Don't count on it: Pragmatic and theoretical concerns and best practices for mapping and quantifying RNA-seq data

Rob Patro

Department of Computer Science

Laufer Center for Physical and Quantitative Biology

@nomad421

rob-p, COMBINE-lab

Stony Brook University

July 7, 2017

RNA-Seq Read Alignment

Given an RNA-seq read, where *might* it come from?

Two main “regimes”

Align to transcriptome

Align reads directly to txps

No “split” alignments — transcripts contain spliced exons directly.

Typically *a lot* of multi-mapping (80-90% of reads may map to multiple places)

Does *not r*equire target *genome*

Can be used in *de novo* context (i.e. after *de novo* assembly)

Align to genome

Align reads to target genome

Reads spanning exons will be “split” (gaps up to 10s of kb)

Typically little multi-mapping (most reads have single genomic locus of origin)

Requires target *genome*

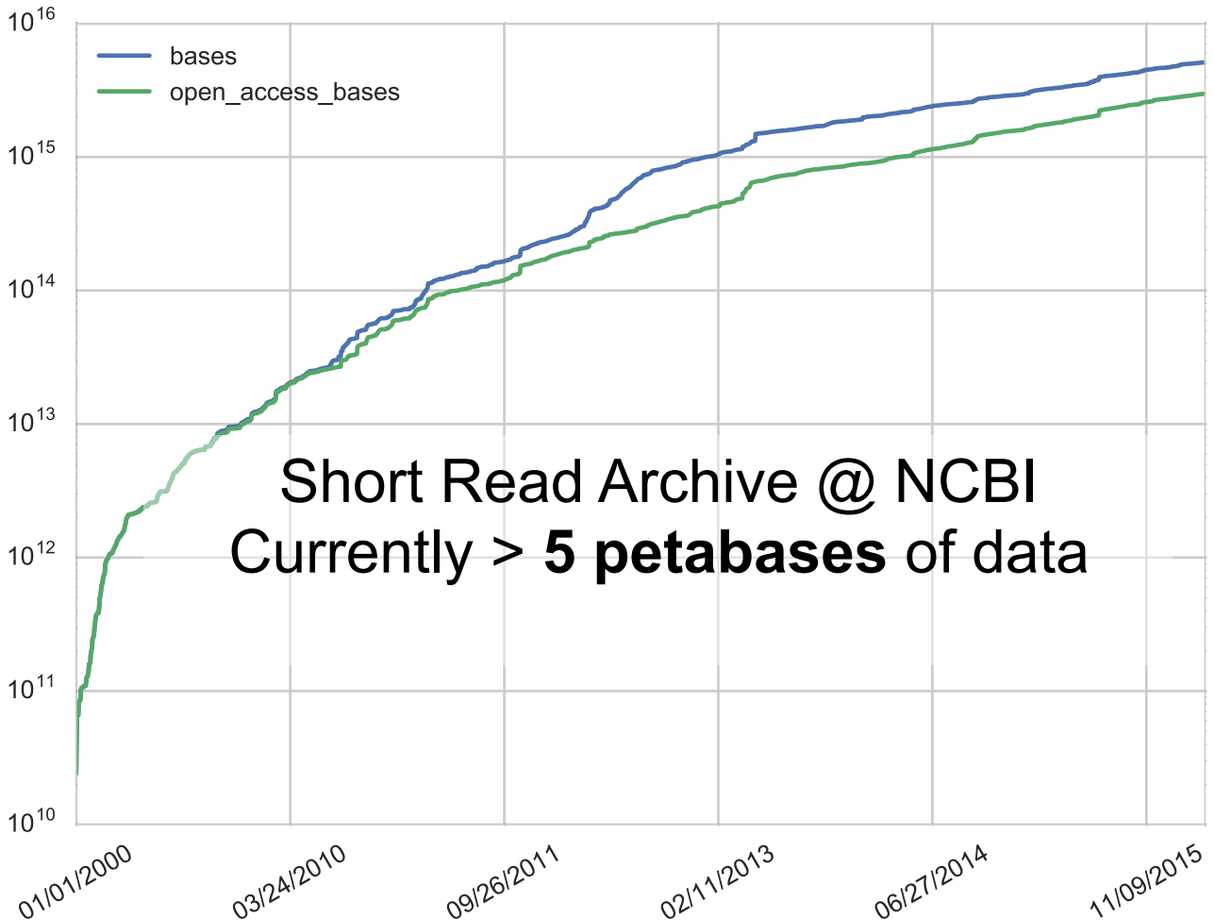
Can be used to find new transcripts

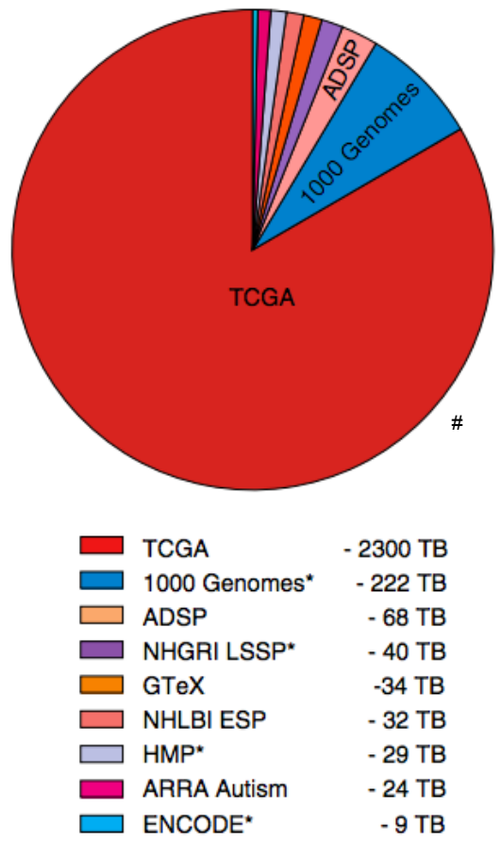
Why do we need faster analysis?

