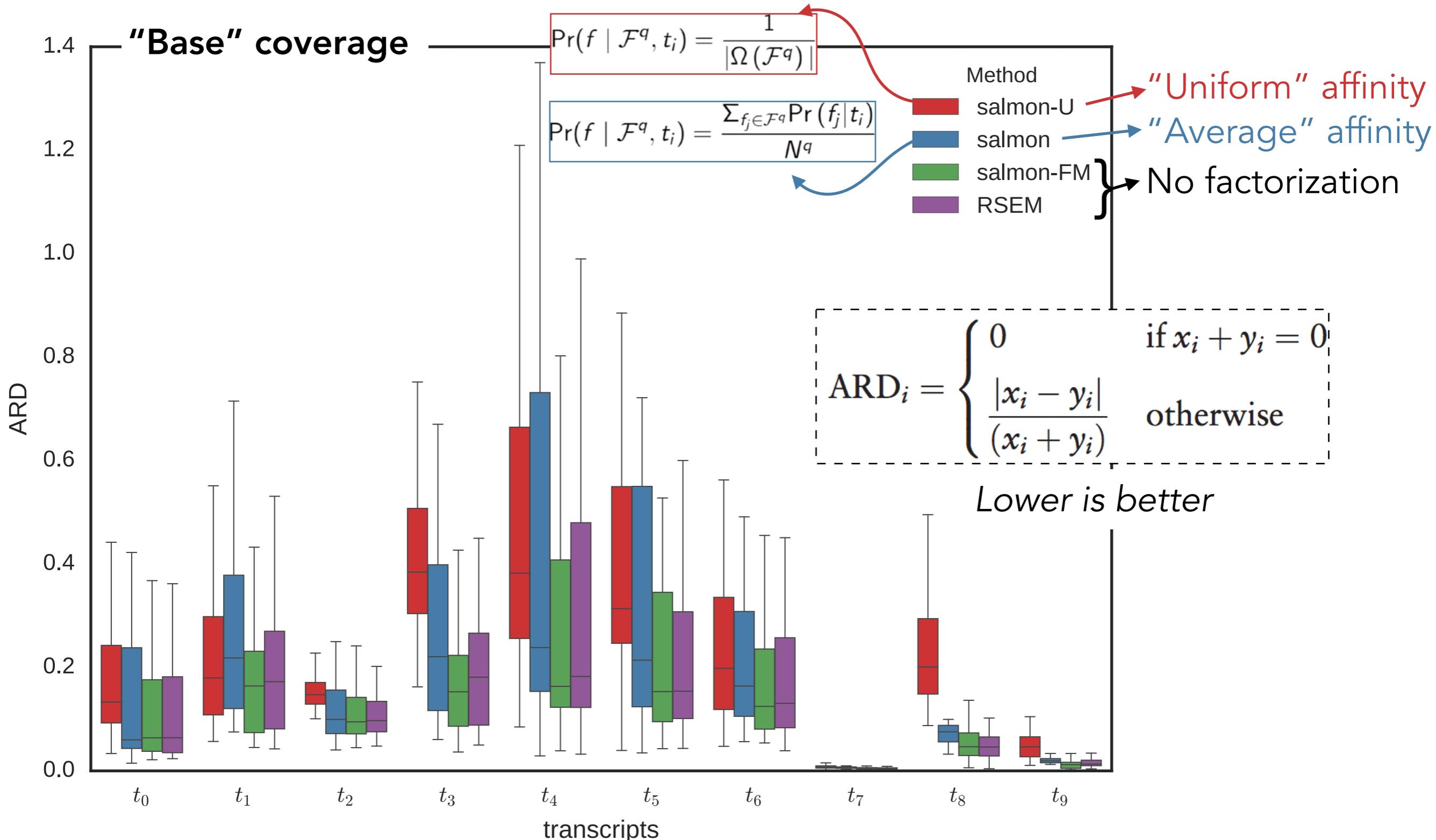
Conditional probabilities can provide valuable information about origin of a fragment! **Potentially different for each transcript/fragment pair.**

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” model¹. e.g. position, orientation, alignment path etc.

