Estimating Gene Expression

Lecture 27
Nov 4, 2016

GO VOTE!

If you're a U.S. citizen and 18 or older, you have the right to vote. Registering to vote is simpler and faster than you may think!

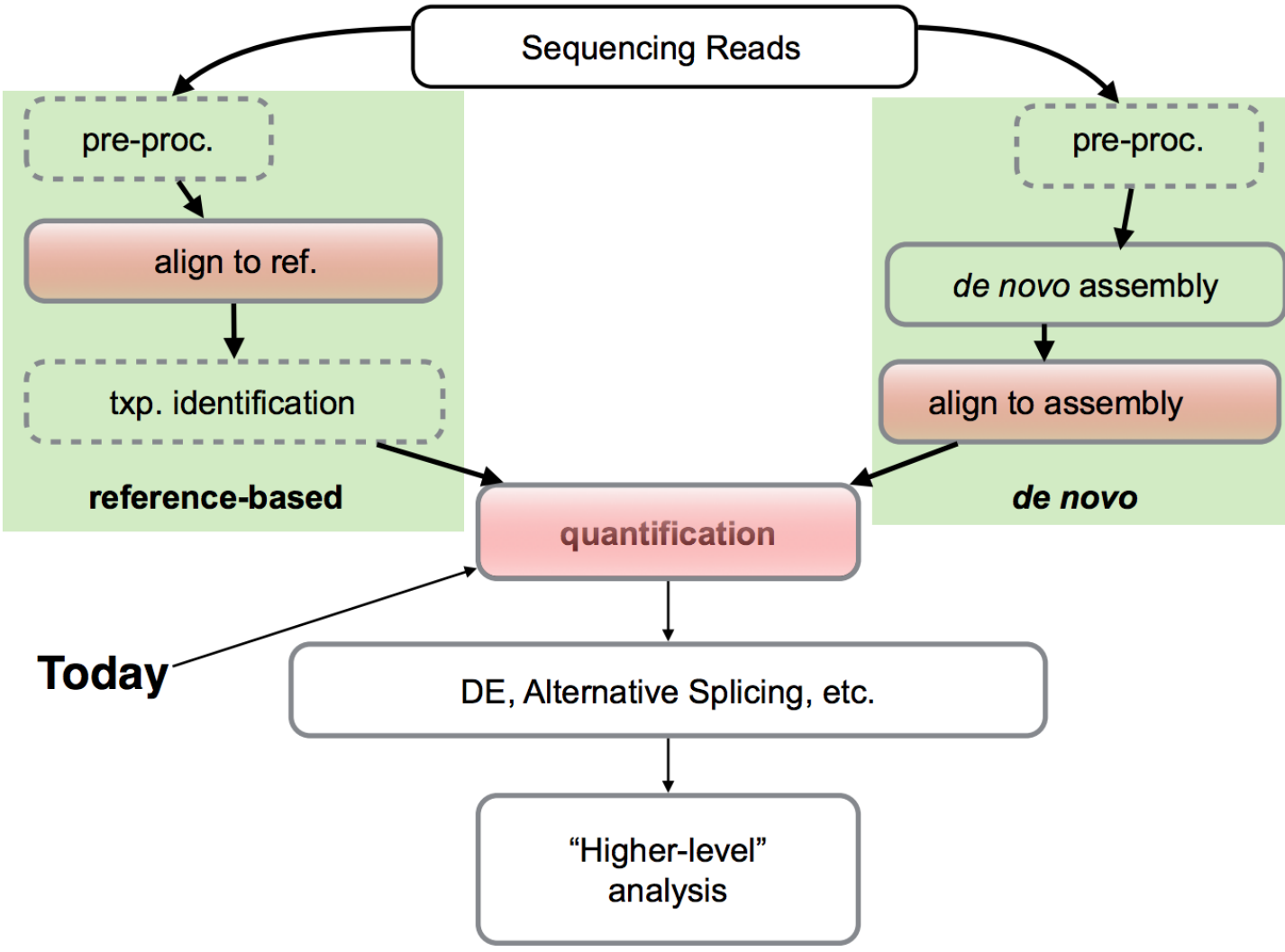
Important Dates

General Election: Tuesday, November 8, 2016

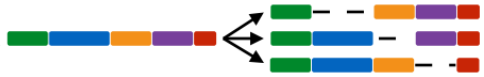
For Durham students, shuttle service will be running from the Holloway Commons bus stop on Main St. to the polling location at Oyster River High School from 7 a.m. to 7 p.m. on Election Day. Please contact the Dean of Students office with any questions.

