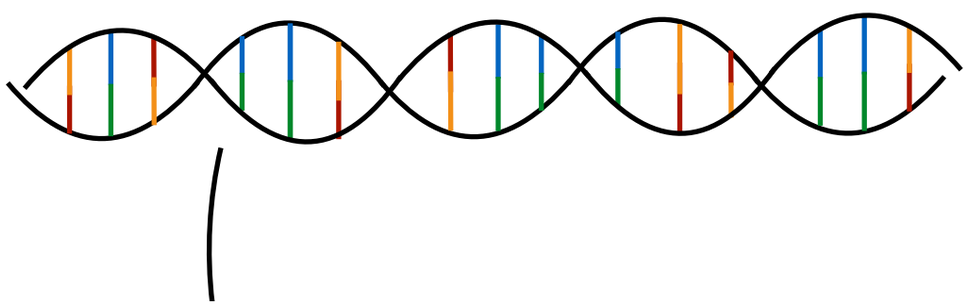
RNA Sequencing & Gene Expression

02-251
Slides by Carl Kingsford

The Central Dogma

DNA



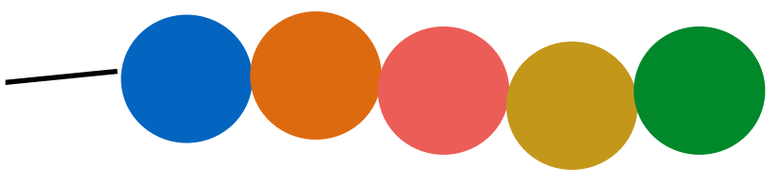
RNA Polymerase
(transcription)

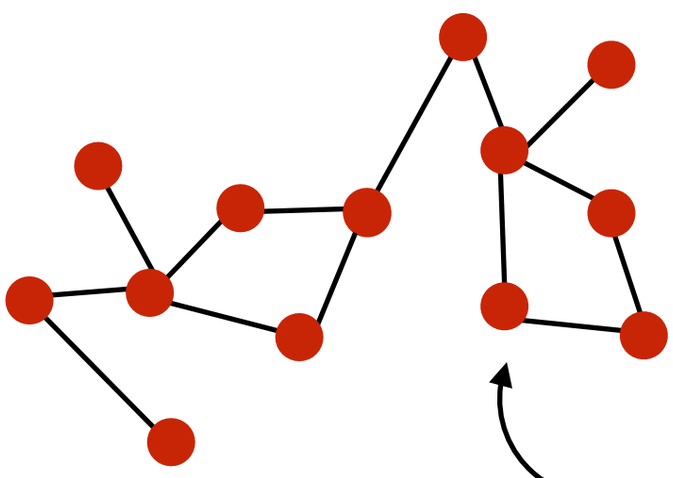
Changes in mRNA expression can have vast effects on cellular function "downstream"

RNA

Ribosomes
(translation)