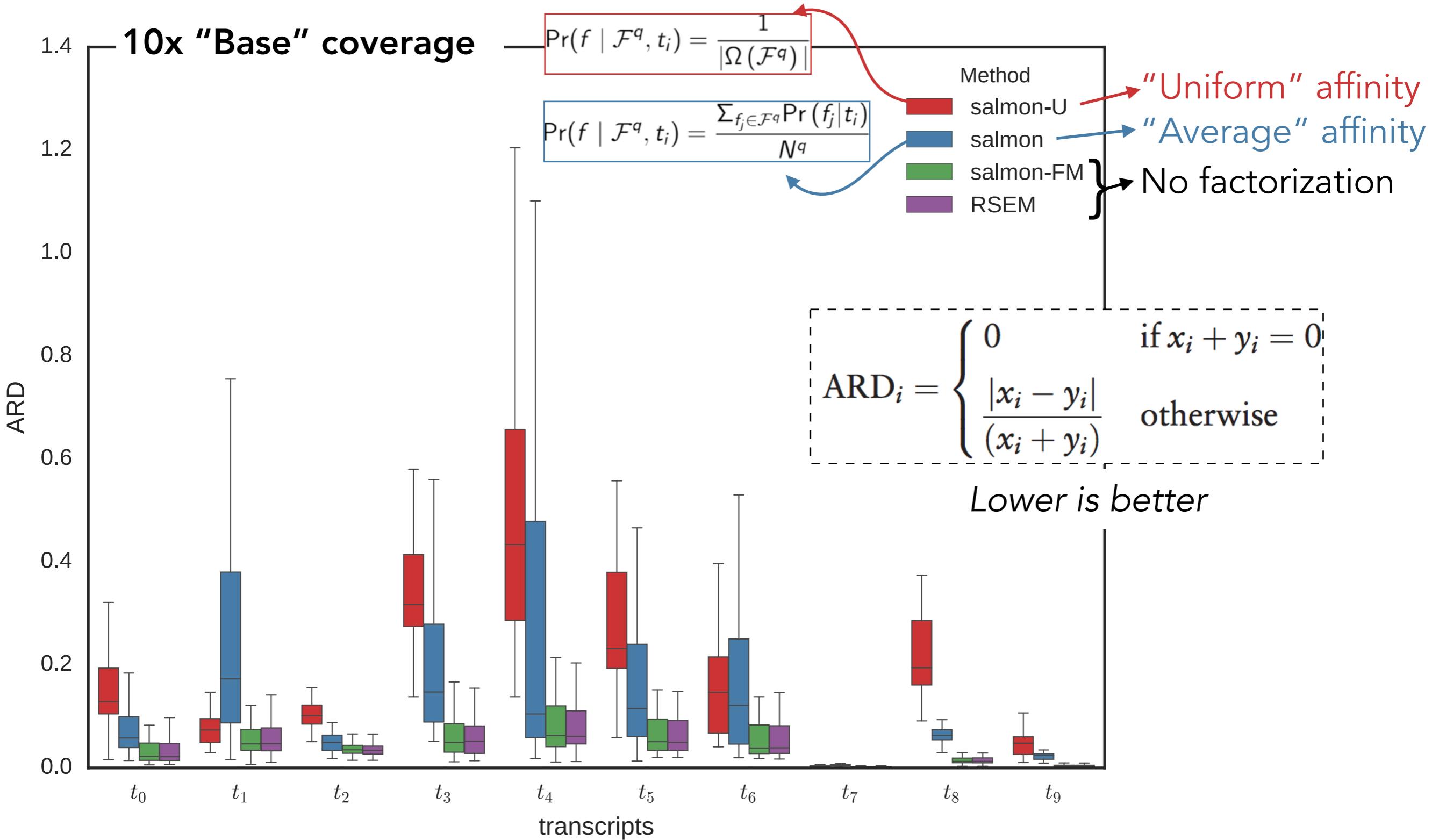
¹ “Salmon provides fast and bias-aware quantification of transcript expression”, Nature Methods 2017

Does this term matter?



