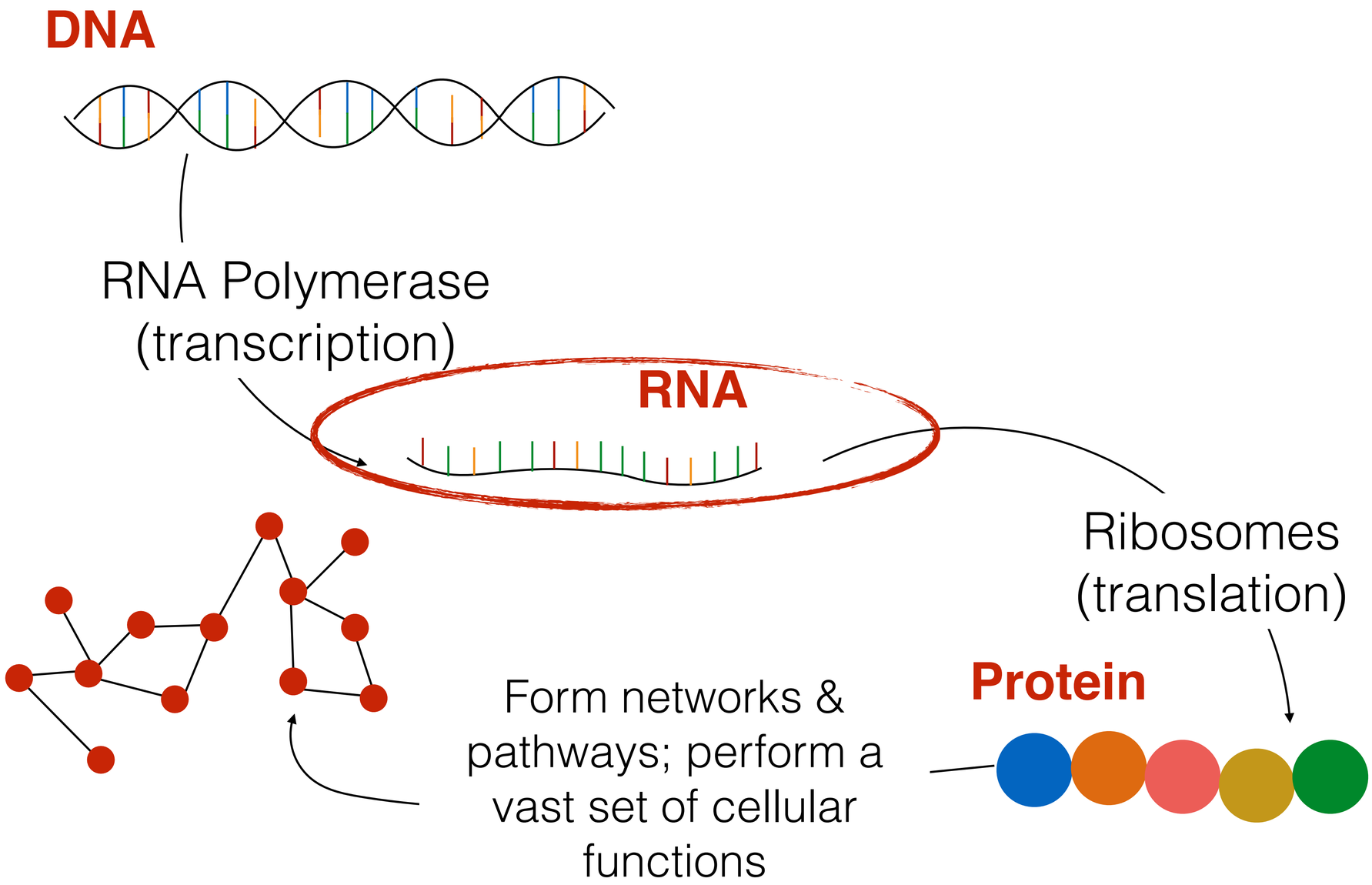
Analyzing gene and transcript expression using RNA-seq

UNIVERSITY OF MARYLAND

“Flow” of information in the cell



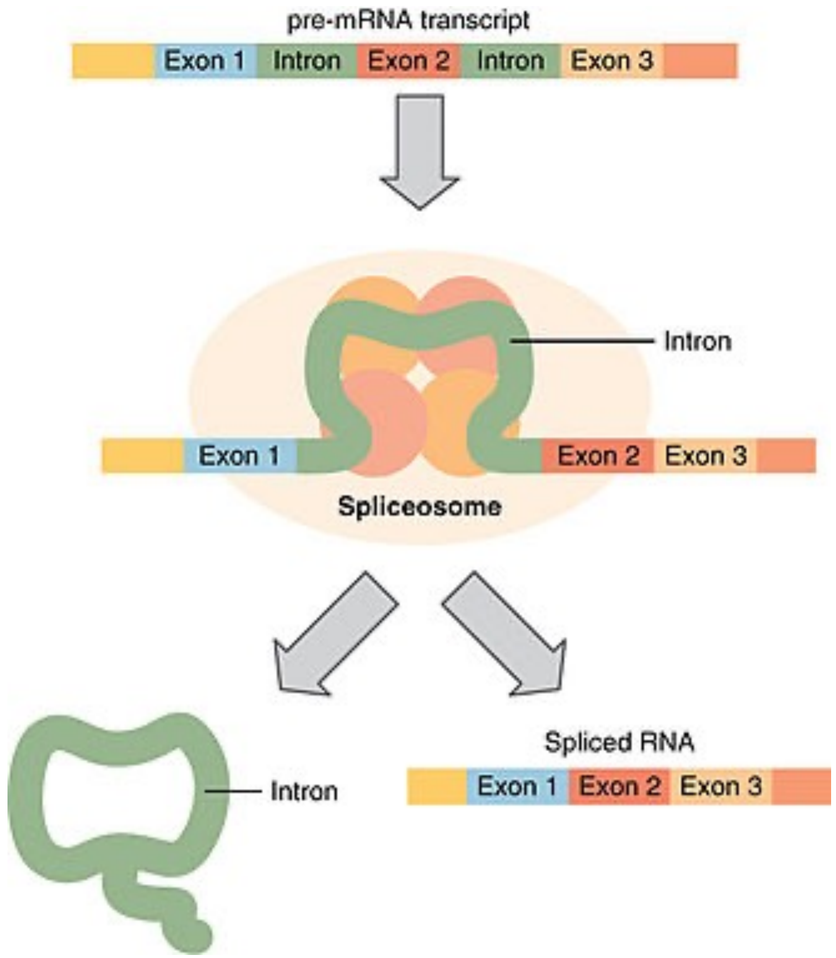
RNA Splicing

DNA transcribed into pre-mRNA

Some “processing occurs”
capping & **polyadenylation**

Introns removed from pre-mRNA

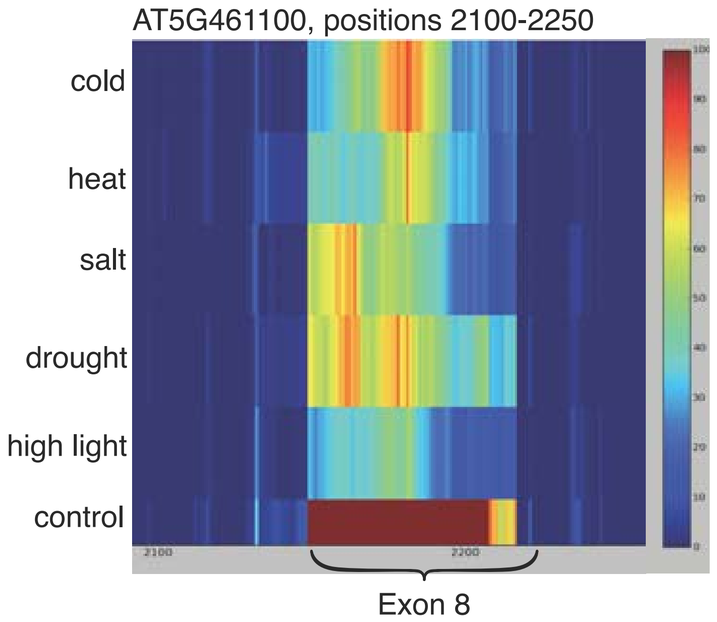
Introns removed resulting in
mature mRNA

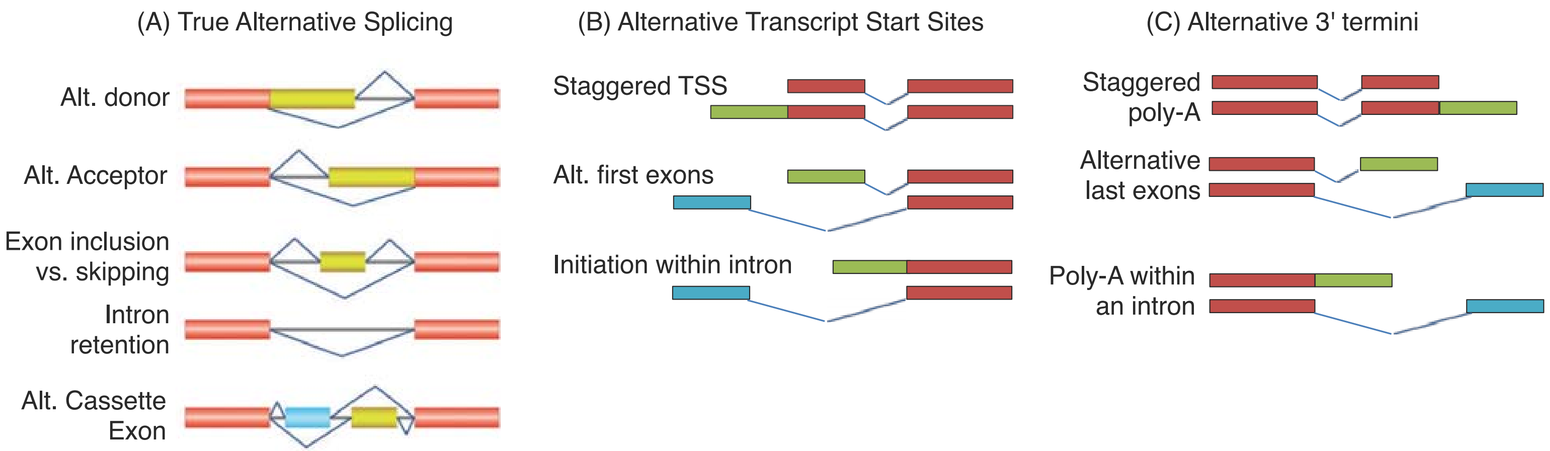


en.wikipedia.org

Alternative Splicing & Isoform Expression

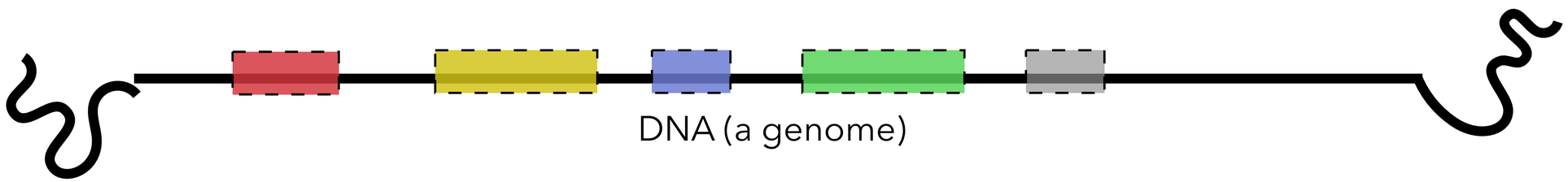
- Expression of genes can be measured via RNA-seq (sequencing transcripts)
- Sequencing gives you short (35-300bp length reads)





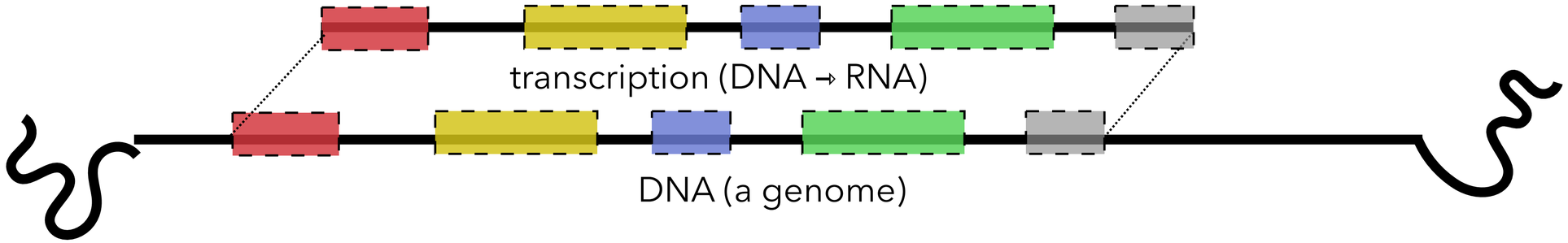
slide courtesy of Carl Kingsford

What is RNA sequencing



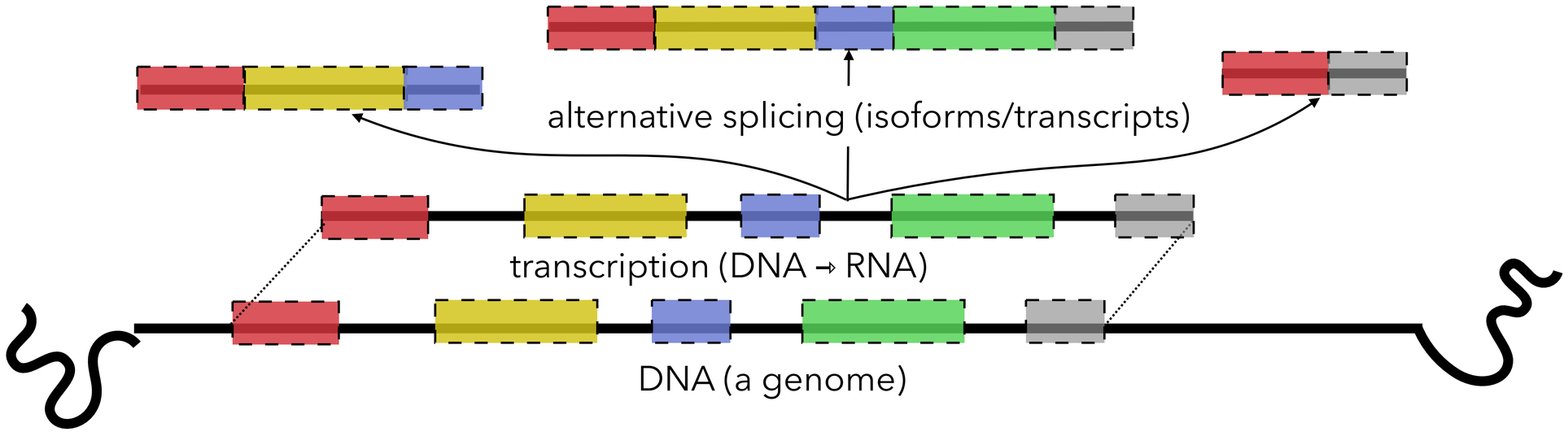
* most protocols actually sequence complementary DNA (cDNA), not RNA directly

What is RNA sequencing



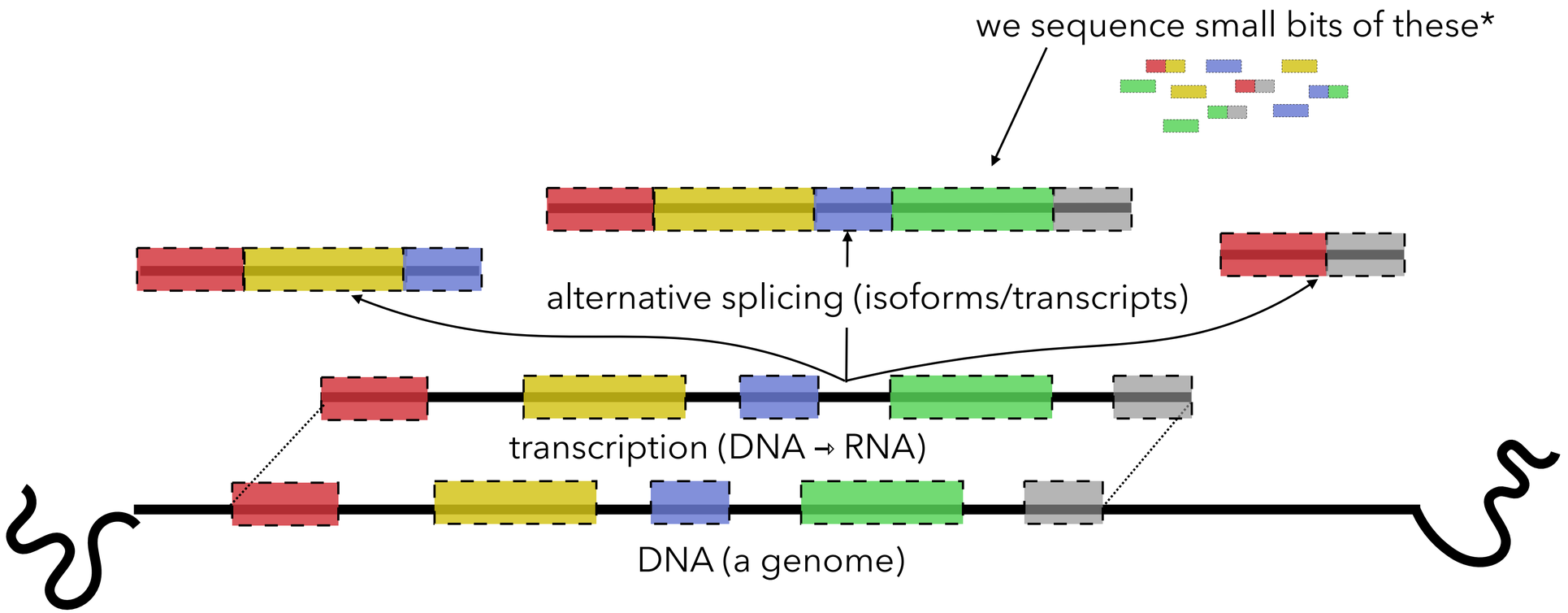
* most protocols actually sequence complementary DNA (cDNA), not RNA directly

What is RNA sequencing



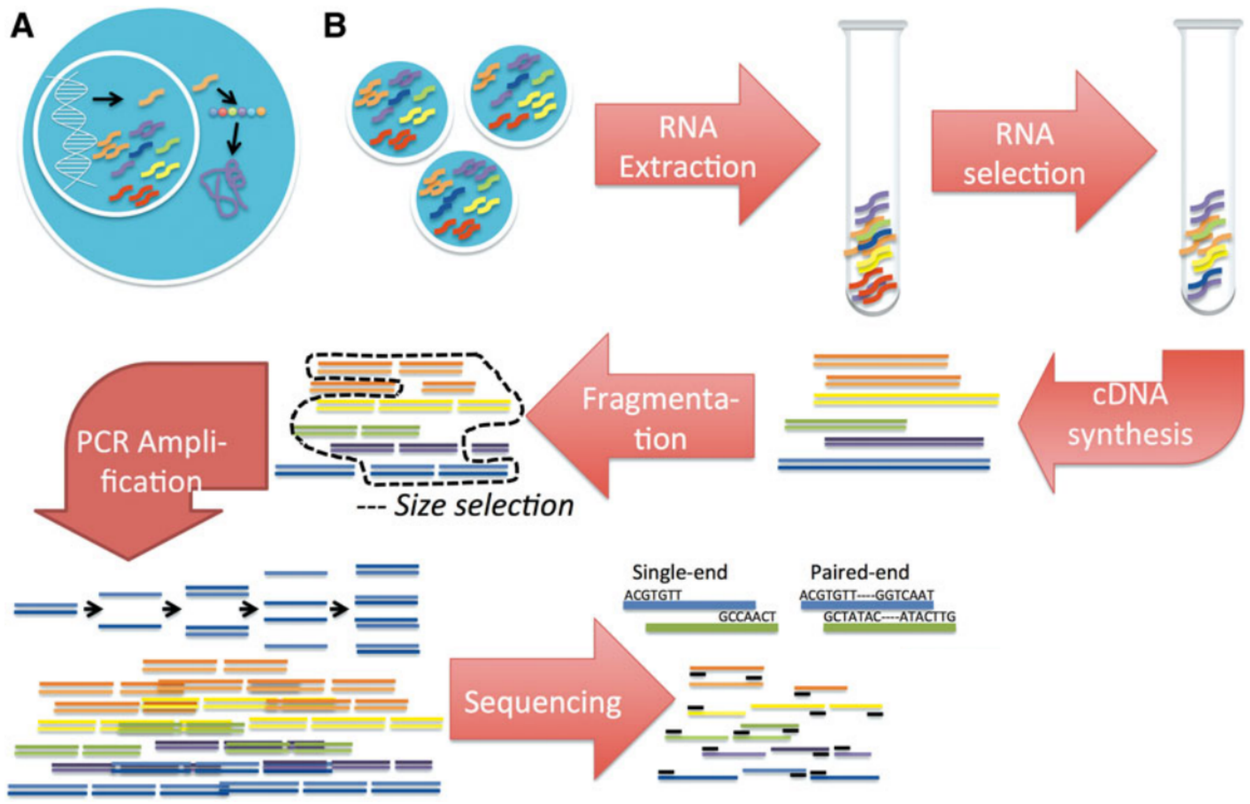
* most protocols actually sequence complementary DNA (cDNA), not RNA directly

What is RNA sequencing



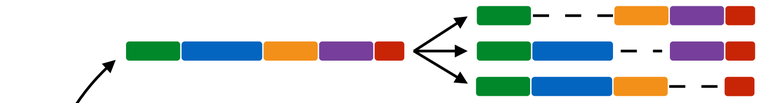
* most protocols actually sequence complementary DNA (cDNA), not RNA directly

Actual protocols are much more involved



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 a
of Computational Biology 24.3 (2017): 200-212.

