Transcript Quantification using RNA-Seq

Dr. Jared Simpson
Ontario Institute for Cancer Research
&
Department of Computer Science
University of Toronto

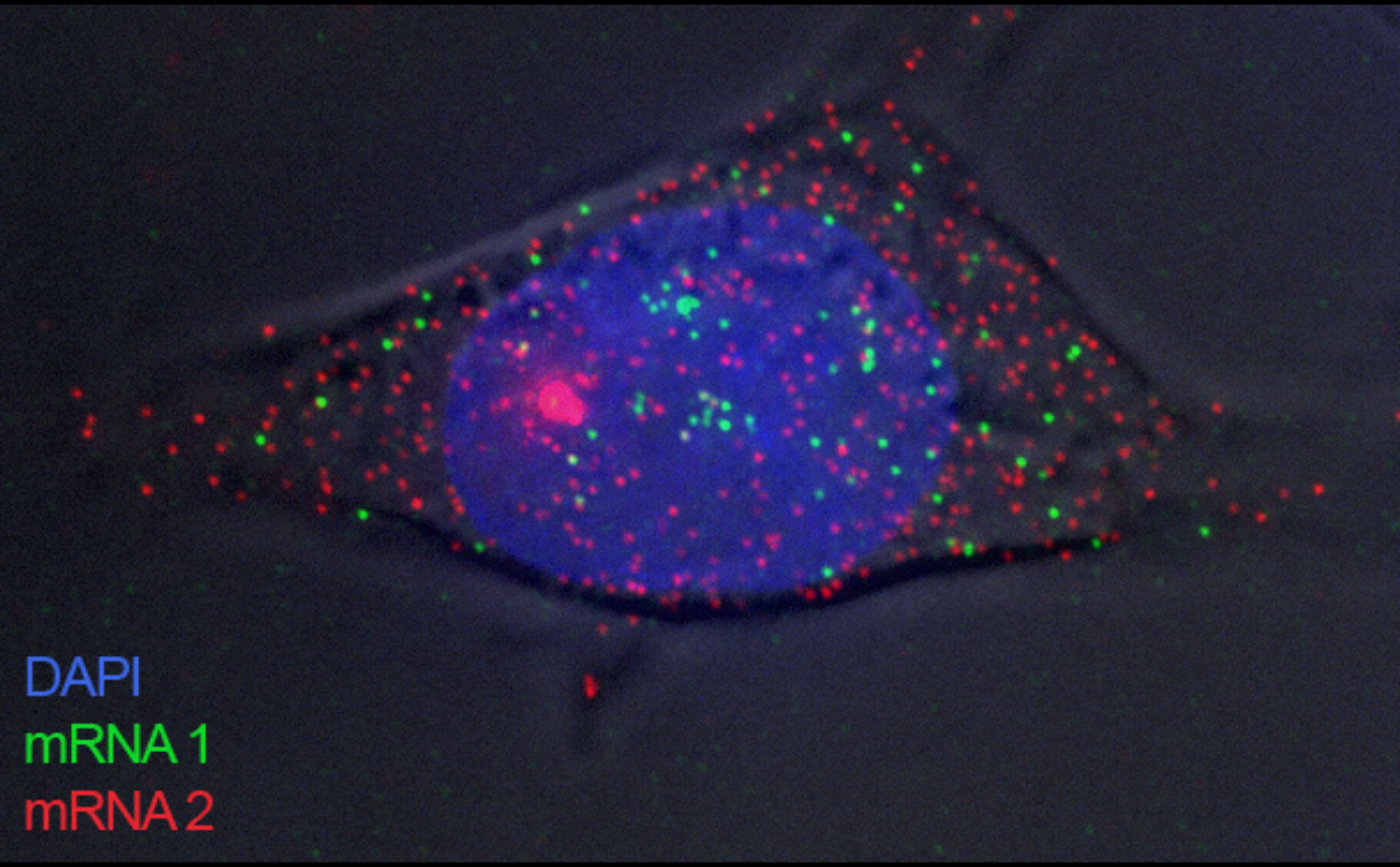
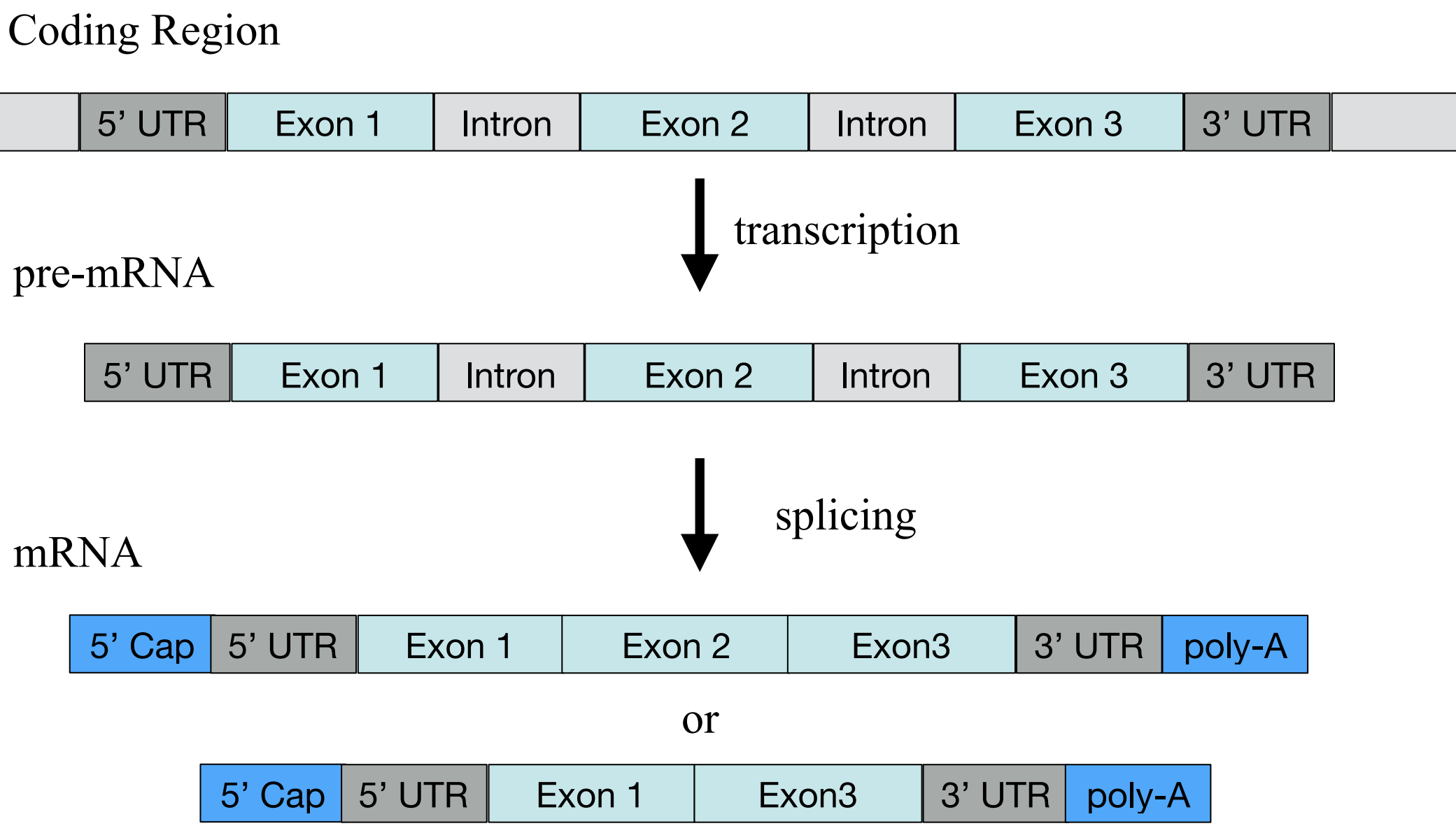


Image: https://homepages.warwick.ac.uk/staff/D.Hebenstreit/

Gene Expression

- What genes are expressed in cell type X and in what quantities?
- What genes are expressed in all cells? ("housekeeping genes")
- What genes are *differentially* expressed in cells A relative to B?
- How do gene expression patterns change over time, during development, as a result of treatment with chemical Y, …
- Answering all of these questions requires methods to quantify the abundance of transcripts

Transcription and Splicing



RNA Sequencing

Most sequencing instruments can't directly sequence RNA molecules. Must convert to cDNA first.

mRNA GATAGCTACTATATACGCCCATCGATTGAAAAAAAAAAAAAA

↓ *reverse* transcription

cDNA CTATCGATGATATATGCGGGTAGCTAACTTTTTTTTTTTTTTT
GATAGCTACTATATACGCCCATCGATTGAAAAAAAAAAAAAA

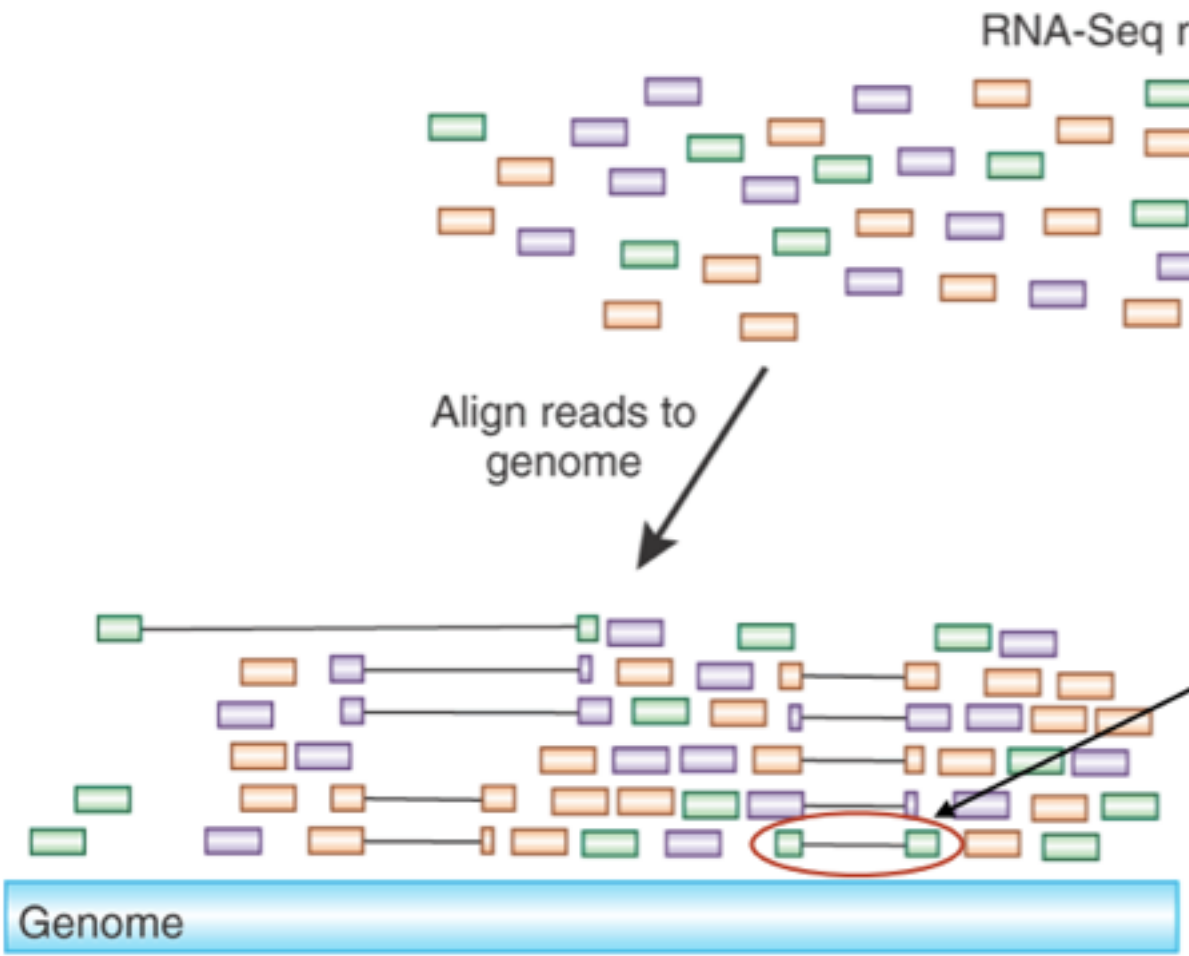
↓ sequence

Analysis Problems

- Can we use the alignment methods we've already seen to align RNA-Seq reads to a reference genome?