OPINION Open Access

The real cost of sequencing: scaling computation to keep pace with data generation

Paul Muir[1,2,3], Shantao Li[4], Shaoke Lou[4,5], Daifeng Wang[4,5], Daniel J Spakowicz[4,5], Leonidas Salichos[4,5], Jing Zhang[4,5], George M. Weinstock[6], Farren Isaacs[1,2], Joel Rozowsky[4,5] and Mark Gerstein[4,5,7*]

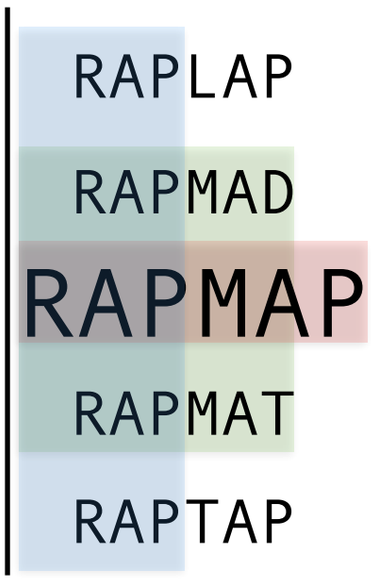




In addition to new data, re-analysis of existing experiments often desired: In light of new annotations, discoveries, and methodological advancements.

From Fig 1 of Muir et al.

RapMap: A Rapid, Sensitive and Accurate Tool for Mapping RNA-seq Reads to Transcriptomes



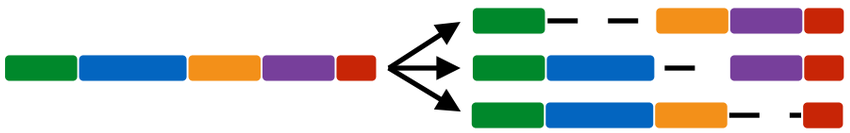
GitHub repository: https://github.com/COMBINE-lab/RapMap (C++11, GPL v3)

Paper: http://bioinformatics.oxfordjournals.org/content/32/12/i192.full.pdf

(appeared at ISMB 16)

co-authors (students): Avi Srivastava, Hirak Sarkar, Nitish Gupta

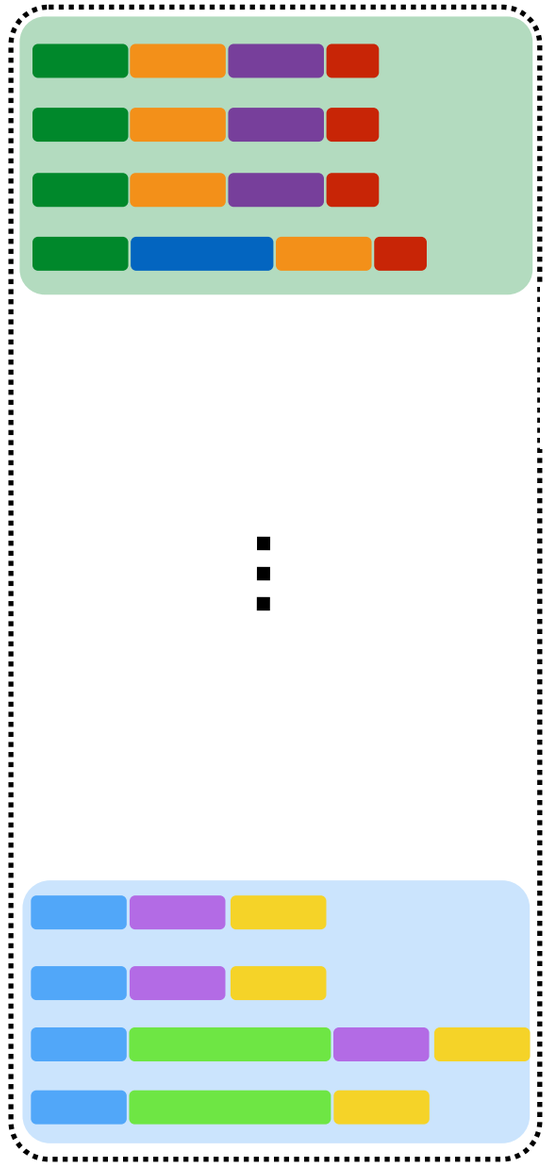
Transcript Quantification: An Overview



1 gene ⇒ many variants (isoforms)

Gene 1

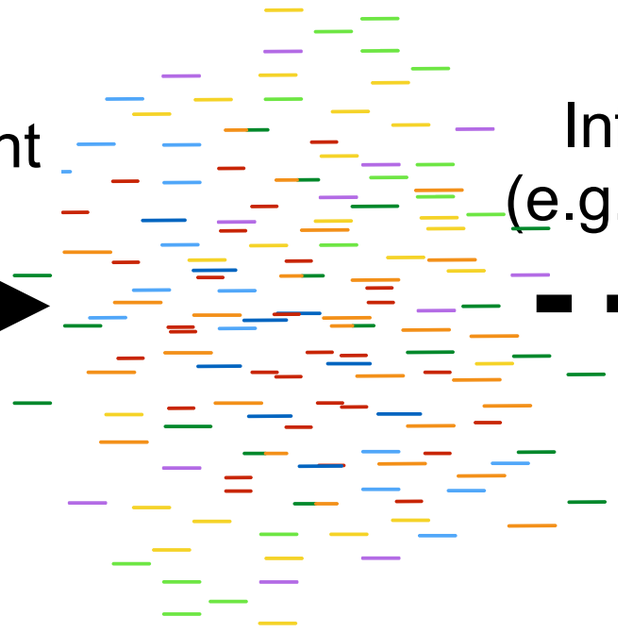
Gene M



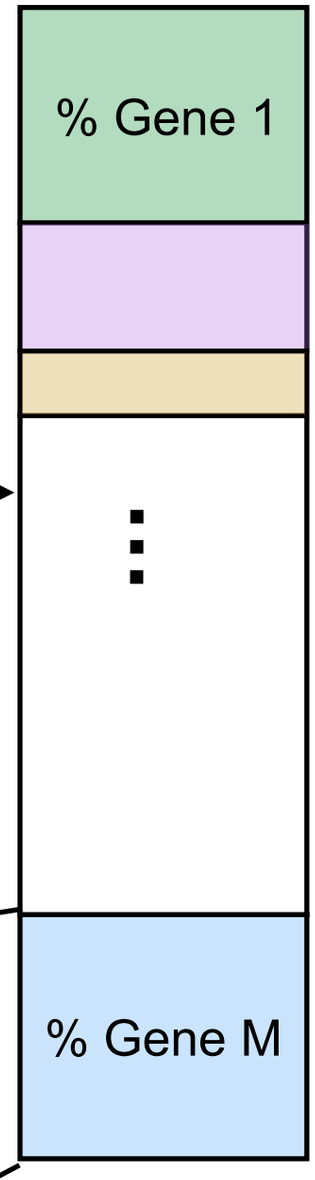
Sample

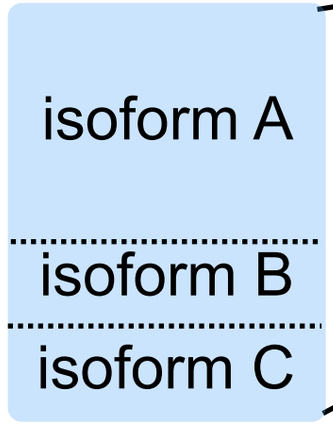
10s-100s of millions of
short (35-300 character) “fragments”