Transcript Quantification: An Overview



1 gene ⇒ many variants (isoforms)

Gene 1

Gene M

Sample

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

Measurement
(RNA-seq)

Inference
(e.g. Salmon)

% Gene 1

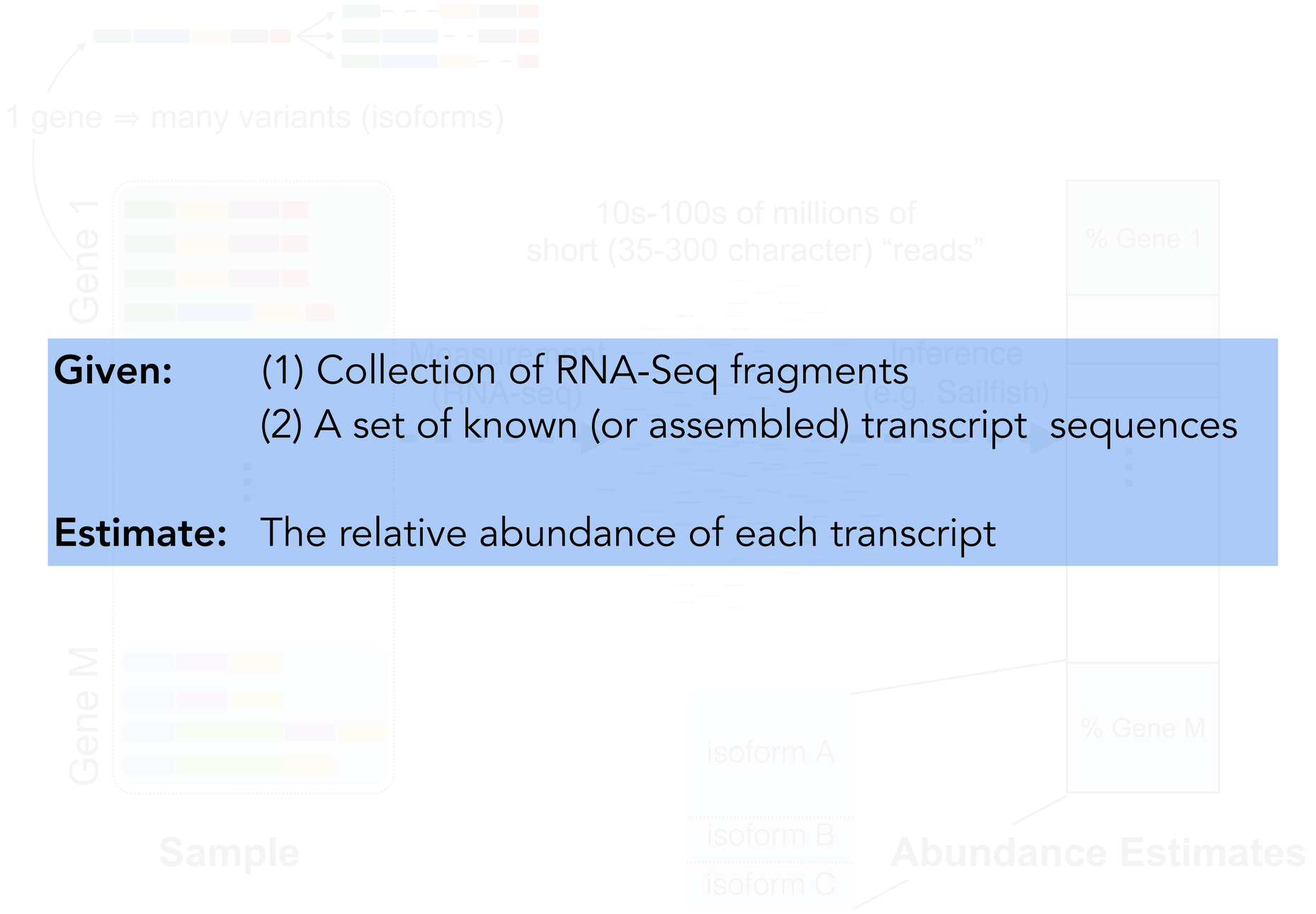
% Gene M

isoform A

isoform B

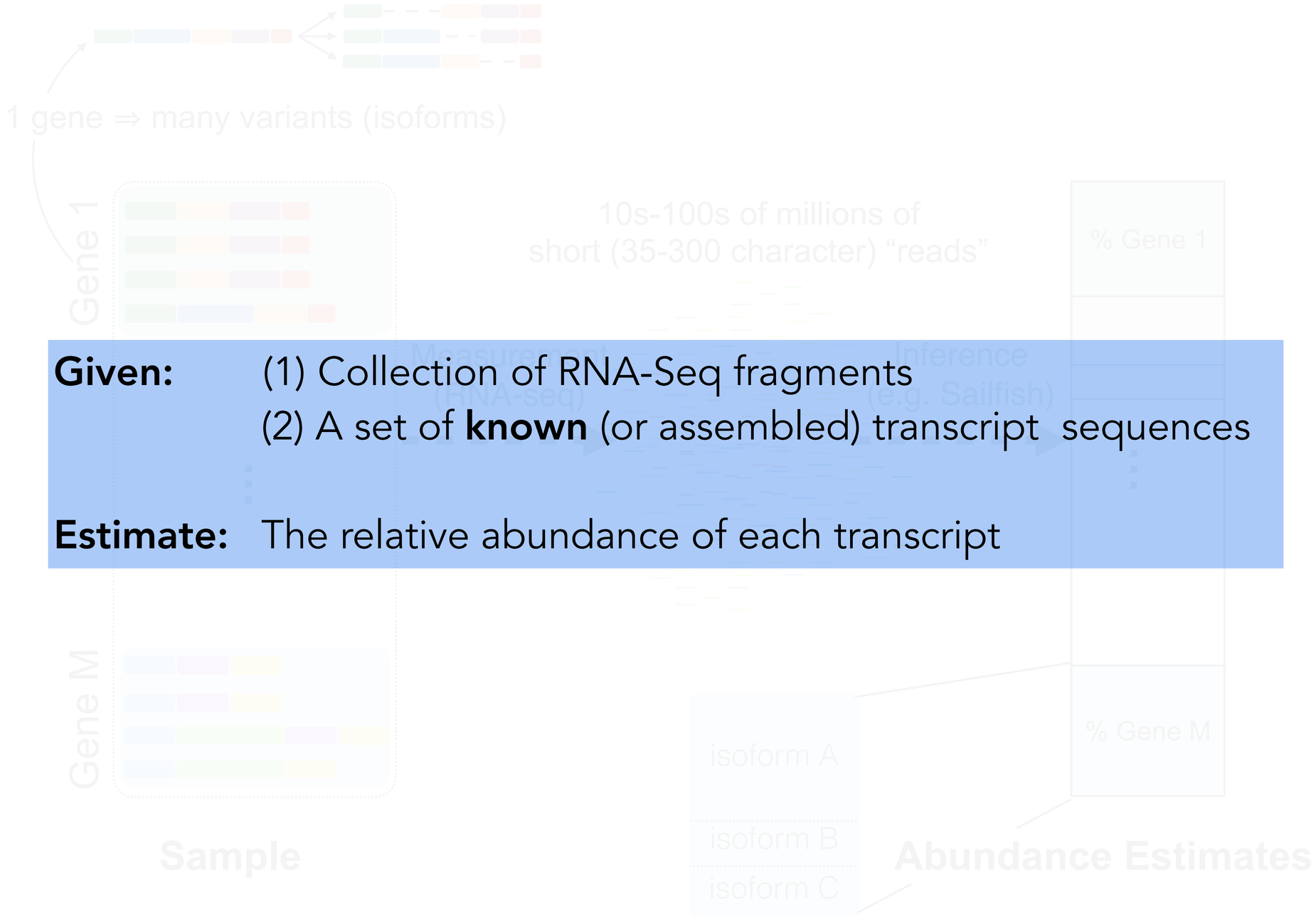
isoform C

Abundance Estimates



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

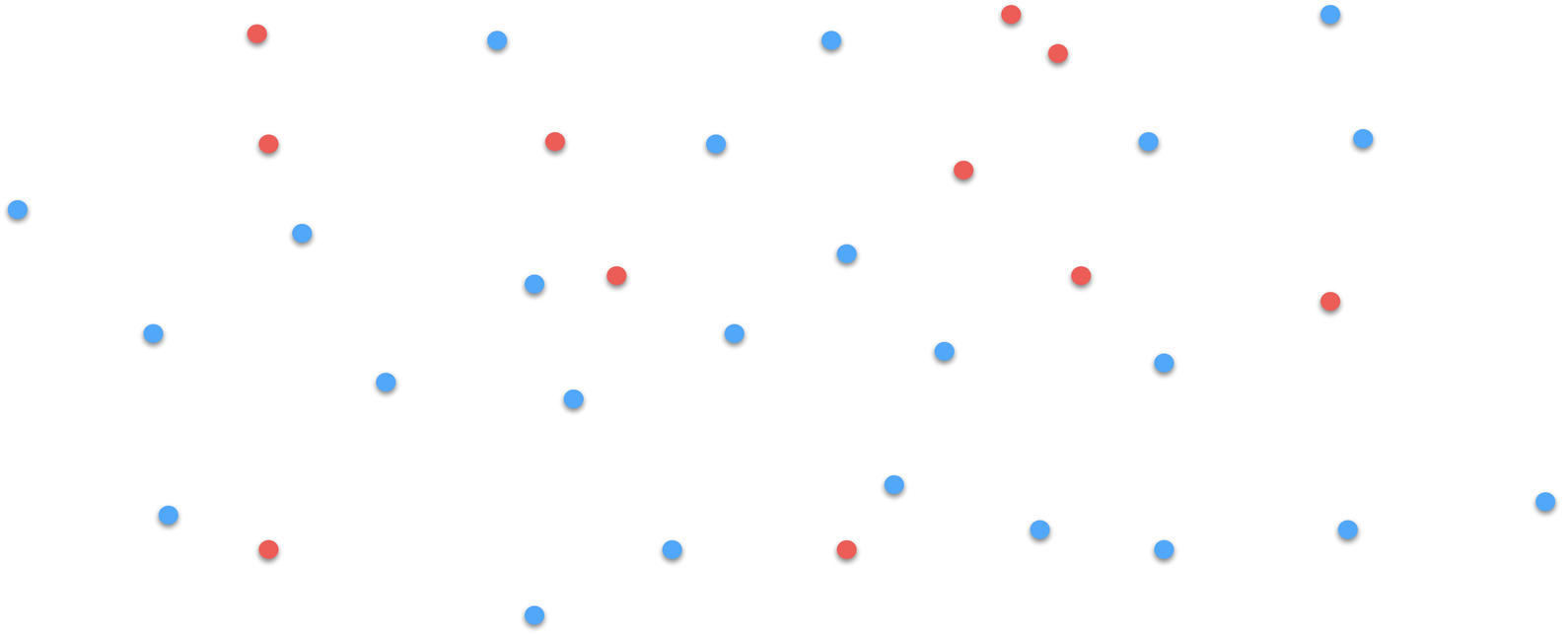
The RNA-seq reads are drawn from transcripts, and our (spliced) aligners let us map them back to the transcripts on the genome from which they originate.

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.



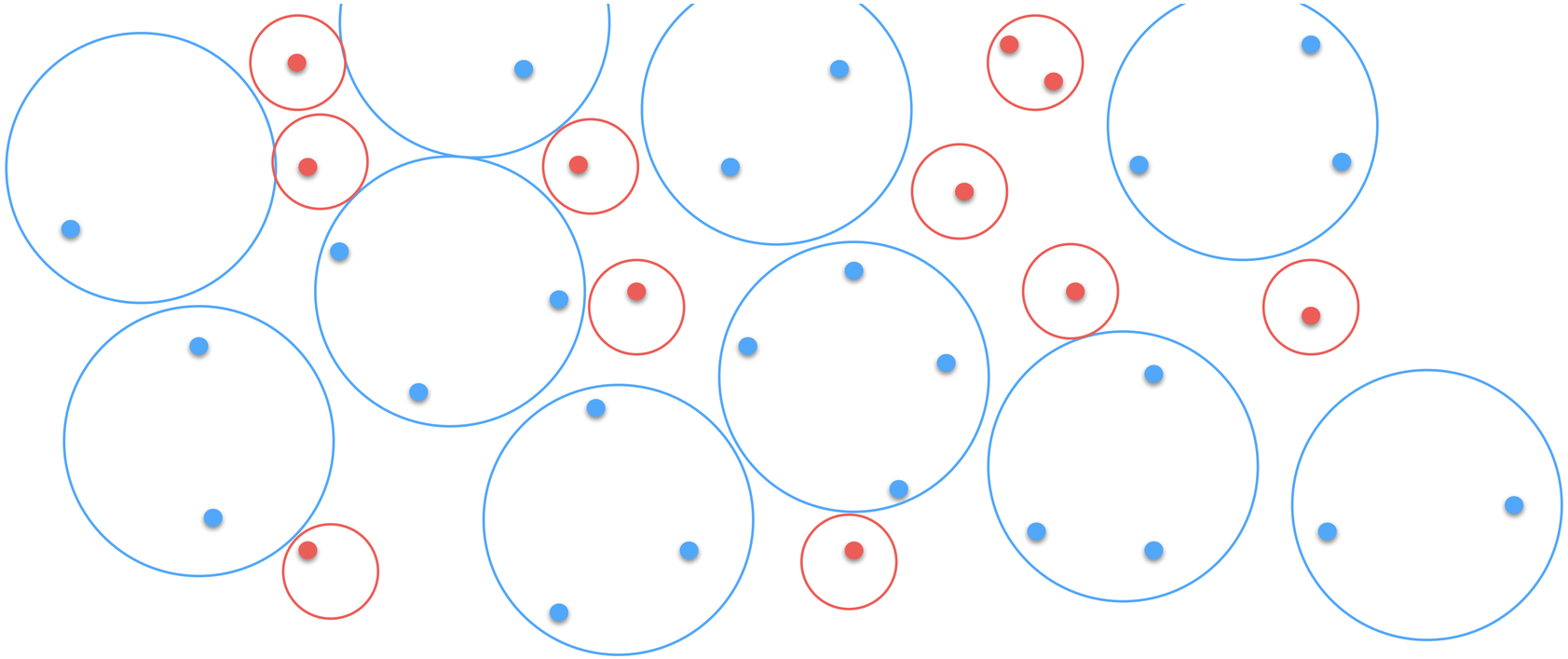
Here, a dot of a color means I hit a circle of that color.

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.

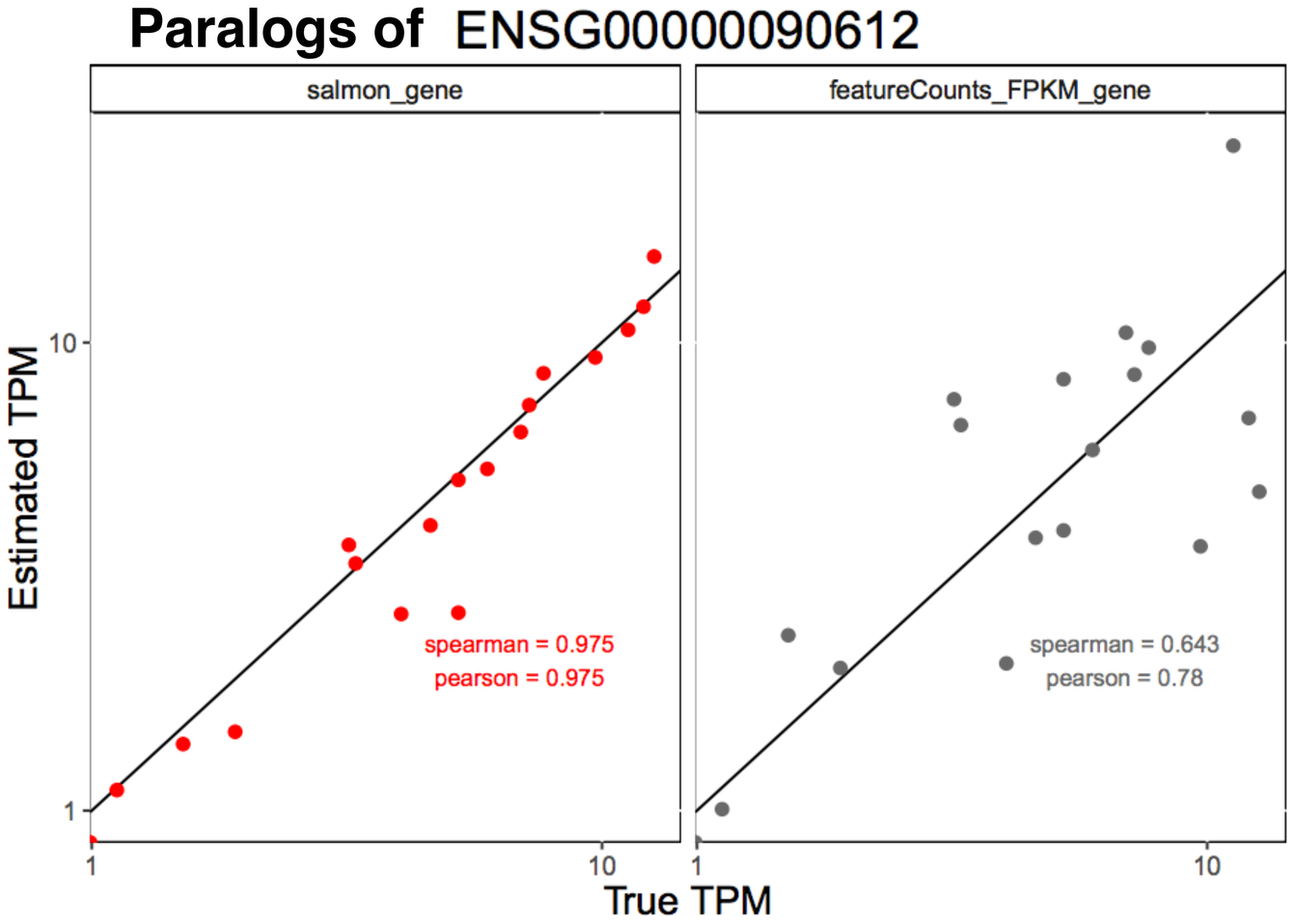
You're missing a **crucial piece of information!**

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

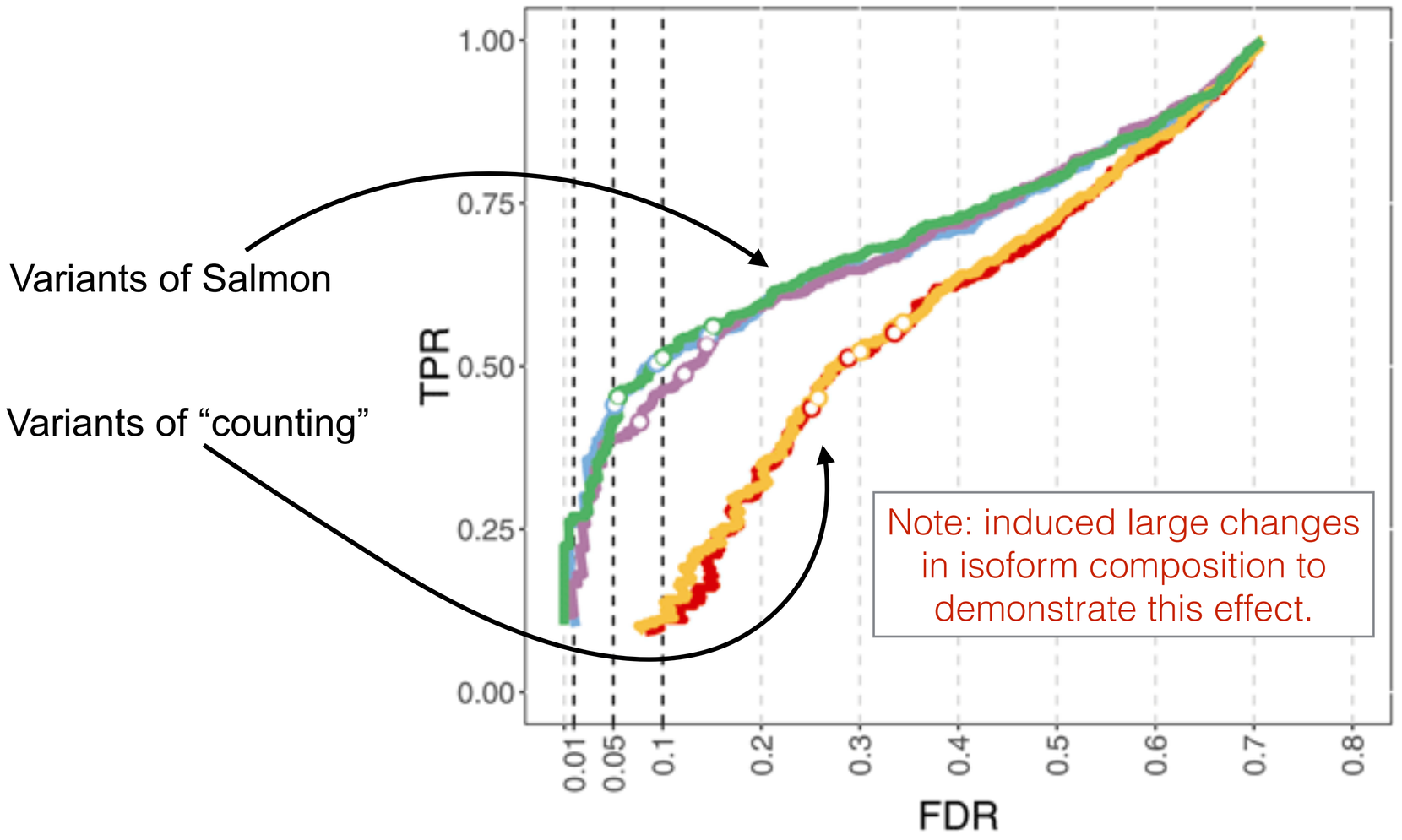
Can even affect abundance estimation in **absence** of alternative-splicing
(e.g. paralogous genes)



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

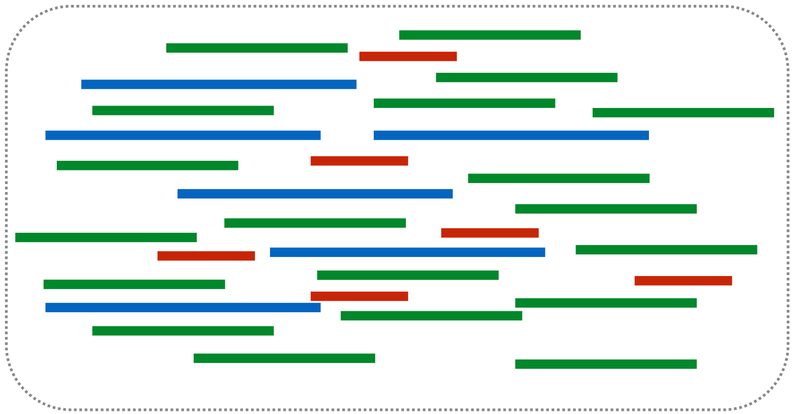
These errors can affect DGE calls



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

How can we perform inference from sequenced fragments?