  Yes but require modifications to allow large gaps for introns

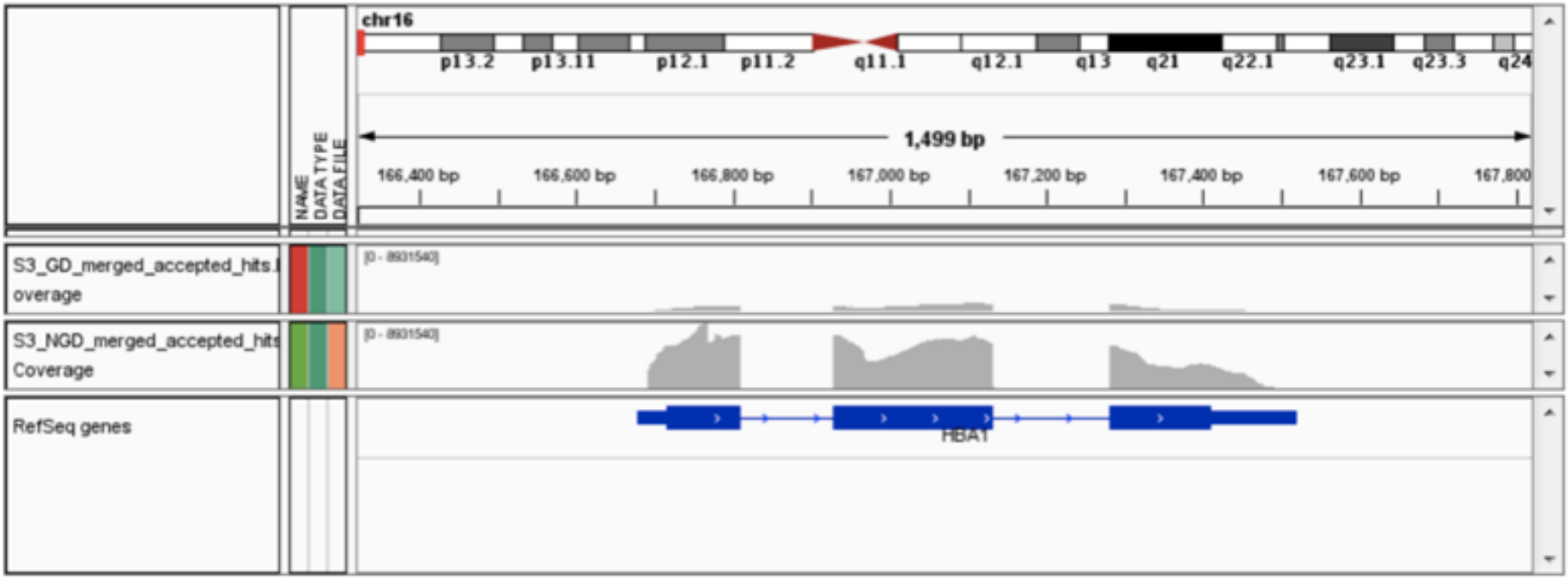
RNA-Seq Alignment



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

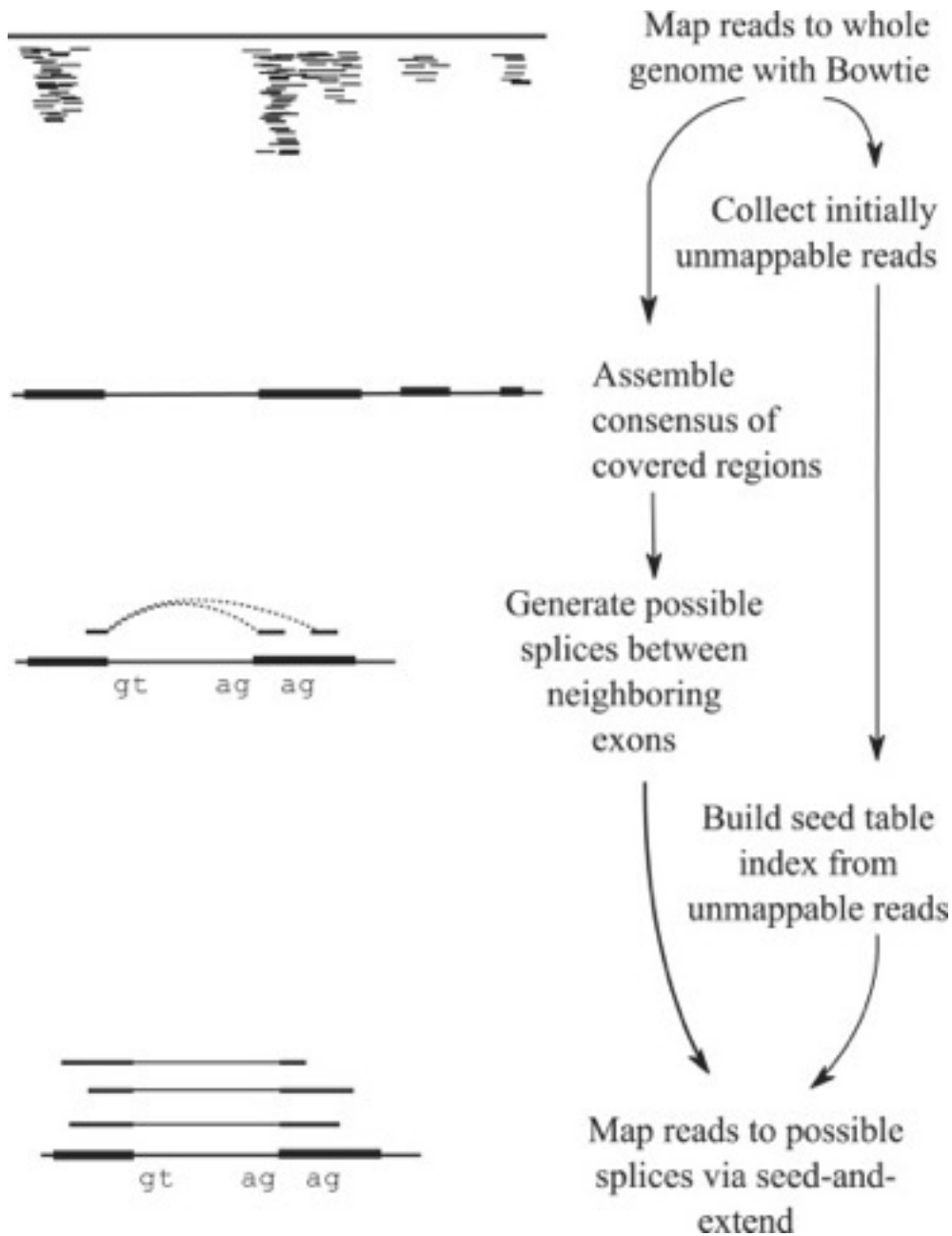
From Rob Patro

RNA-Seq Alignment



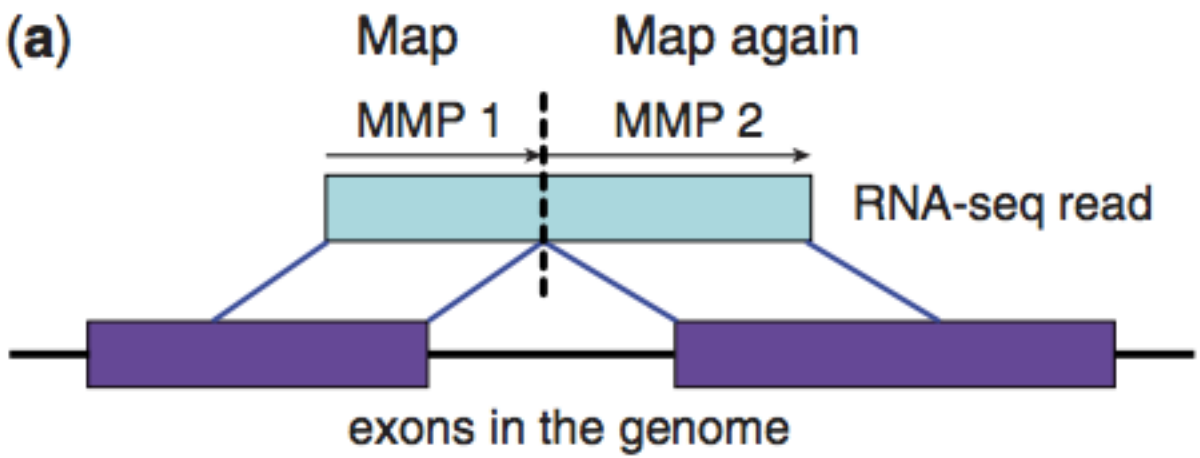
From Rob Patro

RNA-Seq Alignment using Tophat



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2672628

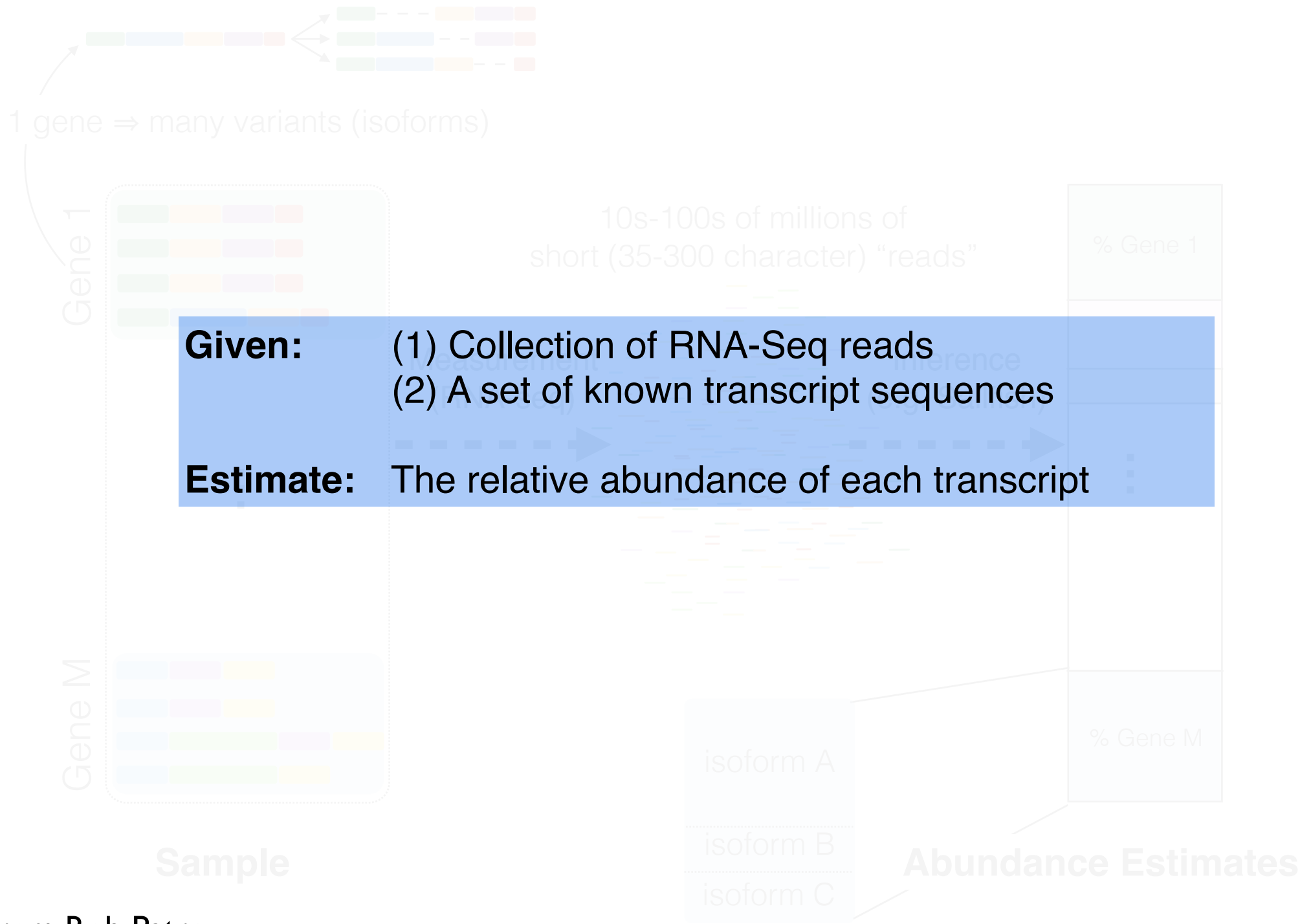
RNA-Seq Alignment using STAR



Step 1: find maximal matching prefix of read
Step 2: repeat for the suffix of the read, while making sure partial matches consistent with splice donors/acceptors

From STAR: ultrafast universal RNA-Seq aligner

Abundance Estimation: An Overview

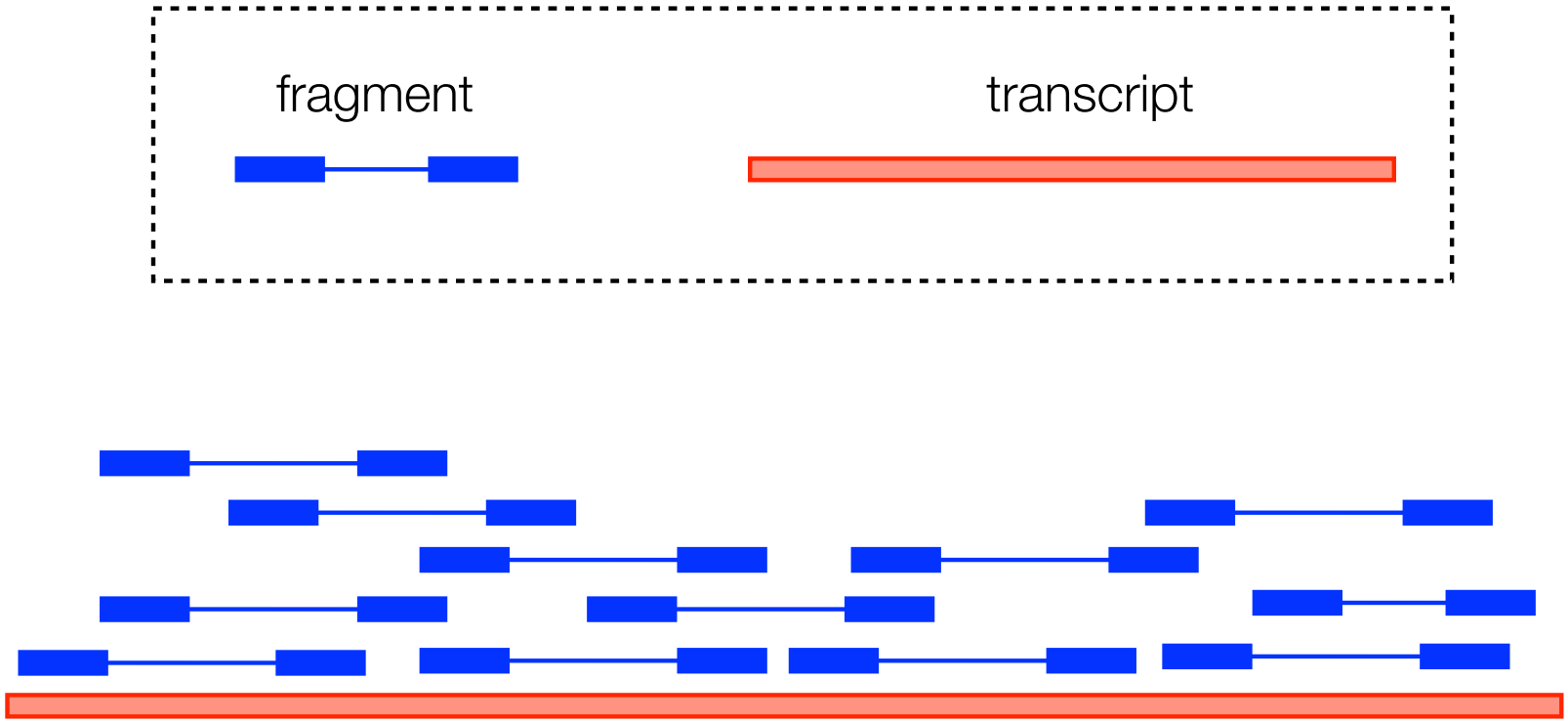


Given: (1) Collection of RNA-Seq reads
(2) A set of known transcript sequences

Estimate: The relative abundance of each transcript

From Rob Patro

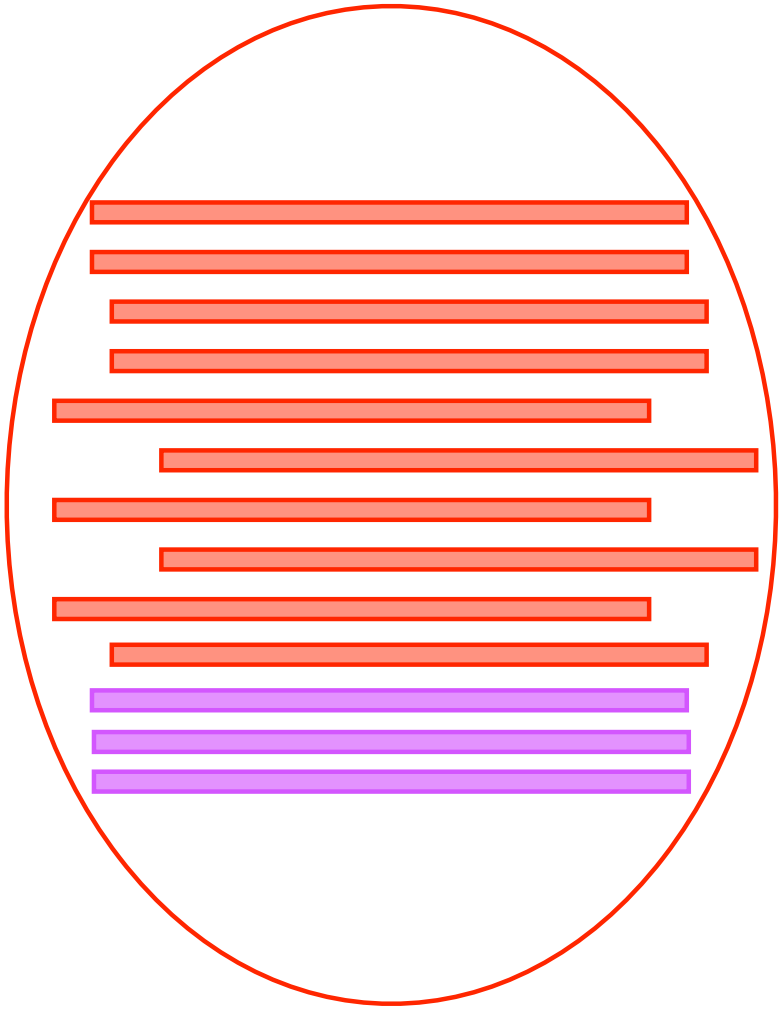
Basic principles of quantification



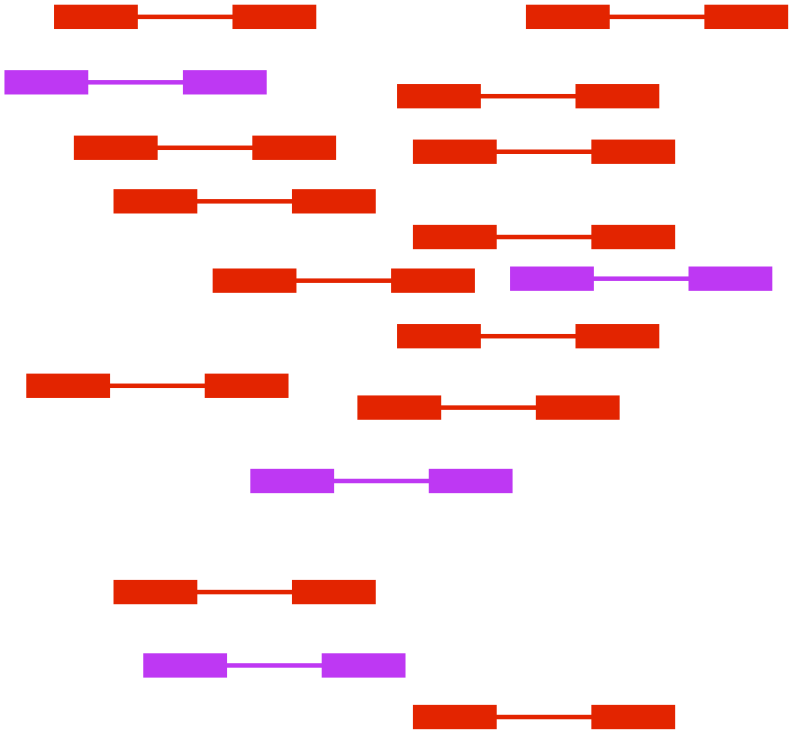
slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf) via Rob Patro

Basic principles of quantification

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



Transcriptome

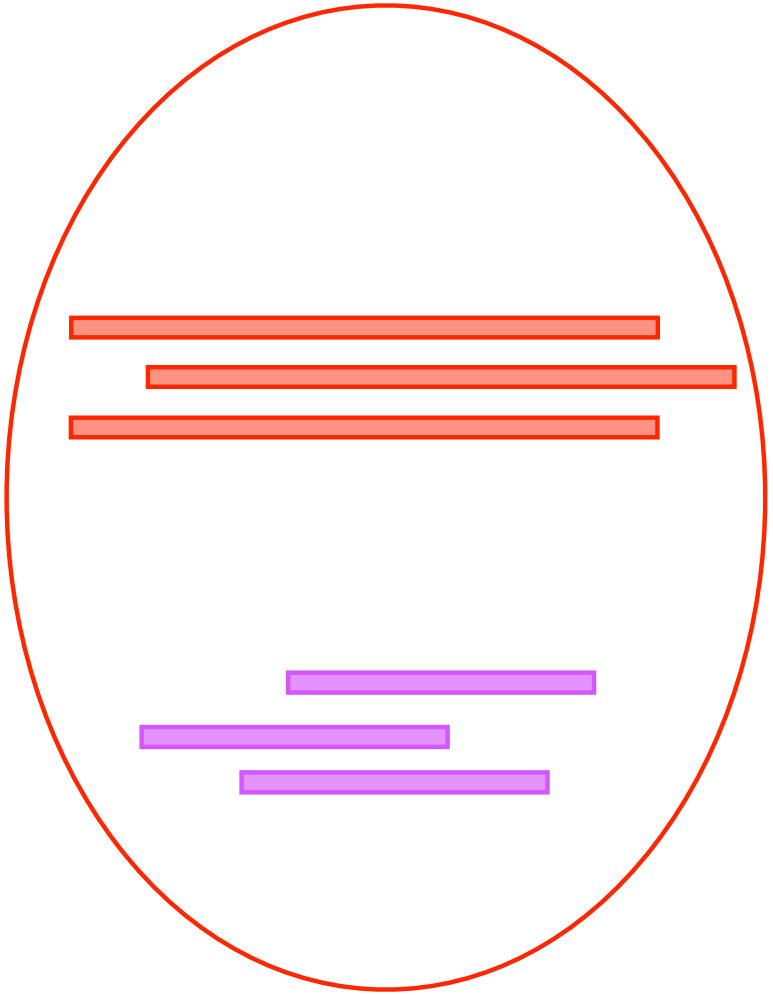


Sequenced fragments

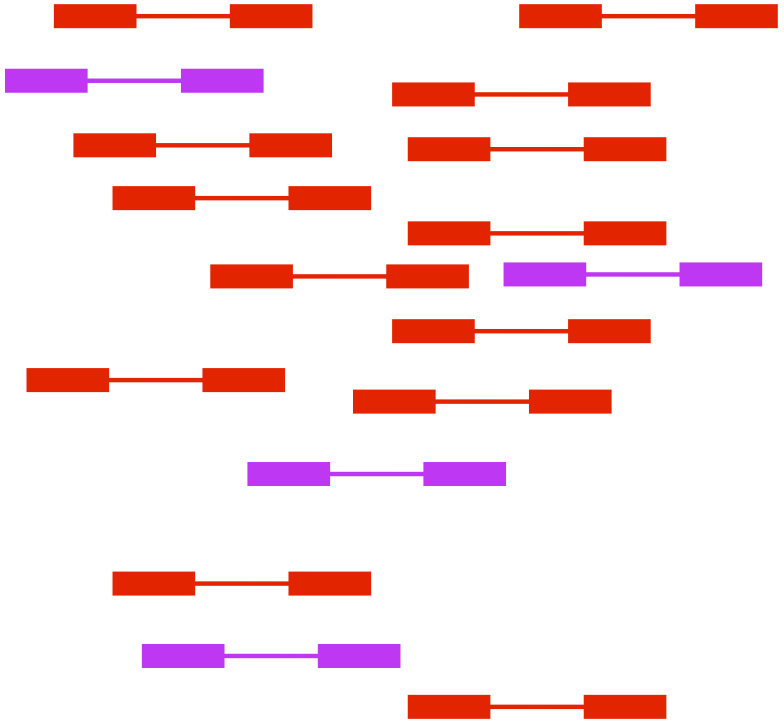
slide from Cole Trapnell (http://www.stat115.org/lectures/stat115_rnaseq.pdf) via Rob Patro

