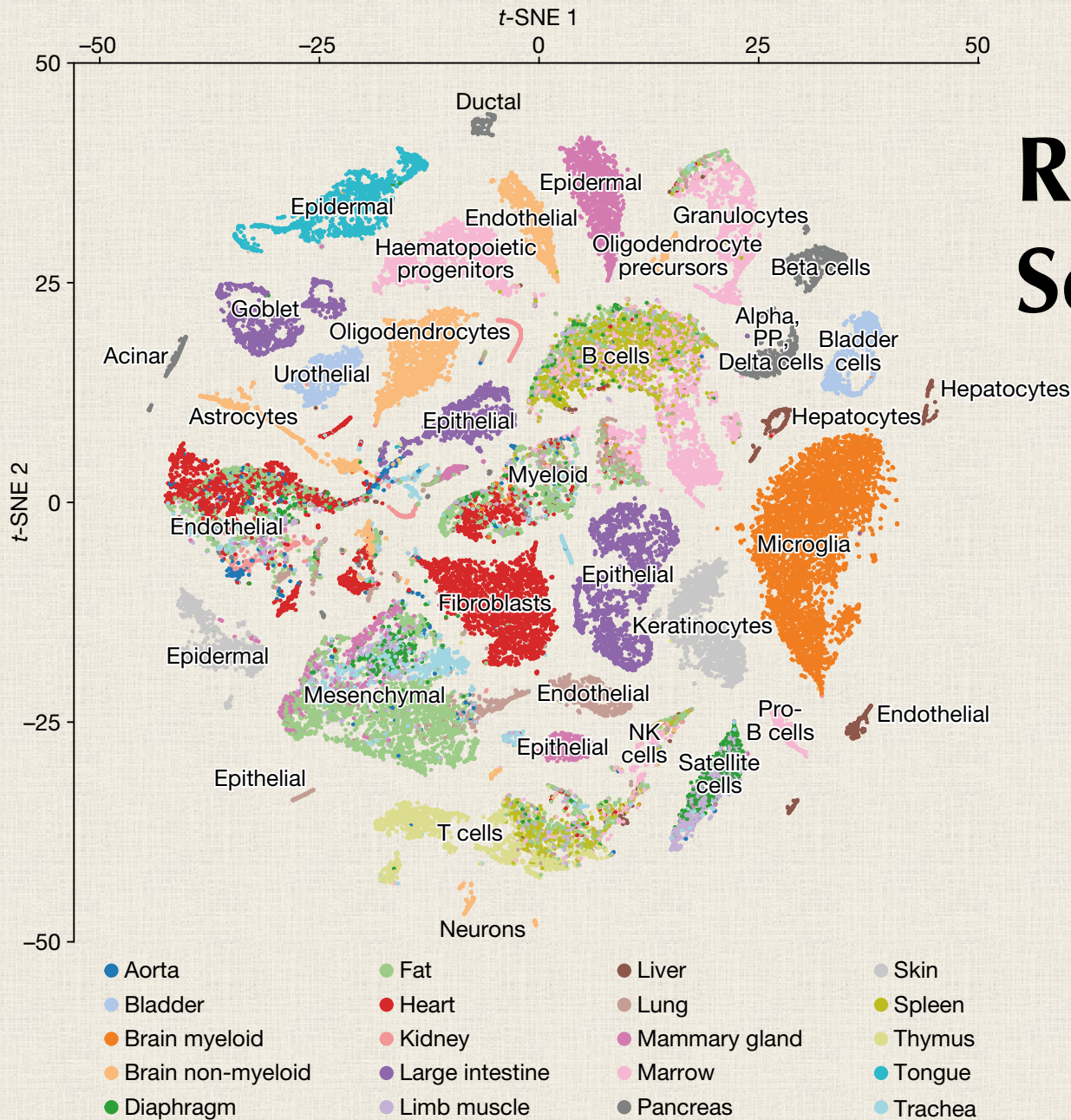


RNA Sequencing



Central Dogma of Molecular Biology:

DNA → RNA → Protein

DNA

5' GTGAAACTTTTTCTTGGTTTAATCAATAT 3'
3' CACTTTGAAAAAGGAACCAAATTAGTTATA 5'

Central Dogma of Molecular Biology: DNA → RNA → Protein

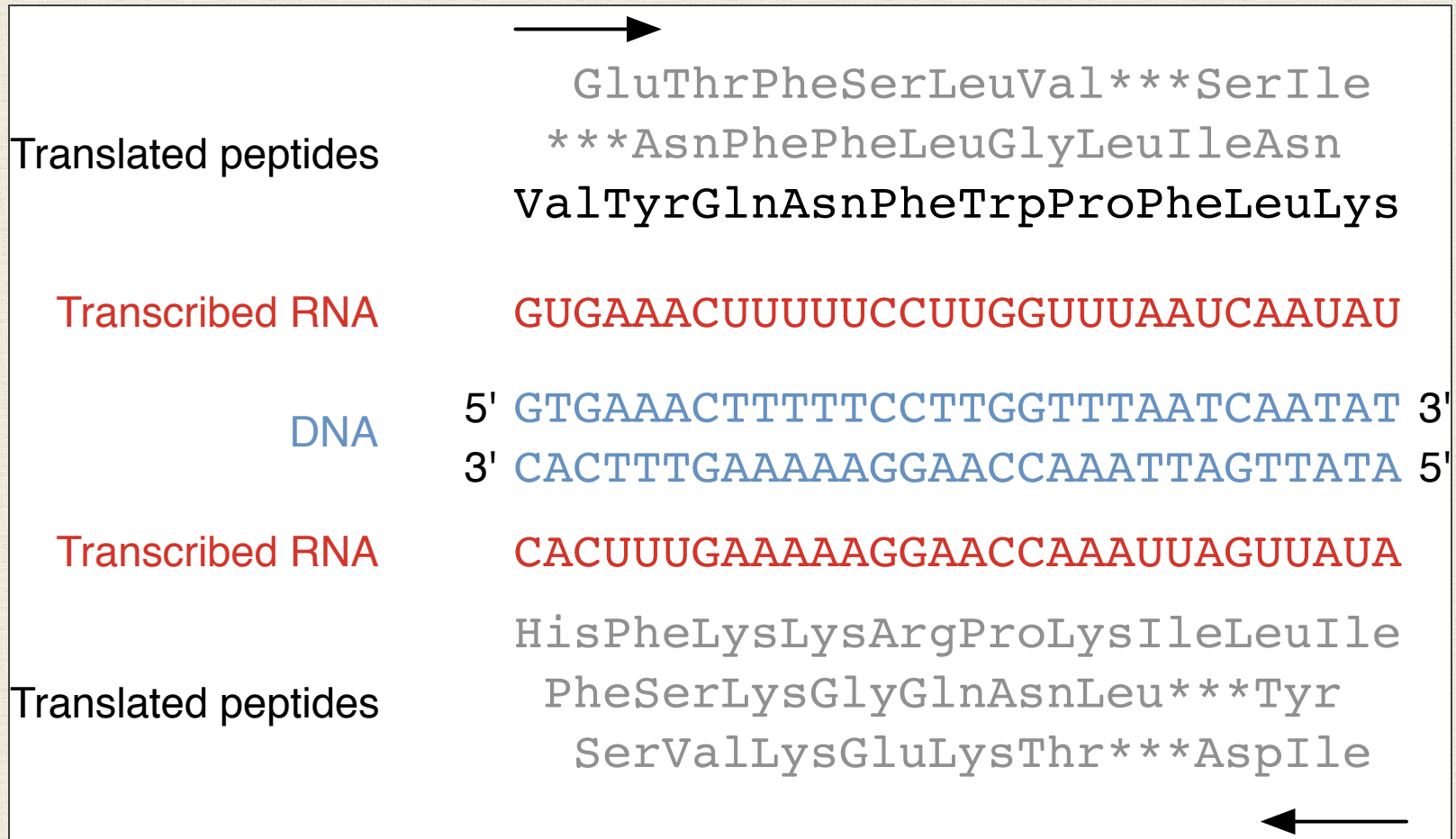
Transcribed RNA GUGAAACUUUUUCCUUGGUUUAUAU

DNA

5' GTGAACTTTTTCTTGGTTAATCAATAT 3'
3' CACTTTGAAAAAGGAACCAAATTAGTTATA 5'

Transcribed RNA CACUUUGAAAAAGGAACCAAUUAGUUAUA

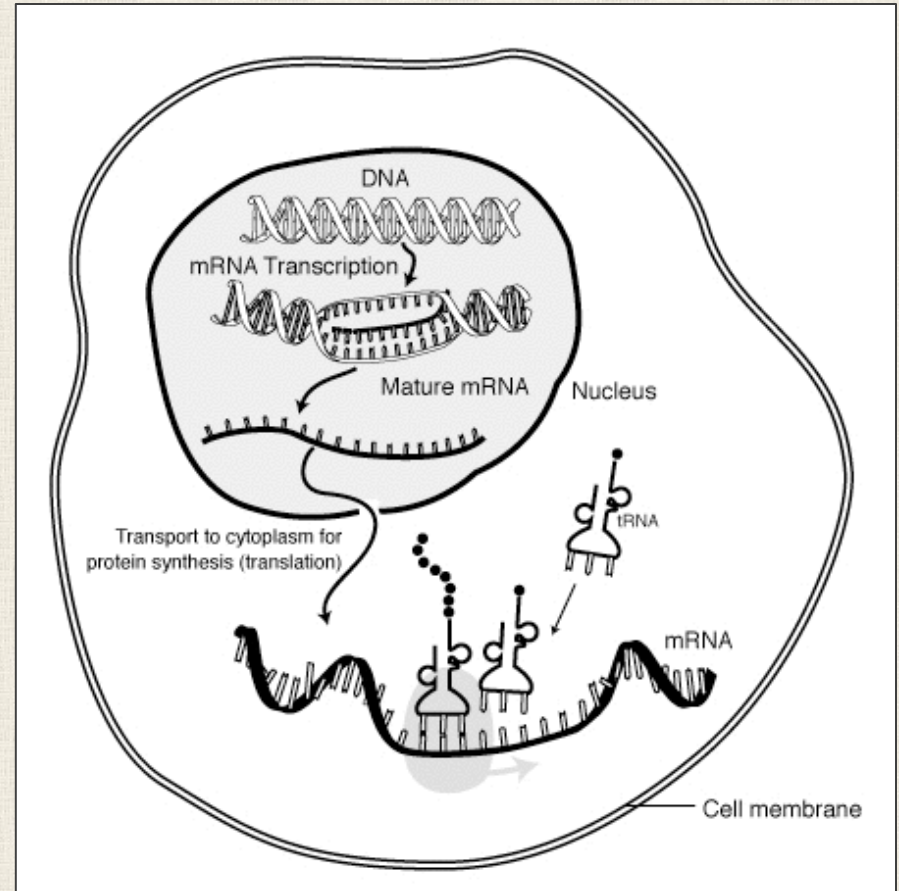
Central Dogma of Molecular Biology: DNA → RNA → Protein



The Central Dogma in Action

DNA is transcribed into **messenger RNA** (mRNA), which then leaves the nucleus.

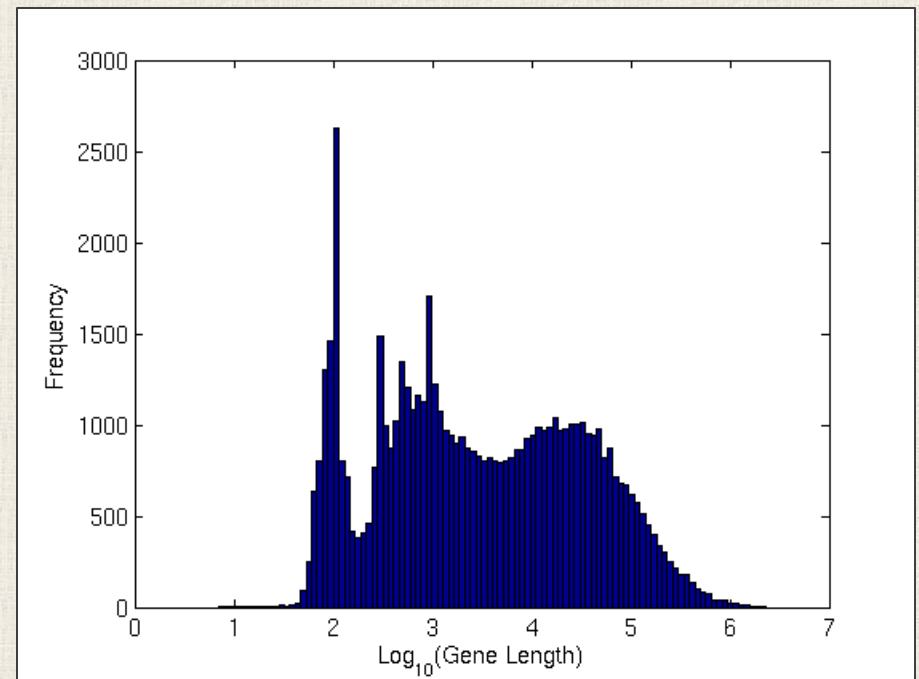
Ribosomes pass down the mRNA strand and build a growing strand of amino acids based on **codons** (triplets of nucleosides).



Distribution of Human Protein Lengths

Length heavily skews toward shorter proteins (much like synteny block fragment lengths).

- Range: 50 – 34000 amino acids.
- Median length: 375 amino acids (= 1125 base pairs of DNA).

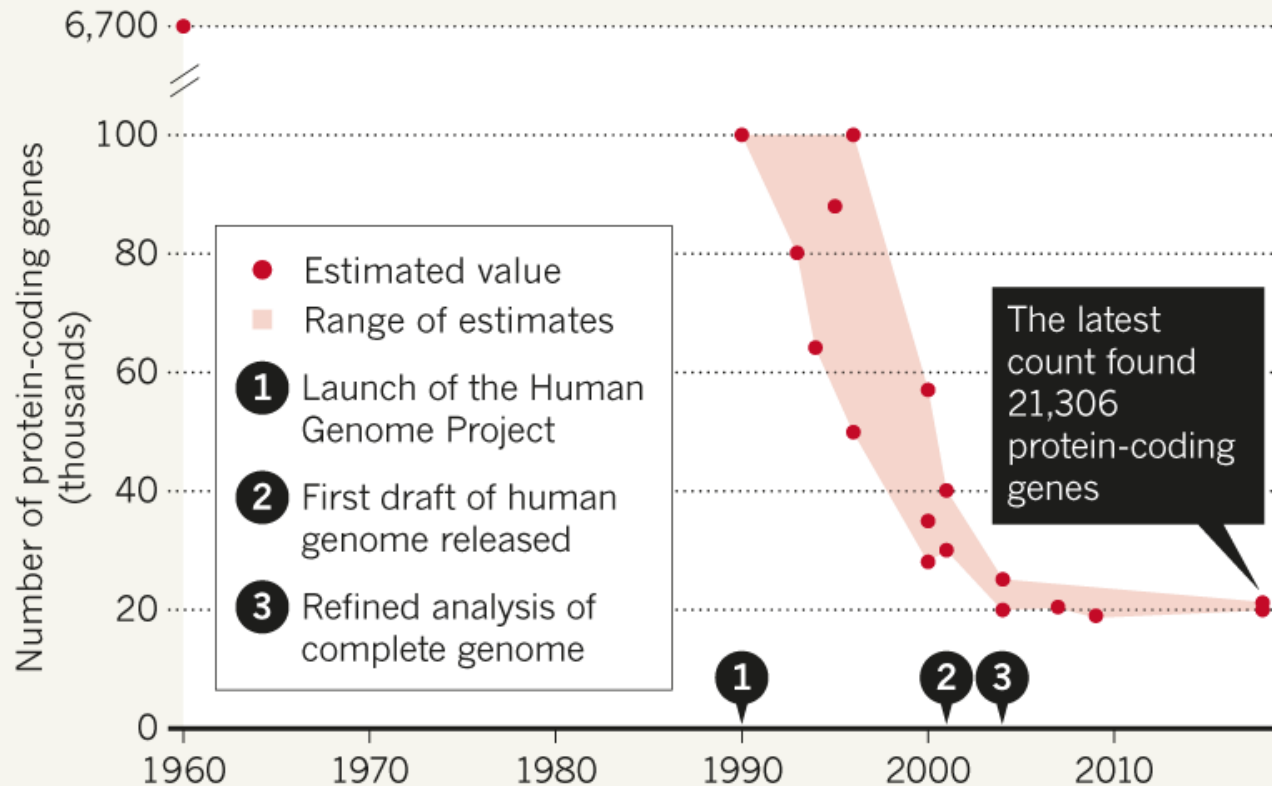


<https://biology.stackexchange.com/questions/48110/how-is-the-size-of-a-gene-defined/48117#48117>

The Estimate of Human Genes Has Decreased Over Time

GENE TALLY

Scientists still don't agree on how many protein-making genes the human genome holds, but the range of their estimates has narrowed in recent years.



©nature



Fruit Fly
44%



Mouse
92%



Chimp
98%



Yeast
26%



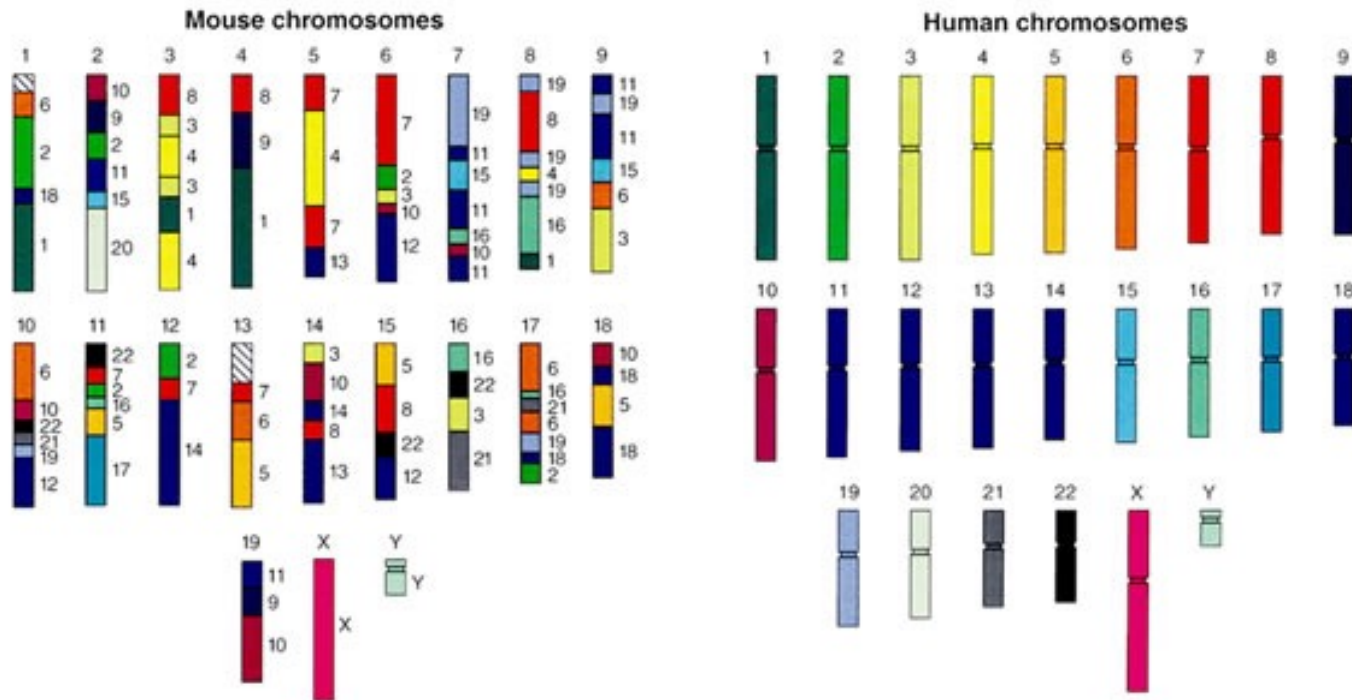
Plant
18%



**What percent
of your genes
do you share?**

This is Misleading

Mouse and Human Genetic Similarities



Courtesy Lisa Stubbs
Oak Ridge National Laboratory

YGA 98-075R2

Three Questions

STOP: What practical purpose might rearranging genes serve for an organism?

Three Questions

STOP: What practical purpose might rearranging genes serve for an organism?

STOP: Your cells all have (essentially) the same genome, so how can they perform different functions?

Three Questions

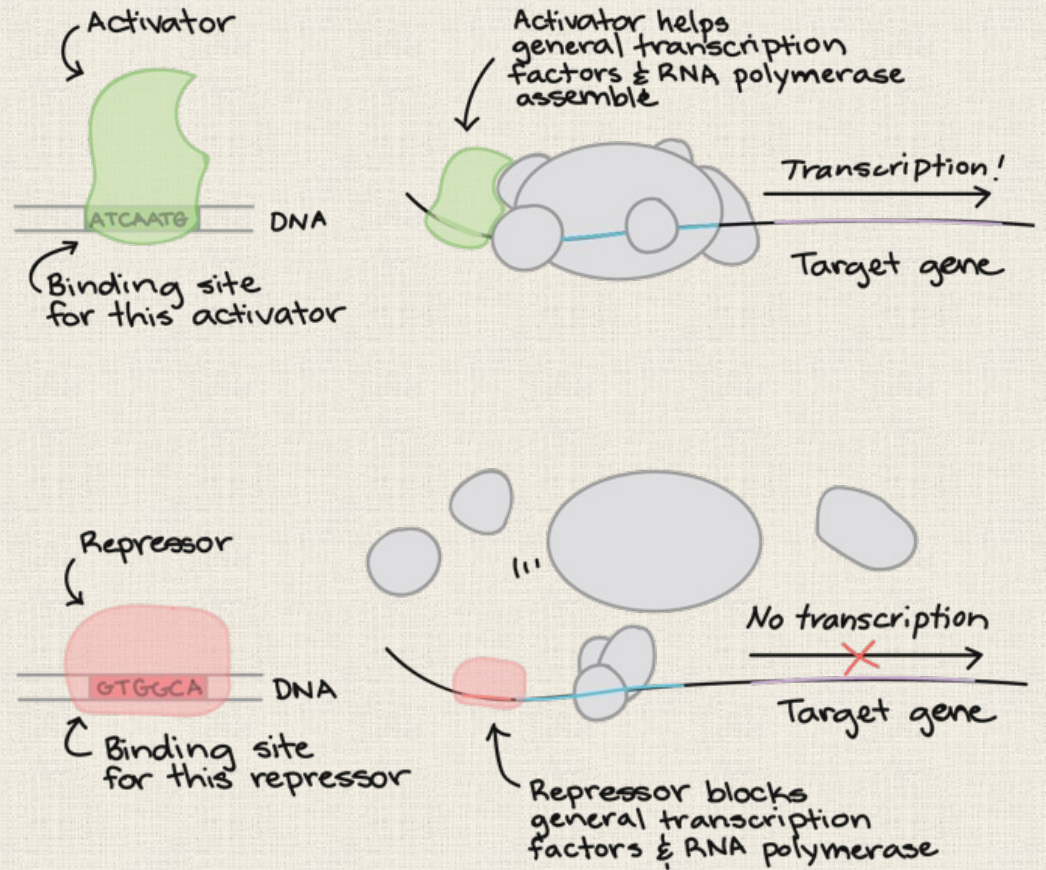
STOP: What practical purpose might rearranging genes serve for an organism?

STOP: Your cells all have (essentially) the same genome, so how can they perform different functions?

STOP: How can the same cell perform different functions at different times?

One Answer to Three Questions: Gene Regulation (a.k.a. "Expression")

Gene regulation: the ability of the cell to increase (activate) or decrease (repress) the production of RNA/protein corresponding to a given gene.



<https://www.khanacademy.org/science/biology/gene-regulation/gene-regulation-in-eukaryotes/a/eukaryotic-transcription-factors>

From Genomes to Protein Analysis

Classic analogy:

- **Genome:** sum total of a cell's DNA = cookbook
- **Transcriptome:** a cell's mRNA = photocopied recipe
- **Proteome:** set of proteins present in given cell = today's menu

From Genomes to Protein Analysis

Classic analogy:

- **Genome:** sum total of a cell's DNA = cookbook
- **Transcriptome:** a cell's mRNA = photocopied recipe
- **Proteome:** set of proteins present in given cell = today's menu

Our question: we have worked largely with genomes, but how can we measure the amount of each type of protein in a cell at a given time?

Genome Sequencing Had a Revolution, But Proteins are Still Waiting

Although we can read long genomes with 10 billion base pairs, isolating and reading proteins is very difficult.

Genome Sequencing Had a Revolution, But Proteins are Still Waiting

Although we can read long genomes with 10 billion base pairs, isolating and reading proteins is very difficult. *For now...*



Genome Sequencing Had a Revolution, But Proteins are Still Waiting

Although we can read long genomes with 10 billion base pairs, isolating and reading proteins is very difficult. *For now...*

Instead, we will take a middle ground and use **RNA-sequencing**: reading the RNA present in a given biological sample as a proxy for protein levels.

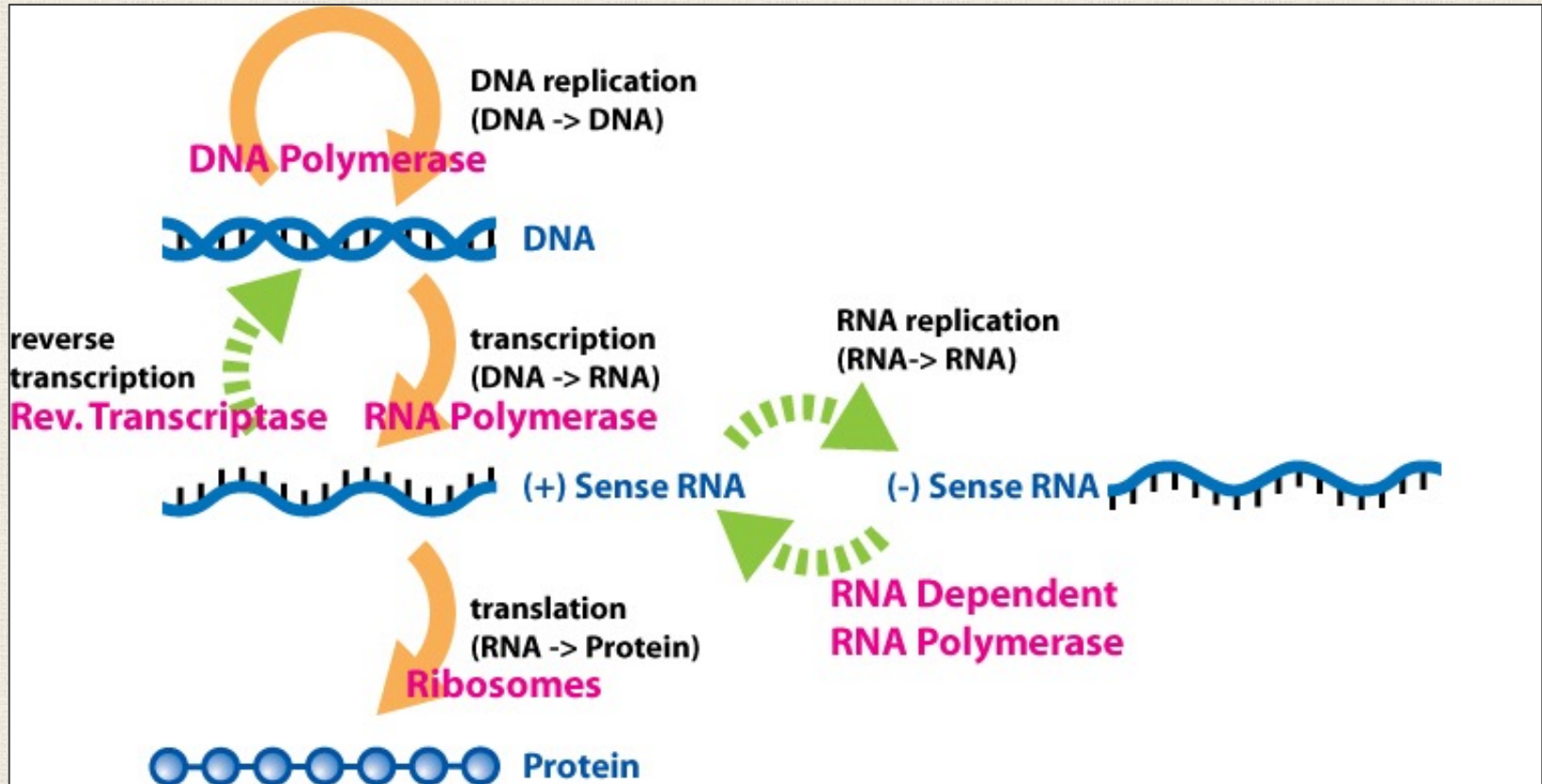
Genome Sequencing Had a Revolution, But Proteins are Still Waiting

Although we can read long genomes with 10 billion base pairs, isolating and reading proteins is very difficult. *For now...*

Instead, we will take a middle ground and use **RNA-sequencing**: reading the RNA present in a given biological sample as a proxy for protein levels.