Measurement
(RNA-seq)



Inference
(e.g. Salmon)





Abundance Estimates

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

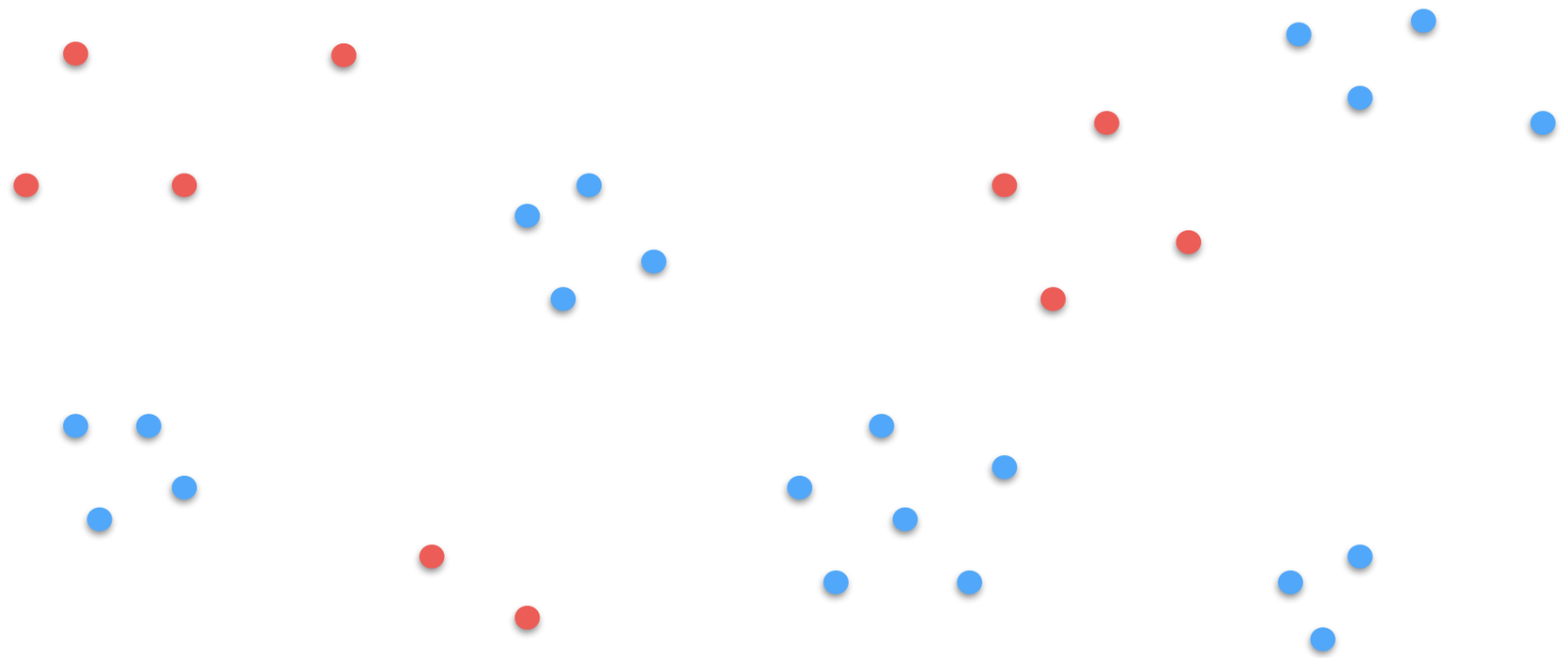
Estimate: The relative abundance of each transcript

Question: If we only care about “gene" abundance, can’t we just count the number of reads mapping / aligning to each gene?

Answer: **No.** I’ll show a general argument (and a few examples) why!

First, consider this non-Biological example

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



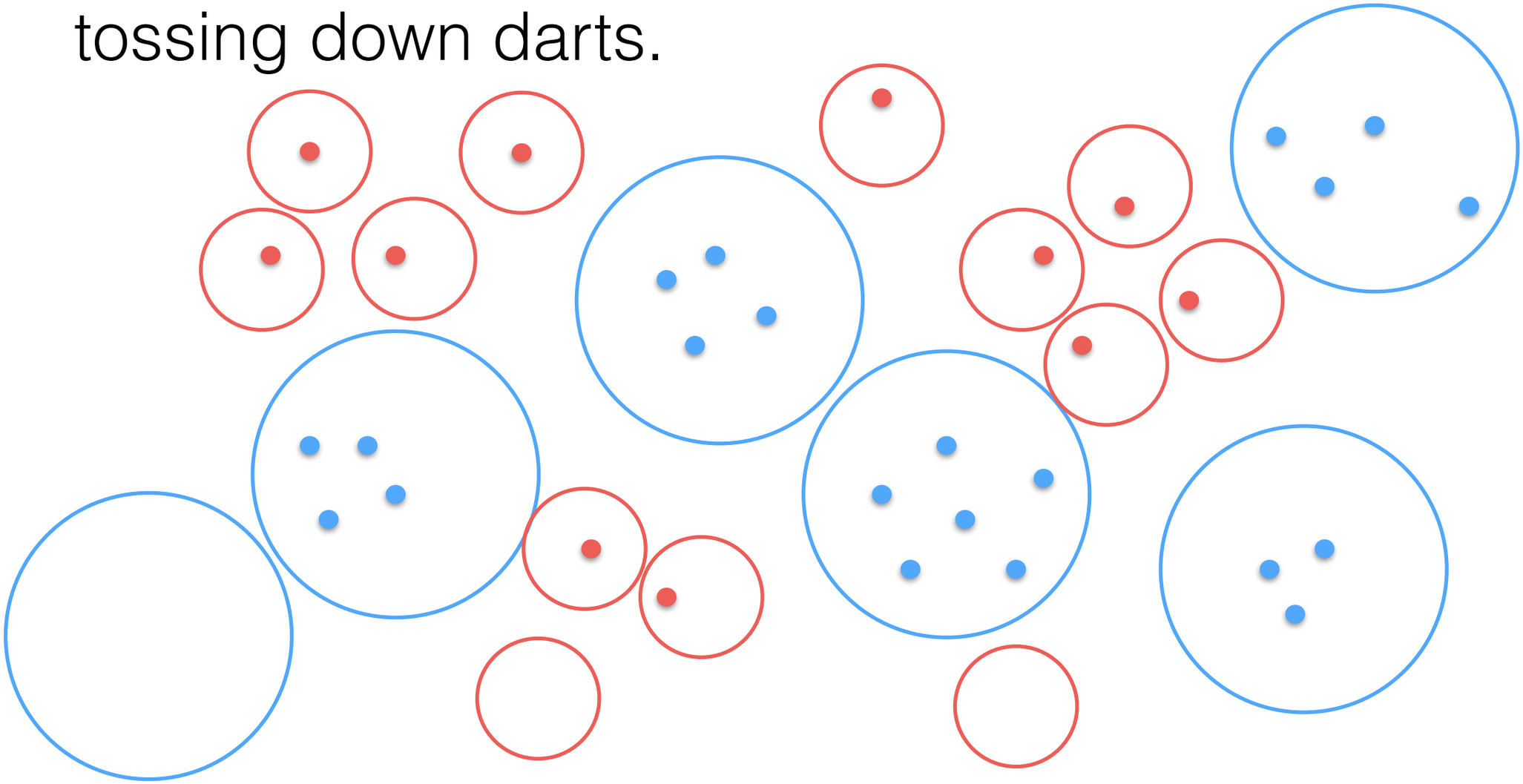
Here, a dot of a color means I hit a circle of that color.