Basic principles of quantification

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



Transcriptome



Sequenced fragments

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

Units for Relative Abundance

TPM (Transcripts Per Million)

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

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

Reads coming from transcript i

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

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

From Rob Patro

Units for Relative Abundance

TPM (Transcripts Per Million)

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

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

Reads coming from transcript i

Length of transcript i

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

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

From Rob Patro

Units for Relative Abundance

TPM (Transcripts Per Million)

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

abundance of i
as fraction of all
measured transcripts

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

Reads coming from
transcript i

Length of transcript i

FPKM (Fragments Per Kilobase Per Million Mapped Reads)

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

From Rob Patro

Units for Relative Abundance

TPM (Transcripts Per Million)

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

abundance of i
as fraction of all
measured transcripts

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

Reads coming from
transcript i

Length of transcript i

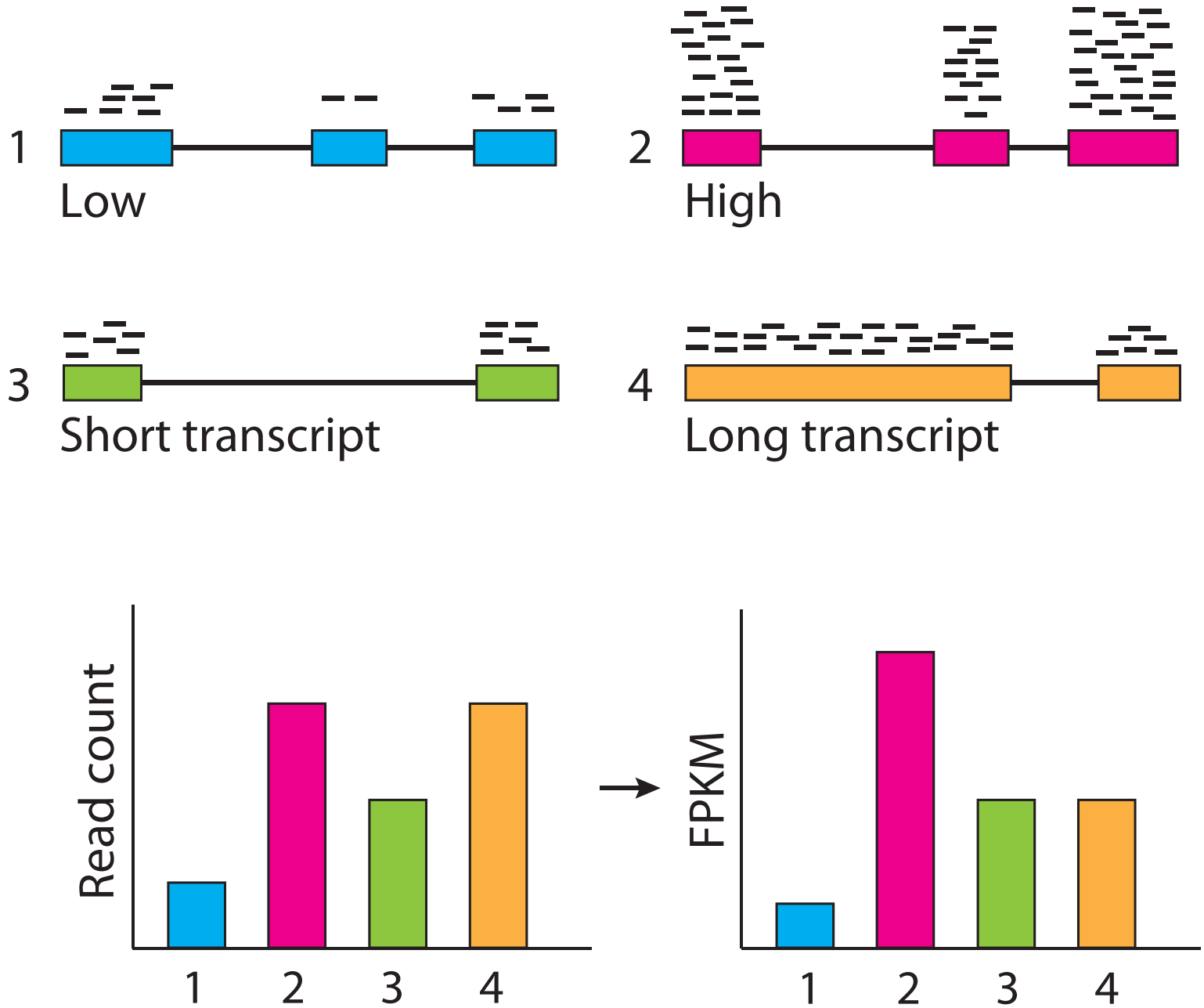
FPKM (Fragments Per Kilobase Per Million Mapped Reads)

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

Total number of mapped reads

From Rob Patro

Calculating expression of genes and transcripts



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

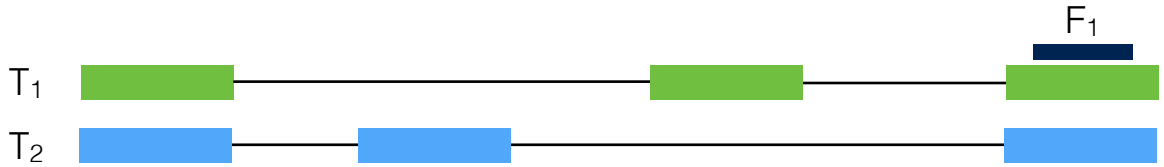
The difficulty is in estimating X_i

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

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

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

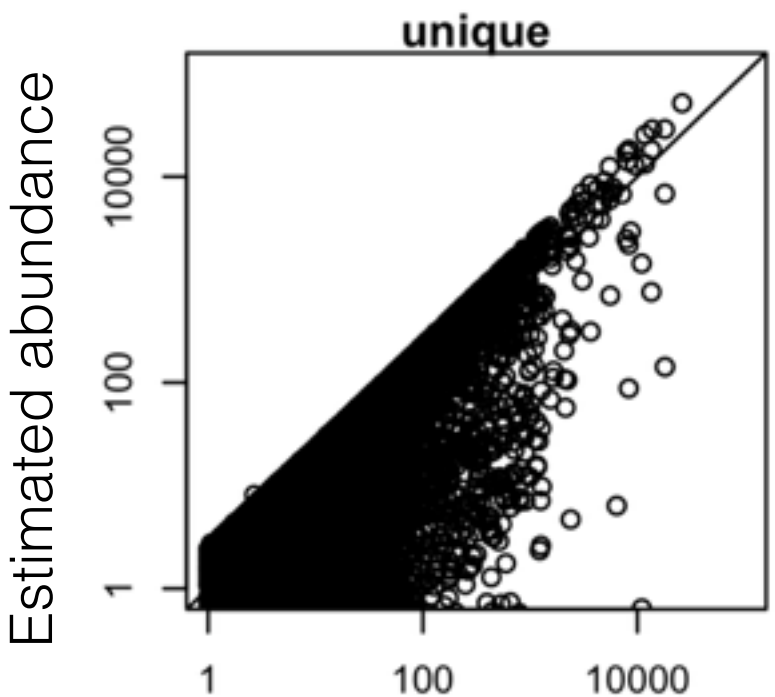
Even without isoforms, such problems arise from similar / related gene families.



From Rob Patro

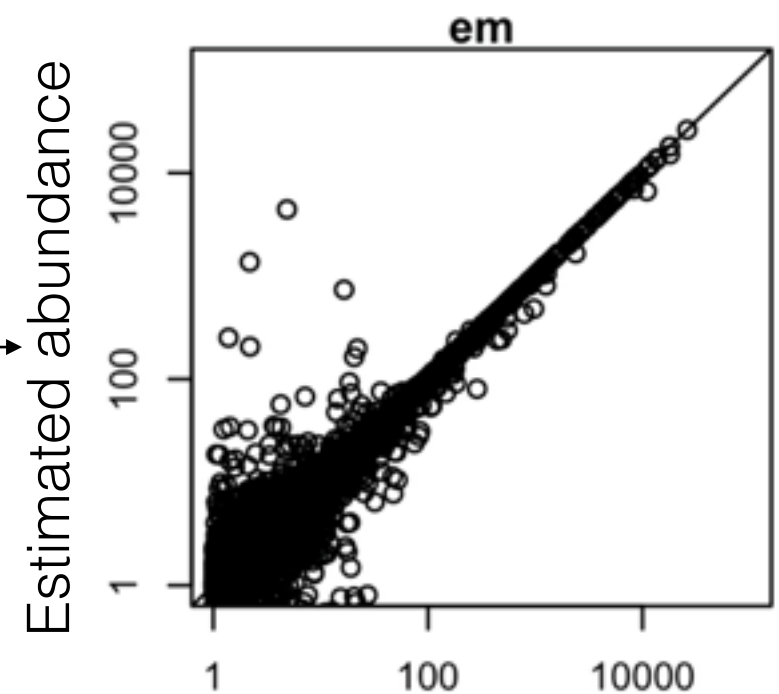
A simple (and wrong) approach

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



True abundance

The "right" approach achieves this



True abundance

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

From Rob Patro

Considering Genes Doesn't Solve the Problem

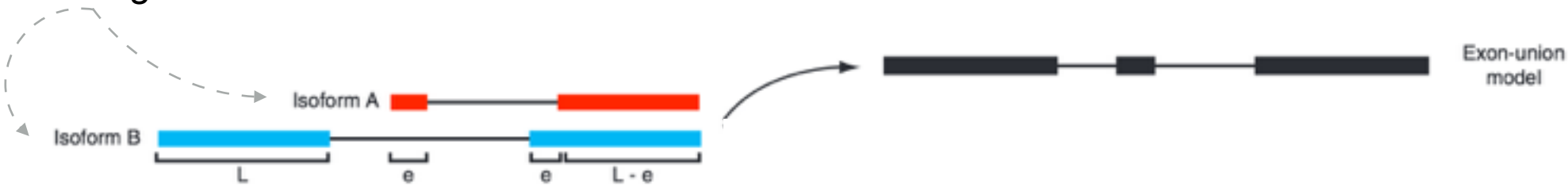
But, what if we only care about expression at the level of "genes" (i.e. we don't care about how much of each isoform there is, only how much of each overall gene)?

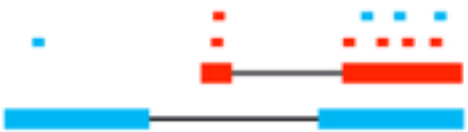
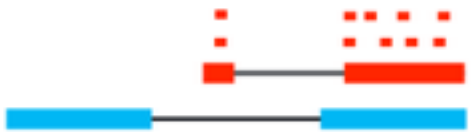
People commonly consider 2 models for this problem (union & intersection); Trapnell et al. show that both models can lead to completely incorrect results → **accurate gene-level estimates require isoform-level abundance computations**!

From Rob Patro

Resolving multi-mapping is fundamental to quantification

Isoform A is half
as long as isoform B



Condition 1	Condition 2	union-model fold-change		true fold-change
		$\log_2\left(\frac{10}{10}\right) = 0$	0 < 0.32	$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
		$\log_2\left(\frac{6}{8}\right) = -0.41$	-0.41 < 0.58	$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
		$\log_2\left(\frac{5}{10}\right) = -1$	-1 < 0	$\log_2\left(\frac{\frac{5}{L}}{\frac{10}{2L}}\right) = 0$

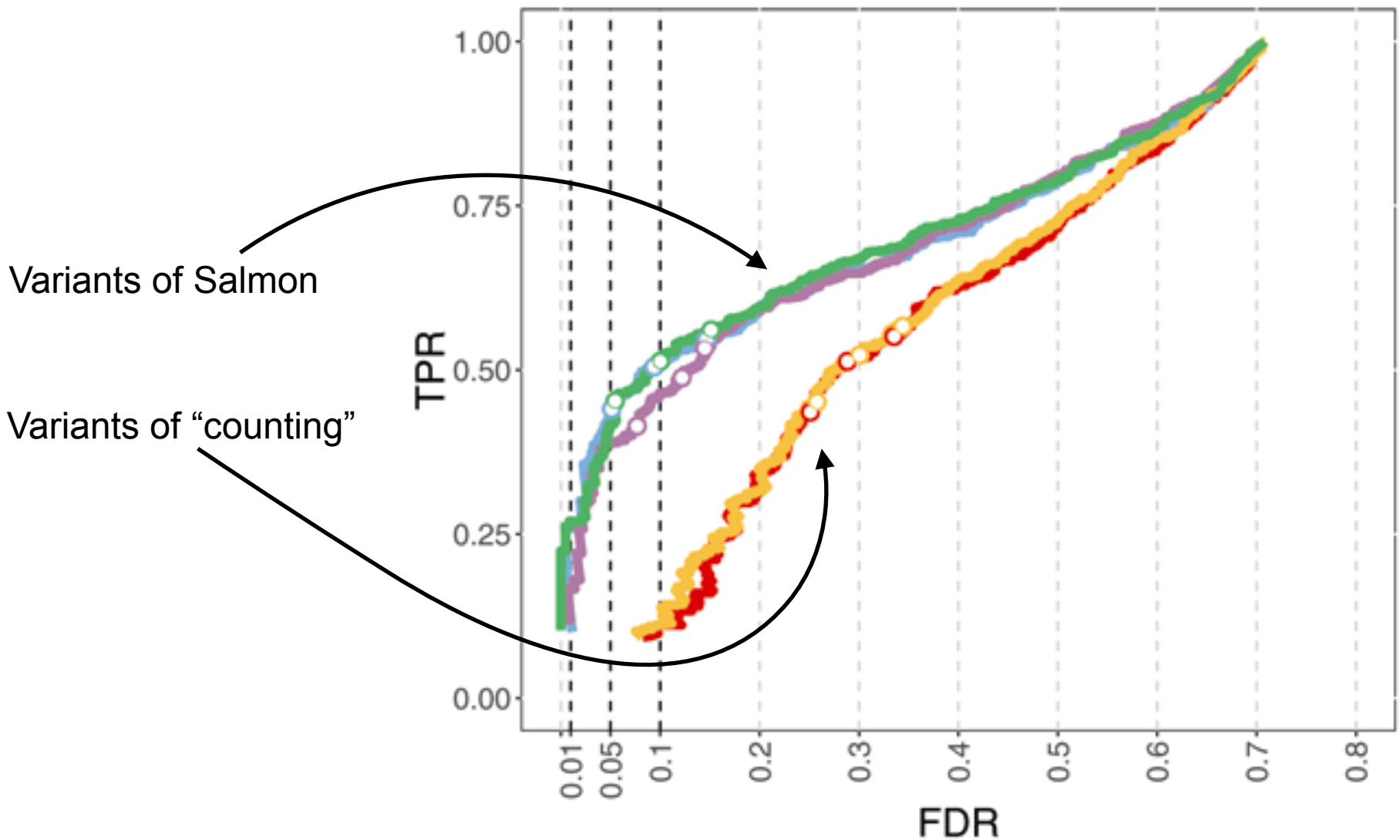
Key point : The length of the *actual molecule* from which the fragments derive is crucially important to obtaining accurate abundance estimates.

Adapted from: Trapnell, Cole, et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq." Nature biotechnology 31.1 (2013): 46-53.

From Rob Patro

Resolving multi-mapping is fundamental to quantification

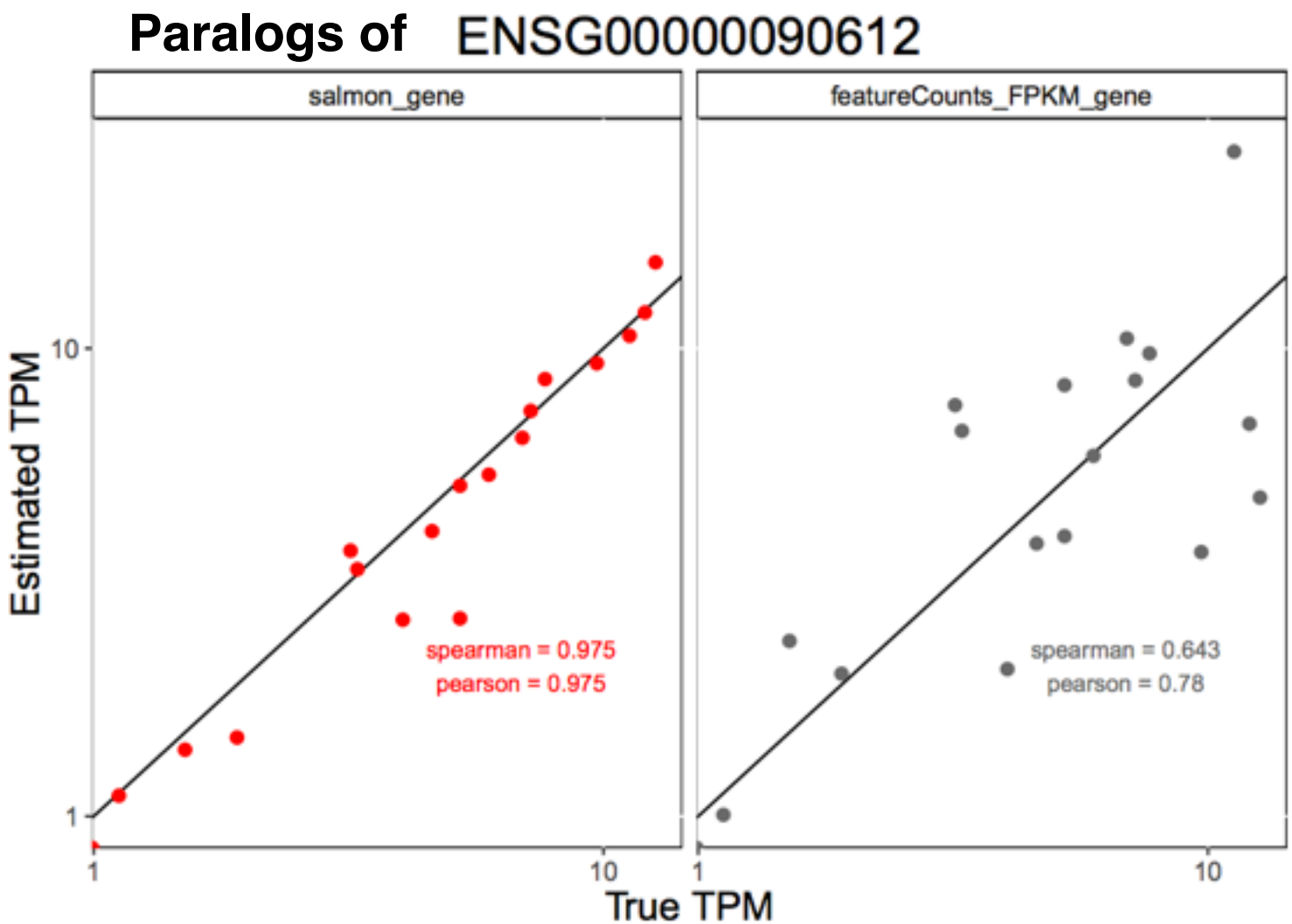
These errors can affect DGE calls



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

Resolving multi-mapping is fundamental to quantification

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



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

So how do we estimate abundance "correctly"?