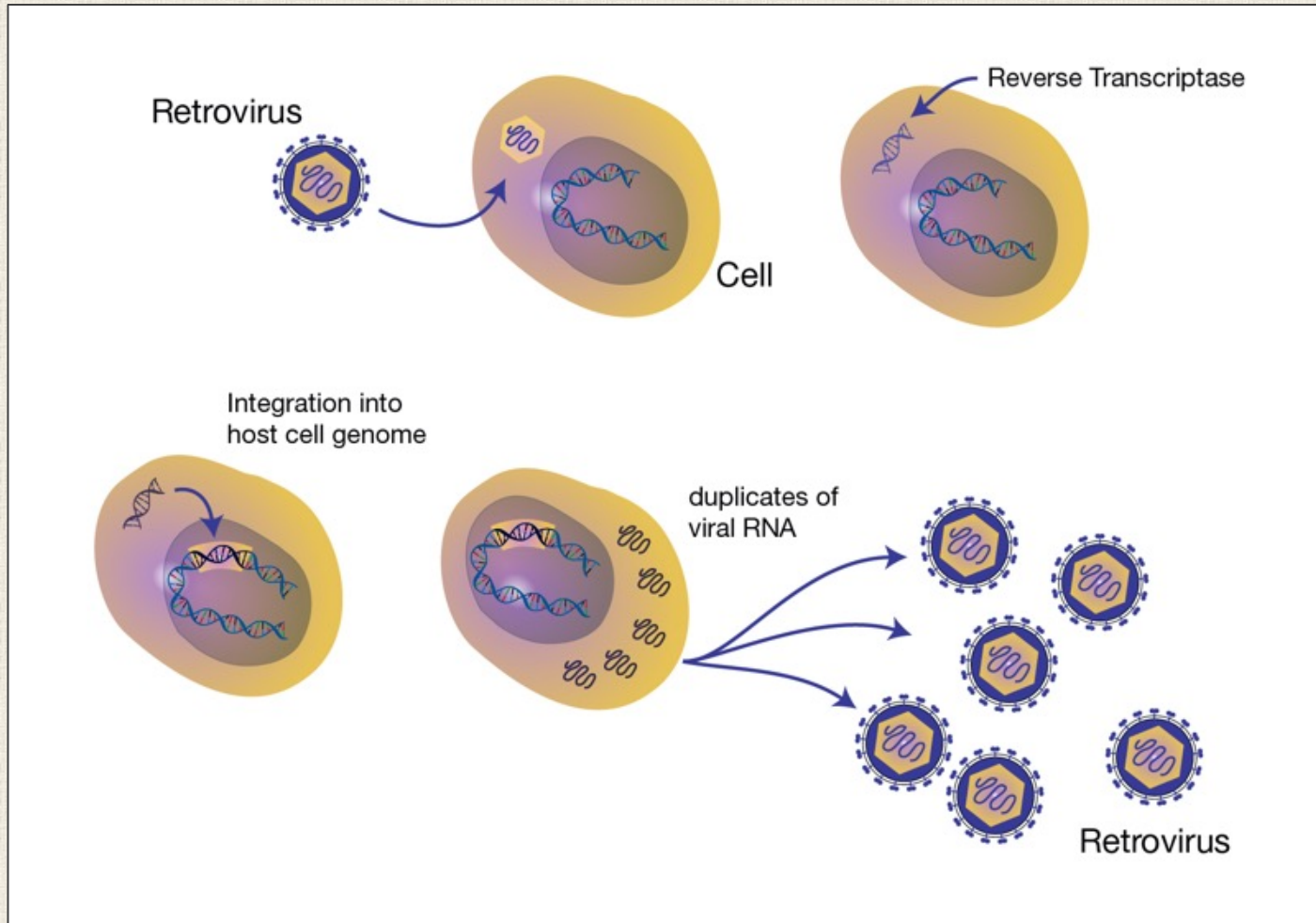
But why is reading RNA easier than reading the protein that it produces?

It's Called a "Dogma" for a Reason



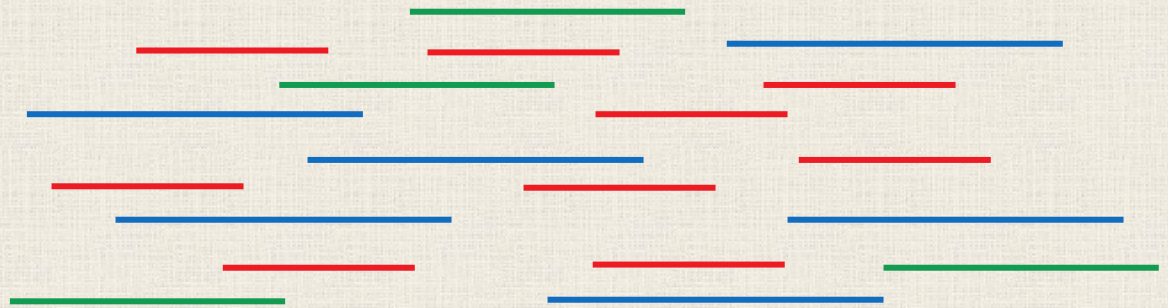
https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology#/media/File:Extended_Central_Dogma_with_Enzymes.jpg

Retroviruses Use Reverse Transcriptase to Convert their RNA to DNA



RNA Sequencing = RNA fragments + DNA Transcriptase + DNA Sequencing

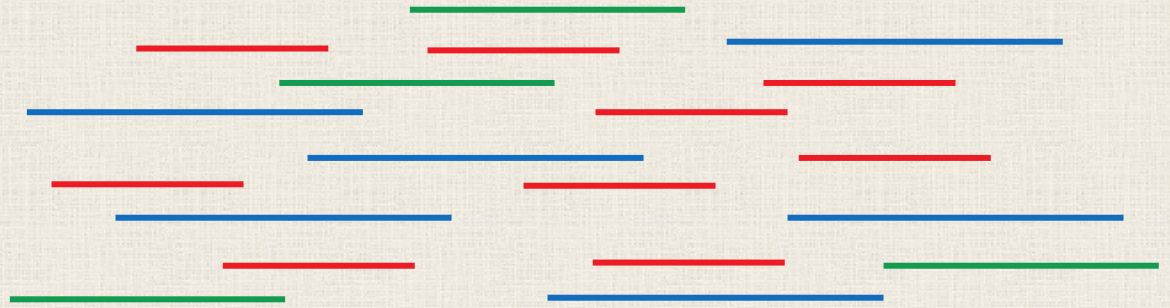
Extract many copies of
different RNA transcripts
from a sample



Note: The lengths of transcripts vary, and the amount of each transcript varies due to expression.

RNA Sequencing = RNA fragments + DNA Transcriptase + DNA Sequencing

Extract many copies of
different RNA transcripts
from a sample



Fragment into smaller
pieces (to match length
demanded by sequencer)



RNA Sequencing = RNA fragments + DNA Transcriptase + DNA Sequencing

Extract many copies of different RNA transcripts from a sample



Fragment into smaller pieces (to match length demanded by sequencer)



Apply reverse transcriptase, sequence, and infer RNA fragments by complementarity

...ACGGATCAT...

...TACGAGCT...

...UGCCUAGUA...

...AUGCUCGA...

RNA Sequencing = RNA fragments + DNA Transcriptase + DNA Sequencing

So now we have a bunch of RNA fragments corresponding to our sample. What do we do?

Apply reverse transcriptase, sequence, and infer RNA fragments by complementarity

...ACGGATCAT...

...TACGAGCT...

...UGCCUAGUA...

...AUGCUCGA...

PART 1: SPLICE JUNCTION IDENTIFICATION

We Have RNA ... So What Do We Do?

Once again, we use DNA to help us ...

- **Input:** a collection of RNA strings.
- **Output:** for each RNA string, a collection of locations where the reverse transcription of these strings (or their reverse complements) “align well” against the reference genome.

We Have RNA ... So What Do We Do?

Once again, we use DNA to help us ...

- **Input:** a collection of RNA strings.
- **Output:** for each RNA string, a collection of locations where the reverse transcription of these strings (or their reverse complements) “align well” against the reference genome.

STOP: Where have we seen this problem before?

We Have RNA ... So What Do We Do?

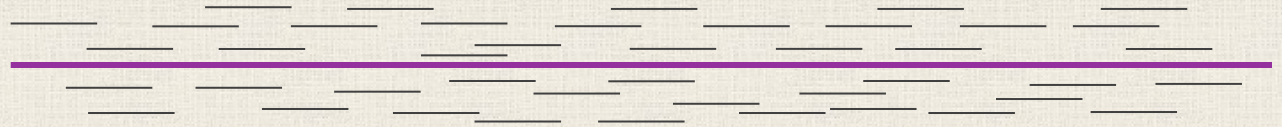
Once again, we use DNA to help us ...

- **Input:** a collection of RNA strings.
- **Output:** for each RNA string, a collection of locations where the reverse transcription of these strings (or their reverse complements) “align well” against the reference genome.

Answer: It seems like it is just read mapping!

Aligning Sequenced Fragments to a Reference Genome

Aligning fragments
against reference
genome



STOP (biologists): There is a major flaw in this picture ... what is it?

Aligning Sequenced Fragments to a Reference Genome

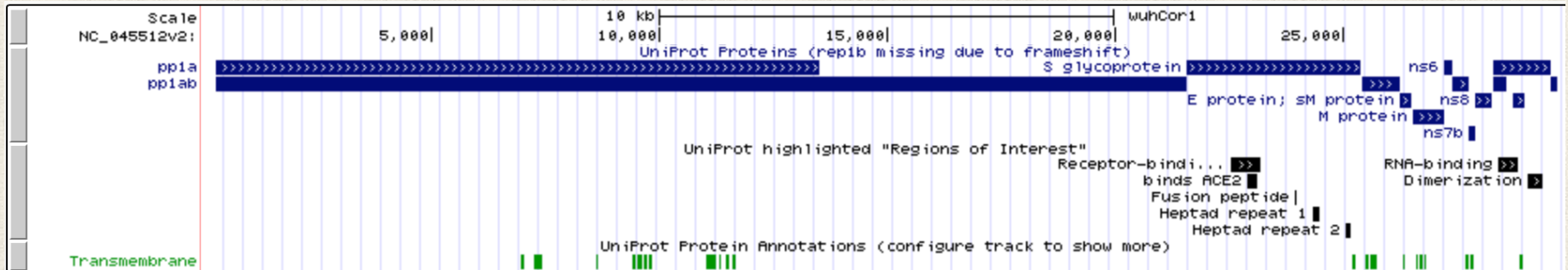
Aligning fragments
against **reference**
genome



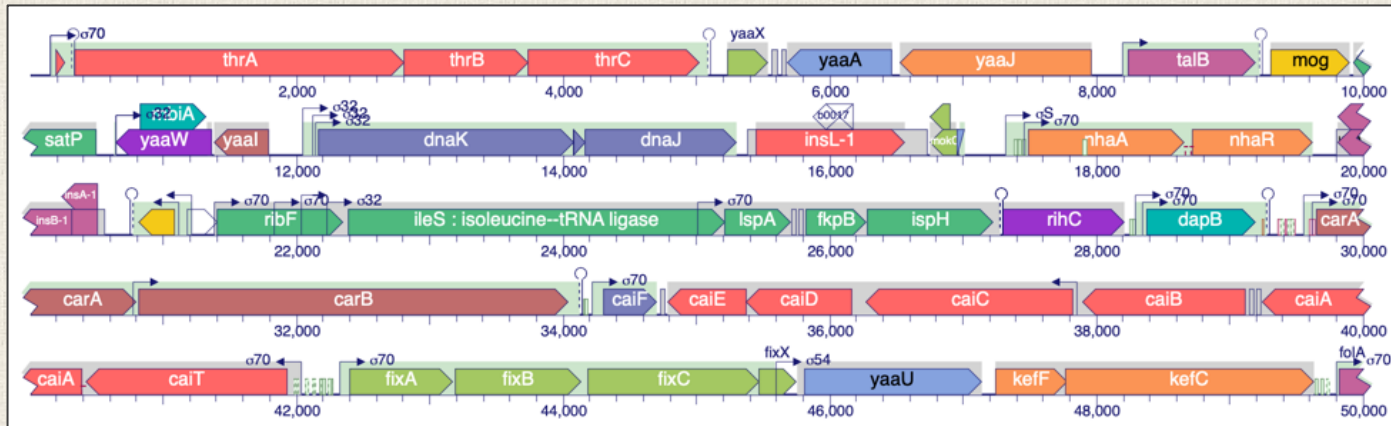
STOP (biologists): There is a major flaw in this picture ... what is it?

Answer: Most of the human genome (98-99%) is not made of genes!

Viruses and Prokaryotes Have Dense Genomes



SARS-CoV

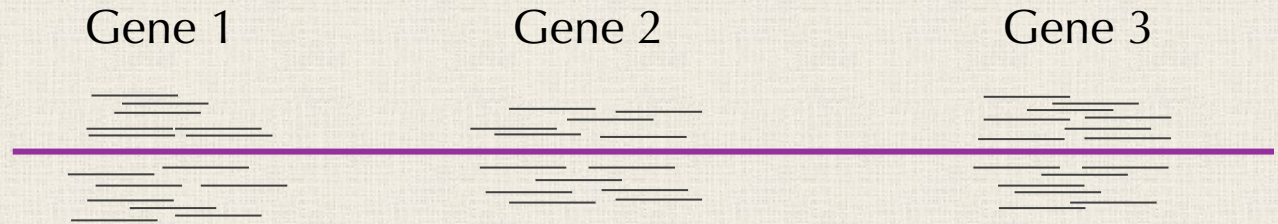


E. Coli (first 50k bp)

Courtesy: EcoCyc

Human Genes are Sparse, So We Need an Updated Picture

Aligning fragments
against **reference**
genome

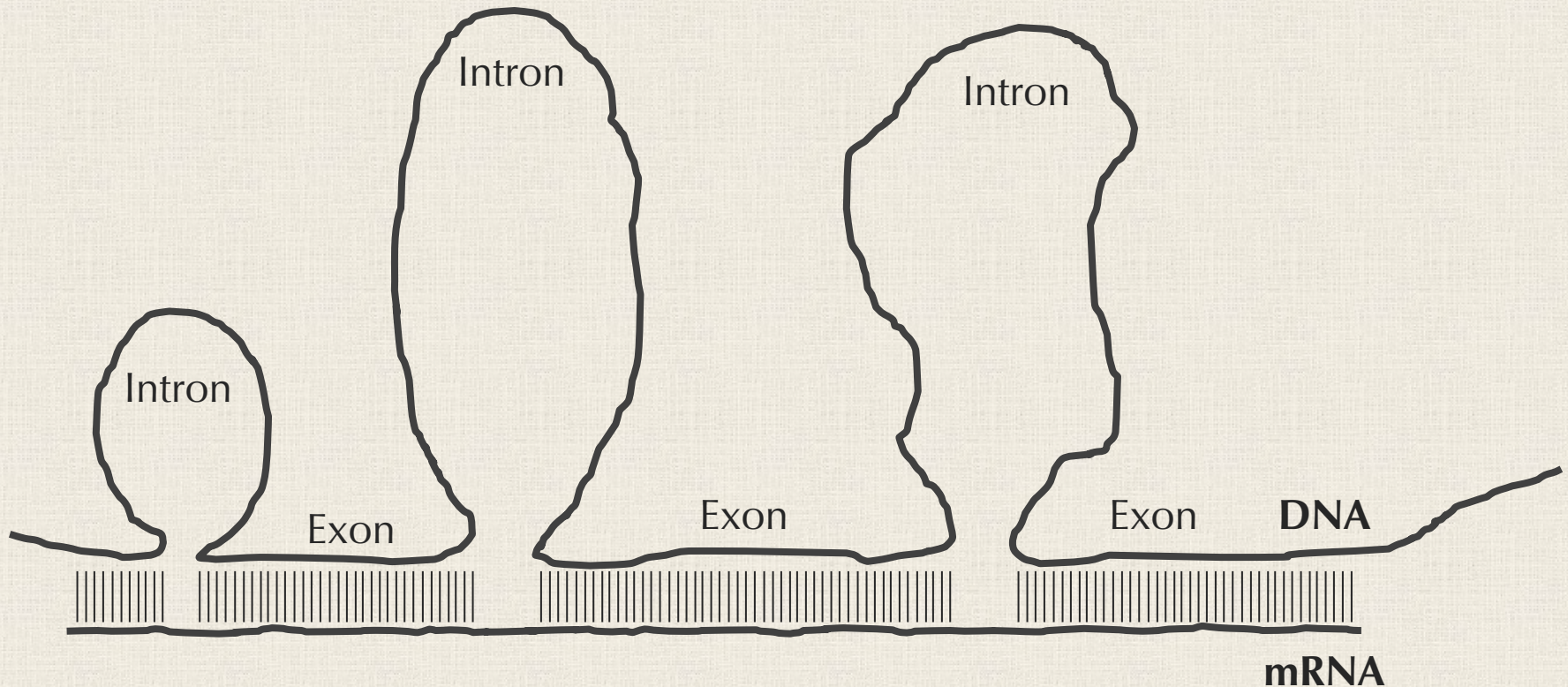


First ~100M bp of human chromosome 1

STOP (biologists): This is still totally wrong. Why?

The Problem is "Split Genes"

1993 Nobel: in eukaryotes, most genes are split between **exons** (coding) and **introns** (non-coding).



Borrowing a Slide from Carl Kingsford

Prokaryotic (bacterial) genes look like this:



Eukaryotic genes usually look like this:



Introns are
thrown away

mRNA:

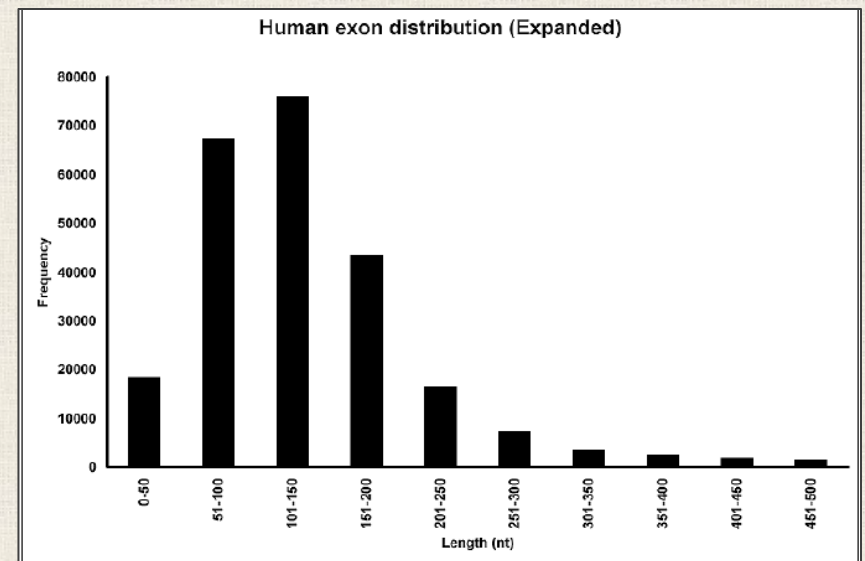
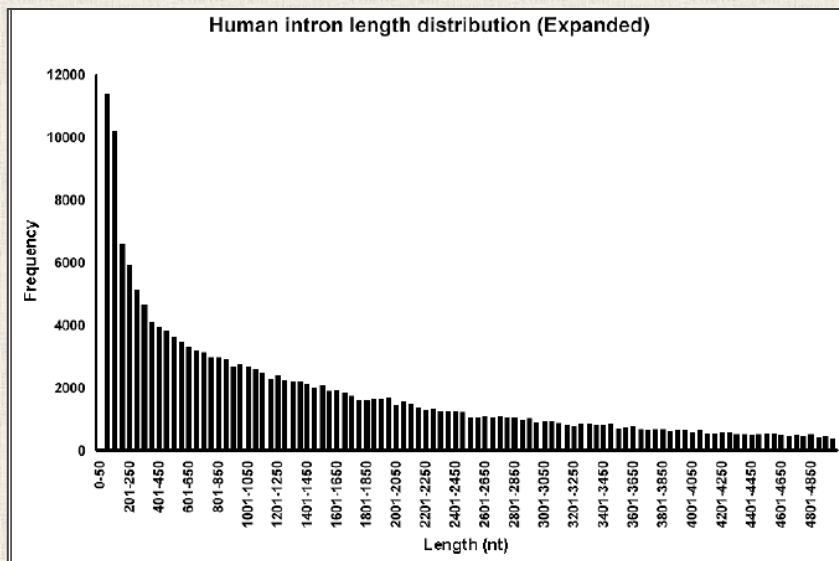


Exons are concatenated together

This spliced RNA is what is
translated into a protein.

Exon/Intron Statistics

- Genes have on average ~9 exons (and ~8 introns).
- Introns tend to be longer than exons.
- Exon lengths are also skewed shorter.



https://www.researchgate.net/figure/a-Frequency-of-intron-length-distributions-for-human-genome-a-and-its-expansion-b_fig4_7498905

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome

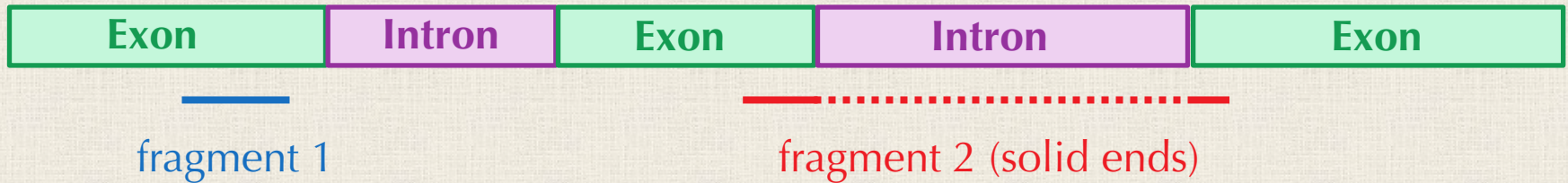


So, we have two possibilities for an RNA fragment.

1. The fragment falls entirely within an exon.

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome

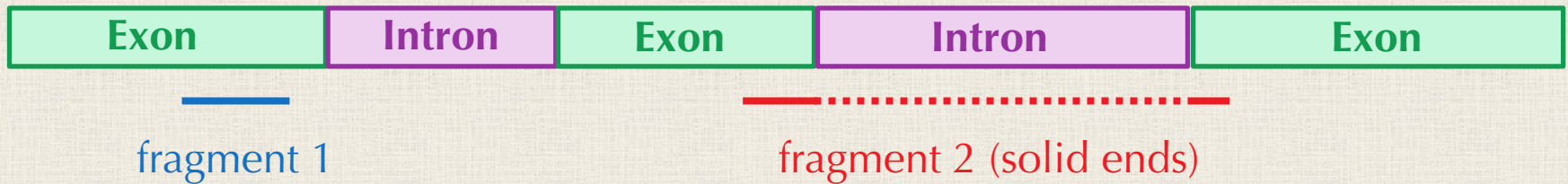


So, we have two possibilities for an RNA fragment.

1. The fragment falls entirely within an exon.
2. The fragment spans exons across an intron(s).

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome



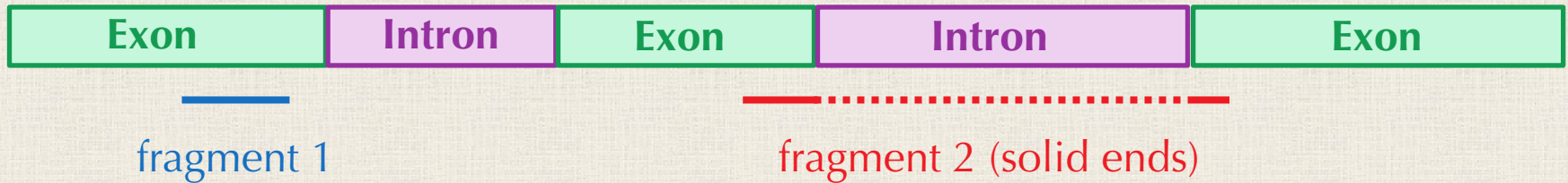
So, we have two possibilities for an RNA fragment.

1. The fragment falls entirely within an exon.
2. The fragment spans exons across an intron(s).

STOP: Which of these will align well against the reference genome?

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome



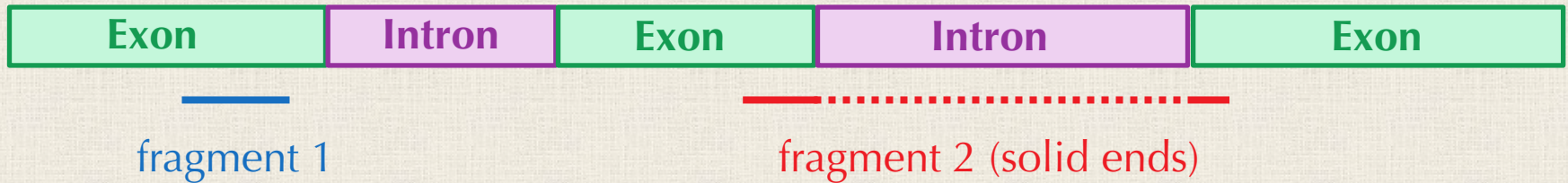
So, we have two possibilities for an RNA fragment.

1. The fragment falls entirely within an exon.
2. The fragment spans exons across an intron(s).

Answer: Type 1 will align against the reference, but type 2 does not occur contiguously in the genome.

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome



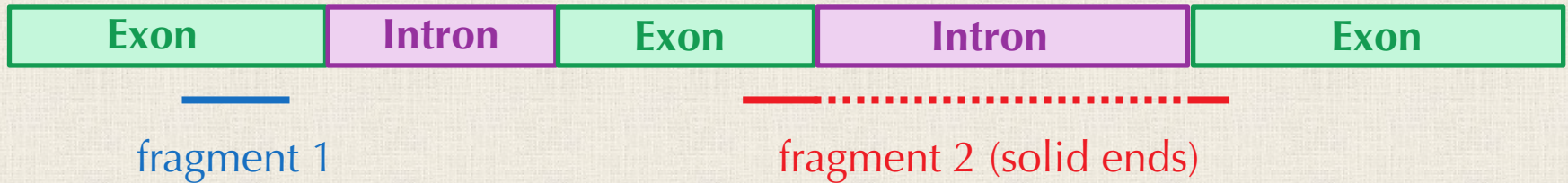
So, we have two possibilities for an RNA fragment.

1. The fragment falls entirely within an exon.
2. The fragment spans exons across an intron(s).

Splice junction: the boundary between an exon and an intron.

Two Possibilities for Where Fragment Aligns

Hypothetical gene in reference genome



This is a feature, not a bug – after finding all the “type 1” reads that align well, the remaining fragments can help us find splice junctions!

Splice junction: the boundary between an exon and an intron.

An Overview of “TopHat” for Splice Junction Discovery

www.ncbi.nlm.nih.gov › [pmc](#) › [articles](#) › [PMC2672628](#) ▼

TopHat: discovering splice junctions with RNA-Seq - NCBI

by C Trapnell - 2009 - [Cited by 9942](#) - [Related articles](#)

Mar 16, 2009 - **TopHat** maps reads to splice sites in a mammalian genome at a rate of ~2.2 million reads per CPU hour. Rather than filtering out possible splice ...

[INTRODUCTION](#) · [METHODS](#) · [RESULTS](#) · [DISCUSSION](#)

Step 1: Assemble Exons

1. Align everything that aligns to the reference genome (and form a consensus of fragments).

An Overview of “TopHat” for Splice Junction Discovery

www.ncbi.nlm.nih.gov › [pmc](#) › [articles](#) › [PMC2672628](#) ▼

TopHat: discovering splice junctions with RNA-Seq - NCBI

by C Trapnell - 2009 - [Cited by 9942](#) - [Related articles](#)

Mar 16, 2009 - **TopHat** maps reads to splice sites in a mammalian genome at a rate of ~2.2 million reads per CPU hour. Rather than filtering out possible splice ...

