Exploring the Transcriptome using RNA-Seq

Genomes to Genes

Once we have a genome (or an approximation thereof), how do we find the genes?

In many genomes (e.g. the human genome), only a small fraction of the genetic material codes for proteins — much of the DNA is regulatory in nature, and much has no known function.

Finding Genes

There are 3 main categories of gene finding (with a genome)

ab initio

latin — “from the beginning”

w/o experimental evidence

based on predictive modeling

how well do genomic sequences score under our “gene model”?

comparative

make use of knowledge across species

a known human gene is strong evidence for a chimp gene

many “housekeeping” genes are incredibly similar across highly divergent species

combined / extrinsic

Make use of experimental evidence (e.g. RNA-seq)

Evidence highlights transcribed regions

Gene structure extracted from evidence (potentially combined with model predictions)

Other Ways of Using Experimental Evidence

Experimental evidence (RNA-seq, in particular) is a great help in improving gene prediction. However, its uses stretch **far** beyond assisting *ab initio* gene prediction.

Transcript quantification

Differential expression, alternative splicing analysis

Fusion/chimera detection

Variant (SNP, SV, CNV) detection

Transcript assembly

Genome guided & de novo

Build higher-level models of transcription

co-expression networks -> regulatory networks

What is “experimental” data?

There are many ways we can obtain “experimental” evidence of a gene.

Sequencing of protein product (mass spec.)

Expensive & slow, but provides direct evidence of protein coding genes (reverse translation)

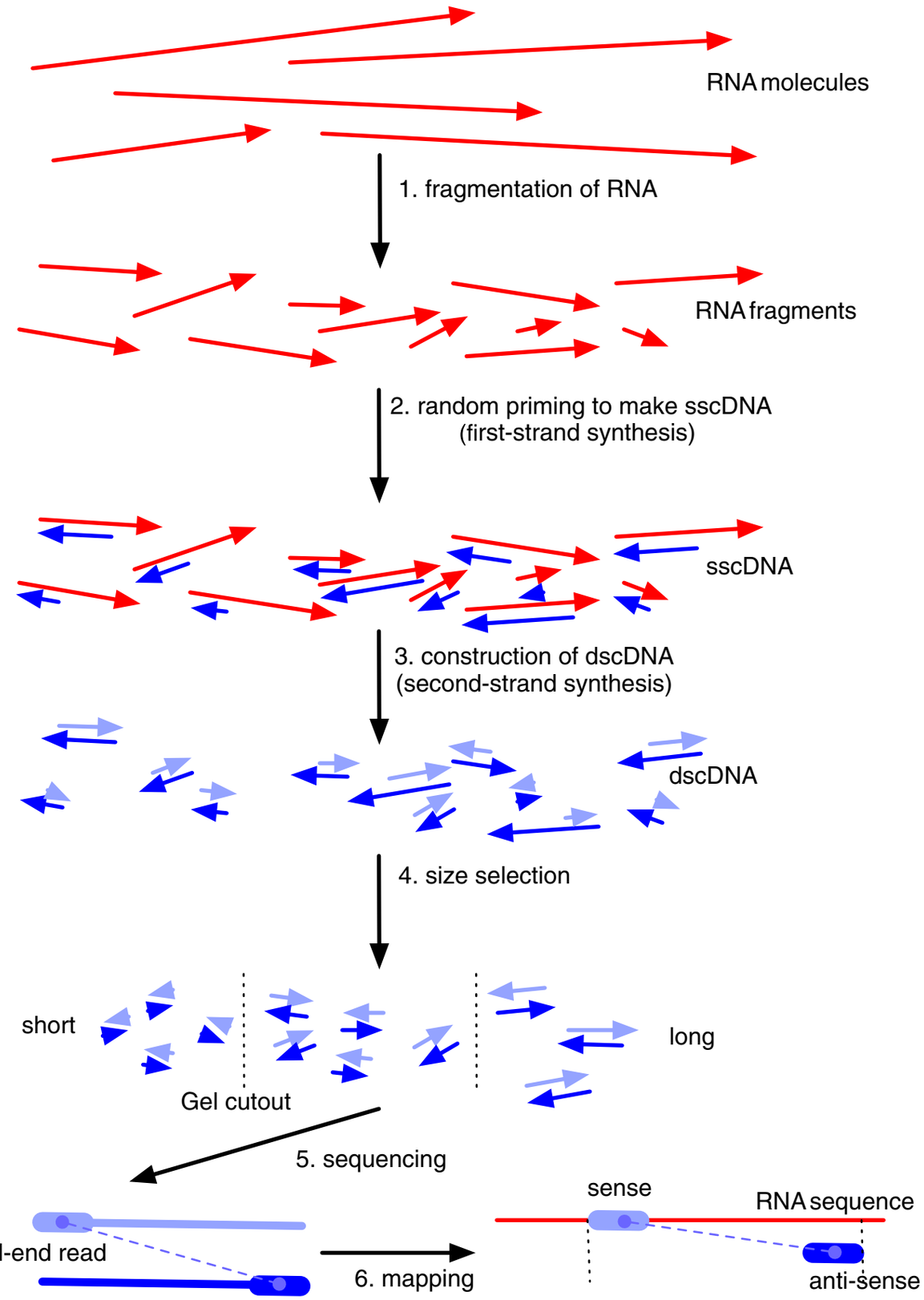
Expressed Sequence Tags (EST)

Targeted sequencing (typically Sanger sequencing) of expressed transcripts

RNA-Seq

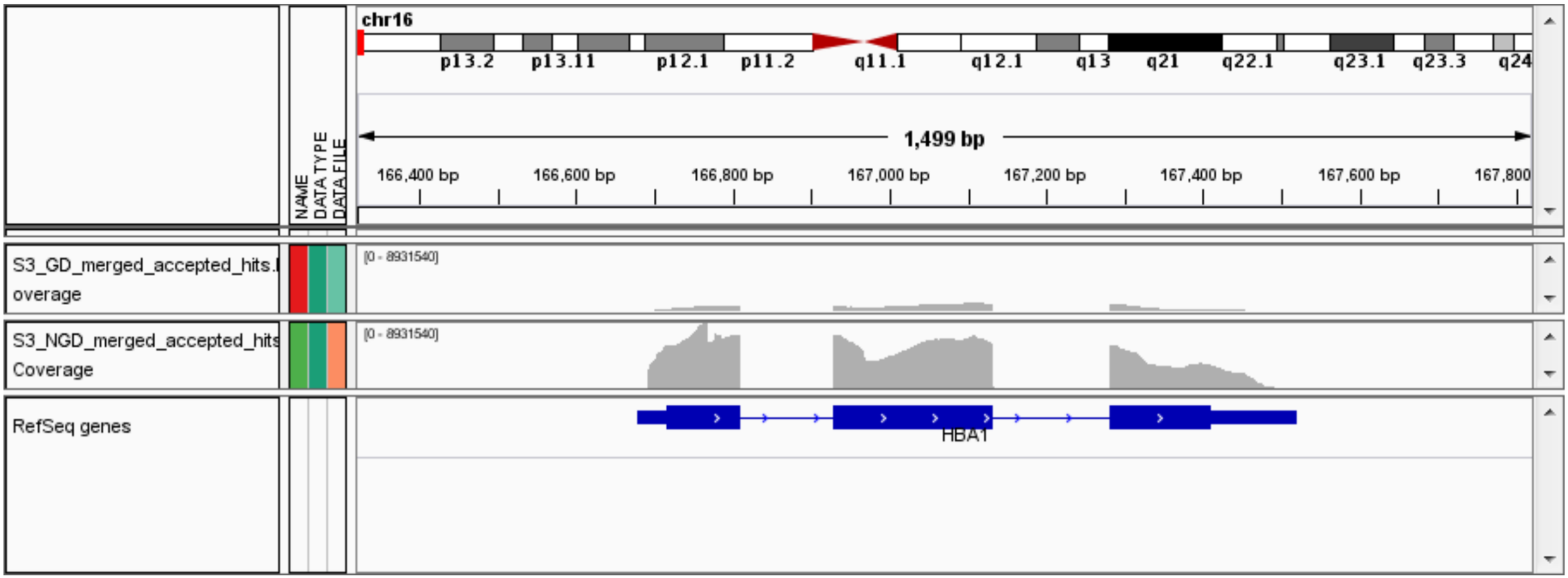
High throughput sequencing of the “transcriptome”

RNA-Seq — Sequencing Transcripts



Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): R22.

How is RNA-seq Useful?

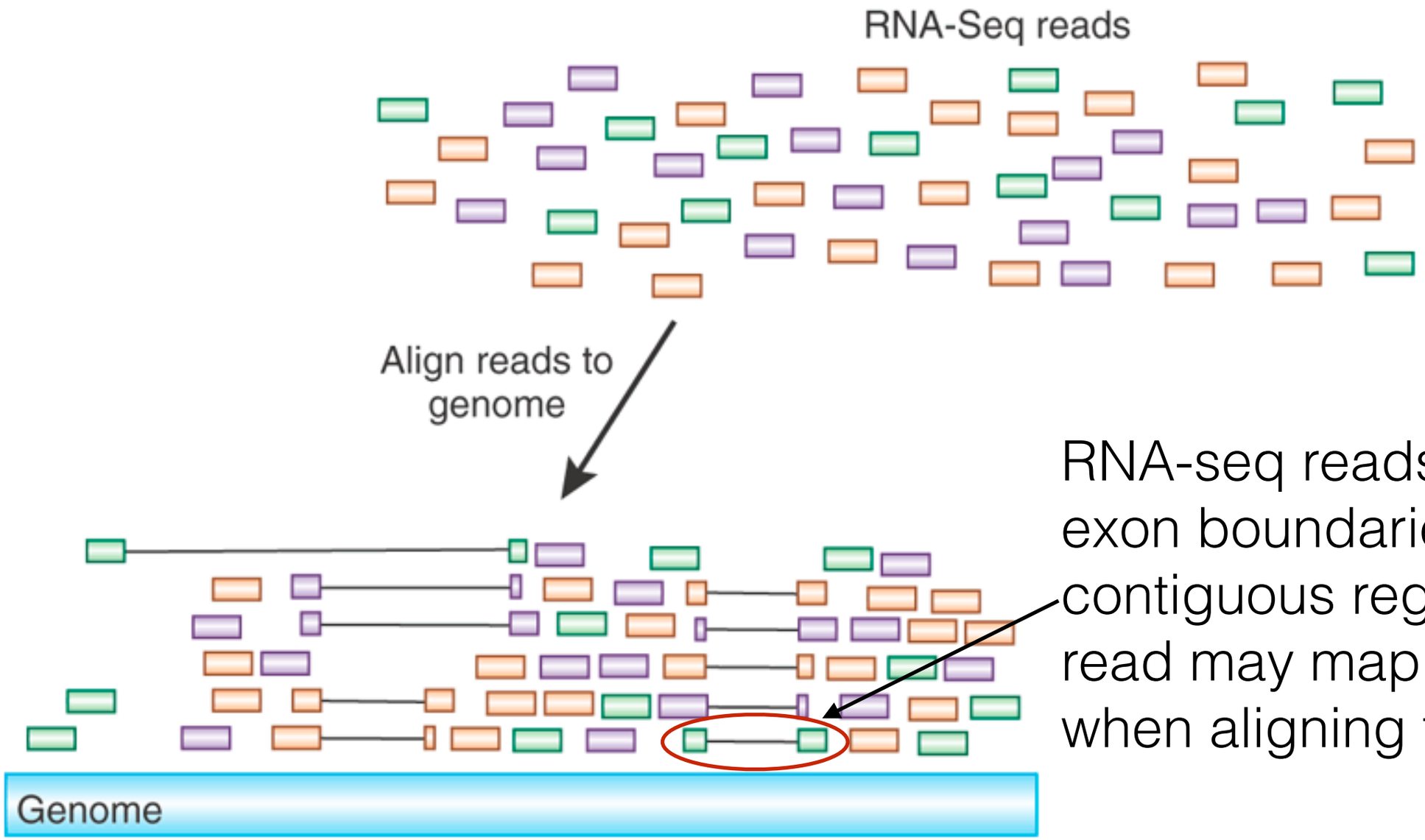


RNA-Seq reads come from a spliced transcript — if we can map them back to the genome, they give us evidence of **transcribed** regions.

Human genome contains > 14,000 pseudogenes [Pei et al. Genome Biology 2012]

Image from: Shin, Heesun, et al. "Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion." PloS one 9.3 (2014): e91041.

RNA-seq Alignment is Hard

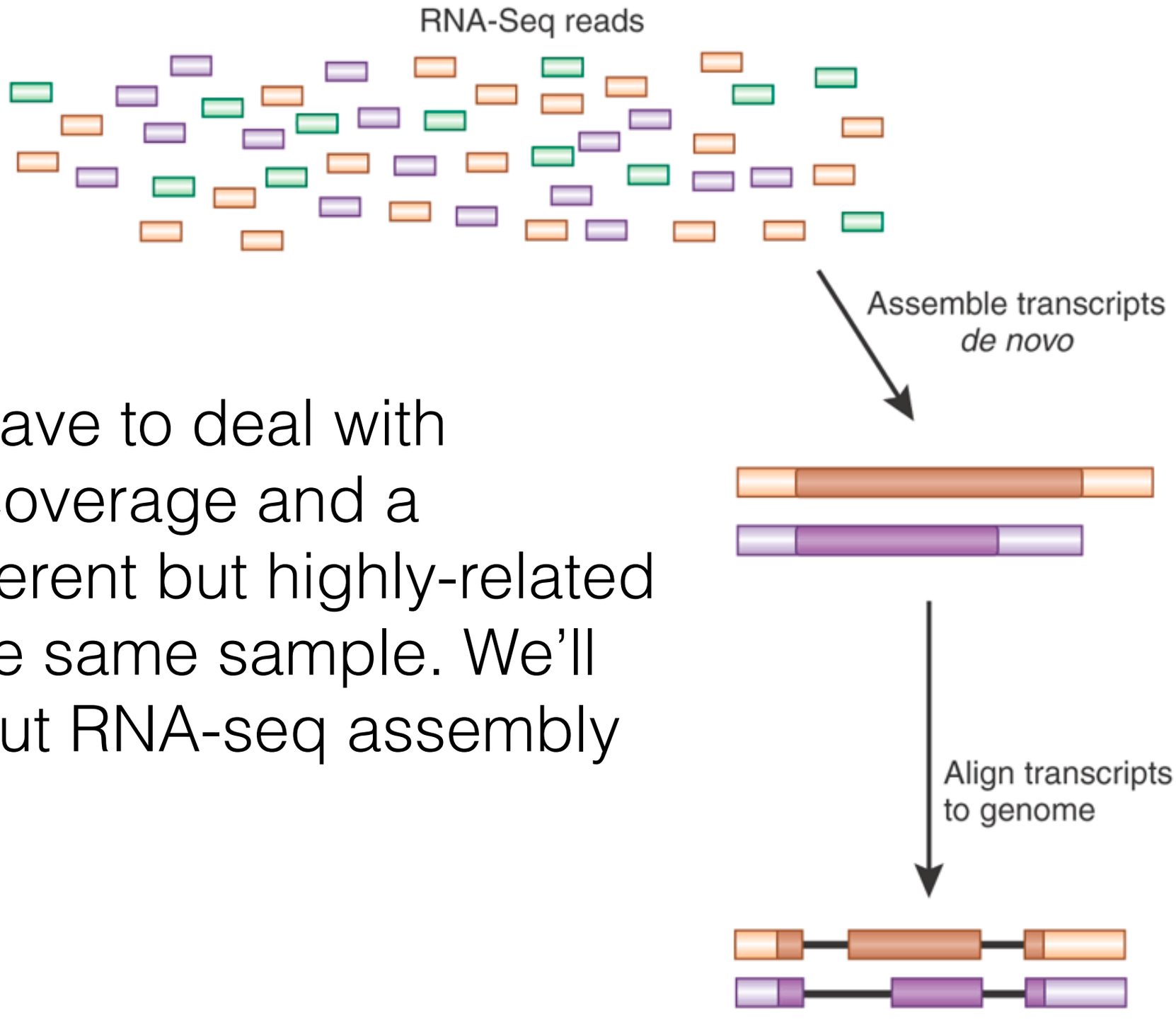


RNA-seq reads can span exon boundaries. Thus, contiguous regions of the read may map kilobases apart when aligning to the genome.

Improving the quality, sensitivity and speed of “spliced” alignment is still an active area of research. How can we be confident in a spliced-alignment when only a small portion of a read maps to an exon?

Haas, Brian J., and Michael C. Zody. "Advancing RNA-seq analysis." Nature biotechnology 28.5 (2010): 421.

RNA-seq Assembly is Harder



Assemblers have to deal with non-uniform coverage and a mixture of different but highly-related isoforms in the same sample. We'll talk more about RNA-seq assembly later.

Does Experimental Evidence Help?

There are many uses of RNA-seq apart from helping *ab initio* gene prediction.

Nonetheless, such evidence may be a powerful tool in helping us predict the existence of new genes.

Next, we'll compare the unsupervised (no experimental evidence) and semi-supervised (with the help of RNA-seq) versions of a popular eukaryotic gene predictor — GeneMark

Effect of Using Spliced-Alignments

Table 4. Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively

		D. melanogaster		*A. aegypti*		*A. gambiae*		*A. stephensi*		*Culex q.*	
		ES	ET	ES	ET	ES	ET	ES	ET	ES	ET
Internal exon	Sn	86.7	**87.2**	69.3	**91.7**	77.6	**80.4**	82.7	**85.1**	77.4	**81.8**
	Sp	76.9	**82.9**	60.7	**75.9**	70.3	**78.6**	76.5	**77.0**	54.7	**65.7**
Intron	Sn	82.6	**84.8**	67.9	**89.6**	77.6	**81.0**	85.2	**88.1**	70.2	**81.1**
	Sp	75.3	**79.2**	64.6	**80.3**	73.4	**80.5**	79.4	**81.7**	59.8	**72.7**
Donor site	Sn	85.3	**87.0**	74.6	**92.8**	81.9	**84.1**	88.2	**90.4**	74.3	**83.5**
	Sp	84.5	**86.5**	76.2	**86.8**	82.9	**88.1**	87.3	**88.1**	74.3	**80.7**
Acceptor site	Sn	86.2	**88.2**	74.3	**94.1**	83.0	**86.0**	90.7	**92.8**	83.9	**88.7**
	Sp	85.5	**87.0**	79.0	**89.6**	83.6	**88.9**	87.7	**89.2**	78.0	**84.6**
Initiation site	Sn	71.0	**75.1**	62.5	**79.6**	63.8	**68.1**	65.0	**66.9**	60.8	**76.7**
	Sp	**83.1**	81.5	77.1	**83.9**	**80.0**	79.9	73.6	**76.3**	77.4	**85.7**
Termination site	Sn	77.3	**84.2**	68.1	**88.0**	72.9	**81.0**	83.0	**84.9**	78.9	**82.8**
	Sp	**90.7**	90.0	91.3	**96.0**	89.7	**91.6**	86.5	**92.4**	89.3	**90.9**
Nucleotide	Sn	91.5	**92.1**	87.0	**98.1**	91.4	**92.9**	97.0	**97.3**	93.9	**94.4**
	Sp	**98.3**	97.4	95.2	**96.2**	98.6	**98.8**	98.5	**98.7**	92.0	**93.0**
Gene	Sn	57.9	**63.6**	40.3	**66.7**	43.8	**53.1**	43.2	**48.6**	46.1	**65.0**
	Sp	57.3	**61.0**	42.6	**64.3**	44.0	**53.0**	39.9	**47.0**	44.3	**62.6**
Partial gene	Sn	59.9	**67.2**	41.2	**69.0**	46.2	**56.0**	48.6	**54.3**	48.1	**66.1**
	Sp	59.3	**64.5**	43.6	**66.5**	46.4	**55.8**	44.9	**52.4**	46.1	**63.6**

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts.
Spliced alignments for GeneMark-ET were produced by UnSplicer.

Lomsadze et al. 2014

Effect of Using Spliced-Alignments

Table 4. Assessment of gene prediction accuracy of GeneMark-ES (ES) and GeneMark-ET (ET) gene finders using unsupervised (genomic based) and semi-supervised (genomic and transcriptomic based) training, respectively

		D. melanogaster		*A. aegypti*		*A. gambiae*		*A. stephensi*		*Culex q.*	
		ES	ET	ES	ET	ES	ET	ES	ET	ES	ET
Internal exon	Sn	86.7	**87.2**	69.3	**91.7**	77.6	**80.4**	82.7	**85.1**	77.4	**81.8**
	Sp	76.9	**82.9**	60.7	**75.9**	70.3	**78.6**	76.5	**77.0**	54.7	**65.7**
Intron	Sn	82.6	**84.8**	67.9	**89.6**	77.6	**81.0**	85.2	**88.1**	70.2	**81.1**
	Sp	75.3	**79.2**	64.6	**80.3**	73.4	**80.5**	79.4	**81.7**	59.8	**72.7**
Donor site	Sn	85.3	**87.0**	74.6	**92.8**	81.9	**84.1**	88.2	**90.4**	74.3	**83.5**
	Sp	84.5	**86.5**	76.2	**86.8**	82.9	**88.1**	87.3	**88.1**	74.3	**80.7**
Acceptor site	Sn	86.2	**88.2**	74.3	**94.1**	83.0	**86.0**	90.7	**92.8**	83.9	**88.7**
	Sp	85.5	**87.0**	79.0	**89.6**	83.6	**88.9**	87.7	**89.2**	78.0	**84.6**
Initiation site	Sn	71.0	**75.1**	62.5	**79.6**	63.8	**68.1**	65.0	**66.9**	60.8	**76.7**
	Sp	**83.1**	81.5	77.1	**83.9**	**80.0**	79.9	73.6	**76.3**	77.4	**85.7**
Termination site	Sn	77.3	**84.2**	68.1	**88.0**	72.9	**81.0**	83.0	**84.9**	78.9	**82.8**
	Sp	**90.7**	90.0	91.3	**96.0**	89.7	**91.6**	86.5	**92.4**	89.3	**90.9**
Nucleotide	Sn	91.5	**92.1**	87.0	**98.1**	91.4	**92.9**	97.0	**97.3**	93.9	**94.4**
	Sp	**98.3**	97.4	95.2	**96.2**	98.6	**98.8**	98.5	**98.7**	92.0	**93.0**
Gene	Sn	57.9	**63.6**	40.3	**66.7**	43.8	**53.1**	43.2	**48.6**	46.1	**65.0**
	Sp	57.3	**61.0**	42.6	**64.3**	44.0	**53.0**	39.9	**47.0**	44.3	**62.6**
Partial gene	Sn	59.9	**67.2**	41.2	**69.0**	46.2	**56.0**	48.6	**54.3**	48.1	**66.1**
	Sp	59.3	**64.5**	43.6	**66.5**	46.4	**55.8**	44.9	**52.4**	46.1	**63.6**

Bold font highlights the higher accuracy value in a given category and given species. Partial gene level accuracy is computed without taking into account a difference in annotation and prediction of translation starts.
Spliced alignments for GeneMark-ET were produced by UnSplicer.