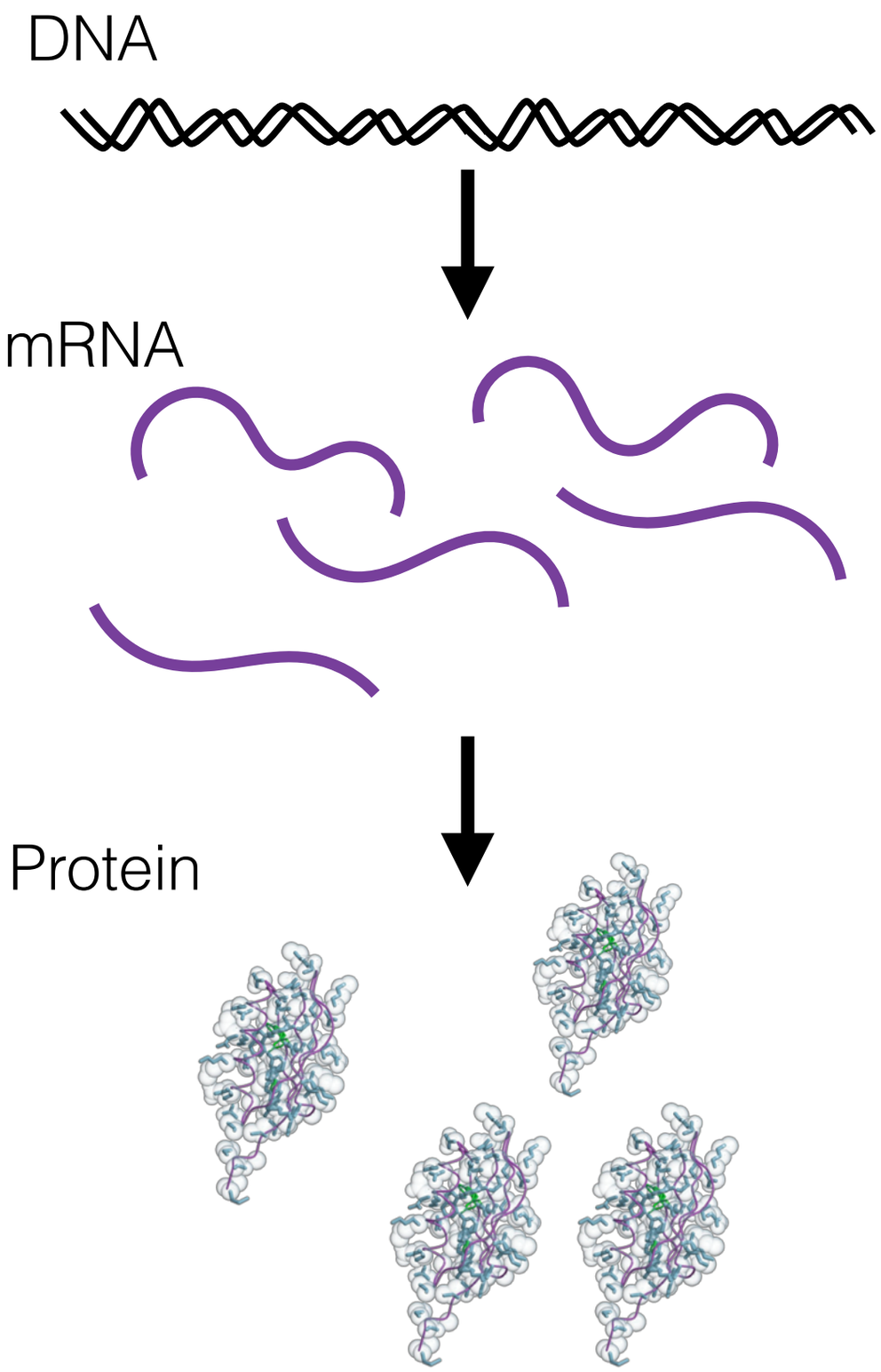
Protein

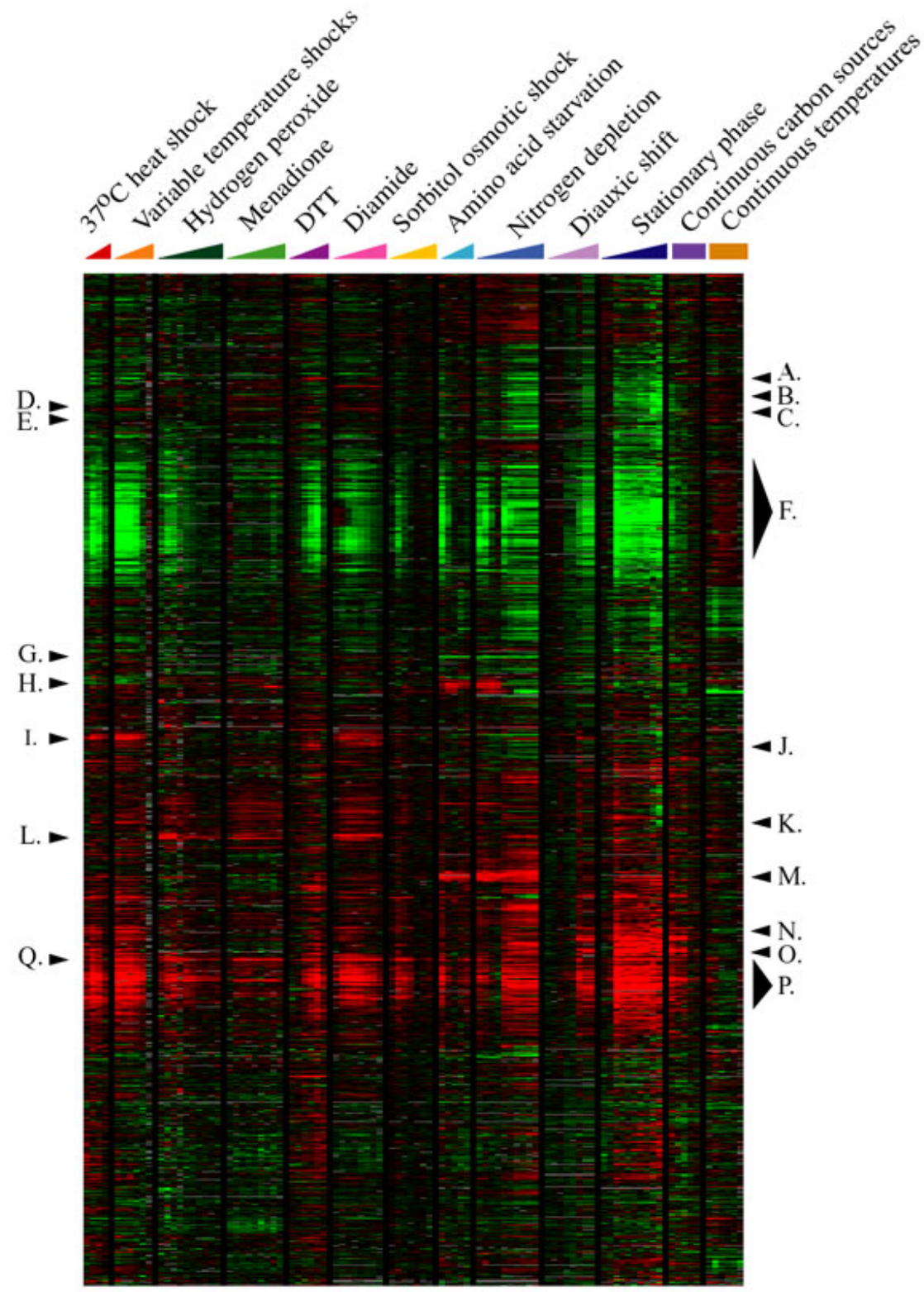




Form networks & pathways; perform a vast set of cellular functions

Gene Expression Varies By Condition





The Genetic Code

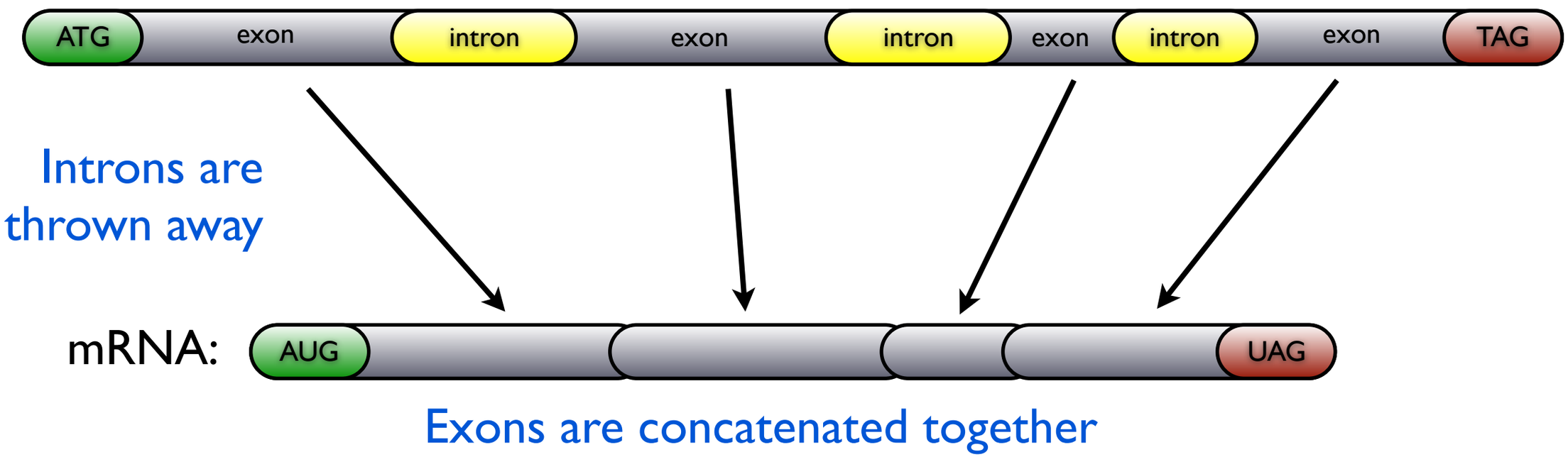
1st base	2nd base: U		2nd base: C		2nd base: A		2nd base: G	
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine
U	UUC	(Phe/F) Phenylalanine	UCC	(Ser/S) Serine	UAC	(Tyr/Y) Tyrosine	UGC	(Cys/C) Cysteine
U	UUA	(Leu/L) Leucine	UCA	(Ser/S) Serine	UAA	Ochre *Stop*	UGA	Opal *Stop*
U	UUG	(Leu/L) Leucine	UCG	(Ser/S) Serine	UAG	Amber *Stop*	UGG	(Trp/W) Tryptophan
C	CUU	(Leu/L) Leucine	CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine
C	CUC	(Leu/L) Leucine	CCC	(Pro/P) Proline	CAC	(His/H) Histidine	CGC	(Arg/R) Arginine
C	CUA	(Leu/L) Leucine	CCA	(Pro/P) Proline	CAA	(Gln/Q) Glutamine	CGA	(Arg/R) Arginine
C	CUG	(Leu/L) Leucine	CCG	(Pro/P) Proline	CAG	(Gln/Q) Glutamine	CGG	(Arg/R) Arginine
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine