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

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

$$\arg\max_{\boldsymbol{\rho} \in \mathbf{P}} \mathcal{L}(\boldsymbol{\rho}; x_1, \ldots, x_n)$$

abundances — parameters of a generative model

observations — alignments of reads to transcripts

From Rob Patro

So how do we estimate abundance “correctly”?

Finding the maximum likelihood estimates first requires defining the likelihood:

We’ll define it in terms of parameters alpha

$$\alpha_t \quad := \quad \mathbb{P}(f \in t) \quad = \quad \frac{\rho_t \tilde{l}_t}{\sum_{r \in T} \rho_r \tilde{l}_r}$$

which are relatable, directly, to the rhos

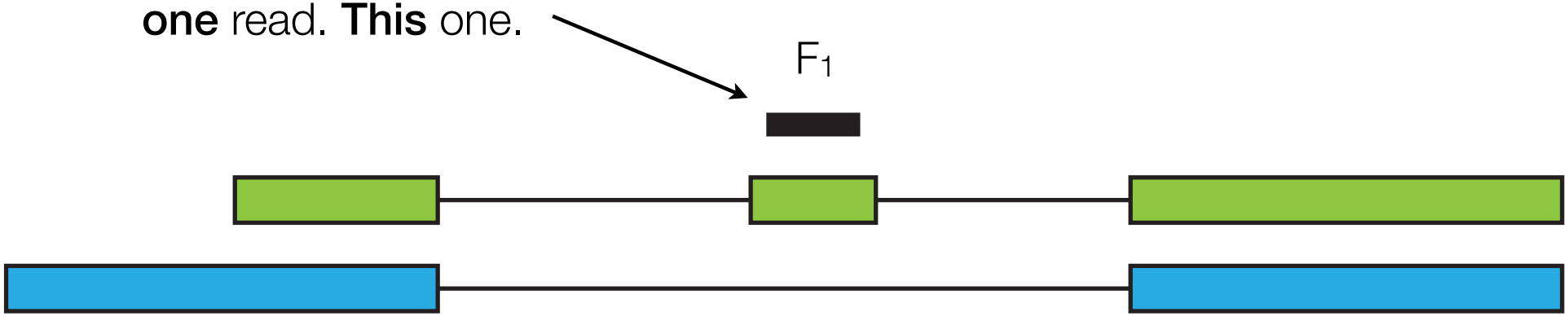
$$\rho_t \quad = \quad \frac{\frac{\alpha_t}{\tilde{l}_t}}{\sum_{r \in T} \frac{\alpha_r}{\tilde{l}_r}}$$

*Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

From Rob Patro

Defining the likelihood function

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



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

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

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

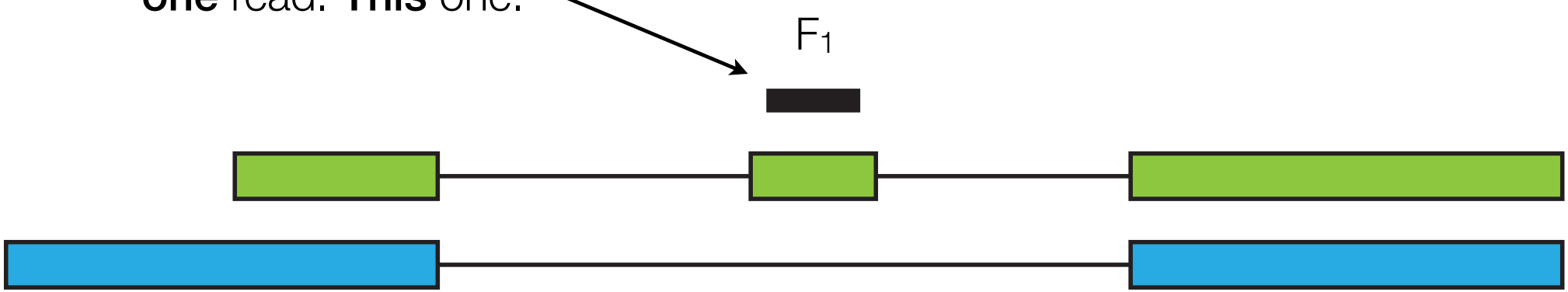
Then, we need to pick this read from that transcript over all the others.

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

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

Defining the likelihood function

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



not normalized by length

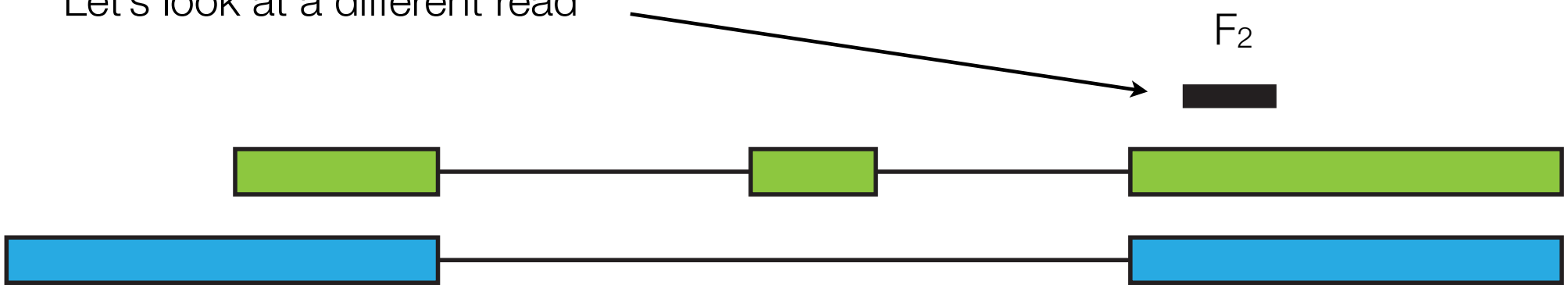
So given a relative abundance for the green transcript, which we'll call α_{green} we can calculate the probability of getting F_1.

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

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

Defining the likelihood function

Let's look at a different read



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

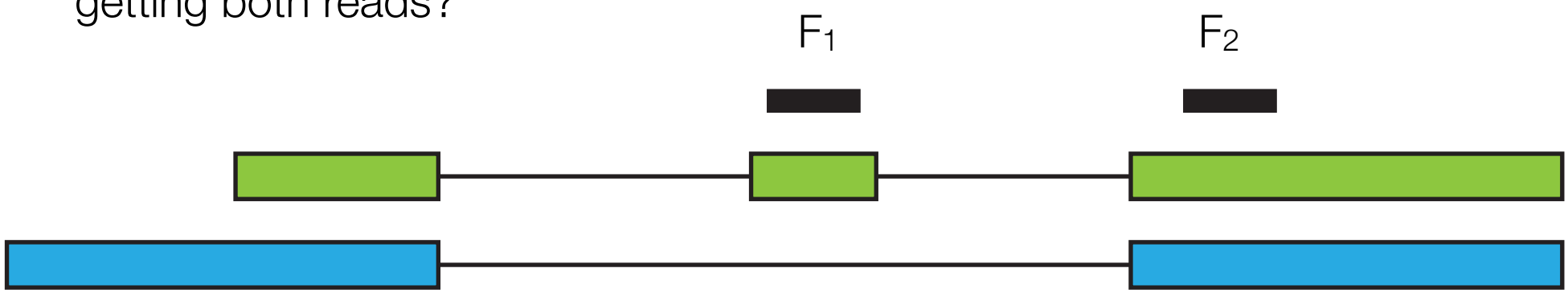
$$\Pr(F_2 \in T_{\text{green}} \text{ or } F_2 \in T_{\text{blue}}) = \Pr(F_2 \mid \alpha) = \frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}$$

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

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

Defining the likelihood function

What are the chances of getting both reads?



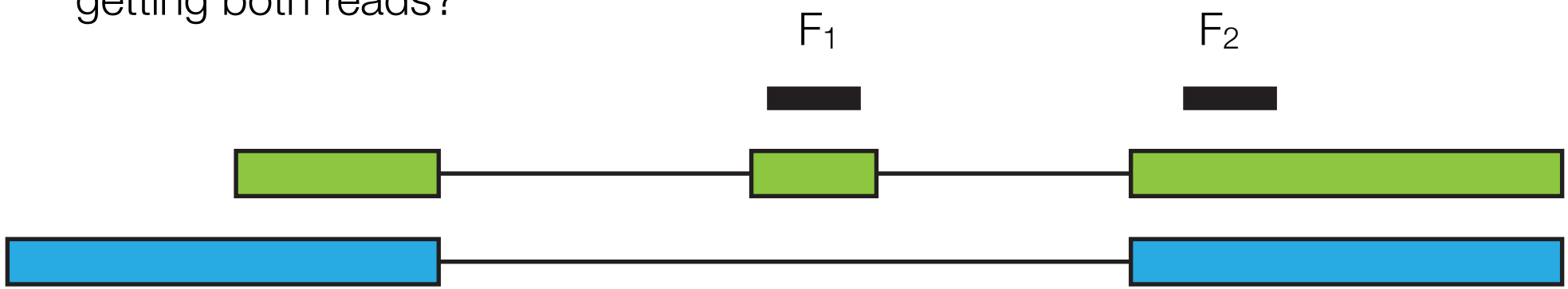
To get both F_1 and F_2, we just need to multiply the two probabilities!

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

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

Defining the likelihood function

What are the chances of getting both reads?



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

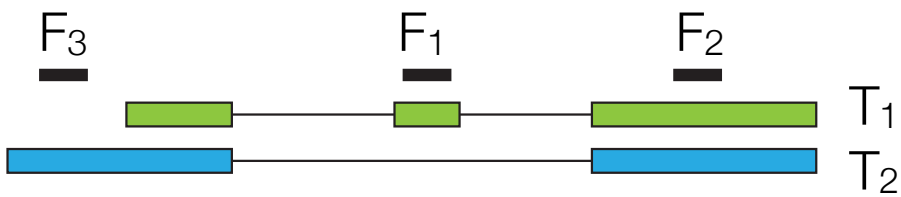
$$\mathcal{L}(\alpha; F) = \mathcal{L}(\alpha) = \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}}\right) \cdot \left(\frac{\alpha_{\text{green}}}{\ell_{\text{green}}} + \frac{\alpha_{\text{blue}}}{\ell_{\text{blue}}}\right)$$

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

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

Defining the likelihood function

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



Compatibility table

	F_1	F_2	F_3
T_1	✓	✓	✗
T_2	✗	✓	✓

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

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

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

Likelihood Function

With the simplest generative model, we get a likelihood function that looks like this:

$$\mathcal{L}(\alpha) = \prod_{t \in T} \left(\frac{\alpha_t}{\tilde{l}_t} \right)^{x_t}$$

$$\propto \prod_{t \in T} \alpha_t^{x_t},$$

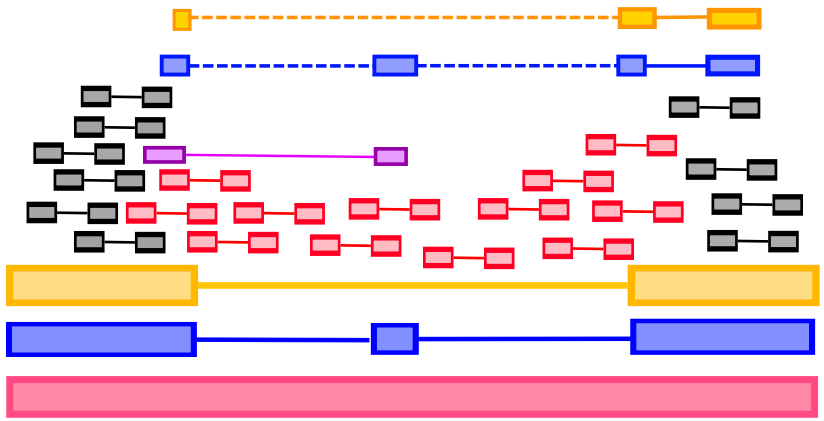
fragments compatible w/ transcript t

*Pachter, Lior. "Models for transcript quantification from RNA-Seq." arXiv preprint arXiv:1104.3889 (2011).

From Rob Patro

Assigning reads to isoforms

Problem: infer which transcript each fragment came from



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

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

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

Finding the MLE

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

Essentially:

While not converged:

E-step Assign fragments to transcripts (probabilistically) using current estimates of transcript abundance.

M-step Re-estimate transcript abundance using probabilistic fragment assignments.

From Rob Patro

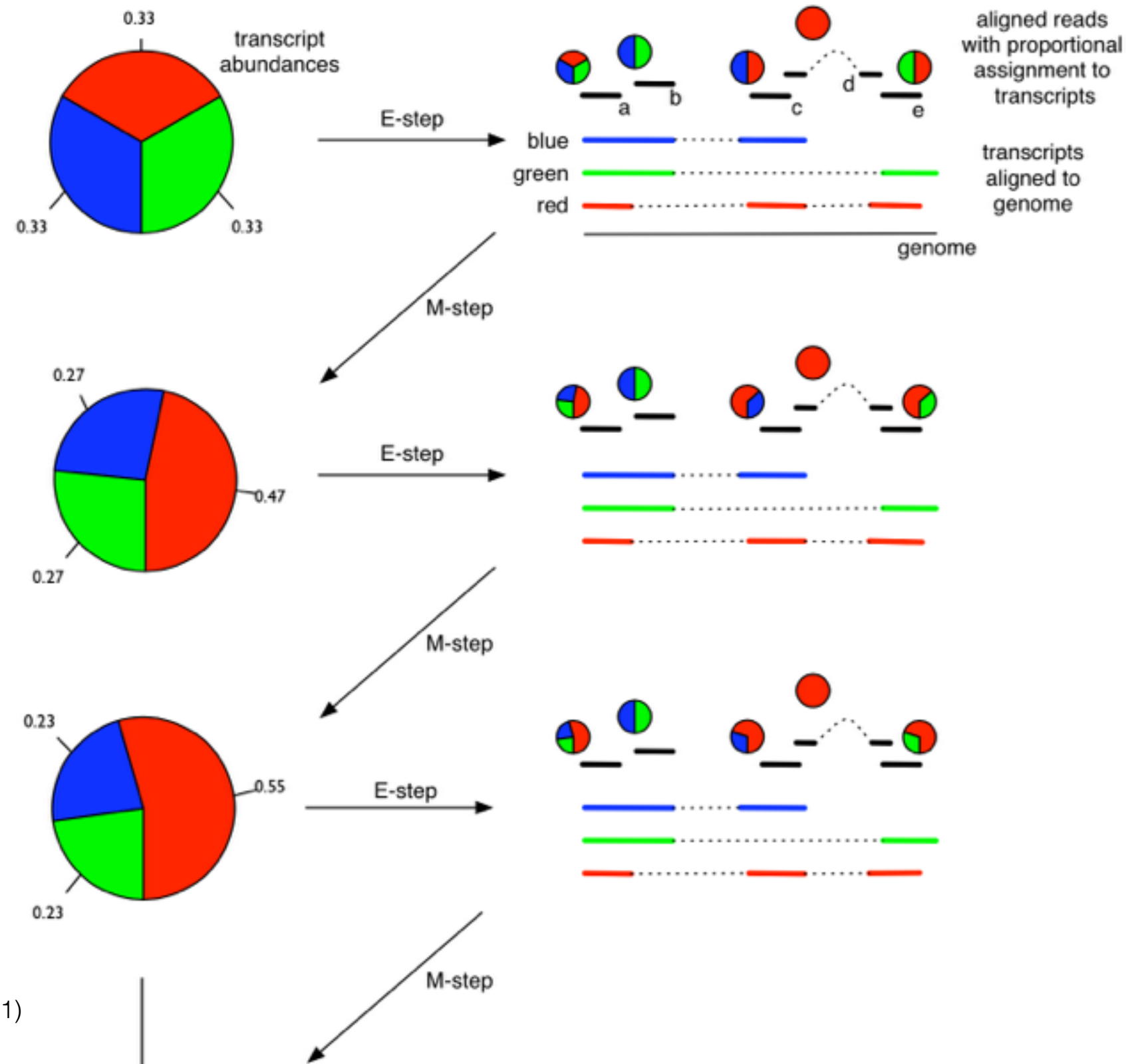


Figure from Pachter (2011)
via Rob Patro

Performance of RSEM (one of the first methods to use EM for RNA-seq quant.)