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

Does this term matter?



- 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:

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



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.

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

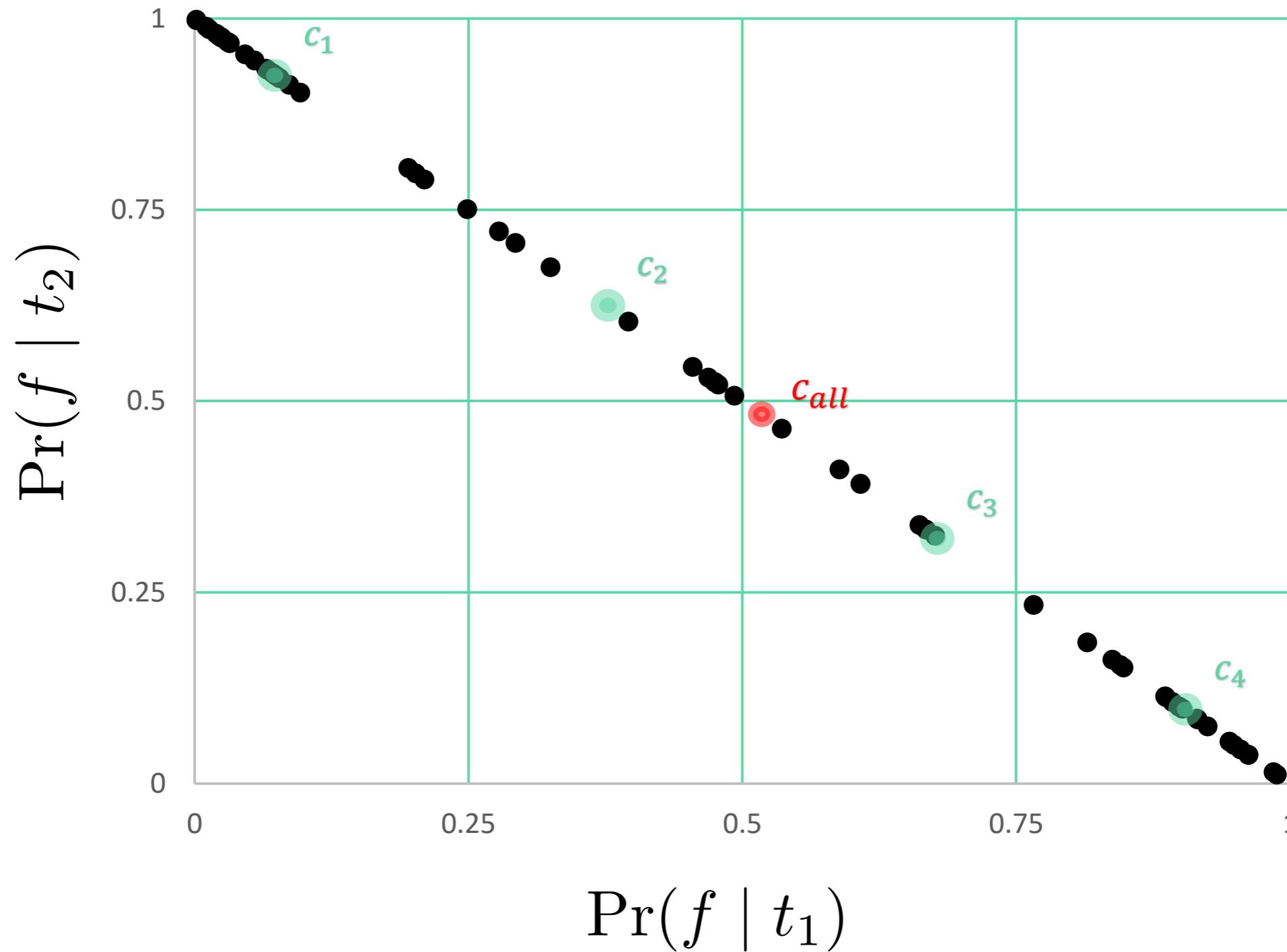
$$f \sim_r f' \iff \Omega(f) = \Omega(f') \wedge b_k(f, \Omega(f)) = b_k(f', \Omega(f'))$$