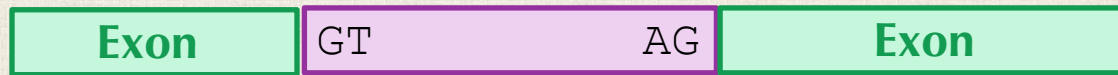
[INTRODUCTION](#) · [METHODS](#) · [RESULTS](#) · [DISCUSSION](#)

Step 1: Assemble Exons

1. Align everything that aligns to the reference genome (and form a consensus of fragments).
2. If we see a gap $< \sim 70$ nt, then join the two fragments, since odds are that this is not an intron.

An Overview of “TopHat” for Splice Junction Discovery

Step 2: Find splice junctions with “type 2” fragments

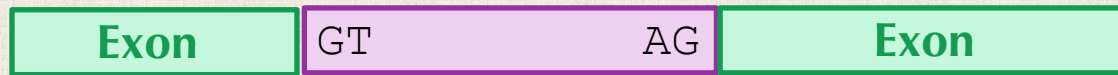


Type 2 fragment alignment

98% of introns start with GT and end with AG, so we can find all such candidate introns between exons and try to align type 2 fragments against them.

An Overview of “TopHat” for Splice Junction Discovery

Step 2: Find splice junctions with “type 2” fragments

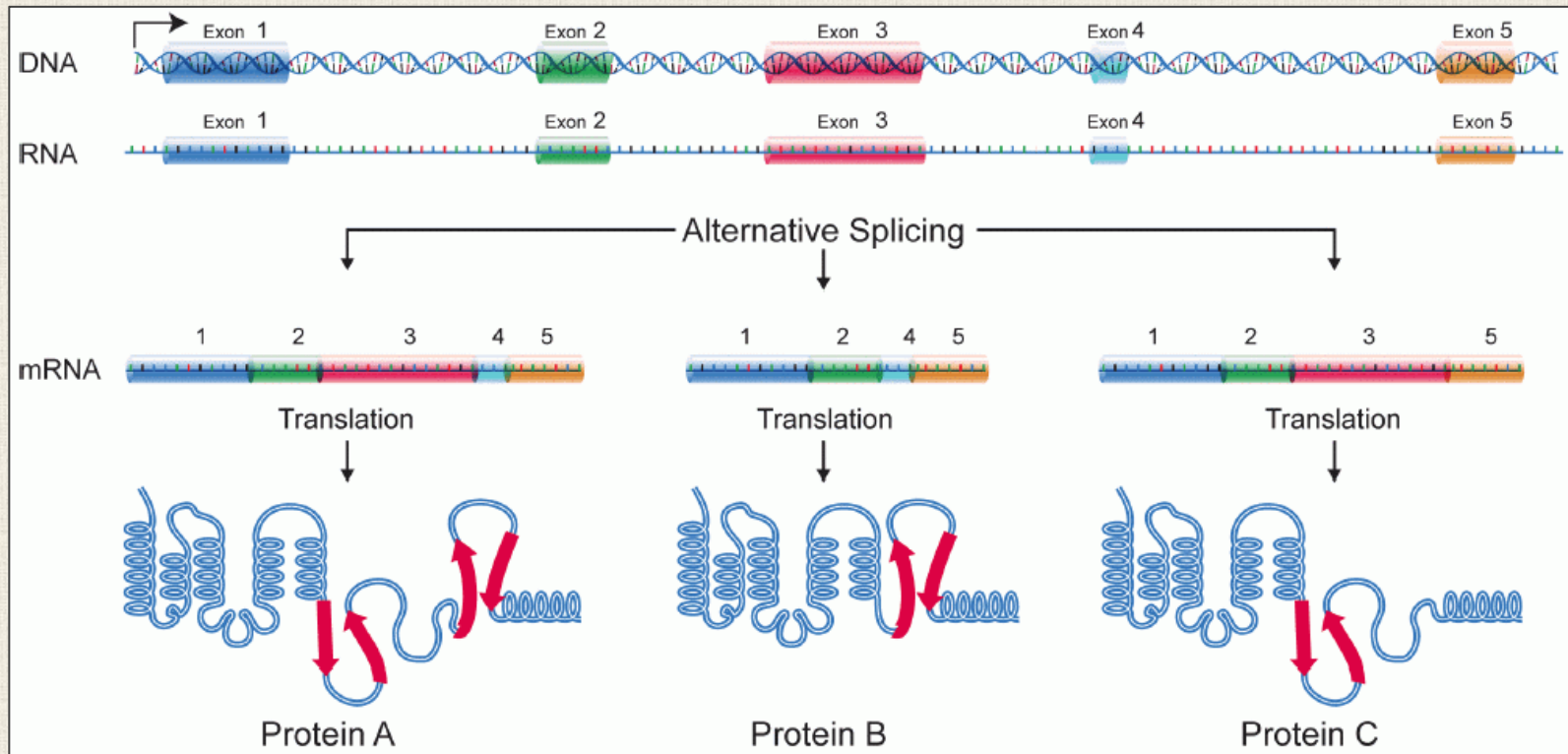


Type 2 fragment alignment

98% of introns start with GT and end with AG, so we can find all such candidate introns between exons and try to align type 2 fragments against them.

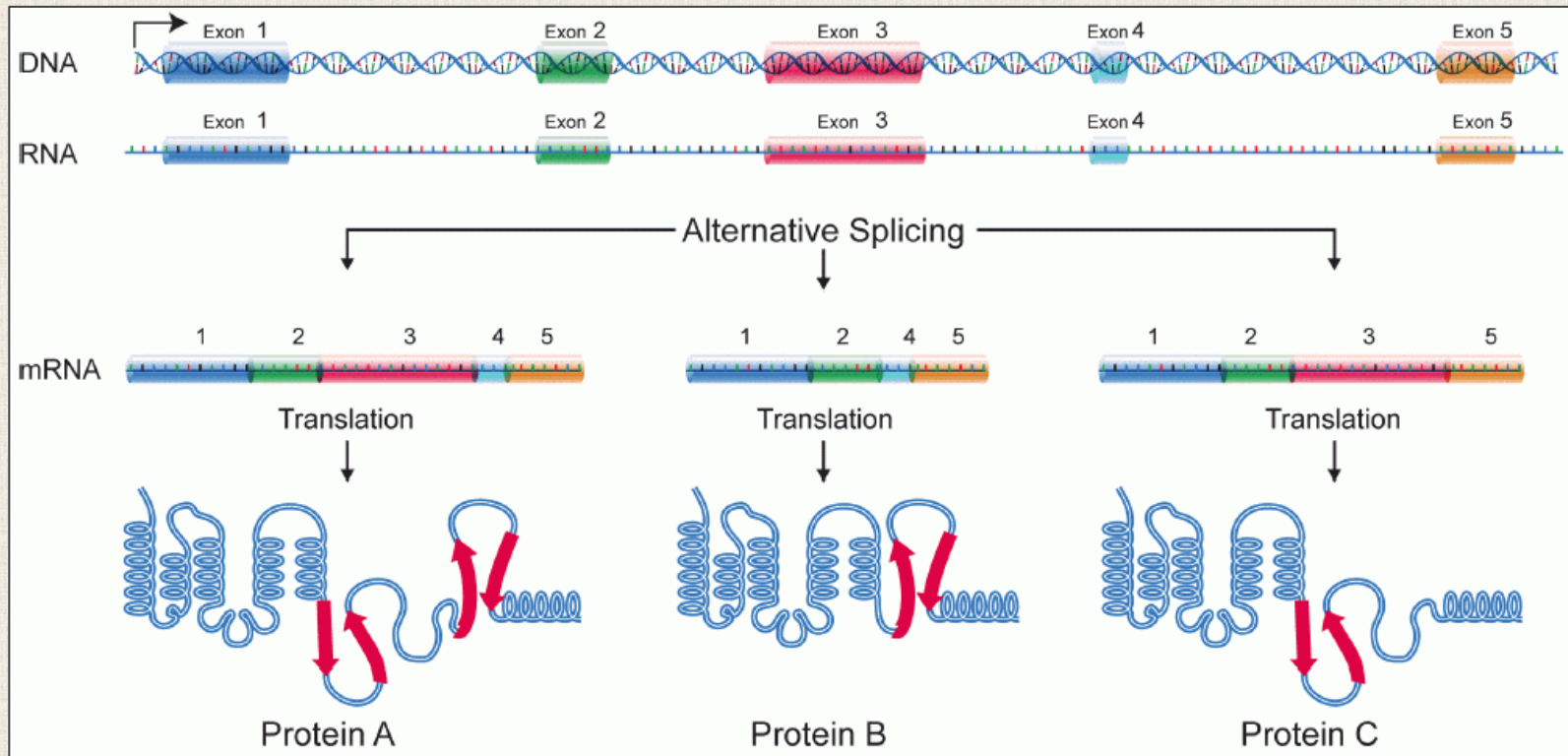
STOP (biologists): Why is this wrong?

Just Because Exons are Consecutive Doesn't Mean They Are Spliced Together



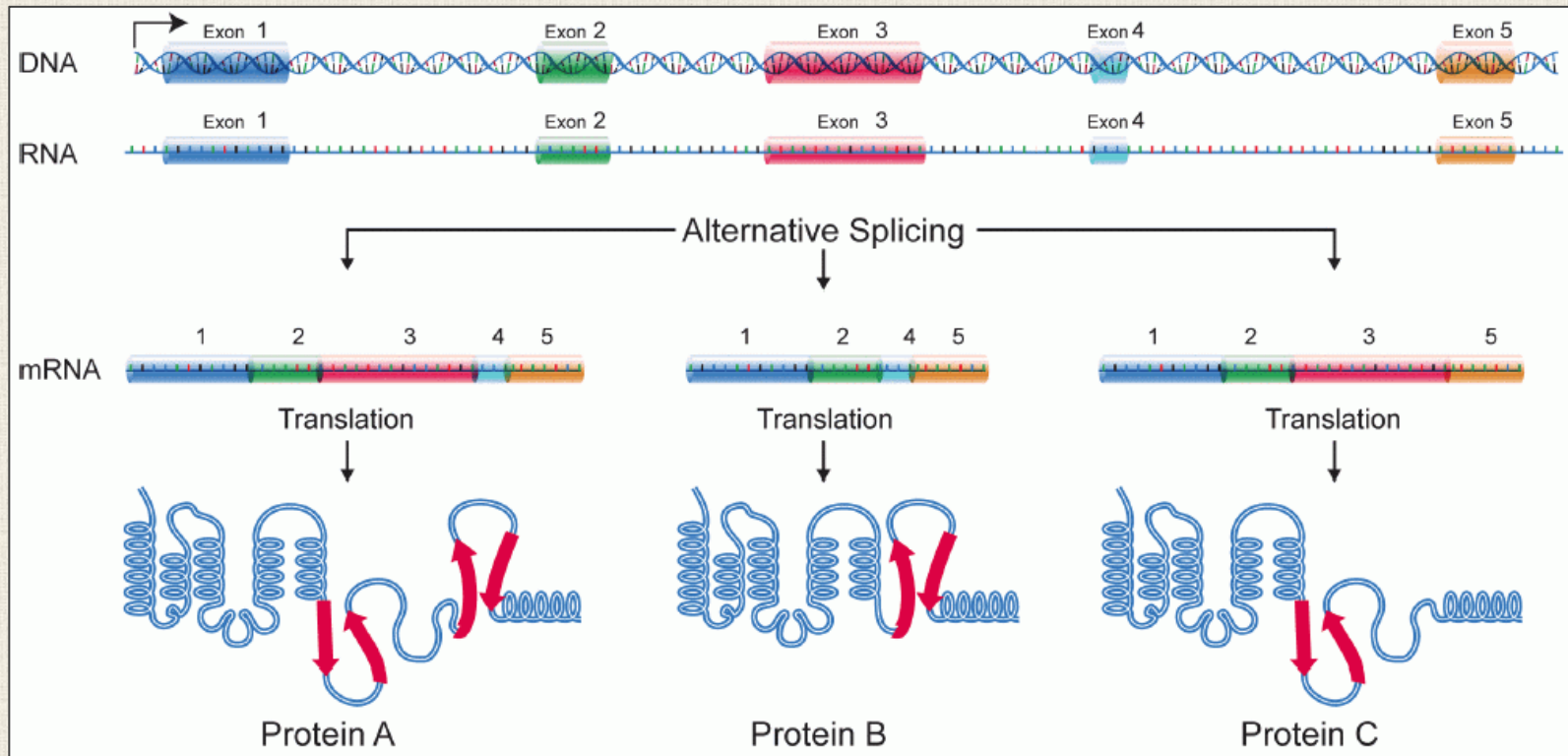
Alternative splicing: exons can be chained in different ways to produce multiple protein **isoforms**.

Just Because Exons are Consecutive Doesn't Mean They Are Spliced Together



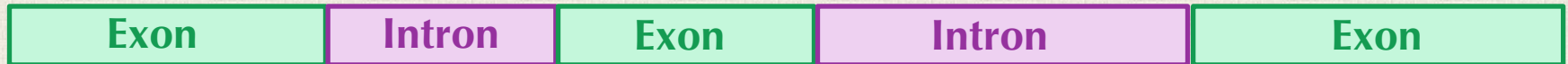
Wang et al., 2008: alternative splicing may affect as many as 95% of human genes.

Just Because Exons are Consecutive Doesn't Mean They Are Spliced Together



Ponomarenko et al., 2016: there could be between 600,000 and 6 million human isoforms.

Hunting for Splice Junctions



STOP: How can we use our RNA fragments to find splicing junctions?

Hunting for Splice Junctions



STOP: How can we use our RNA fragments to find splicing junctions?

Answer: Perform a special “spliced” alignment of type 2 fragments against the ends of “nearby” exons.

Hunting for Splice Junctions



STOP: In the above picture, which exon pairs do we conclude are splice junctions?

Hunting for Splice Junctions



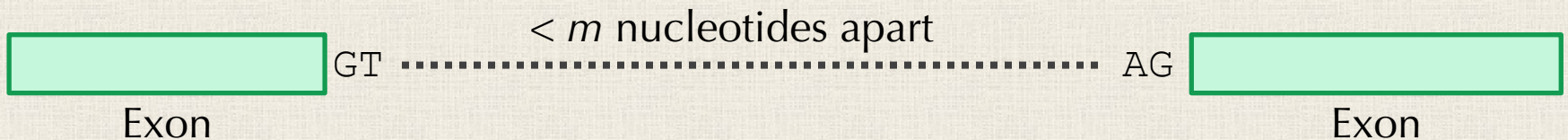
STOP: In the above picture, which exon pairs do we conclude are splice junctions?

Answer: Exons 1 and 3, as well as exons 2 and 3. But exons 1 and 2 aren't a splice junction.

Performing a Spliced Alignment

Step 2: Find splice junctions with “type 2” fragments

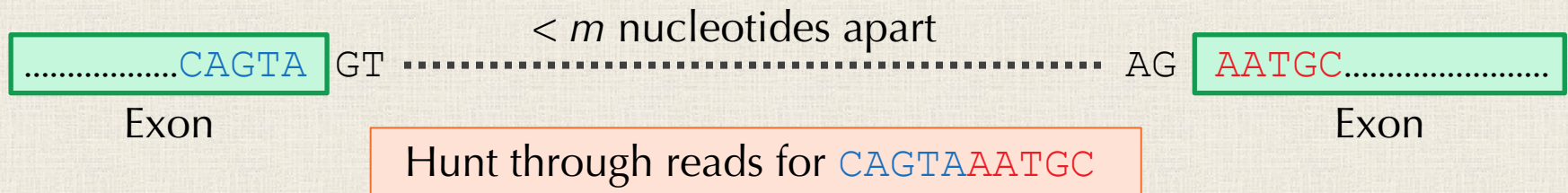
1. For every exon produced in step 1, use **GT-AG** rule to find all potential neighbor exons up to m nucleotides downstream ($m \sim 20\text{k bp}$ in practice).



Performing a Spliced Alignment

Step 2: Find splice junctions with “type 2” fragments

1. For every exon produced in step 1, use **GT-AG** rule to find all potential neighbor exons up to m nucleotides downstream ($m \sim 20k$ bp in practice).
2. Form $2k$ -mer x by joining k -mers ($k \sim 5$ bp) at ends of two exons and search through all type 2 RNA-seq reads for exact matches against x .



Performing a Spliced Alignment

Step 2: Find splice junctions with “type 2” fragments

1. For every exon produced in step 1, use **GT-AG** rule to find all potential neighbor exons up to m nucleotides downstream ($m \sim 20\text{k bp}$ in practice).
2. Form $2k$ -mer x by joining k -mers ($k \sim 5 \text{ bp}$) at ends of two exons and search through all type 2 RNA-seq reads for exact matches against x .

STOP: Once we find these exact matches, what do we do?

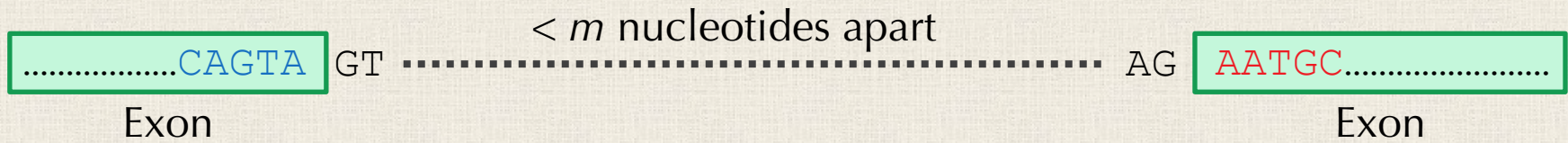
Performing a Spliced Alignment

Step 2: Find splice junctions with “type 2” fragments

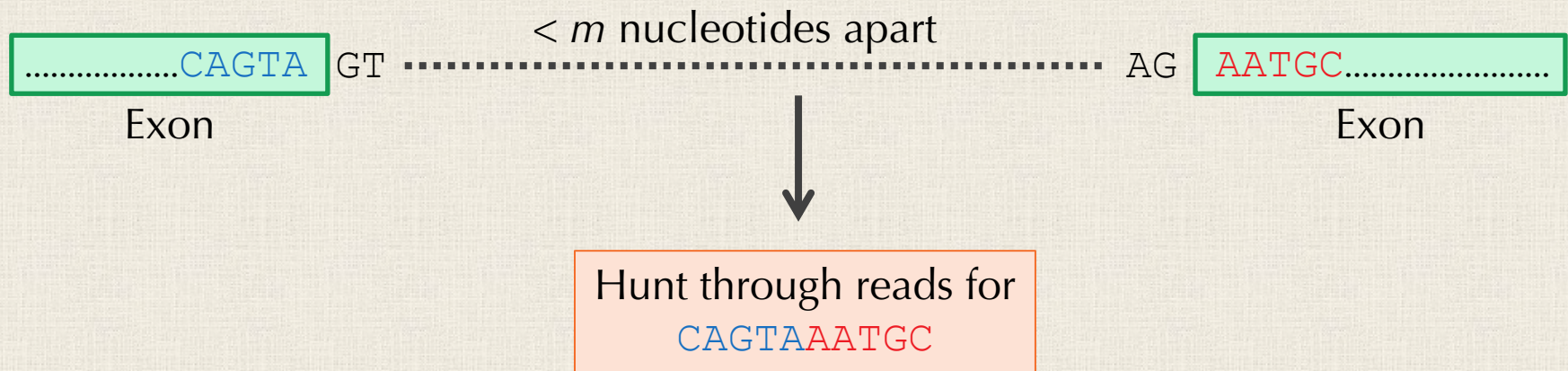
1. For every exon produced in step 1, use **GT-AG** rule to find all potential neighbor exons up to m nucleotides downstream ($m \sim 20\text{k bp}$ in practice).
2. Form $2k$ -mer x by joining k -mers ($k \sim 5 \text{ bp}$) at ends of two exons and search through all type 2 RNA-seq reads for exact matches against x .

Answer: We have found *seeds*, so now we just need to *extend*.

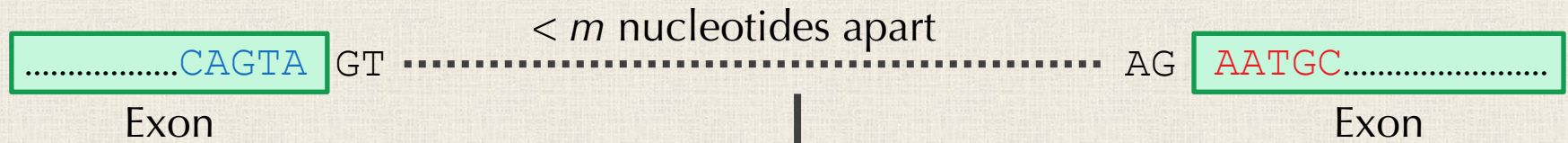
Extending Seed Alignments



Extending Seed Alignments



Extending Seed Alignments



Hunt through reads for
CAGTA AATGC

Perform alignment of concatenated exons against any fragments that match seed. Keep if above a threshold.



Tophat Step 2 in Summary

Step 2: Find splice junctions with “type 2” fragments

1. For every exon produced in step 1, use GT-AG rule to find all potential neighbor exons up to m nucleotides downstream ($m \sim 20\text{k bp}$ in practice).
2. Form $2k$ -mer x by joining k -mers ($k \sim 5 \text{ bp}$) at ends of two exons, and search through all type 2 RNA-seq reads for exact *seed* matches against x .
3. Determine whether any of the seed hits are valid by *extending* these seeds in either direction.

TopHat Step 2 in Summary

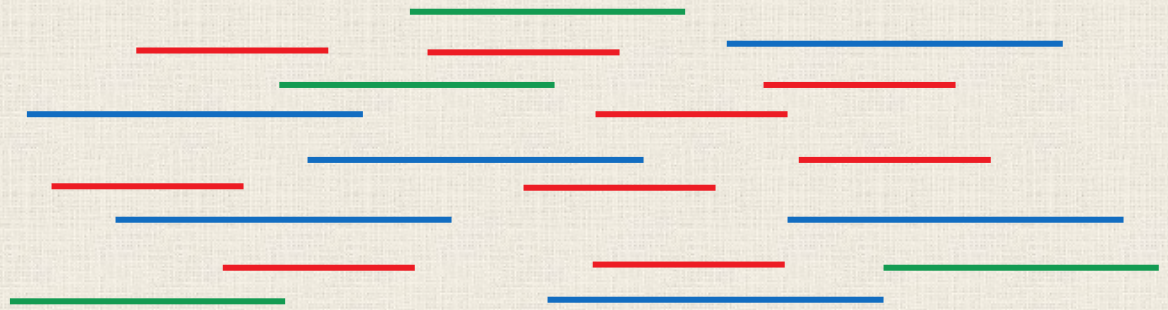
Results: We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20 000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. We describe several challenges unique to *ab initio* splice site discovery from RNA-Seq reads that will require further algorithm development.



PART 2: TRANSCRIPT ASSEMBLY

Recall Our Original Problem

Extract many copies of different RNA transcripts from a sample



Fragment into smaller pieces (to match length demanded by sequencer)



Apply reverse transcriptase, sequence, and infer RNA fragments by complementarity

...ACGGATCAT...

...TACGAGCT...

...UGCCUAGUA...

...AUGCUCGA...

Recall Our Original Problem

Extract many copies of different RNA transcripts from a sample



Goal: Can we re-assemble these transcripts?

Recall Our Original Problem

Extract many copies of different RNA transcripts from a sample



Goal: Can we re-assemble these transcripts?

- **Given:** A collection of RNA-sequencing reads.
- **Find:** The RNA transcripts present in the dataset.

Another Way of Asking this Question

Note that we have already learned two things from the sequencing reads.

- Sequence identity of exons (and location in genome).
- Splice junctions between exons in dataset.

- **Given:** A collection of RNA-sequencing reads.
- **Find:** The RNA transcripts present in the dataset.

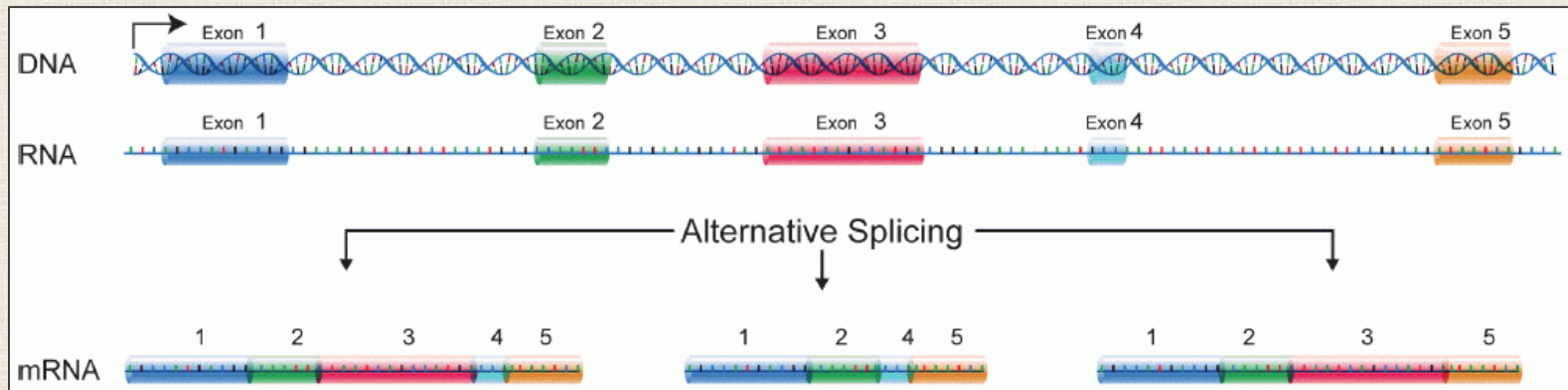
Another Way of Asking this Question

Note that we have already learned two things from the sequencing reads.

- Sequence identity of exons (and location in genome).
- Splice junctions between exons in dataset.

- **Given:** The exons and splice junctions produced from a collection of RNA-sequencing reads.
- **Find:** The RNA transcripts present in the dataset.

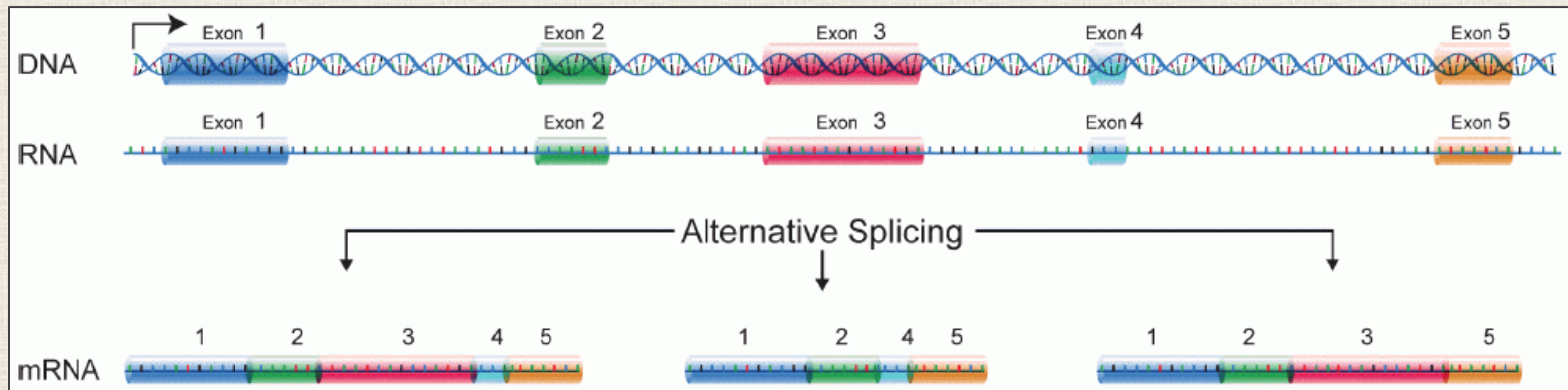
Another Way of Asking this Question



Also, inferring transcripts = knowing exon order.

- **Given:** The exons and splice junctions produced from a collection of RNA-sequencing reads.
- **Find:** The RNA transcripts present in the dataset.

Another Way of Asking this Question



Also, inferring transcripts = knowing exon order.