MPE = median percentage error

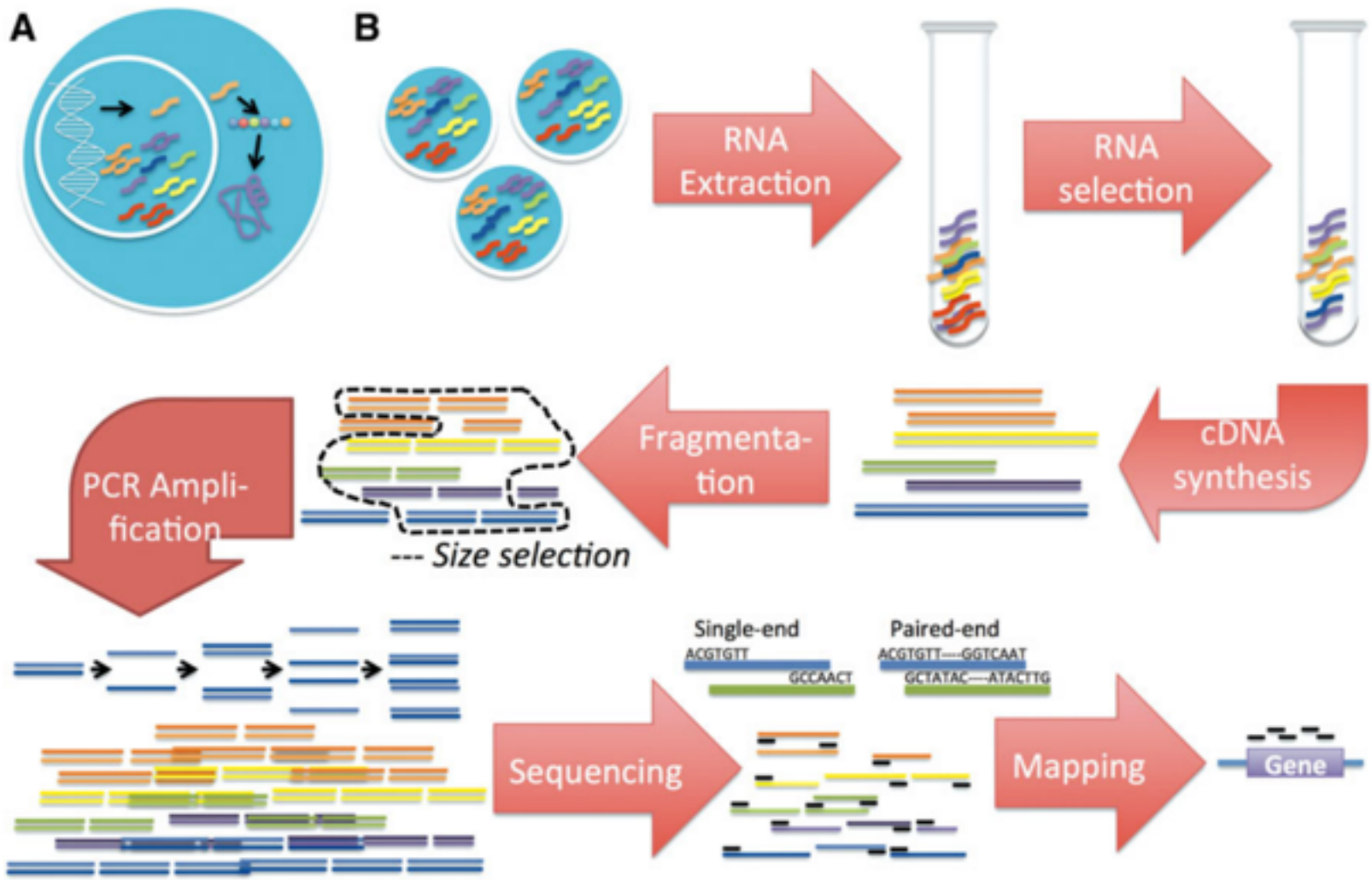
EF = error fraction (% error > 5)

Table 2. Error of the `unique`, `rescue` and `em` estimated gene expression levels with respect to sample expression values from simulations of mouse and maize RNA-Seq data

			Sample gene expression in NPM (ν) or TPM (τ)					
			$[1, 10)$	$[10, 10^2)$	$[10^2, 10^3)$	$[10^3, 10^4)$	$[10^4, 10^5)$	All
Simulation of mouse RNA-Seq data								
		N	6279	4025	886	111	15	11316
	MPE	`unique`	29.6	29.2	30.9	32.8	32.1	29.6
		`rescue`	12.6	6.8	6.1	5.9	5.8	8.2
τ		`em`	**2.6**	**1.0**	**0.4**	**0.3**	**0.4**	**1.5**
	EF	`unique`	93.7	93.9	95.6	99.1	100.0	94.0
		`rescue`	79.5	73.2	72.2	69.4	66.7	76.6
		`em`	27.8	6.2	1.1	0.0	0.0	17.7
Simulation of maize RNA-Seq data								
		N	9210	4931	1040	113	12	15306
	MPE	`unique`	86.1	84.2	85.2	80.5	96.3	85.5
		`rescue`	21.3	11.8	8.9	8.5	7.7	16.0
τ		`em`	**4.6**	**1.5**	**0.6**	**0.5**	0.3	**2.8**
	EF	`unique`	97.2	96.7	97.1	98.2	100.0	97.0
		`rescue`	89.4	88.3	85.8	82.3	91.7	88.8
		`em`	47.5	18.8	6.1	4.4	16.7	35.1

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

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



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

Prakash Celine and Von Haeseler Arndt. Journal of Computational Biology. September 2016, ahead of print. doi:10.1089/cmb.2016.0096.

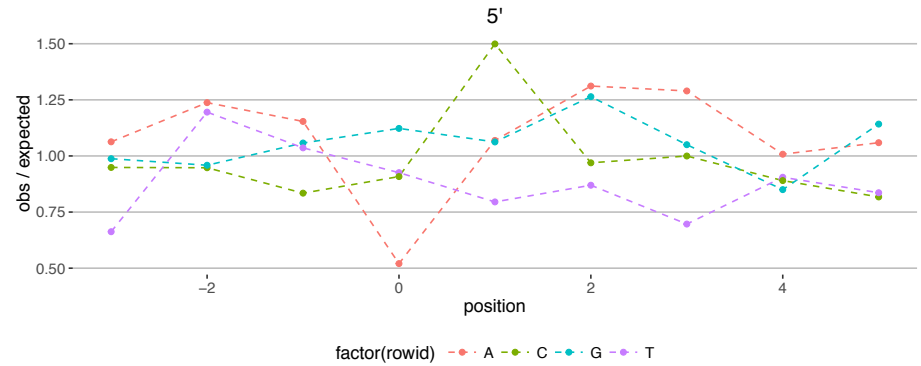
Biases abound in RNA-seq data

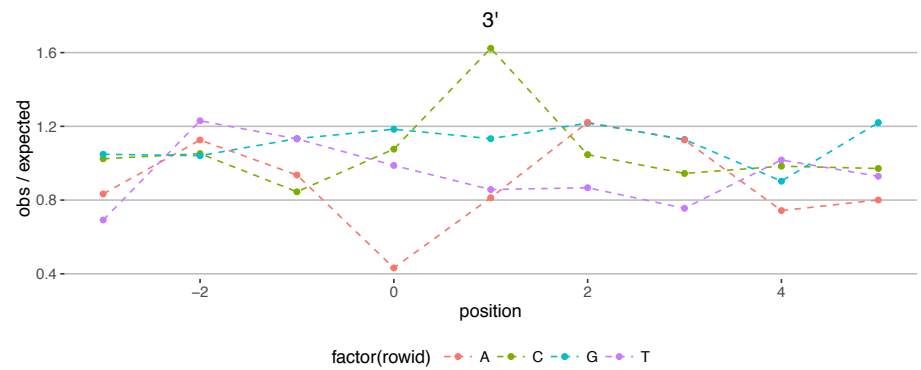
Biases in prep & sequencing can have a significant effect on the fragments we see.

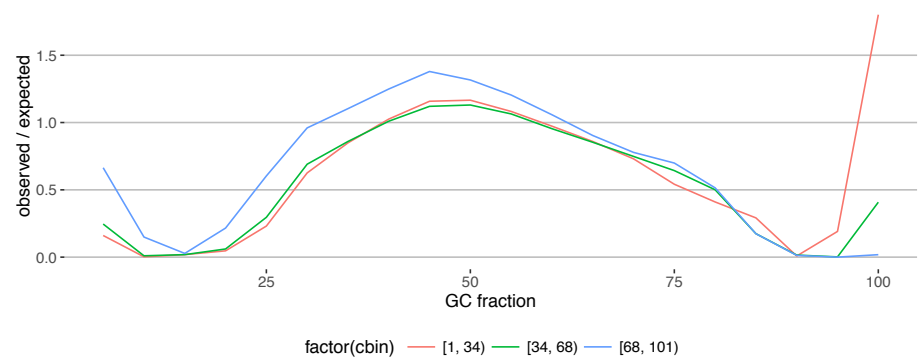
Fragment gc-bias[1]—
The GC-content of the fragment affects the likelihood of sequencing

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

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







1:Love, Michael I., John B. Hogenesch, and Rafael A. Irizarry. "Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation." bioRxiv (2015): 025767.
2:Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

From Rob Patro

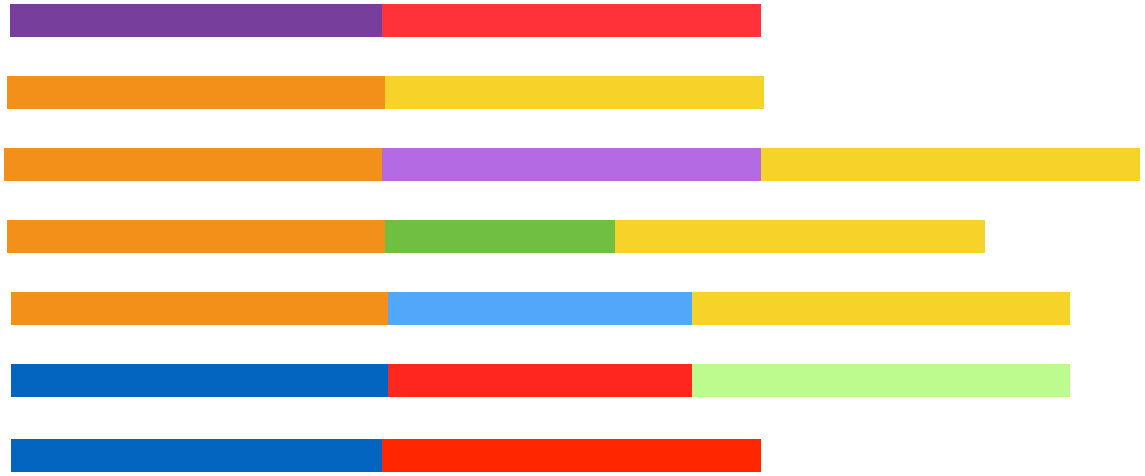
Fast abundance estimation

- Standard RNA-Seq pipeline:
  - align reads to reference genome/transcriptome
  - process BAM file, produce transcript abundance estimates
- New algorithms greatly reduce runtime
  - Key idea: full alignment of read-to-reference is unnecessary
- All we care about is if a read is *compatible* with a certain transcript
- We'll discuss:
  - rapmap (Rob Patro)
  - kallisto (Pall Melsted, Lior Pachter, Nicolas Bray)

Mapping reads to a Transcriptome

Consider the following scenario:

Transcripts



Read

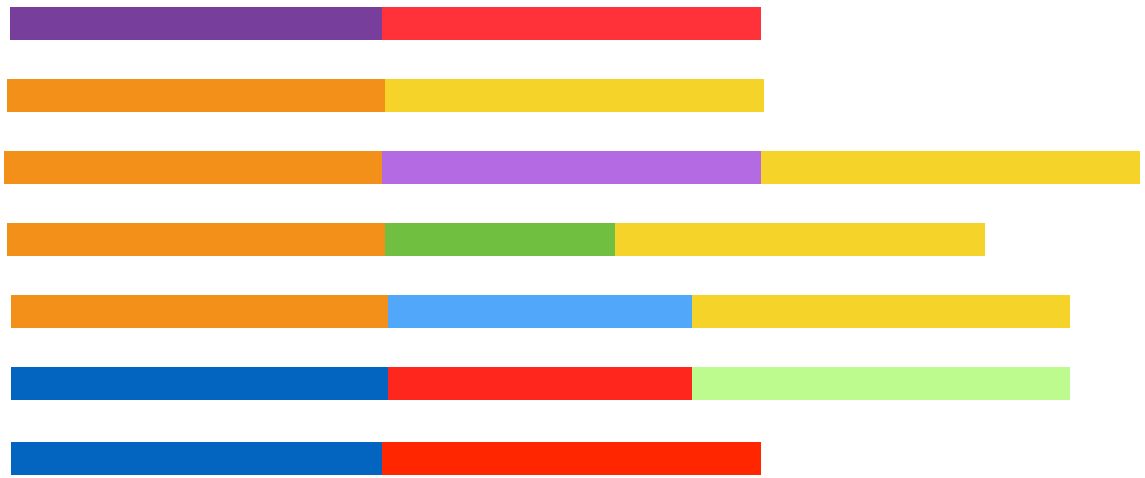
From Rob Patro

Mapping reads to a Transcriptome

Consider the following scenario:

Say that colors represent exonic sequence.
Intuitively, from where does the read originate?

Transcripts



Read

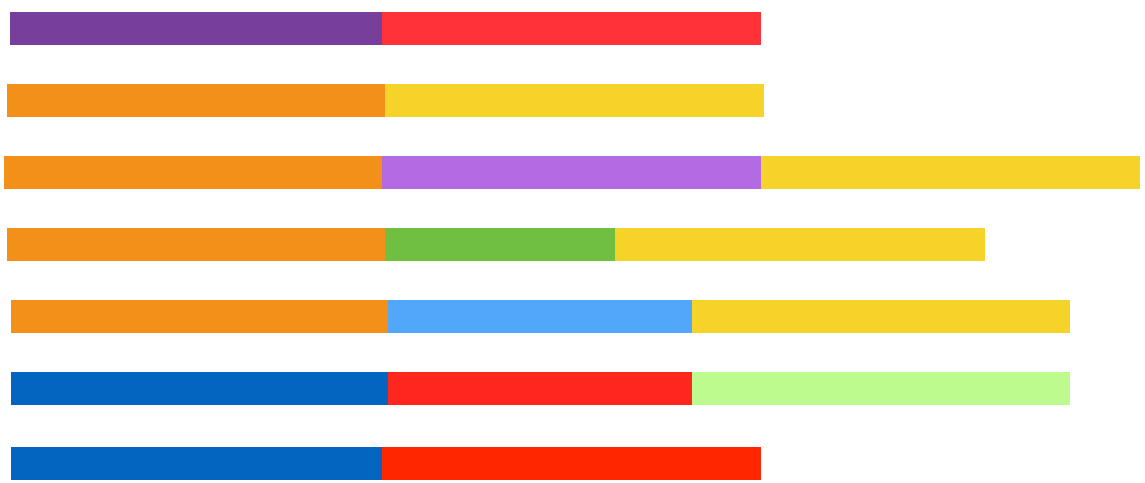
From Rob Patro

Mapping reads to a Transcriptome

Consider the following scenario:

Say that colors represent exonic sequence.
Intuitively, from where does the read originate?
What about *this* read?

Transcripts

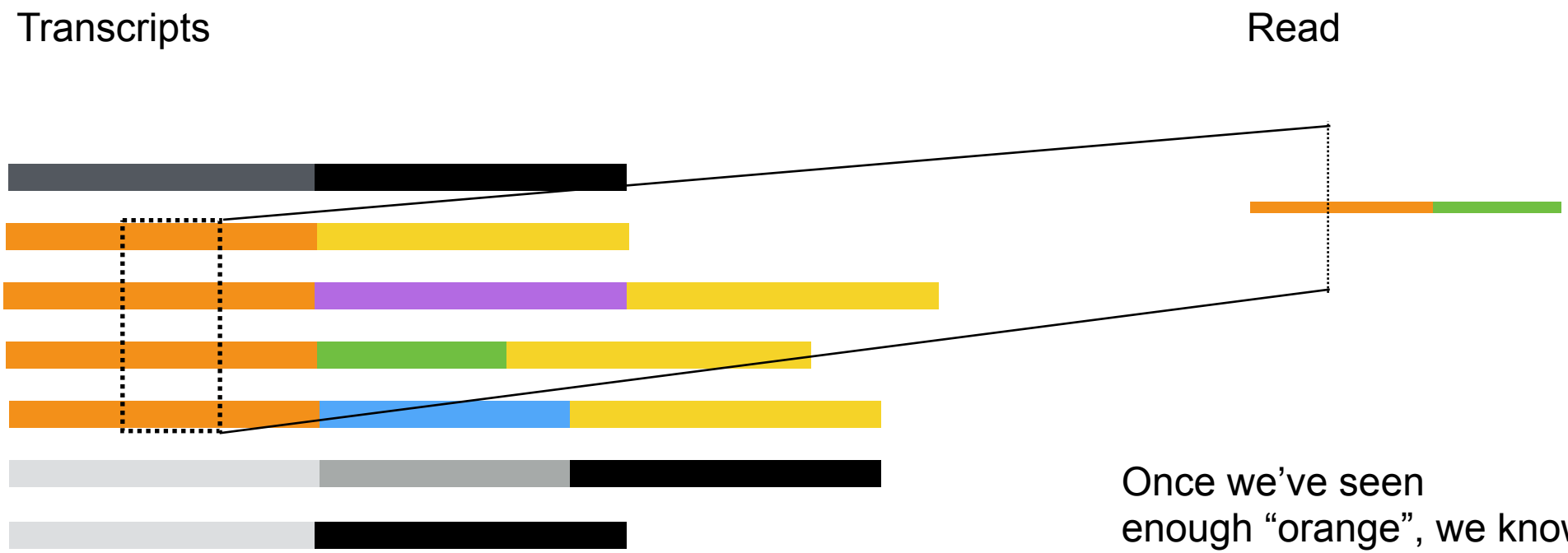


Read

From Rob Patro

Mapping reads to a Transcriptome

Consider the following scenario:

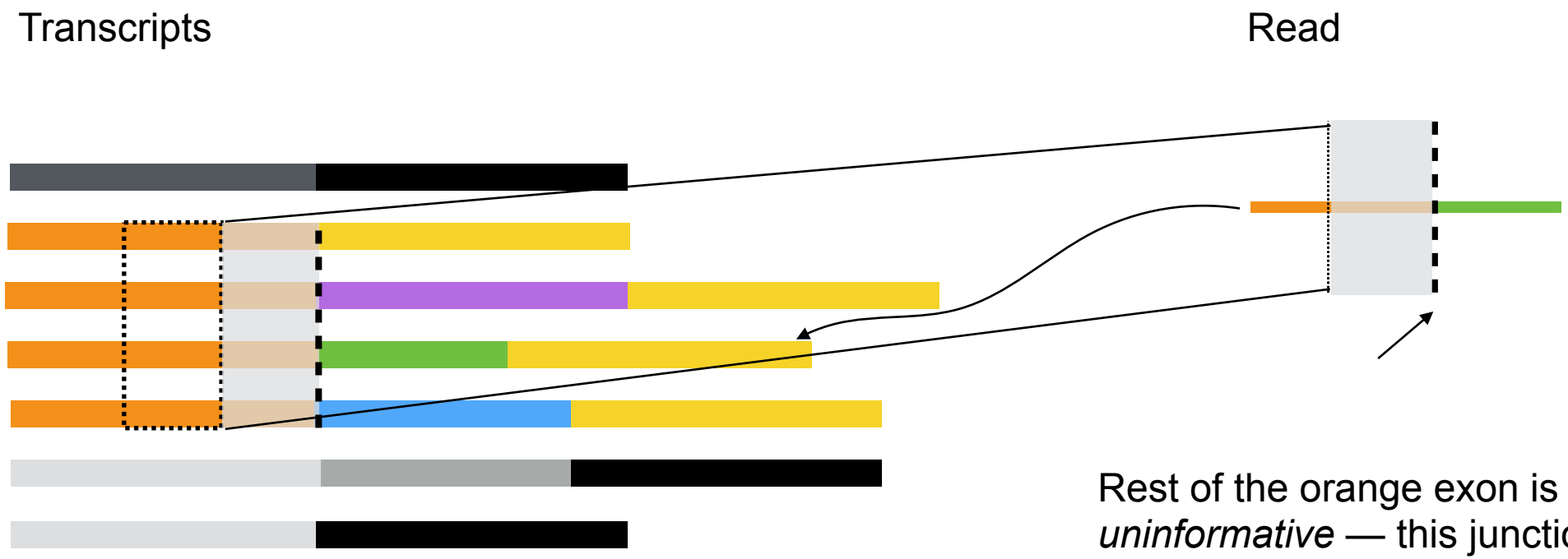


Once we've seen enough "orange", we know the read must map to txps with this exon; but which one(s)?