Maps to the same set of transcripts

Has the same binned cond. prob vector

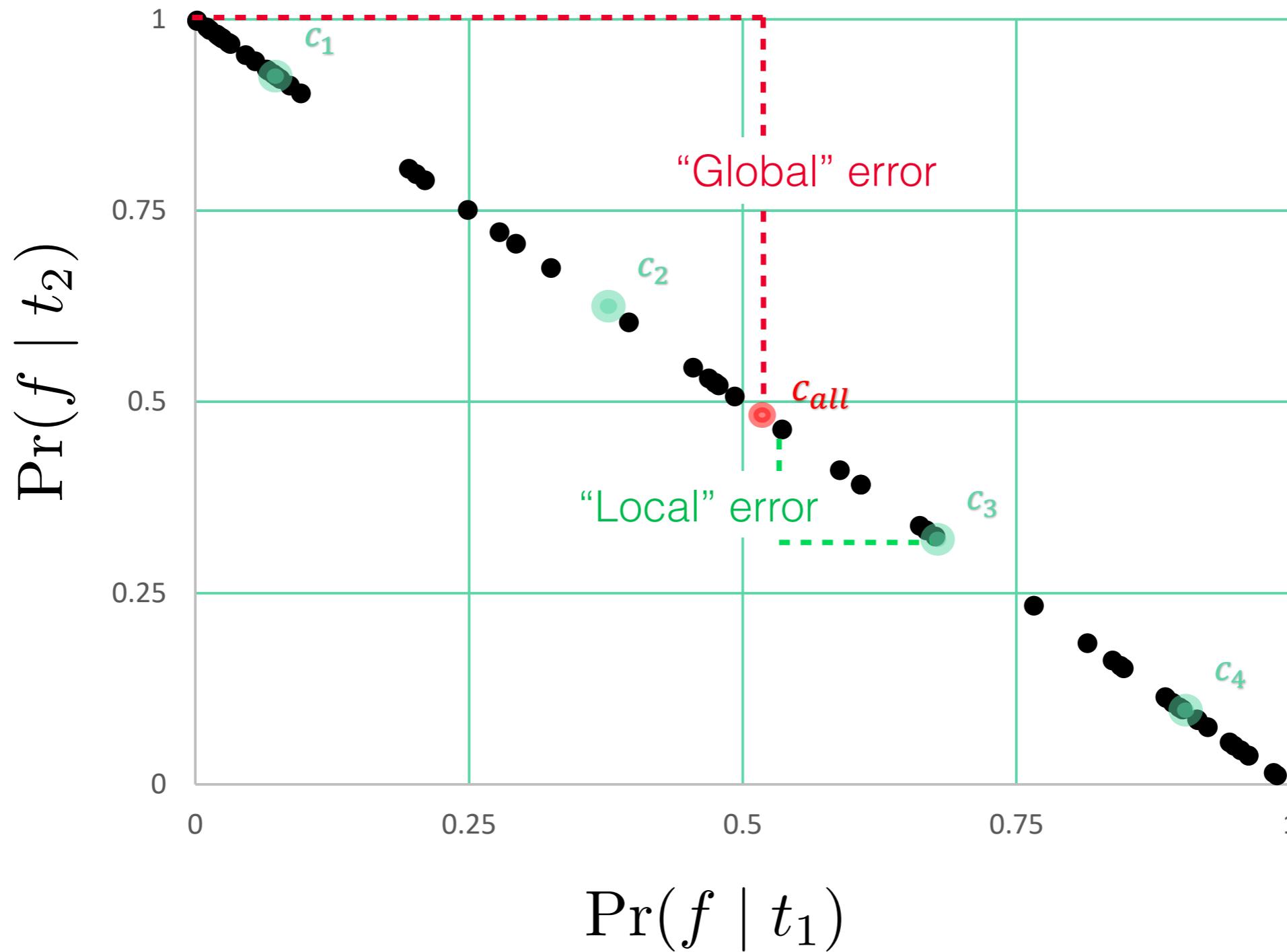
Range-based factorization

60 fragments in equivalence class $\{t_1, t_2\}$