- **Given:** The exons and splice junctions produced from a collection of RNA-sequencing reads.
- **Find:** The *ordering* of exons for each transcript present in the data.

Another Way of Asking this Question

That is, the following two problems are equivalent (although they aren't well-defined computationally).

- **Given:** A collection of RNA-sequencing reads.
- **Find:** The RNA transcripts present in the dataset.

- **Given:** The exons and splice junctions produced from a collection of RNA-sequencing reads.
- **Find:** The *ordering* of exons for each transcript present in the data.

Cufflinks Uses a Splice Graph to Assemble Transcripts

www.nature.com › nature biotechnology › letters

Transcript assembly and quantification by RNA-Seq reveals ...

by C Trapnell - 2010 - [Cited by 9562](#) - [Related articles](#)

May 2, 2010 - To test **Cufflinks**, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation ...

Given the exons and splice junctions we have inferred, we can form a **splice graph** for each gene:

- **Nodes:** exons
- **Edges:** connect exon x to y with a directed edge if there is a splice junction $x \mid y$.

Cufflinks Uses a Splice Graph to Assemble Transcripts

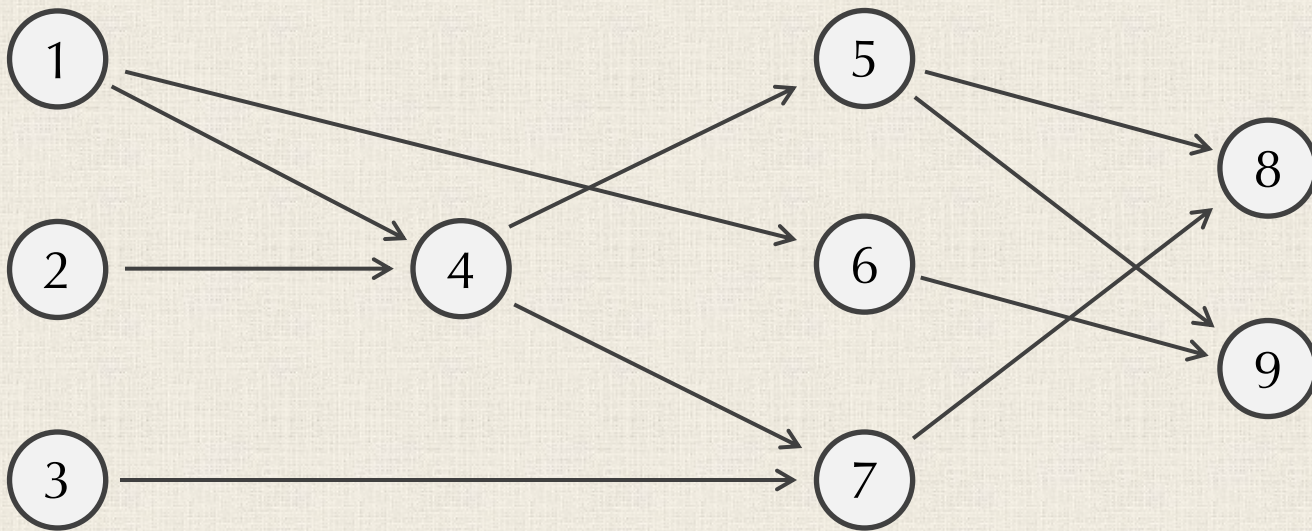
STOP: What type of graph is the splice graph?

Given the exons and splice junctions we have inferred, we can form a **splice graph** for each gene:

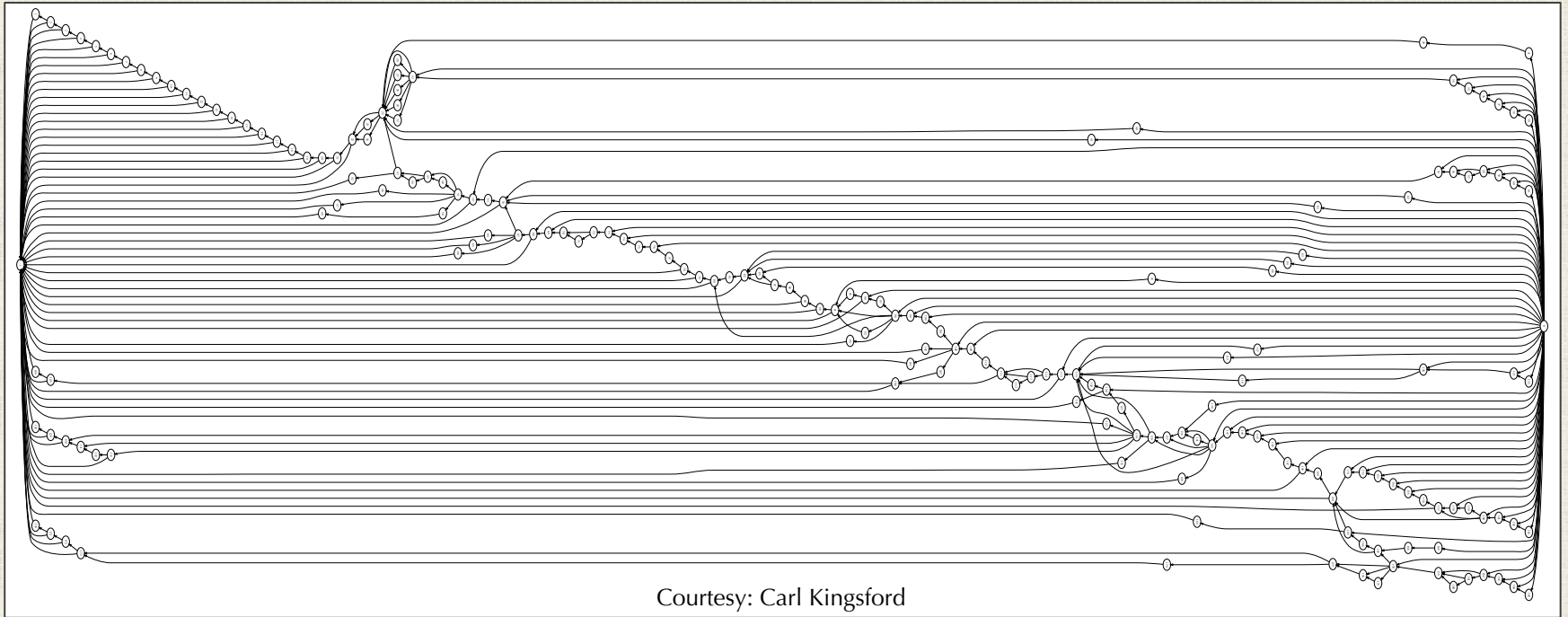
- **Nodes:** exons
- **Edges:** connect exon x to y with a directed edge if there is a splice junction $x \mid y$.

Cufflinks Uses a Splice Graph to Assemble Transcripts

Answer: A DAG – a cycle would mean that order of exons in original gene isn't preserved in RNA.

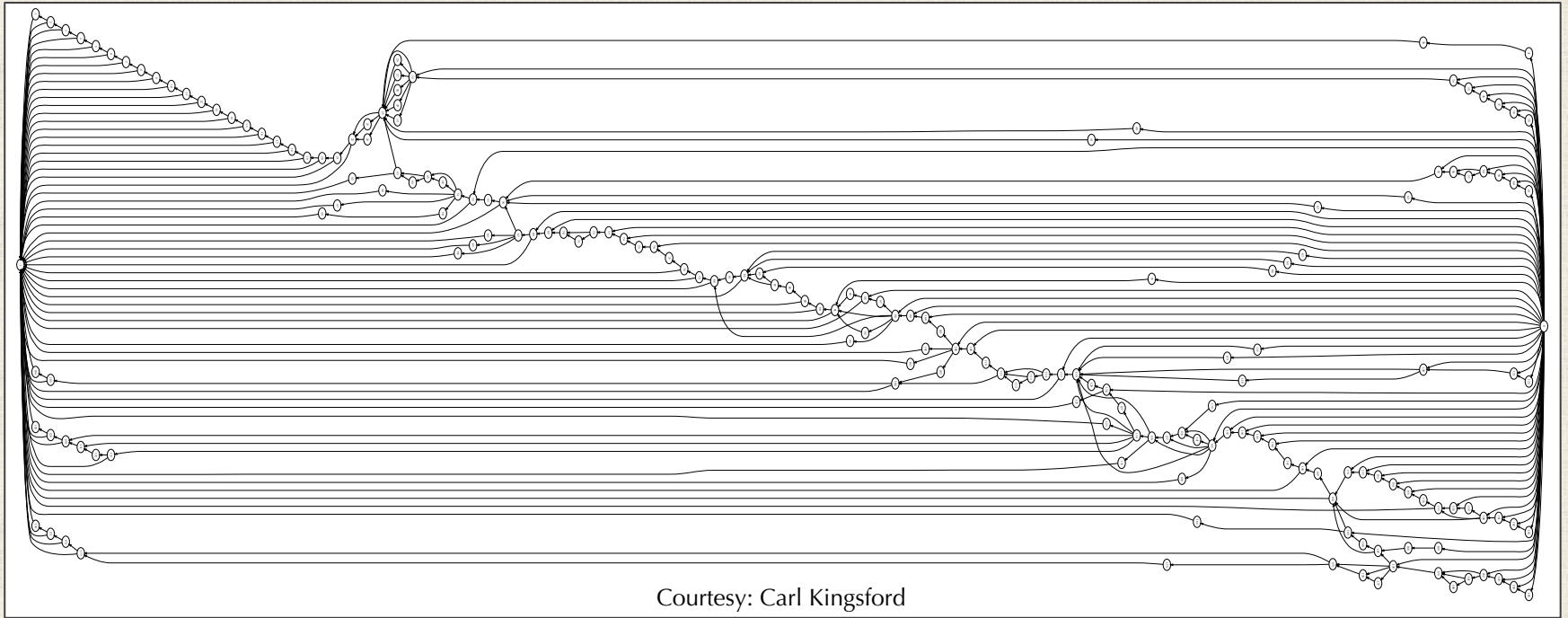


Example Splice Graph



Splice graphs can be complicated for real genes.

Example Splice Graph

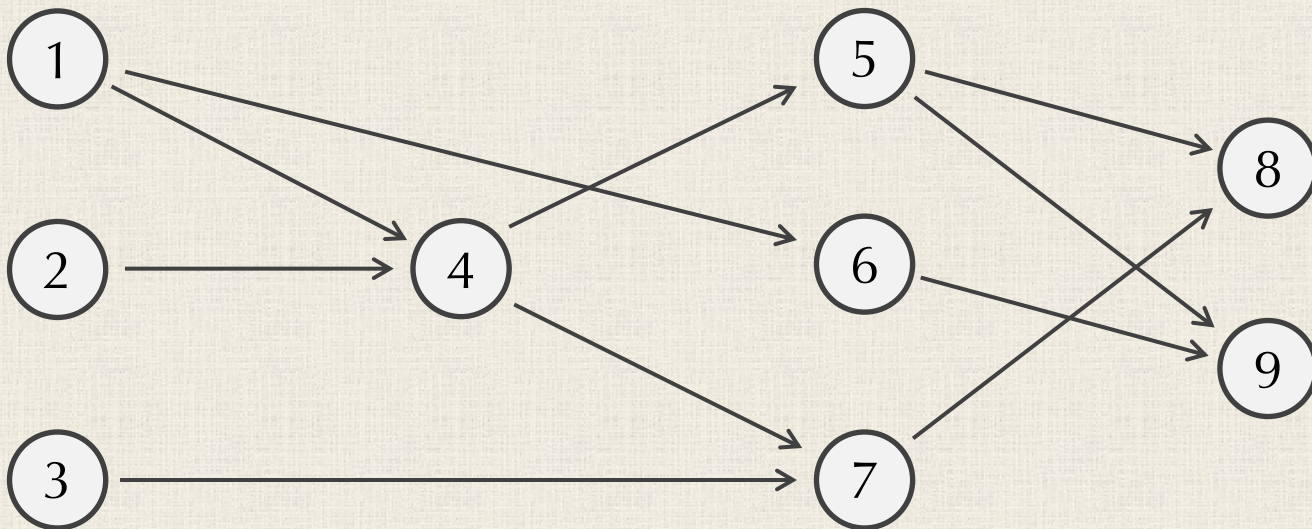


STOP: What are we looking for in this graph if we are trying to reconstruct all transcripts?

Path Edge Covers = Sets of Transcripts

Given a DAG, a **path edge cover** is a collection of paths whose union contains all edges.

Example: The paths $(2, 4, 5, 8)$, $(1, 6, 9)$, $(1, 4, 5, 9)$, $(1, 4, 7, 8)$, $(3, 7, 8)$ form a path edge cover below.



Path Edge Covers = Sets of Transcripts

Given a DAG, a **path edge cover** is a collection of paths whose union contains all edges.

STOP: What kind of path edge cover are we looking for in a splice graph?

- **Given:** The exons and splice junctions produced from a collection of RNA-sequencing reads.
- **Find:** The *ordering* of exons for each transcript present in the data.

Path Edge Covers = Sets of Transcripts

Given a DAG, a **path edge cover** is a collection of paths whose union contains all edges.

Answer: If we follow *parsimony*, then we want a path edge cover to have as few paths as possible!

Minimum Path Edge Cover Problem

- **Input:** A directed graph.
- **Output:** A path edge cover of the graph having as few paths as possible.

Path Edge Covers = Sets of Transcripts

Given a DAG, a **path edge cover** is a collection of paths whose union contains all edges.

Unfortunately, this problem is NP-Hard ... ☹️

Minimum Path Edge Cover Problem

- **Input:** A directed graph.
- **Output:** A path edge cover of the graph having as few paths as possible.

Path Edge Covers = Sets of Transcripts

Given a DAG, a **path edge cover** is a collection of paths whose union contains all edges.

Unfortunately, this problem is NP-Hard ... but it is polynomial-time solvable for a DAG (Dilworth's theorem).

Minimum Path Edge Cover Problem

- **Input:** A directed **acyclic** graph.
- **Output:** A path edge cover of the graph having as few paths as possible.

This Might Seem Simplistic, but ...

www.nature.com › nature biotechnology › letters

Transcript assembly and quantification by RNA-Seq reveals ...

by C Trapnell - 2010 - [Cited by 9562](#) - [Related articles](#)

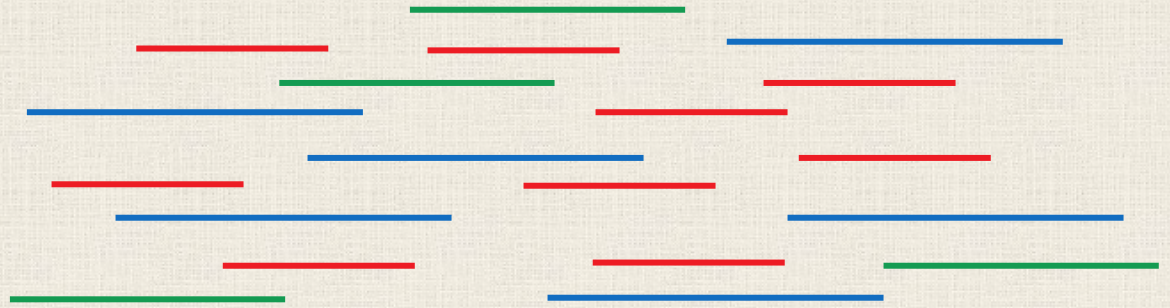
May 2, 2010 - To test **Cufflinks**, we sequenced and analyzed >430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation ...

... (a version of) this approach became the software program Cufflinks, which found over 3,000 new putative mouse transcripts in 2010.

PART 3: TRANSCRIPT QUANTIFICATION

Recall our Original Figure

Extract many copies of different RNA transcripts from a sample



Fragment into smaller pieces (to match length demanded by sequencer)



Apply reverse transcriptase, sequence, and infer RNA fragments by complementarity

...ACGGATCAT...

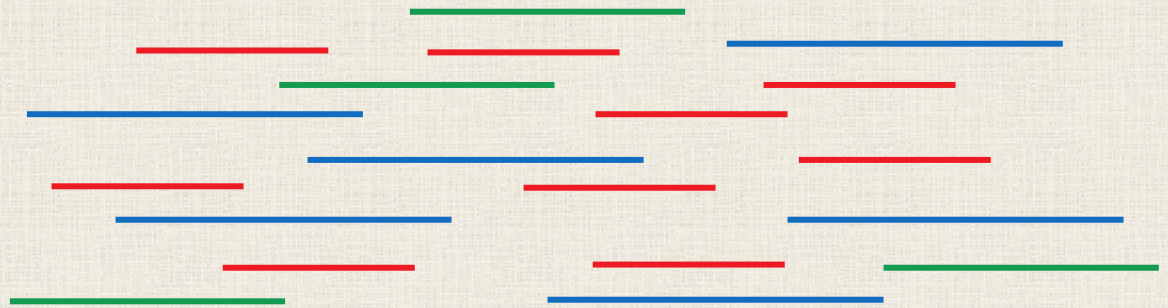
...TACGAGCT...

...UGCCUAGUA...

...AUGCUCGA...

Now That We Know the Transcripts, Can We Determine Their Abundances?

Extract many copies of different RNA transcripts from a sample



Fragment into smaller pieces (to match length demanded by sequencer)



- **Given:** A collection of RNA-sequencing reads and a collection of transcripts inferred from them.
- **Find:** The abundance of each transcript present.

Let's Quantify What We Want to Infer

Extract many copies of different RNA transcripts from a sample



- 9 red transcripts x 500 nt = 4500 nt
- 4 green transcripts x 750 nt = 3000 nt
- 6 blue transcripts x 1000 nt = 6000 nt

Let's Quantify What We Want to Infer

Extract many copies of different RNA transcripts from a sample



- 9 red transcripts x 500 nt = 4500 nt
- 4 green transcripts x 750 nt = 3000 nt
- 6 blue transcripts x 1000 nt = 6000 nt

As percentage of the total, we have

$$\theta = (4500/13500, 3000/13500, 6000/13500)$$

Let's Quantify What We Want to Infer

Extract many copies of different RNA transcripts from a sample



- 9 red transcripts x 500 nt = 4500 nt
- 4 green transcripts x 750 nt = 3000 nt
- 6 blue transcripts x 1000 nt = 6000 nt

As percentage of the total, we have

$$\theta = (4500/13500, 3000/13500, 6000/13500)$$
$$= (0.333, 0.222, 0.444)$$

Tweaking our Problem a Bit

Extract many copies of different RNA transcripts from a sample



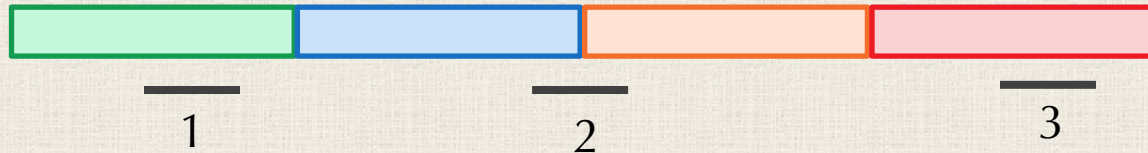
- **Given:** A collection of RNA-sequencing reads and a collection of transcripts inferred from them.
- **Find:** The “abundance vector” θ of the transcripts.

As percentage of the total, we have

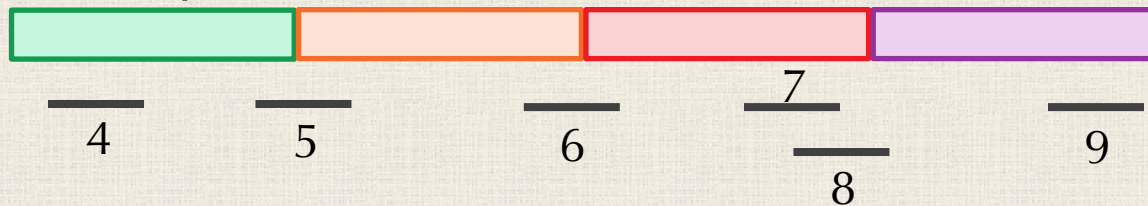
$$\begin{aligned}\theta &= (4500/13500, 3000/13500, 6000/13500) \\ &= (0.333, 0.222, 0.444)\end{aligned}$$

A Simple Example with Three Isoforms

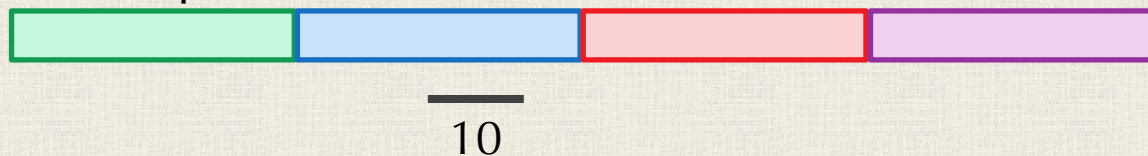
Transcript (a)



Transcript (b)



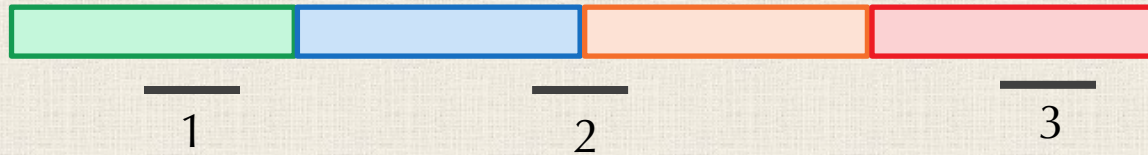
Transcript (c)



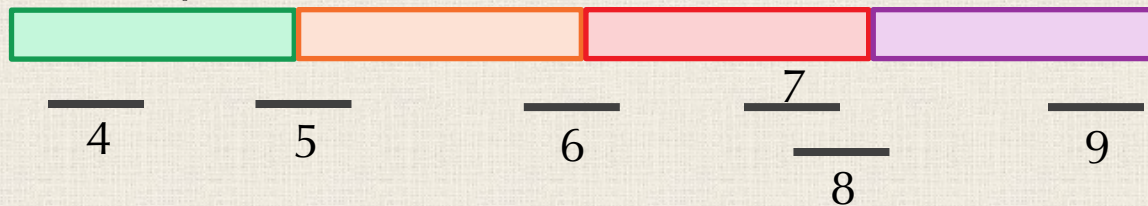
STOP: If we know which fragment each read came from, what is θ ?

A Simple Example with Three Isoforms

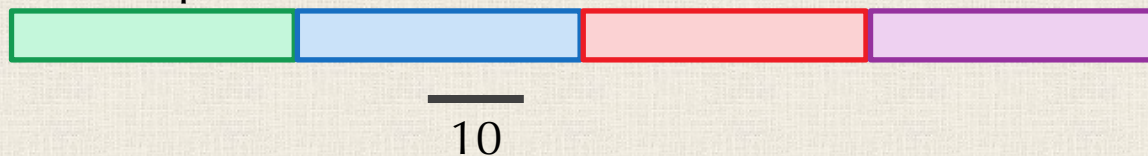
Transcript (a)



Transcript (b)



Transcript (c)



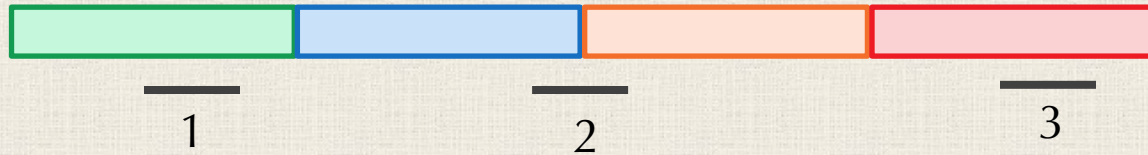
Z

	(a)	(b)	(c)
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	0
9	0	1	0
10	0	0	1

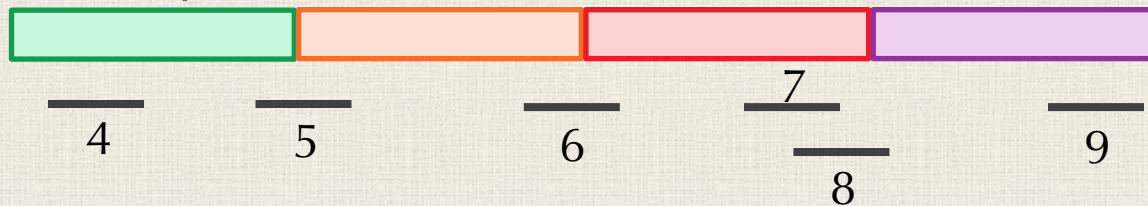
We can log each read's assignment to a transcript in matrix Z.

A Simple Example with Three Isoforms

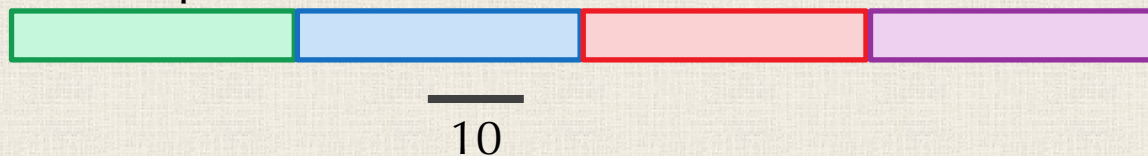
Transcript (a)



Transcript (b)



Transcript (c)



We can log each read's assignment to a transcript in matrix Z .

Z

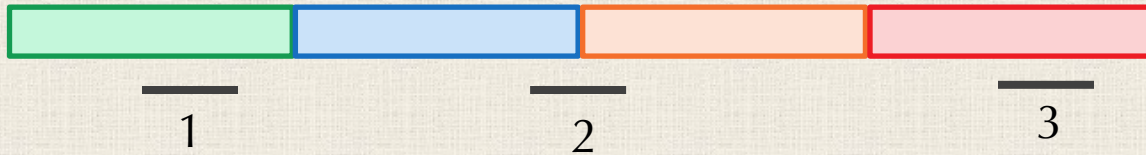
	(a)	(b)	(c)
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	0
9	0	1	0
10	0	0	1

Totals 3 6 1

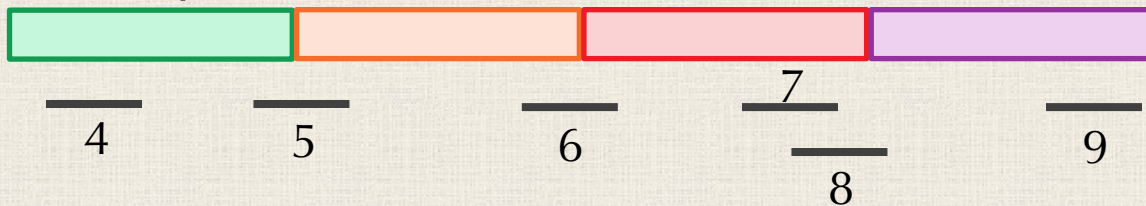
$\theta = (0.3, 0.6, 0.1)$

A Simple Example with Three Isoforms

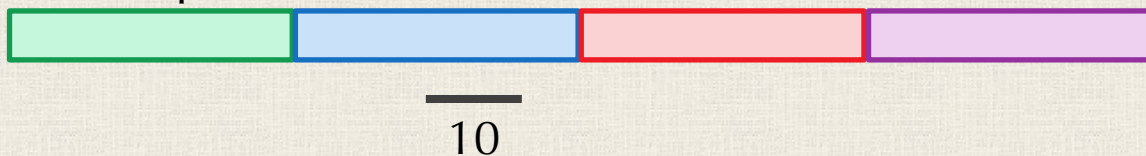
Transcript (a)



Transcript (b)



Transcript (c)



Z

	(a)	(b)	(c)
1	1	0	0
2	1	0	0
3	1	0	0
4	0	1	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	0
9	0	1	0
10	0	0	1

Totals 3 6 1

Key Point: Inferring θ from Z is “trivial”.

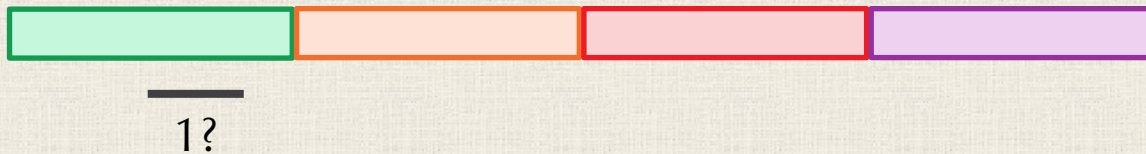
$\theta = (0.3, 0.6, 0.1)$

But Z is *Hidden* from U s ...