Lomsadze et al. 2014

Other Ways of Using Experimental Evidence

Experimental evidence (RNA-seq, in particular) is a great help in improving gene prediction. However, its uses stretch **far** beyond assisting *ab initio* gene prediction.

Transcript quantification

Differential expression, alternative splicing analysis

Fusion/chimera detection

Variant (SNP, SV, CNV) detection

Transcript assembly

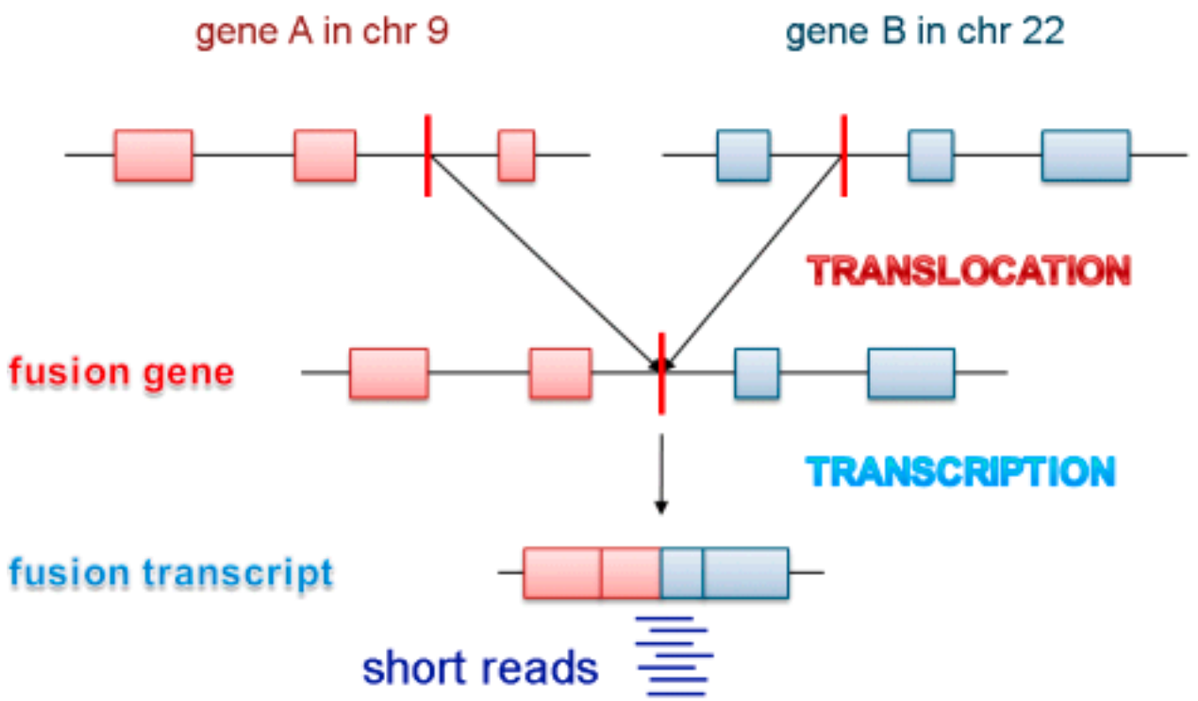
Genome guided & de novo

Build higher-level models of transcription

co-expression networks -> regulatory networks

Other Ways of Using Experimental Evidence

Fusion/chimera detection



Variant (SNP, SV, CNV) detection

Find small (SNP) or large (SV) variation in how read map back to their genes of origin

Find differences in the number of copies of a gene in the DNA (CNV)

image from: http://biome.ewha.ac.kr:8080/FusionGene/

Other Ways of Using Experimental Evidence

Transcript assembly

With sufficiently deep sequencing, we can hope to assemble transcripts present in an experiment in a manner similar to how we assemble DNA.

This often lets us find previously undiscovered genes, as well as novel splice variants (combination of exons that make up an isoform of the gene).

Genome guided and *de novo*

Assembly can either rely on knowing the reference genome (making the problem much easier), or can be done directly from the RNA-seq reads without the reference (or via hybrid approaches).

Other Ways of Using Experimental Evidence

Build higher-level models of transcription

Co-expression networks -> regulatory networks

By looking at how the expression of different genes covaries across many different experimental conditions and tissue types, we can begin to view the set of genes as a network.

Which genes tend to be “turned on” when others are “turned off”.

Use such information to try and determine regulatory relationships between genes — which genes control others, and how.

Transcript quantification

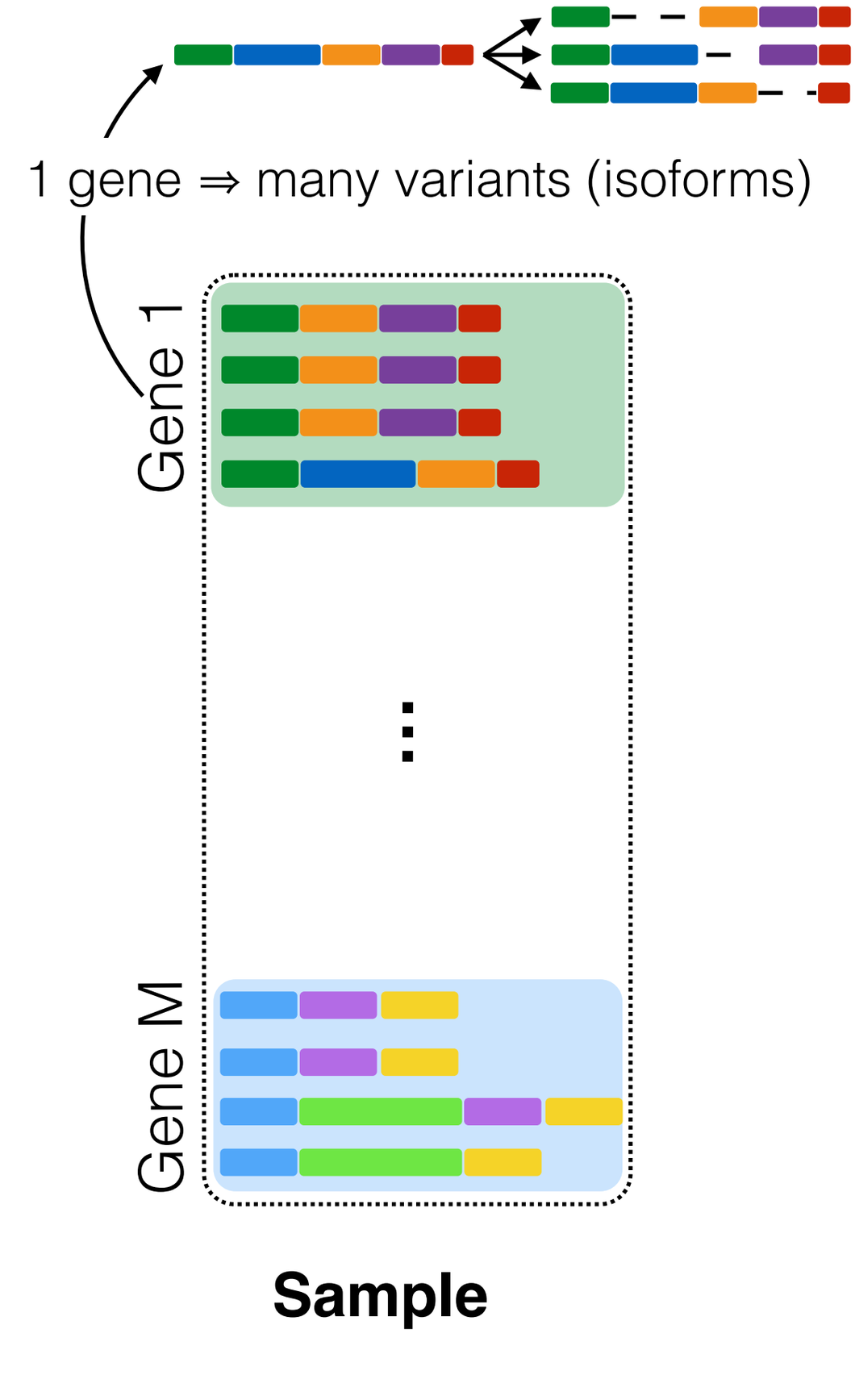
There are a number of tools that assume a set of known “targets”, and then quantify the abundance of these targets given RNA-seq data.

Such tools are very useful in organisms with well characterized transcriptomes, or when dealing with *de novo* assembly, since we know the set of targets (the output of the assembly) that we wish to quantify.

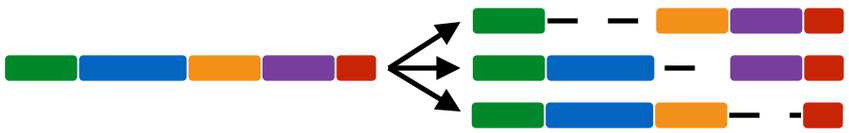
In this setting, we align / map reads directly to these targets. Multi-mapping is prevalent among transcripts of a gene, and among paralogous genes.

This multi-mapping must be resolved — ignoring it leads to incorrect results!

Abundance Estimation: An Overview



Abundance Estimation: An Overview



1 gene ⇒ many variants (isoforms)

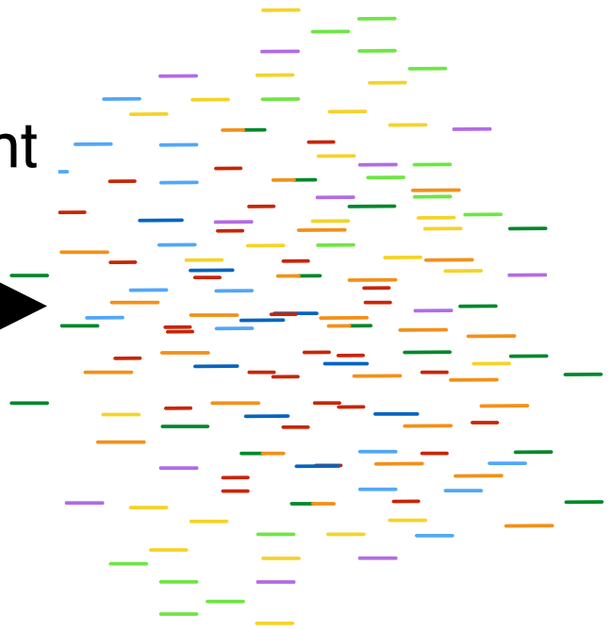
Gene 1

Gene M

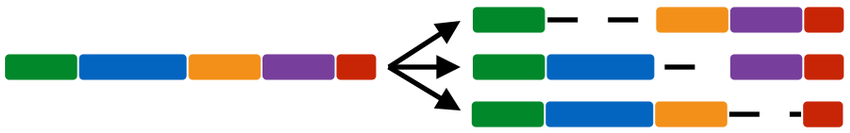
Sample

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

Measurement
(RNA-seq)



Abundance Estimation: An Overview



1 gene ⇒ many variants (isoforms)

Gene 1

Gene M

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

Measurement
(RNA-seq)

Inference

% Gene 1

% Gene M

isoform A

isoform B

isoform C

Sample

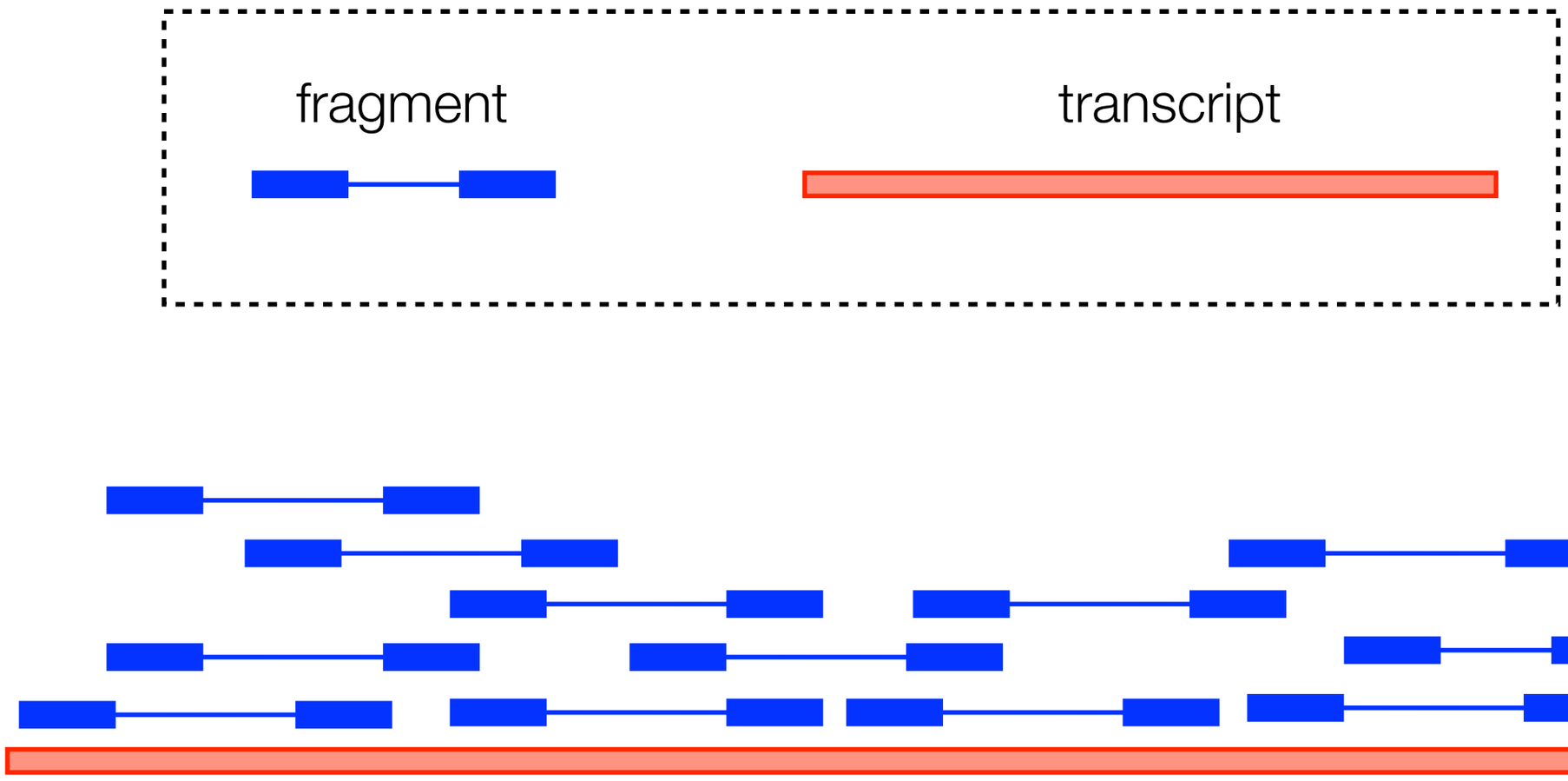
Abundance Estimates

Abundance Estimation: An Overview

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

Estimate: The relative abundance of each transcript

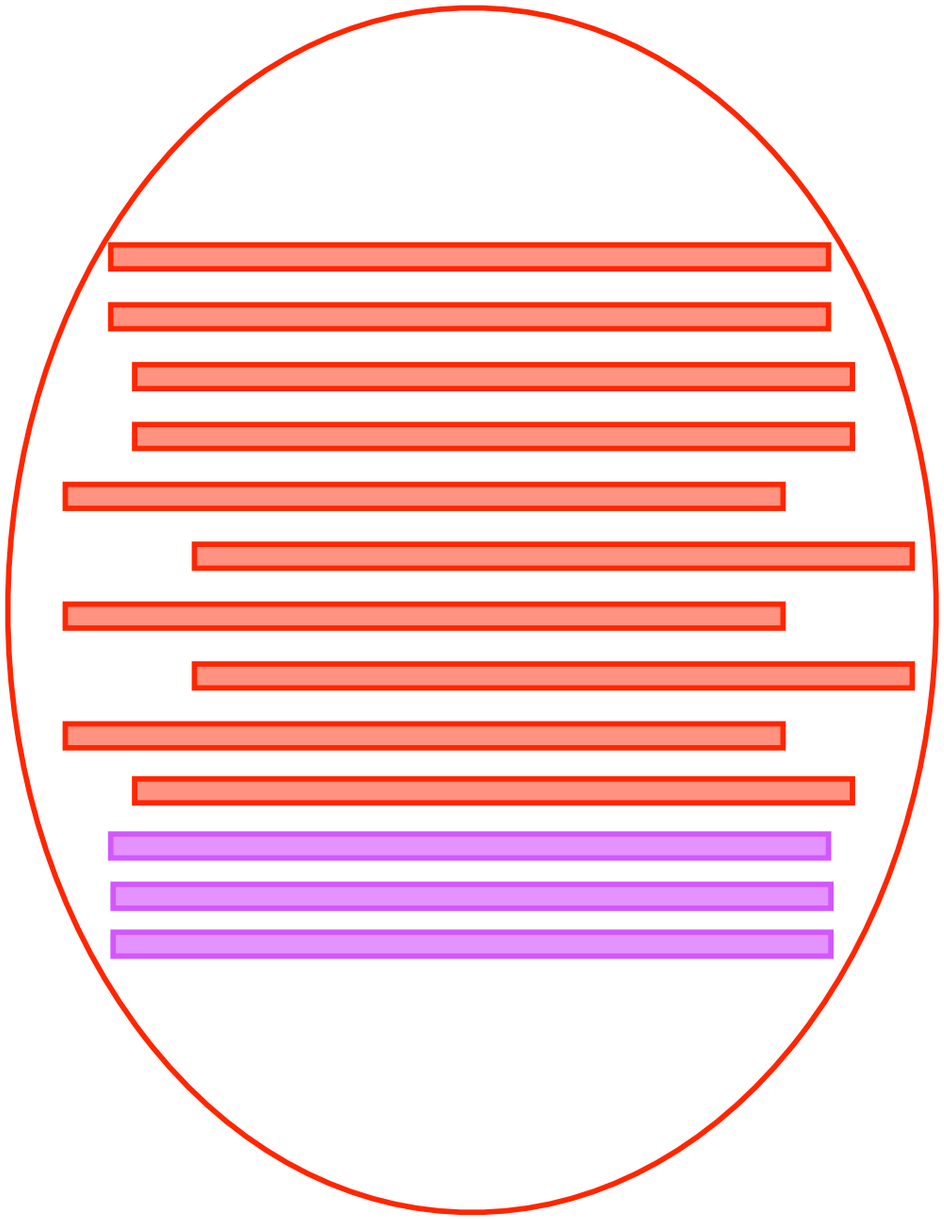
Basic principles of quantification



slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Basic principles of quantification

The **more abundant** a transcript is, the more fragments we'll sequence from it



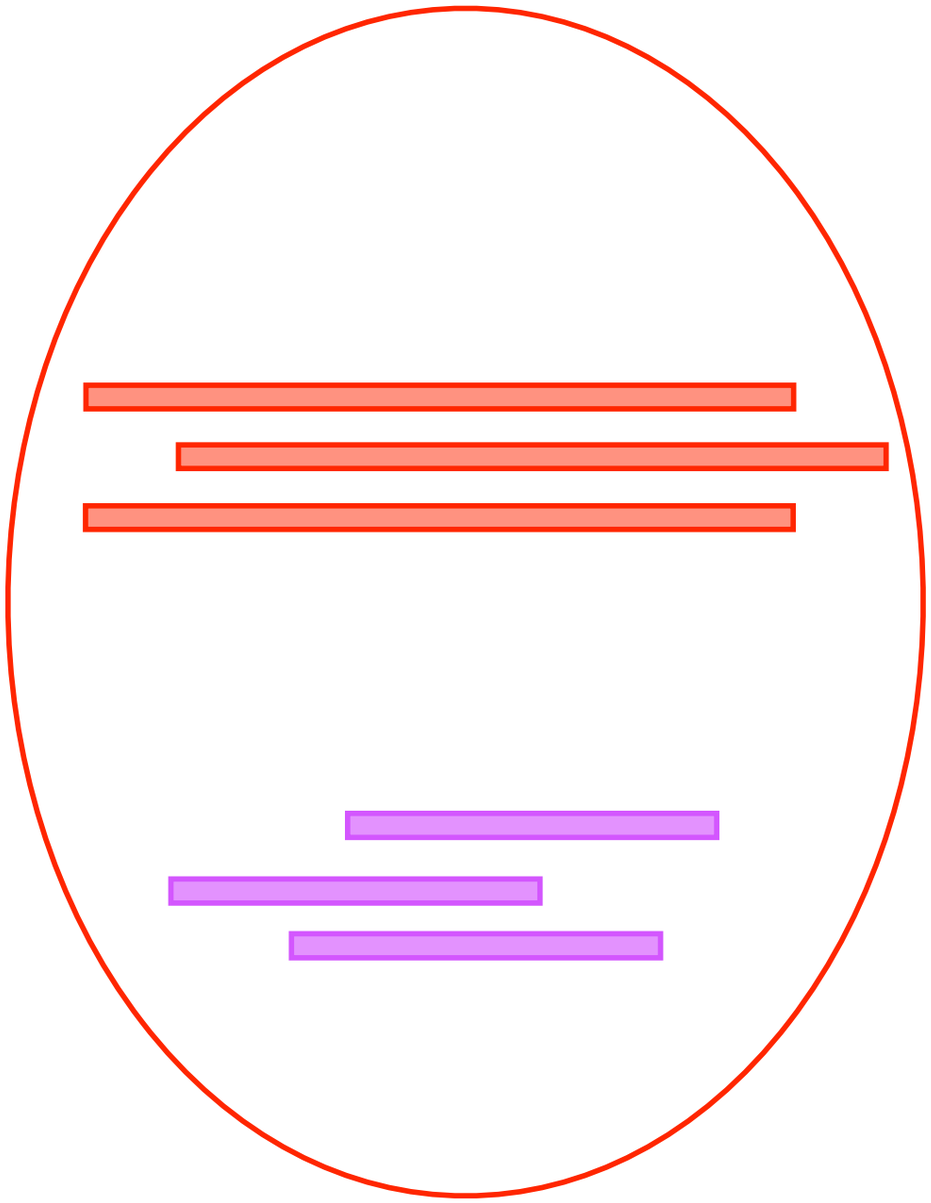
Transcriptome

Sequenced fragments

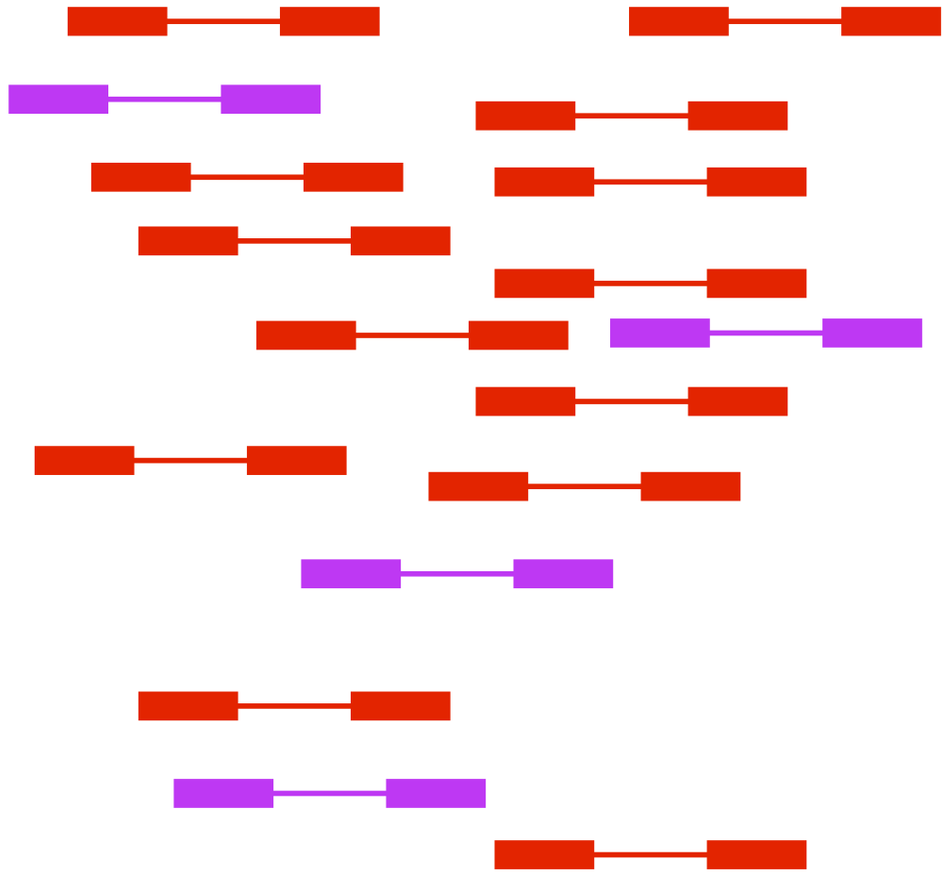
slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Basic principles of quantification