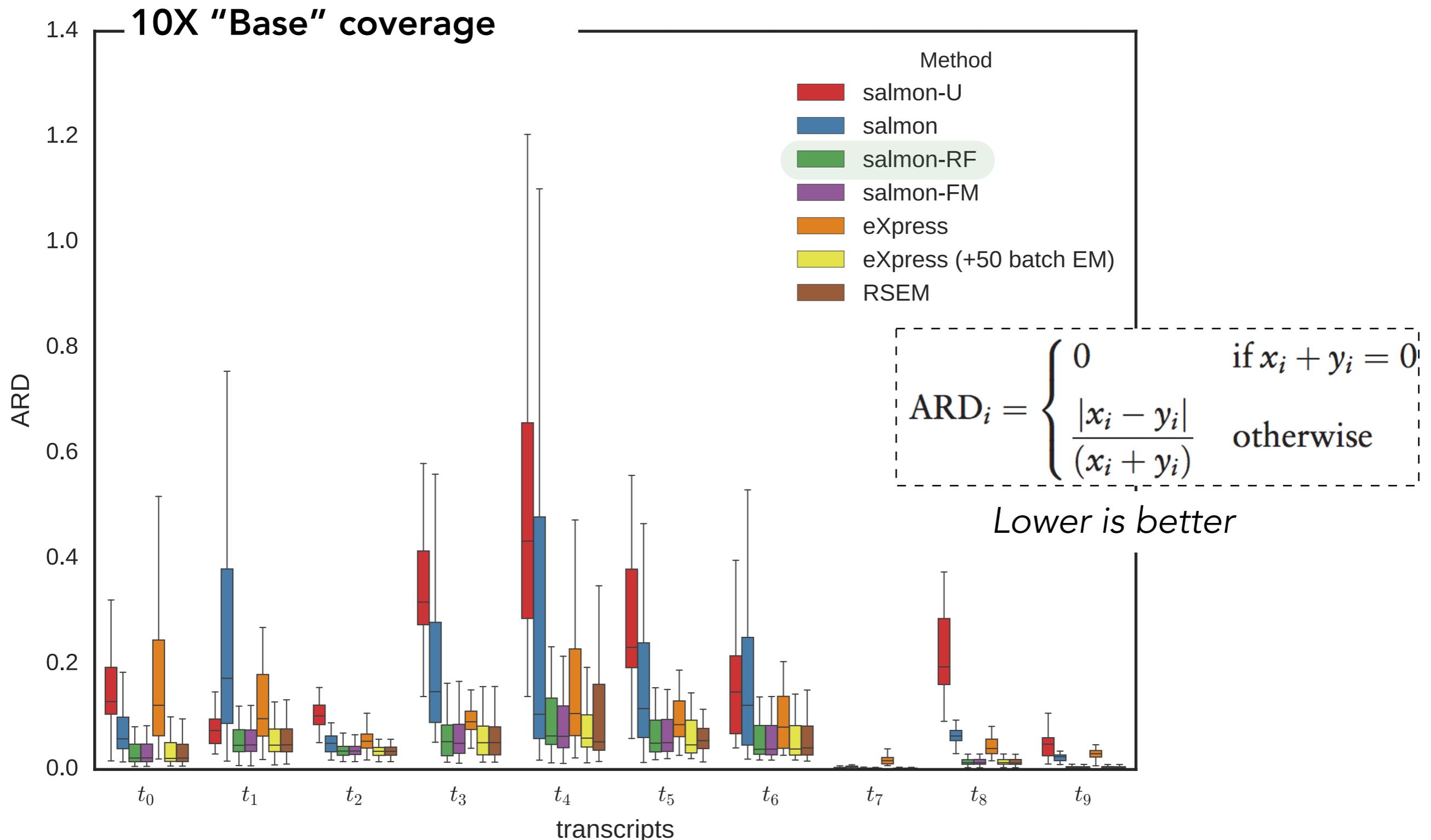
Range-based factorization improves approximation

60 fragments in equivalence class $\{t_1, t_2\}$



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

How well does this work?