What type of circle is more prevalent?

What is the fraction of red / blue circles?

First, consider this non-Biological example

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



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

The areas!

First, consider this non-Biological example

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

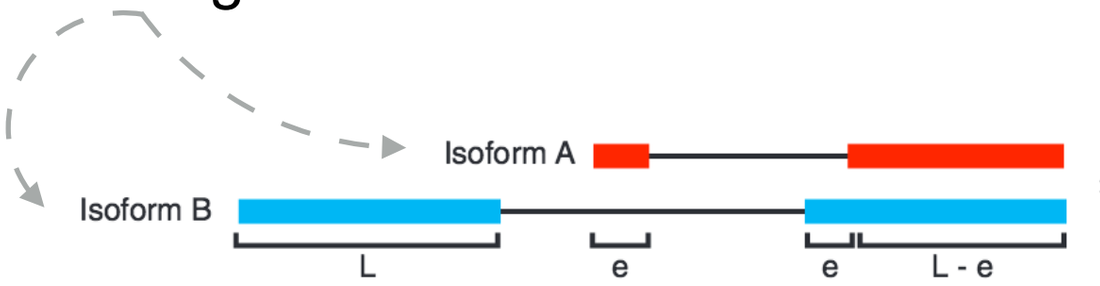
You're missing a **crucial piece of information!**

The areas!

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

Resolving multi-mapping is fundamental to quantification

Isoform A is half
as long as isoform B



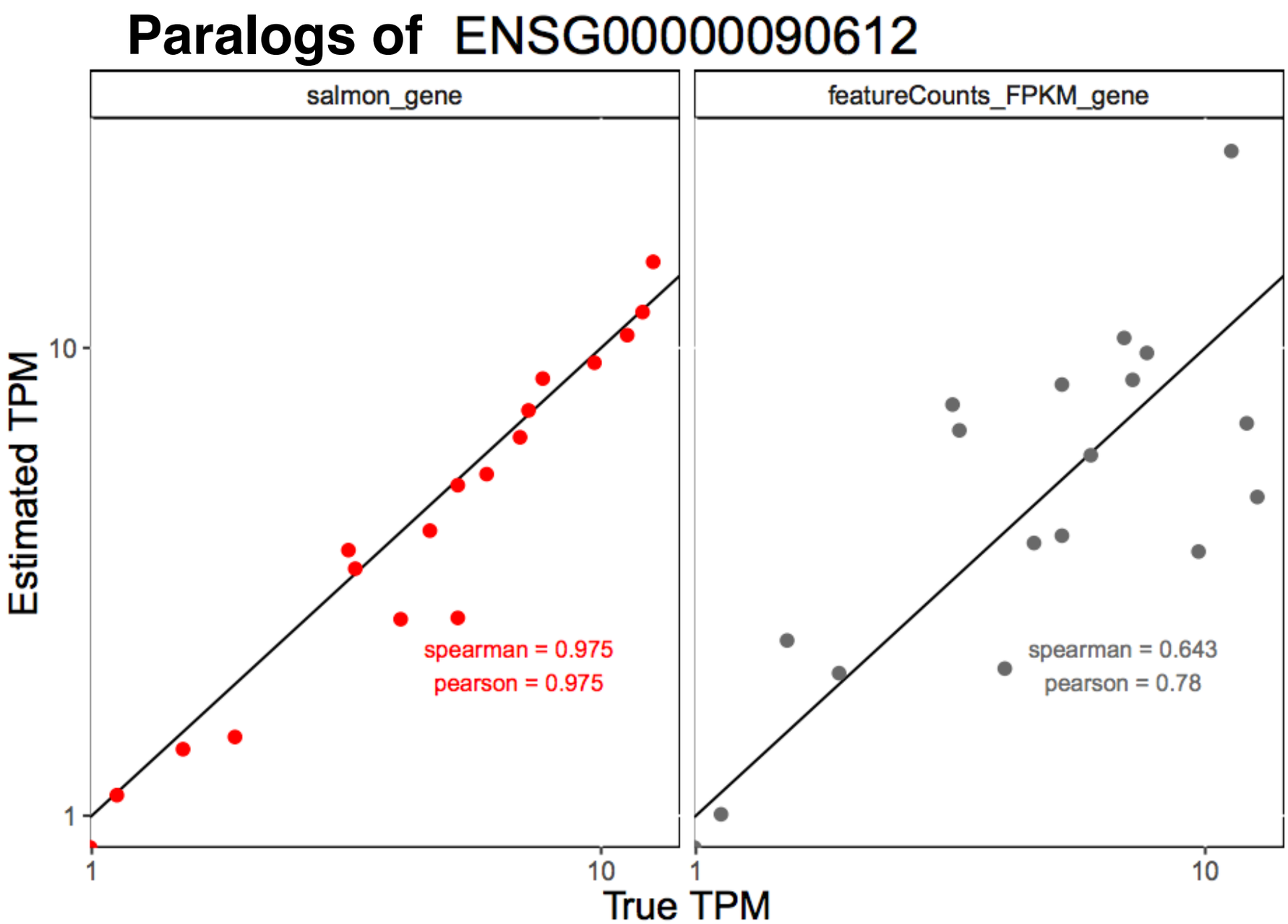
Condition 1	Condition 2	union-model fold-change		true fold-change
		$\log_2\left(\frac{10}{10}\right) = 0$	$0 < 0.32$	$\log_2\left(\frac{\frac{10}{L}}{\frac{6}{L}+\frac{4}{2L}}\right) = 0.32$
		$\log_2\left(\frac{6}{8}\right) = -0.41$	$-0.41 < 0.58$	$\log_2\left(\frac{6/L}{8/2L}\right) = 0.58$
		$\log_2\left(\frac{5}{10}\right)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

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.