Experimental Mixture

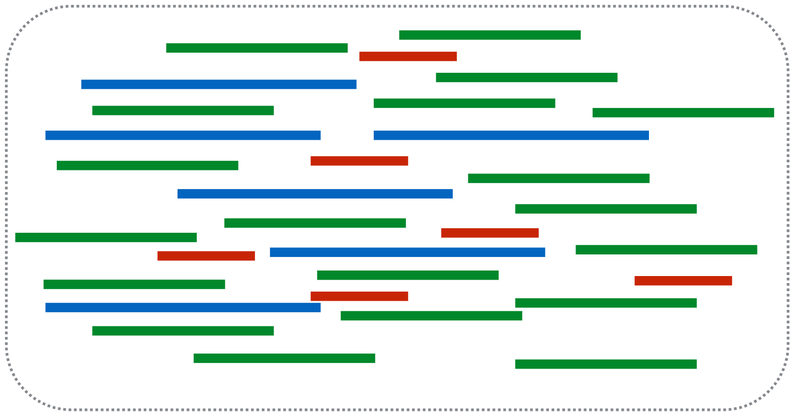


In an unbiased experiment, sampling fragments depends on:

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

How can we perform inference from sequenced fragments?

Experimental Mixture



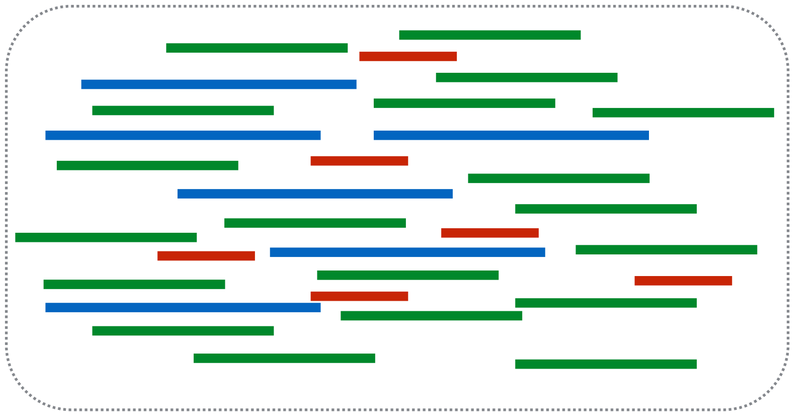
length( ) = 100

In an unbiased experiment, sampling fragments depends on:

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

How can we perform inference from sequenced fragments?

Experimental Mixture



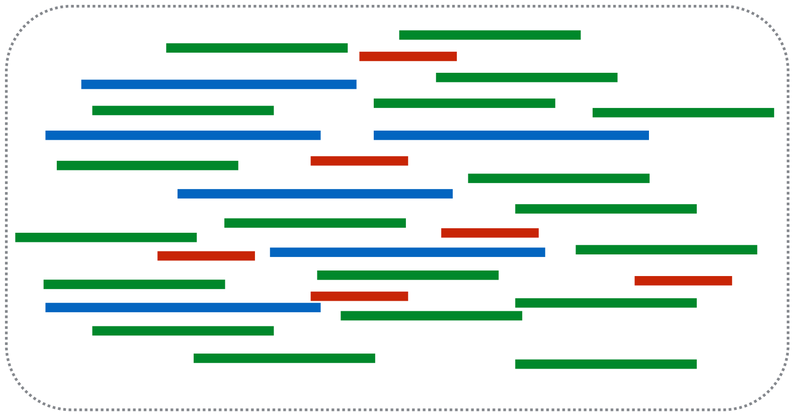
In an unbiased experiment, sampling fragments depends on:

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

length( ) = 100 x 6 copies

How can we perform inference from sequenced fragments?

Experimental Mixture



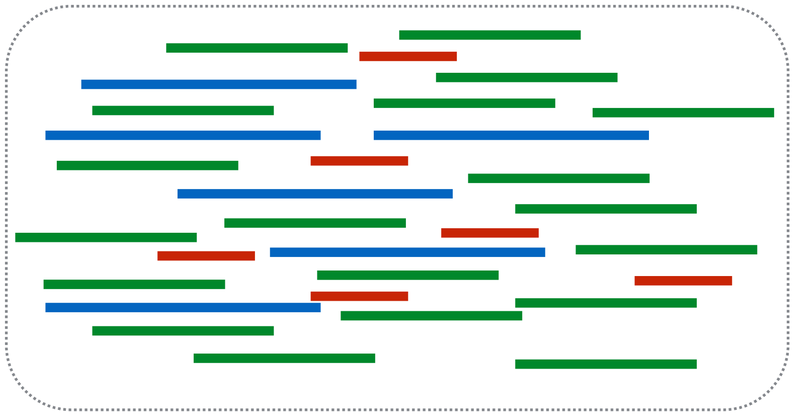
In an unbiased experiment, sampling fragments depends on:

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

length( ) = 100 x 6 copies = 600 nt

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

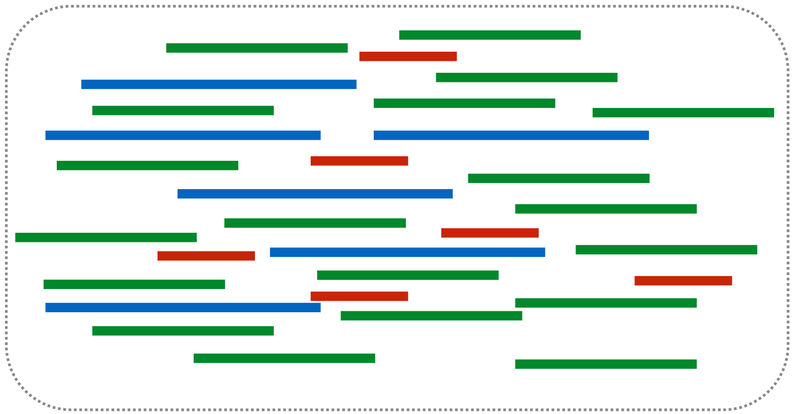
length( ) = 100 x 6 copies = 600 nt

length( ) = 66 x 19 copies = 1254 nt

length( ) = 33 x 6 copies = 198 nt

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

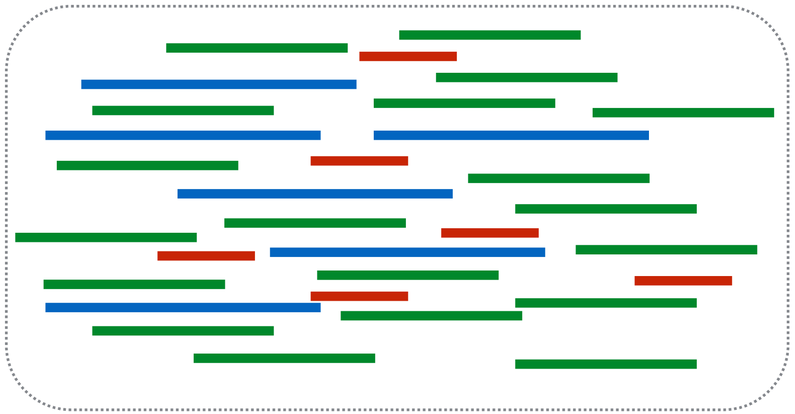
length( ) = 100 x 6 copies = 600 nt ~ 30% blue

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

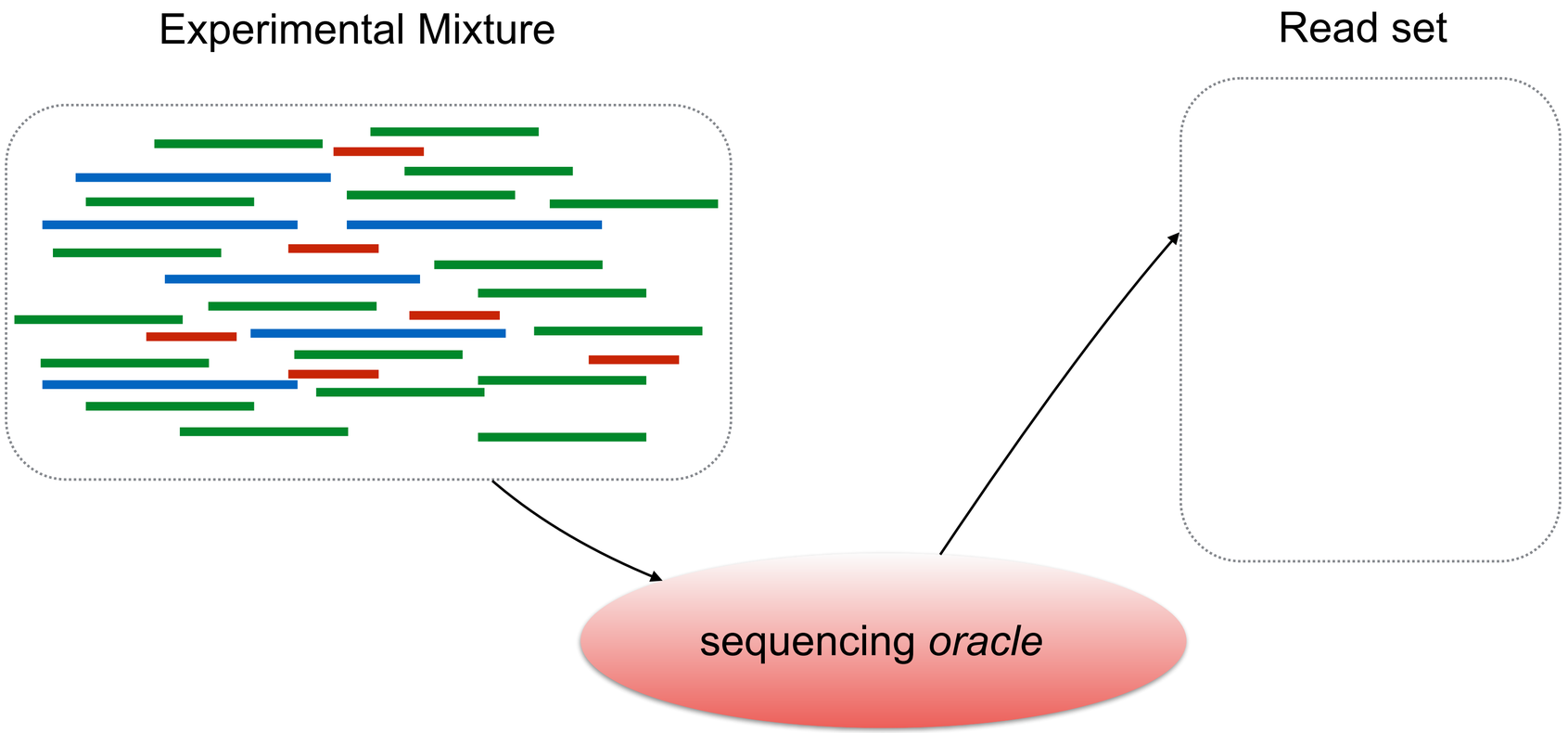
length( ) = 66 x 19 copies = 1254 nt ~ 60% green

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 . . .