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

Transcriptome-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
<i>Salmon-U</i>	0.24	0.80
<i>Salmon</i>	0.22	0.81
<i>Salmon-RF</i>	0.21	0.83
<i>Salmon-FM</i>	0.21	0.83
<i>eXpress</i>	0.29	0.78
<i>eXpress (+50)</i>	0.23	0.83
<i>RSEM</i>	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.

Transcriptome-wide assessment can mask important differences

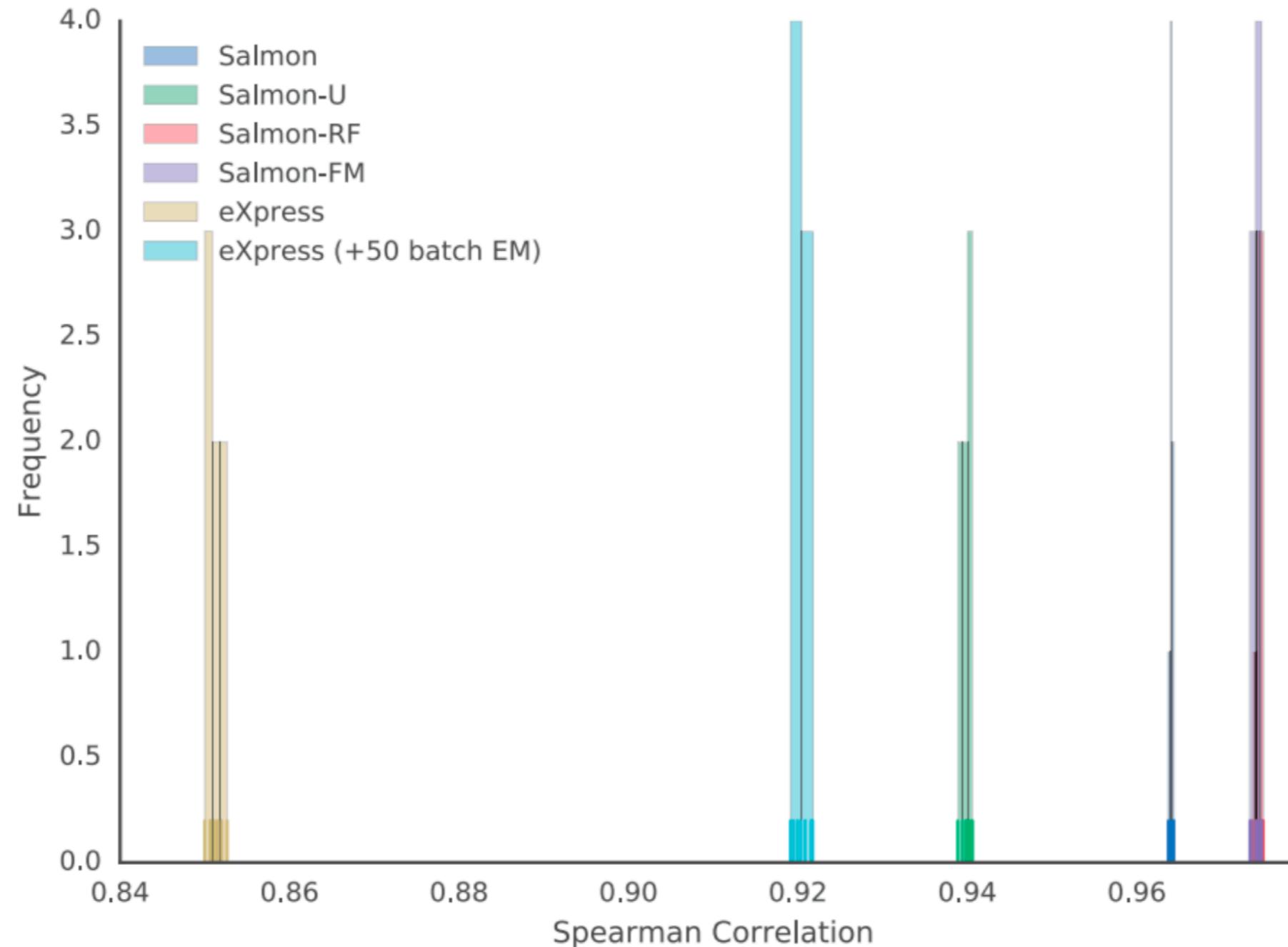
- Focus on a subset of “critical” transcripts (not too easy, not intractable)
- Transcripts where RSEM yields an ARD in [0.25,0.75]