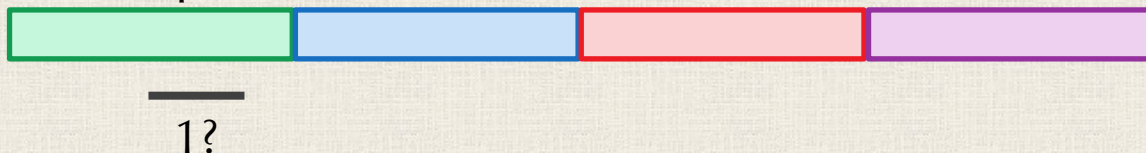
Transcript (a)



Transcript (b)



Transcript (c)



Y

	(a)	(b)	(c)
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			

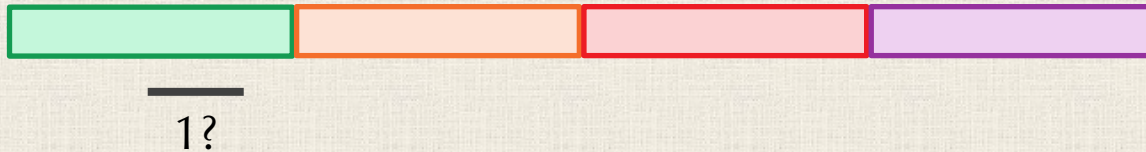
A read is **consistent** with a transcript if it maps well to the transcript.

But Z is *Hidden* from U s ...

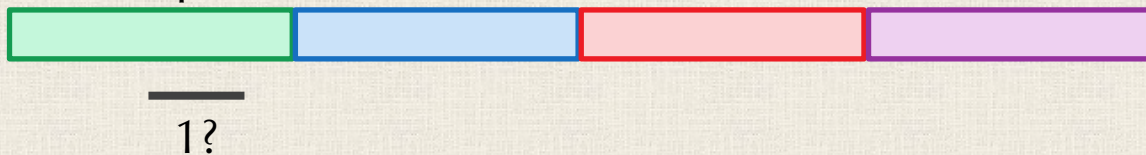
Transcript (a)



Transcript (b)



Transcript (c)



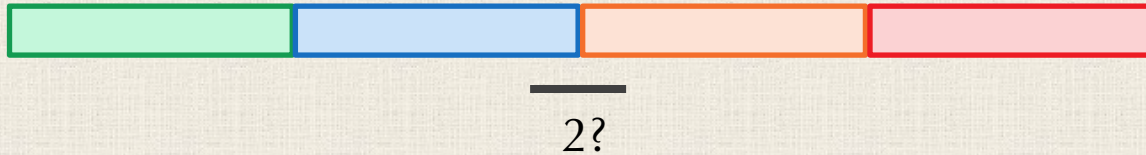
Y

	(a)	(b)	(c)
1	1	1	1
2			
3			
4			
5			
6			
7			
8			
9			
10			

We form matrix Y , where $Y_{i,k} = 1$ if read i is consistent with transcript k .

Identifying Consistent Transcripts for Each Read

Transcript (a)



Transcript (b)



Transcript (c)



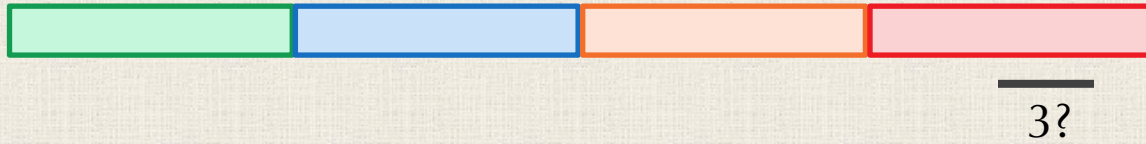
Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3			
4			
5			
6			
7			
8			
9			
10			

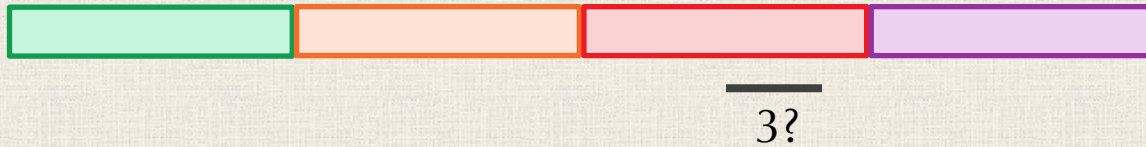
We form matrix Y , where $Y_{i,k} = 1$ if read i is consistent with transcript k .

Identifying Consistent Transcripts for Each Read

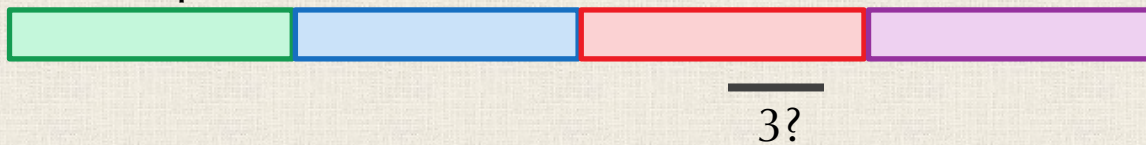
Transcript (a)



Transcript (b)



Transcript (c)



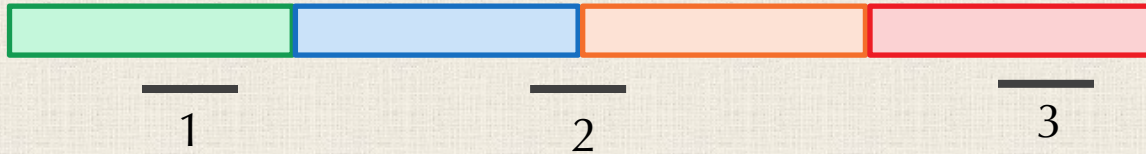
Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4			
5			
6			
7			
8			
9			
10			

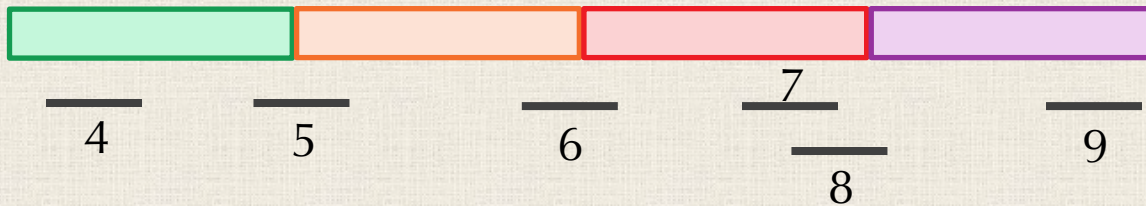
We form matrix Y , where $Y_{i,k} = 1$ if read i is consistent with transcript k .

Identifying Consistent Transcripts for Each Read

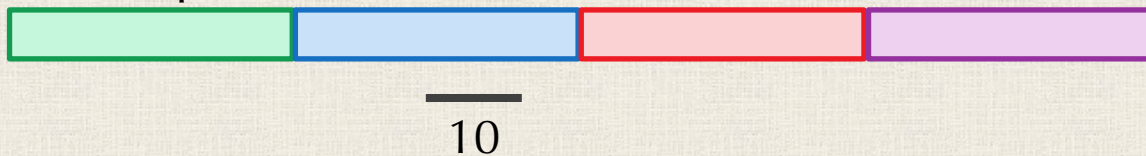
Transcript (a)



Transcript (b)



Transcript (c)



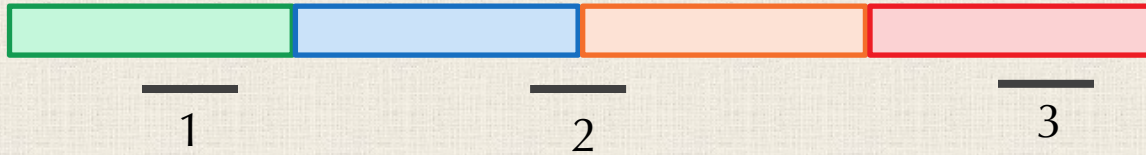
Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4			
5			
6			
7			
8			
9			
10			

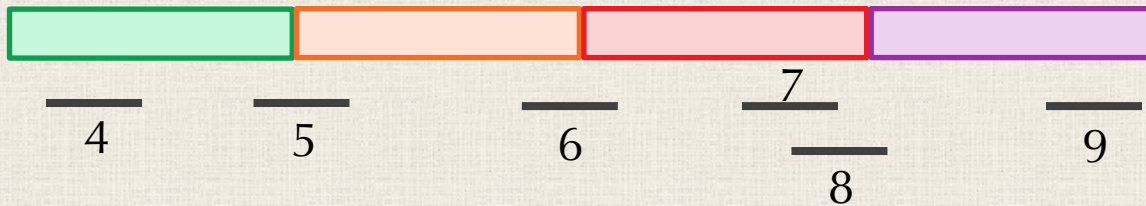
Exercise: Enter the remaining values.

Identifying Consistent Transcripts for Each Read

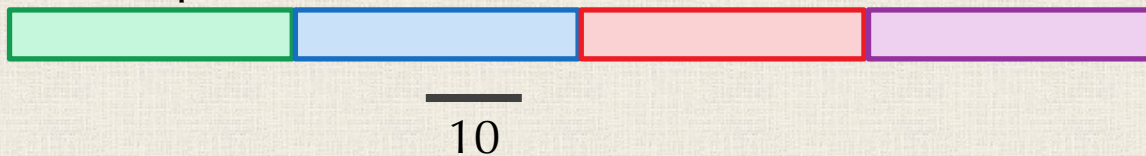
Transcript (a)



Transcript (b)



Transcript (c)



Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

From an Initial Guess of θ to Z

Let's start with an initial guess of $\theta^{(0)} = (1/3, 1/3, 1/3)$ since we know nothing *a priori* about the correct parameters.

STOP: How would we estimate Z from θ ?

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Initial Guess of $\theta \rightarrow Z$

Answer: assign "confidence" of each transcript to each read, based on weighted average of θ :

$$Z^1_{i,k} = Y_{i,k} * \theta^0_k / s_i$$

$$(s_i = \sum_{\text{transcripts } j} Y_{i,j} * \theta^0_j)$$

Z^1

	(a)	(b)	(c)
1	1/3	1/3	1/3
2	1	0	0
3	1/3	1/3	1/3
4	1/3	1/3	1/3
5	0	1	0
6			
7			
8			
9			
10			

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Initial Guess of $\theta \rightarrow Z$

Exercise: Fill in the remaining values of Z^1 .

Z^1

	(a)	(b)	(c)
1	1/3	1/3	1/3
2	1	0	0
3	1/3	1/3	1/3
4	1/3	1/3	1/3
5	0	1	0
6			
7			
8			
9			
10			

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Initial Guess of $\theta \rightarrow Z$

Exercise: Fill in the remaining values of Z^1 .

	Z^1			Y		
	(a)	(b)	(c)	(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	1
2	1	0	0	1	0	0
3	1/3	1/3	1/3	1	1	1
4	1/3	1/3	1/3	1	1	1
5	0	1	0	0	1	0
6	1/2	1/2	0	1	1	0
7	1/3	1/3	1/3	1	1	1
8	0	1/2	1/2	0	1	1
9	0	1/2	1/2	0	1	1
10	1/2	0	1/2	1	0	1

Initial Guess of $\theta \rightarrow Z$

STOP: Is this a reasonable estimate of the real Z ? How can we tell?

	Z^1			Z		
	(a)	(b)	(c)	(a)	(b)	(c)
1	1/3	1/3	1/3	1	0	0
2	1	0	0	1	0	0
3	1/3	1/3	1/3	1	0	0
4	1/3	1/3	1/3	0	1	0
5	0	1	0	0	1	0
6	1/2	1/2	0	0	1	0
7	1/3	1/3	1/3	0	1	0
8	0	1/2	1/2	0	1	0
9	0	1/2	1/2	0	1	0
10	1/2	0	1/2	0	0	1

Initial Guess of $\theta \rightarrow Z$

STOP: Is this a reasonable estimate of the real Z ? How can we tell?

Answer: The totals follow the same pattern as the correct matrix Z ...

Z^1				Z			
	(a)	(b)	(c)		(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	0	0
2	1	0	0	2	1	0	0
3	1/3	1/3	1/3	3	1	0	0
4	1/3	1/3	1/3	4	0	1	0
5	0	1	0	5	0	1	0
6	1/2	1/2	0	6	0	1	0
7	1/3	1/3	1/3	7	0	1	0
8	0	1/2	1/2	8	0	1	0
9	0	1/2	1/2	9	0	1	0
10	1/2	0	1/2	10	0	0	1
Totals	20/6	23/6	17/6	Totals	3	6	1

Recomputing $\theta^{(t)}$ from $Z^{(t)}$

STOP: Now that we have our estimate of Z , how can we improve our guess for θ ?

Z^1				Z			
	(a)	(b)	(c)		(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	0	0
2	1	0	0	2	1	0	0
3	1/3	1/3	1/3	3	1	0	0
4	1/3	1/3	1/3	4	0	1	0
5	0	1	0	5	0	1	0
6	1/2	1/2	0	6	0	1	0
7	1/3	1/3	1/3	7	0	1	0
8	0	1/2	1/2	8	0	1	0
9	0	1/2	1/2	9	0	1	0
10	1/2	0	1/2	10	0	0	1
Totals	20/6	23/6	17/6	Totals	3	6	1

Recomputing $\theta^{(t)}$ from $Z^{(t)}$

STOP: Now that we have our estimate of Z , how can we improve our guess for θ ?

Answer: Normalize the totals in each column by the number of transcripts.

	Z^1			Z		
	(a)	(b)	(c)	(a)	(b)	(c)
1	1/3	1/3	1/3	1	0	0
2	1	0	0	1	0	0
3	1/3	1/3	1/3	1	0	0
4	1/3	1/3	1/3	0	1	0
5	0	1	0	0	1	0
6	1/2	1/2	0	0	1	0
7	1/3	1/3	1/3	0	1	0
8	0	1/2	1/2	0	1	0
9	0	1/2	1/2	0	1	0
10	1/2	0	1/2	0	0	1

Totals 20/6 23/6 17/6 Totals 3 6 1

$$\theta^{(1)} = (.333, .383, .283)$$

Working with a Simpler Example

So if we have a guess for θ , we can make a guess for Z .

	Z^1			Y		
	(a)	(b)	(c)	(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	1
2	1	0	0	1	0	0
3	1/3	1/3	1/3	1	1	1
4	1/3	1/3	1/3	1	1	1
5	0	1	0	0	1	0
6	1/2	1/2	0	1	1	0
7	1/3	1/3	1/3	1	1	1
8	0	1/2	1/2	0	1	1
9	0	1/2	1/2	0	1	1
10	1/2	0	1/2	1	0	1

Totals 20/6 23/6 17/6

$$\theta^{(1)} = (.333, .383, .283)$$

Working with a Simpler Example

So if we have a guess for θ , we can make a guess for Z .

And if we have a guess for Z , we can make a guess for θ .

Z^1				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	1	1
2	1	0	0	2	1	0	0
3	1/3	1/3	1/3	3	1	1	1
4	1/3	1/3	1/3	4	1	1	1
5	0	1	0	5	0	1	0
6	1/2	1/2	0	6	1	1	0
7	1/3	1/3	1/3	7	1	1	1
8	0	1/2	1/2	8	0	1	1
9	0	1/2	1/2	9	0	1	1
10	1/2	0	1/2	10	1	0	1

Totals 20/6 23/6 17/6

$$\theta^{(1)} = (.333, .383, .283)$$

Working with a Simpler Example

So if we have a guess for θ , we can make a guess for Z .

And if we have a guess for Z , we can make a guess for θ .

STOP: What does this remind you of?

	Z^1			Y		
	(a)	(b)	(c)	(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	1
2	1	0	0	1	0	0
3	1/3	1/3	1/3	1	1	1
4	1/3	1/3	1/3	1	1	1
5	0	1	0	0	1	0
6	1/2	1/2	0	1	1	0
7	1/3	1/3	1/3	1	1	1
8	0	1/2	1/2	0	1	1
9	0	1/2	1/2	0	1	1
10	1/2	0	1/2	1	0	1

Totals 20/6 23/6 17/6

$$\theta^{(1)} = (.333, .383, .283)$$

Working with a Simpler Example

So if we have a guess for θ , we can make a guess for Z .

And if we have a guess for Z , we can make a guess for θ .

Answer: Expectation maximization!

Z^1				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1	1/3	1/3	1/3	1	1	1	1
2	1	0	0	2	1	0	0
3	1/3	1/3	1/3	3	1	1	1
4	1/3	1/3	1/3	4	1	1	1
5	0	1	0	5	0	1	0
6	1/2	1/2	0	6	1	1	0
7	1/3	1/3	1/3	7	1	1	1
8	0	1/2	1/2	8	0	1	1
9	0	1/2	1/2	9	0	1	1
10	1/2	0	1/2	10	1	0	1

Totals 20/6 23/6 17/6

$\theta^{(1)} = (.333, .383, .283)$

Carrying out a Few More Steps

E-step: compute $Z^{(t)}$
from $\theta^{(t-1)}$ using

$$Z^{(t)}_{i,k} = Y_{i,k} * \theta^{(t-1)}_k / s_i$$

$$s_i = \sum_{\text{transcripts } j} Y_{i,j} * \theta^{(t-1)}_j$$

Z^2

	(a)	(b)	(c)
1	.333	.383	.283
2	1	0	0
3	.333	.383	.283
4	.333	.383	.283
5	0	1	0
6	.465	.535	0
7	.333	.383	.283
8	0	.575	.425
9	0	.575	.425
10	.541	0	.459

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

$$\theta^{(1)} = (.333, .383, .283)$$

Carrying out a Few More Steps

M-step: sum each column of $Z^{(t)}$ and normalize by the number of rows (reads) to produce $\theta^{(t)}$.

Z^2

	(a)	(b)	(c)
1	.333	.383	.283
2	1	0	0
3	.333	.383	.283
4	.333	.383	.283
5	0	1	0
6	.465	.535	0
7	.333	.383	.283
8	0	.575	.425
9	0	.575	.425
10	.541	0	.459

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.338 4.217 2.441

$$\theta^{(2)} = (.334, .422, .244)$$

Carrying out a Few More Steps

M-step: sum each column of $Z^{(t)}$ and normalize by the number of rows (reads) to produce $\theta^{(t)}$.

Exercise: Apply one more E-step and one more M-step to find Z^3 and θ^3 .

Z^3				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1				1	1	1	1
2				2	1	0	0
3				3	1	1	1
4				4	1	1	1
5				5	0	1	0
6				6	1	1	0
7				7	1	1	1
8				8	0	1	1
9				9	0	1	1
10				10	1	0	1

$$\theta^{(2)} = (.334, .422, .244)$$

Carrying out a Few More Steps

M-step: sum each column of $Z^{(t)}$ and normalize by the number of rows (reads) to produce $\theta^{(t)}$.

Exercise: Apply one more E-step and one more M-step to find Z^3 and θ^3 .

Z^3				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1	.334	.422	.244	1	1	1	1
2	1	0	0	2	1	0	0
3	.334	.422	.244	3	1	1	1
4	.334	.422	.244	4	1	1	1
5	0	1	0	5	0	1	0
6	.442	.558	0	6	1	1	0
7	.334	.422	.244	7	1	1	1
8	0	.634	.366	8	0	1	1
9	0	.634	.366	9	0	1	1
10	.578	0	.422	10	1	0	1

$$\theta^{(2)} = (.334, .422, .244)$$

Carrying out a Few More Steps

M-step: sum each column of $Z^{(t)}$ and normalize by the number of rows (reads) to produce $\theta^{(t)}$.

Exercise: Apply one more E-step and one more M-step to find Z^3 and θ^3 .

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Carrying out a Few More Steps

M-step: sum each column of $Z^{(t)}$ and normalize by the number of rows (reads) to produce $\theta^{(t)}$.

Exercise: Apply one more E-step and one more M-step to find Z^3 and θ^3 .

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Convergence of the Algorithm

STOP: When will we stop this algorithm?

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Convergence of the Algorithm

STOP: When will we stop this algorithm?

Answer: When the difference between $\theta^{(t)}$ and $\theta^{(t-1)}$ sinks beneath some threshold ϵ .

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Convergence of the Algorithm

STOP: Any guesses on what you think θ might converge to in this case?

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Convergence of the Algorithm

STOP: Any guesses on what you think θ might converge to in this case?

Answer (thanks Eric Xu): $\theta = (.4, .6, 0)$.

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Running EM Multiple Times

STOP: EM is run multiple times on different inputs. What are our inputs, and how would we change them?

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Y

	(a)	(b)	(c)
1	1	1	1
2	1	0	0
3	1	1	1
4	1	1	1
5	0	1	0
6	1	1	0
7	1	1	1
8	0	1	1
9	0	1	1
10	1	0	1

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Running EM Multiple Times

STOP: EM is run multiple times on different inputs. What are our inputs, and how would we change them?

Answer: This example used $\theta^{(0)} = (1/3, 1/3, 1/3)$, but we could run multiple times with different possible $\theta^{(0)}$.

Z^3				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1	.334	.422	.244	1	1	1	1
2	1	0	0	2	1	0	0
3	.334	.422	.244	3	1	1	1
4	.334	.422	.244	4	1	1	1
5	0	1	0	5	0	1	0
6	.442	.558	0	6	1	1	0
7	.334	.422	.244	7	1	1	1
8	0	.634	.366	8	0	1	1
9	0	.634	.366	9	0	1	1
10	.578	0	.422	10	1	0	1

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Running EM Multiple Times

But how would we choose the “best” possible final θ and Z over all these runs?
What are we optimizing?!

Z^3				Y			
	(a)	(b)	(c)		(a)	(b)	(c)
1	.334	.422	.244	1	1	1	1
2	1	0	0	2	1	0	0
3	.334	.422	.244	3	1	1	1
4	.334	.422	.244	4	1	1	1
5	0	1	0	5	0	1	0
6	.442	.558	0	6	1	1	0
7	.334	.422	.244	7	1	1	1
8	0	.634	.366	8	0	1	1
9	0	.634	.366	9	0	1	1
10	.578	0	.422	10	1	0	1

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Expectation Maximization Has Same Structure in Different Contexts

In both problems, we want to find something hidden in the data that best “explains” the data.

Motif Finding

- **Given:** set of strings
- **Want:** profile matrix
- **Hidden:** starting position of motif in each string

RNA-Seq Quantification

- **Given:** RNA reads
- **Want:** abundance vector θ
- **Hidden:** matrix Z containing assignment of reads to transcripts

Expectation Maximization Has Same Structure in Different Contexts

In both problems, we want to find something hidden in the data that best “explains” the data.

Motif Finding

- **Given:** set of strings
- **Want:** profile matrix
- **Hidden:** starting position of motif in each string

Scoring motifs gives us a way of comparing different results.

RNA-Seq Quantification

- **Given:** RNA reads
- **Want:** abundance vector θ
- **Hidden:** matrix Z containing assignment of reads to transcripts

Expectation Maximization Has Same Structure in Different Contexts

In both problems, we want to find something hidden in the data that best “explains” the data.

Motif Finding

- **Given:** set of strings
- **Want:** profile matrix
- **Hidden:** starting position of motif in each string

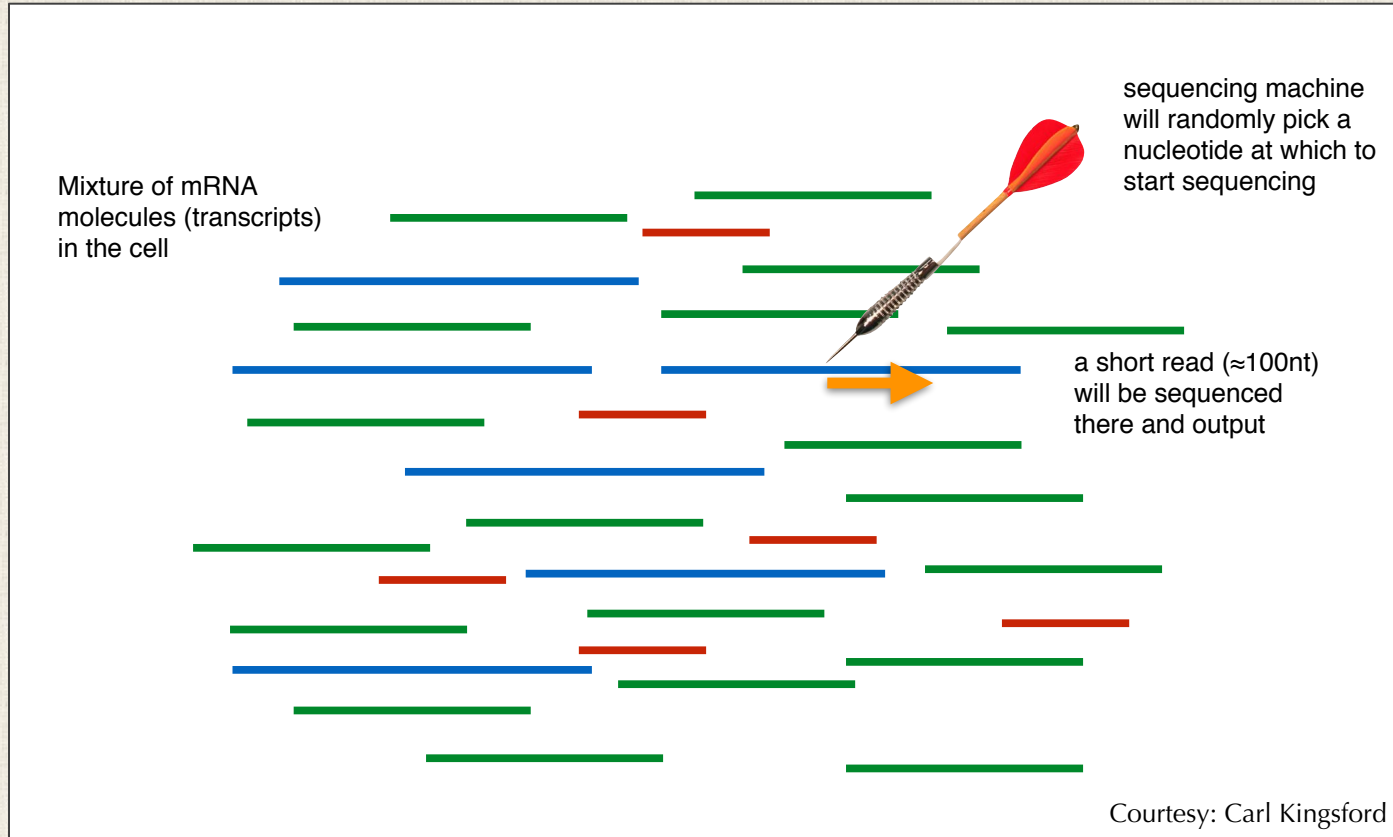
Scoring motifs gives us a way of comparing different results.