From Rob Patro

Mapping reads to a Transcriptome

Consider the following scenario:



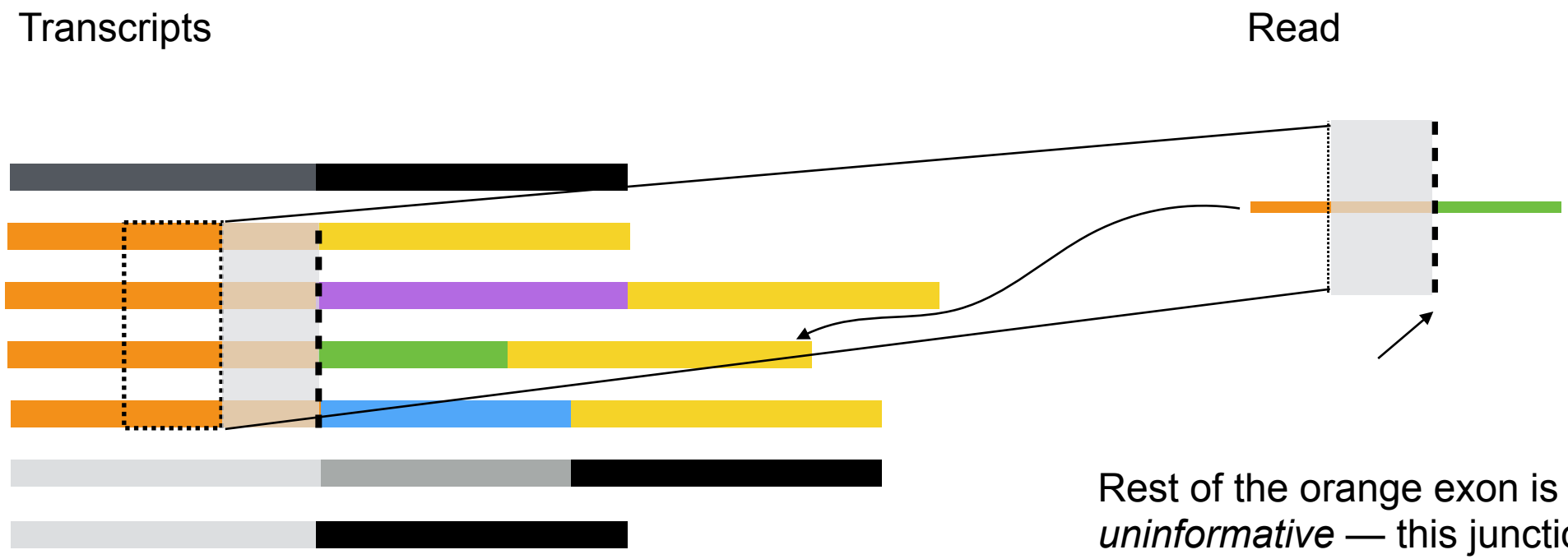
Rest of the orange exon is *uninformative* — this junction is the *next informative position*.

From Rob Patro

Mapping reads to a Transcriptome

Consider the following scenario:

Is there some ***general/formal*** way to always find the next informative position (NIP) when mapping a read?



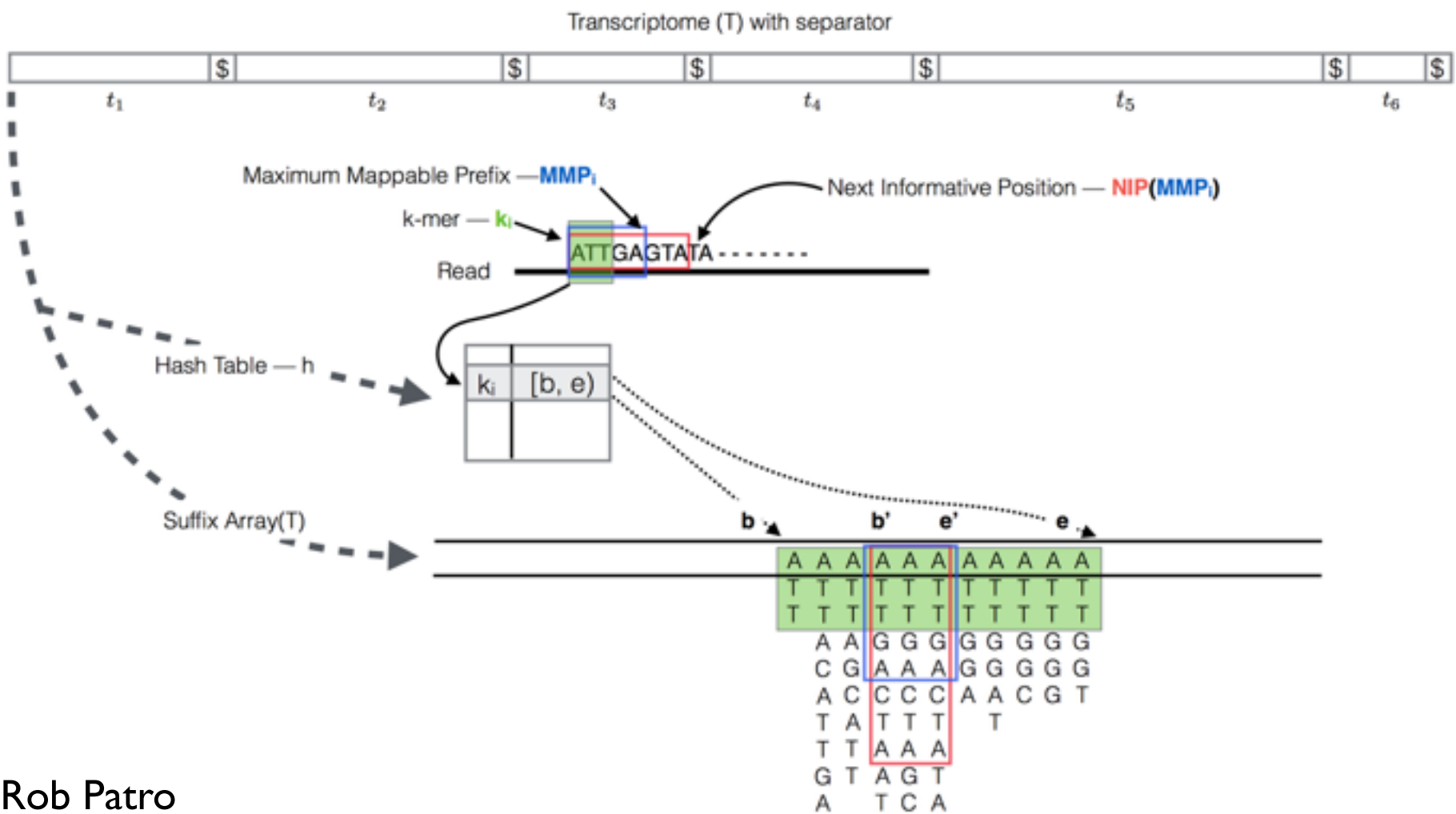
Rest of the orange exon is *uninformative* — this junction is the *next informative position*.

From Rob Patro

An algorithm for quasi-mapping

Move from left to right along read, until we find a k-mer with non-empty SA interval.

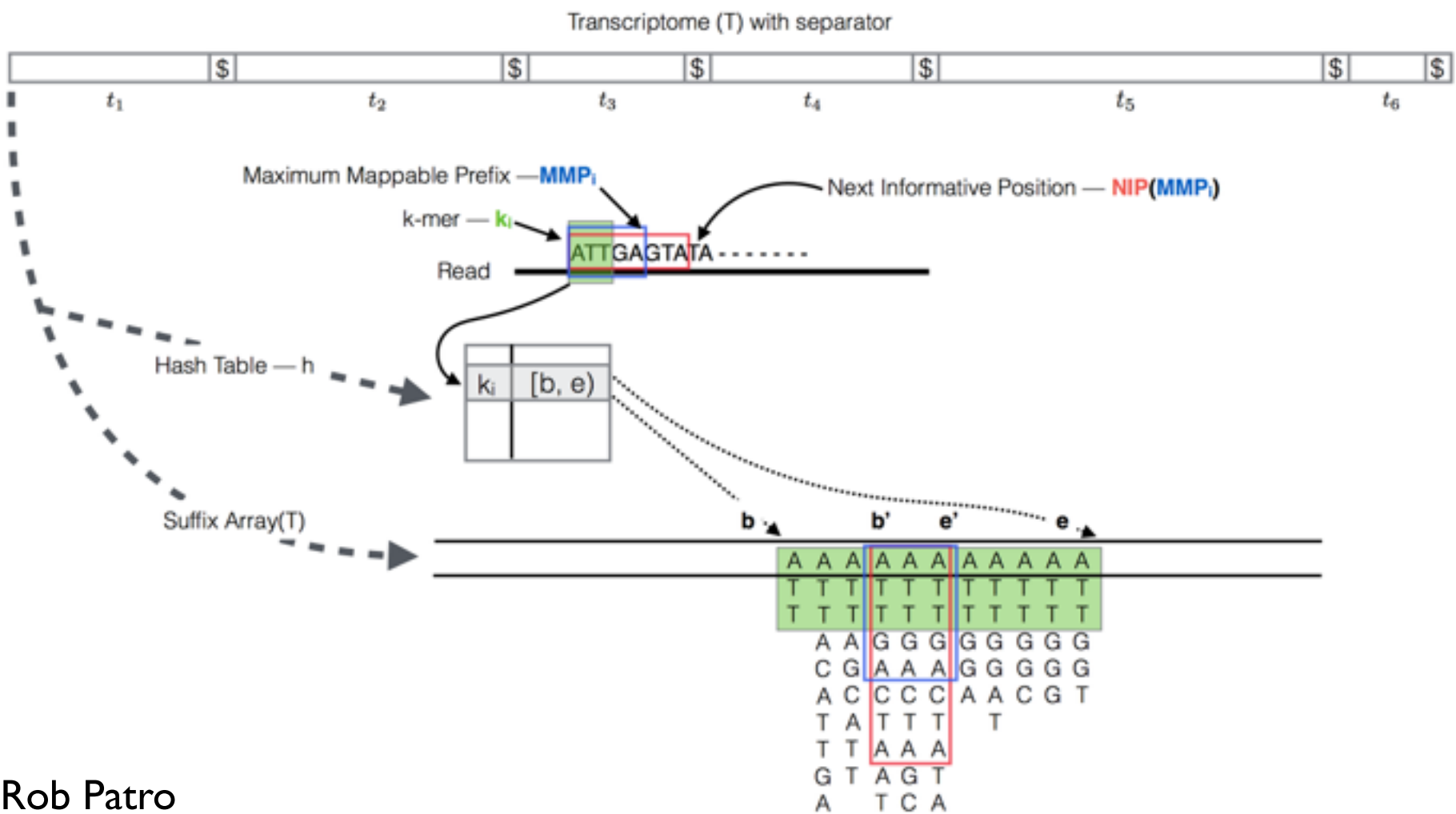
Compute Maximum Mappable Prefix (**MMP**) starting with this k-mer — logarithmic in k-mers SA interval



From Rob Patro

An algorithm for quasi-mapping

Compute **NIP** of this **MMP** — (fast) linear in read length

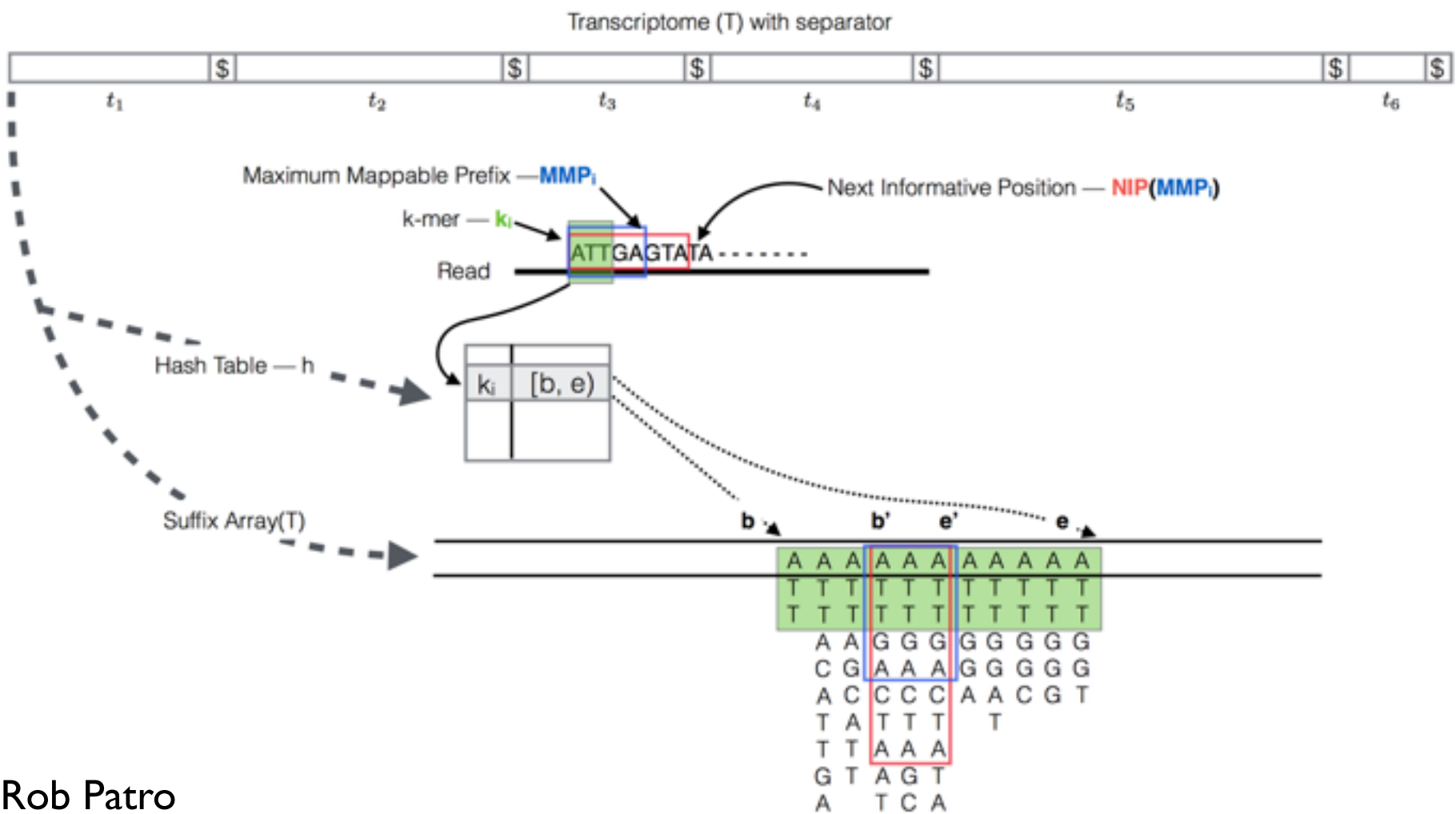


From Rob Patro

An algorithm for quasi-mapping

Compute **NIP** of this **MMP** — (fast) linear in read length

intuitively: **NIP** jumps you to the next exon boundary overlapping the read (need not be an actual exon boundary)



From Rob Patro

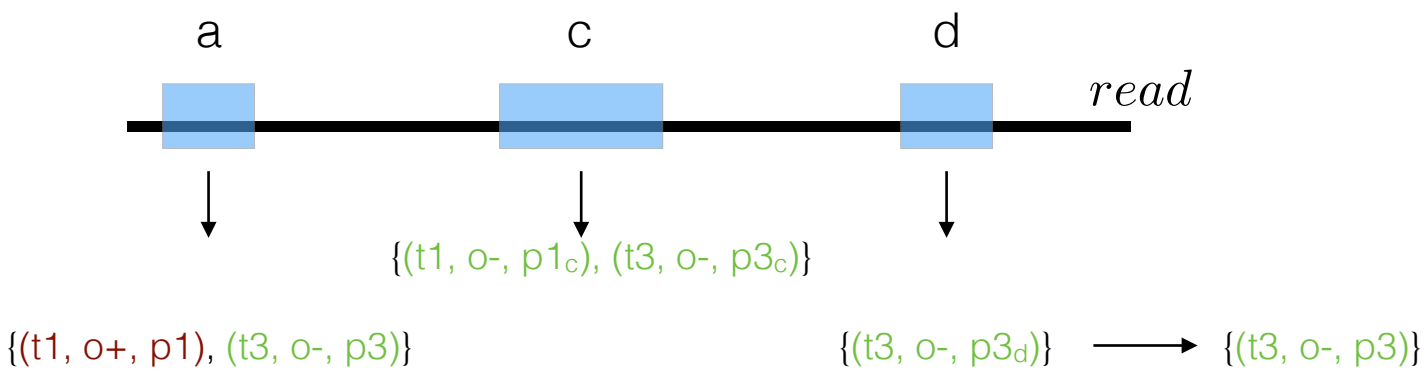
An algorithm for quasi-mapping

Produces a set of disjoint *hits* over each query (read).