Method	MARD	Spearman
<i>Salmon-U</i>	0.46	0.56
<i>Salmon</i>	0.43	0.58
<i>Salmon-RF</i>	0.41	0.64
<i>Salmon-FM</i>	0.41	0.65
<i>eXpress</i>	0.53	0.54
<i>eXpress (+50)</i>	0.48	0.59
<i>RSEM</i>	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

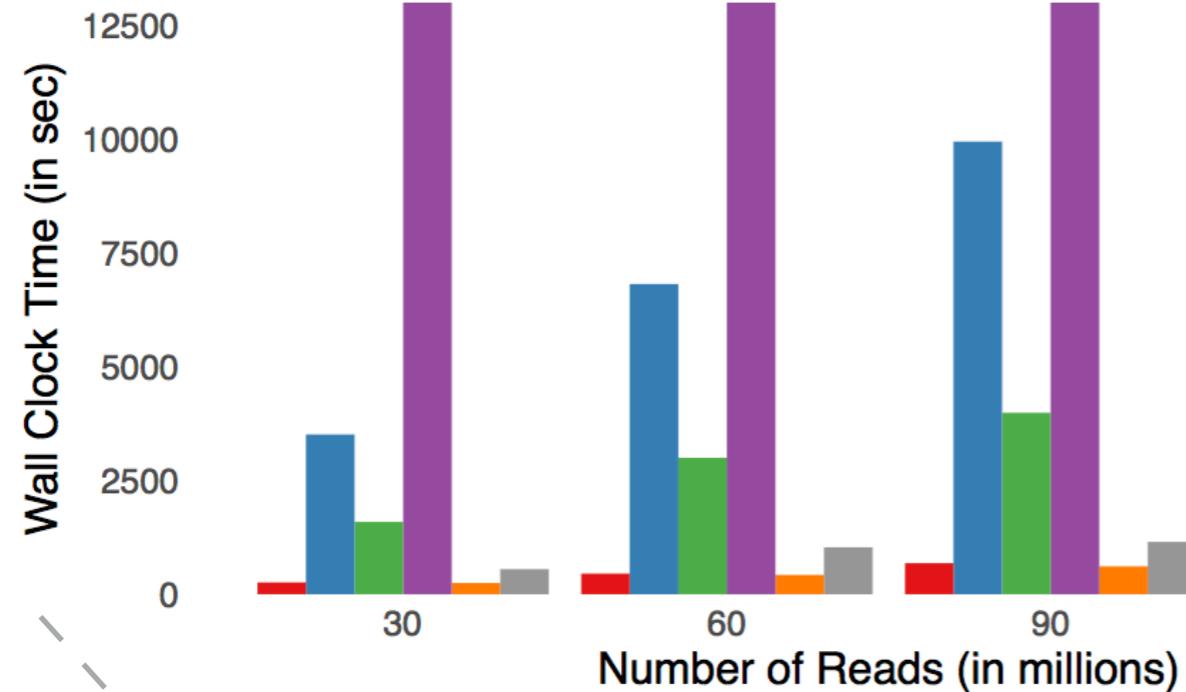
	<i>Salmon-U</i>	<i>Salmon</i>	<i>Salmon-RF</i>	<i>Salmon-FM</i>
# eq. classes	438,393	438,393	625,638	29,447,710
# hits	5,986,371	5,986,371	8,212,669	103,663,423