Resolving multi-mapping is fundamental to quantification

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



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

Main challenges of fast & accurate quantification

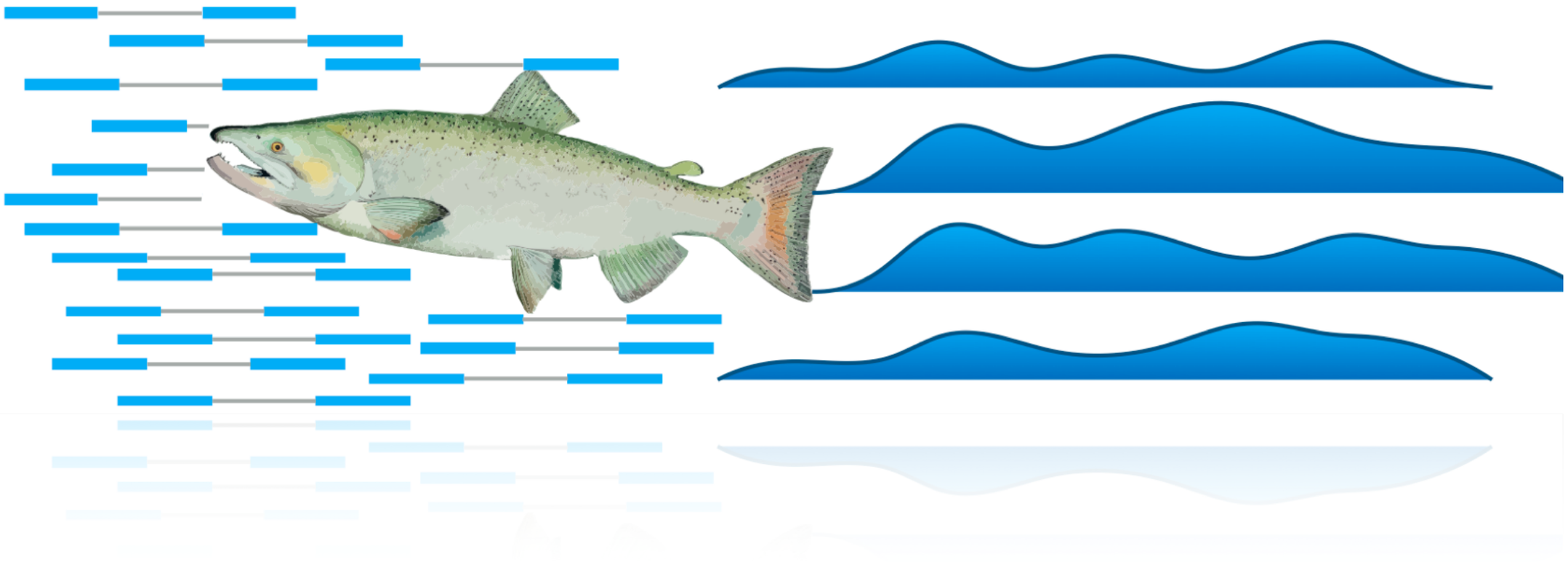
- finding locations of reads (alignment) is slower than necessary → simply aligning reads in a sample **can take hours**

- alternative splicing and related sequences creates ambiguity about where reads came from → **multi-mapping reads** *cannot* be ignored / discarded or assigned naïvely

- sampling of reads is not uniform or idealized, exhibits multiple types of bias → RNA-seq can exhibit ***extensive* and *sample-specific bias***

- uncertainty in ML estimate of abundances → There is both technical (shot noise) and **inherent *inferential* uncertainty** in abundance estimates

Salmon provides fast and bias-aware quantification of transcript expression



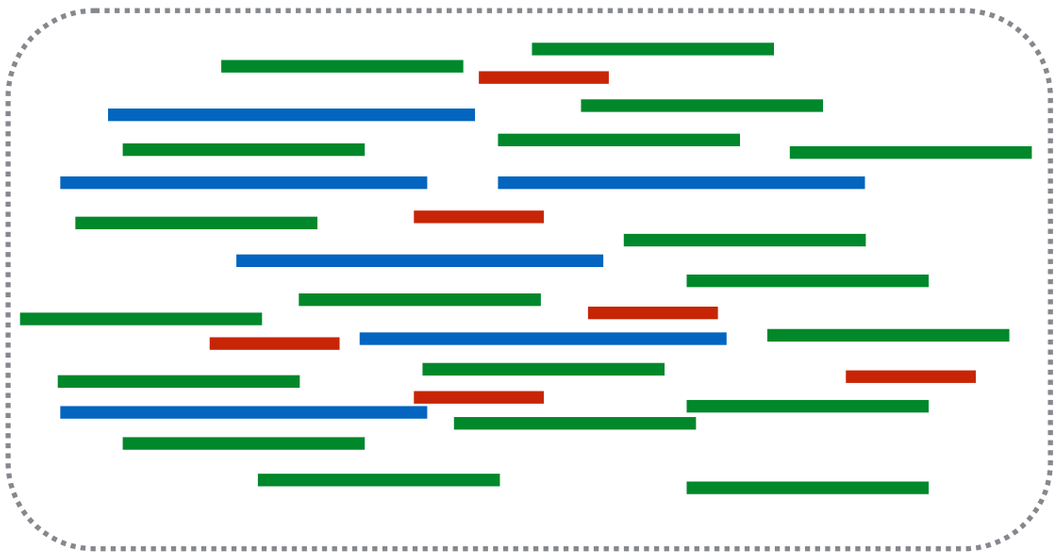
Official website: http://combine-lab.github.io/salmon/

GitHub repository: https://github.com/COMBINE-lab/salmon (C++11, GPL v3)

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017).
Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*.

How can we perform inference from sequenced fragments?