A	AUC	(Ile/I) Isoleucine	ACC	(Thr/T) Threonine	AAC	(Asn/N) Asparagine	AGC	(Ser/S) Serine
A	AUA	(Ile/I) Isoleucine	ACA	(Thr/T) Threonine	AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine
A	AUG [A]	(Met/M) Methionine	ACG	(Thr/T) Threonine	AAG	(Lys/K) Lysine	AGG	(Arg/R) Arginine
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine
G	GUC	(Val/V) Valine	GCC	(Ala/A) Alanine	GAC	(Asp/D) Aspartic acid	GGC	(Gly/G) Glycine
G	GUA	(Val/V) Valine	GCA	(Ala/A) Alanine	GAA	(Glu/E) Glutamic acid	GGA	(Gly/G) Glycine
G	GUG	(Val/V) Valine	GCG	(Ala/A) Alanine	GAG	(Glu/E) Glutamic acid	GGG	(Gly/G) Glycine

- There are 20 different amino acids & 64 different codons.
- Lots of different ways to encode for each amino acid.
- The 3rd base is typically less important for determining the amino acid
- Three different "stop" codons that signal the end of the gene
- Start codons differ depending on the organisms, but AUG is often used.

Eukaryotic Genes & Exon Splicing

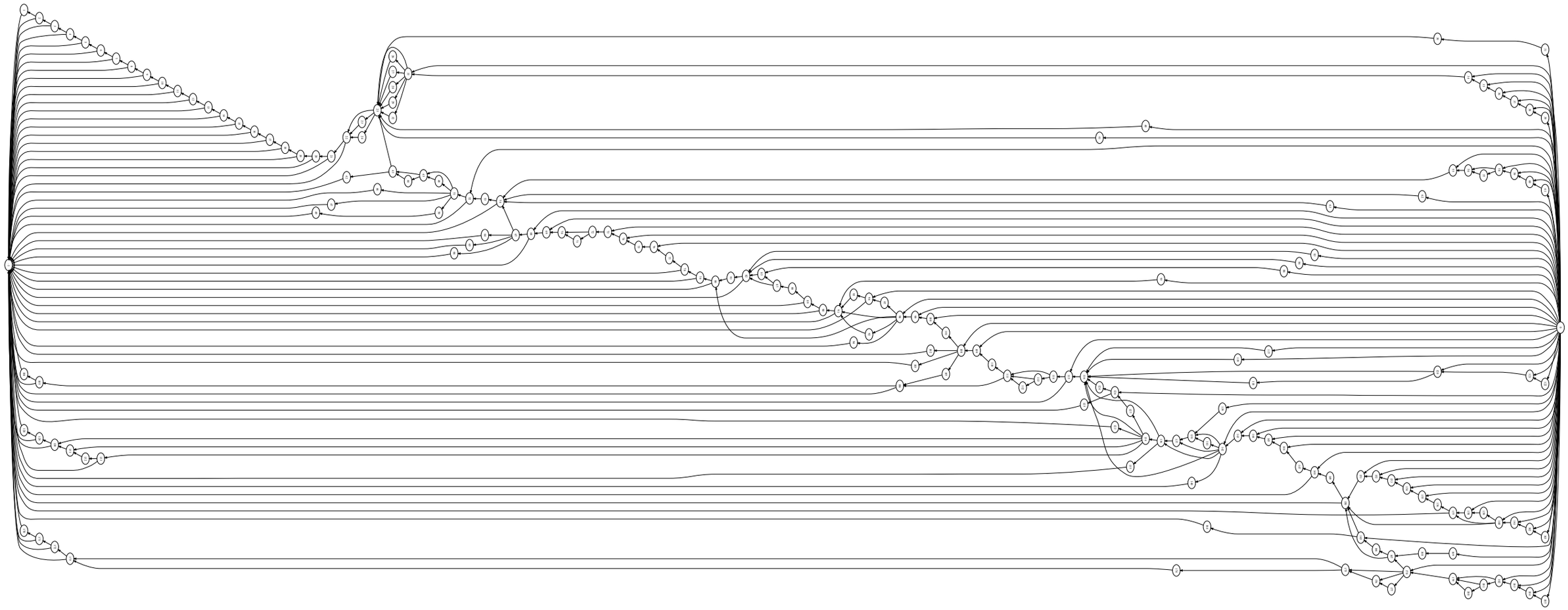
Prokaryotic (bacterial) genes look like this:

Eukaryotic genes usually look like this:



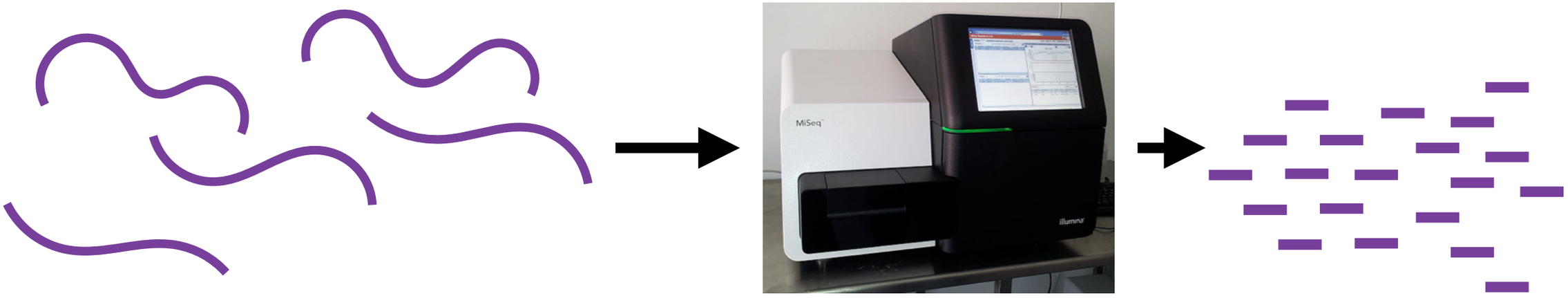
This spliced RNA is what is translated into a protein.

Alternative Splicing: It can get pretty complex…



- Splice graph of a nurexin, which is a presynaptic protein that helps to connect neurons at the synapse.
  - Node = exon (or part of an exon)
  - Edge (a,b) = sequence b can follow sequence a in some transcript

Transcriptome Sequencing

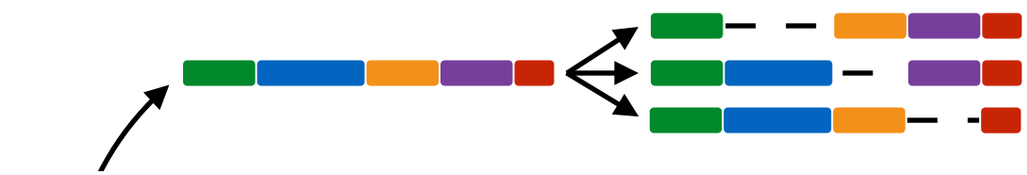


mRNA in a cell under
a given condition

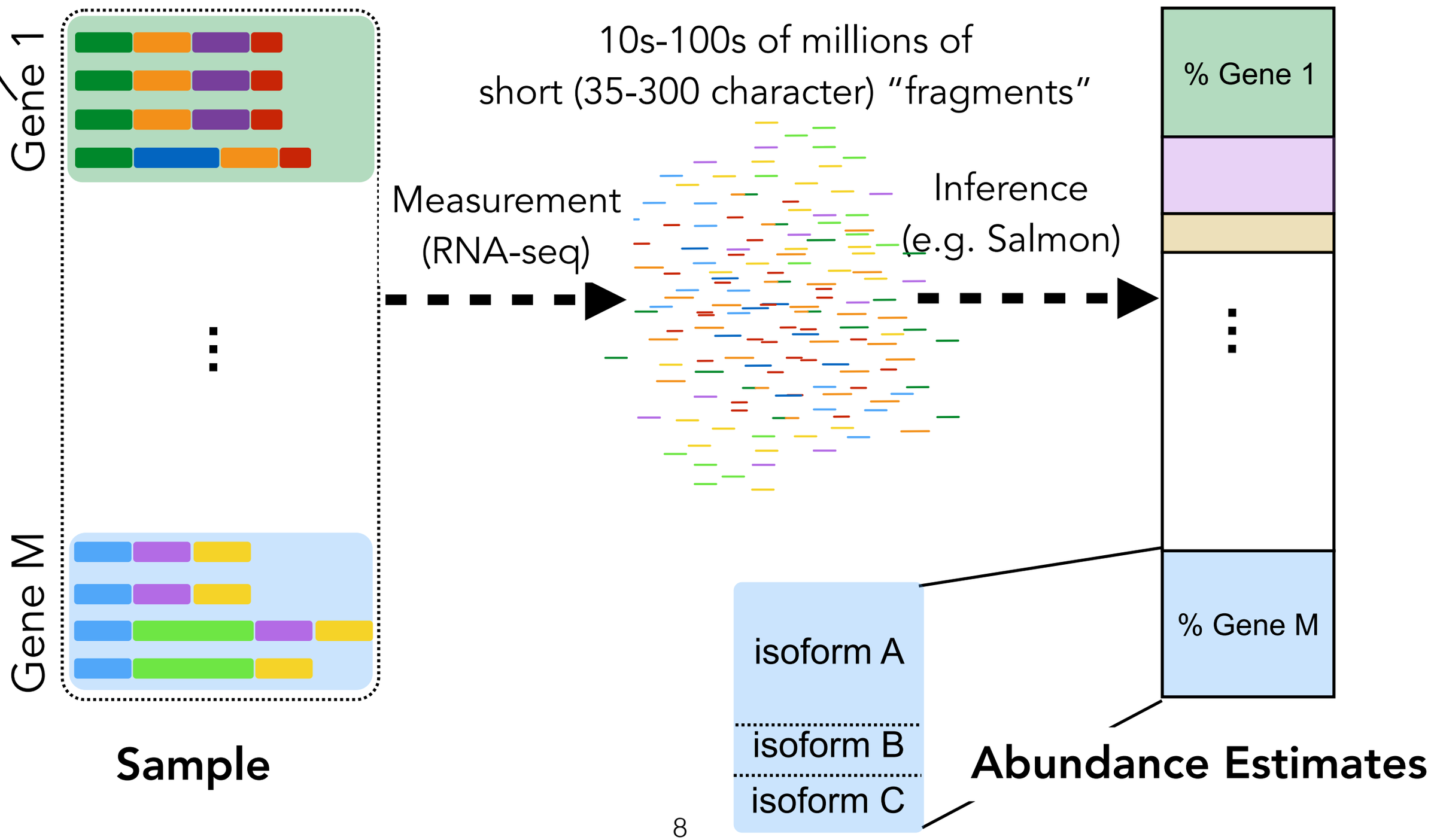
Given:
- Collection of short reads
- A set of known transcript sequences