RNA-Seq Quantification

- **Given:** RNA reads
- **Want:** abundance vector θ
- **Hidden:** matrix Z containing assignment of reads to transcripts

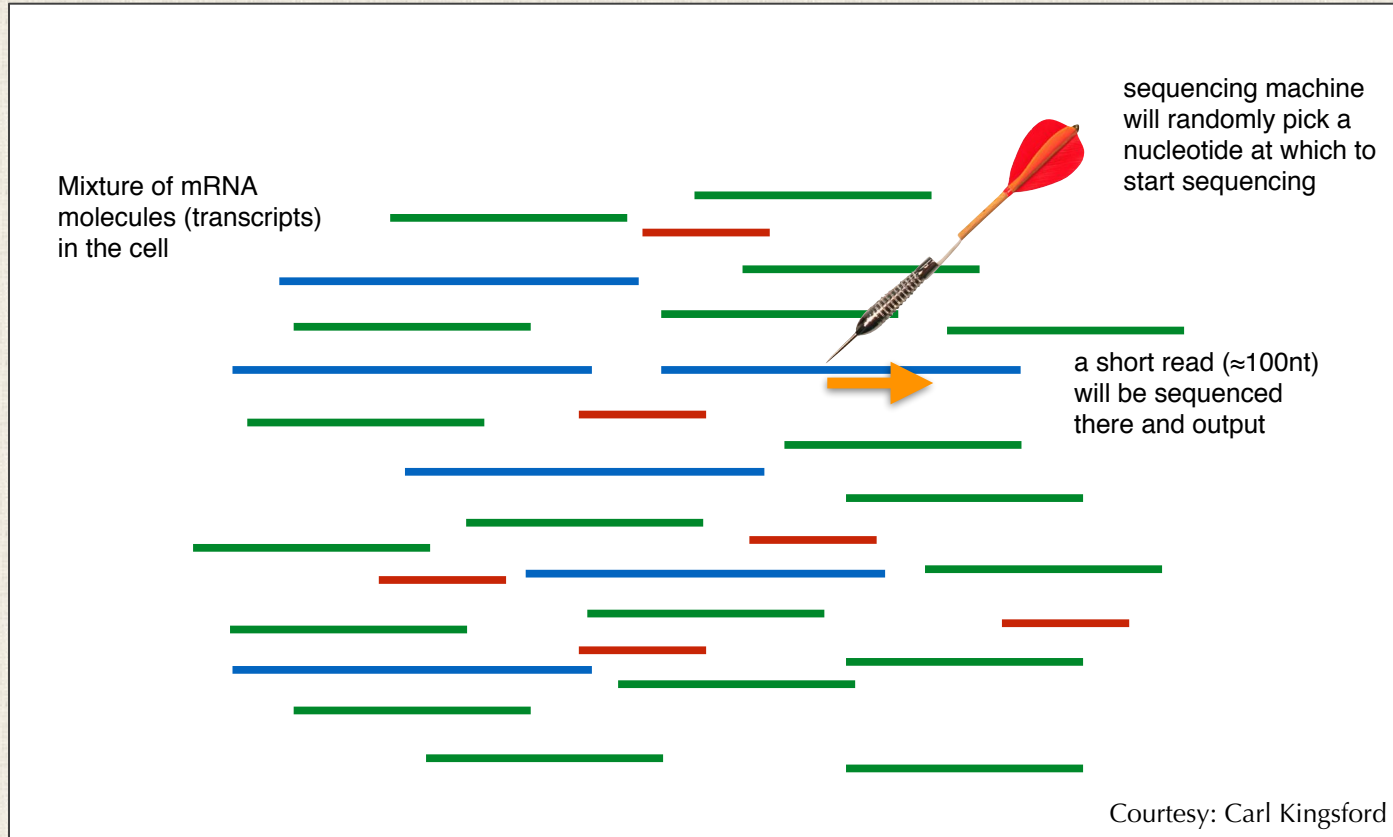
How do we “score” different abundance vectors?

A Probabilistic Model for RNA-Seq



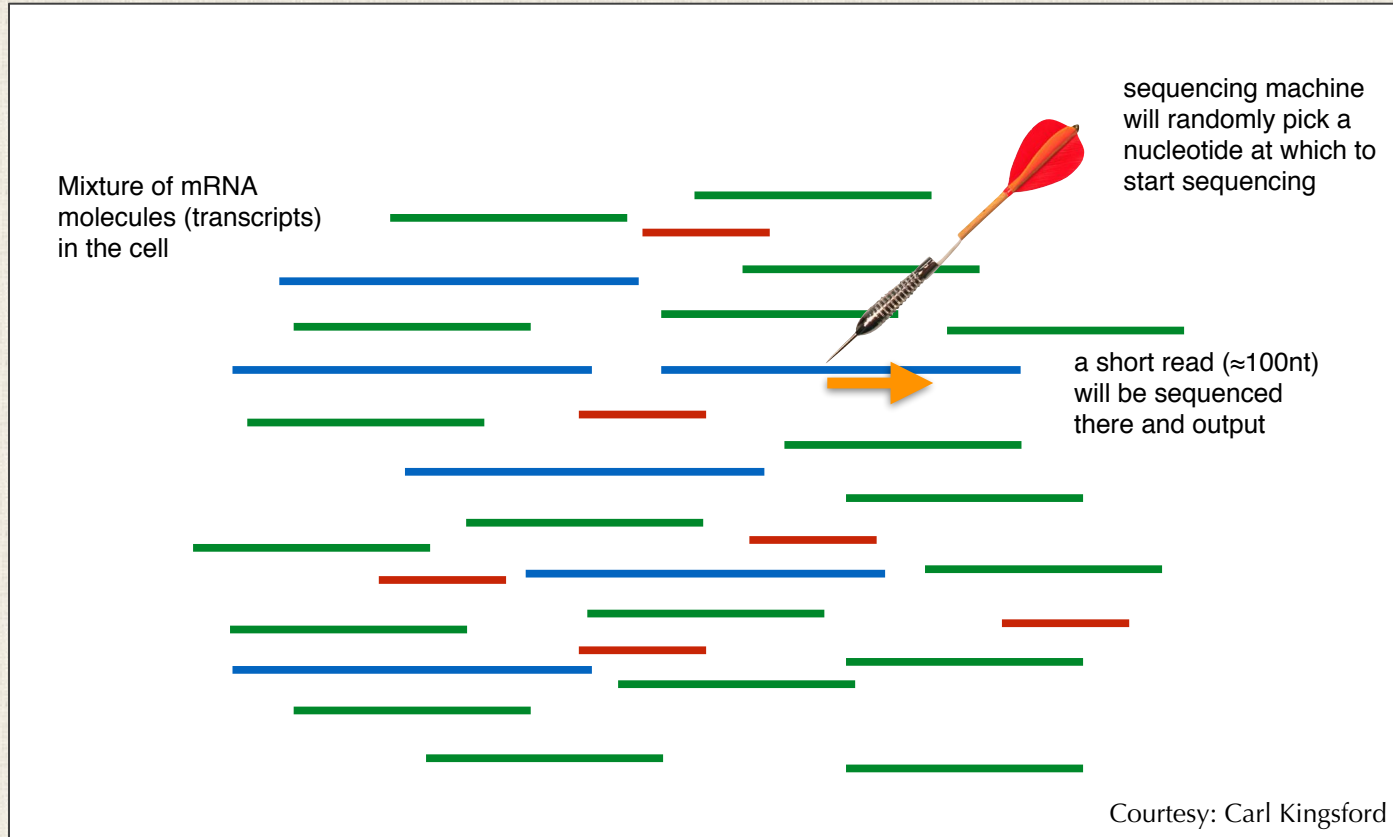
Given a fixed abundance vector θ , $\Pr(\mathbf{x}|\theta)$ is the probability that this model would have generated the RNA-sequencing reads \mathbf{x} that we observe.

A Probabilistic Model for RNA-Seq



Key Point: If θ were heavily weighted toward red, then $\Pr(\mathbf{x}|\theta)$ would be much lower than if θ were heavily weighted toward blue.

A Probabilistic Model for RNA-Seq



Determining $\Pr(\mathbf{x}|\theta)$ is beyond our work here, but it allows us to compare abundance vectors resulting from running EM on different initial vectors θ .

Finally, A Point about Timing

Cufflinks uses this EM approach for quantification prediction, but an earlier paper described the method and perhaps was too early to get the credit that it deserves.

www.ncbi.nlm.nih.gov › [pmc](#) › [articles](#) › [PMC1475746](#) ▼

[An expectation-maximization algorithm for probabilistic ...](#)

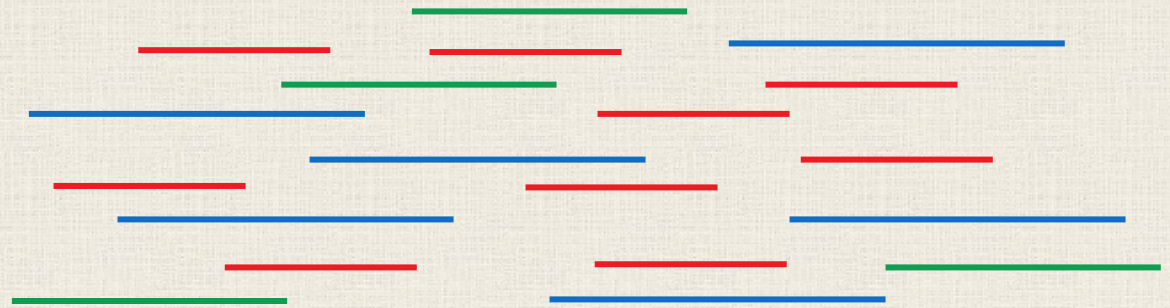
by Y Xing - 2006 - [Cited by 107](#) - [Related articles](#)

Jun 6, 2006 - and **Christopher Lee** ... In fact, over 80% of alternative splicing events in the human **transcriptome** are detected ... We describe an **EM** algorithm to estimate the probability for each ... Probabilistic formulation and **EM** algorithm ...

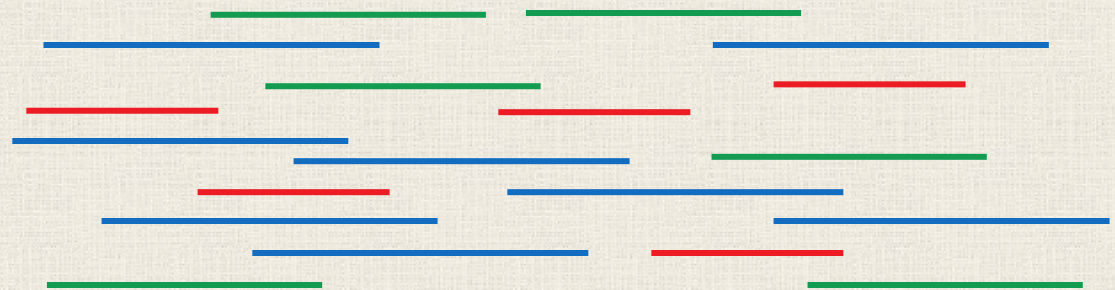
PART 4: COMPARING EXPRESSION ACROSS SAMPLES

How do we compare RNA-seq samples?

Sample 1



Sample 2



Say that we want to compare the gene expression in two samples. How can we infer this difference from the *fragments* resulting from these samples?

How do we compare RNA-seq samples?

Sample 1



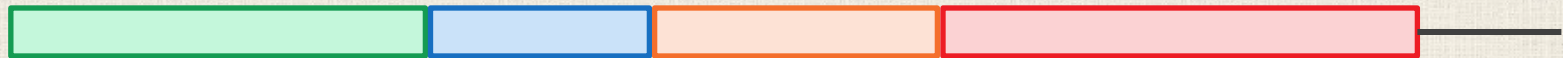
Sample 2



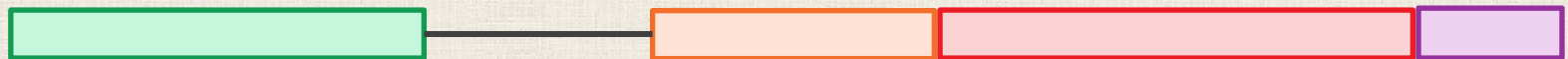
Key point: We need to use what we have already learned about inferring information from a sample's fragments in order to differentiate the samples.

Comparing two samples gene by gene

Let's focus on a single gene, which may have multiple isoforms with exons of differing lengths.



Isoform (a)



Isoform (b)



Isoform (c)

The Exon Union Model is a Simple Way of Quantifying Expression of a Gene

Exon union model: Chain all exons of a gene together, even if no isoform contains them all.



Chained exons

The Exon Union Model is a Simple Way of Quantifying Expression of a Gene

Exon union model: Chain all exons of a gene together, even if no isoform contains them all.

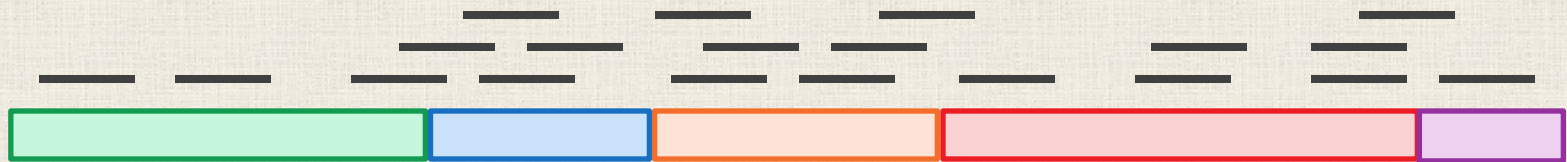


Chained exons

We can set the **expression** of a gene in a sample equal to the number of reads from the sample mapping to the gene.

Let's Consider an Example

Exercise: What is the expression of a gene in a sample where fragments map as below?



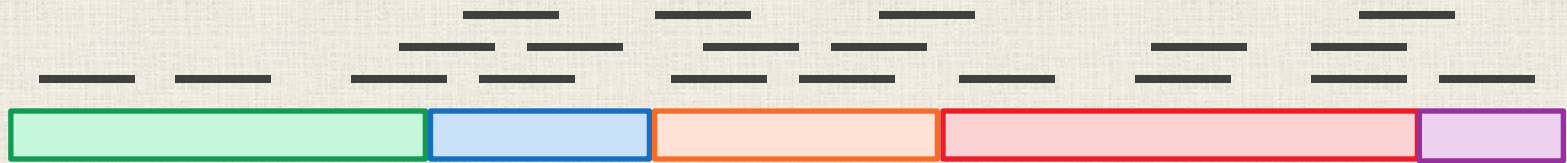
Chained exons

We can set the **expression** of a gene in a sample equal to the number of reads from the sample mapping to the gene.

Let's Consider an Example

Answer: 20 reads mapped.

STOP: Why is this metric flawed?

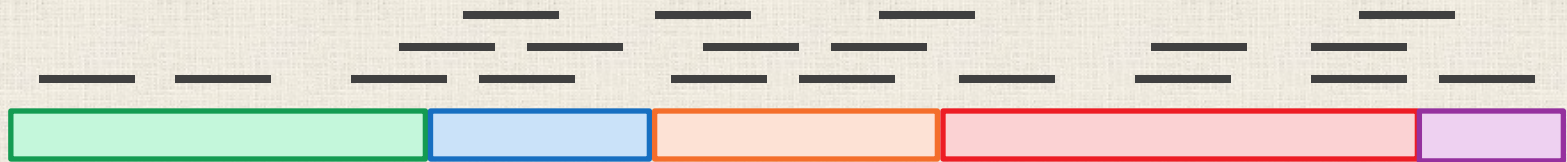


Chained exons

We can set the **expression** of a gene in a sample equal to the number of reads from the sample mapping to the gene.

Let's Consider an Example

Key point: long genes will receive more reads, so we should normalize expression by *gene length*.

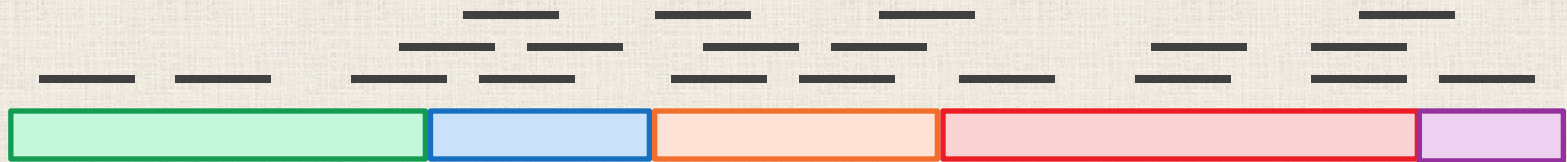


Chained exons

We can set the **expression** of a gene in a sample equal to the number of reads from the sample mapping to the gene, ***per kilobase***.

Let's Consider an Example

Exercise: What is the expression of a gene of length 800 bp in a sample where fragments map as below?

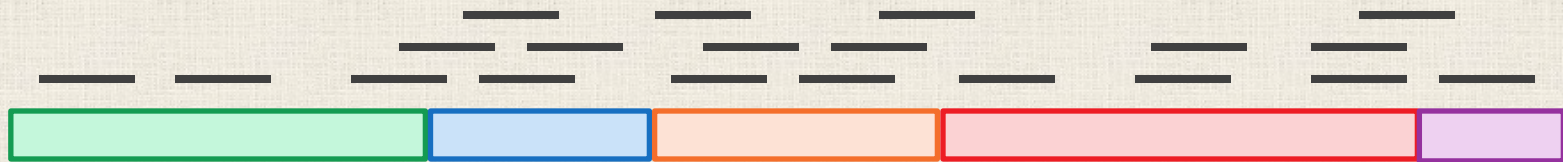


Chained exons

We can set the **expression** of a gene in a sample equal to the number of reads from the sample mapping to the gene, *per kilobase*.

Let's Consider an Example

Answer: (20 reads mapped)/(0.8 kilobases) = 25 reads per kilobase.

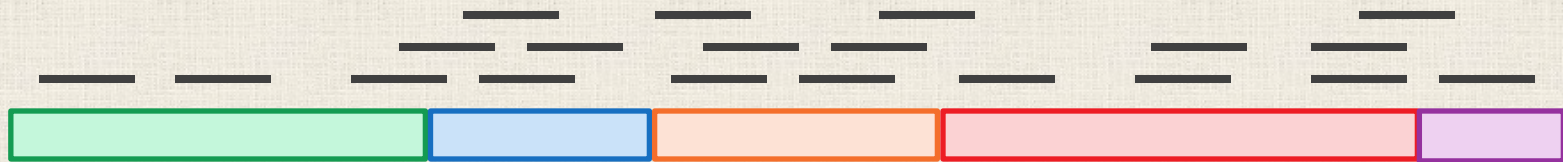


Chained exons

We set the *expression* of a gene in a sample equal to the number of reads from the sample mapping to the gene, divided by the total length of all exons.

Let's Consider an Example

Answer: (20 reads mapped)/(0.8 kilobases) = 25 reads per kilobase.



Chained exons

STOP: How could we compare the expression of a gene *across* two different samples?

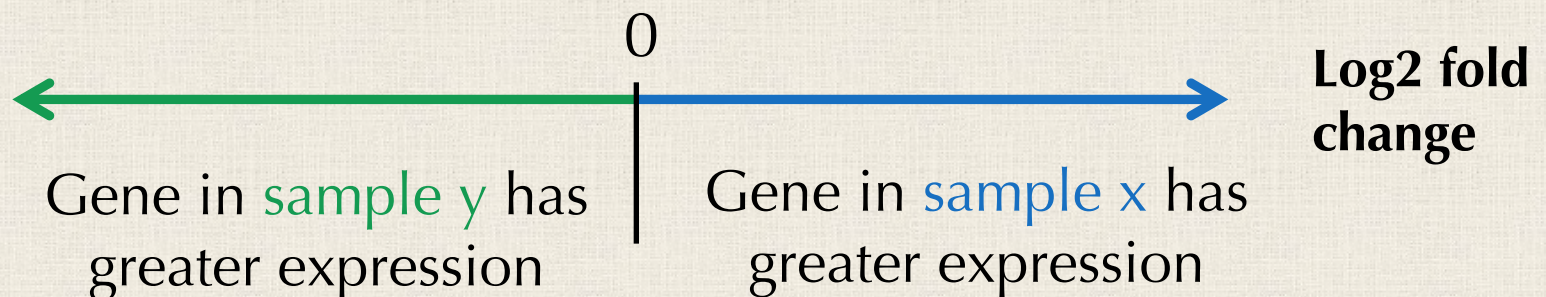
Log2 Fold Change Compares Expression of a Gene in Two Samples

To compare the expression of a gene in two samples, we use **log2 fold change**: the base-2 logarithm of the ratio of the expression values.

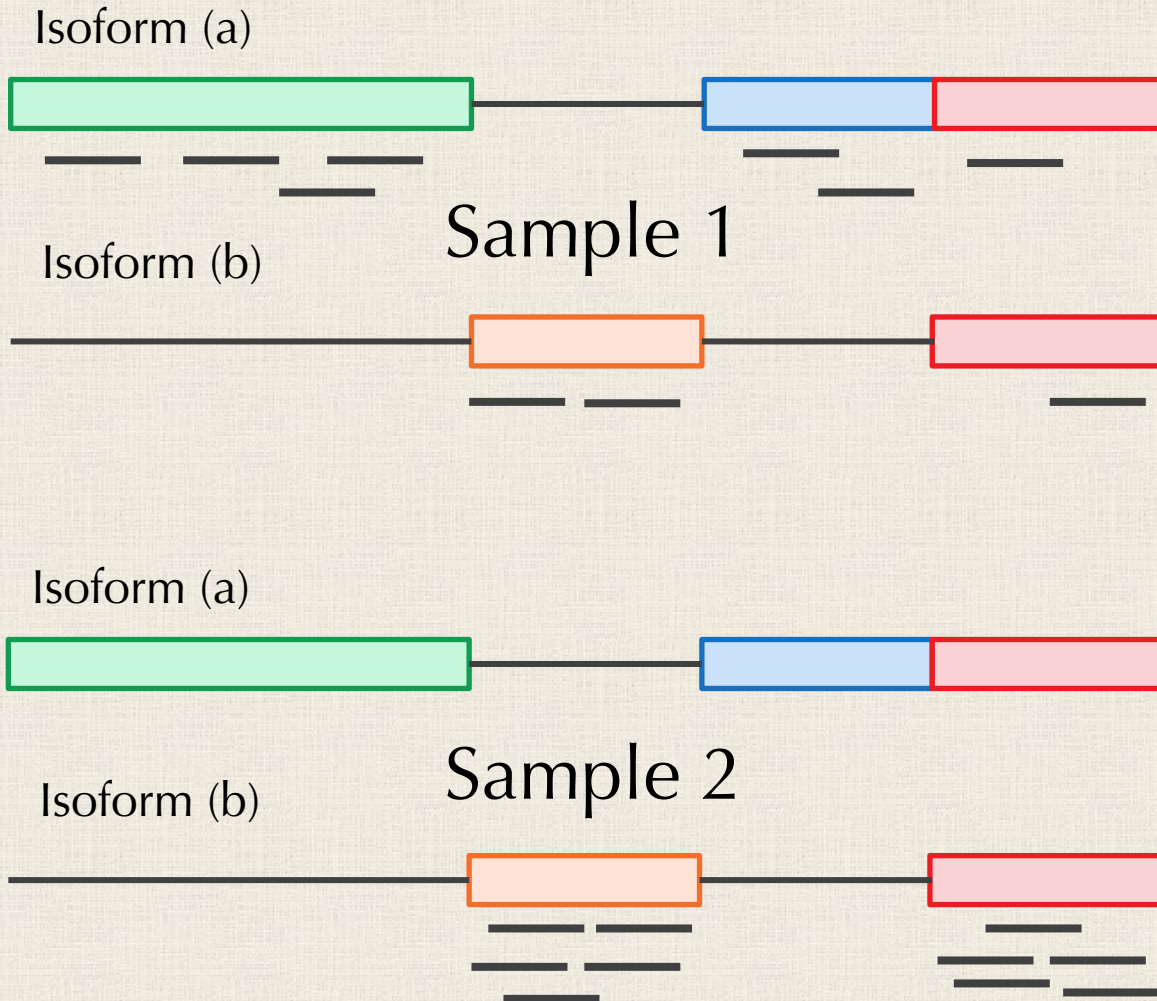
Log₂ Fold Change Compares Expression of a Gene in Two Samples

To compare the expression of a gene in two samples, we use **log₂ fold change**: the base-2 logarithm of the ratio of the expression values.

If the expression x of a gene in sample 1 is greater than the expression y of this gene in sample 2, then $\log_2(x / y)$ will be > 0 .

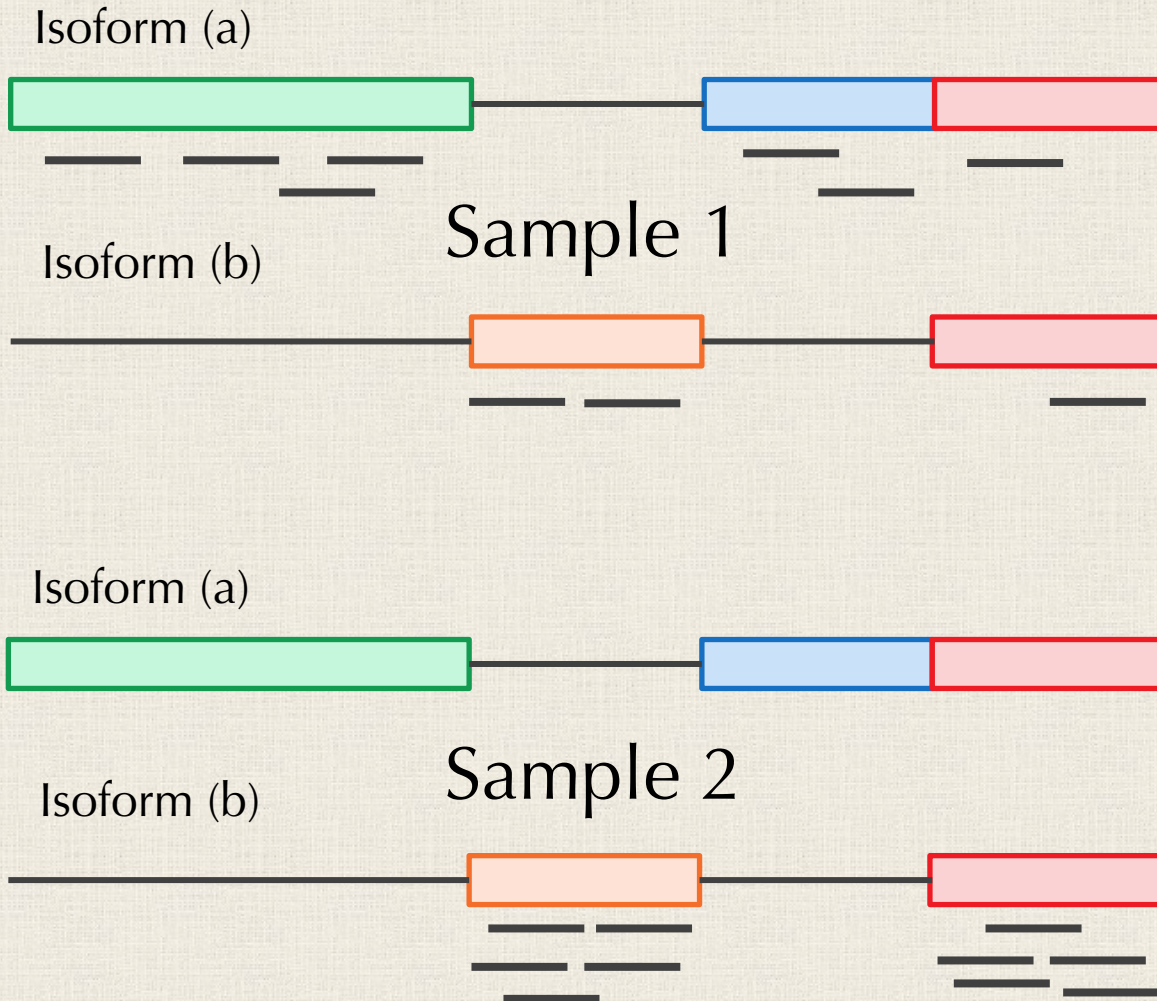


Problems with the current model



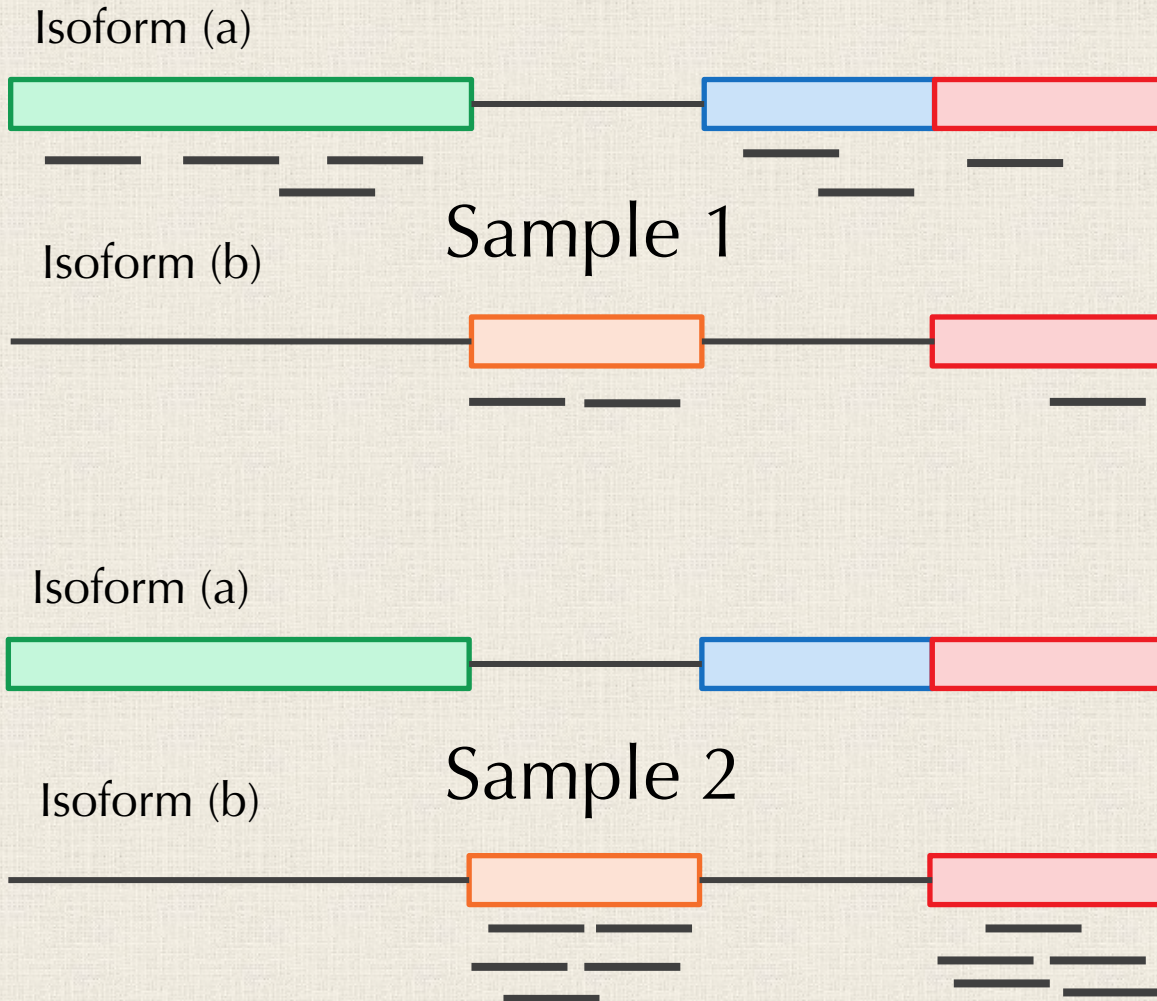
STOP: What is the \log_2 fold change of this gene in the two samples under the exon union model?