The **longer** a transcript is, the more fragments we'll sequence from it



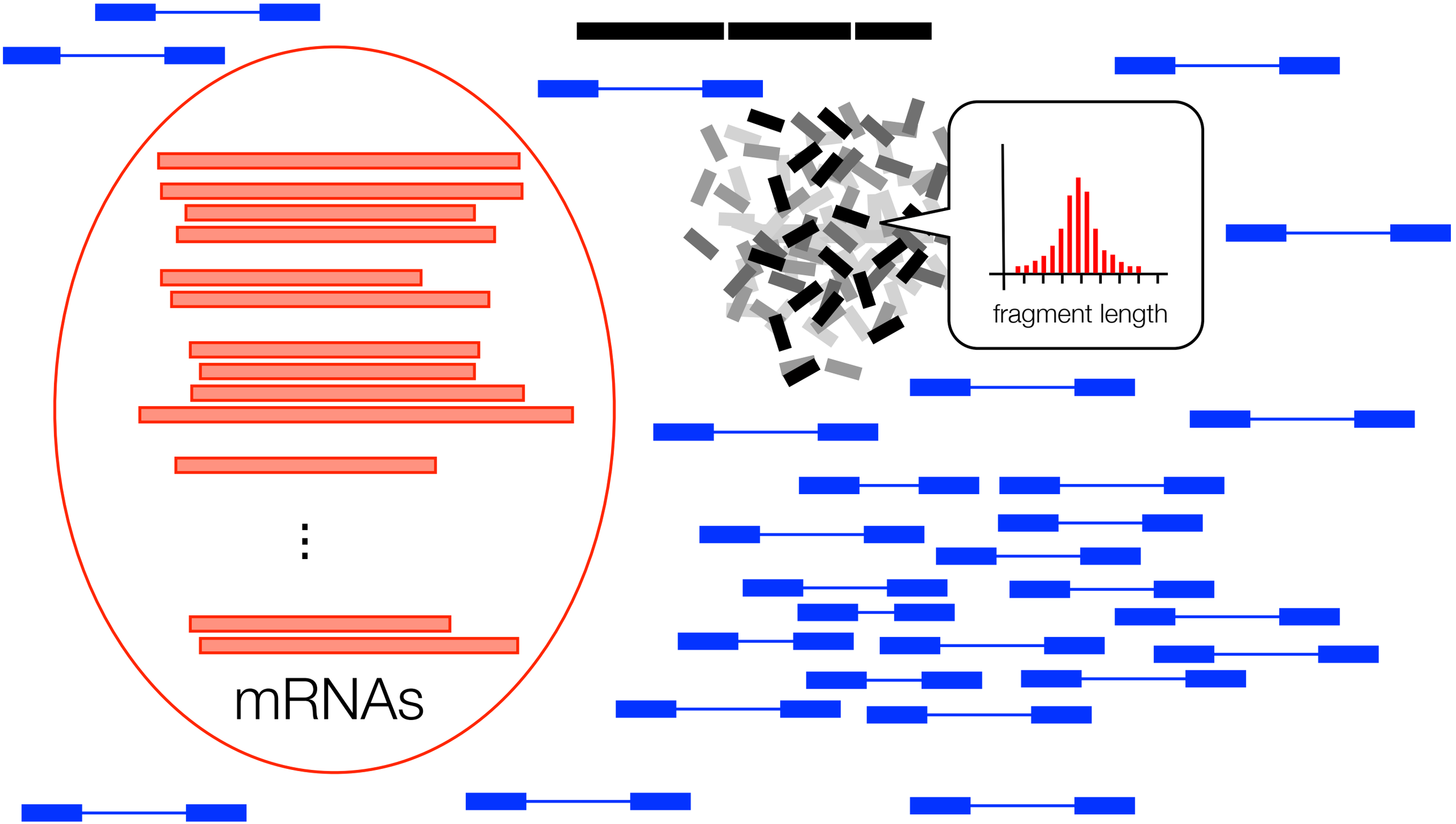
Transcriptome



Sequenced fragments

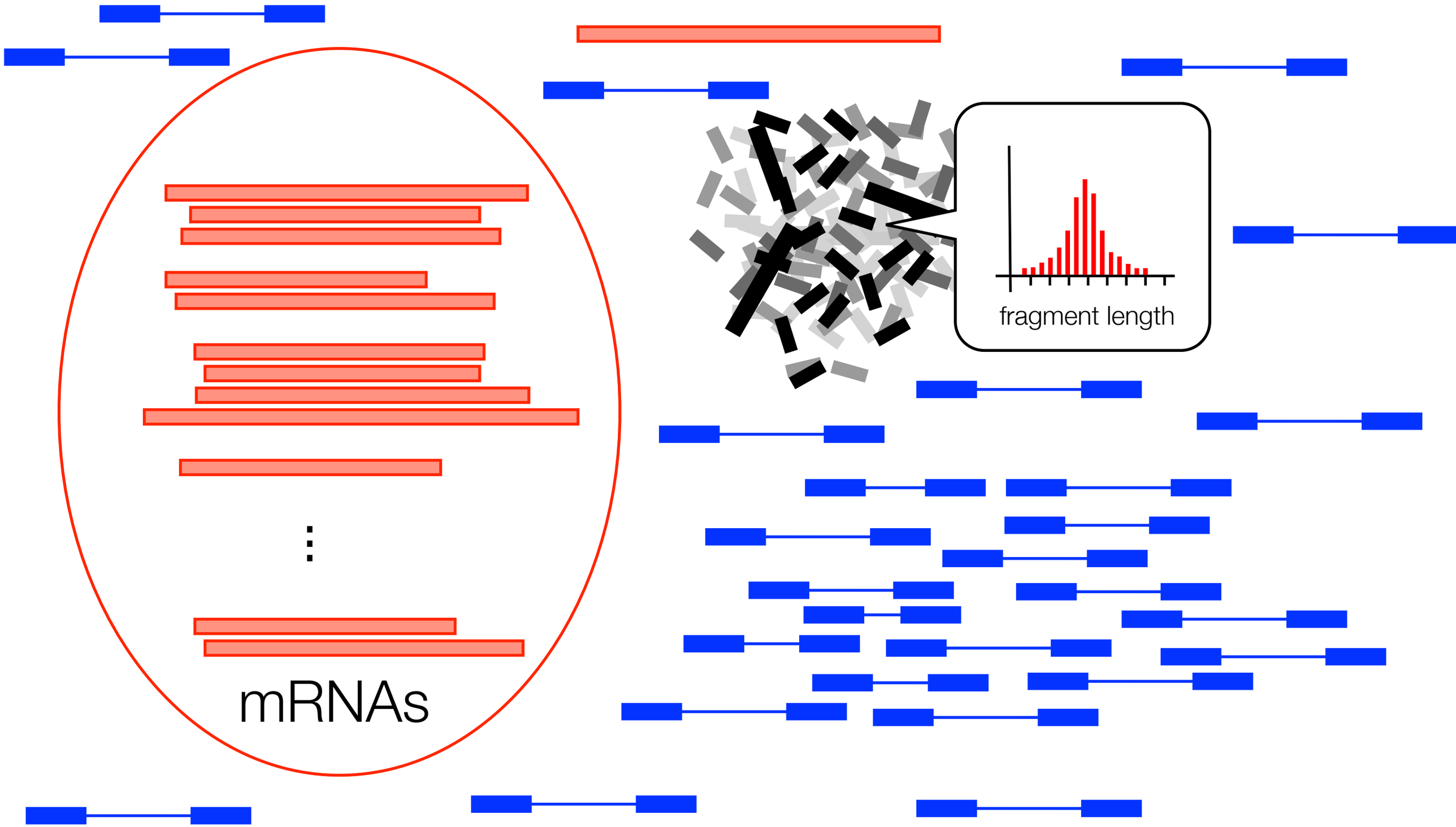
slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Basic principles of quantification



slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Basic principles of quantification



slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

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

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\ell_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\ell_i N} \cdot 10^9$$

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

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\ell_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\ell_i N} \cdot 10^9$$

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

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\ell_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\ell_i N} \cdot 10^9$$

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

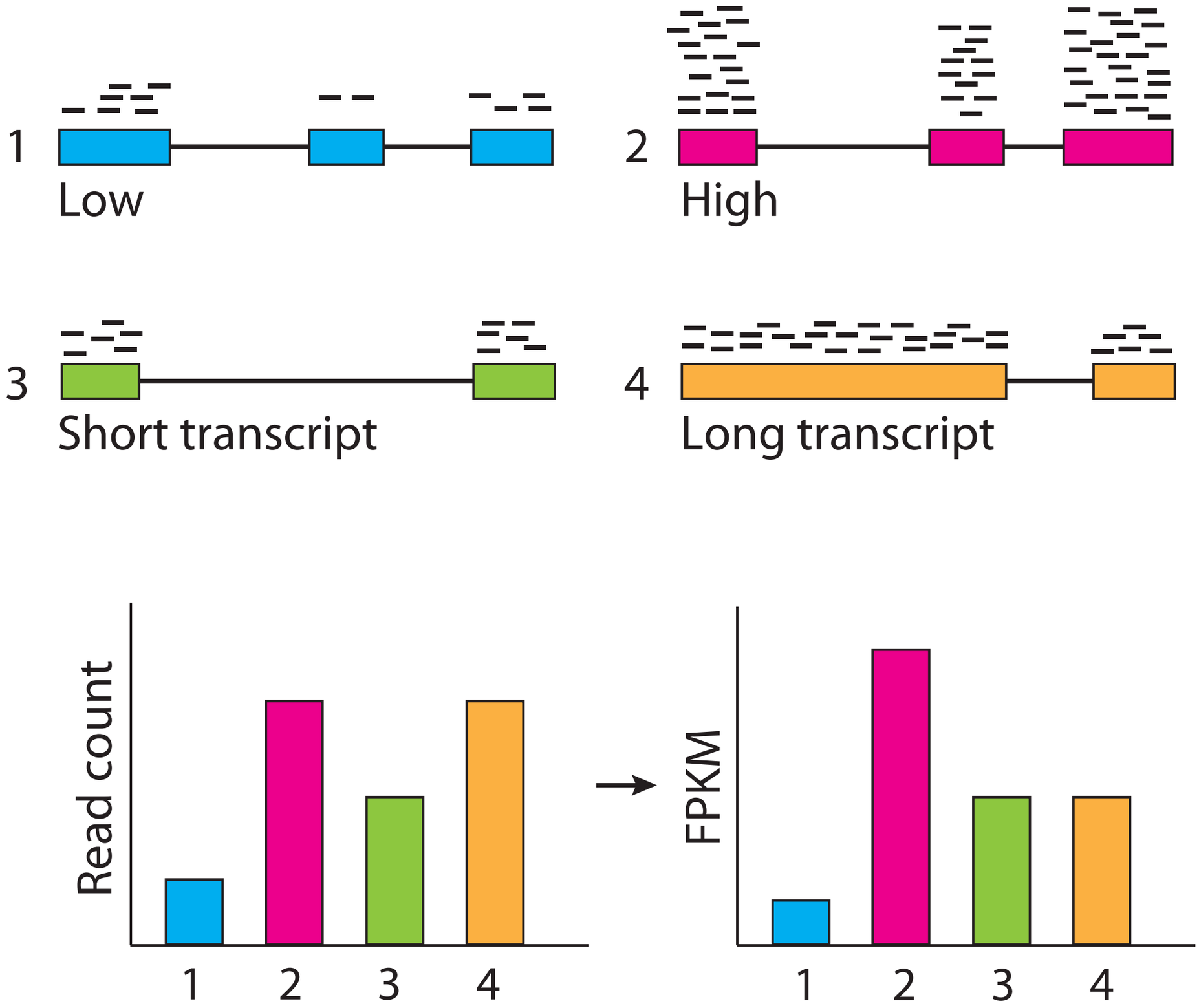
Length of transcript i

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

$$\mathrm{FPKM}_i = \frac{X_i}{\left(\frac{\ell_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\ell_i N} \cdot 10^9$$

Total number of mapped reads

Calculating expression of genes and transcripts



slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

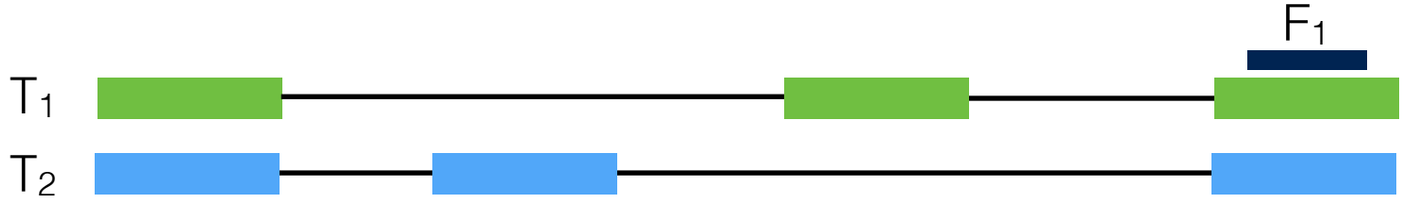
The difficulty is in estimating X_i

All equations on the previous slides assume we know the value of X_i — the number of reads originating from transcript i.

This is not as easy as it seems; multi-mapping reads are a major confounder:

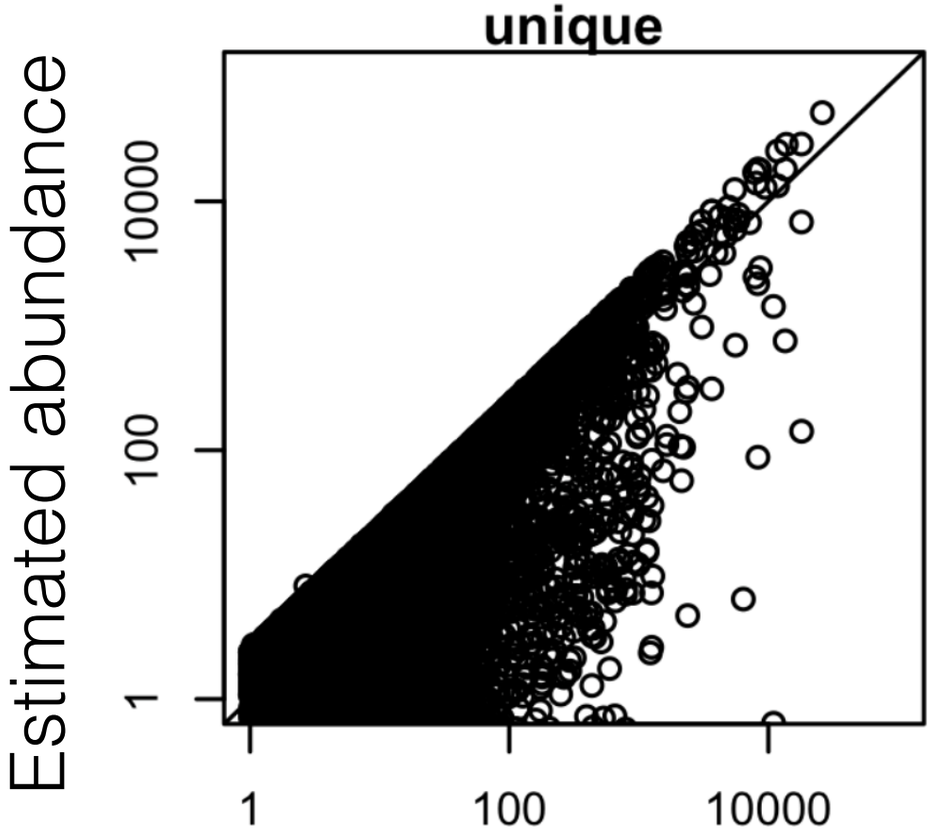
Different transcripts may share much of the same sequence (e.g. shared exons) — how do we assign a fragment in such situations?

Even without isoforms, such problems arise from similar / related gene families.

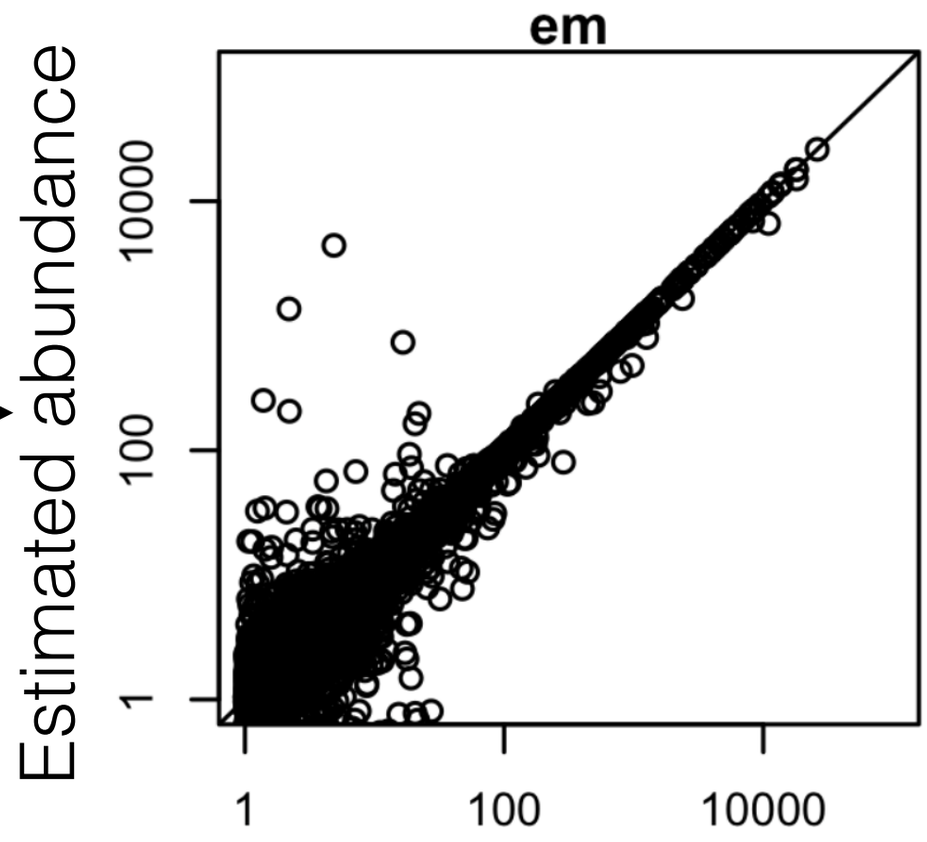


A simple (and wrong) approach

What if we consider only reads that map uniquely to a single transcript?



The "right" approach achieves this

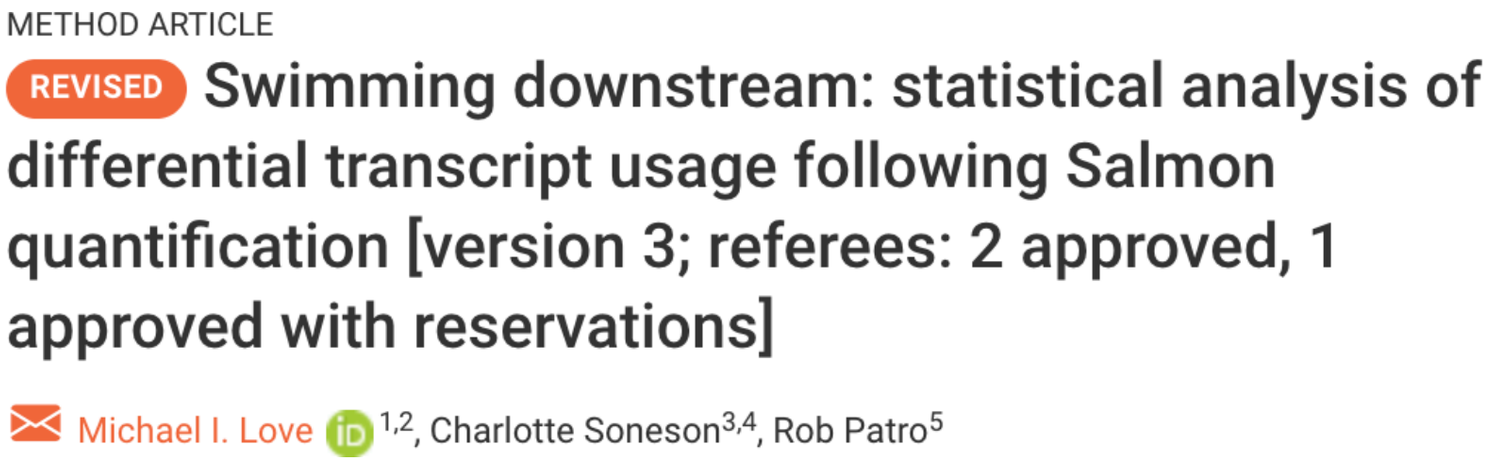


Li, Bo, et al. "RNA-Seq gene expression estimation with read mapping uncertainty." *Bioinformatics* 26.4 (2010): 493-500.

Transcript Quantification: Why?