Estimate:
- The relative abundance of each transcript

Transcript Quantification: An Overview

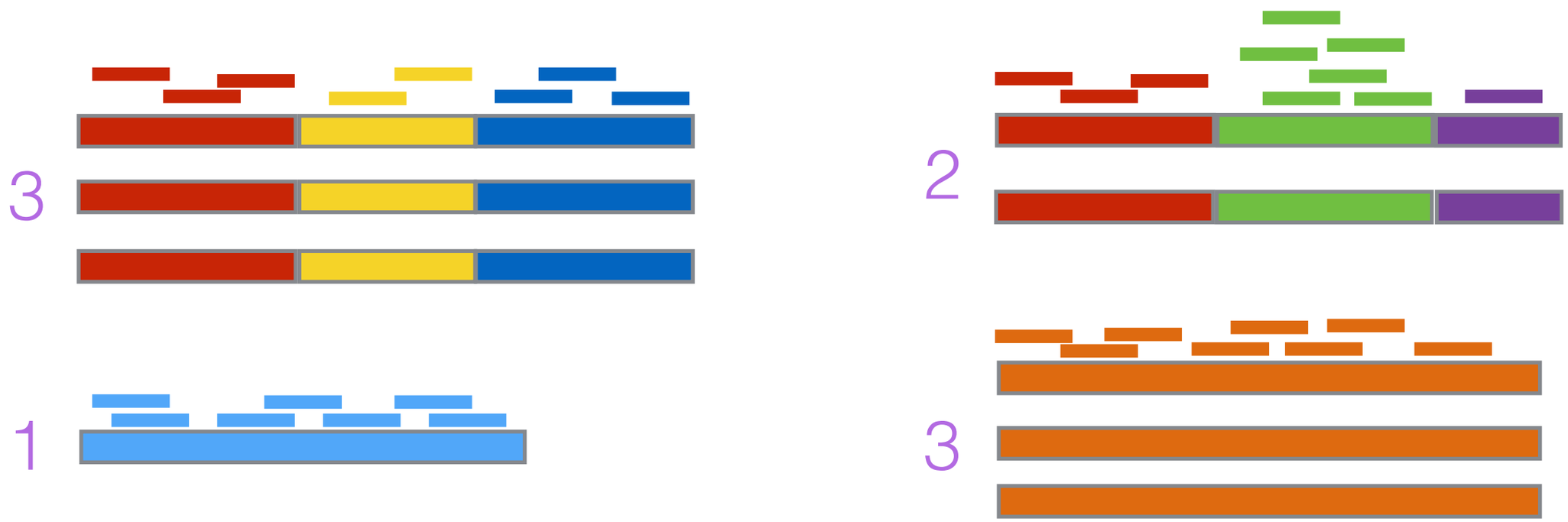


1 gene ⇒ many variants (isoforms)



Alternative Splicing is the Main Challenge

Goal: estimate the abundance of each kind of transcript given short reads sampled from the expressed transcripts.

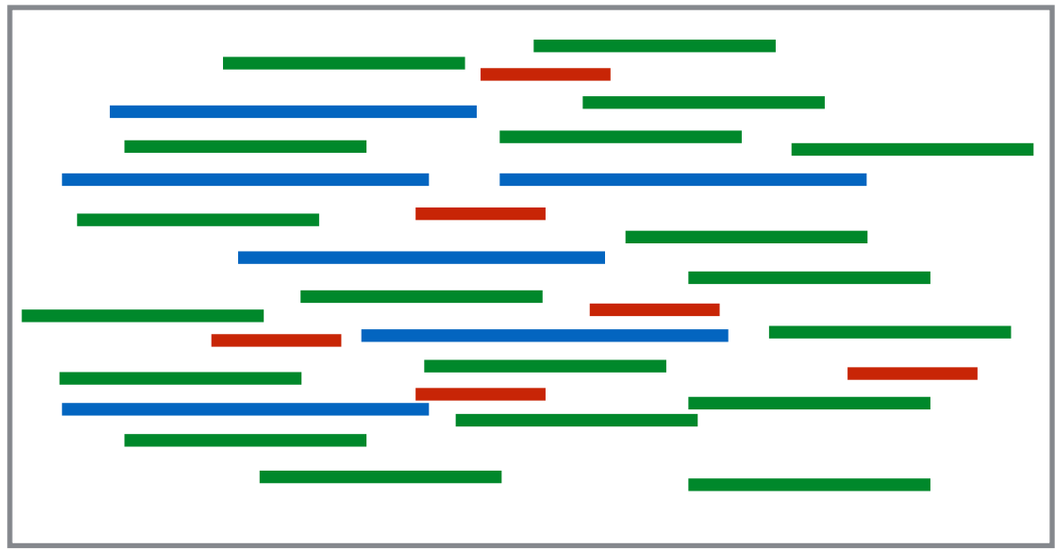


<u>Challenges</u>:

- hundreds of millions of short reads per experiment
- finding locations of reads (mapping) is traditionally slow
- alternative splicing creates ambiguity about where reads came from
- sampling of reads is not uniform

Inference Problem

Experimental mixture:



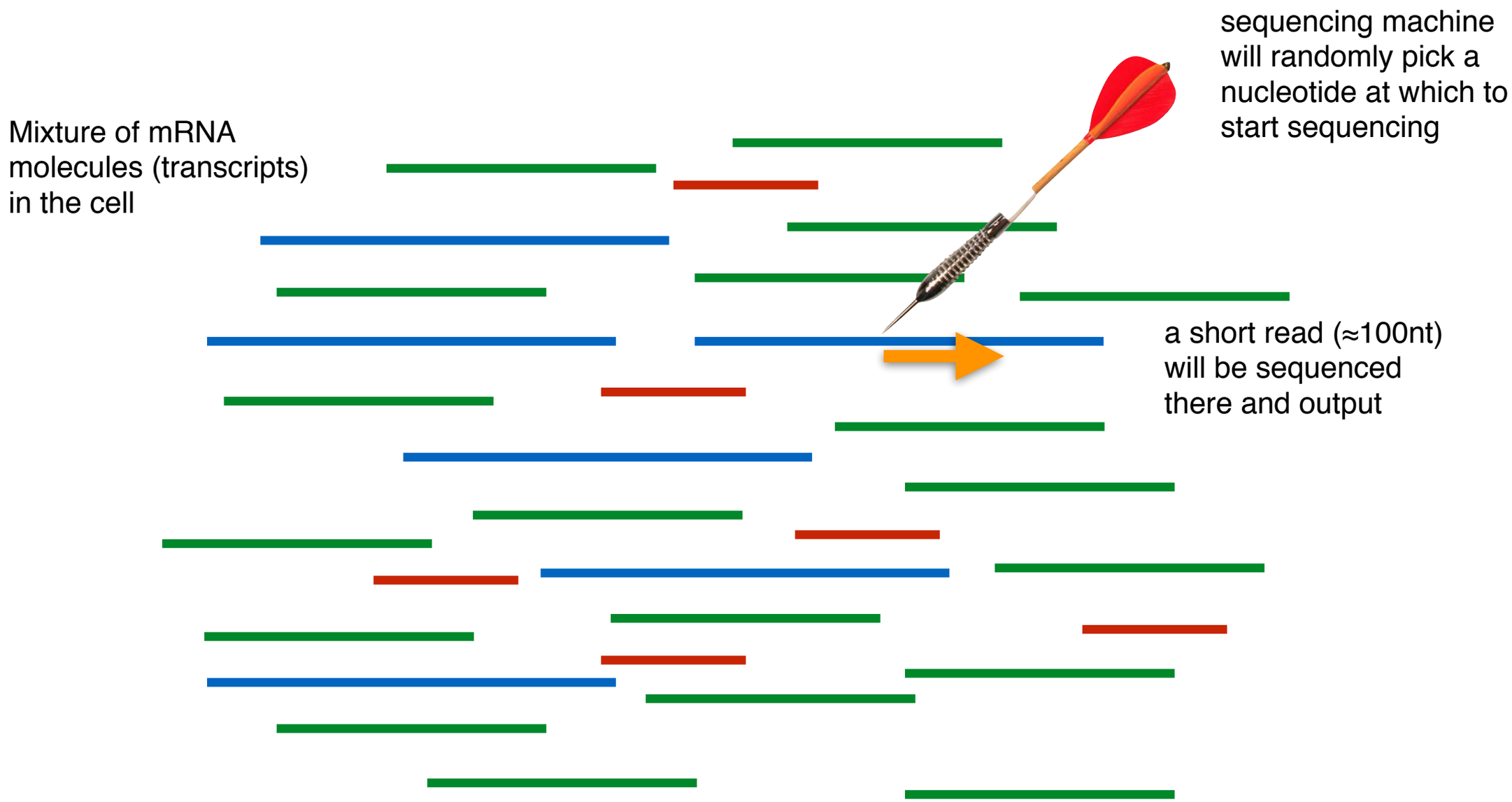
length( ) = 100 x 6 copies = 600 nt ~ 30% blue

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

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

These values η = [0.3, 0.6, 0.1] are the *nucleotide fractions*; they are the quantities we want to infer

Model of Sequencing



Motif Finding ↔ Gene Expression

Gene Expression		Motif Finding
• **Have**: reads	x, "data", f_i	• **Have**: sequences
• **Want**: abundance vector (model)	η or M	• **Want**: sequence profile matrix (model)
• **Hidden data**: which transcript each read came from	z	• **Hidden data**: where the motifs start in each sequence

⇒ perhaps a similar EM approach will work

To Apply EM

- Recall the main EM equation:

$$\underbrace{\mathbb{E}_z \log \Pr(\text{data} \mid M_{\text{new}})}_{Q(M_{\text{new}} \mid M_{\text{old}})} = \underbrace{\sum_z \Pr(z \mid \text{data}, M_{\text{old}}) \overbrace{\log \Pr(\text{data}, z \mid M_{\text{new}})}^{\text{quality of new model}}}_{\text{in expectation over choices of hidden variables (drawn according to the old model)}}$$

data	sequenced fragments
M (model)	transcript abundances (η)
z	which transcript each fragment came from

- Let's write Pr(data | model) in terms of gene expression notation (next slide)

Gene Expression Inference

- **Want**: expression η(t) for each transcript t.
- **Observed**: Sequence fragments f_i sampled from molecules in the cell.
- **Hidden variables:**

indicator variable: if we knew this, answer would be easy to compute

$$z_{ti} = \begin{cases} 1 & \text{if fragment } i \text{ came from transcript } t \\ 0 & \text{otherwise} \end{cases}$$

- Goal:

hidden variables

$$\underset{\eta}{\text{maximize}} \;\; g(\eta) = \Pr(\{f_i\} \mid \eta) = \sum_z \boxed{\Pr(\{f_i\}, z \mid \eta)}$$

observed data

unknown model

green box is our focus

Why do we introduce these hidden variables?