Problems with the current model



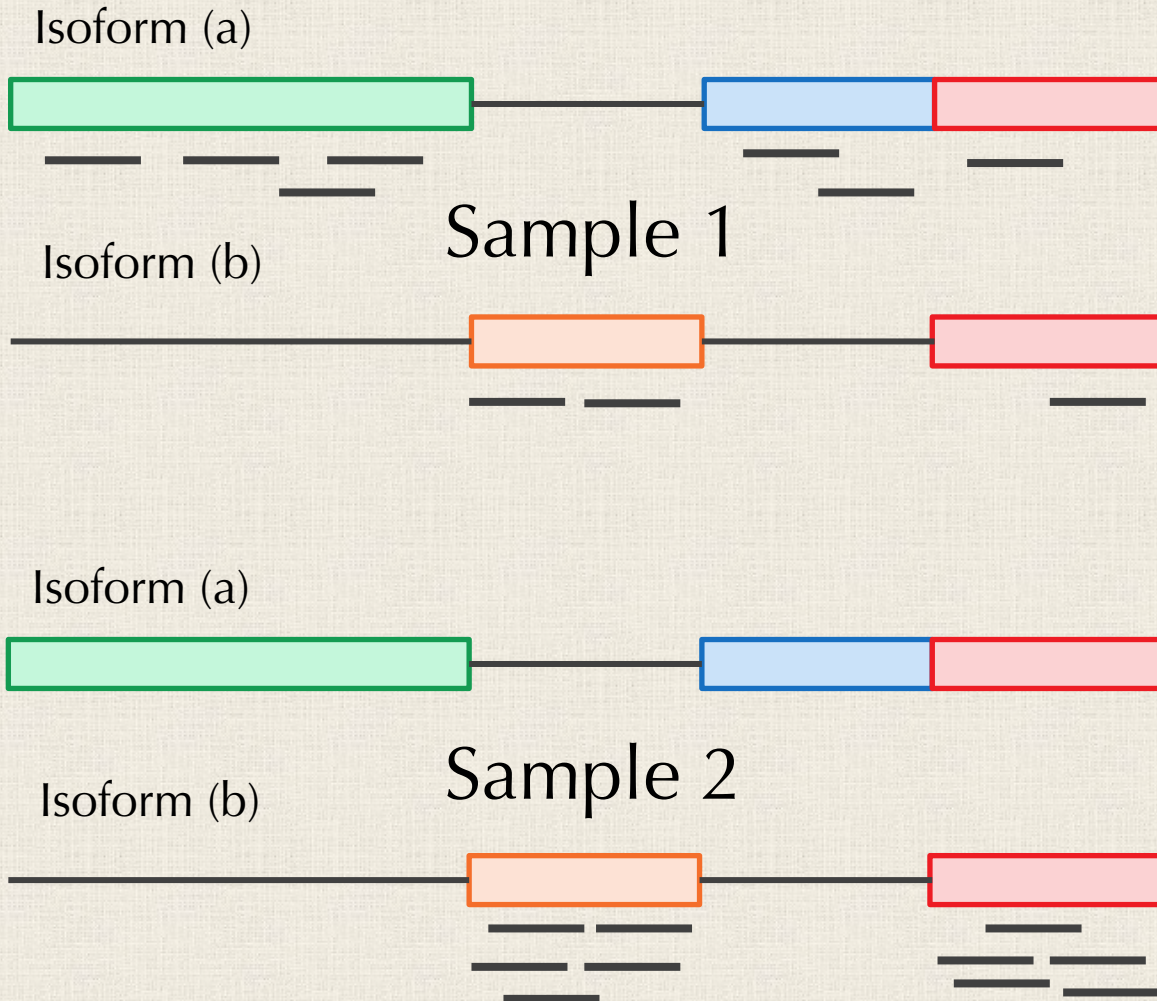
Answer: Zero, because they have the same expression under the exon union model.

Problems with the current model



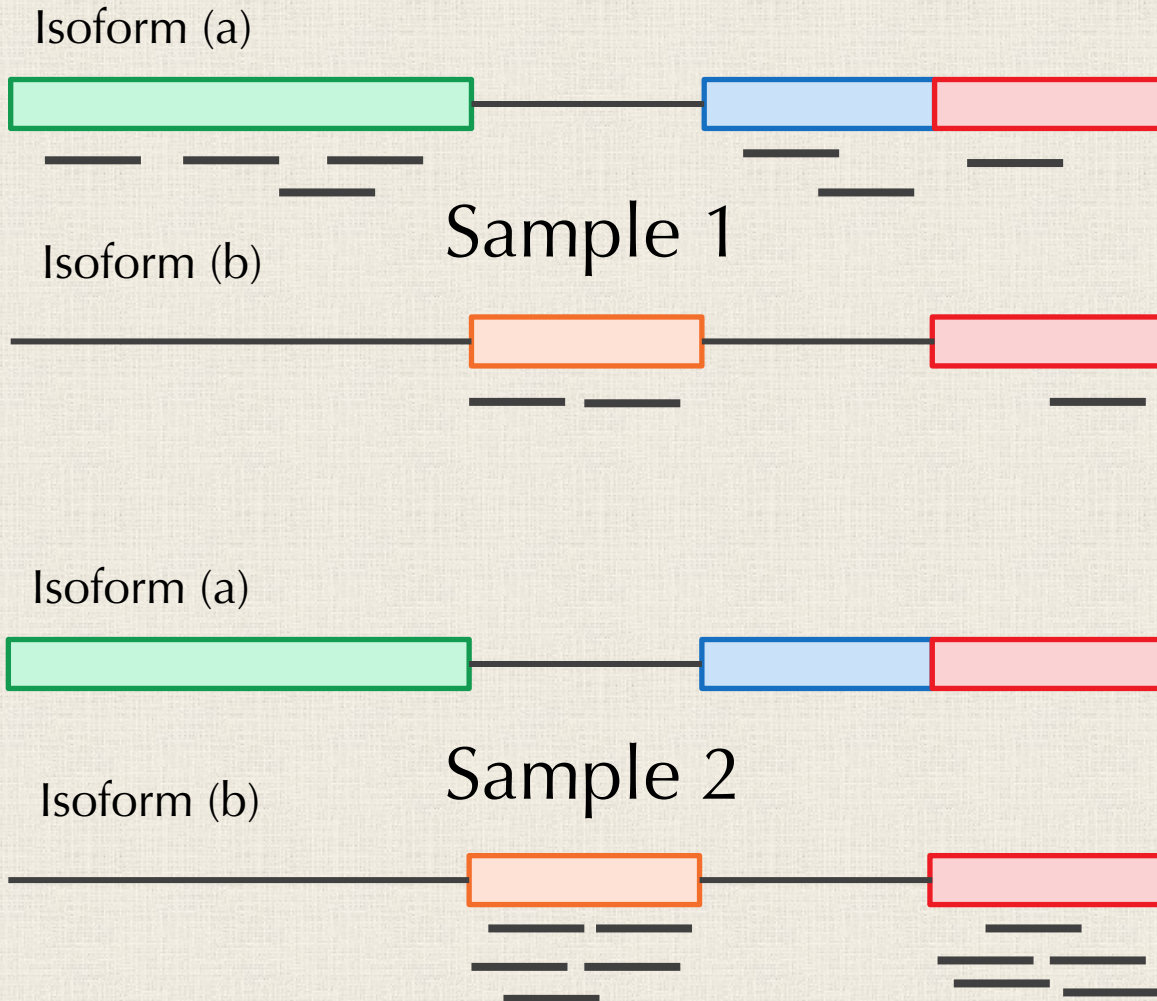
STOP: Why is this an issue? What biological fact have we missed in these samples?

Problems with the current model



Answer:
Reads map only to one isoform in sample 1, and this isoform's expression is far greater than in sample 1.

Problems with the current model



Key point:
We need *transcript* level comparison of expression.

Fortunately, Cufflinks gives us abundance estimates for each *transcript*

Recall that the EM algorithm gives us θ , which estimates the fraction of reads that map to each individual transcript.

If EM estimates that 33.6% of 1000 reads mapping to a gene come from one transcript, we get a simple expression value of 336 fragments mapped.

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

$$\theta^{(3)} = (.336, .451, .213)$$

Fortunately, Cufflinks gives us abundance estimates for each *transcript*

STOP: How can we improve this metric for expression?

If EM estimates that 33.6% of 1000 reads mapping to a gene come from one transcript, we get a simple expression value of 336 fragments mapped.

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

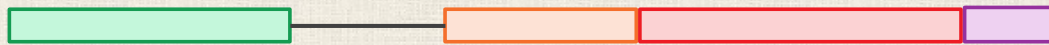
$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Answer: Take number of reads mapped to a transcript *per kilobase* of the transcript.



Isoform (a)



Isoform (b)



Isoform (c)

Z^3

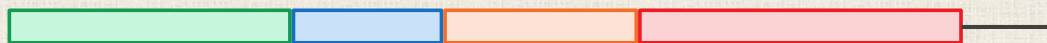
	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

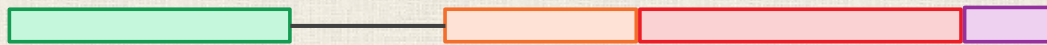
$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Exercise: Using $\theta^{(3)}$, what is each isoform's expression for 1000 reads, if (a), (b), (c) have respective lengths 1200 bp, 1000 bp, and 800 bp?



Isoform (a)



Isoform (b)



Isoform (c)

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Answer:

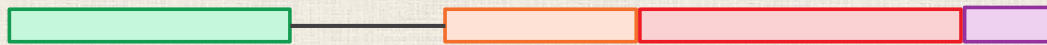
a) 336 reads / 1.2 kbp = 280 reads/kbp

b) 451 reads / 1 kbp = 451 reads/kbp

c) 213 reads / 0.8 kbp = 266.25 reads/kbp



Isoform (a)



Isoform (b)



Isoform (c)

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Answer:

a) 336 reads / 1.2 kbp = 280 reads/kbp

b) 451 reads / 1 kbp = 451 reads/kbp

c) 213 reads / 0.8 kbp = 266.25 reads/kbp

STOP: Say experiment 2 gives us these values. Are the values very different from experiment 1?

a) 29000 reads/kbp

b) 46000 reads/kbp

c) 26000 reads/kbp

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Answer:

a) 336 reads / 1.2 kbp = 280 reads/kbp

b) 451 reads / 1 kbp = 451 reads/kbp

c) 213 reads / 0.8 kbp = 266.25 reads/kbp

STOP: But what if I told you that experiment 2 generated 100x as many reads as experiment 1?

a) 29000 reads/kbp

b) 46000 reads/kbp

c) 26000 reads/kbp

Z^3

	(a)	(b)	(c)
1	.334	.422	.244
2	1	0	0
3	.334	.422	.244
4	.334	.422	.244
5	0	1	0
6	.442	.558	0
7	.334	.422	.244
8	0	.634	.366
9	0	.634	.366
10	.578	0	.422

Totals 3.356 4.514 2.130

$\theta^{(3)} = (.336, .451, .213)$

Improving our simple expression metric

Key point: Expression of every gene will be higher on average in experiments that generate more reads, so we need to *normalize* by the number of reads sequenced.

Improving our simple expression metric

Key point: Expression of every gene will be higher on average in experiments that generate more reads, so we need to *normalize* by the number of reads sequenced.

The expression value used by Cufflinks is **RPKM: reads mapped per kilobase of transcript, per million mapped reads.**

Comparing our improved expression metric for two samples

Experiment 1 (1M reads):

- a) 280 reads/kbp
- b) 451 reads/kbp
- c) 266.25 reads/kbp

Experiment 2 (100M reads):

- a) 29000 reads/kbp
- b) 46000 reads/kbp
- c) 26000 reads/kbp

The expression value used by Cufflinks is **RPKM: reads mapped per kilobase of transcript, per million mapped reads.**

Exercise: What is the RPKM of each isoform in each of the two experiments?

Comparing our improved expression metric for two samples

Experiment 1 (1M reads):

a) 280 reads/kbp

b) 451 reads/kbp

c) 266.25 reads/kbp

Answer:

a) $(280 \text{ reads/kbp}) / (1 \text{ M reads}) = 280 \text{ RPKM}$

b) $(451 \text{ reads/kbp}) / (1 \text{ M reads}) = 451 \text{ RPKM}$

c) $(266.25 \text{ reads/kbp}) / (1 \text{ M reads}) = 266.25 \text{ RPKM}$

Experiment 2 (100M reads):

a) 29000 reads/kbp

b) 46000 reads/kbp

c) 26000 reads/kbp

Answer:

a) $(29000 \text{ reads/kbp}) / (100 \text{ M reads}) = 290 \text{ RPKM}$

b) $(46000 \text{ reads/kbp}) / (100 \text{ M reads}) = 460 \text{ RPKM}$

c) $(26000 \text{ reads/kbp}) / (100 \text{ M reads}) = 260 \text{ RPKM}$

Comparing our improved expression metric for two samples

Now we can make a fair comparison of the resulting expression levels with log2foldchange!

STOP: Are these RPKMs similar? What's missing?

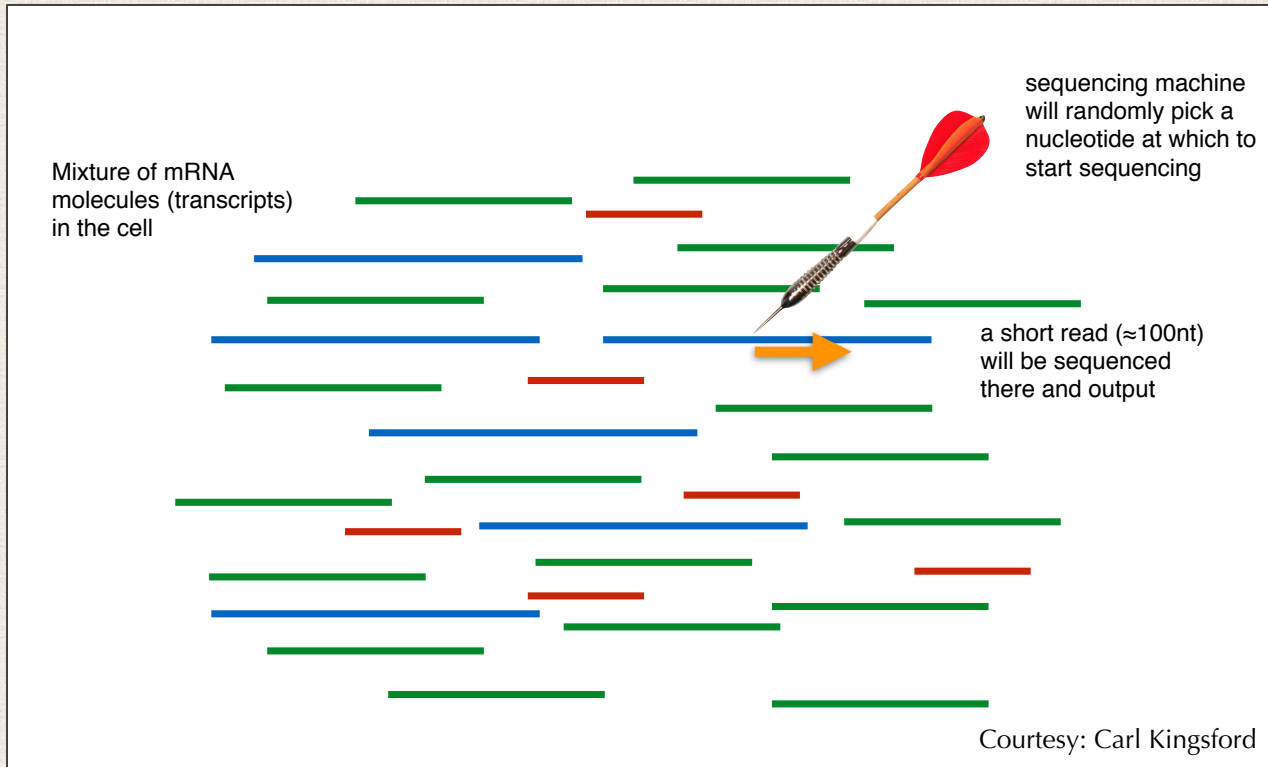
Answer:

- a) $(280 \text{ reads/kbp}) / (1 \text{ M reads}) = 280 \text{ RPKM}$
- b) $(451 \text{ reads/kbp}) / (1 \text{ M reads}) = 451 \text{ RPKM}$
- c) $(266.25 \text{ reads/kbp}) / (1 \text{ M reads}) = 266.25 \text{ RPKM}$

Answer:

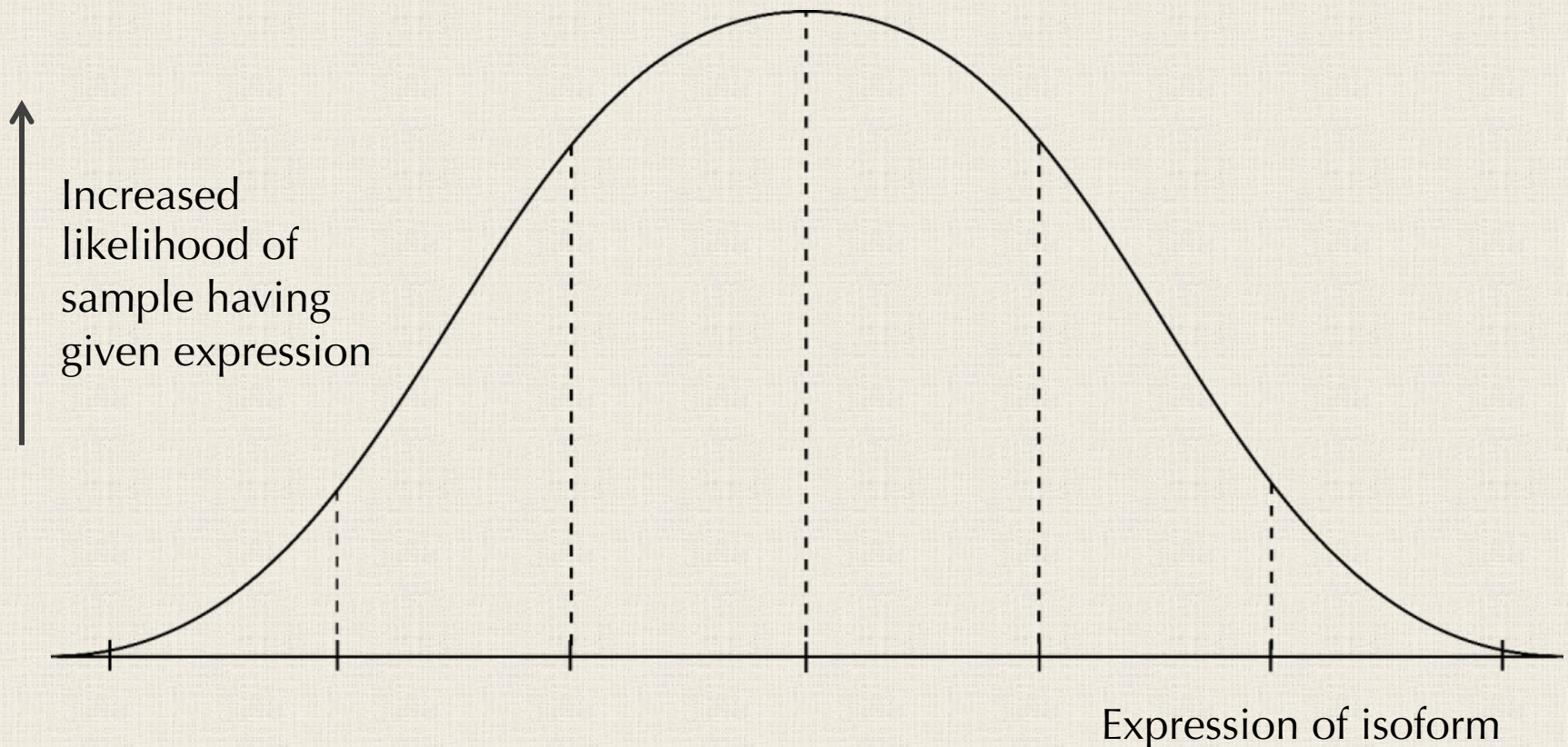
- a) $(29000 \text{ reads/kbp}) / (100 \text{ M reads}) = 290 \text{ RPKM}$
- b) $(46000 \text{ reads/kbp}) / (100 \text{ M reads}) = 460 \text{ RPKM}$
- c) $(26000 \text{ reads/kbp}) / (100 \text{ M reads}) = 260 \text{ RPKM}$

We need to incorporate *stochasticity* into differential expression



Key Point: We should not expect the same result from different RNA-seq runs on the same sample.

We use high-powered statistics to build a curve around expression estimate



We use high-powered statistics to build a curve around expression estimate

Instead of “Is the expression of two transcripts different?” we ask “How likely would *random chance* have caused the difference we see?”

We use high-powered statistics to build a curve around expression estimate

Instead of “Is the expression of two transcripts different?” we ask “How likely would *random chance* have caused the difference we see?”

STOP: What does this remind us of?

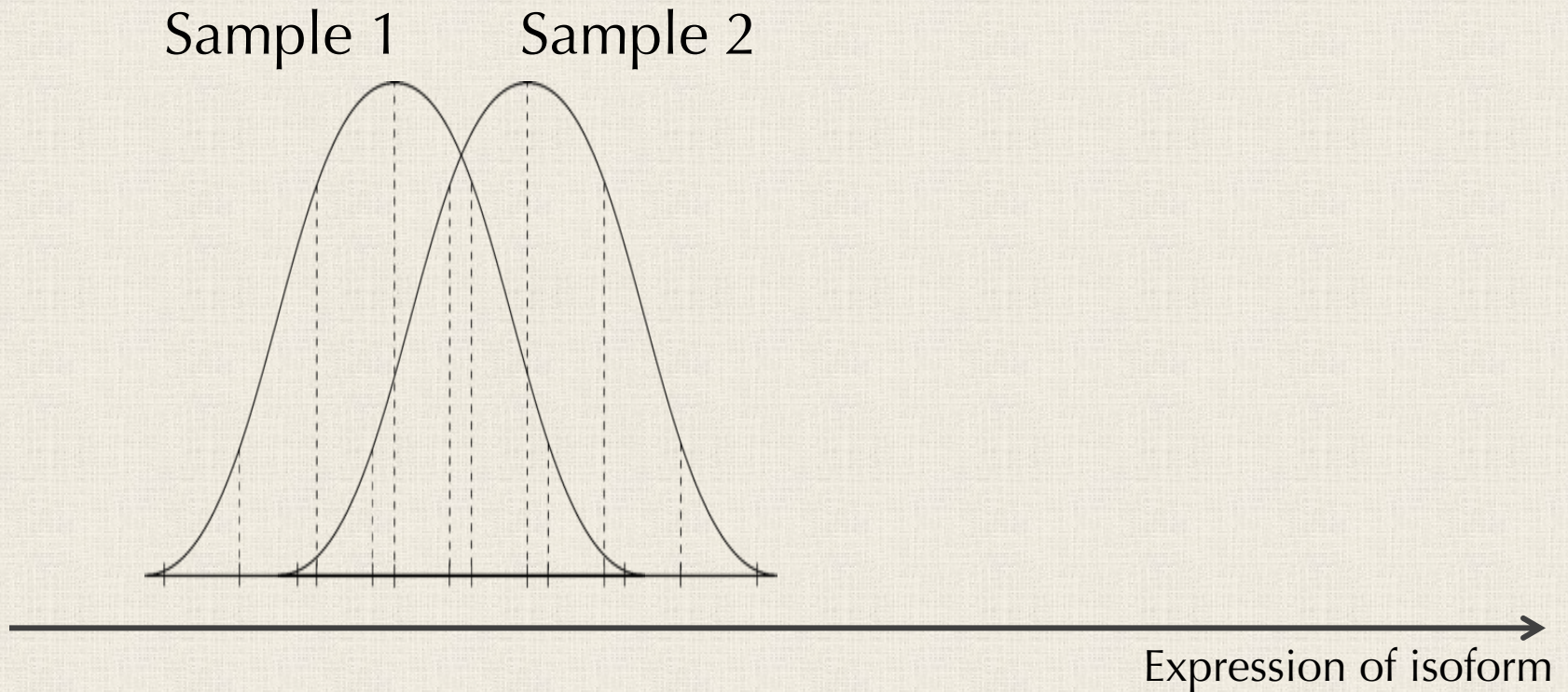
We use high-powered statistics to build a curve around expression estimate

Instead of “Is the expression of two transcripts different?” we ask “How likely would *random chance* have caused the difference we see?”

STOP: What does this remind us of?

Answer: BLAST!

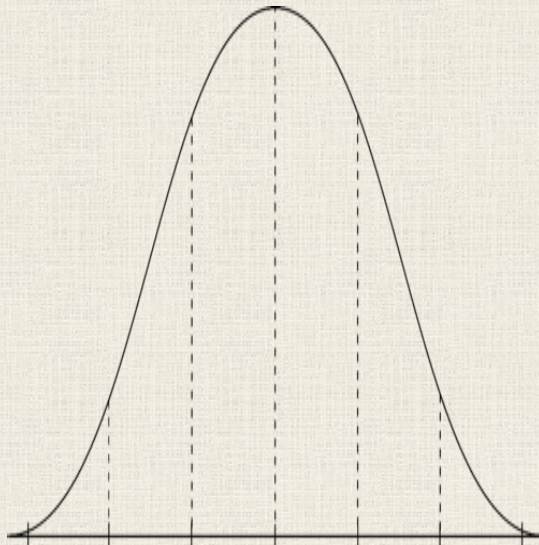
Our problem then reduces to curve comparison



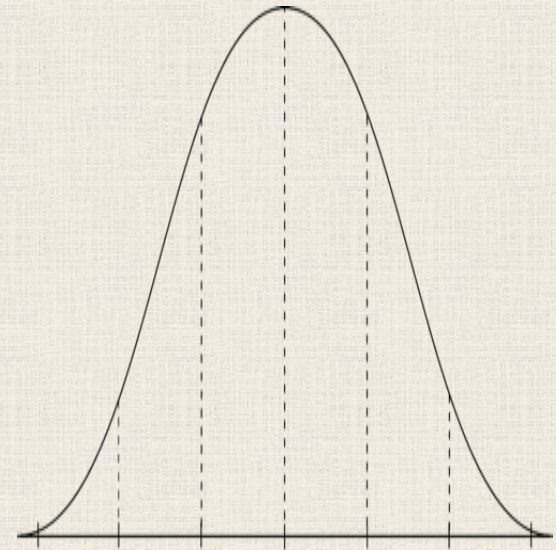
STOP: How sure are we that the isoform is differentially expressed in the two samples?