- Even if a gene's total transcriptional output remains the same, the distributions of expressed isoforms may change (Differential Transcript Usage, DTU)

- DTU is common, and *can* be biologically meaningful.
  - response to condition / disease / stimulus



METHOD ARTICLE

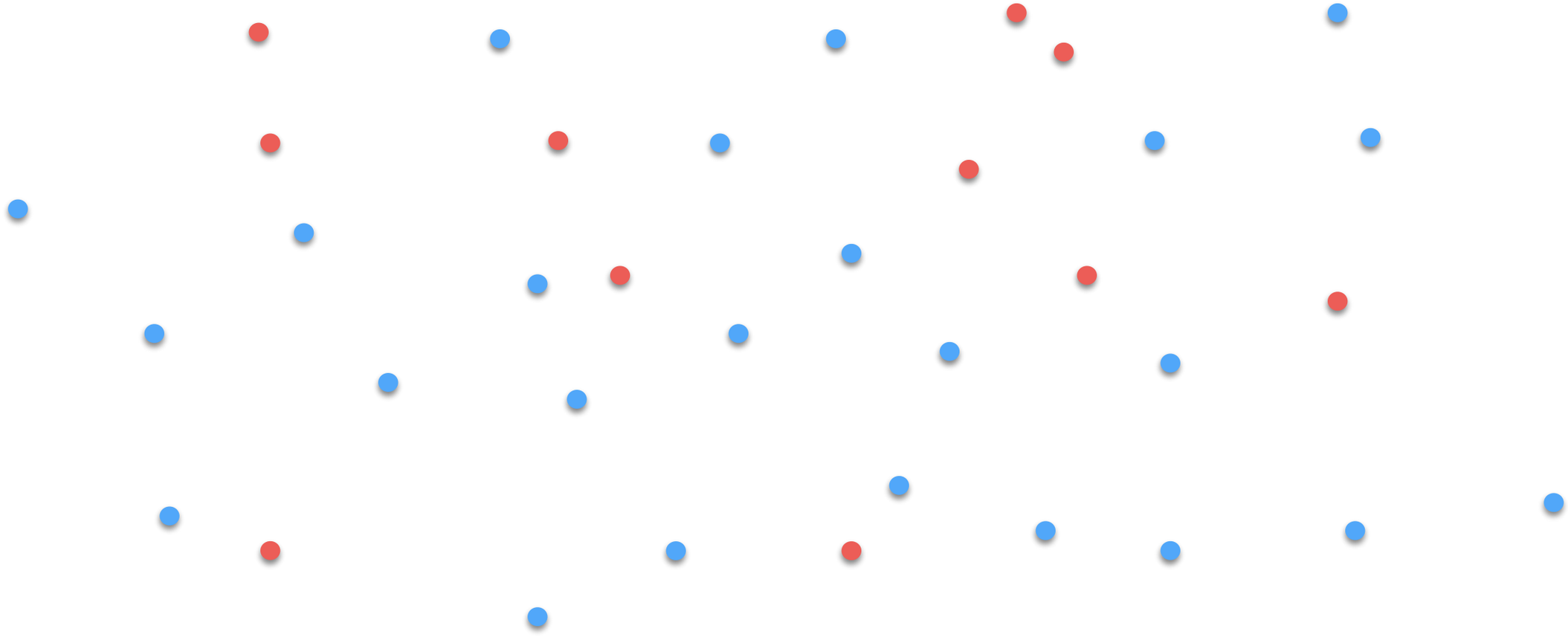
REVISED **Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification [version 3; referees: 2 approved, 1 approved with reservations]**

Michael I. Love [1,2], Charlotte Soneson[3,4], Rob Patro[5]

- Even to estimate gene-level expression well, you have to address the transcript quantification problem

First, consider this non-Biological example

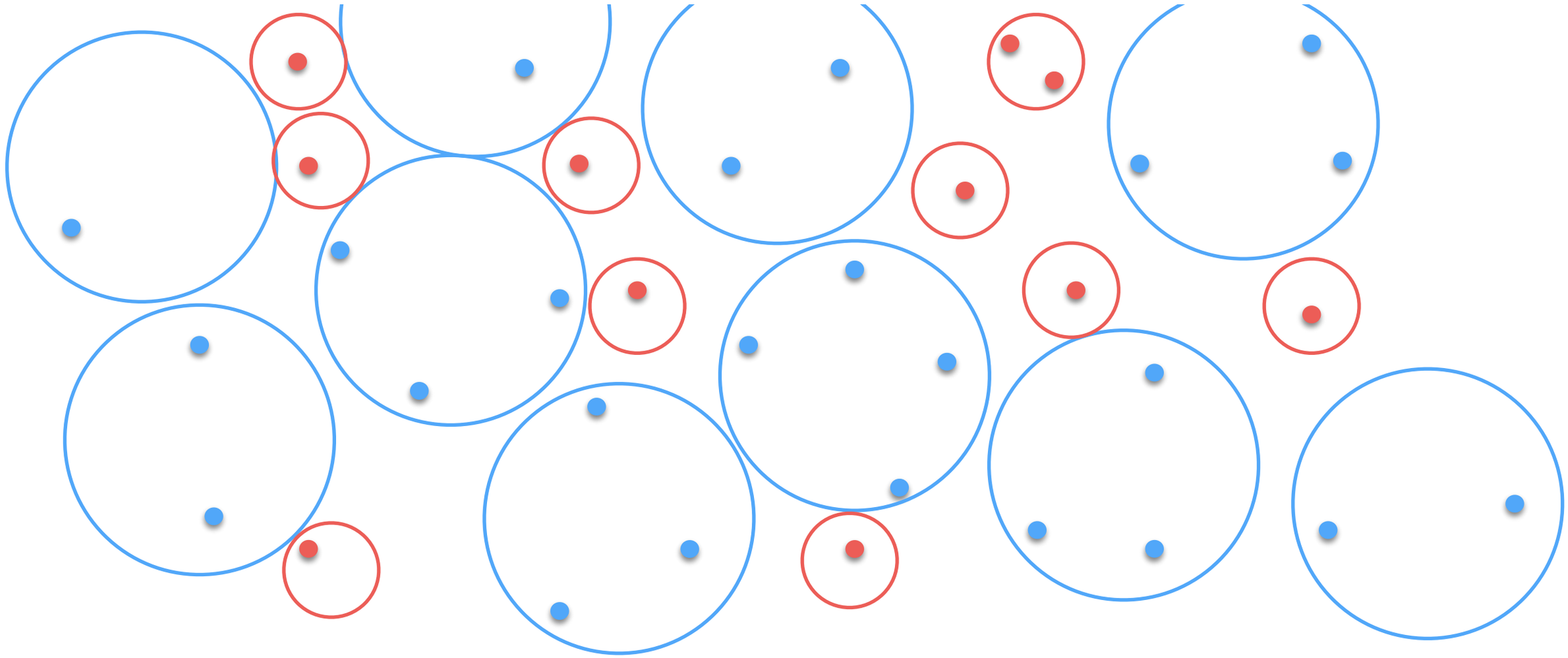
Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



Here, a dot of a color means I hit a circle of that color.
What type of circle is more prevalent?
What is the fraction of red / blue circles?

First, consider this non-Biological example

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



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

The areas!

First, consider this non-Biological example

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

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

The areas!

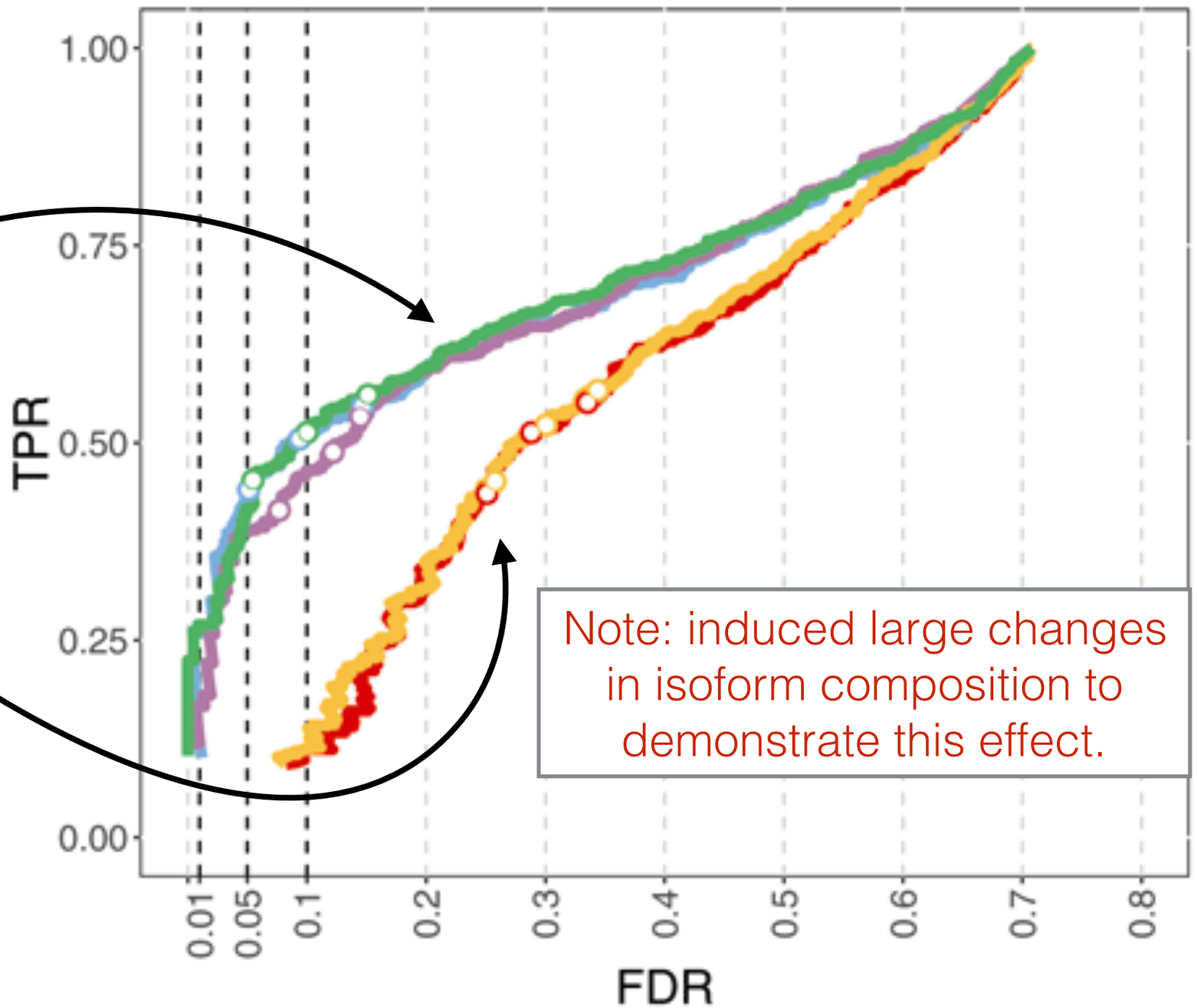
There is an analog in RNA-seq, one needs to know the **length** of the target from which one is drawing to meaningfully assess abundance!

Resolving multi-mapping is fundamental to quantification

These errors can affect DGE calls

Variants of Salmon

Variants of “counting”

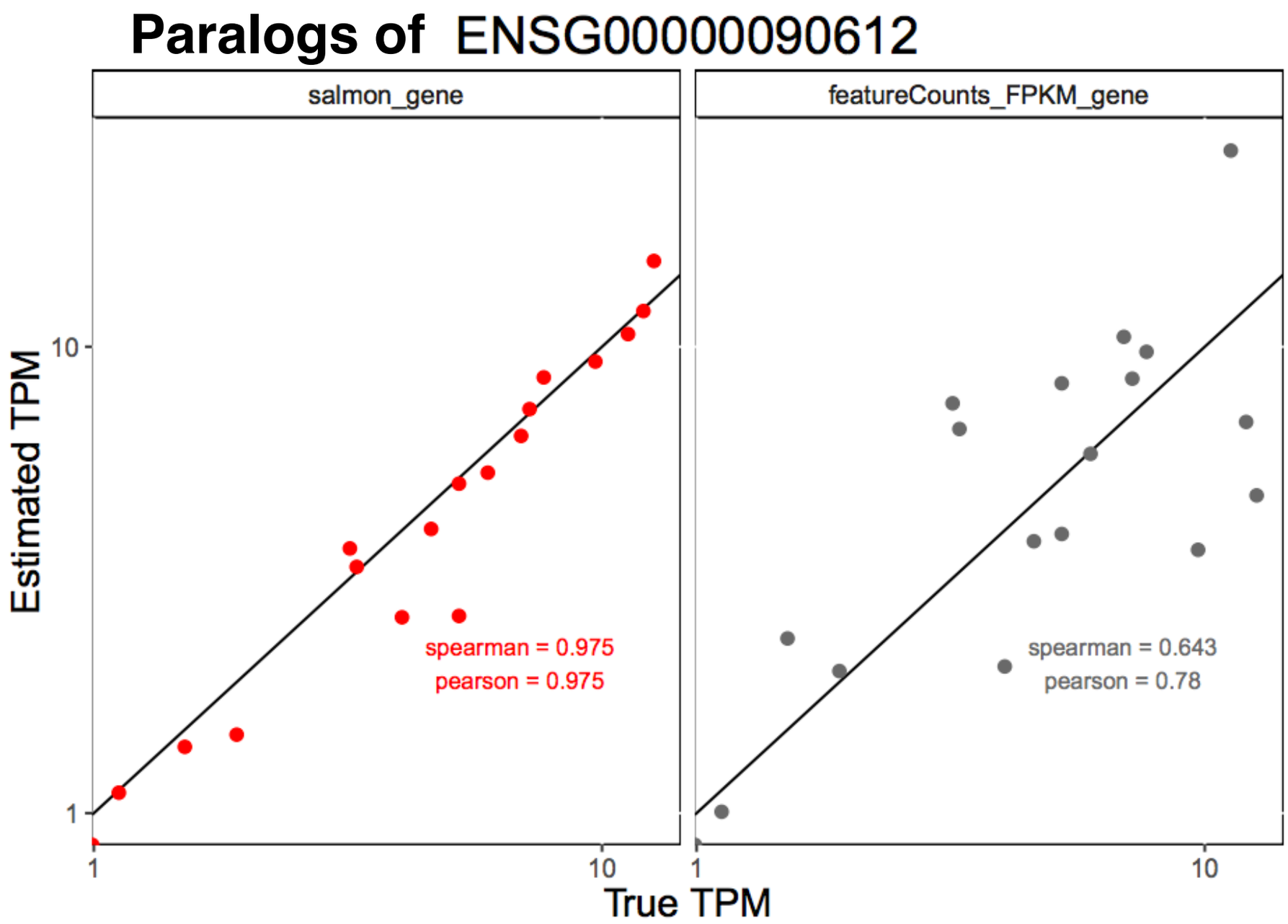


Note: induced large changes in isoform composition to demonstrate this effect.

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

Resolving multi-mapping is fundamental to quantification

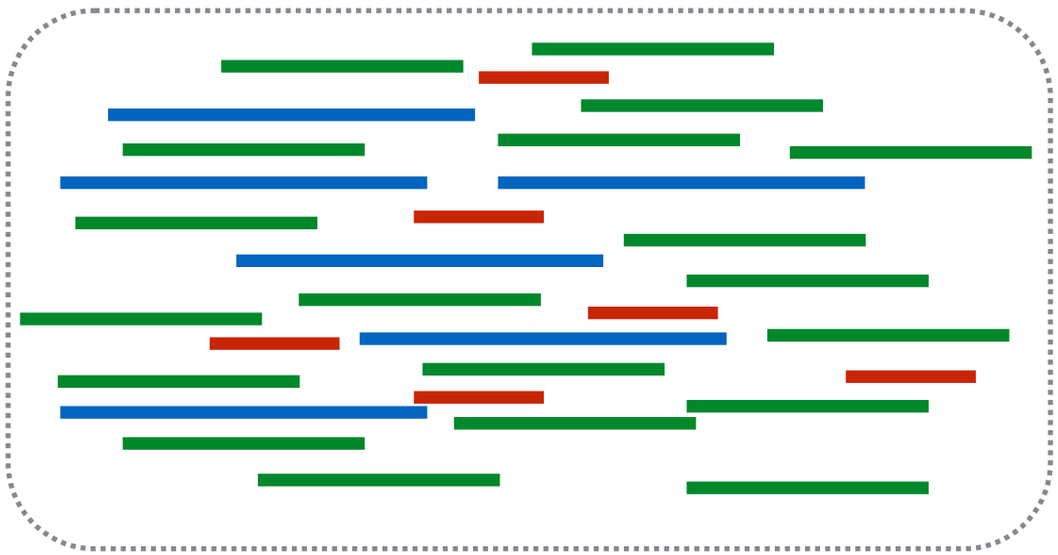
Can even affect abundance estimation in **absence** of alternative-splicing (e.g. paralogous genes)



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

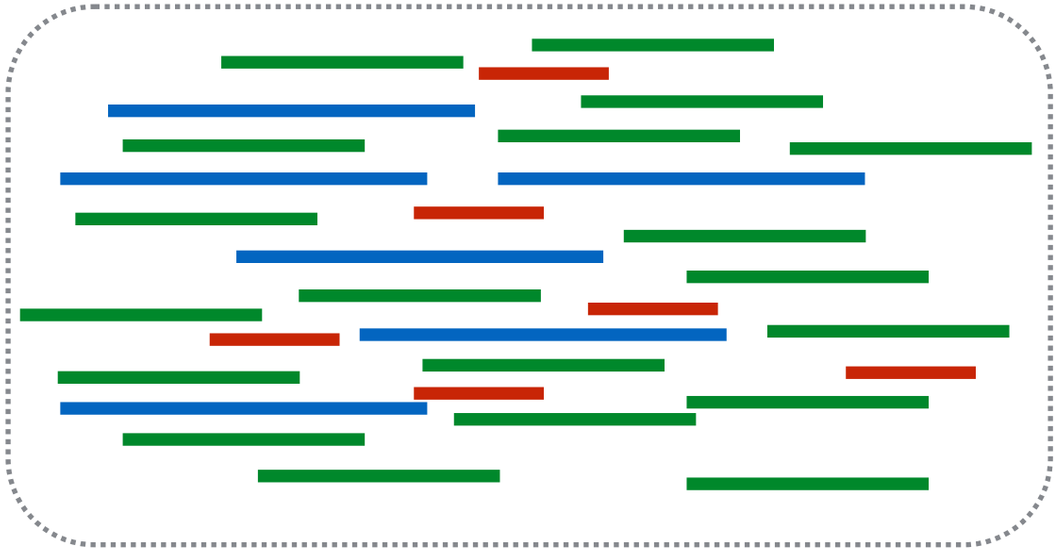


In an unbiased experiment, sampling fragments depends on:

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

How can we perform inference from sequenced fragments?

Experimental Mixture



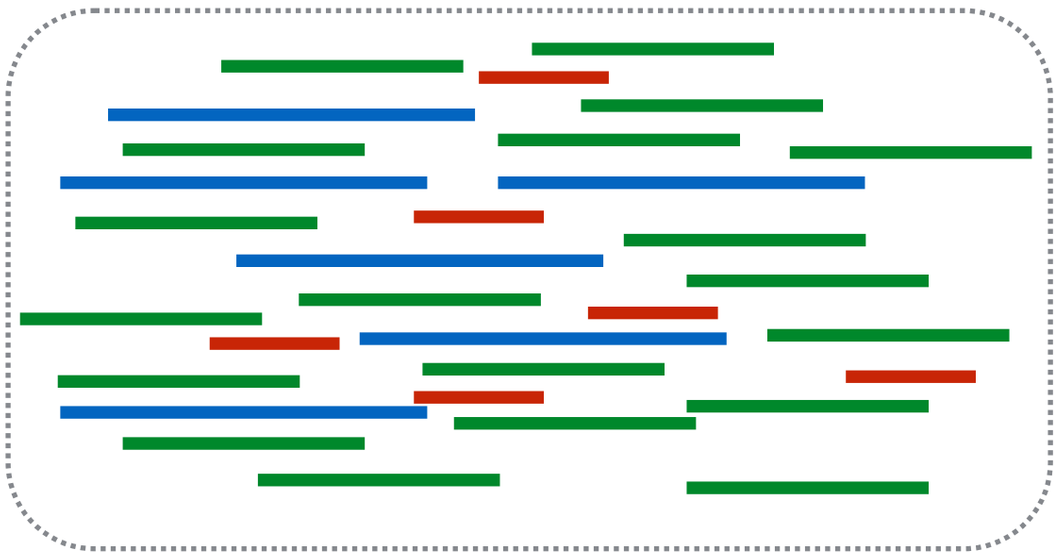
length( ) = 100

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



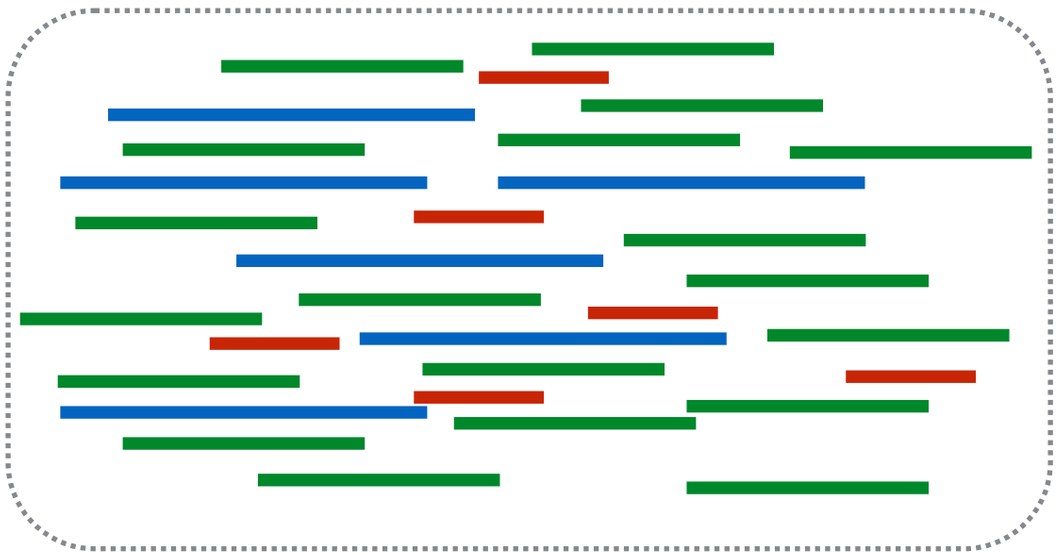
In an unbiased experiment, sampling fragments depends on:

- # 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



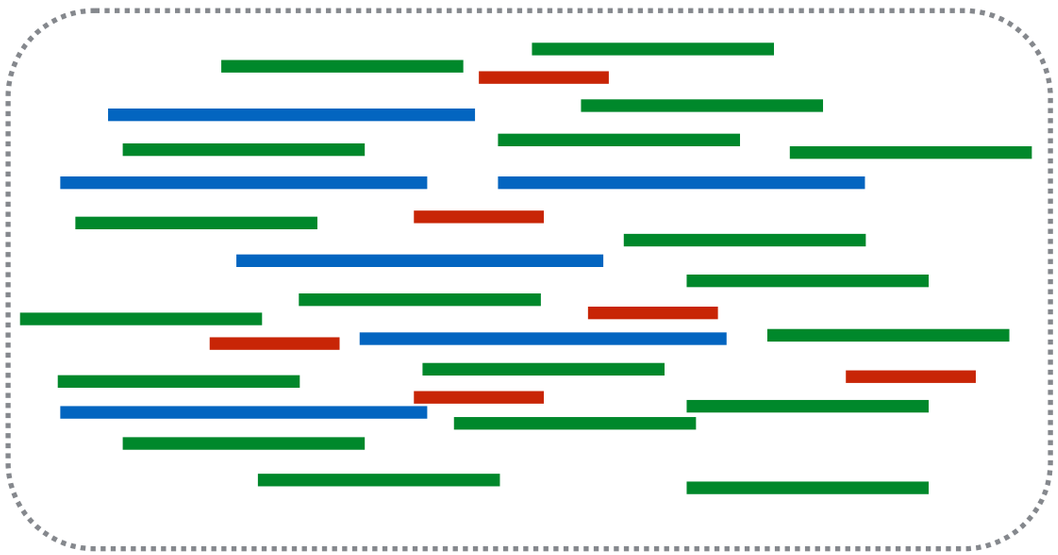
In an unbiased experiment, sampling fragments depends on:

- # 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

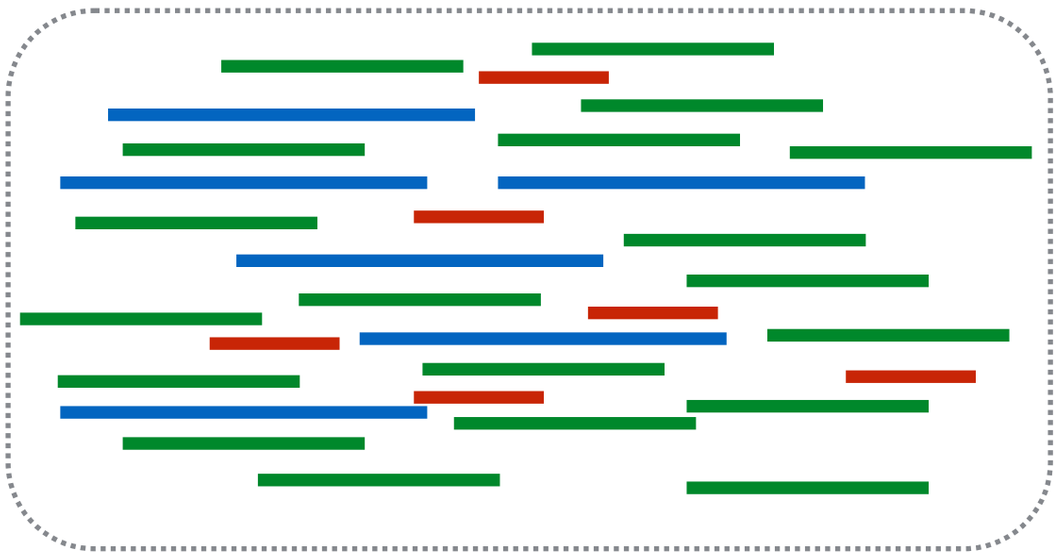
length( ) = 100 x 6 copies = 600 nt

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

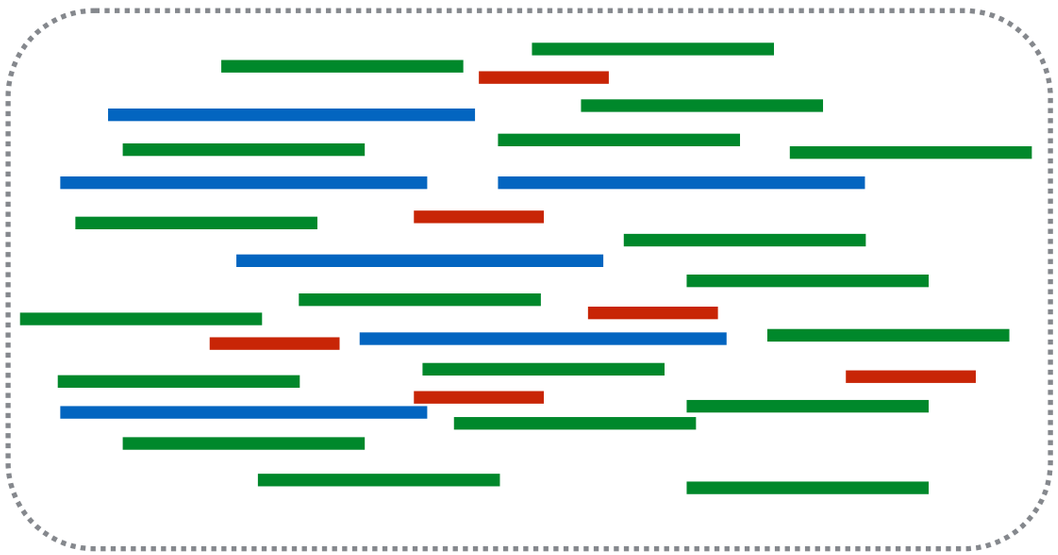
length( ) = 100 x 6 copies = 600 nt ~ 30% blue

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

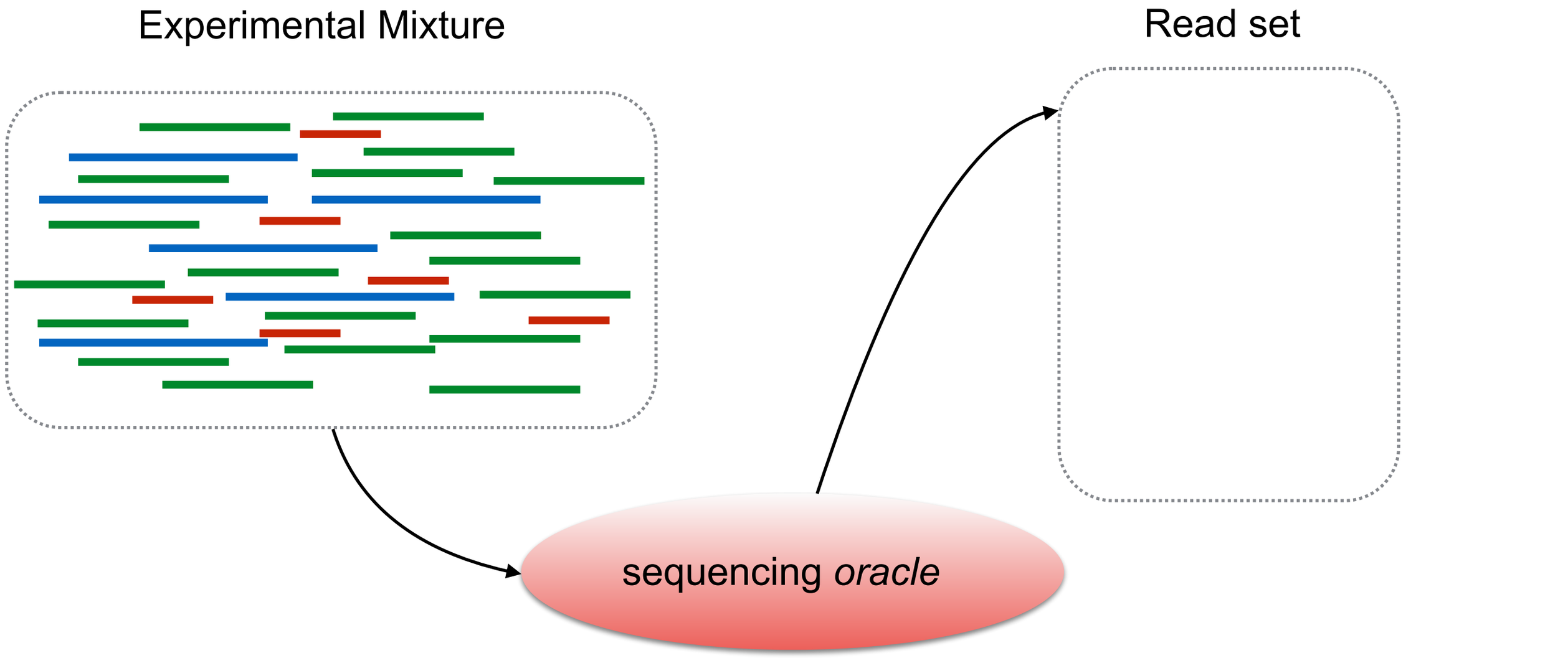
length(———) = 66 x 19 copies = 1254 nt ~ 60% green

length(—) = 33 x 6 copies = 198 nt ~ 10% red

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

How can we perform inference from sequenced fragments?