A *hit* is a tuple — (query offset, orientation, length, SA-interval)

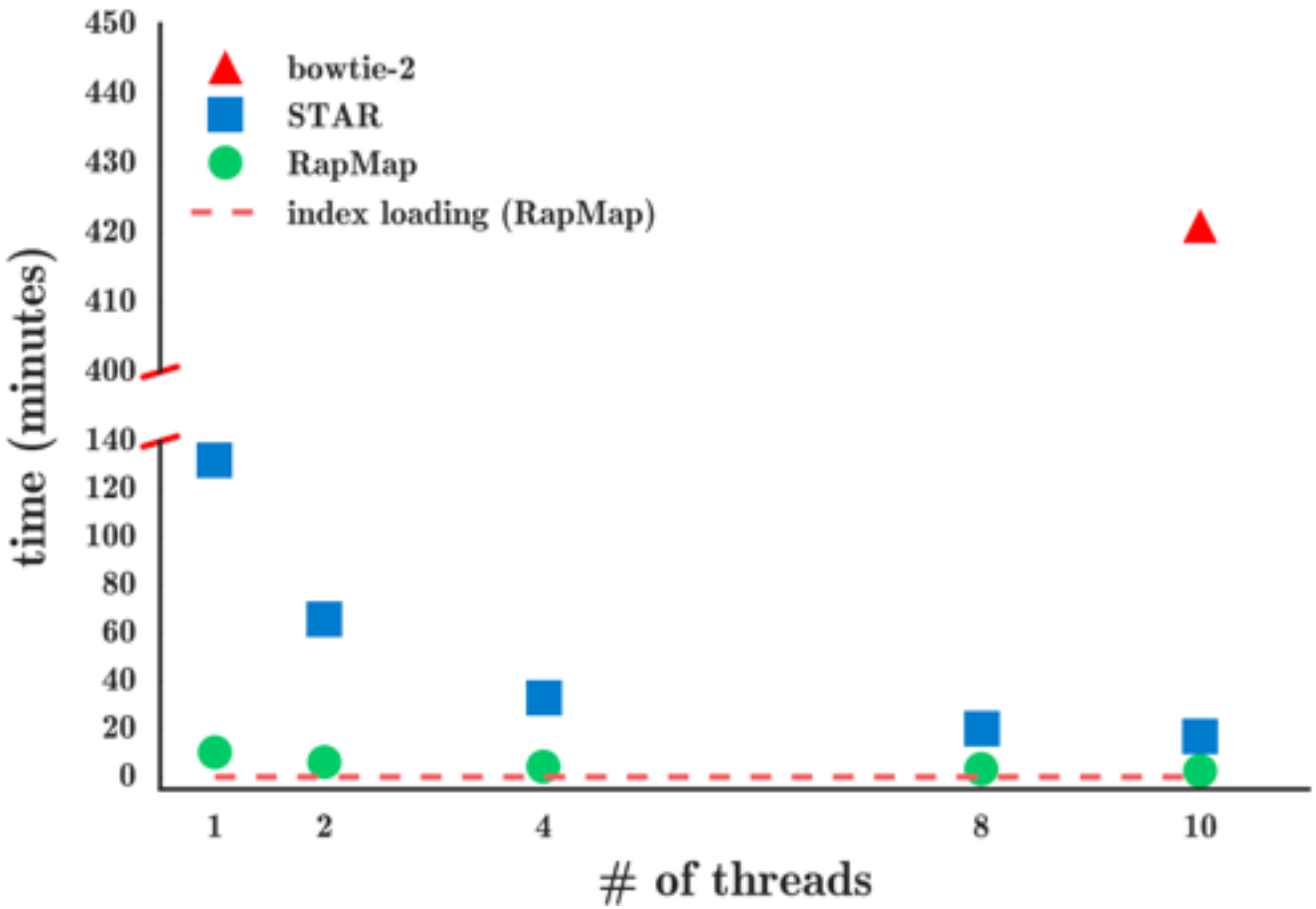
Mappings are determined by a *consensus* mechanism over hits:

- *default:* a read maps to a transcript if that transcript appears in **every hit for that read**.
- other (stricter or looser) mechanisms are trivial to enforce (e.g. co-linearity of hits wrt read & reference).



From Rob Patro

Quasi-mapping is Fast

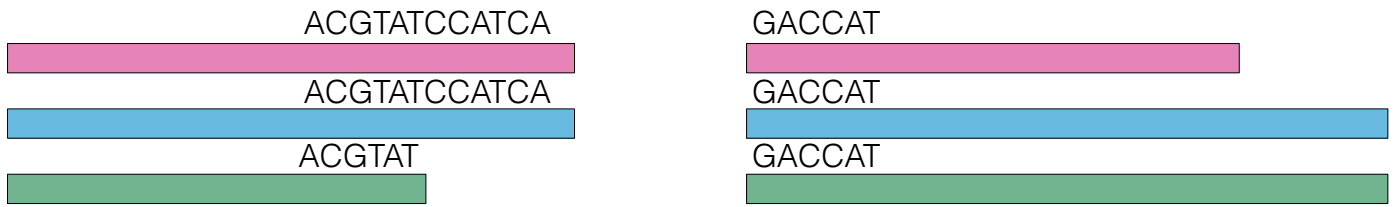


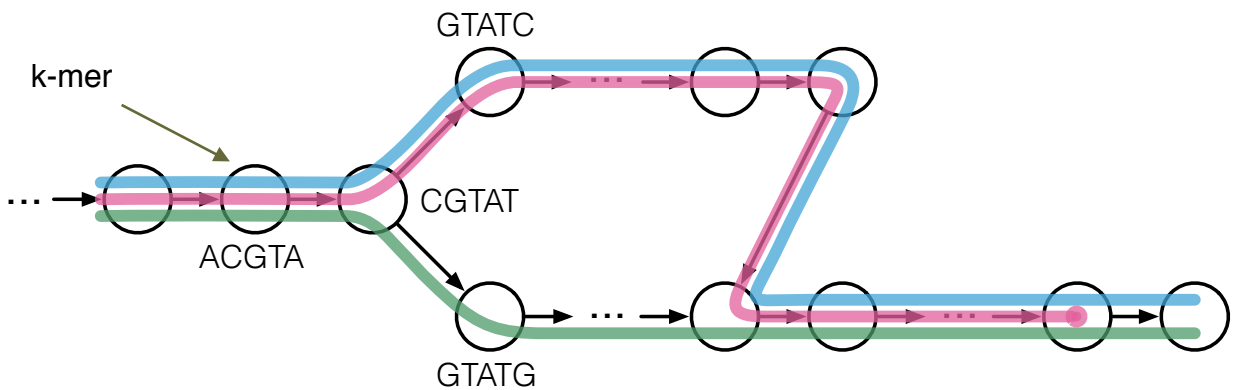
Can map **75 million paired-end reads** (76 bp) to the human transcriptome in matter of **minutes**; even with few threads.

Note: High degree of multi-mapping and inability to report top "**stratum**" means Bowtie2 is often reporting more than the "best" mapping (though it's commonly used in this context).

From Rob Patro

How kallisto computes pseudoaligments

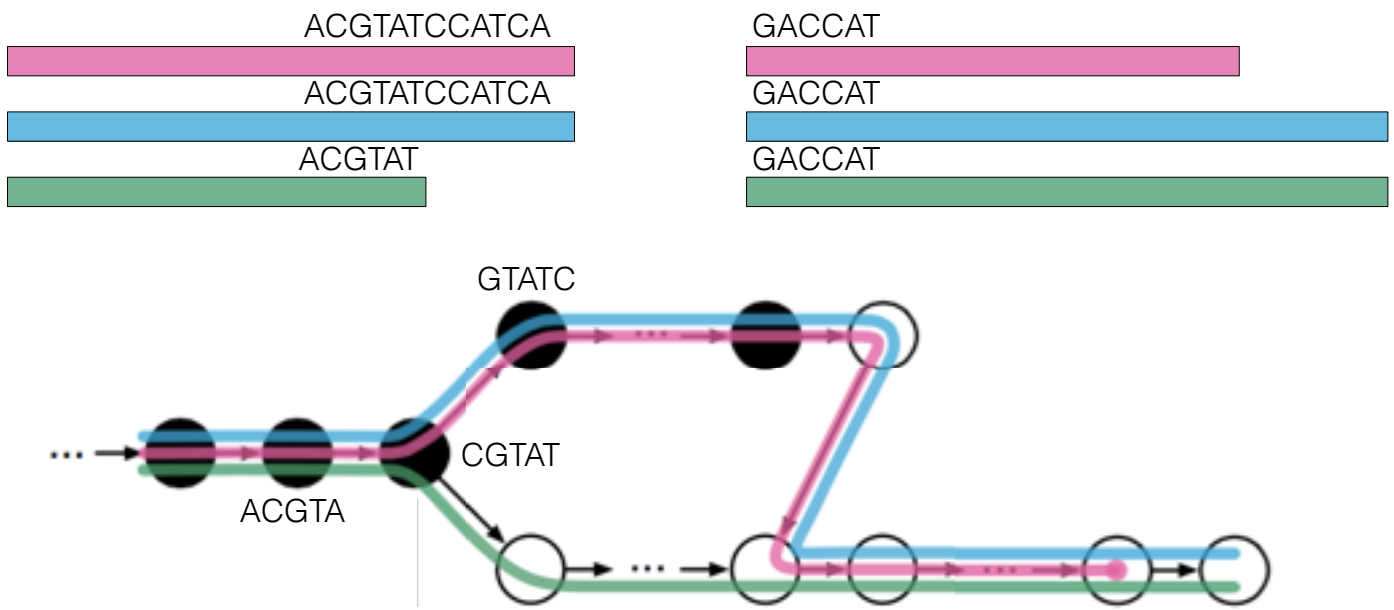




- Given our reference transcriptome, we first construct its *target de Bruijn Graph (T-DBG)*
- This encodes the transcript sequences but also provides information about how they overlap with each other
- Only has to be done *once* per transcriptome (and is fast)

Slides from Pall Melsted and Nicolas Bray

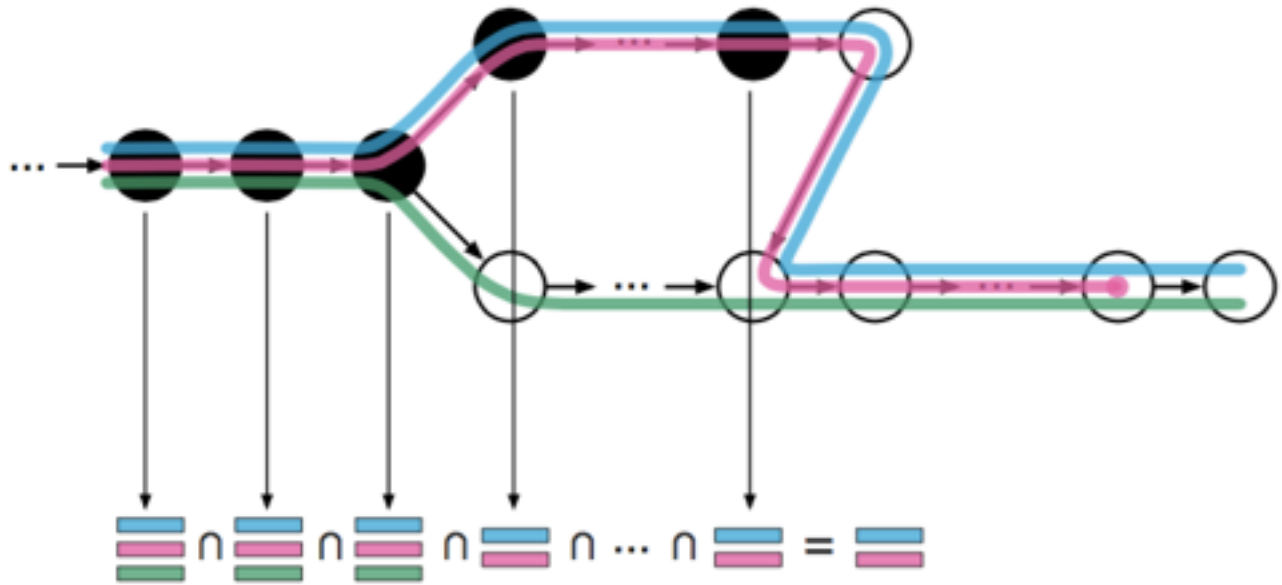
How kallisto computes pseudoaligments



- Given a read, finding its constitutive *k*-mers in the T-DBG gives you information about where the read could have come from
- This can be done *very* fast
- **But individual *k*-mers might be more ambiguous than the read as a whole**

Slides from Pall Melsted and Nicolas Bray

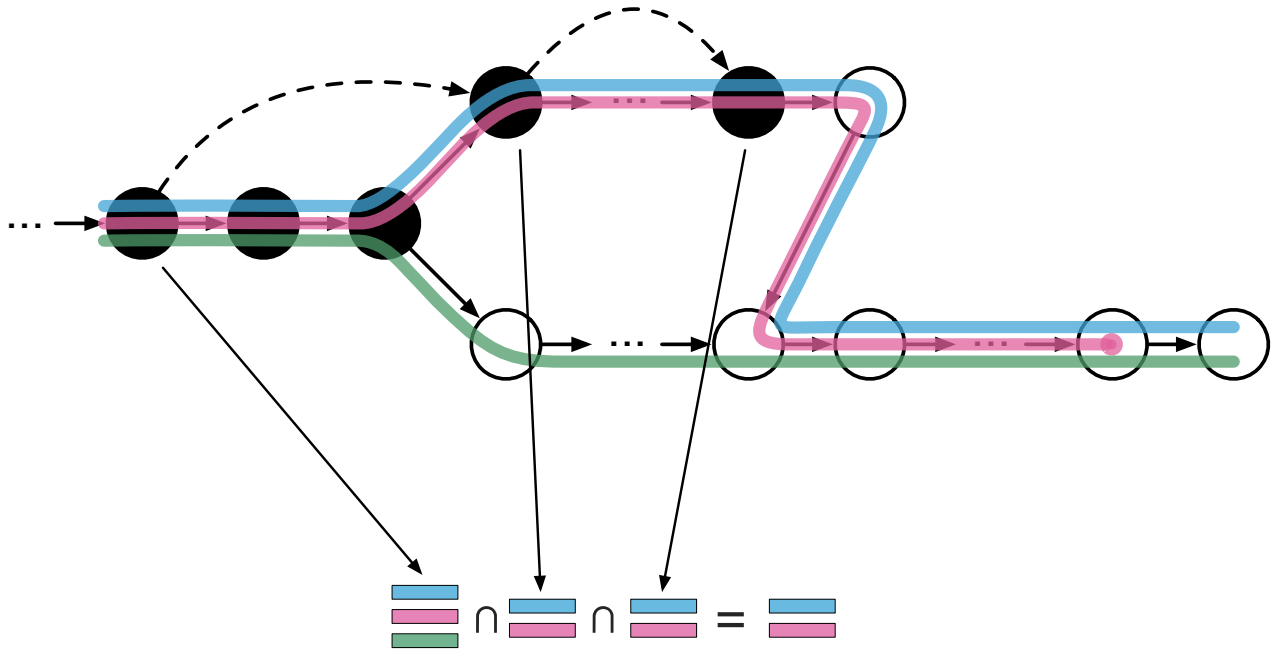
How kallisto computes pseudoaligments



- Combining information across the k-mers can recover lost information
- For each k-mer we have the set of transcripts it could have come from. Intersecting them gives the set of transcripts that *all* k-mers could have come from
- It's possible for their combination to have information equivalent to the entire read, even if no single k-mer does by itself

Slides from Pall Melsted and Nicolas Bray

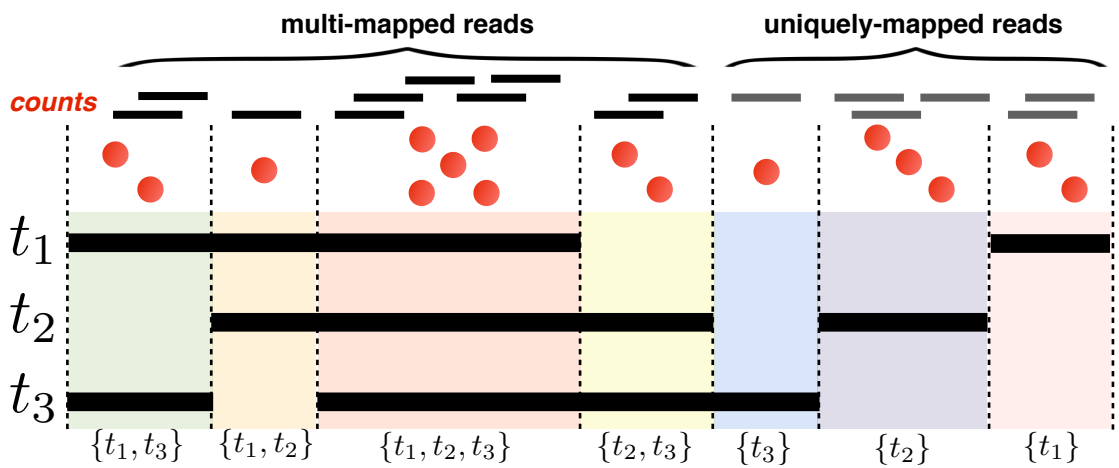
How kallisto computes pseudoaligments

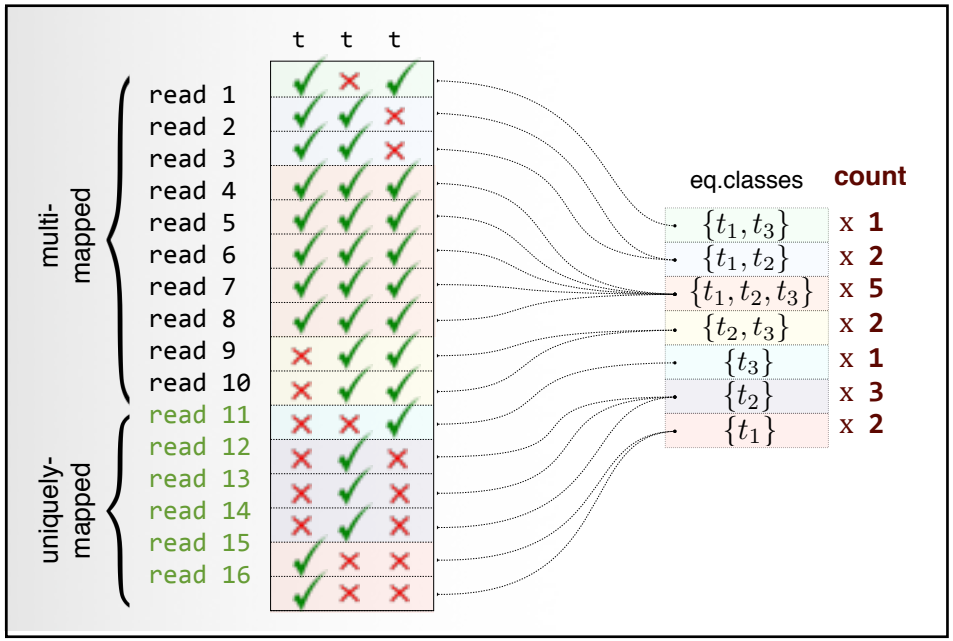


- Knowing the T-DBG, we can predict ahead of time which k-mers will be potentially interesting
- By only processing those k-mers, kallisto runs ~8 times faster