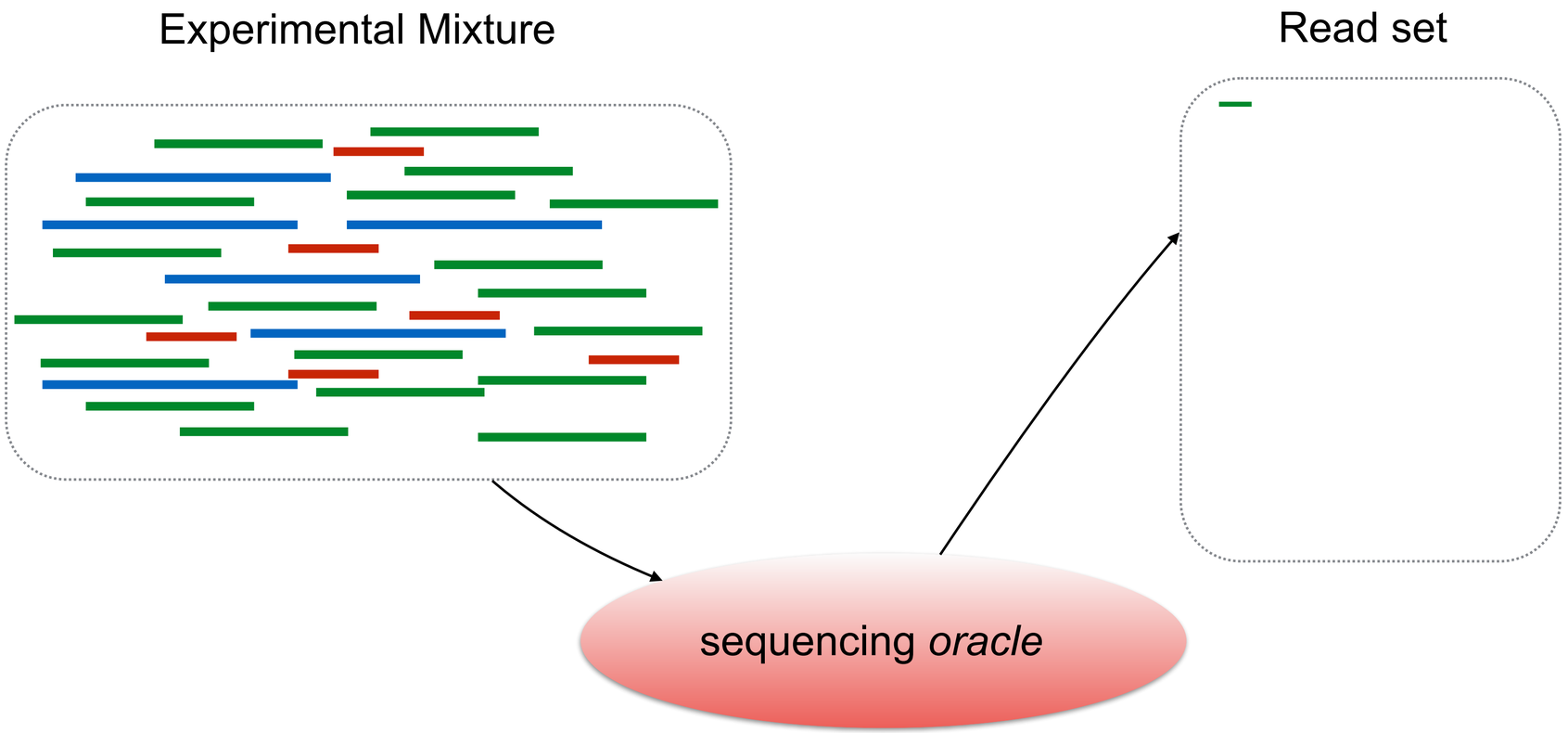


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

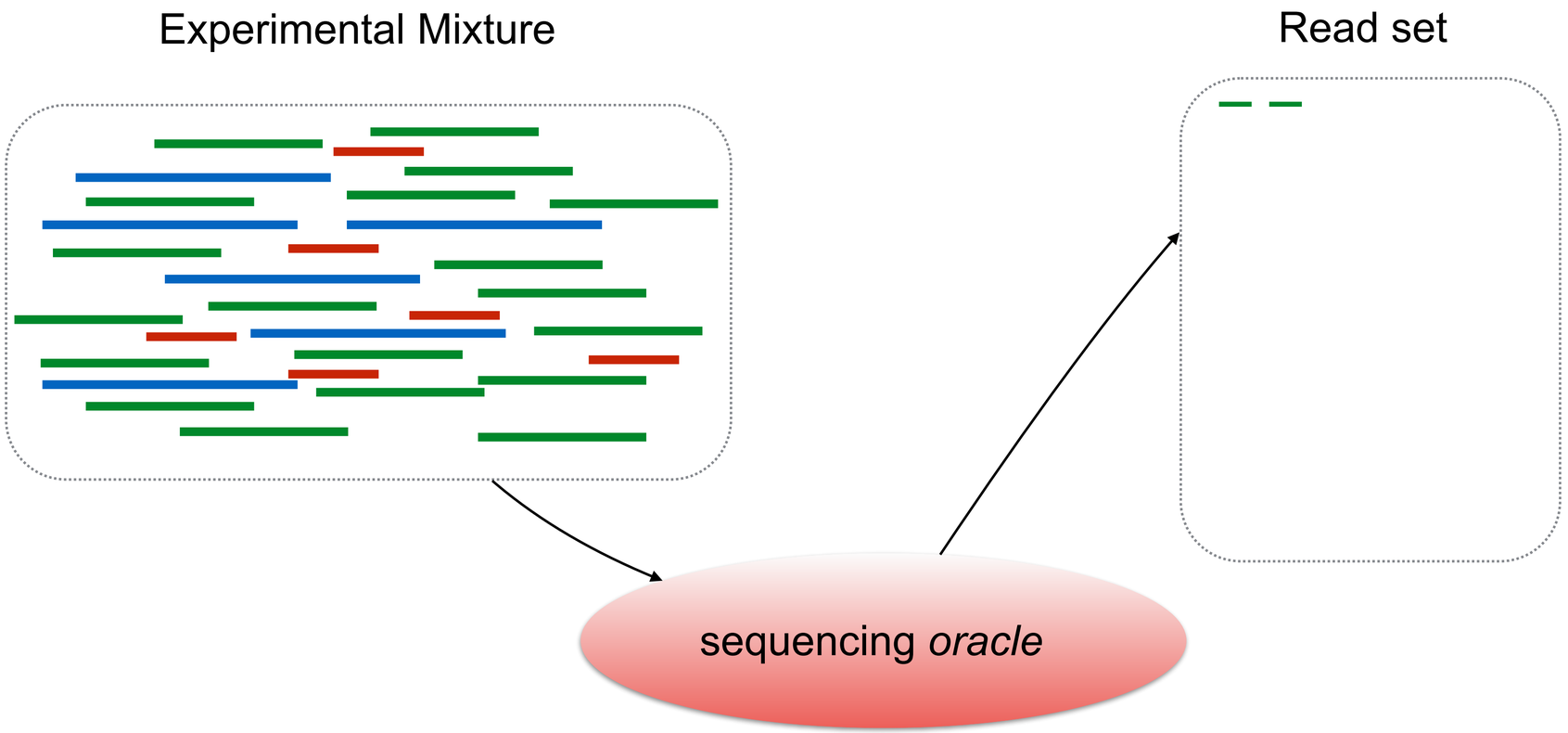


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

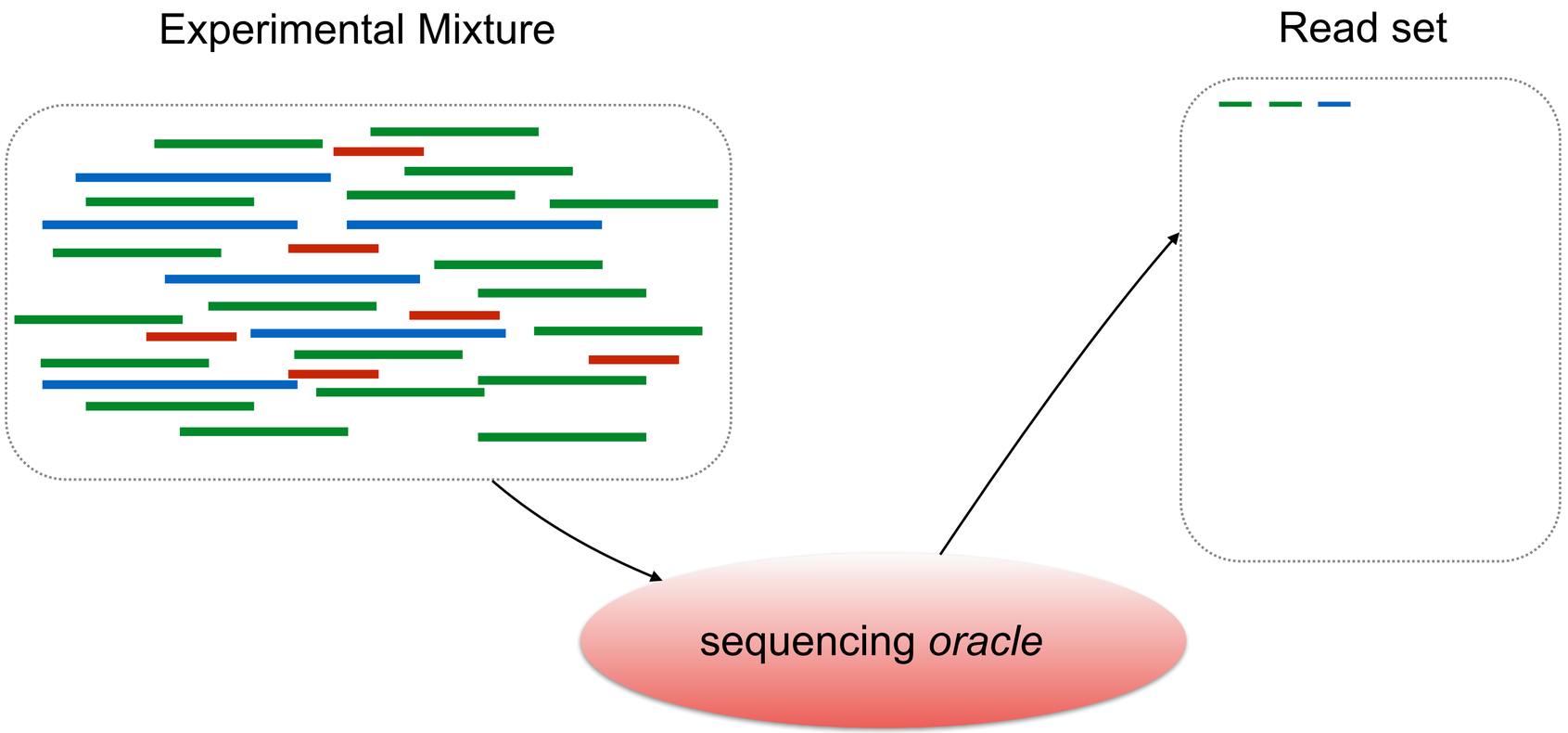


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

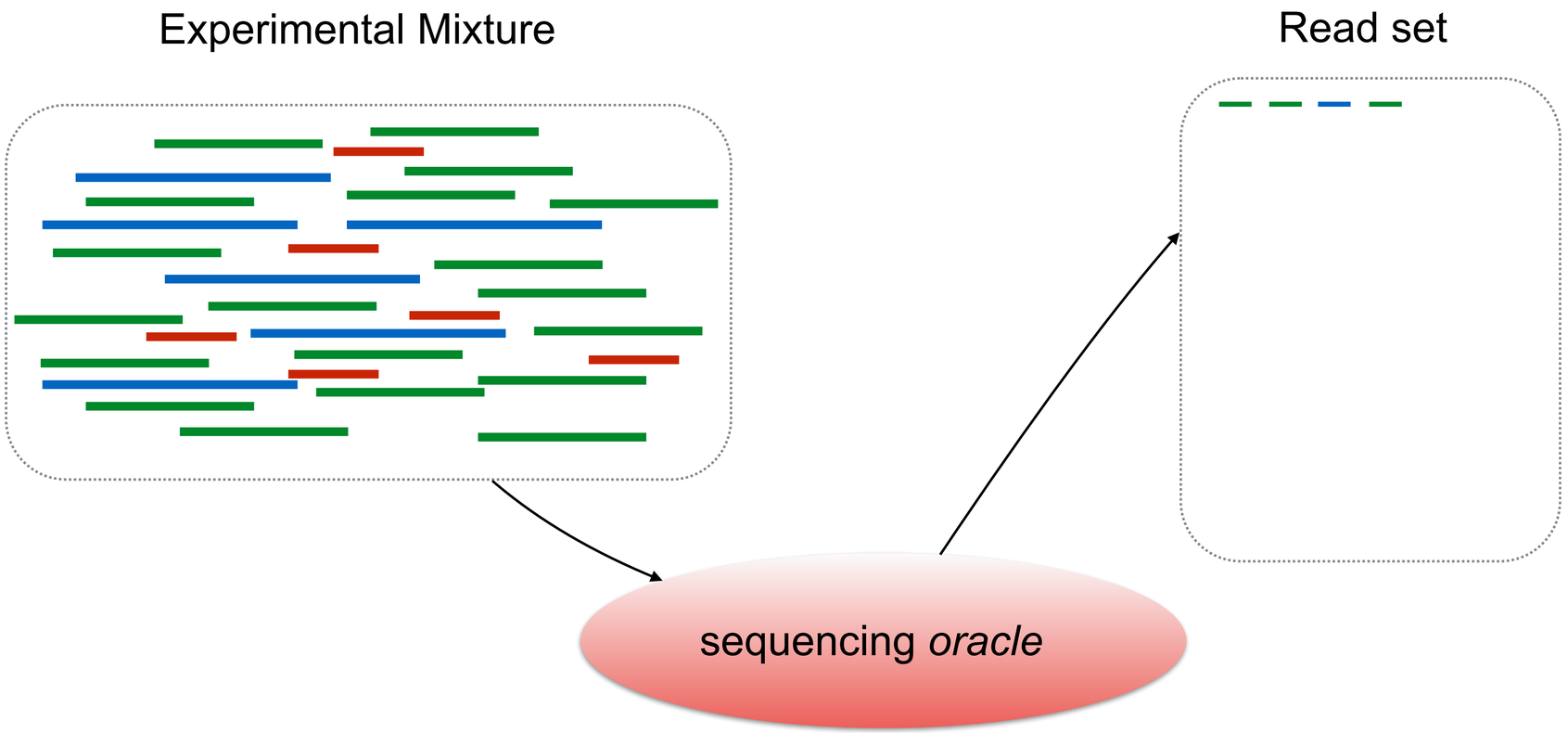


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

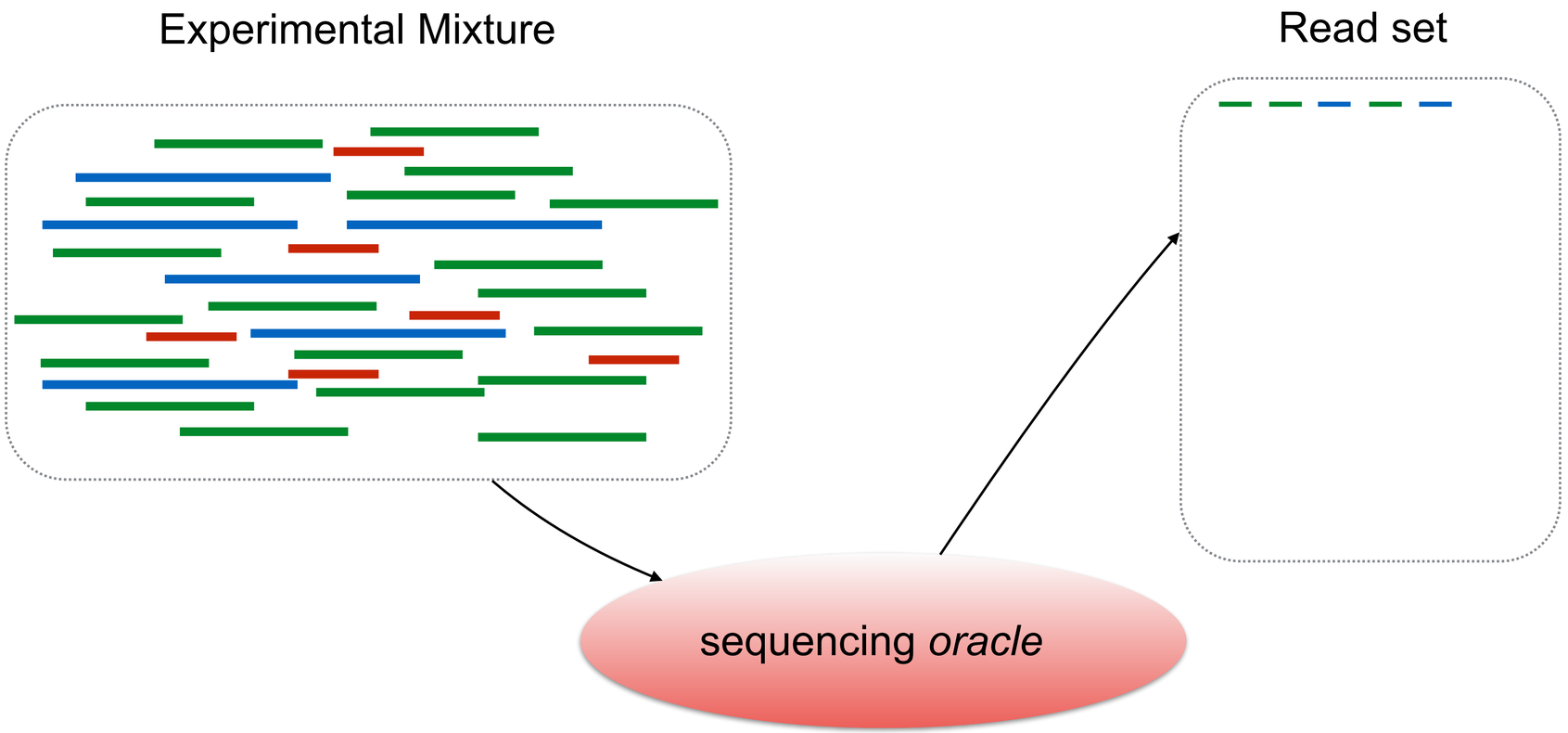


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

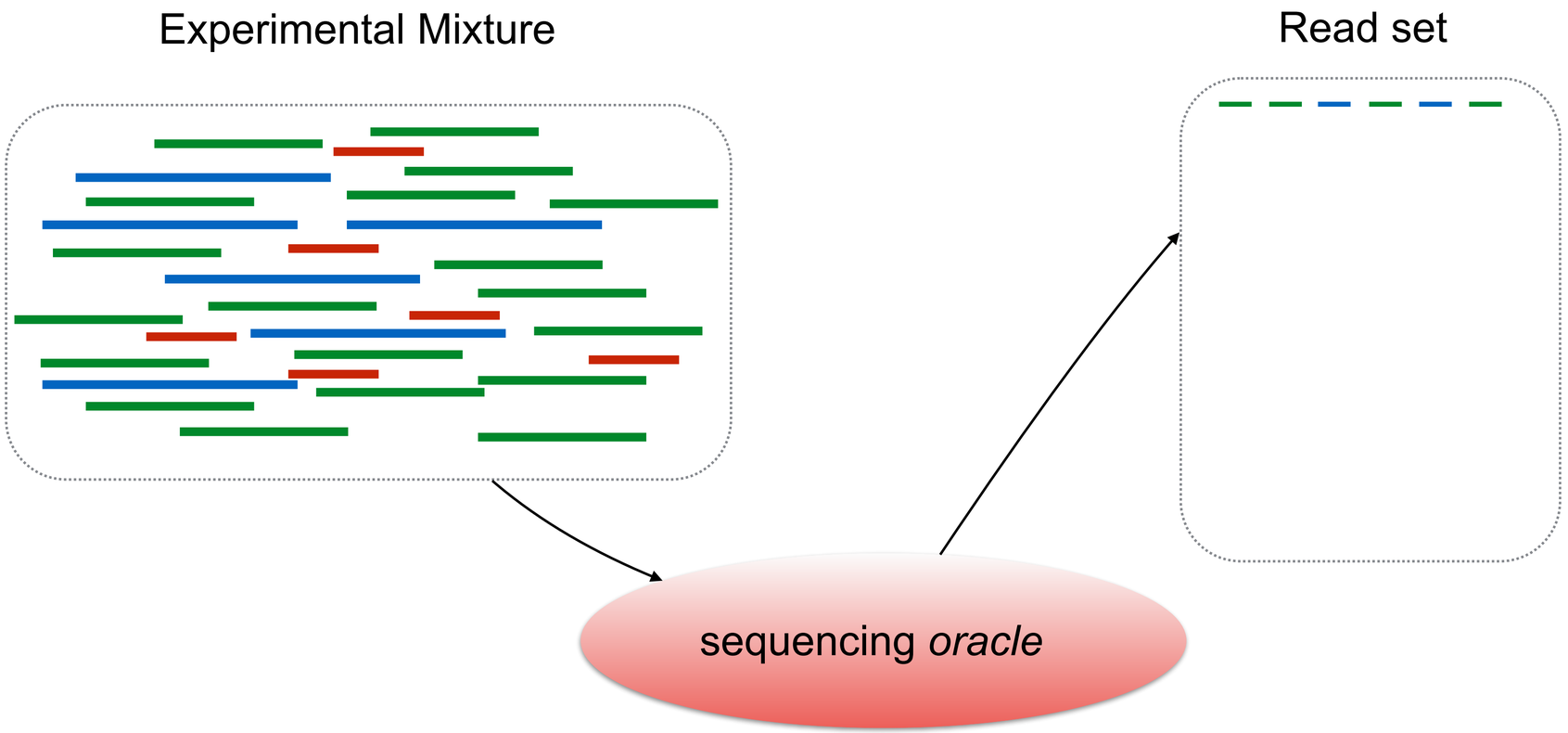


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

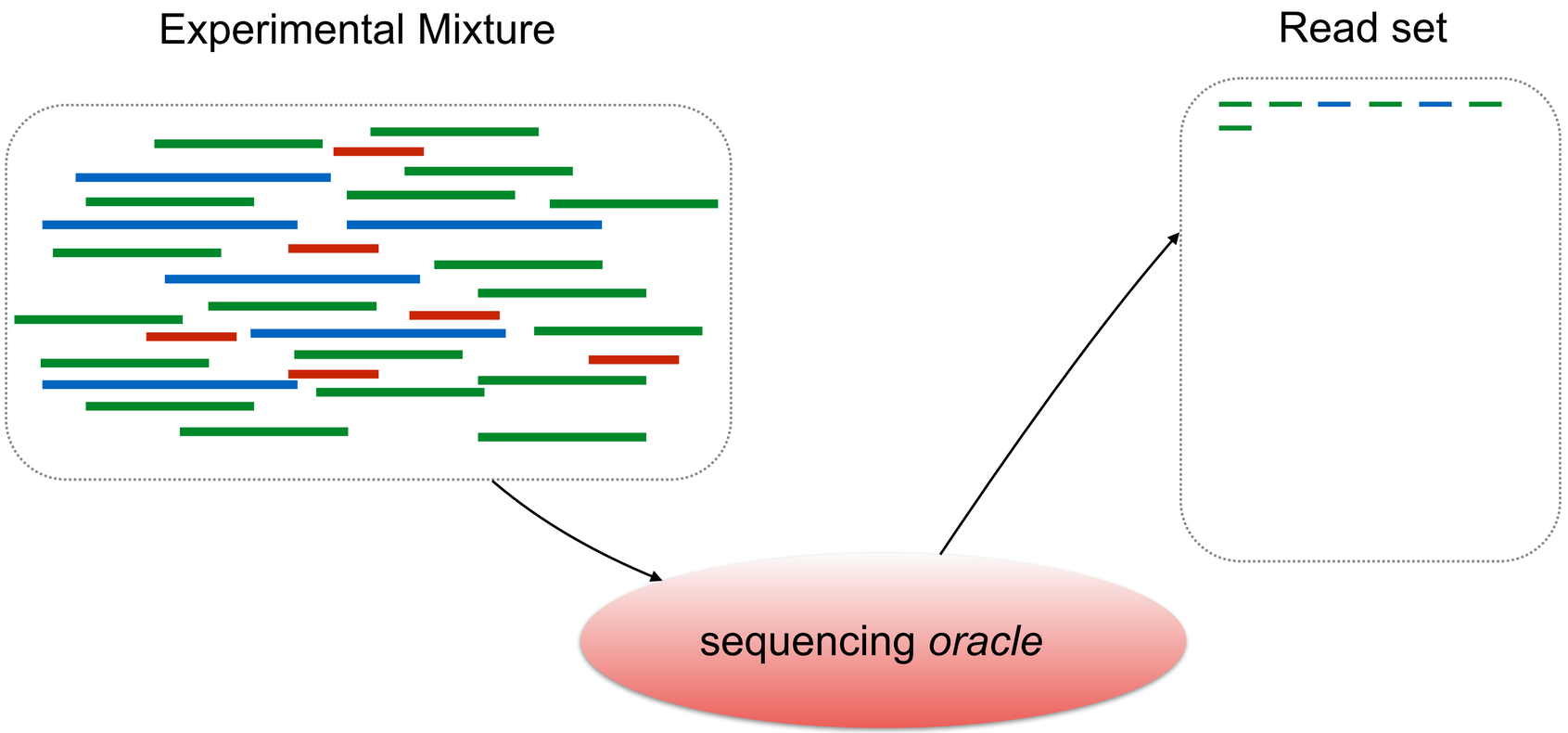


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

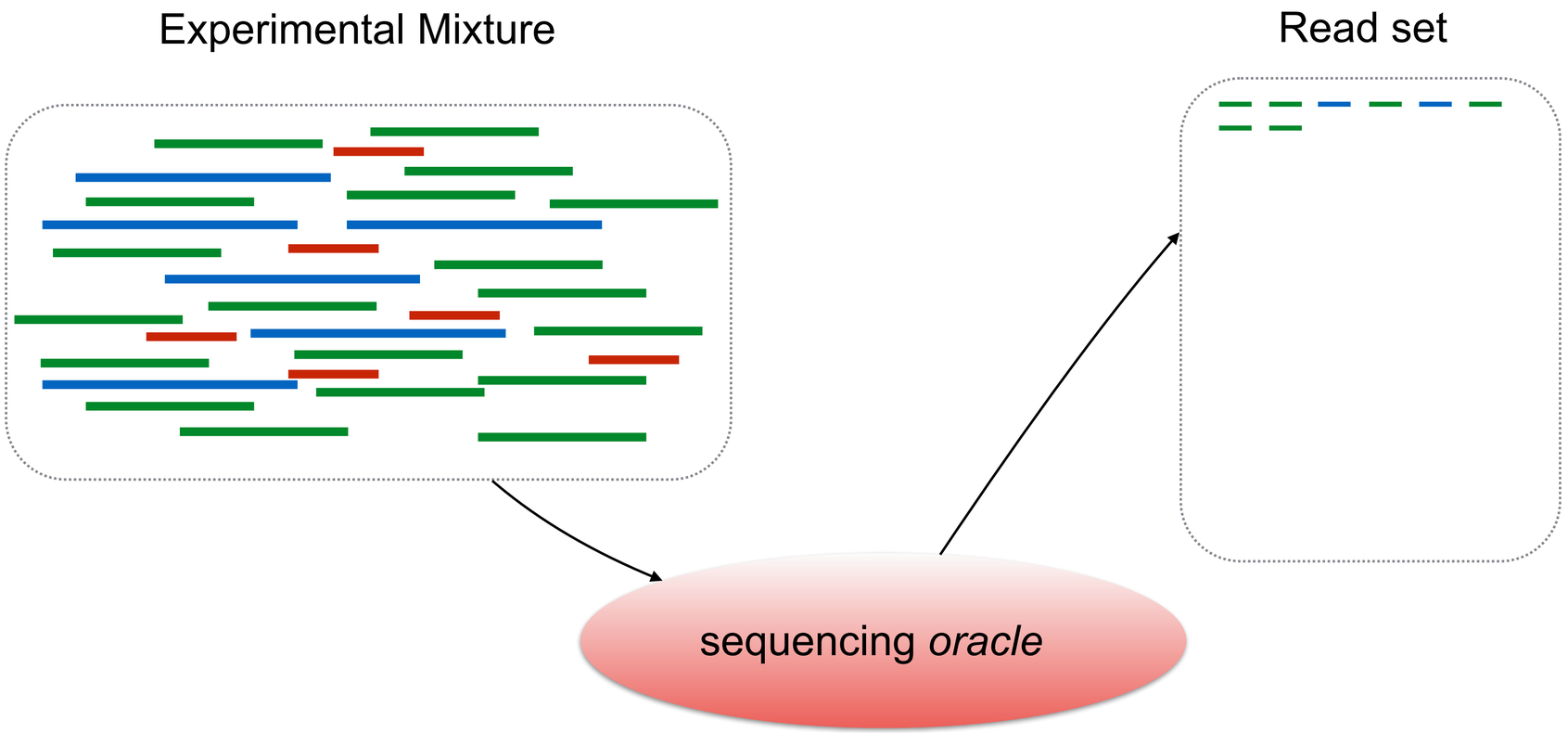


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

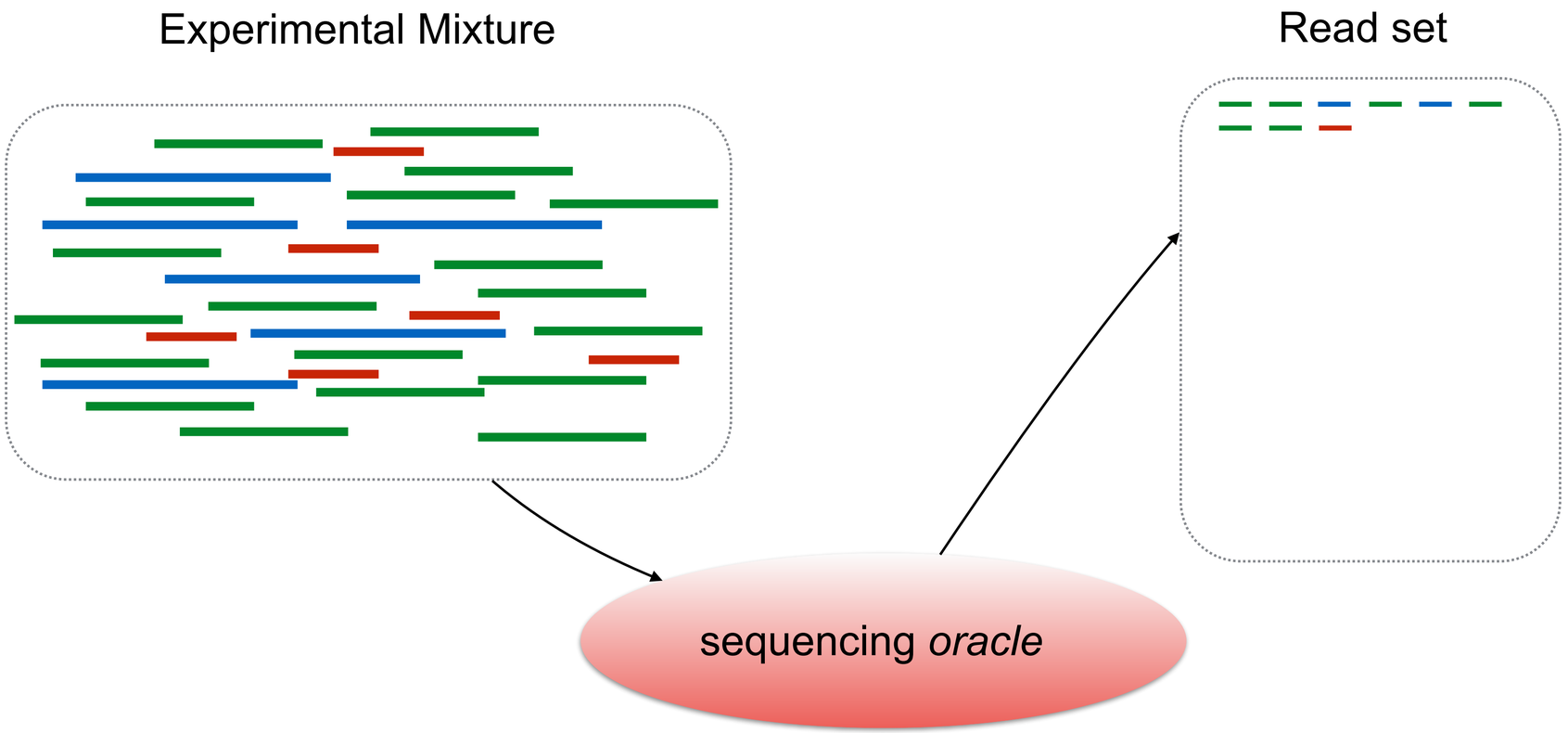


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

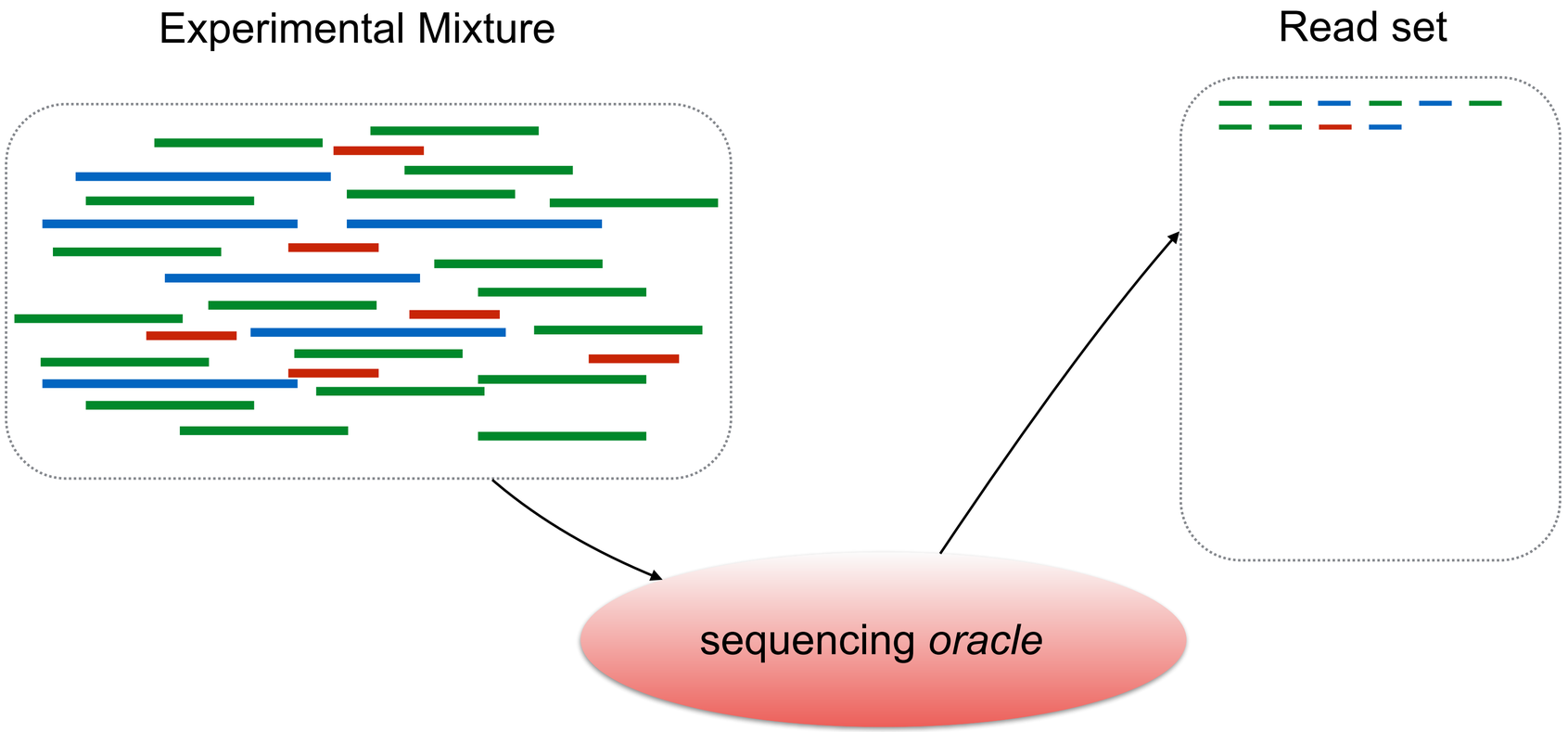


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

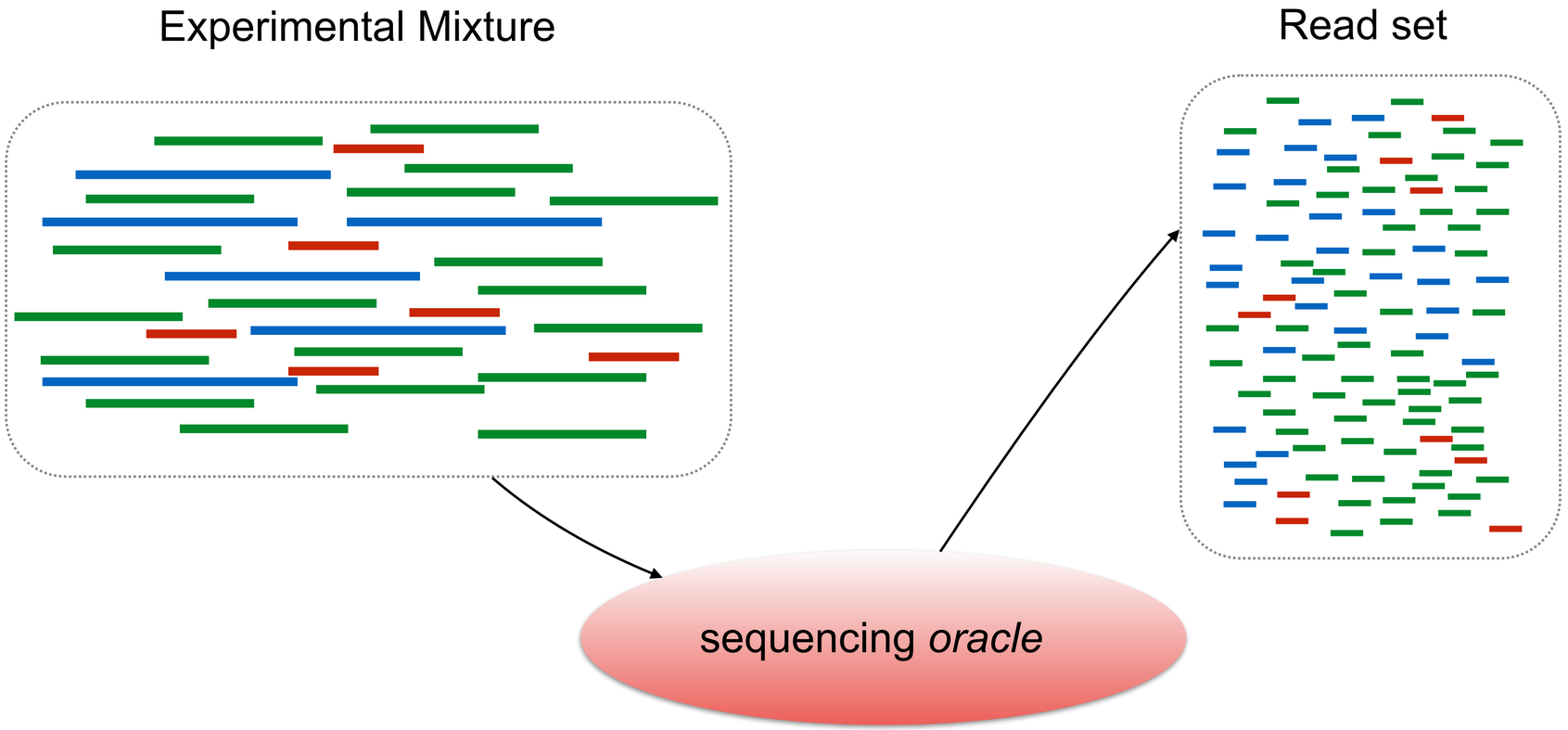


(1) Pick transcript **t** $\propto$ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

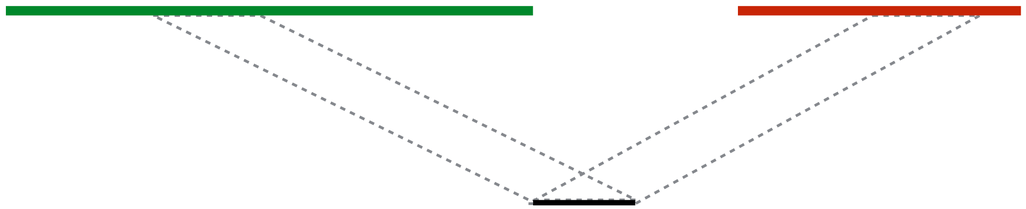
Think about the “ideal” RNA-seq experiment . . .



(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

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)

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

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

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)

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

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)

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

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

Interlude: Maximum Likelihood Estimation & the EM-algorithm