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

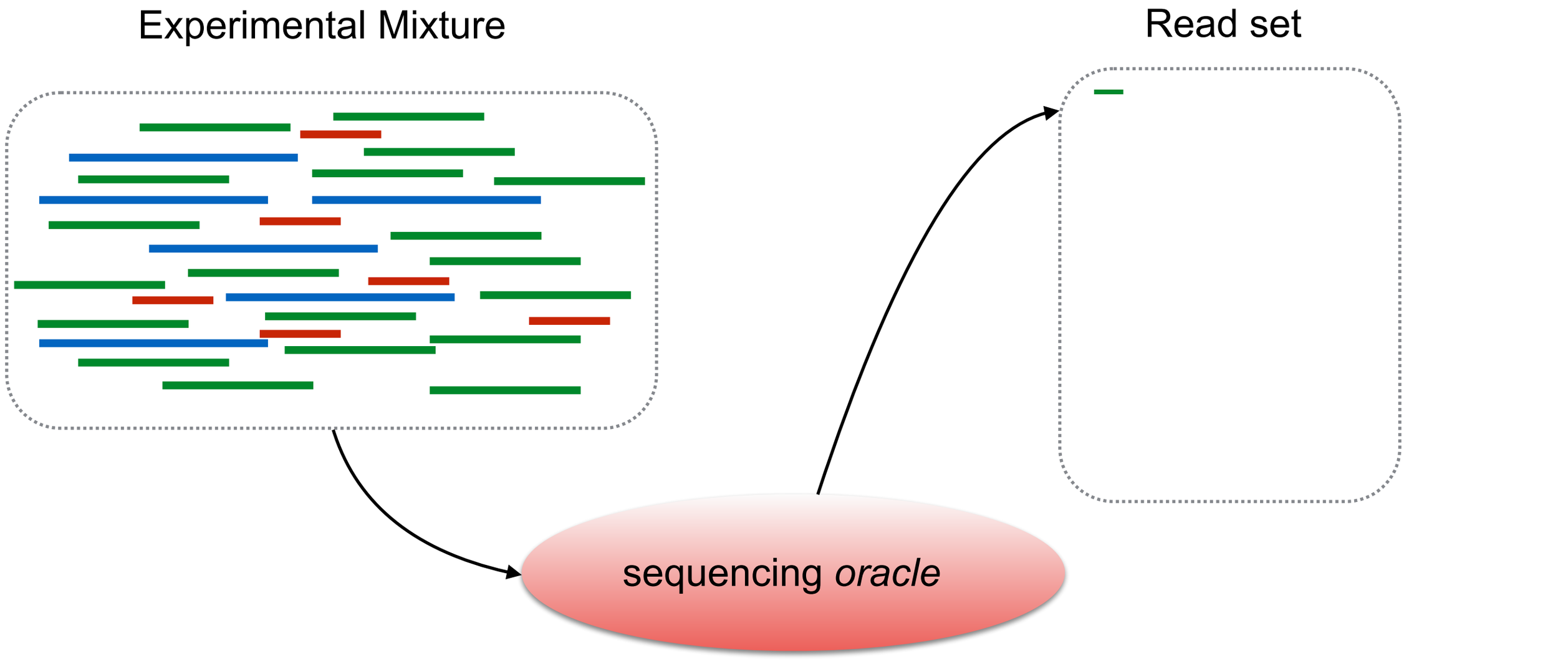


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

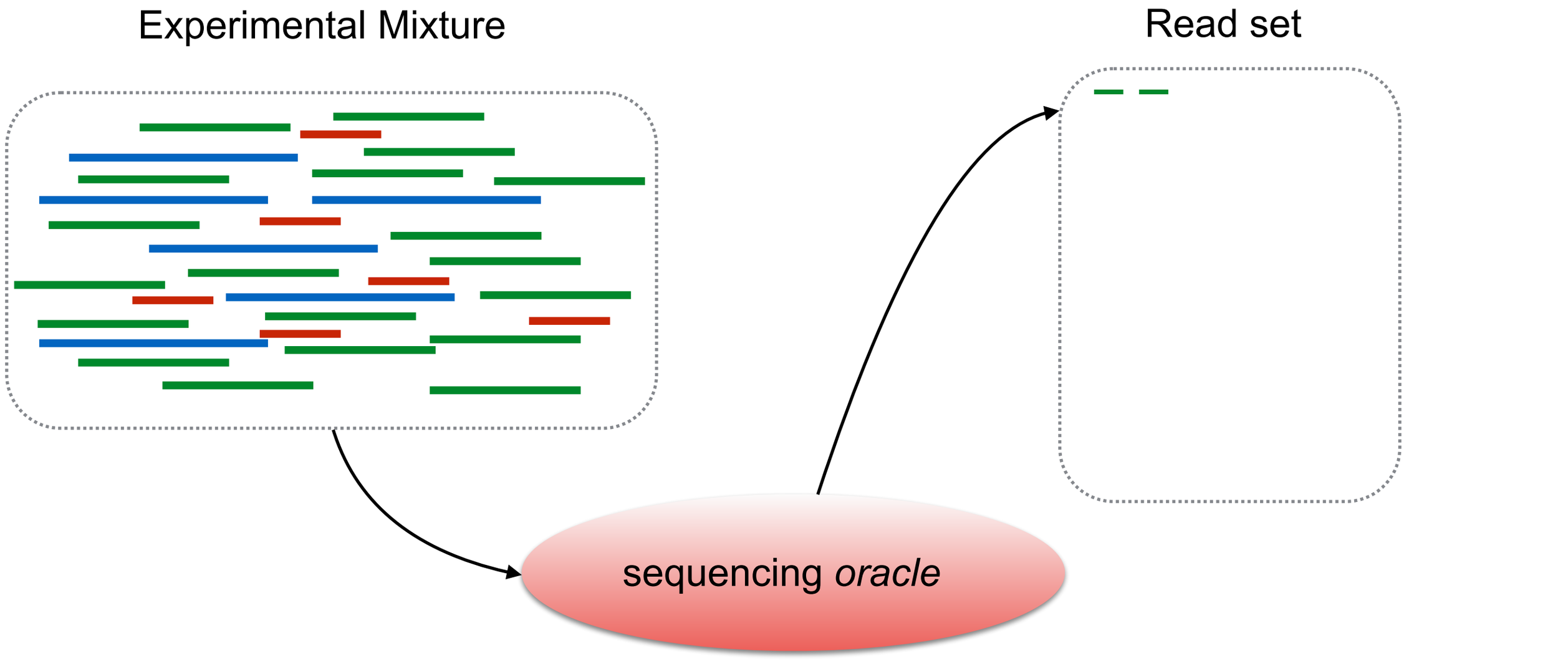


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

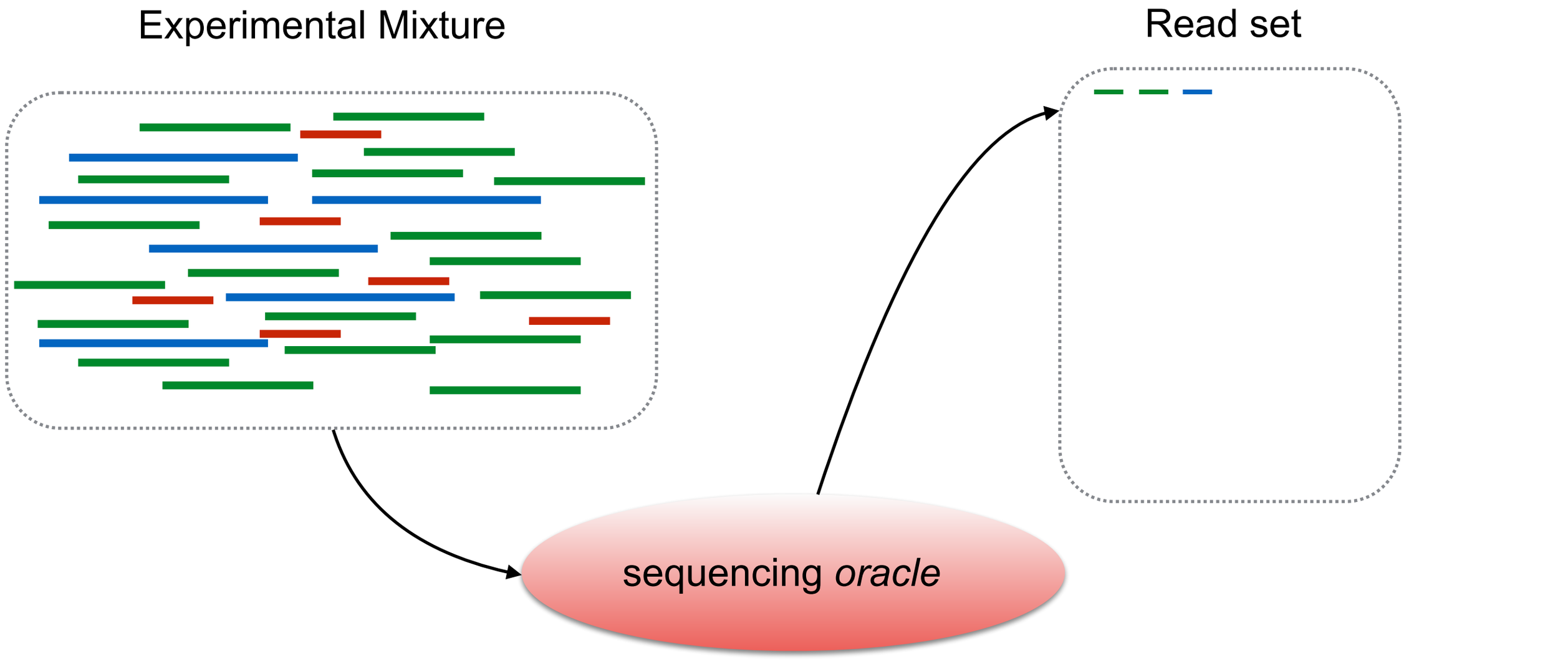


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

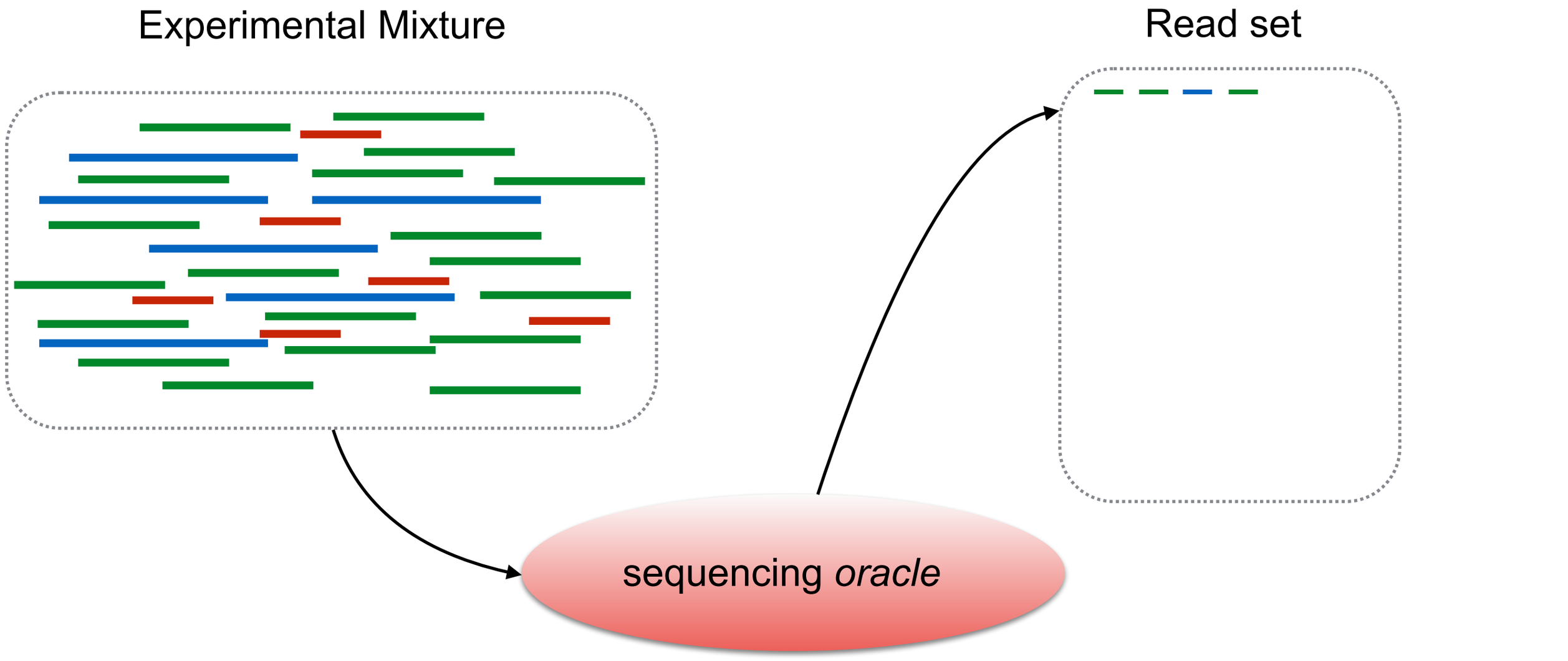


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

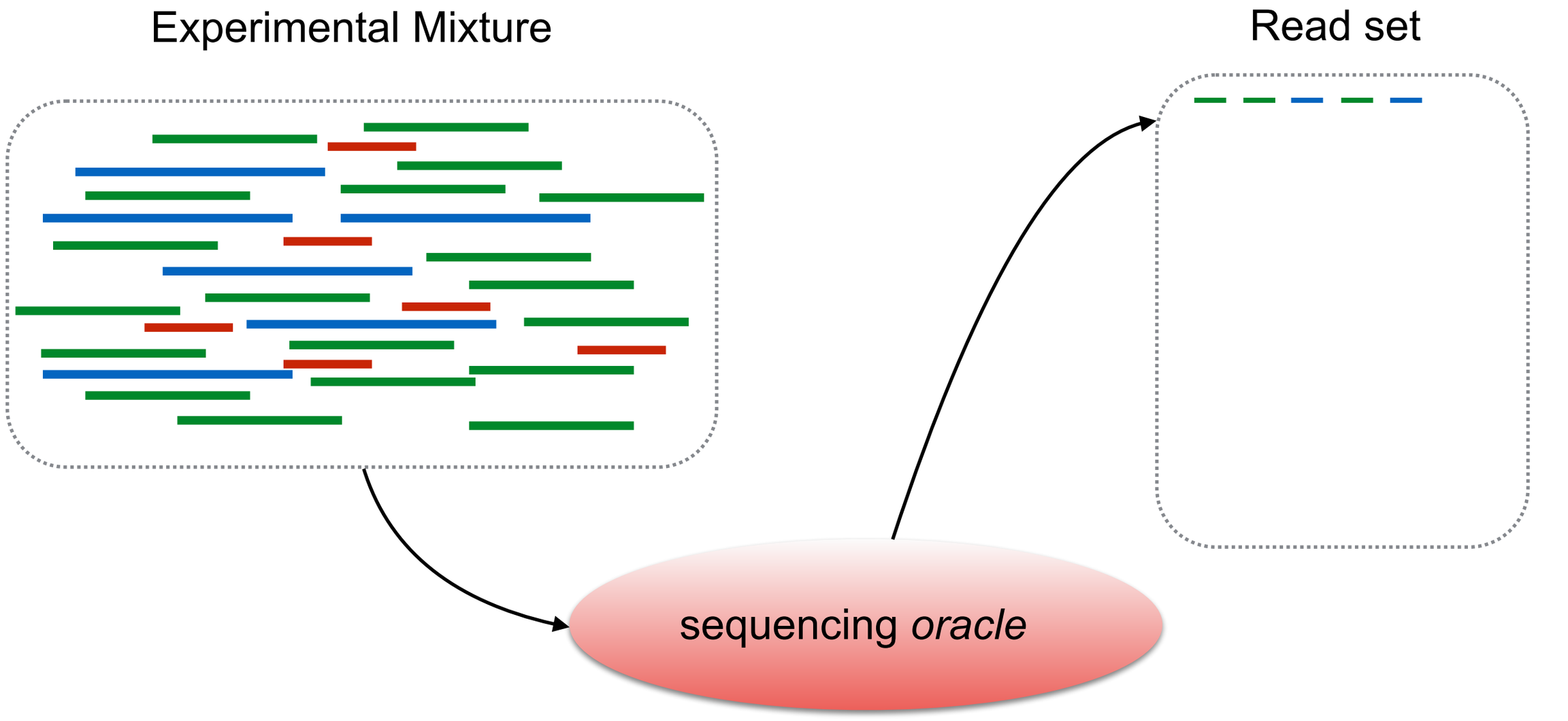


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

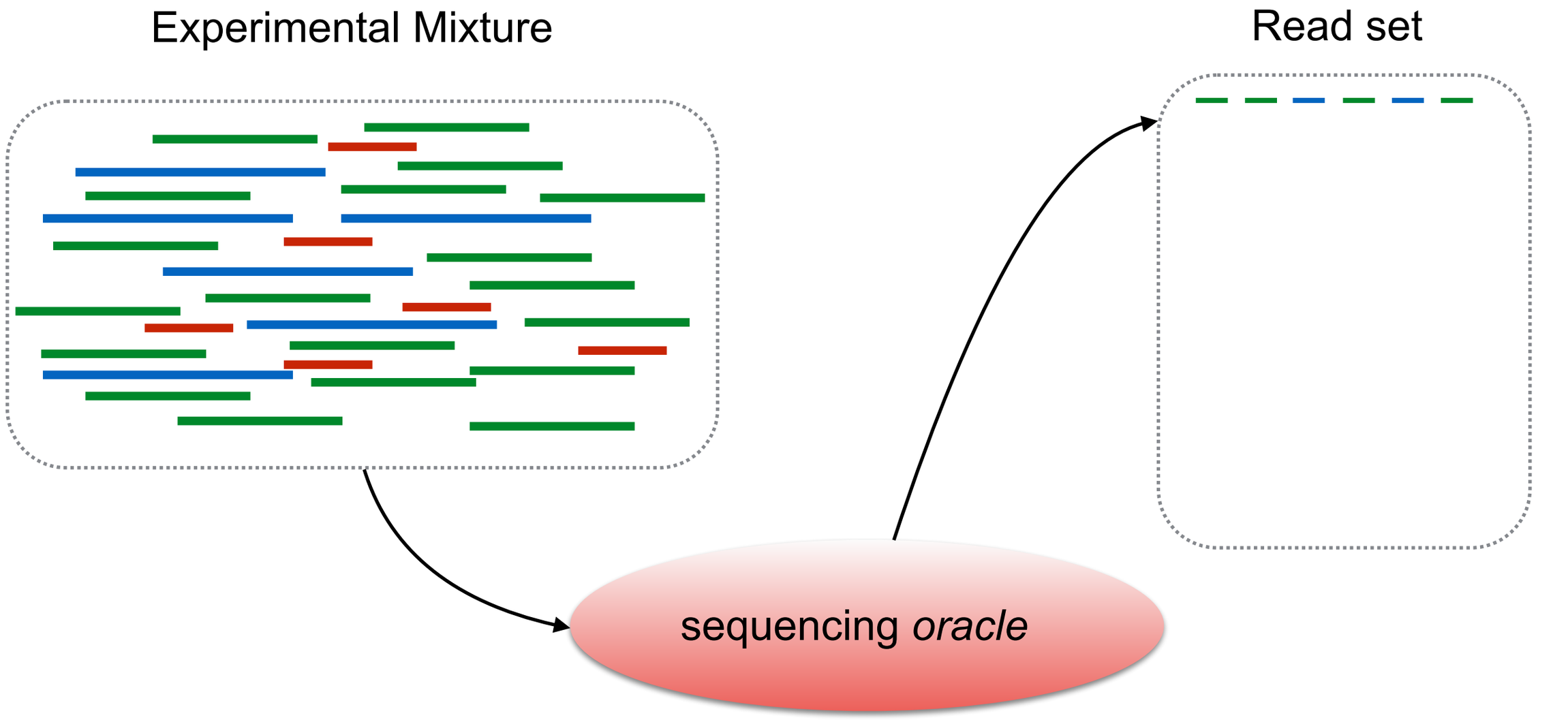


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

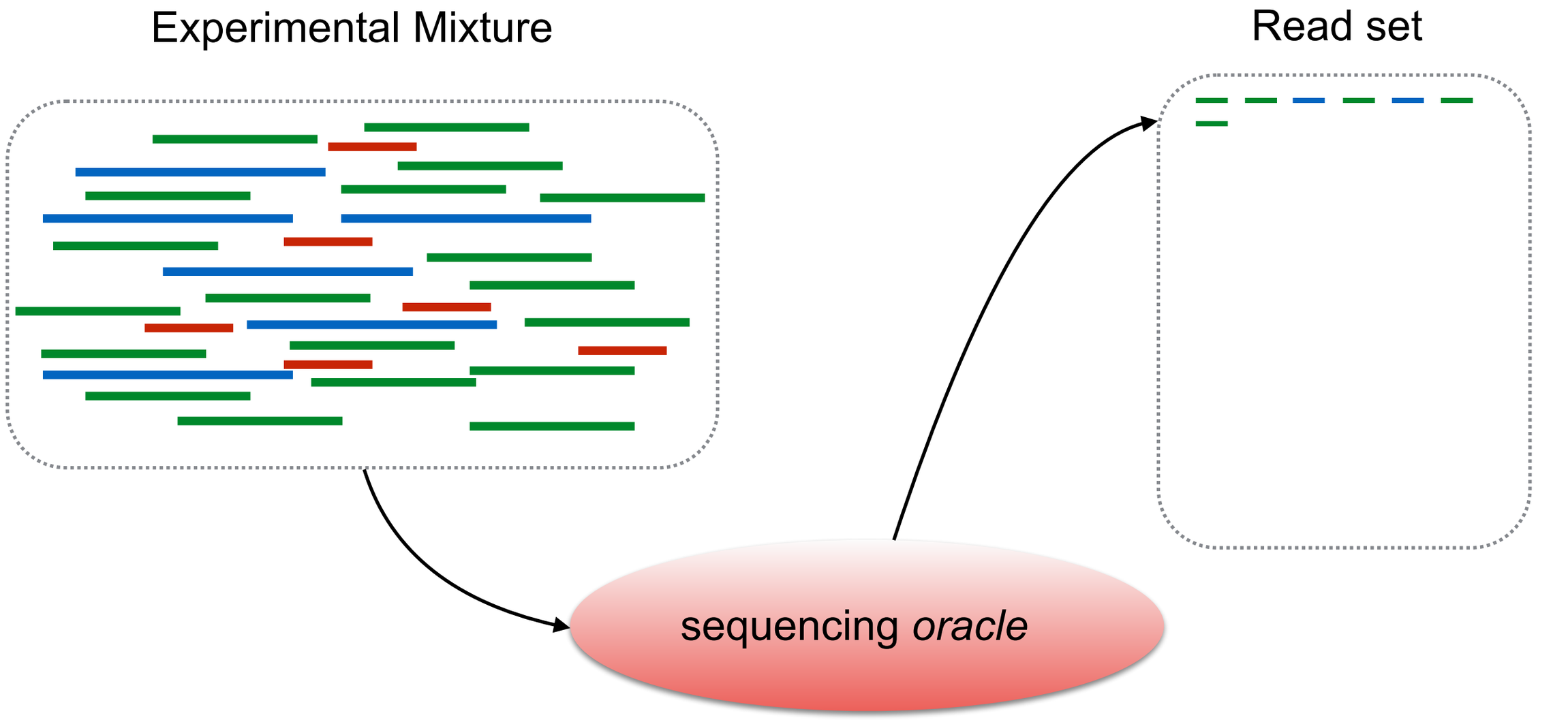


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

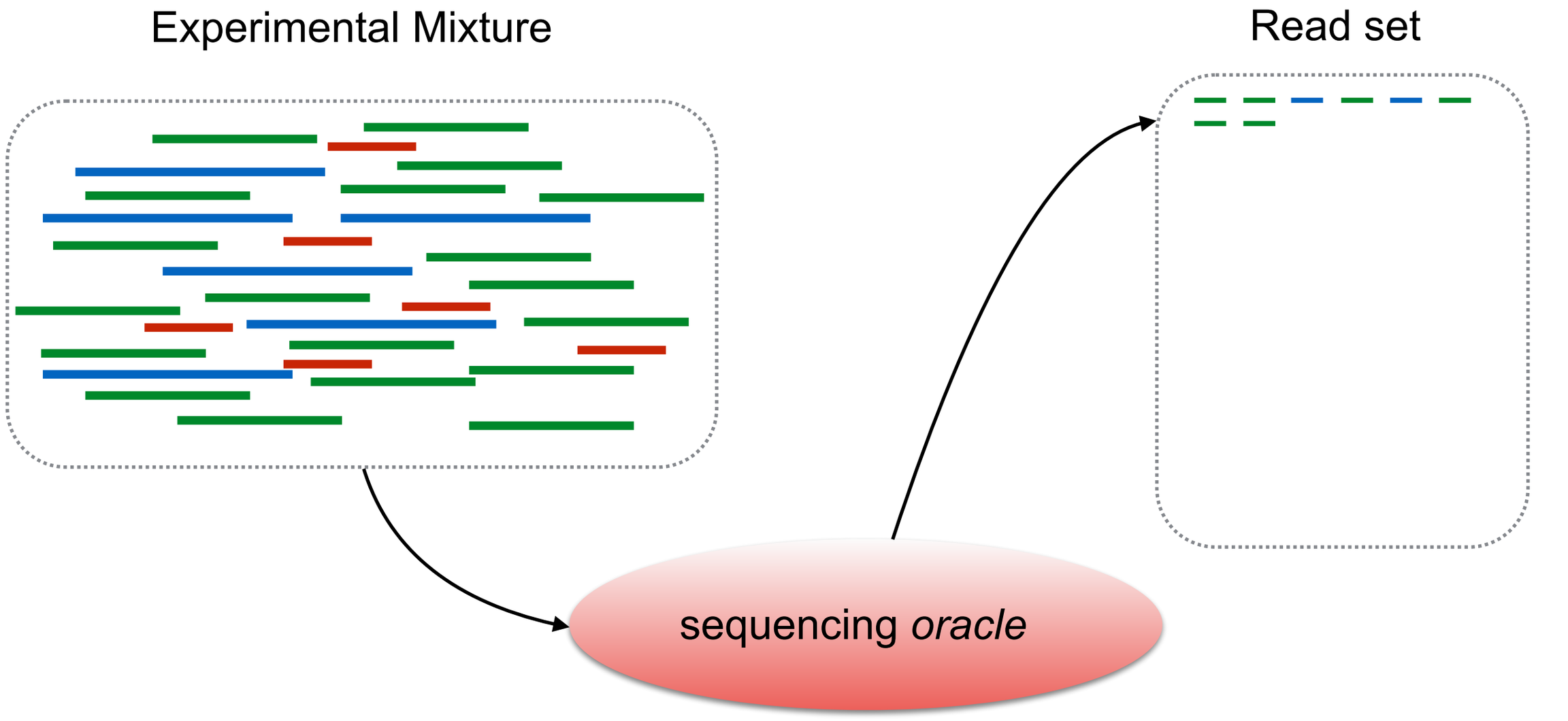


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

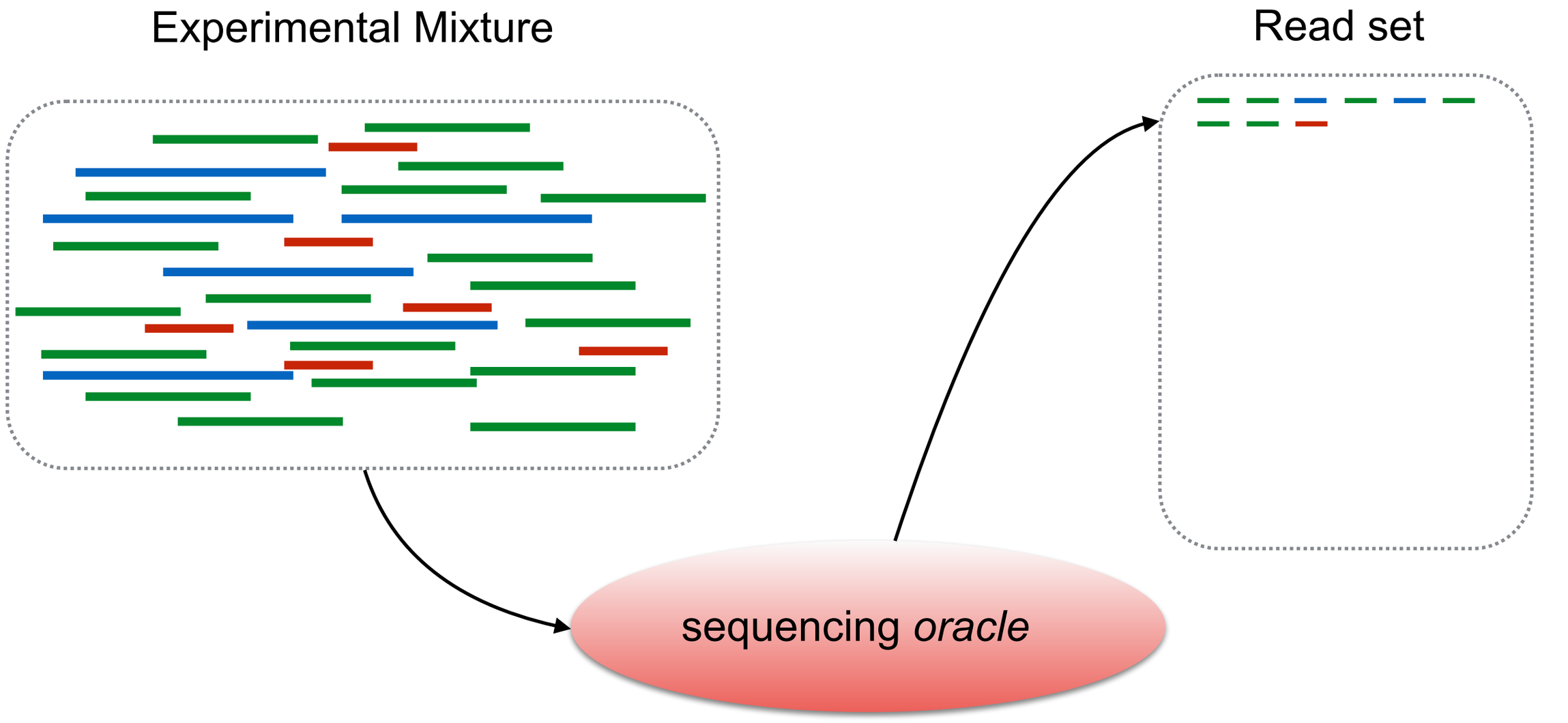


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

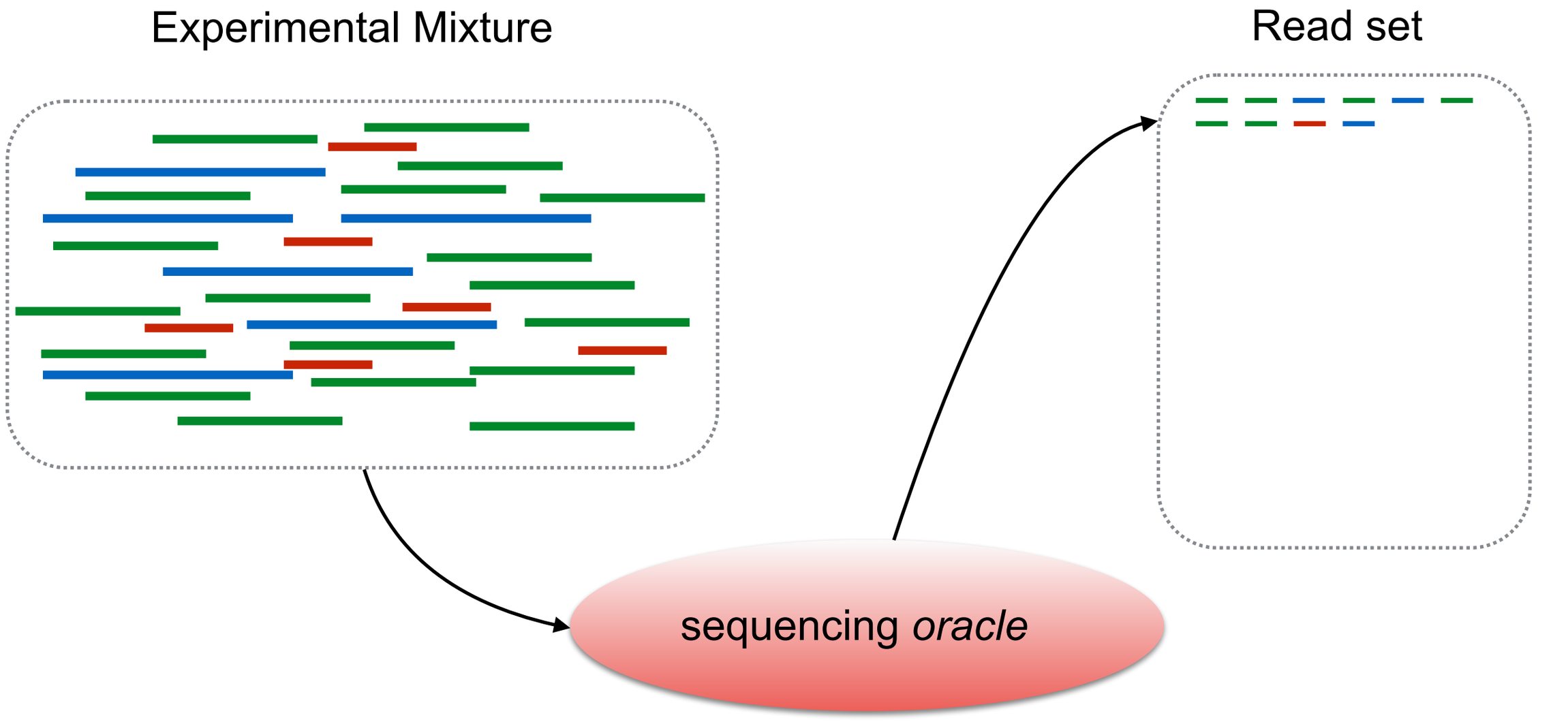


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

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

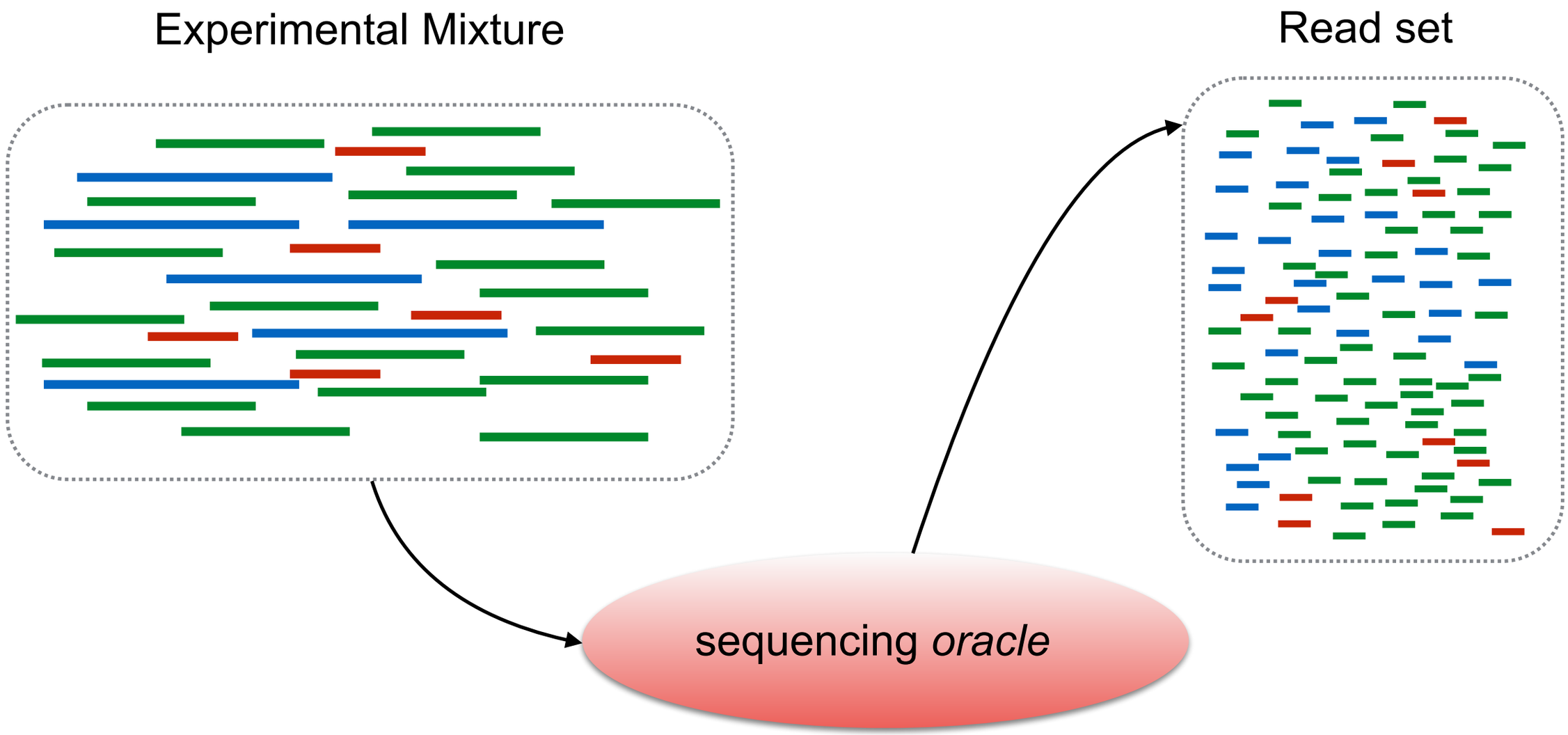


(1) Pick transcript **t** ∝ total available nucleotides = count * length

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

How can we perform inference from sequenced fragments?