Our problem then reduces to curve comparison

Sample 1



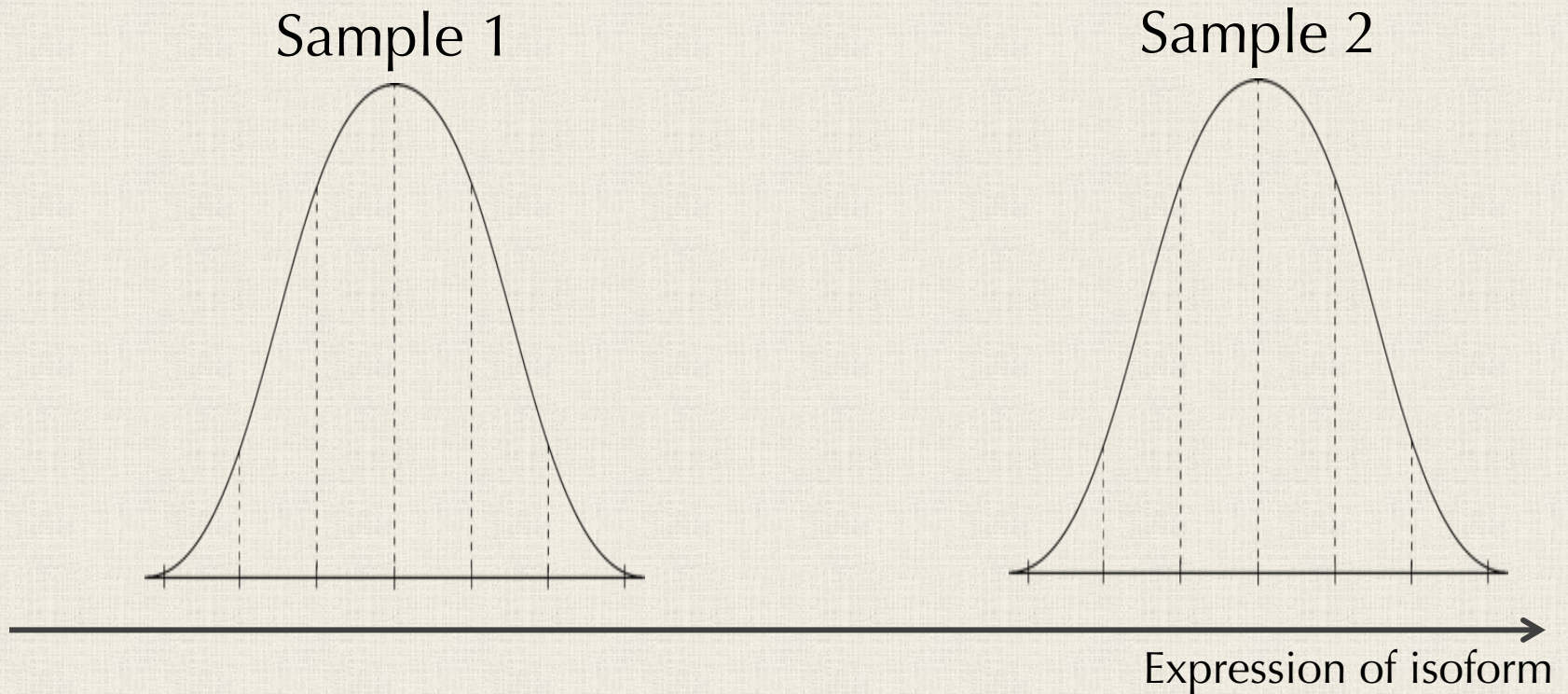
Sample 2



Expression of isoform

STOP: What about now?

Our problem then reduces to curve comparison



Note: This is a big simplification of a very complicated process.

This idea is the engine of “Cuffdiff”

p-value: The likelihood that we observe an outcome due to random chance.

This idea is the engine of “Cuffdiff”

p-value: The likelihood that we observe an outcome due to random chance.

When comparing two samples, we compute a p-value for every transcript in the samples, and focus on isoforms with low p-values.

Differential analysis of gene regulation at transcript resolution with RNA-seq

[C Trapnell, DG Hendrickson, M Sauvageau... - Nature ..., 2013 - nature.com](#)

... Here we introduce **Cuffdiff 2**, which addresses both problems simultaneously by modeling variability in the number of fragments generated by each transcript across replicates ... **Cuffdiff 2** identified genes that Differential analysis of gene regulation at transcript ...

☆  Cited by 2939 [Related articles](#) [All 26 versions](#)

Quick p-value quiz

STOP: Say that we have the following p-values for a differential expression analysis of 20,000 human genes. Which ones would you want to include?

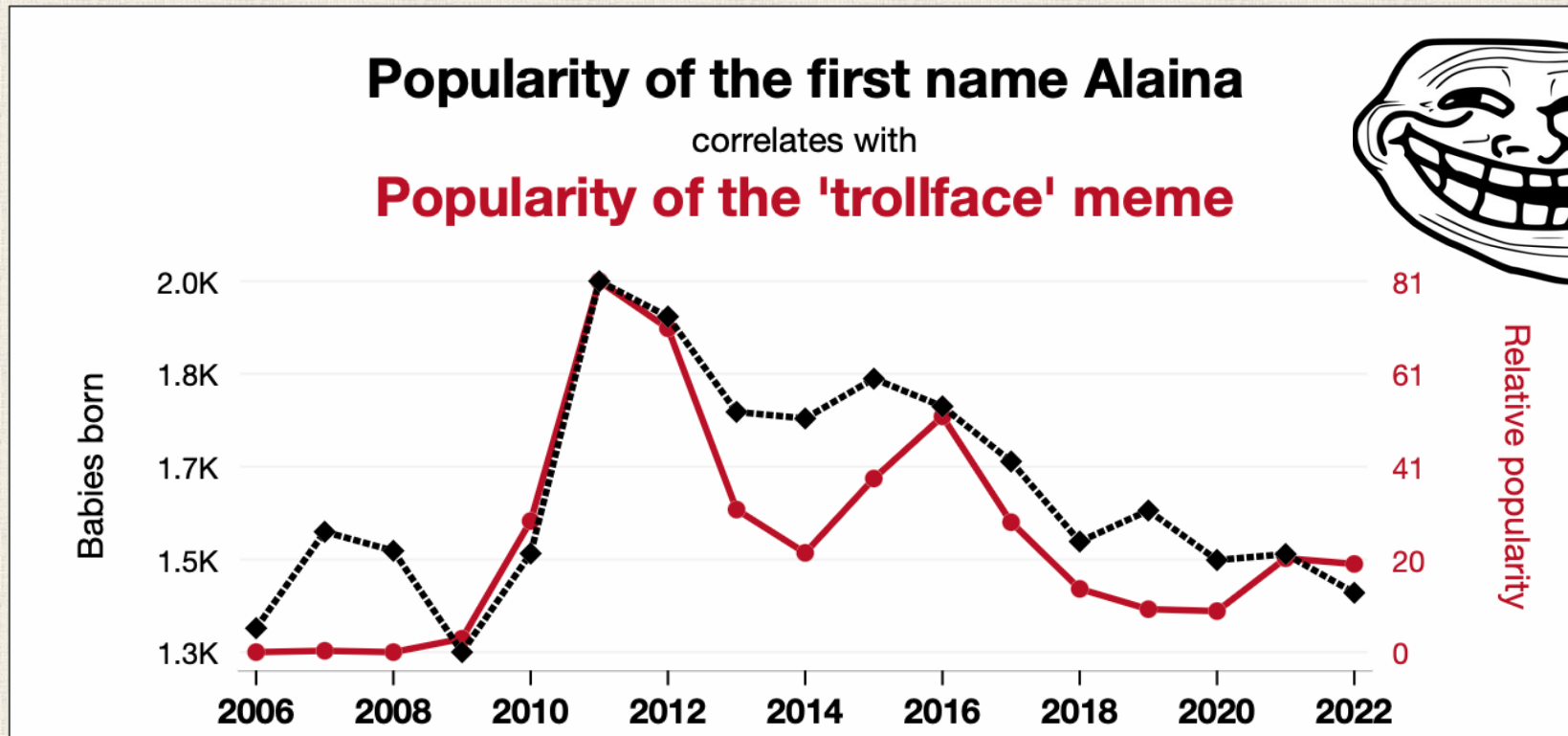
Gene	p-value
A	0.79
B	0.07
C	0.01
D	0.0031
E	0.00000079

Quick p-value quiz

STOP: Say you play a casino game 20,000 times with the following probability of success. Which games would you not expect to win?

Game	Probability
A	0.79
B	0.07
C	0.01
D	0.0031
E	0.00000079

Many trials means many chances for a low probability event to occur



Correcting our p-values with Bonferroni

Bonferroni Correction: When running n statistical tests simultaneously, we multiply all p-values by n .

Gene	p-value	Corrected Value
A	0.79	15800
B	0.07	1400
C	0.01	200
D	0.0031	62
E	0.00000079	0.0158

Correcting our p-values with Bonferroni

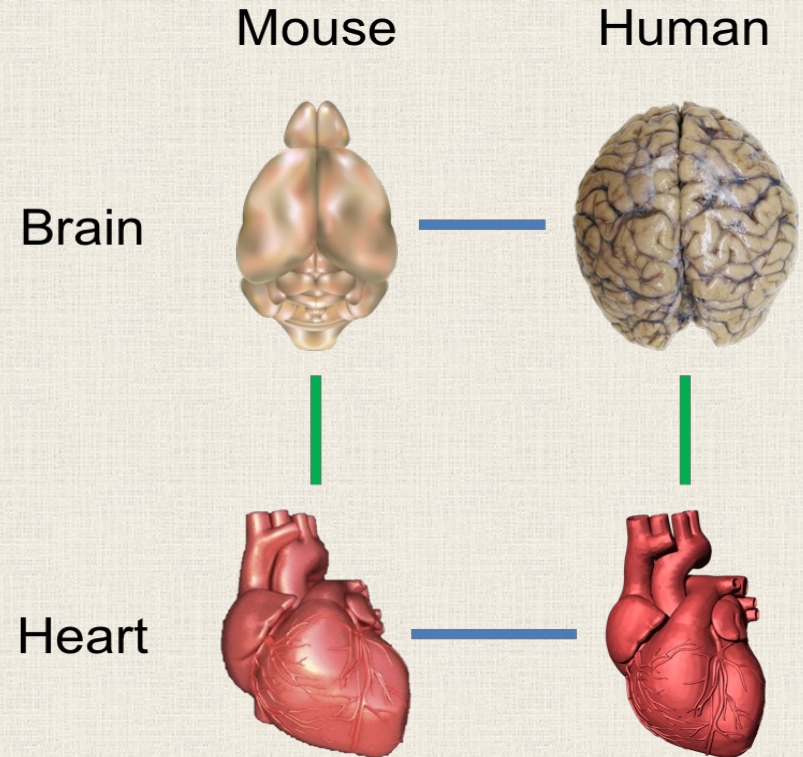
STOP: Now which genes would we report as differentially expressed?

Gene	p-value	Corrected Value
A	0.79	15800
B	0.07	1400
C	0.01	200
D	0.0031	62
E	0.00000079	0.0158

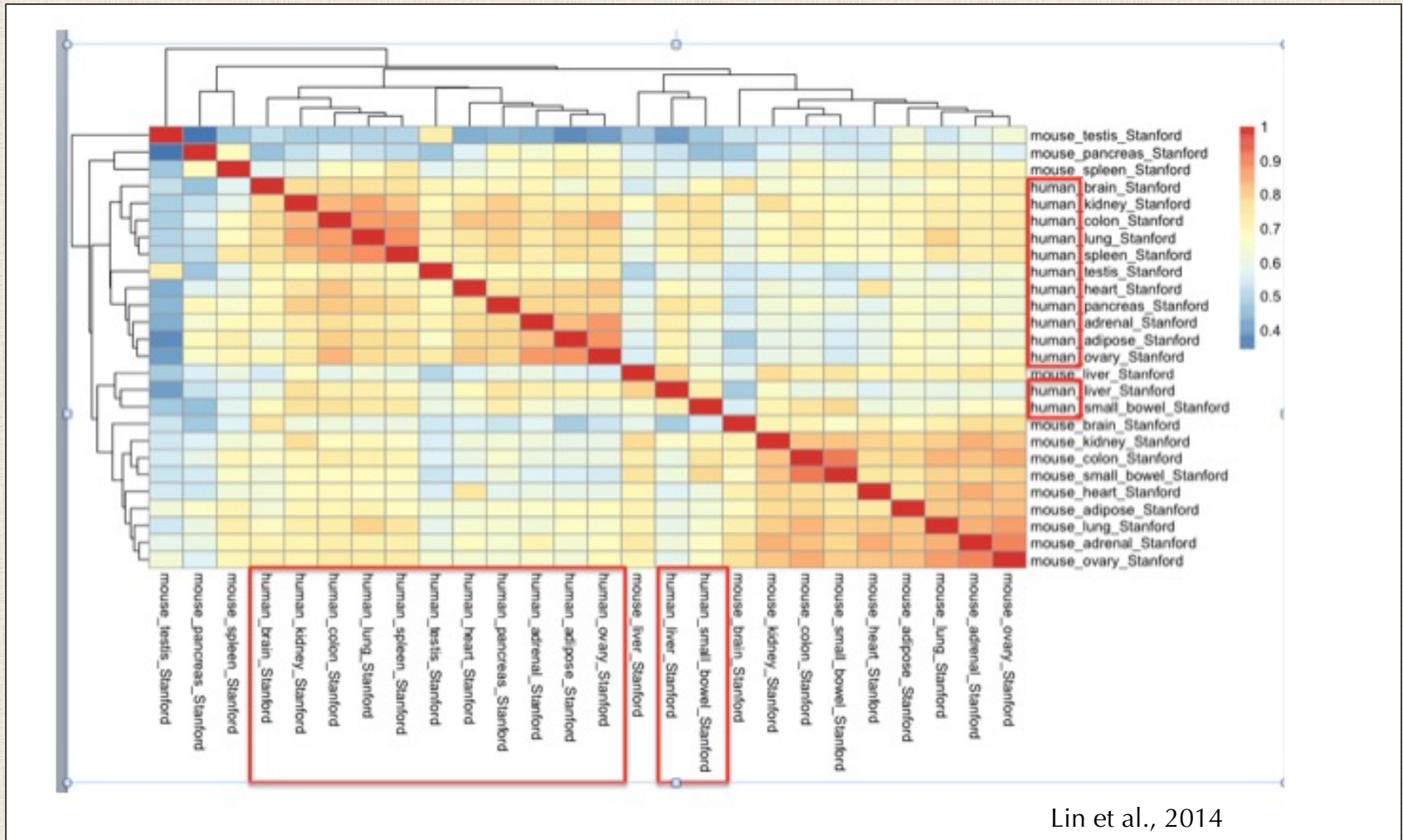
TWO ~~X~~ TWITTER STORIES, AND CLUSTERING CELLS

A short RNA-seq story

STOP: Would you expect the same tissue in two similar species to have more similar gene expression, or different tissues in the same species?



Heatmap of differential expression shows intraspecies similarity across tissue



The problem is batch effects!

RNA-seq is sensitive to **batch effects**, in which experimental conditions can influence the results of the experiment.

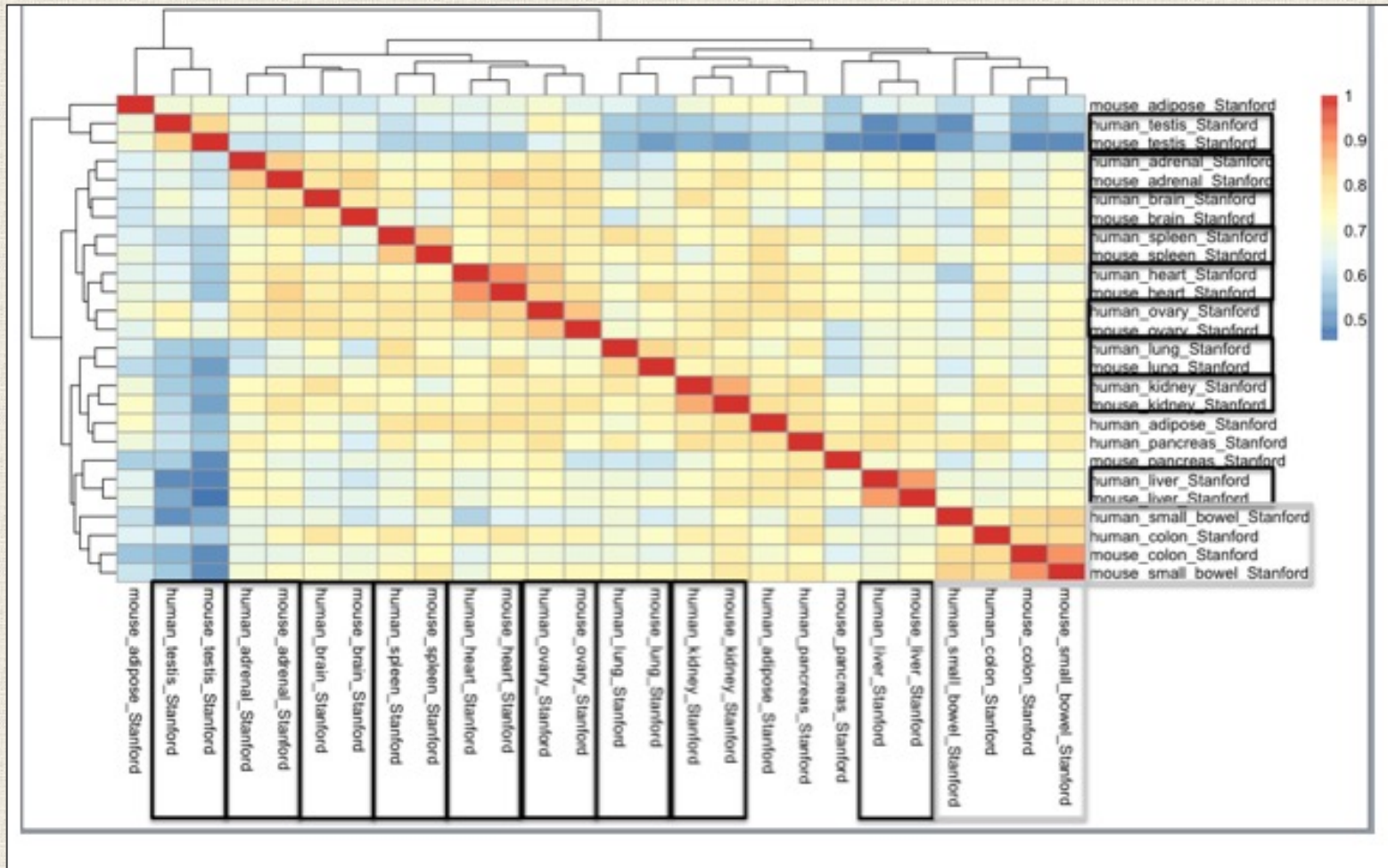
STOP: What should researchers have done instead?

Sequence study design (sequencer ID, run ID, lane number):

D87PMJN1 (run 253, lane 7)	D87PMJN1 (run 253, lane 8)	D4LHBFN1 (run 276, lane 4)	MONK (run 312, lane 6)	HWI- ST373 (run 375, lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● human
testis		pancreas		● mouse

https://twitter.com/Y_Gilad/status/593088451462963202

Heatmap after “batch correction” shows human and mouse cluster by tissue



https://twitter.com/Y_Gilad/status/593088451462963202

From batch RNA-seq to “single cell” RNA-seq

2009: researchers find a way to measure expression of transcripts in a single cell.

From batch RNA-seq to “single cell” RNA-seq

2009: researchers find a way to measure expression of transcripts in a single cell.

For each cell, we obtain a *vector* of expression values x , where x_i is the expression value of the i -th gene/isoform.

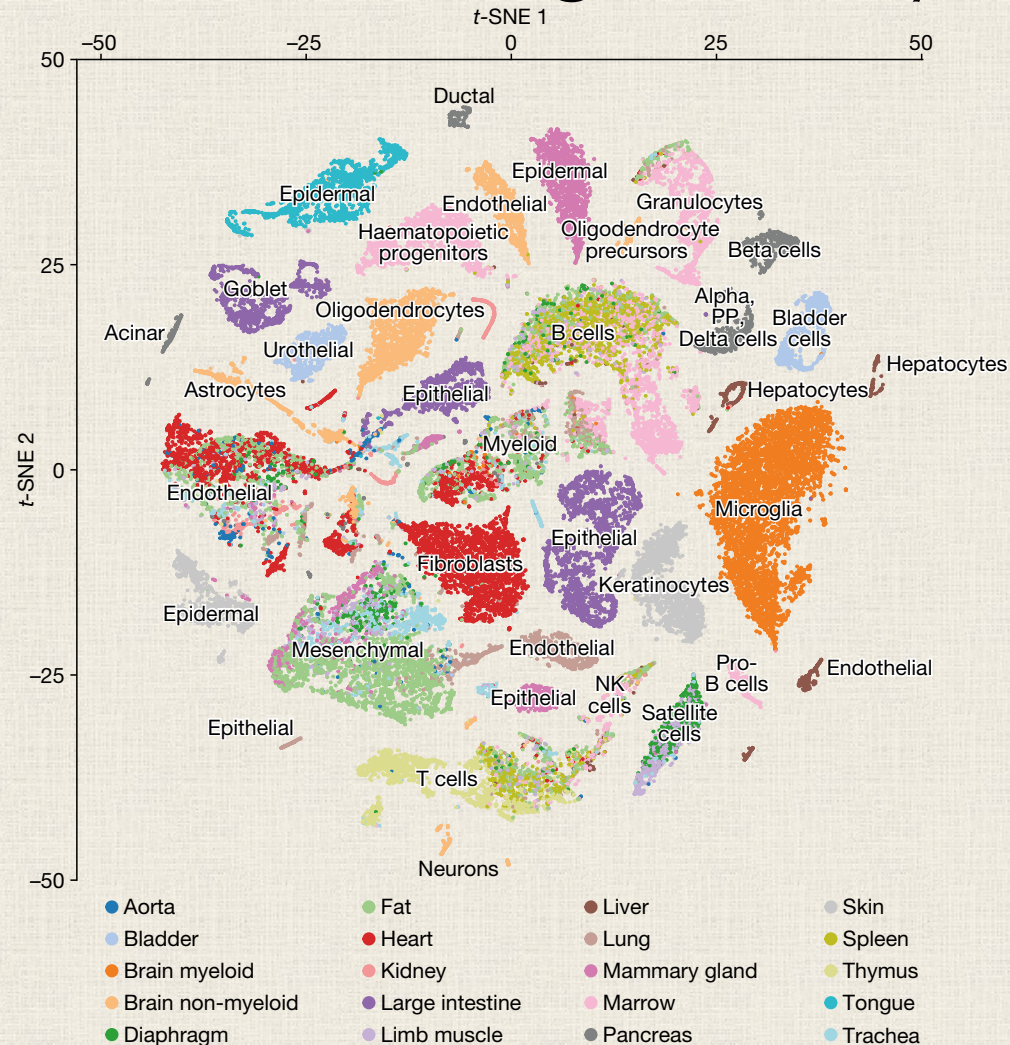
From batch RNA-seq to “single cell” RNA-seq

2009: researchers find a way to measure expression of transcripts in a single cell.

For each cell, we obtain a *vector* of expression values x , where x_i is the expression value of the i -th gene/isoform.

STOP: Having an expression vector for ~ 1 M cells gives us a lot of data. So how do we visualize it?

Dimension Reduction Produces Beautiful Plots Differentiating Cells by Type

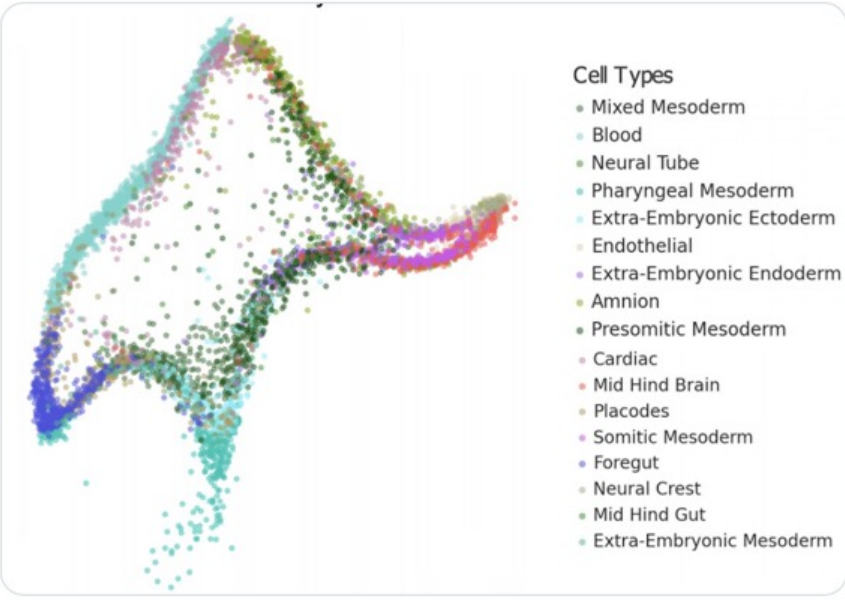


<http://www.nature.com/articles/s41586-018-0590-4>

Dimension reduction is the subject of another controversy ... more later!

Lior Pachter @lpachter

It's time to stop making t-SNE & UMAP plots. In a new preprint w/ Tara Chari we show that while they display some correlation with the underlying high-dimension data, they don't preserve local or global structure & are misleading. They're also arbitrary. [biorxiv.org/content/10.1101...](https://doi.org/10.1101/2021.08.27.454444)



Cell Types

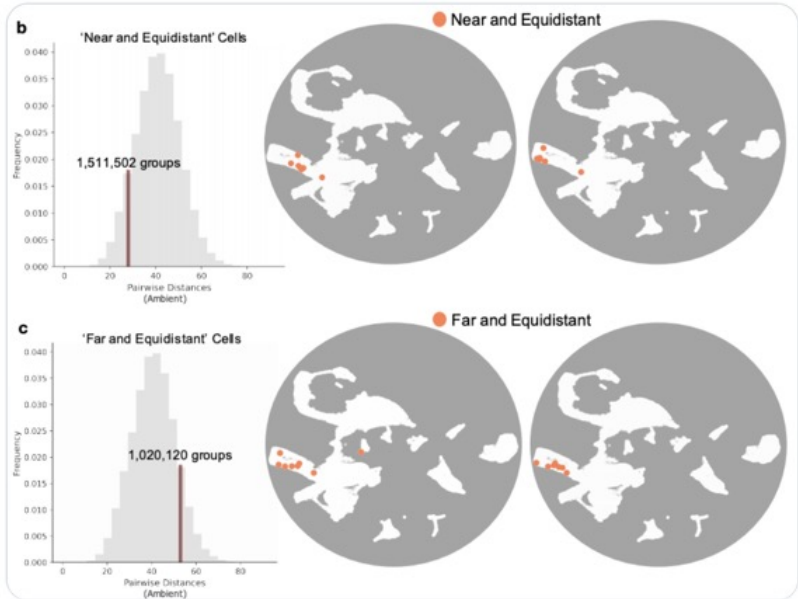
- Mixed Mesoderm
- Blood
- Neural Tube
- Pharyngeal Mesoderm
- Extra-Embryonic Ectoderm
- Endothelial
- Extra-Embryonic Endoderm
- Amnion
- Presomitic Mesoderm
- Cardiac
- Mid Hind Brain
- Placodes
- Somitic Mesoderm
- Foregut
- Neural Crest
- Mid Hind Gut
- Extra-Embryonic Mesoderm

2:41 PM · Aug 27, 2021

95 1.5K 4.3K 1.7K

Lior Pachter @lpachter

On t-SNE & UMAP preserving structure: 1) we show massive distortion by examining what happens to equidistant cells and cell types. 2) neighbors aren't preserved. 3) Biologically meaningful metrics are distorted. E.g., see below:



b 'Near and Equidistant' Cells

Frequency

Pairwise Distances (Ambient)

1,511,502 groups

c 'Far and Equidistant' Cells

Frequency

Pairwise Distances (Ambient)

1,020,120 groups

2:41 PM · Aug 27, 2021

4 37 137 31