Maximum Likelihood & EM slides taken from UW CSE312 (winter '17)

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)

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

Prob. of selecting t_i *given* **η**

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 **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

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)

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

Prob. of selecting t_i *given* **η**

Depends on abundance estimate

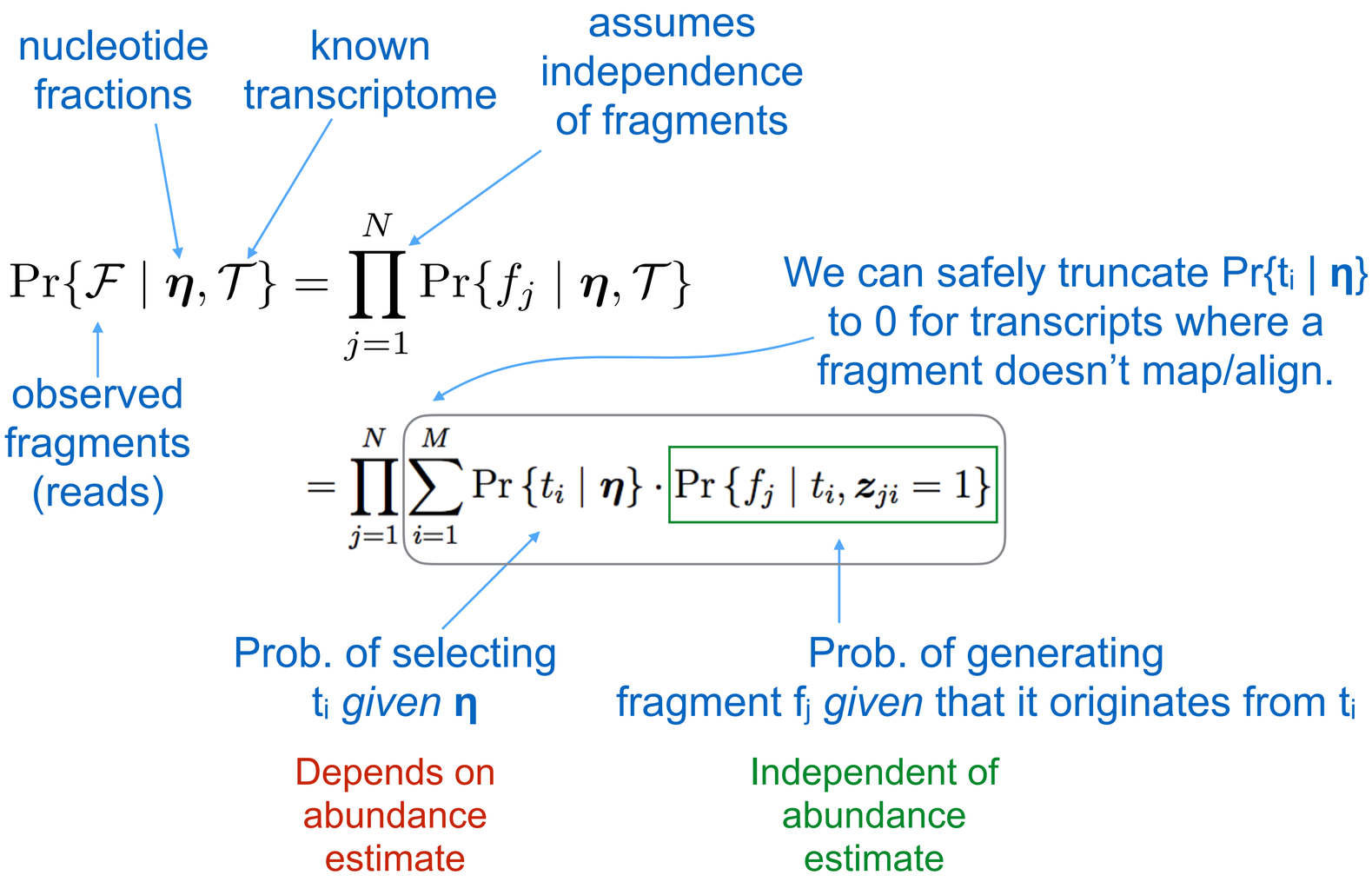
Prob. of generating fragment f_j *given* that it originates from t_i

Independent of abundance estimate

We want to find the values of **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



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)

$$E_{Z|\mathcal{F},\eta^{(t)}}[Z_{nij}] = P(Z_{nij} = 1 \mid \mathcal{F}, \eta^{(t)}) = \frac{(\eta_i^{(t)}/\ell_i)P(f_n \mid Z_{nij} = 1)}{\sum_{i',j'}(\eta_{i'}^{(t)}/\ell_{i'}')P(f_n \mid Z_{ni'j'} = 1)}$$

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

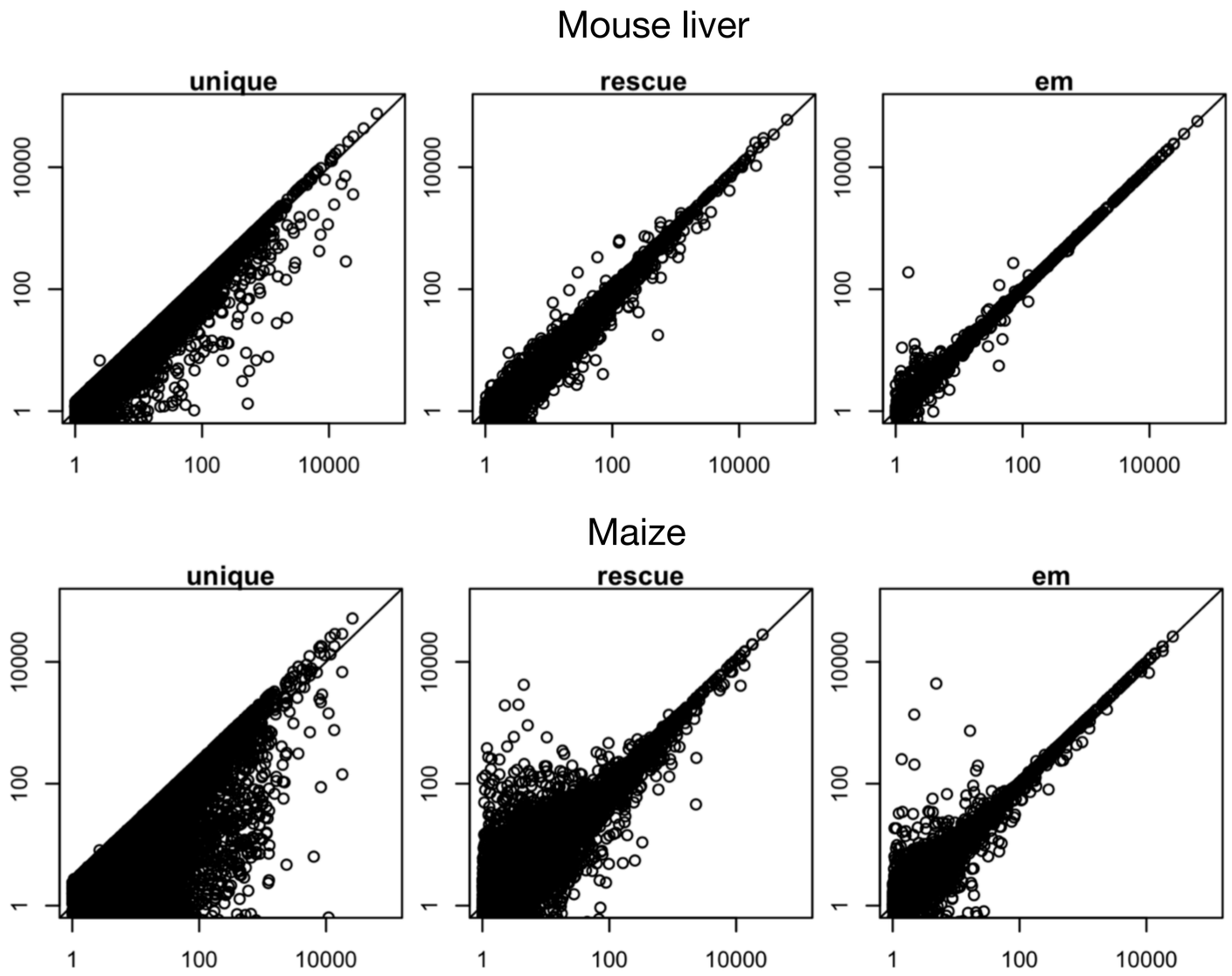
$$\eta_i^{(t+1)} = \frac{E_{Z|\mathcal{F},\eta^{(t)}}\left[C_i\right]}{N},$$

$$\text{where } C_i = \sum_{n,j} Z_{nij}$$

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

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://d
btp692

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

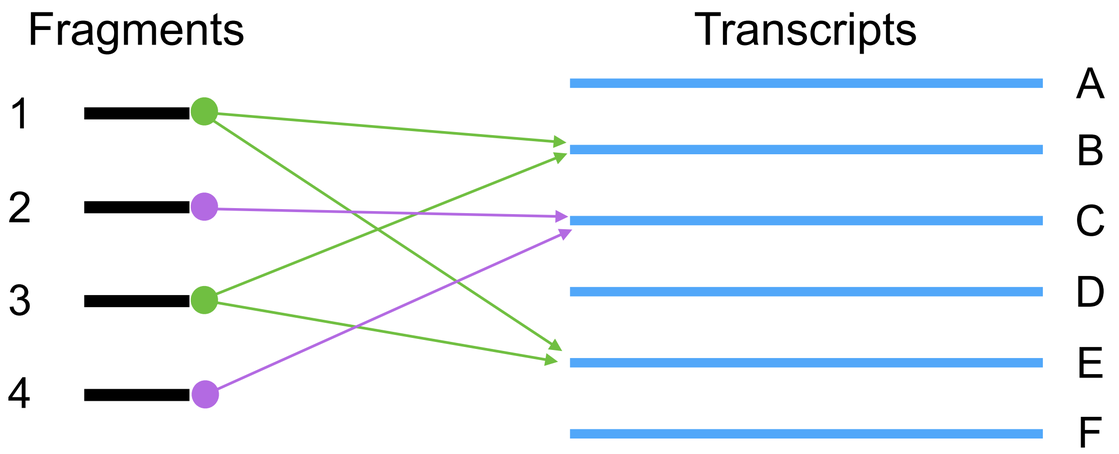
This leads to an iterative EM algorithm where *each iteration* scales in the total number of **alignments** in the sample (typically on the order of 10^7 — 10^8), and typically 10^2—10^3 **iterations**

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

Set of transcripts where f maps/aligns

*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.