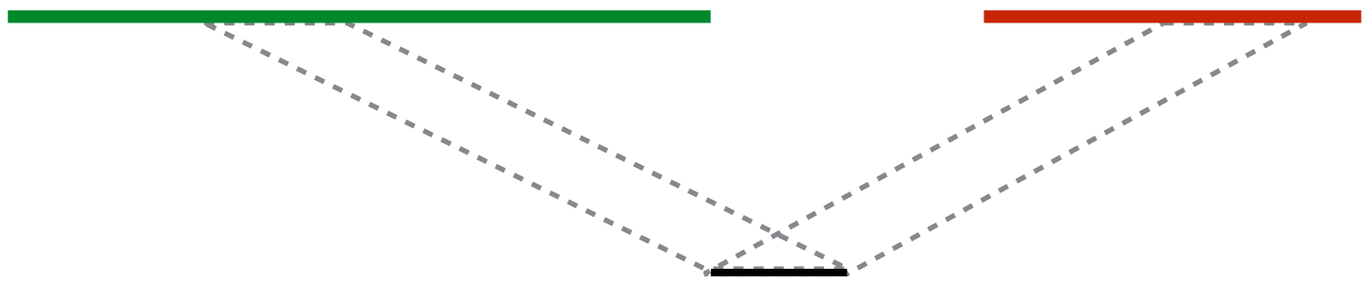
Think about the “ideal” RNA-seq experiment . . .



(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?

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

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

normalization factor

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

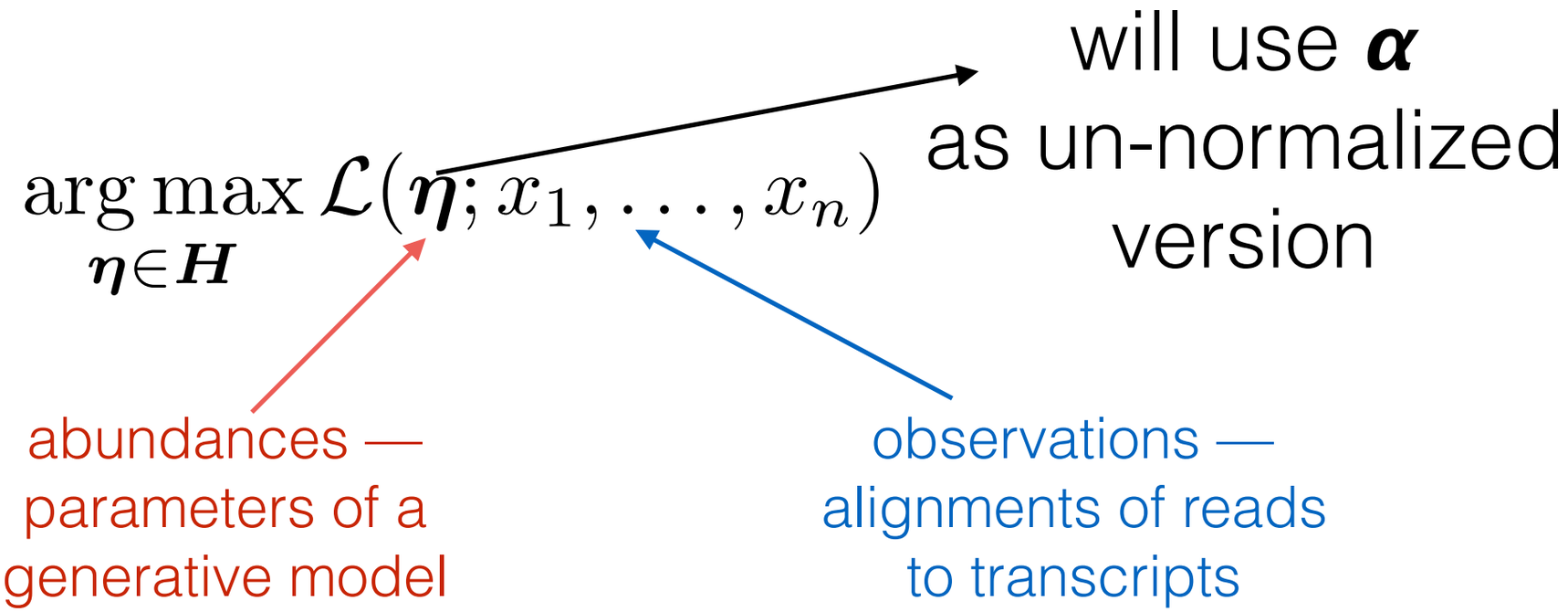
length( ) = 66 x 19 copies = 1254 nt ~ 60% green

length( ) = 33 x 6 copies = 198 nt ~ 10% red

So how do we estimate abundance “correctly”?

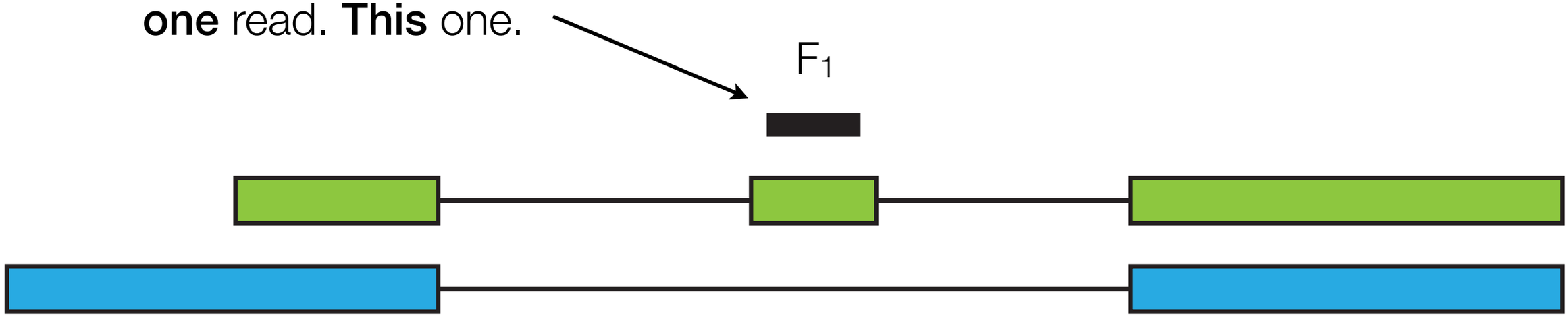
Key idea: Find the set of transcript abundances that maximizes the probability of the observed data — this is done by *probabilistic* assignment of fragments to transcripts.

That is: We’re asking for the maximum likelihood estimates of transcript abundance



Defining the likelihood function

Suppose we sequenced just **one** read. **This** one.



A few things need to happen to get this read as opposed to all the others we could have gotten:

We need to pick out a transcript from the RNA pool that could generate this read:

$$\text{Prob(Picking the green transcript)} = \frac{\text{copies of the green transcript}}{\text{total number of transcripts in the pool}}$$

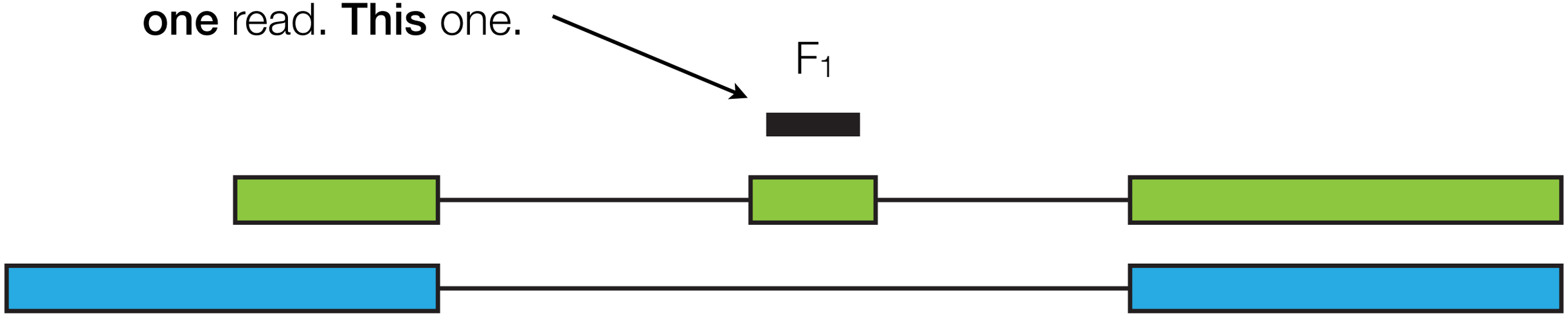
Then, we need to pick this read from that transcript over all the others.

$$\text{Prob(picking this read)} = \frac{1}{\text{length of green transcript}}$$

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Defining the likelihood function

Suppose we sequenced just **one** read. **This** one.



not normalized by length

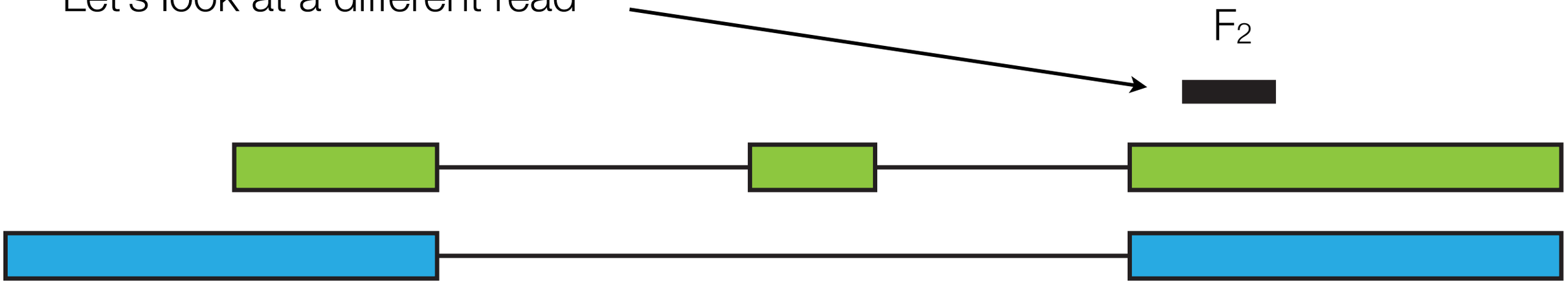
So given a relative abundance for the green transcript, which we'll call α_{green} we can calculate the probability of getting F_1.

$$\Pr(F_1 \in T_{\text{green}}) = \Pr(F_1 \mid \alpha_{\text{green}}) = \frac{\alpha_{\text{green}}}{\ell_{\text{green}}}$$

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Defining the likelihood function

Let's look at a different read



F_2 could have come from either transcript, so we have to consider two ways of getting it:

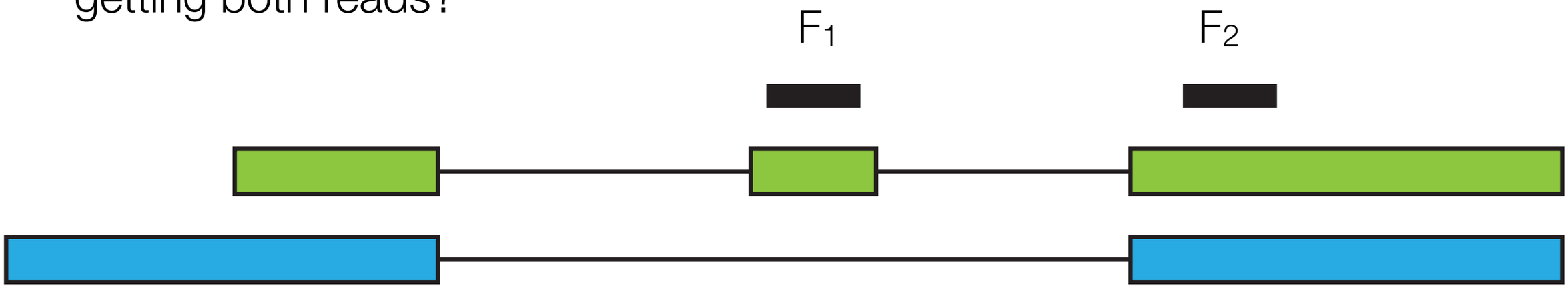
$$\Pr(F_2 \in T_{\text{green}} \text{ or } F_2 \in T_{\text{blue}}) = \Pr(F_2 \mid \alpha) = \frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}$$

That is, in order to know the probability of getting F_2, we need to know the abundances of both the transcripts it might have come from.

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Defining the likelihood function

What are the chances of getting both reads?



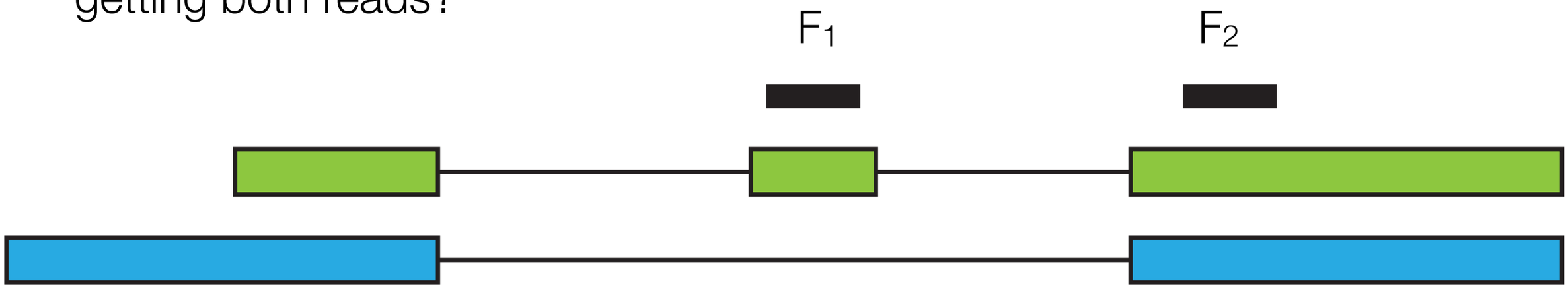
To get both F_1 and F_2, we just need to multiply the two probabilities!

$$\Pr(F_1 \in T_{\text{green}} \text{ and } F_2 \in T_{\text{green}} \text{ or } F_2 \in T_{\text{blue}}) = \Pr(F \mid \alpha) = \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}}\right) \cdot \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}\right)$$

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Defining the likelihood function

What are the chances of getting both reads?



Let's look at this probability as a *function* of alpha:

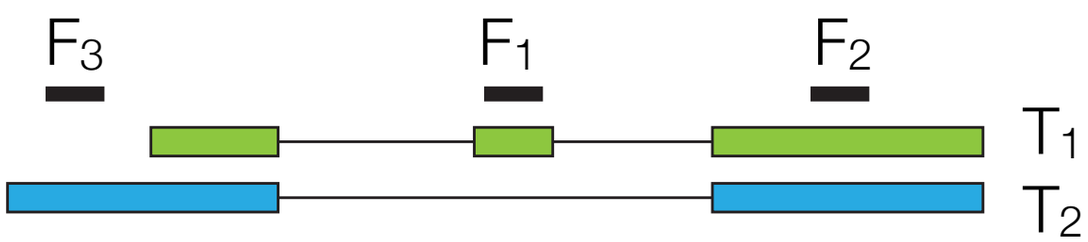
$$\mathcal{L}(\alpha; F) = \mathcal{L}(\alpha) = \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}}\right) \cdot \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}\right)$$

Given a input assignment of abundances to transcripts (the alphas), this function returns a number. The greater the number, the better the chances of seeing the reads we actually see.

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Defining the likelihood function

We can take any set of reads and any set of transcripts, and build one of these likelihood functions:



Compatibility table

	F_1	F_2	F_3
T_1	✓	✓	✗
T_2	✗	✓	✓

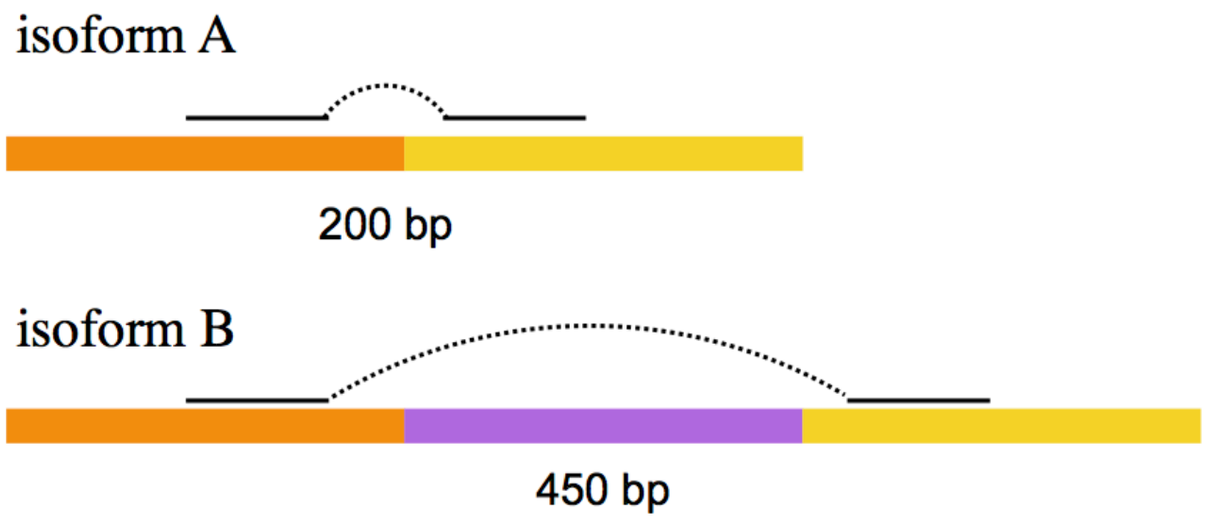
$$\mathcal{L}(\alpha; F) = \mathcal{L}(\alpha) = \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}}\right) \cdot \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}\right) \cdot \left(\frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}\right)$$

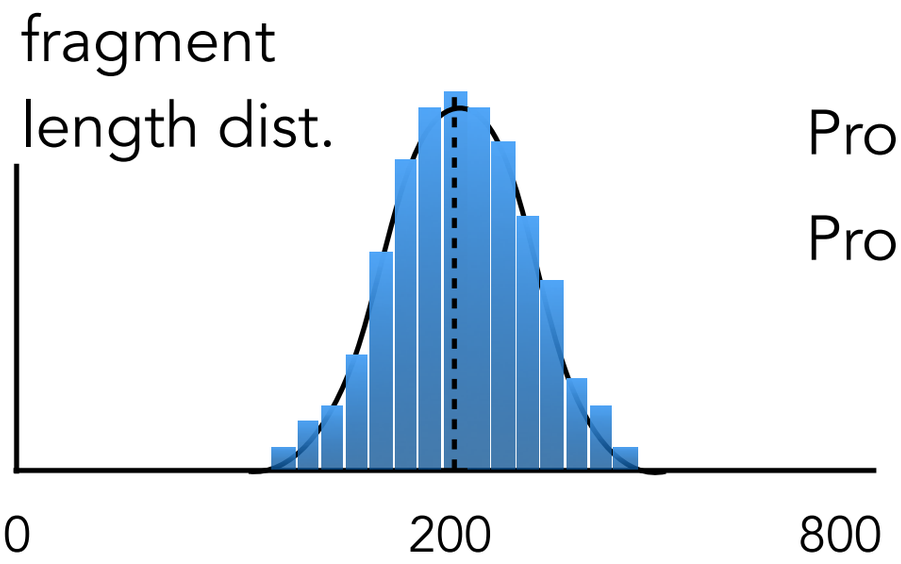
Now we want to find the values of alpha that **maximize** this likelihood function.

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

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

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**

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

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/VBEM algorithm.

but

This leads to an iterative EM/VBEM 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**

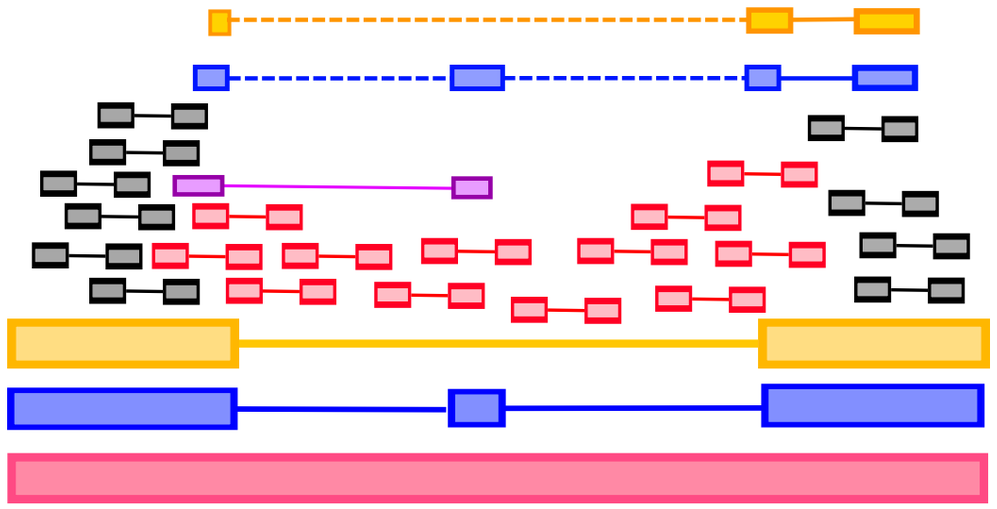
$$\mathcal{L}(\boldsymbol{\eta}; \mathcal{F}, \mathcal{T}) = \prod_{f \in \mathcal{F}} \sum_{t_i \in \Omega(f)} \Pr(t_i \mid \boldsymbol{\eta}) \Pr(f \mid t_i)$$

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.

Assigning reads to isoforms

Problem: infer which transcript each fragment came from



Some fragments could have come from any transcript (black), while others only one (blue, yellow). The purple fragment could have come from either the red or the blue one.