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

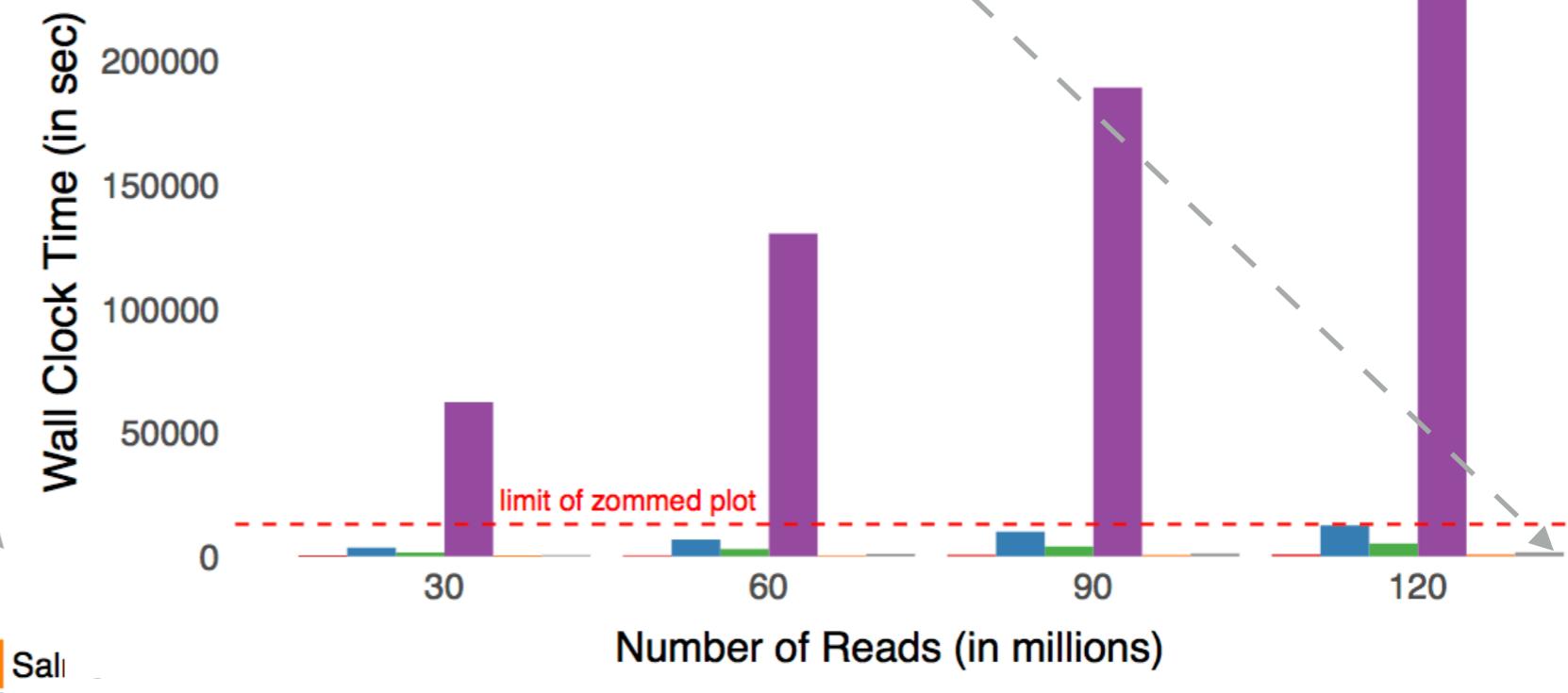
hits : The number of hits is the sum, over each equivalence class, of the number of transcripts in this equivalence class — i.e. $\sum_{\mathcal{F}^q \in \mathcal{C}} |\Omega(\mathcal{F}^q)|$

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

Range-factorization is still very (computationally) efficient



Zooming out



Method

Salmon RF	eXpress	Sal
Red	Green	Orange
RSEM	eXpress (+50 batch EM)	Salmon FM

Range-factorization controls memory requirements