Fragment Equivalence Classes



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

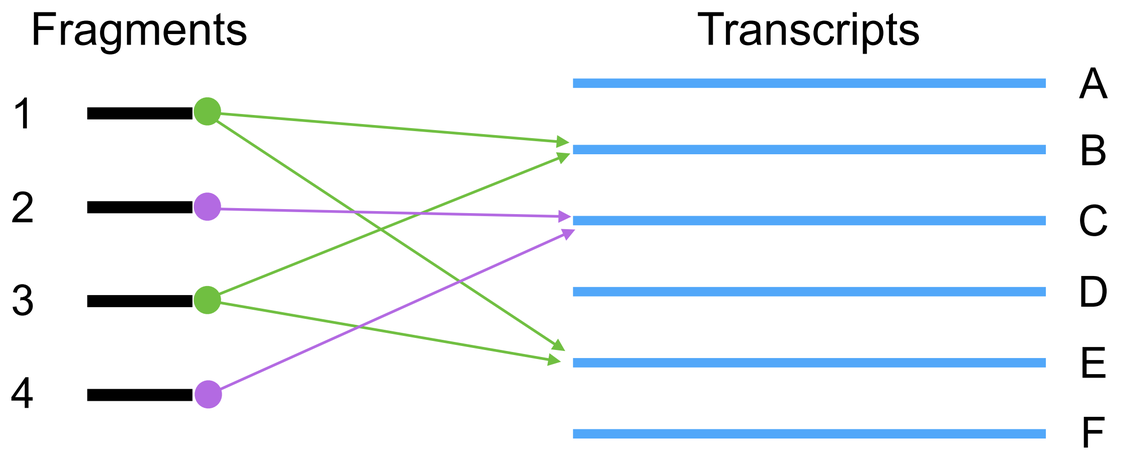
We have 4 reads, but only 2 eq. classes of reads

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.

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

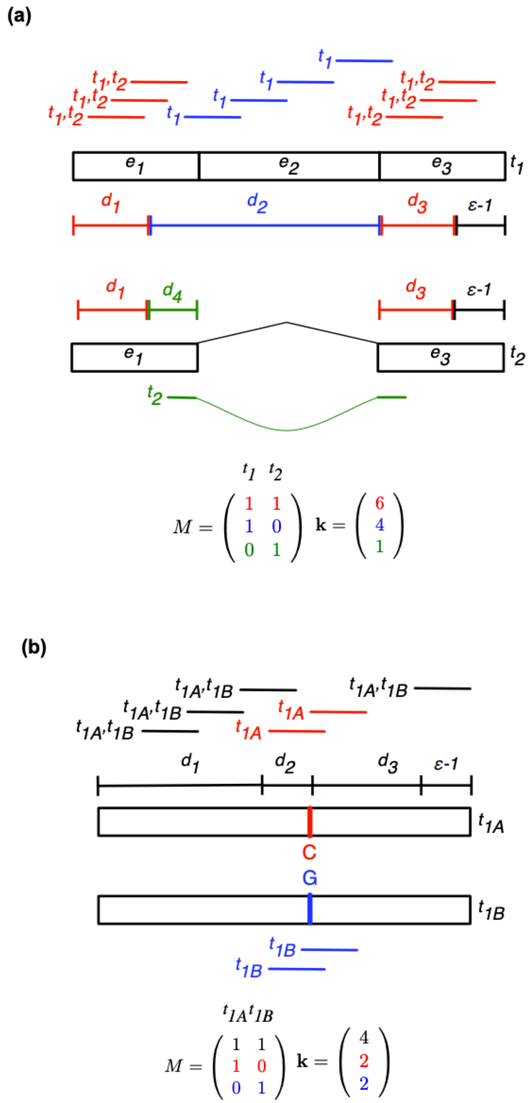


Figure 2 from Turro, Ernest, et al. "Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads." *Genome biology* 12.2 (2011): R13.

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:

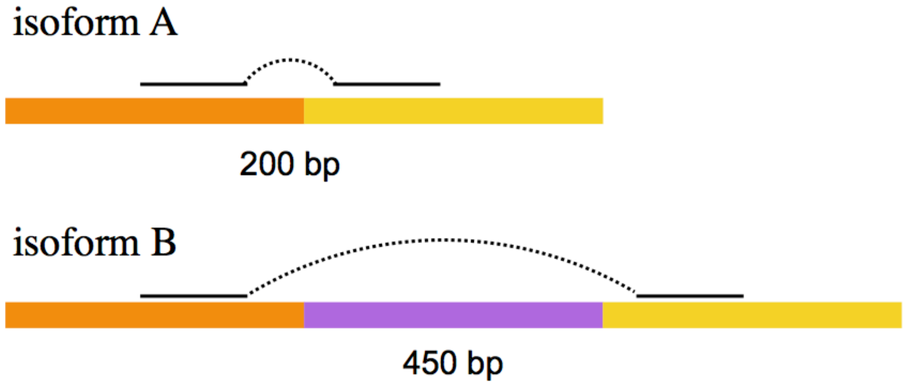
$$\mathcal{L}\left(\boldsymbol{\eta}; \mathcal{F}\right) \approx \prod_{\mathcal{F}^q \in \boldsymbol{\mathcal{C}}} \left(\sum_{\langle i, t_i \rangle \in \Omega(\mathcal{F}^q)} \Pr\left(t_i \mid \boldsymbol{\eta}\right) \cdot \Pr\left(f \mid \mathcal{F}^q, t_i\right) \right)^{N^q}$$

sum over all transcripts labeling this eq. class

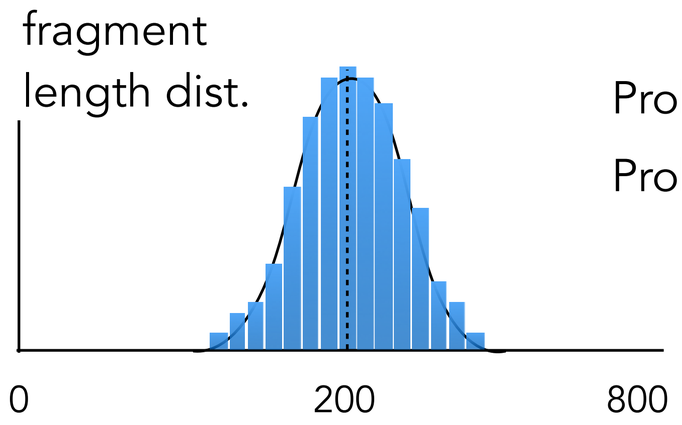
product over all equivalence classes

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.***



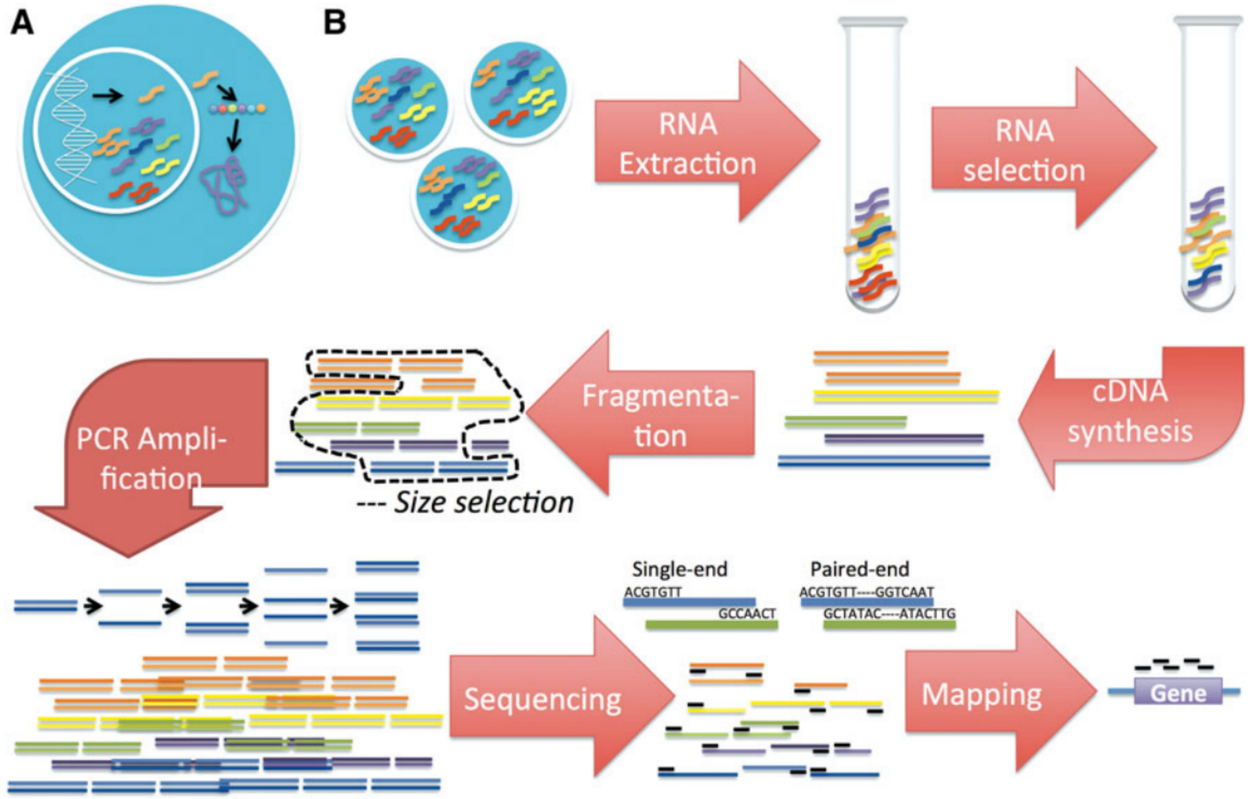
Prob of observing a fragment of size ~200 is **large**

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

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 a
of Computational Biology 24.3 (2017): 200-212.

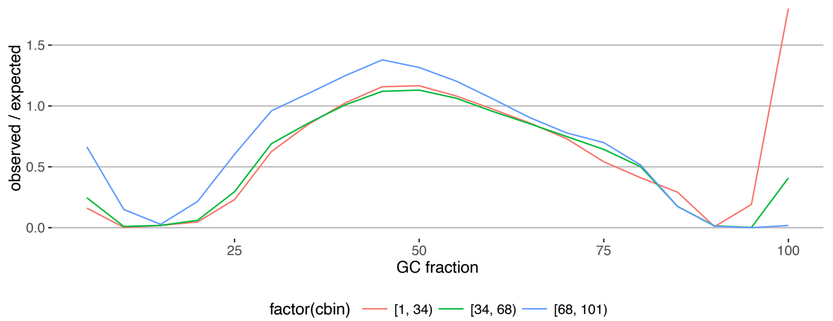
Biases abound in RNA-seq data

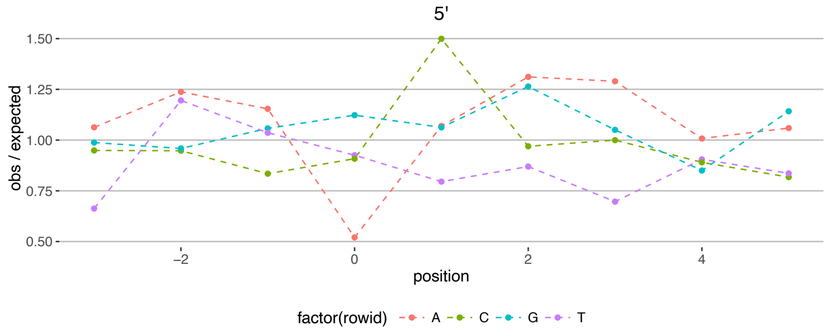
Biases in prep & sequencing
can have a significant effect on the
fragments we see:

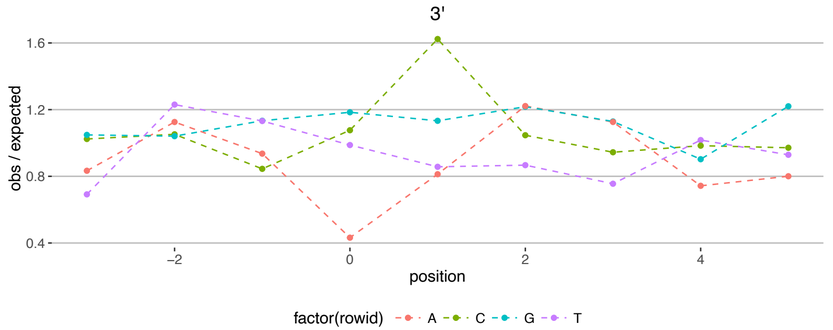
Fragment gc-bias[1]—
The GC-content of the fragment
affects the likelihood of sequencing

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



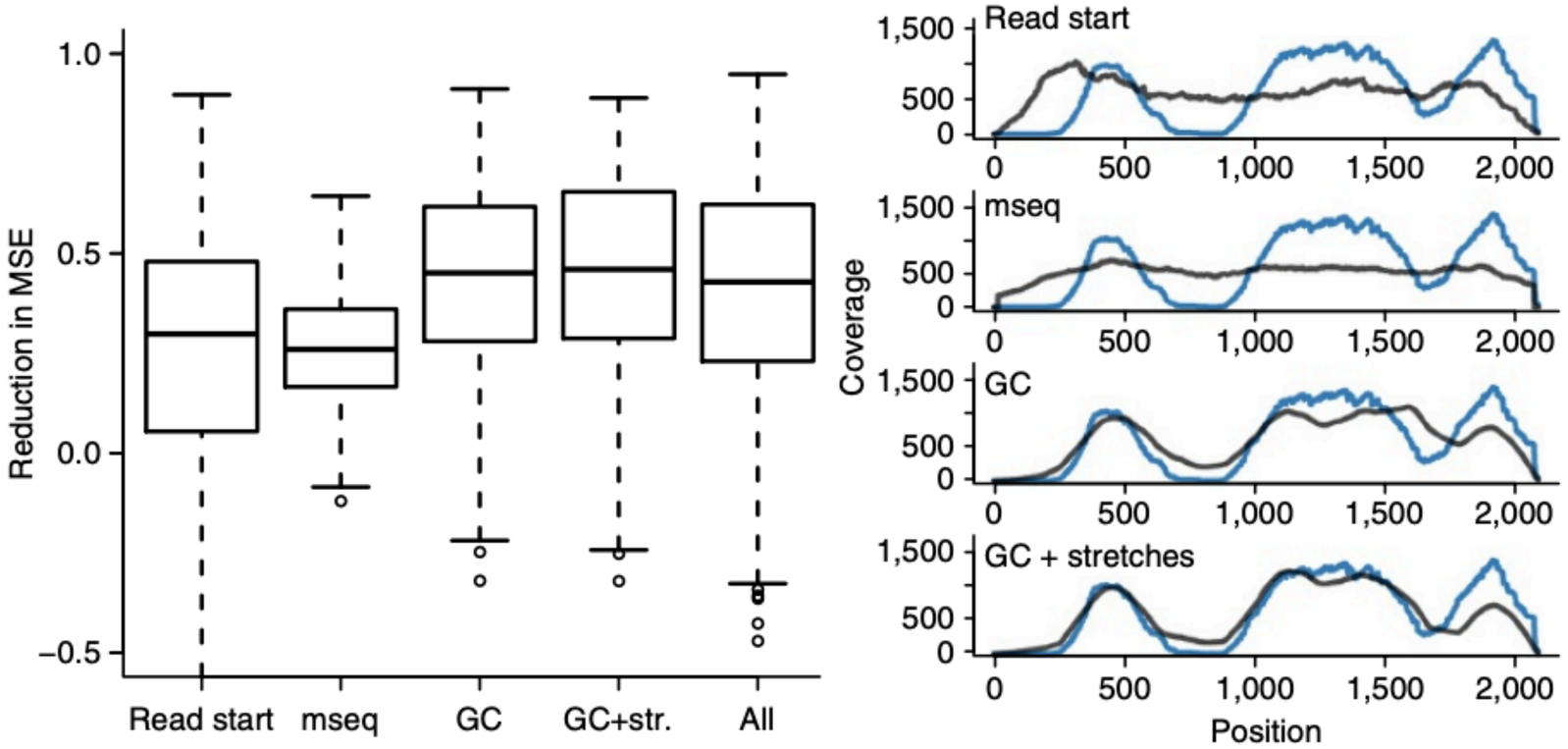




1:Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.

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

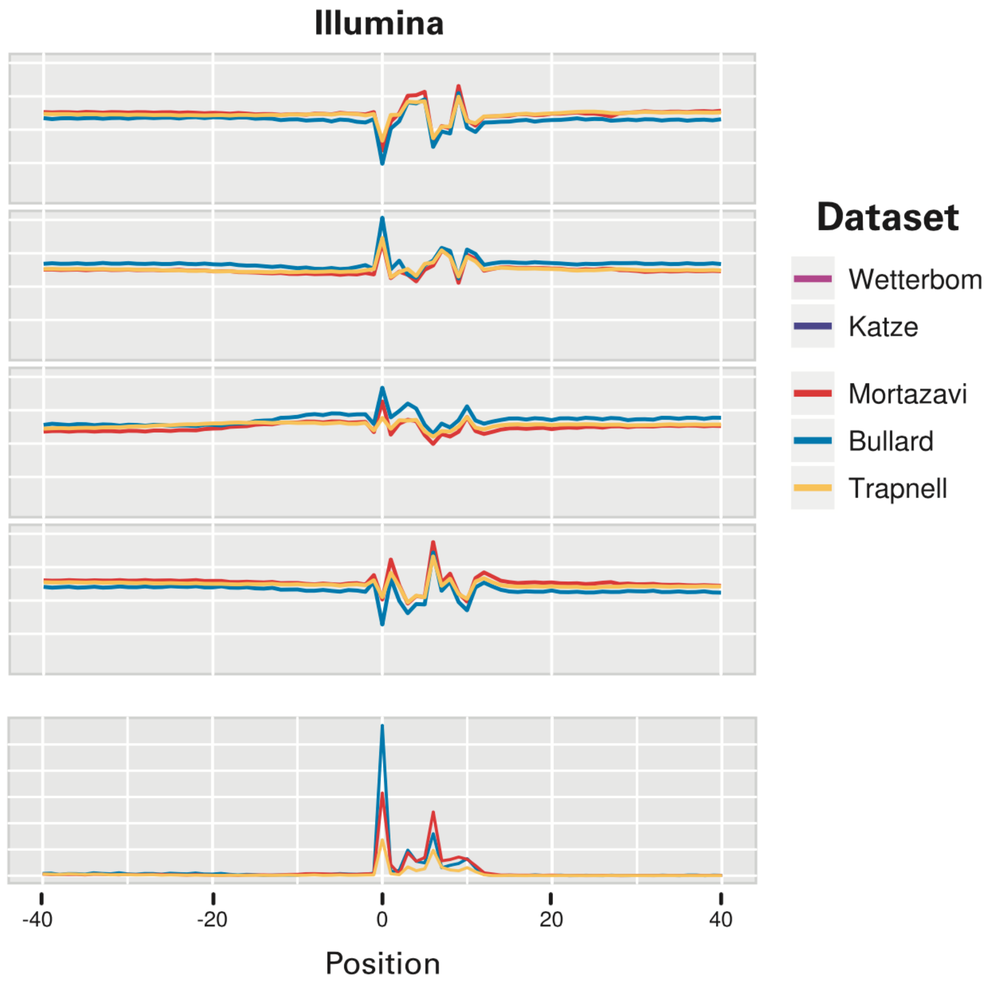
Biases abound in RNA-seq data



Fragment GC-bias is often the most extreme

Love, M. I., Hogenesch, J. B., & Irizarry, R. A. (2016). Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nature biotechnology*, *34*(12), 1287.

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 (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.

1:Love, Michael I., J

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

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.

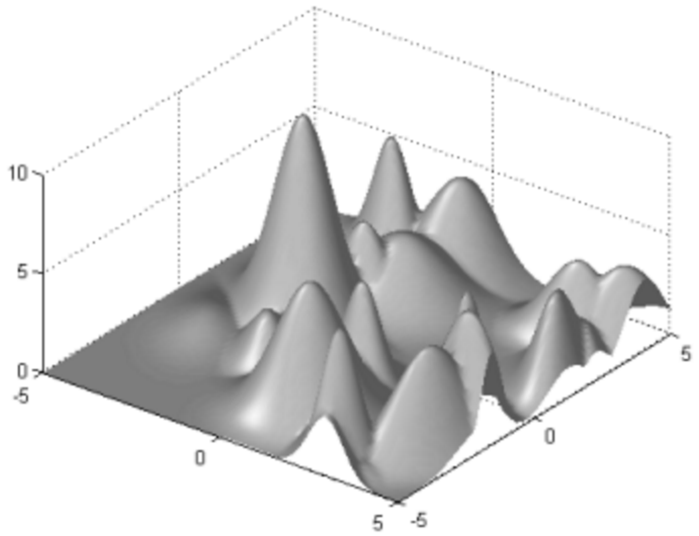
One “issue” with maximum likelihood (ML)

There are multiple sources of uncertainty e.g.