- Computing $\Pr(\{f_i\} \mid \eta)$ directly is complicated

- Once we introduce z, we have, by conditional probability:

green box from last slide

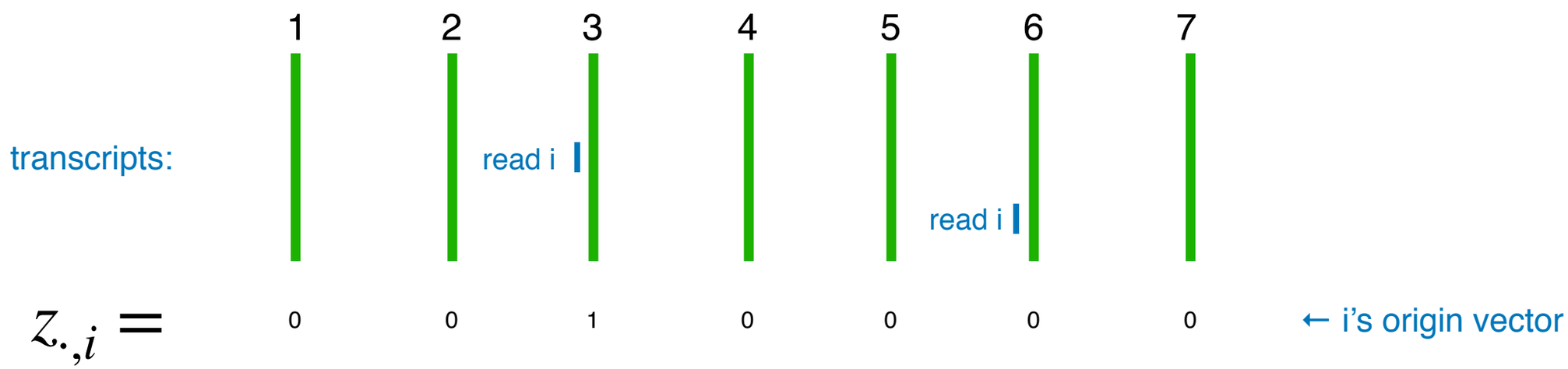
$$\boxed{\Pr(\{f_i\}, z \mid \eta)} = \Pr(z \mid \eta) \Pr(\{f_i\} \mid z, \eta)$$

- and we only need to compute:

$\Pr(z \mid \eta)$ Probability of picking transcript, given abundances

$\Pr(\{f_i\} \mid z, \eta)$ Probability of generating a fragment, given where it came from

Treat Each Sequence Read Independently



$$\Pr(z \mid \eta) = \prod_i \Pr(z_{\cdot,i} \mid \eta)$$

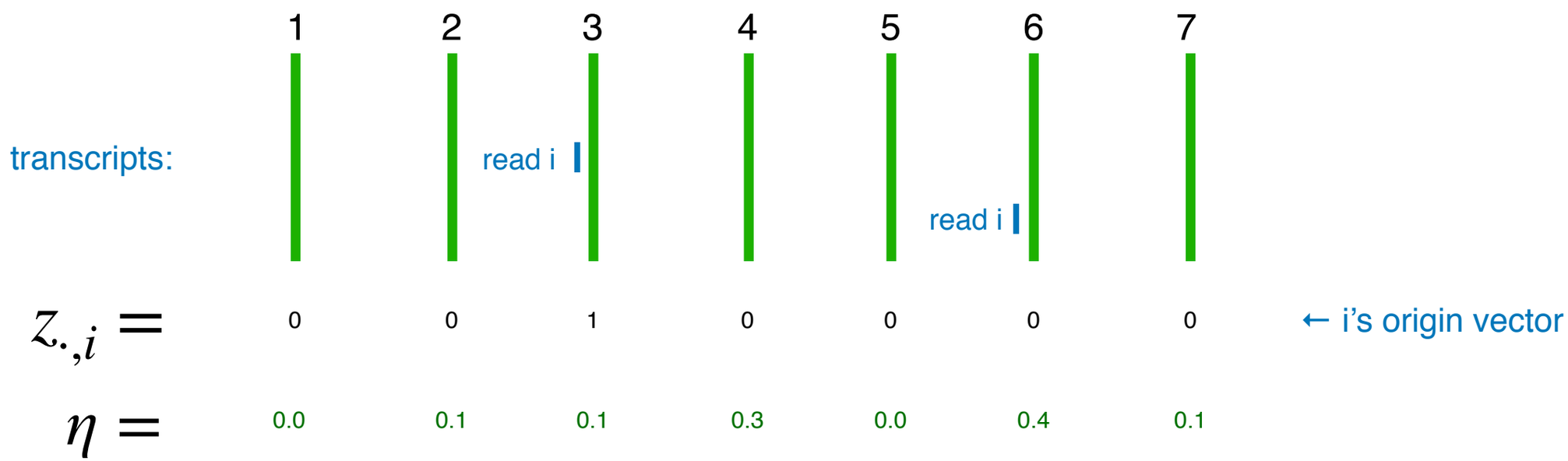
← probability of fragment i's origin vector

$$= \prod_i \sum_t \Pr(z_{ti} = 1 \mid \eta)$$

← because each read has only 1 origin, we can just ask which z_{ti} is 1 in its origin vector.

↑ Probability that read i came from transcript t given the abundances

But this is "easy":



$$\Pr(z_{ti} = 1 \mid \eta) = \frac{\eta(t)}{\sum_q \eta(q)}$$

q sums over the transcripts where the read mapped

In example above:

$\Pr(z_{3i} = 1 \mid \eta) = 0.1 / (0.1 + 0.4) = 0.2$

$\Pr(z_{6i} = 1 \mid \eta) = 0.4 / (0.1 + 0.4) = 0.8$

$\Pr(z_{2i} = 1 \mid \eta)$ because i didn't map to transcript 2

What about the other probability?

Again, treat each fragment independently:

When looking at fragment i, only i's origin vector matters

$$\Pr(\{f_i\} \mid z, \eta) = \prod_i \Pr(f_i \mid z_{\cdot,i}, \eta)$$

Since the "z" is given (after the I in the probability), we can assume we know which transcript f_i came from.