Conditional probability that a fragment came from a given isoform is a function of that isoform's abundance!

slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf)

Finding the MLE

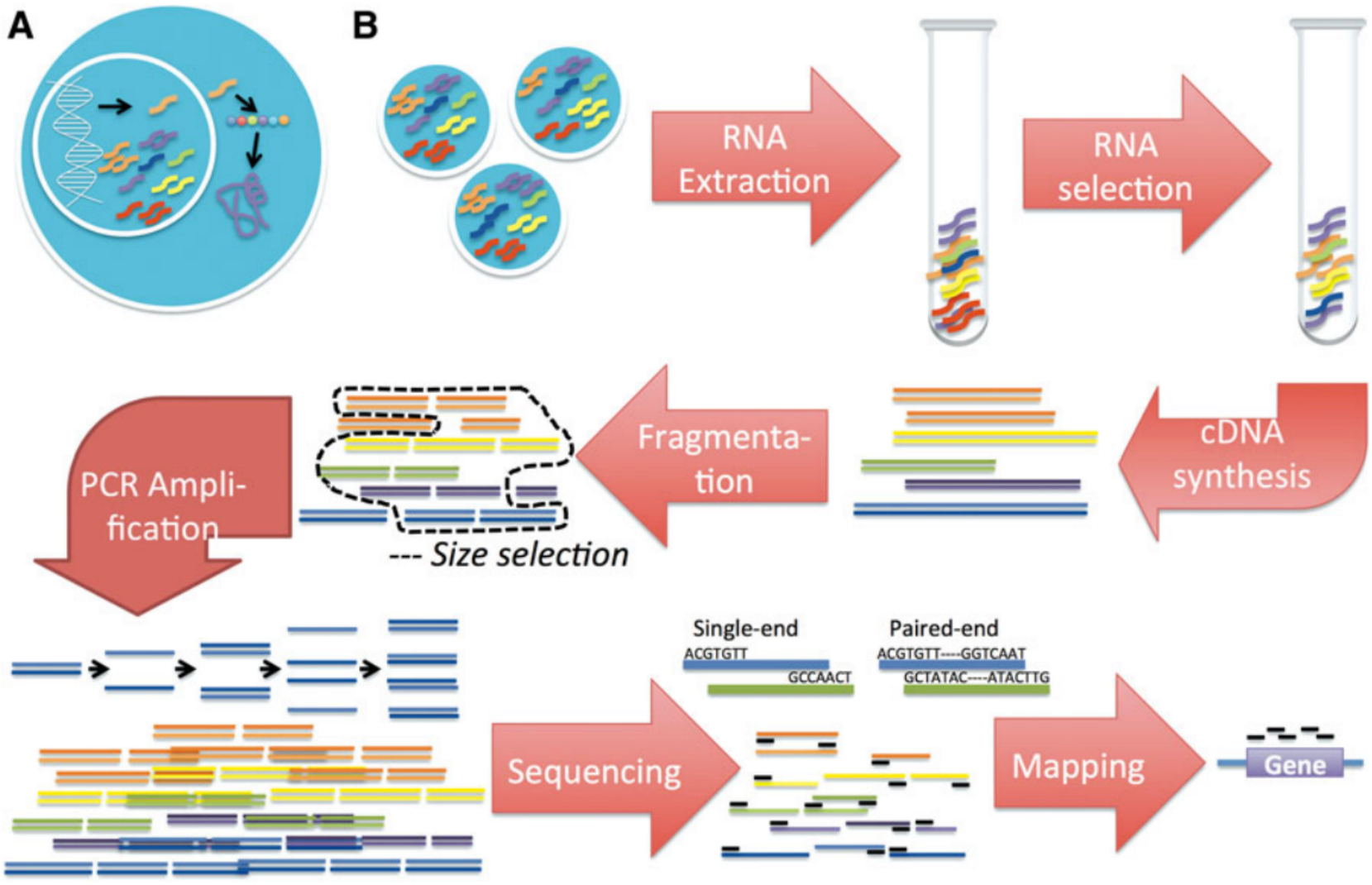
This problem lends itself very well to an Expectation Maximization (EM) approach.

Essentially:

While not converged:

E-step	Assign fragments to transcripts (probabilistically) using current estimates of transcript abundance.
M-step	Re-estimate transcript abundance using probabilistic fragment assignments.

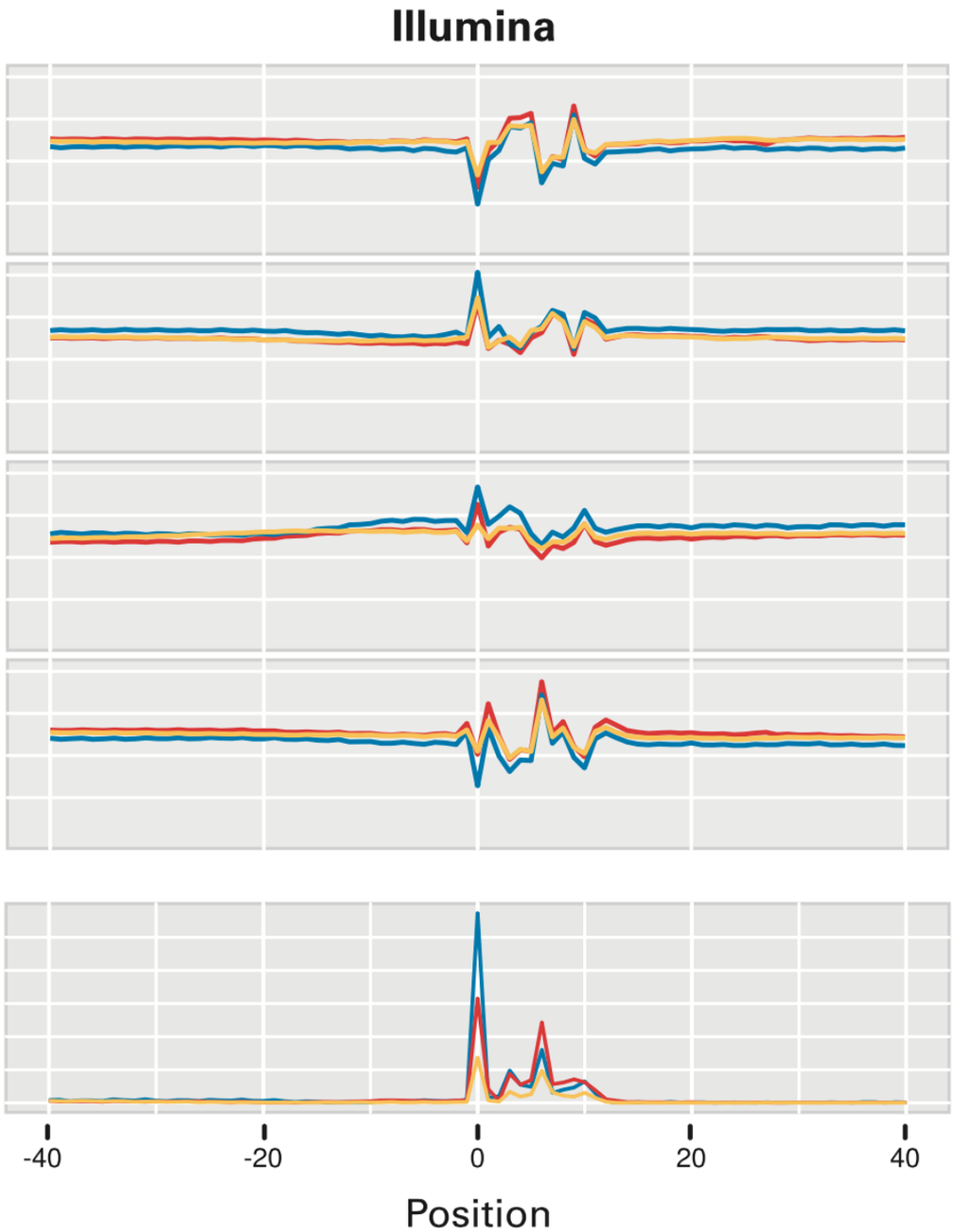
Actual RNA-seq protocols are a bit more “involved”



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

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

Modeling biases can be important



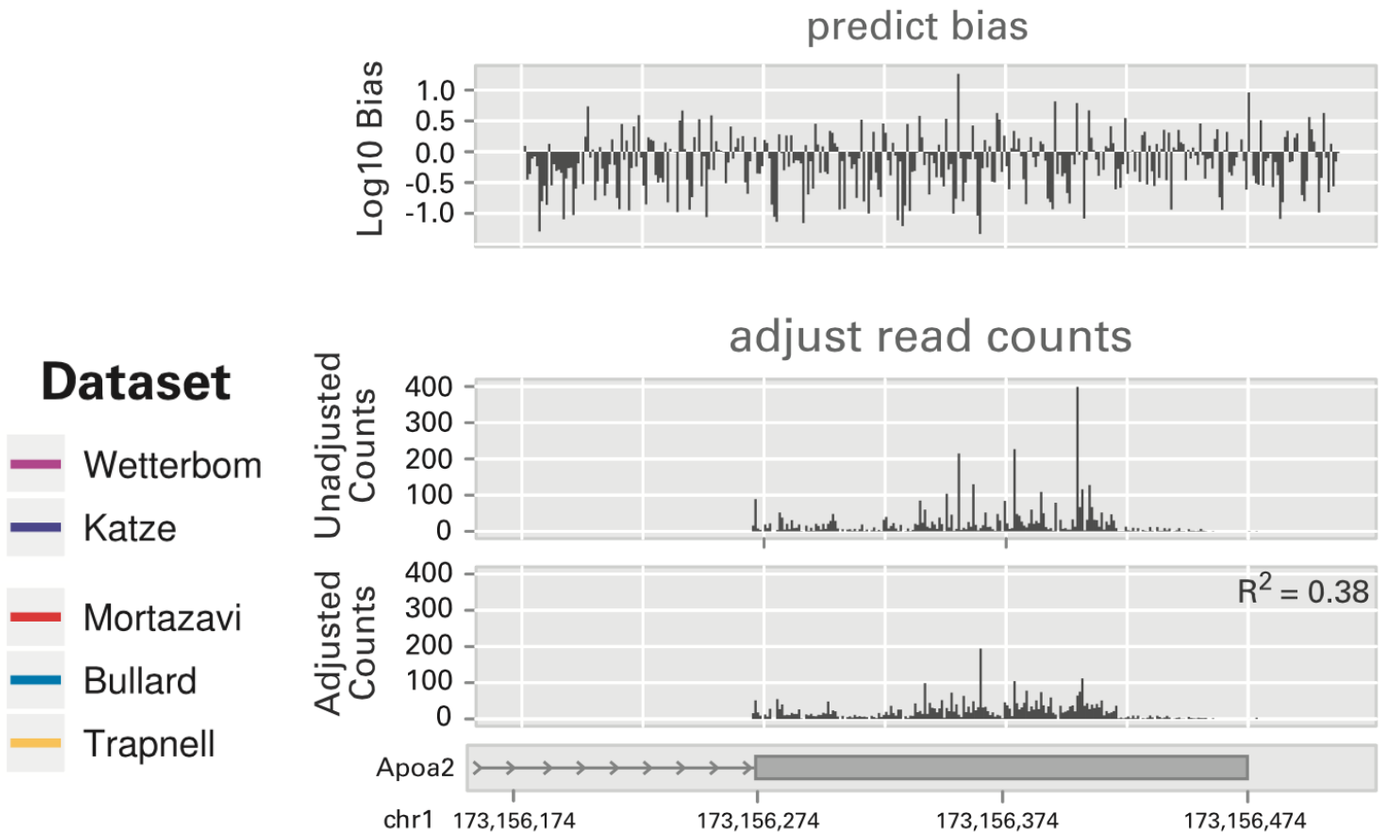


Table 3. The Pearson's correlation coefficient *r* between log-adjusted read counts and log-adjusted TaqMan values

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

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

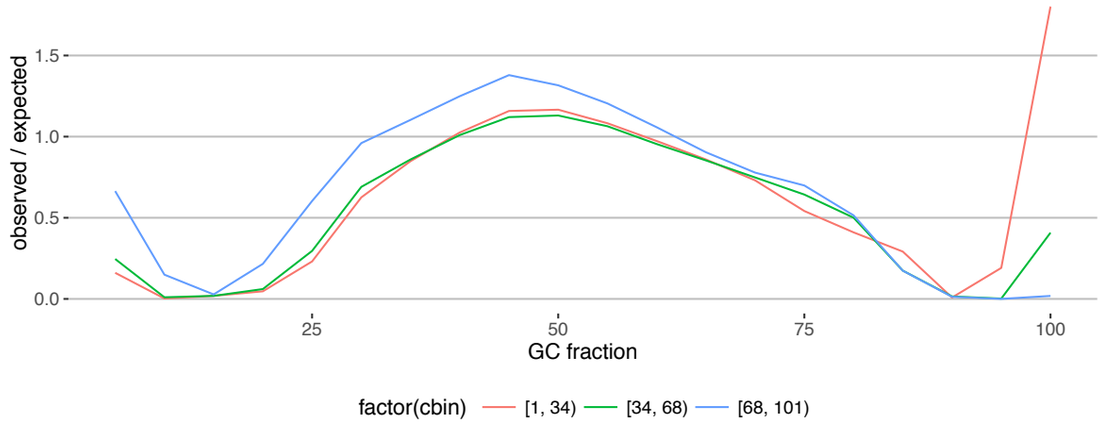
Biases abound in RNA-seq data

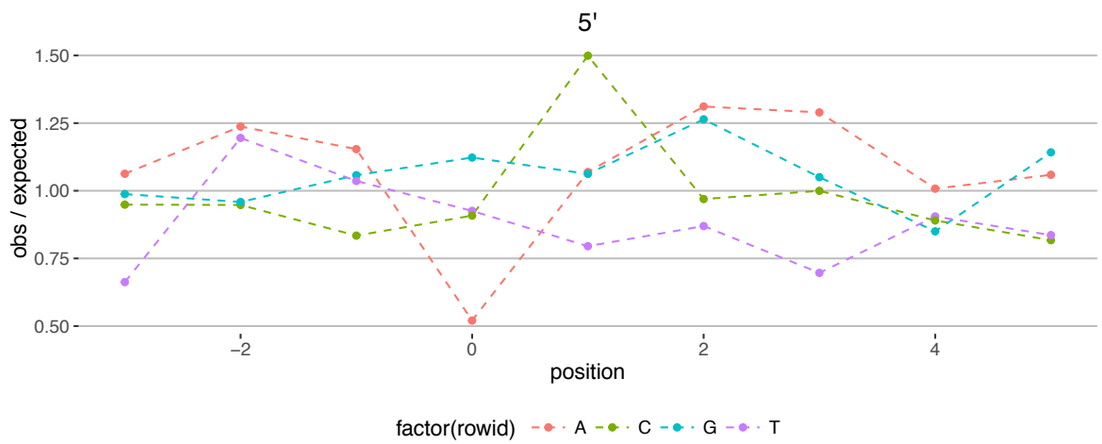
Biases in prep & sequencing can have a significant effect on the fragments we see:

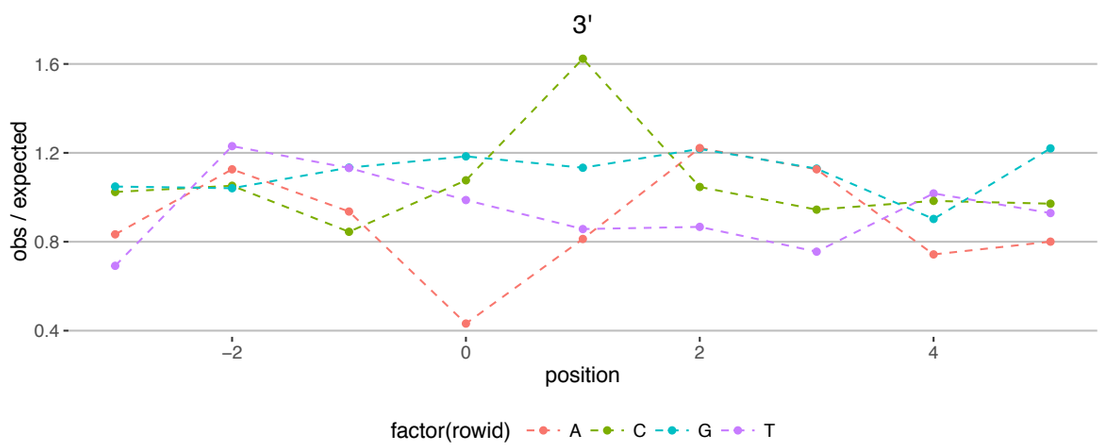
Fragment gc-bias[1]—
The GC-content of the fragment affects the likelihood of sequencing

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

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







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

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

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

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

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

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

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

Bias Modeling

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

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

Fragment GC bias model:

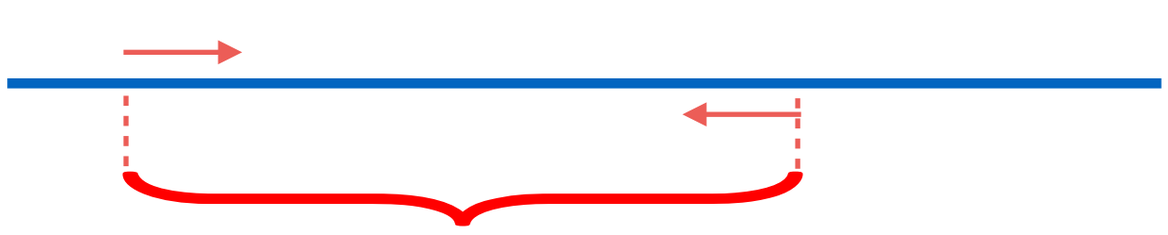
Density of fragments with specific GC content, **conditioned** on GC fraction at read start/end

Foreground:

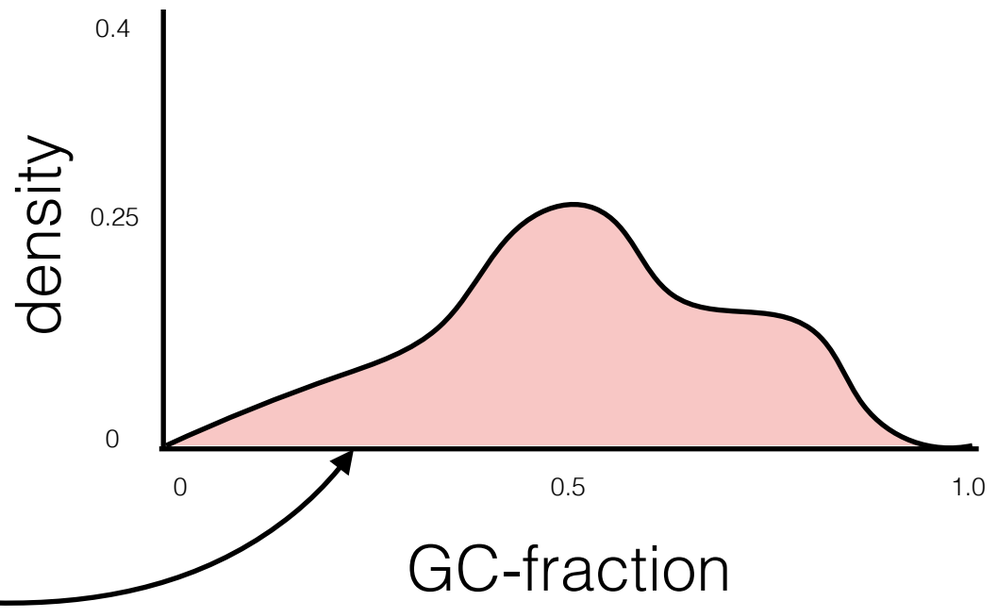
Observed

Background:

Expected given est. abundances



GC-fraction of fragment



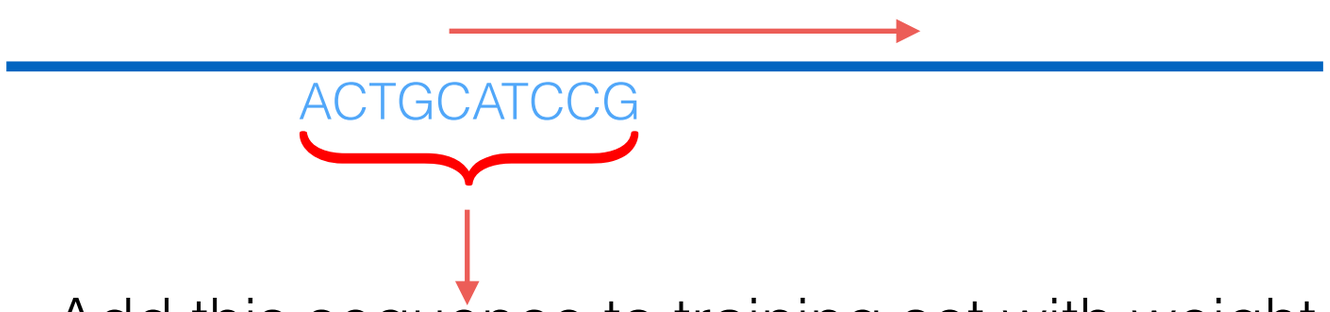
Bias Modeling

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



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

Foreground:

Observed

Background:

Expected given est. abundances

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

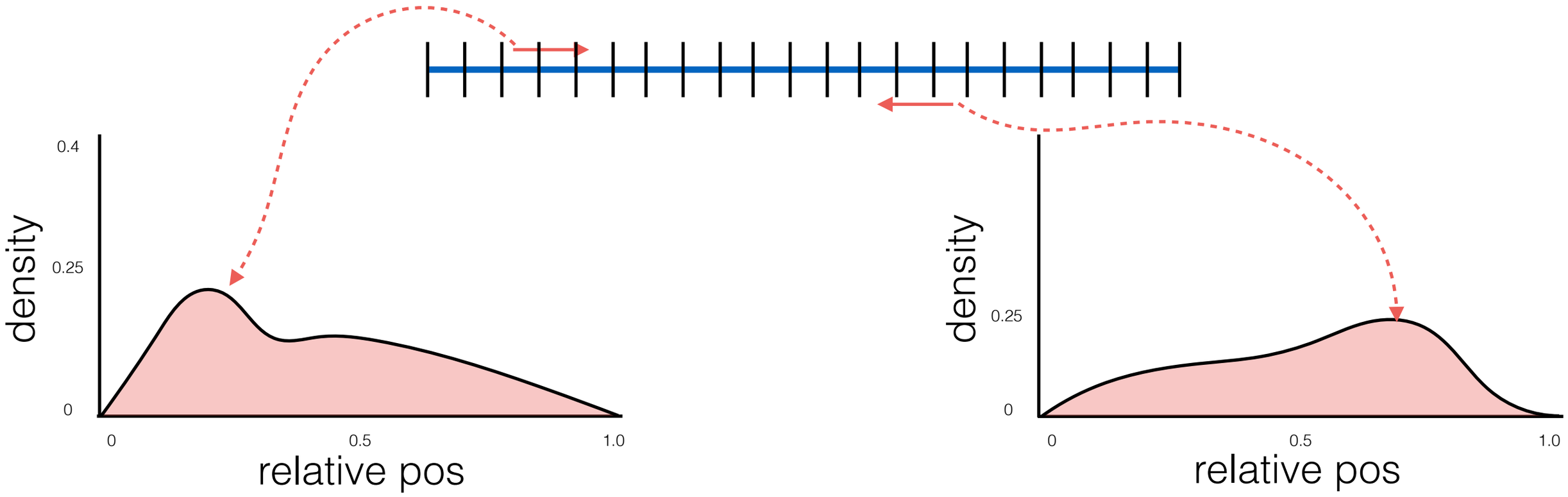
$$\tilde{\ell}'_i = \sum_{j=1}^{j \leq \ell_i} \sum_{k=1}^{k \leq f_i(j,L)} \frac{b_{gc+}(t_i, j, j+k)}{b_{gc-}(t_i, j, j+k)} \cdot \frac{b^{5'}_{s+}(t_i, j)}{b^{5'}_{s-}(t_i, j)} \cdot \frac{b^{3'}_{s+}(t_i, j+k)}{b^{3'}_{s-}(t_i, j+k)} \cdot \frac{b^{5'}_{p+}(t_i, j+k)}{b^{5'}_{p-}(t_i, j+k)} \cdot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

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.

Summary

RNA-seq is a technique that allows us to assay expressed transcripts “directly”.

It can be used to help annotate genes, find variation within genes, discover novel genes / transcripts, measure gene abundance, and study how abundance changes across conditions (among other things).

How the data is processed has a *huge* effect on our ability to draw useful information from it. Incorrect processing of even the best data can lead to missed results or spurious conclusions!