Experimental Mixture



In an unbiased experiment, sampling fragments depends on:

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

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

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

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

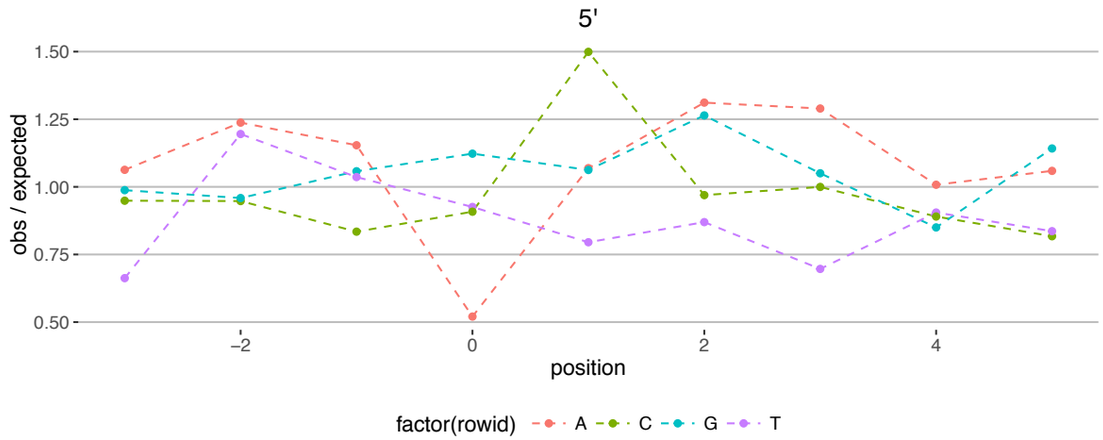
Biases abound in RNA-seq data

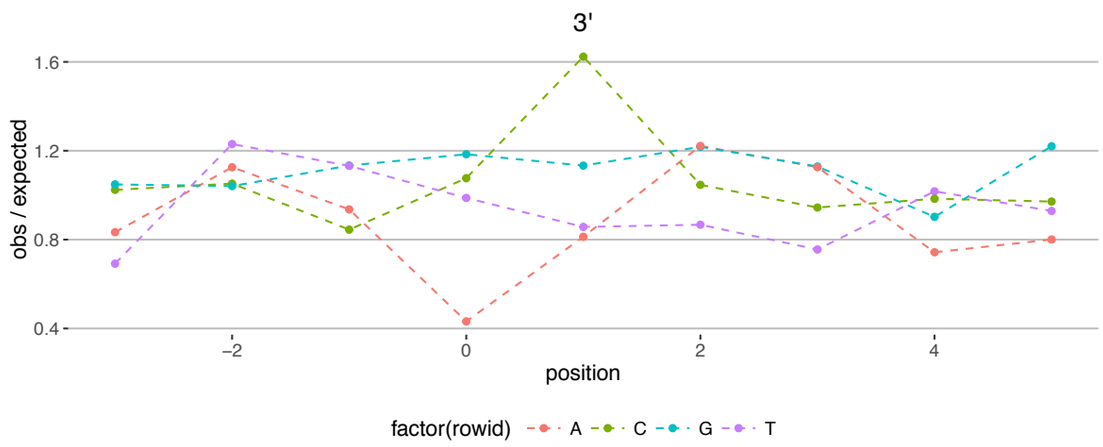
Biases in prep & sequencing can have a significant effect on the fragments we see:

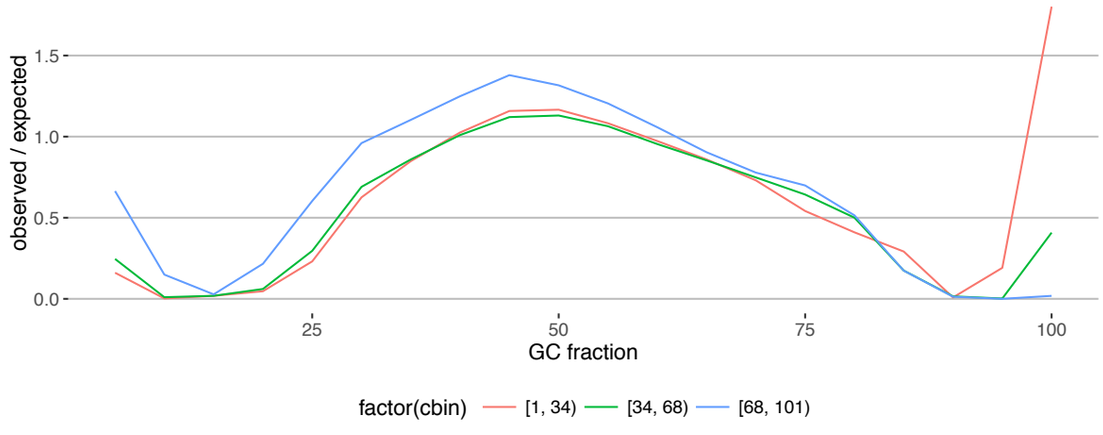
Fragment gc-bias[1]—
The GC-content of the fragment affects the likelihood of sequencing

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

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







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

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

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

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

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

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

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

Bias Modeling

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

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

Fragment GC bias model:

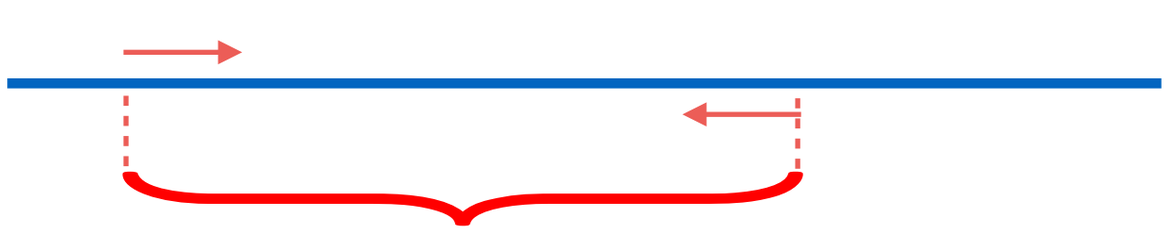
Density of fragments with specific GC content, **conditioned** on GC fraction at read start/end

Foreground:

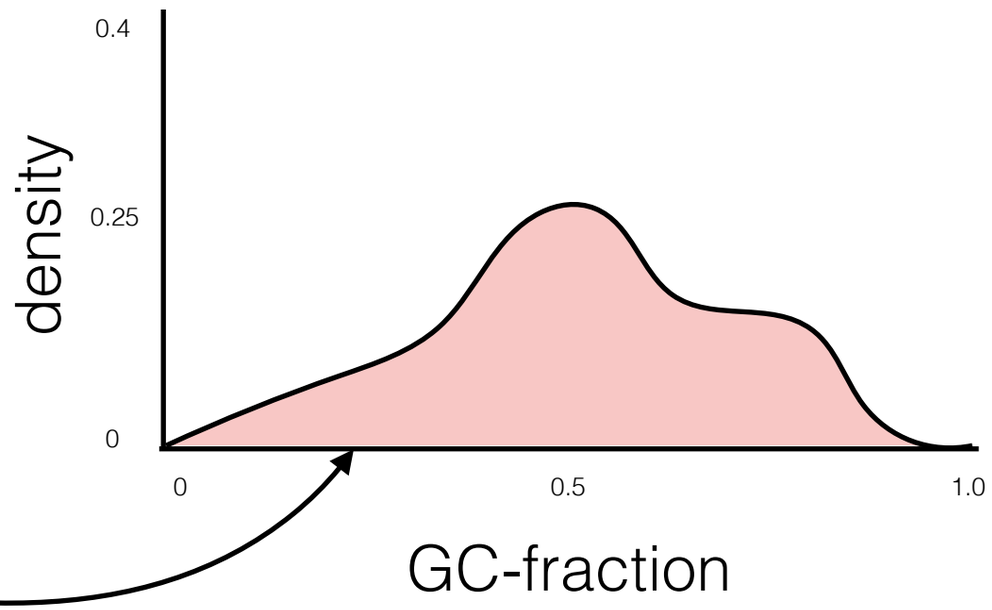
Observed

Background:

Expected given est. abundances



GC-fraction of fragment



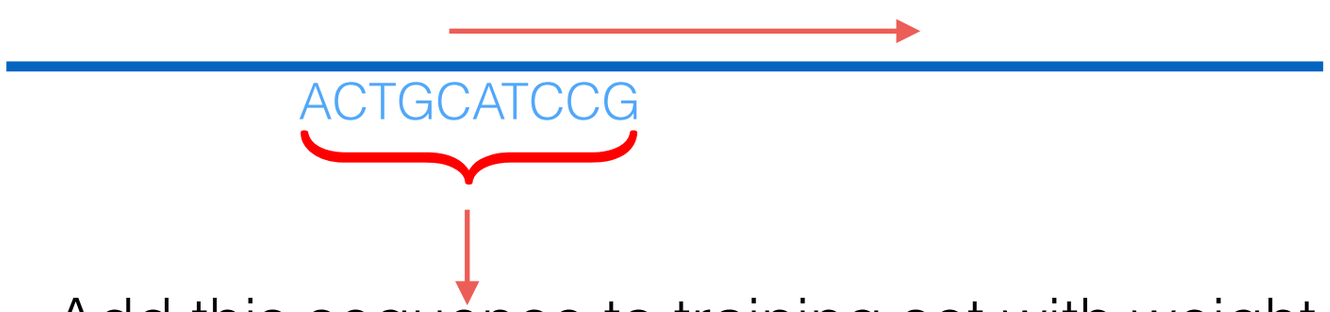
Bias Modeling

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

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

Seq-specific bias model*:

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



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

Foreground:

Observed

Background:

Expected given est. abundances

Same, but independent model for 3’ end

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

Bias Modeling

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

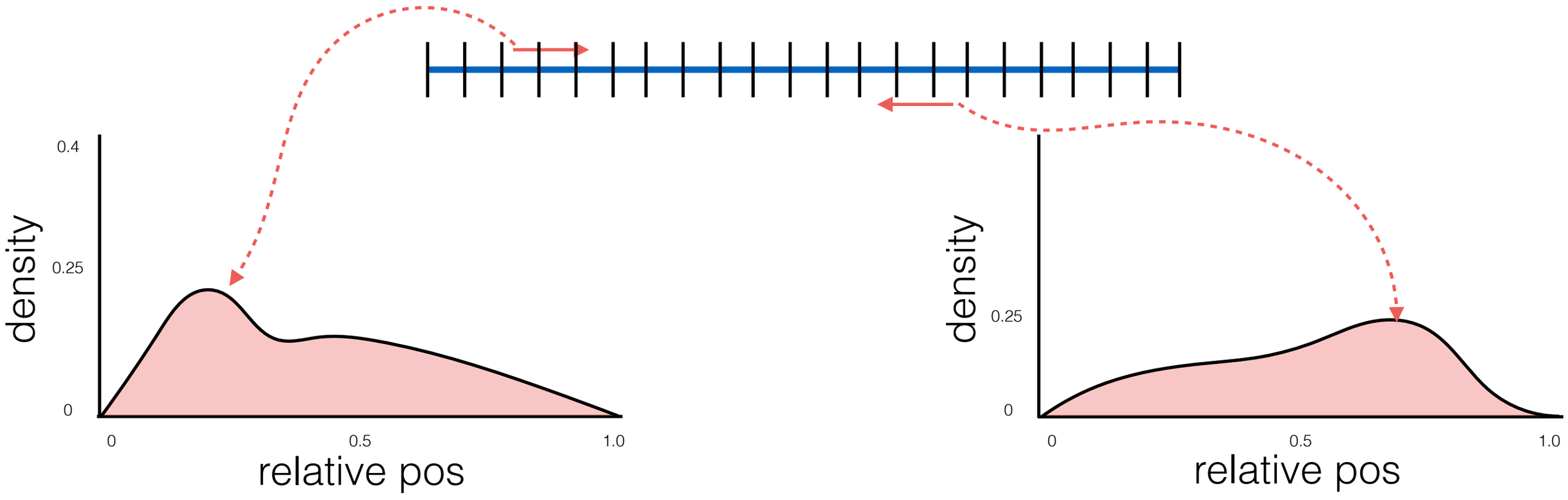
$$\tilde{\ell}'_i = \sum_{j=1}^{j \leq \ell_i} \sum_{k=1}^{k \leq f_i(j,L)} \frac{b_{gc^+}(t_i, j, j+k)}{b_{gc^-}(t_i, j, j+k)} \cdot \frac{b^{5'}_{s^+}(t_i, j)}{b^{5'}_{s^-}(t_i, j)} \cdot \frac{b^{3'}_{s^+}(t_i, j+k)}{b^{3'}_{s^-}(t_i, j+k)} \cdot \frac{b^{5'}_{p^+}(t_i, j+k)}{b^{5'}_{p^-}(t_i, j+k)}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