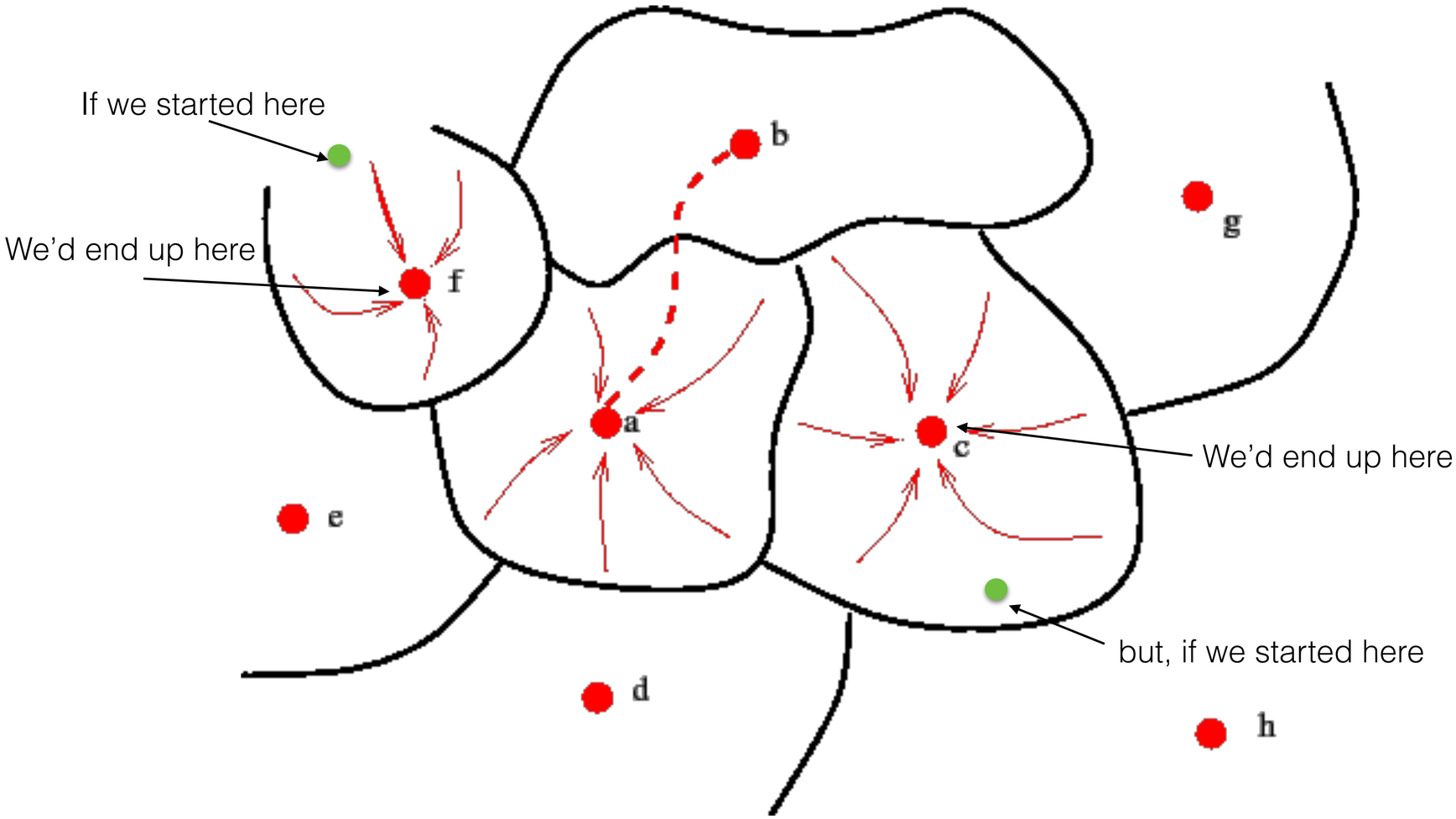
- Technical variance : If we sequenced the *exact* same sample again, we’d get a different set of fragments, and, potentially a different solution.

- 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)



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.

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.

This uncertainty matters

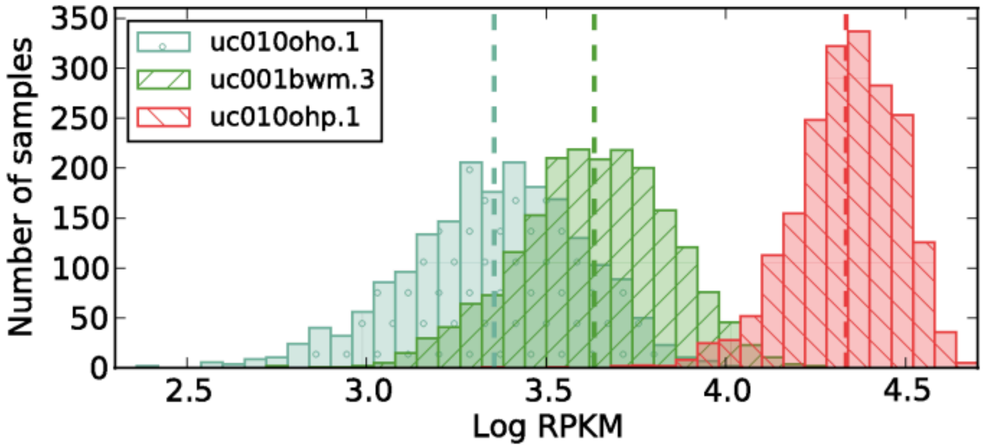
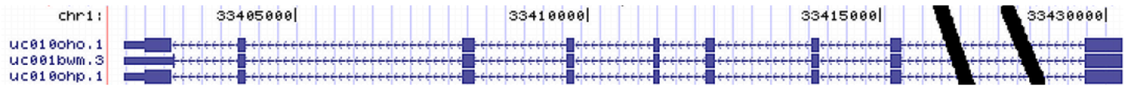


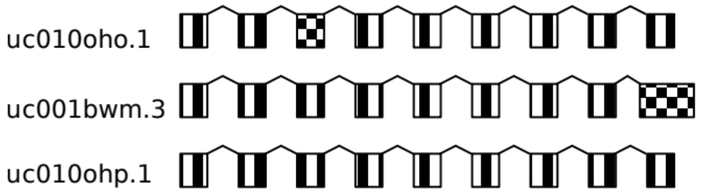
Figure 2.10: **Posterior distribution of expression levels of three transcripts of gene Q6ZMZ0.** The posterior distribution is represented in form of a histogram of expression samples converted into Log RPKM expression measure. The dashed lines mark the mean expression for each transcript.

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

This uncertainty matters

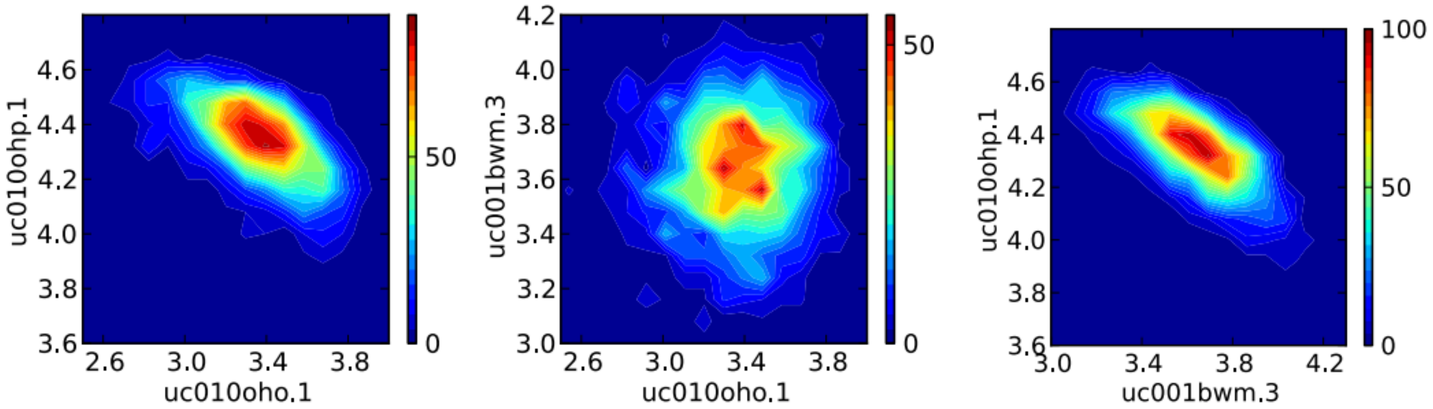


(a) Transcript sequence profile.



(b) Splice variant model.

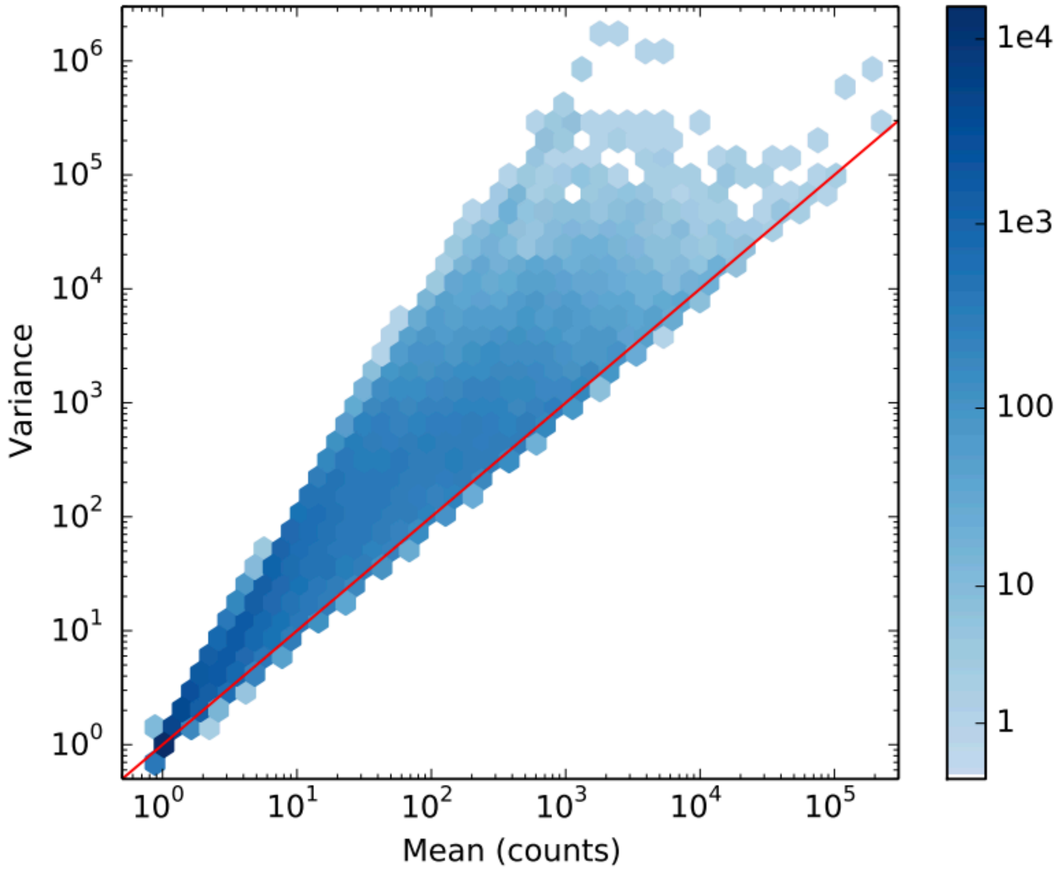
Figure 2.12: **Exon model of transcripts of gene Q6ZMZ0.** (a) transcript sequence profile obtained from the UCSC genome browser (Kuhn et al., 2013). In this annotation, transcript uc001bwm.3 has different 3' untranslated region and transcript uc010oho.1 has extra nucleotides at the end of second exon. 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)!

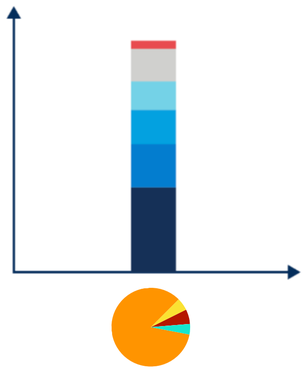
*Glaus, Peter. *Bayesian Methods for Gene Expression Analysis from High-throughput Sequencing Data*. Diss. University of Manchester, 2014.

Bulk RNA-seq →Single-cell RNA-seq

A brave new world

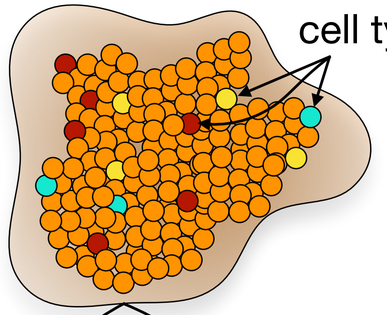
Bulk RNA-seq

- Typically millions or 10s of millions of cells
- High-fidelity & high-sensitivity
- Measure transcript abundance at the population-level



Tissue

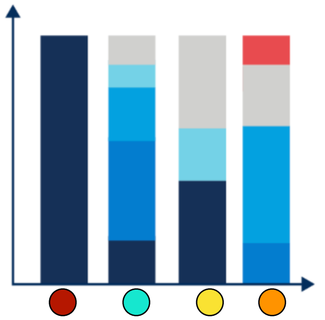
rare cell types



These present **very** different challenges

Single-cell RNA-seq

- Typically tens of thousands of cells
- Low-coverage (reads / cell)
- Measure gene abundance at the single-cell level



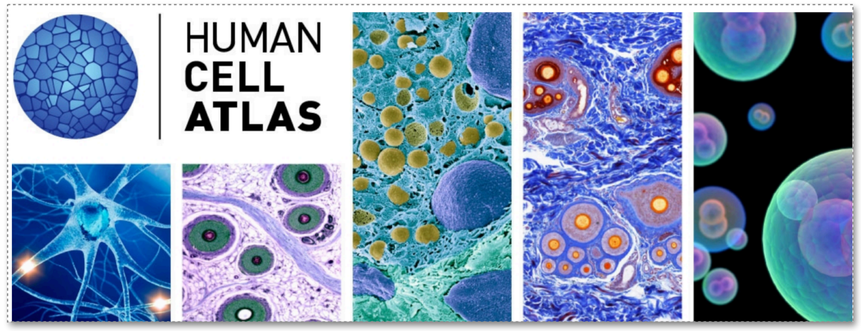
involves cell-type identification (supervised or unsupervised)

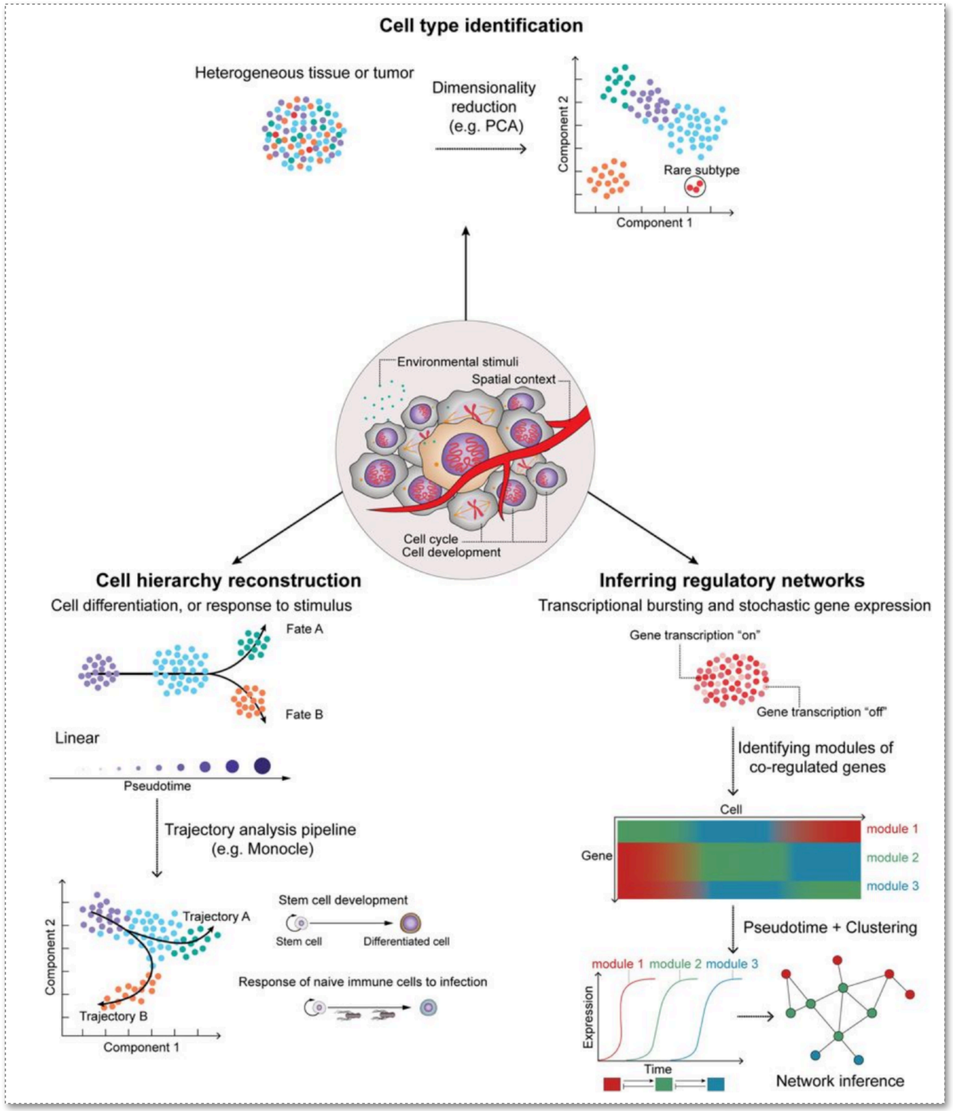
https://www.slideshare.net/QIAGENscience/why-scientist-analyze-single-cells (smoothie)
https://community.10xgenomics.com/t5/10x-Blog/Single-Cell-RNA-Seq-An-Introductory-Overview-and-Tools-for/ba-p/547 (bar chart images)

What are some exciting prospects of single-cell sequencing?

Explore biology at *unprecedented resolution.*
Some *transformative* applications:

- Study *treatment-resistant* cells in disease (cancer)
- Understand tissue / organism *development* (cell fate)
- Understand immune response at the cellular level
- Understand dynamic cellular processes
- Learn how expression is regulated (regulatory net.)
- Characterize new cell types & cell states (HCA)



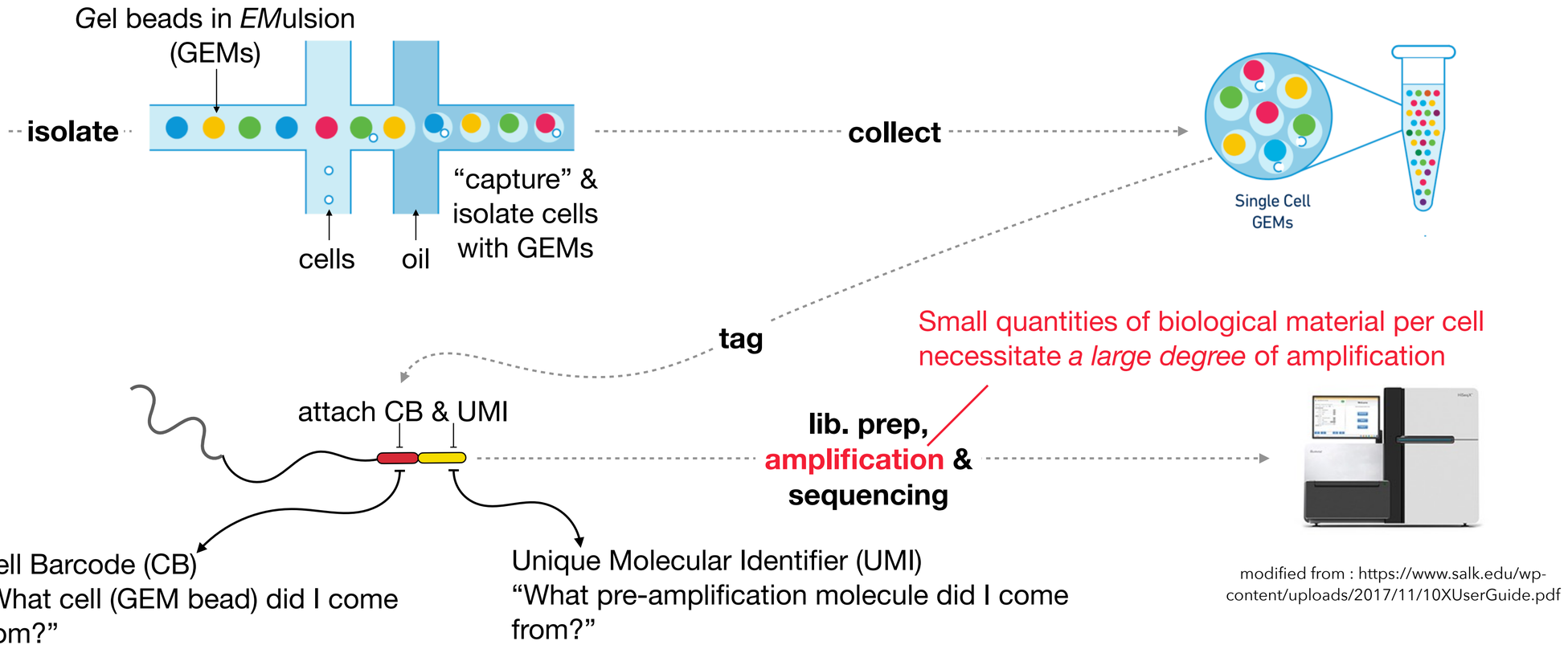


Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, *50*(8), 96.

https://www.broadinstitute.org/research-highlights-human-cell-atlas

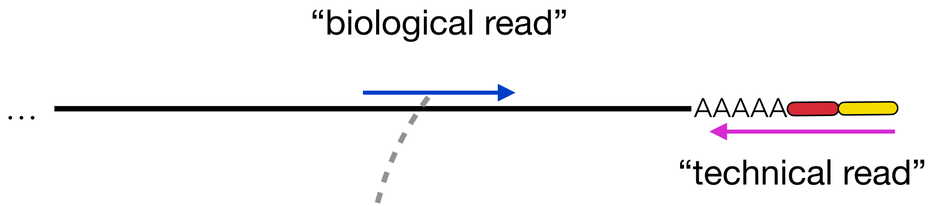
(Droplet-based) Single-cell RNA-seq

Many different protocols — here is a gist of droplet-based (microfluidic) techniques (the general approach used by 10X genomics Chromium platform).