Need to compute only:

$$\Pr(f_i \mid z_{ti} = 1, \eta) = \begin{cases} 0 \text{ if } f_i \text{ doesn't map to transcript } t \\ \text{probability of generating read } f_i \text{ from the sequence of transcript } t \end{cases}$$

(see next slide)

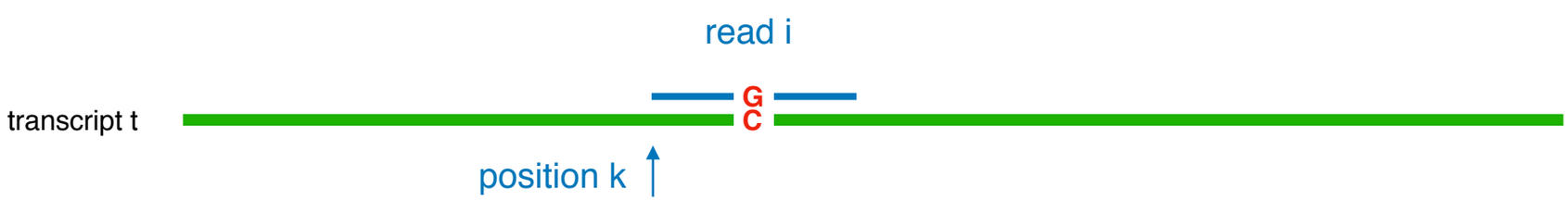
where you can encode "intelligence" about the experimental process

Fragment Generation Probability

Fragment generation probability is independent of abundances *once you know which fragment the transcript came from:*

$$\Pr(f_i \mid z_{ti} = 1, \eta) = \Pr(f_i \mid z_{ti} = 1)$$

Can estimate the above probability using various terms that model the experiment:



$$\Pr(f_i \mid z_{ti} = 1) \approx$$

Prob. of seeing a fragment of length len(f_i)	x	Prob. of sequencing a fragment at position k	x	Prob. of seeing so many differences between read and transcript

Expression Quantification EM Summary

- Want: expression vector; missing knowledge of where sequenced fragments came from

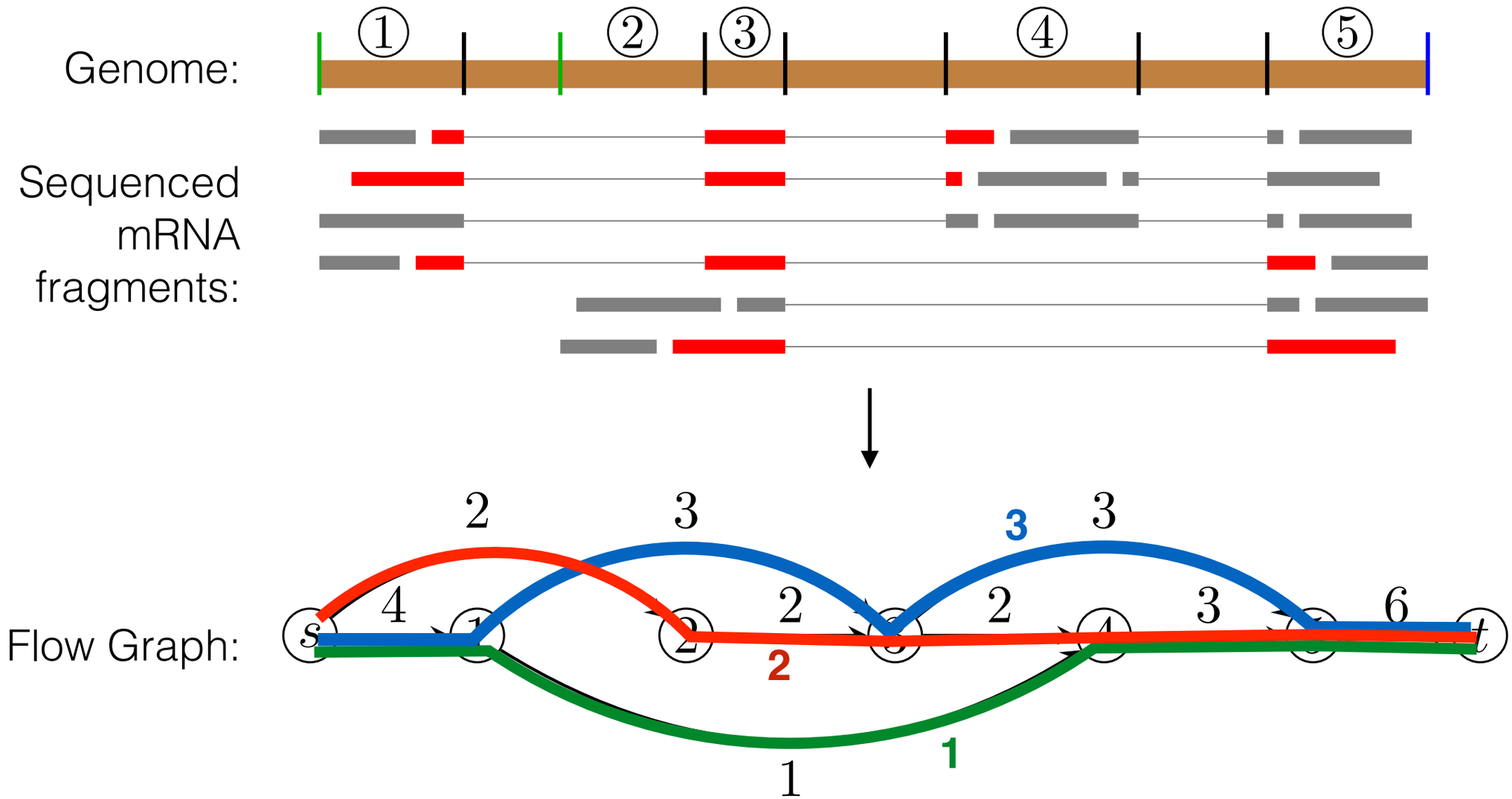
- Solution: add “hidden” (aka latent) variables z, and estimate Pr(fragments, z | model).

- Do this by breaking it into 2 parts:
  Pr(z | model, fragments)
  Pr(fragments | z, model)
  each of which is easier to estimate

- Once we can compute the above probabilities, we can apply EM.

Transcript Assembly

Sequencing Isoforms



Mingfu Shao and Carl Kingsford. Scallop Enables Accurate Assembly Of Transcripts Through Phasing-Preserving Graph Decomposition. *Nature Biotechnology* (2017)