Slides from Pall Melsted and Nicolas Bray

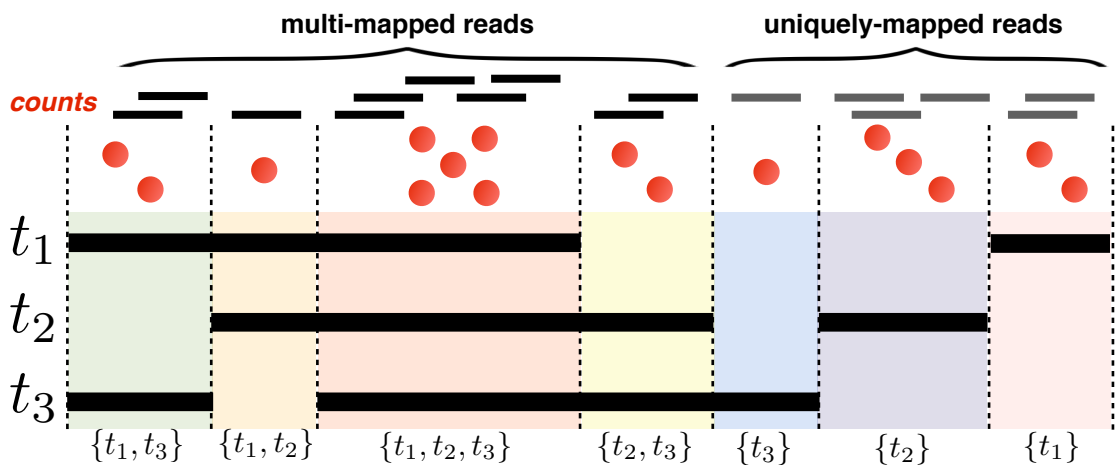
Transcript compatibility counts

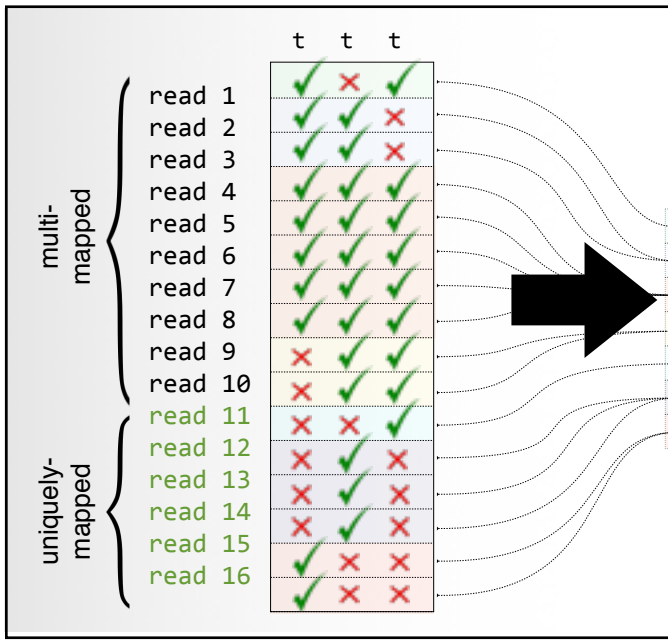


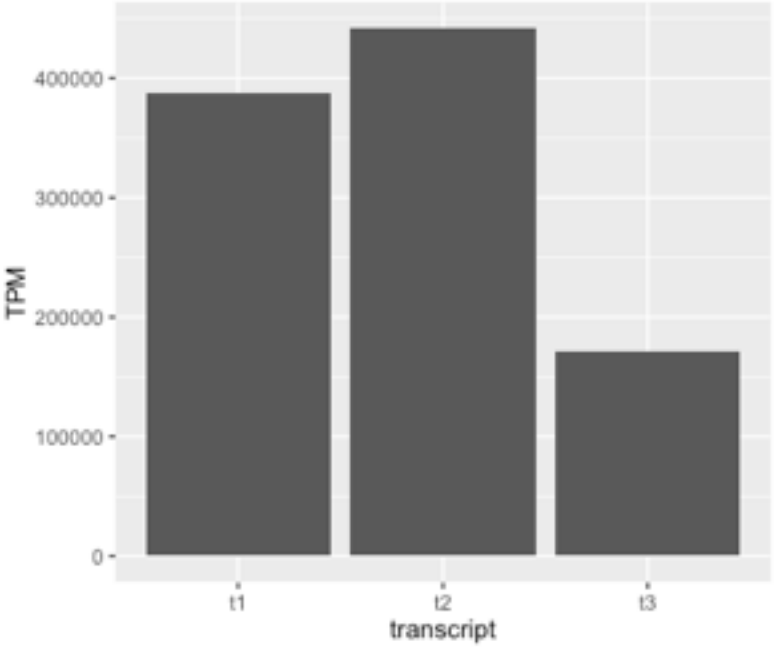


Slides from Pall Melsted and Nicolas Bray

Quantifying transcript abundances







Slides from Pall Melsted and Nicolas Bray

Estimating uncertainty

- “What are the abundances of the different transcripts in my sample?”
- kallisto gives *an* answer but how sure should you be of it?
- In an alternate universe, your sample prep and sequencing might have produced slightly different data for no real biological reason
- What would that data look like?

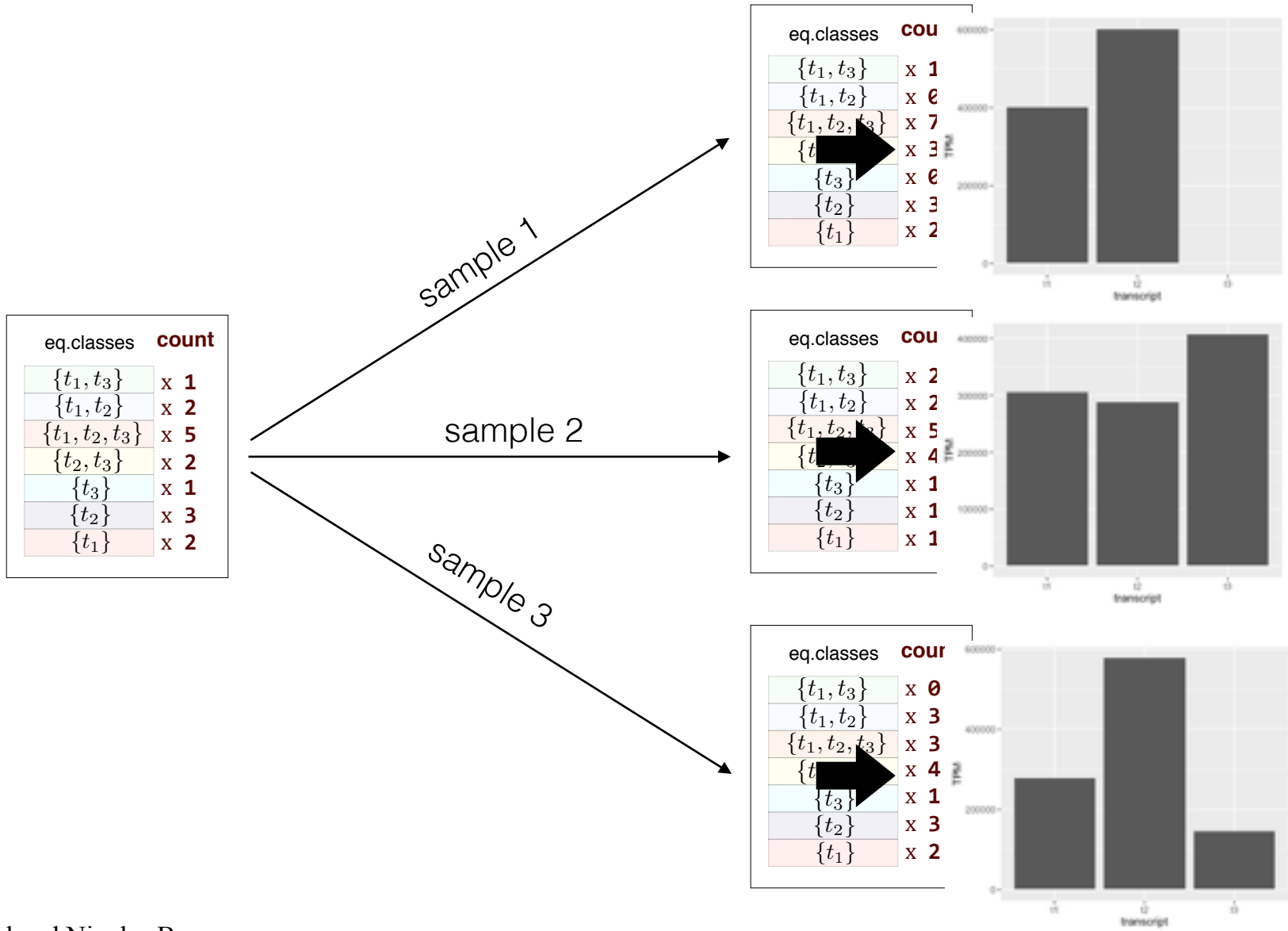
Slides from Pall Melsted and Nicolas Bray

Estimating uncertainty

- The simplicity of the kallisto method allows us to apply a classic statistical tool known as the *bootstrap.*
- We can't access alternate universes, but we can try to simulate them as best we can
- Alternate datasets are constructed by resampling from the original dataset
- Each alternate dataset can then be analyzed with kallisto allowing us to gain some insight into the variability inherent in the data

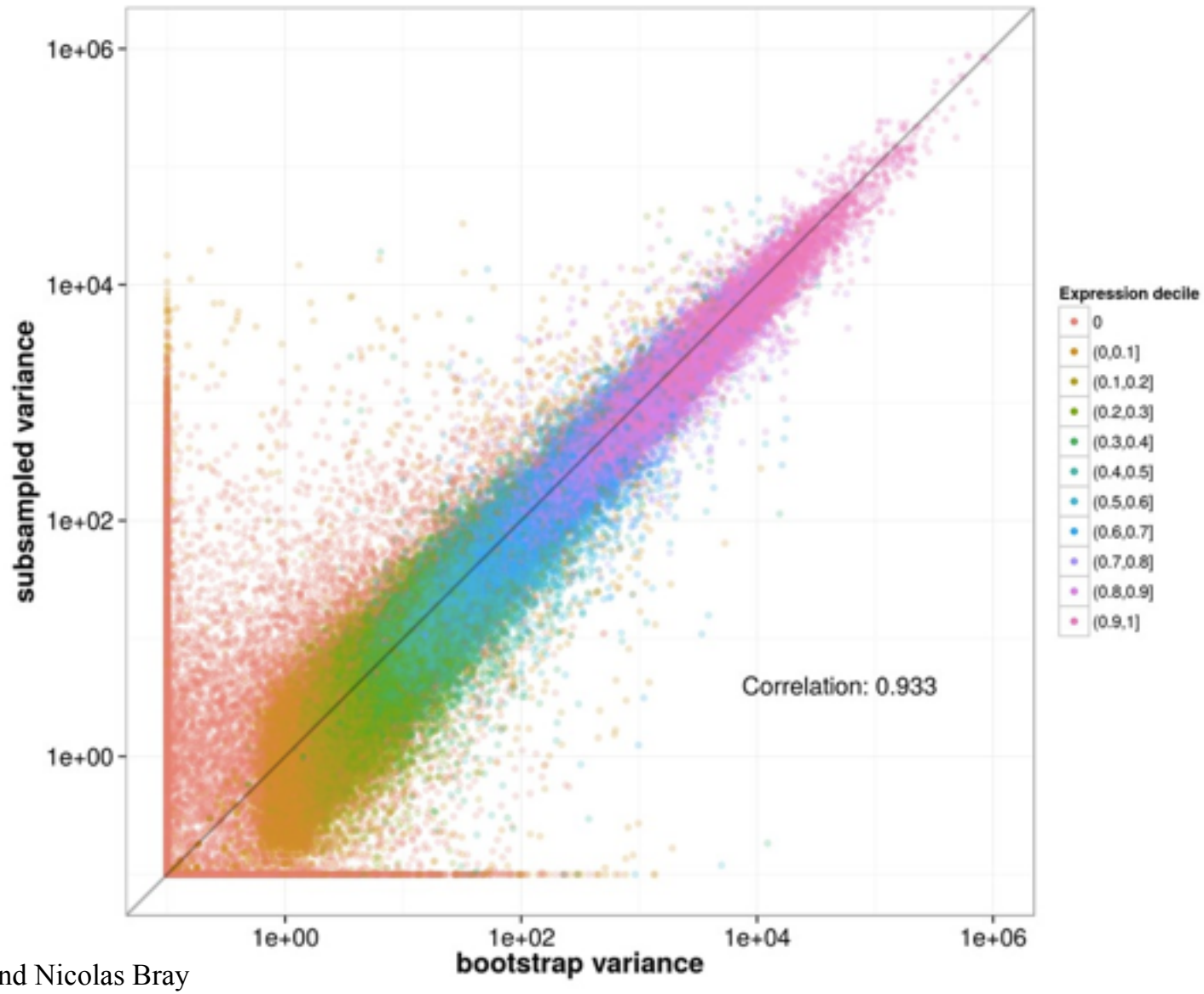
Slides from Pall Melsted and Nicolas Bray

Estimating uncertainty



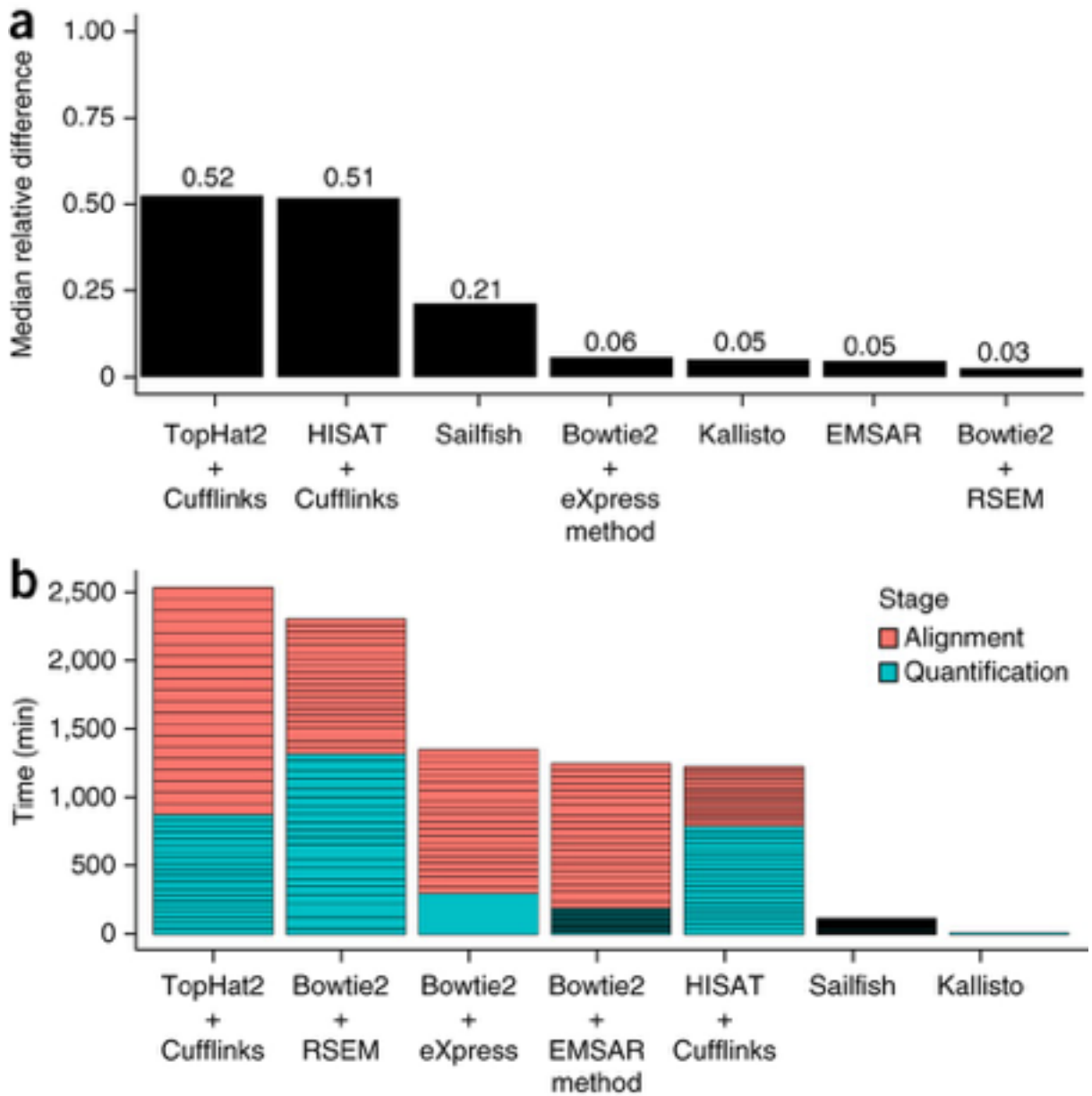
Slides from Pall Melsted and Nicolas Bray

Testing the bootstrap



Slides from Pall Melsted and Nicolas Bray

Kallisto Performance



From http://www.nature.com/nbt/journal/v34/n5/full/nbt.3519.html

Rapmap/Kallisto Summary

- Rapmap: uses suffix array and LCP to avoid uninformative character comparisons
  - Output can be used in rapid abundance inference programs
- Kallisto: uses T-DBG to compute read/transcript compatibility, EM uses equivalence classes to reduce computation
- These approaches *much* faster than previous generation tools
- Field is moving quickly:
  - new Salmon tool from Rob Patro (http://robpatro.com/blog/?p=248)
  - Kallisto has DGE, data exploration extensions (Sleuth)