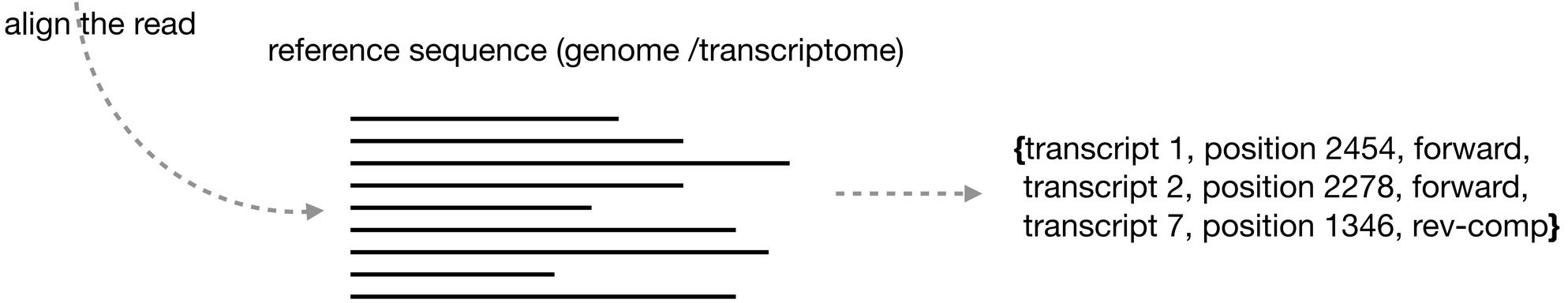


modified from : https://www.salk.edu/wp-content/uploads/2017/11/10XUserGuide.pdf

Determining the origin of “biological” reads



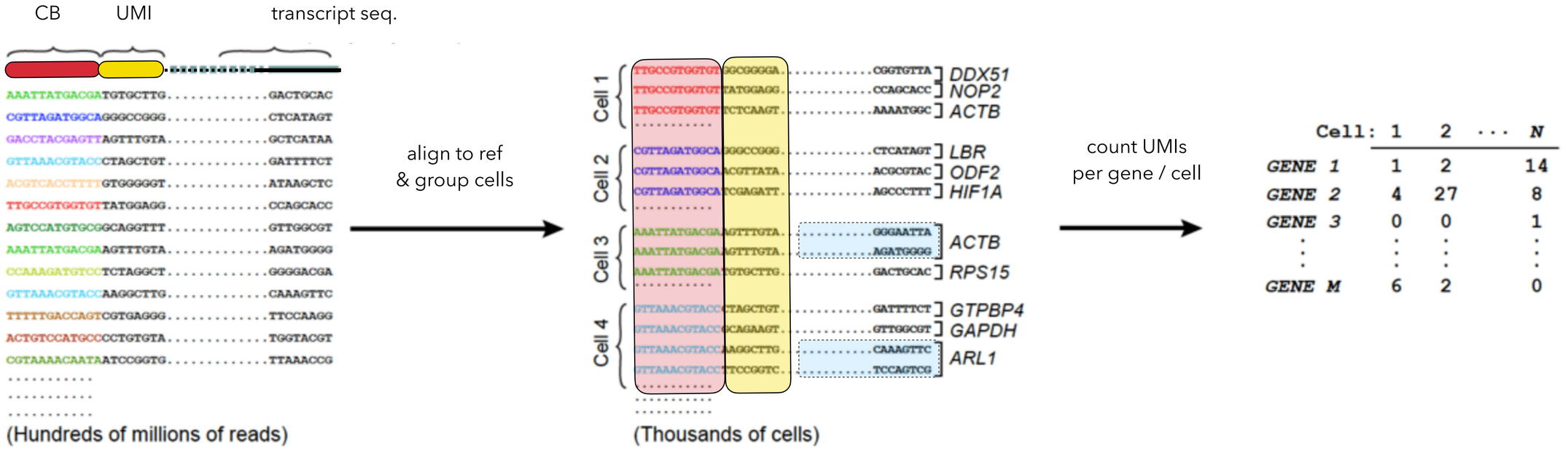
Note: because the primer with cell barcode & UMI binds to polyA, we expect a strong 3’ bias in where the biological reads are sampled



Partition by cell & group by UMI

In the case of perfect alignment & sequencing, and no ambiguity, using the CB & UMI to estimate gene expression we simply:

1. Group sequenced reads by their CB (read → cell)
2. Collapse all reads with the same UMI (read in cell → molecule)
3. Count!



In practice both mapping & sequencing are imperfect and the process is more complicated

http://mccarrolllab.org/wp-content/uploads/2015/05/figure-2D.png

So what are the challenges?

A *non-exhaustive* list

There is a *lot* of data — e.g. PBMC 10k dataset (~10k cells) has 638,901,019 (paired-end) fragments is 44GB (*compressed*) : **methods need to scale just to be usable**

Scalability needed in (at least) 2 dimensions

- Time : we'd like to process data quickly and scale well with many threads
- Memory : we'd like to be able to process data with moderate & stable memory requirements

Resolving the cell, gene and splicing status of origin for a fragment is not always simple : **methods need to be smart**

- Fragments can arise from unexpected locations and should be accounted for
- Cellular barcodes (& UMIs) are subject to sequencing error just like "biological" reads
- Reads are not always clearly attributable to a single gene (e.g. sequence similar genes)

The existing commercial solution (CellRanger) is **computationally & memory intensive**.
Also, since version 3 (now version 7), it is **closed source & method is "opaque"**.

Alevin-fry's USA mode enables accurate RNA-velocity inference

RNA-velocity uses the ratio of predicted spliced to unspliced molecules across cells to model the *dynamics of splicing*.

Allows the prediction of the movement of cells through transcriptional space / developmental time.

Velocyto

Letter | Published: 08 August 2018

RNA velocity of single cells

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson & Peter V. Kharchenko

Nature **560**, 494–498 (2018) | Cite this article

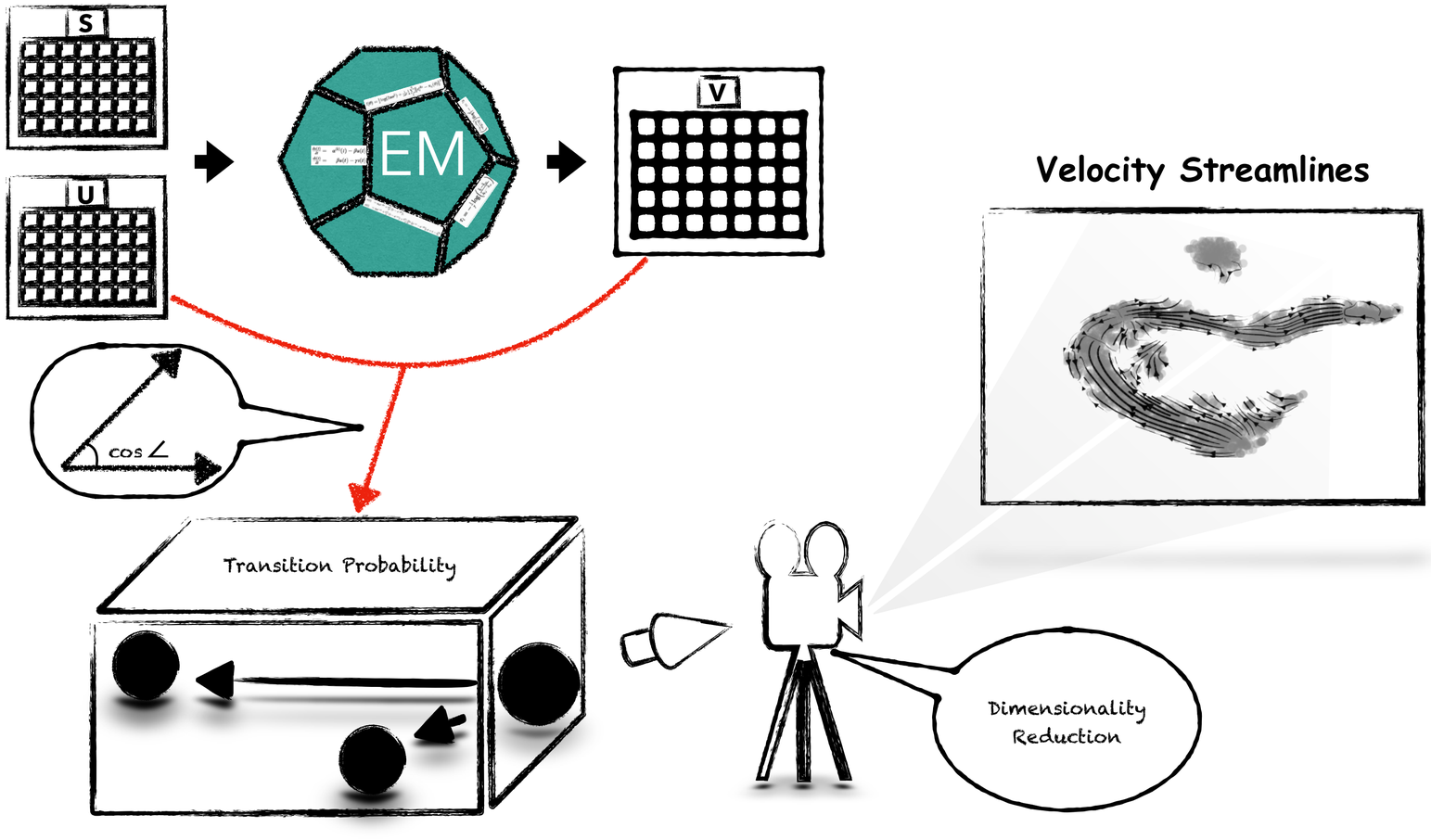
scVelo

Article | Published: 03 August 2020

Generalizing RNA velocity to transient cell states through dynamical modeling

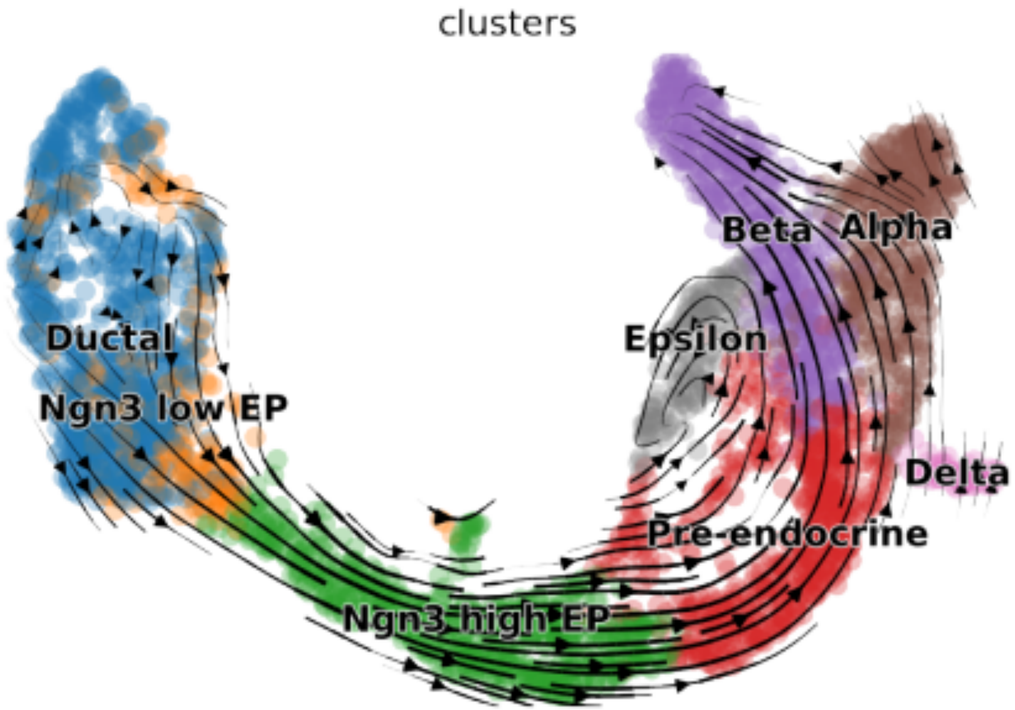
Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf & Fabian J. Theis

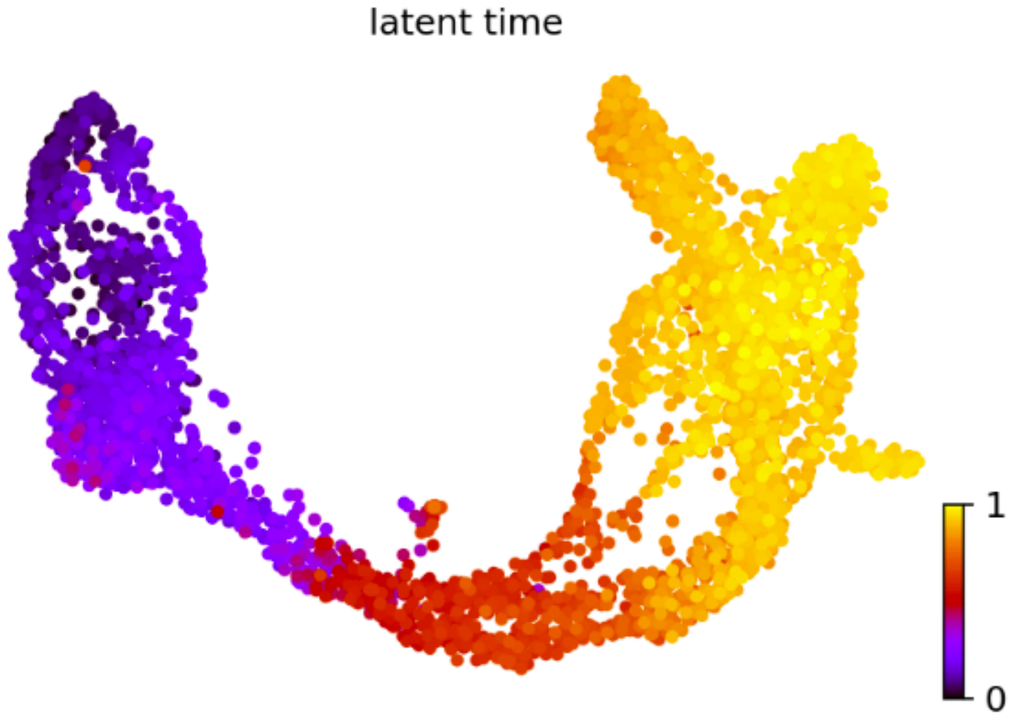
Nature Biotechnology **38**, 1408–1414 (2020) | Cite this article



RNA-velocity inference (a case where it works as expected)

Data from mouse pancreas, showing endocrinogenesis*





*Aimée Bastidas-Ponce, Sophie Tritschler, Leander Dony, Katharina Scheibner, Marta Tarquis-Medina, Ciro Salinno, Silvia Schirge, Ingo Burtscher, Anika Böttcher, Fabian Theis, Heiko Lickert, and Mostafa Bakhti. Massive single-cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development*, January 2019. doi: 10.1242/dev.173849.

Bonus Material

Bias Modeling

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

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

Fragment GC bias model:

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

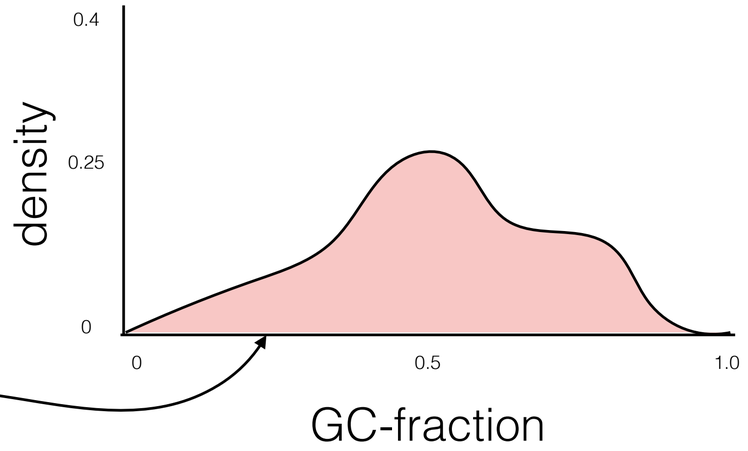
Foreground:

Observed

Background:

Expected given est. abundances

GC-fraction of fragment



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*

Bias Modeling

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

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

Seq-specific bias model*:

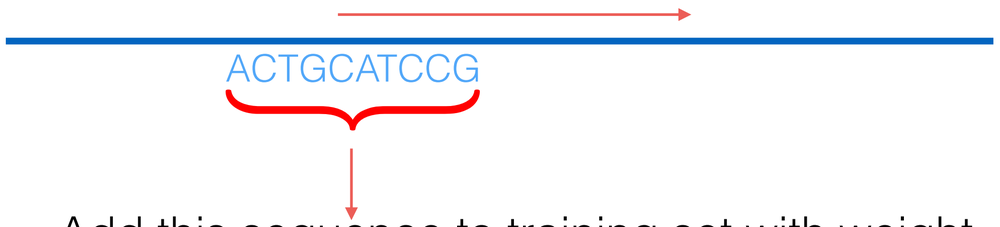
VLMM for the 10bp window surrounding the 5' read start site and the 3' read start site

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.

Bias Modeling

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

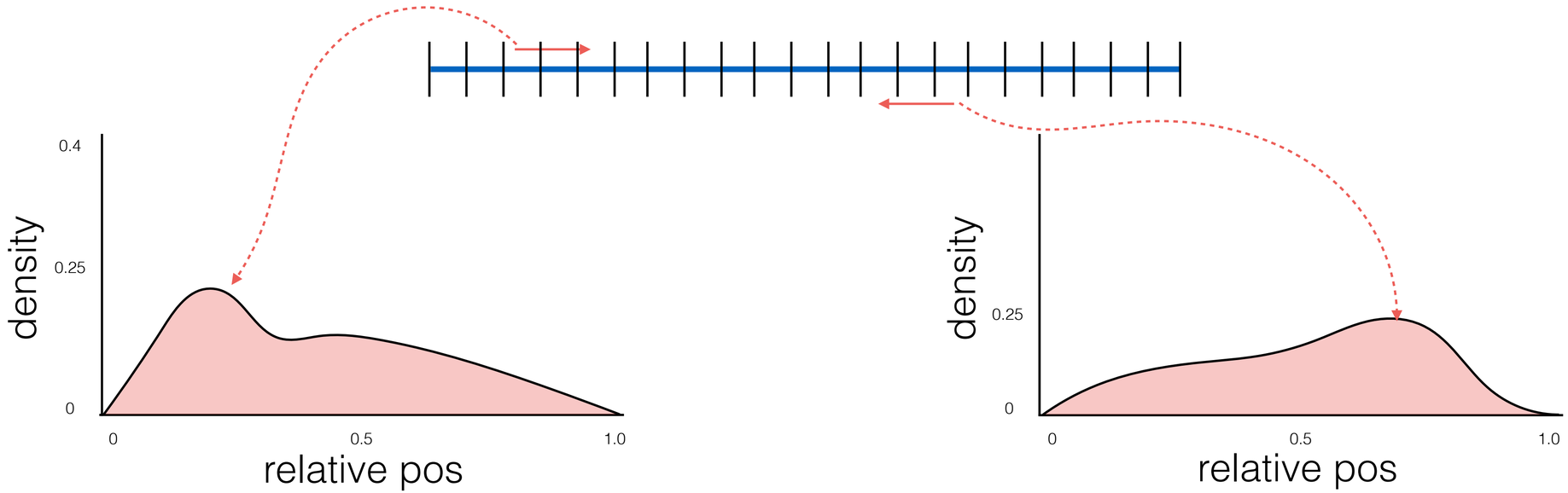
$$\tilde{\ell}'_i = \sum_{j=1}^{j \leq \ell_i} \sum_{k=1}^{k \leq f_i(j,L)} \frac{b_{gc+}(t_i, j, j+k)}{b_{gc-}(t_i, j, j+k)} \cdot \frac{b^{5'}_{s+}(t_i, j)}{b^{5'}_{s-}(t_i, j)} \cdot \frac{b^{3'}_{s+}(t_i, j+k)}{b^{3'}_{s-}(t_i, j+k)} \cdot \frac{b^{5'}_{p+}(t_i, j+k)}{b^{5'}_{p-}(t_i, j+k)} \cdot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

Position bias model*:

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.

A few ways to implement Gibbs Sampling for this problem

The model of MMSeq

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

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

The full conditionals are:

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

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

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

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

[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)

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

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.]