Position bias model*:

Density of 5’ and 3’ read start positions — different models for transcripts of different length

Foreground:
Observed

Background:
Expected given est. abundances

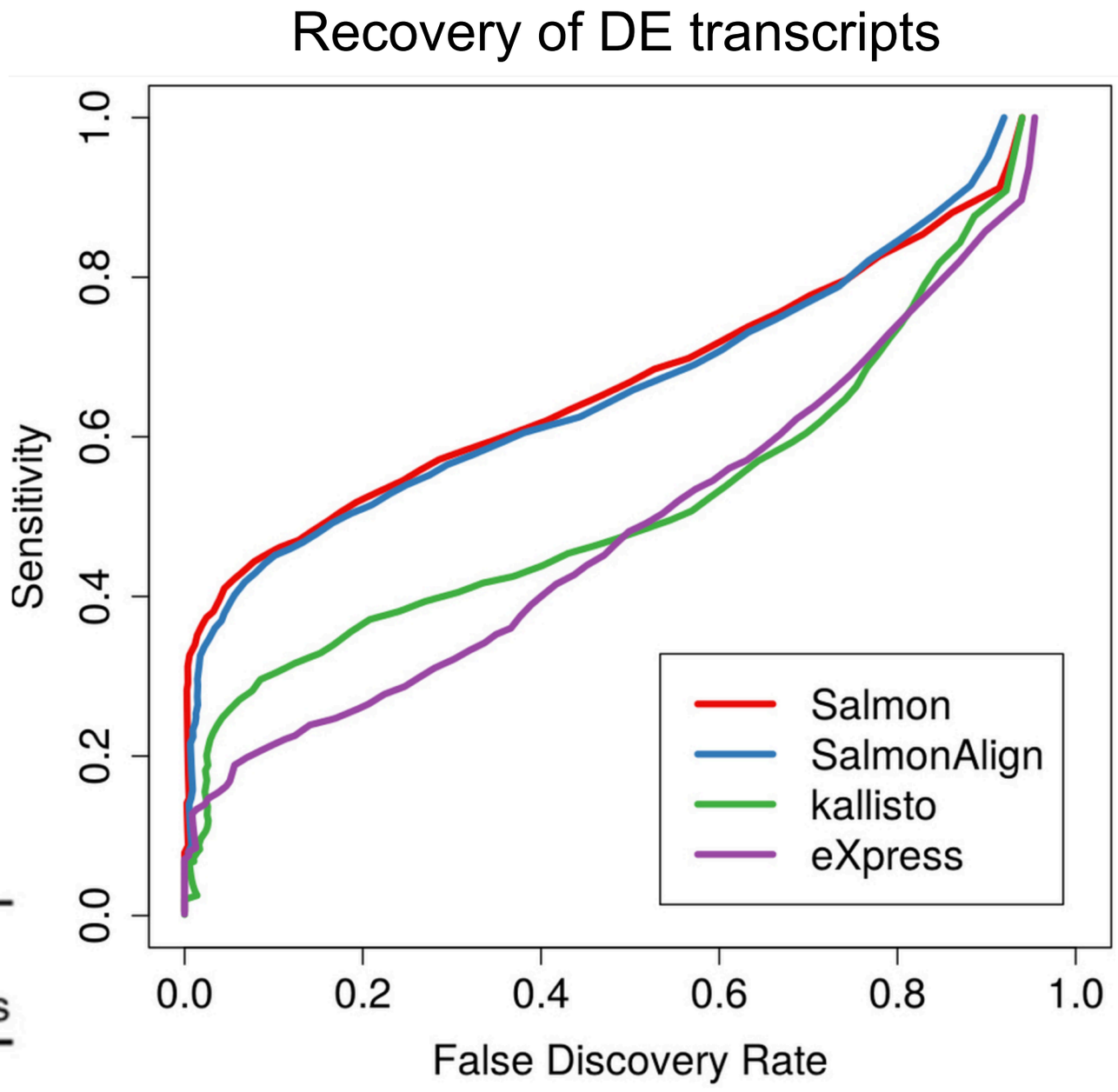


*Roberts, Adam, et al. "Improving RNA-Seq expression estimates by correcting for fragment bias." Genome biology 12.3 (2011): 1.

Mis-estimates confound downstream analysis

Simulated data:
2 conditions; 8 replicates each

- set 10% of txps to have fold change of 1/2 or 2 — rest unchanged.
- How well do we recover true DE?
- Since bias is systematic, effect may be even worse than accuracy difference suggests.



	Sensitivity at given FDR			
FDR	Salmon	Salmon (a)	kallisto	eXpress
0.01	0.326	0.233	0.072	0.128
0.05	0.409	0.379	0.248	0.162
0.1	0.454	0.442	0.296	0.211

At the same FDR, accuracy differences of 53 - 450%

Salmon addresses the main challenges of quantification

- finding locations of reads (alignment) is slow than necessary → Use quasi-mapping

- alternative splicing and related sequences creates ambiguity about where reads came from → Use dual-phase inference algorithm

- sampling of reads is not uniform or idealized,exhibits multiple types of bias → Use bias models learned from data

- uncertainty in ML estimate of abundances → Use posterior Gibbs sampling or bootstraps to assess uncertainty

Some remaining challenges*

What other biases affect quantification? How can we model them?

What is the right way to incorporate / propagate uncertainty *downstream* ?

What if the reference sequence is wrong / incomplete ?

How can we scale fast “mapping” strategies to large reference seqs ?

> The indices for these are currently large; fine for the txome, but bloated for the genome / metagenome / metatranscriptome.
>
> ***New data structure in development that addresses this** (http://robpatro.com/blog/?p=494)

How can exploit fast mapping to generate alignments when we need them (e.g. for variant detection)?

> Fresh pre-print on some ideas in this direction; enables actual “alignment” but still “ultra-fast”
>
> **Towards Selective-Alignment: Producing Accurate And Sensitive Alignments Using Quasi-Mapping**
> Hirak Sarkar, Mohsen Zakeri, Laraib Malik, Rob Patro. bioRxiv 138800; doi: https://doi.org/10.1101/138800

* List of things immediately on our mind; far from comprehensive!

Thanks!

Collaborators on Salmon

Geet Duggal (CMU / DNAnexus)

Carl Kingsford (CMU)

Mike Love (Harvard / UNC)

Rafael Irizarry(Harvard)

Members of COMBINE Lab

Fatemeh Almodaresi

Lariab Malik

Hirak Sarkar

Avi Srivastava

Mohsen Zakeri

BIO-1564917