Voting Info

Voting in New Hampshire is quick and easy, especially when you bring appropriate identification. A valid state driver's license or non-driver ID proves your identity, your age, and your domicile (if it shows the address you are claiming as your voting domicile). A lease or piece of mail will also show your domicile. A birth certificate (or copy) or passport will prove your citizenship. If you lack proper identification, you may fill out an affidavit (a legal document attesting to your identity and domicile). You may register in person at the town or city clerk's office up to 10 days prior to an election, or on Election Day at the polling place.


Sequencing Reads
pre-proc.
align to ref.
txp. identification
reference-based
pre-proc.
de novo assembly
align to assembly
de novo
quantification
Today
DE, Alternative Splicing, etc.
"Higher-level" analysis

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)

% Gene 1

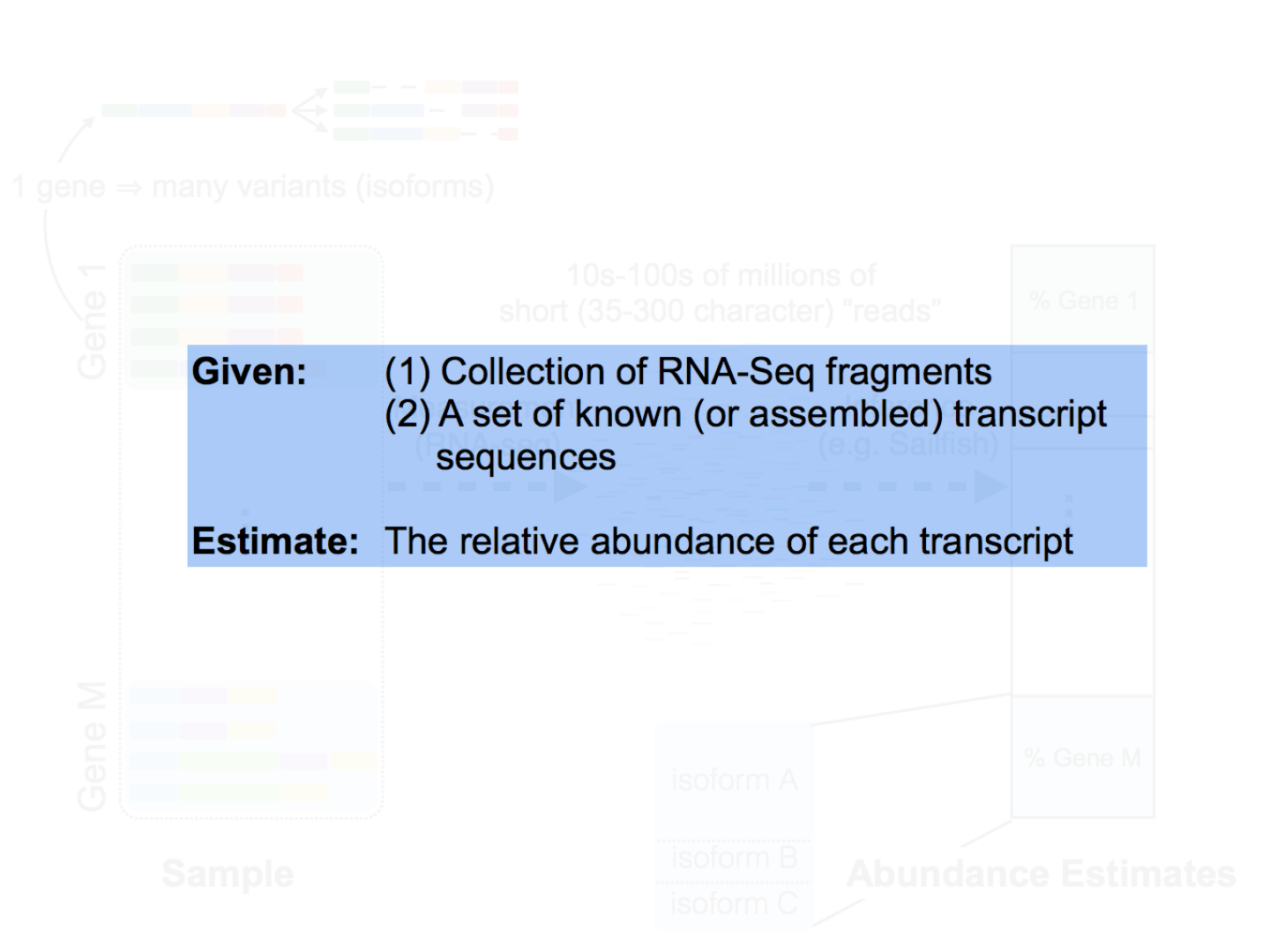
% Gene M

isoform A

isoform B

isoform C

Abundance Estimates

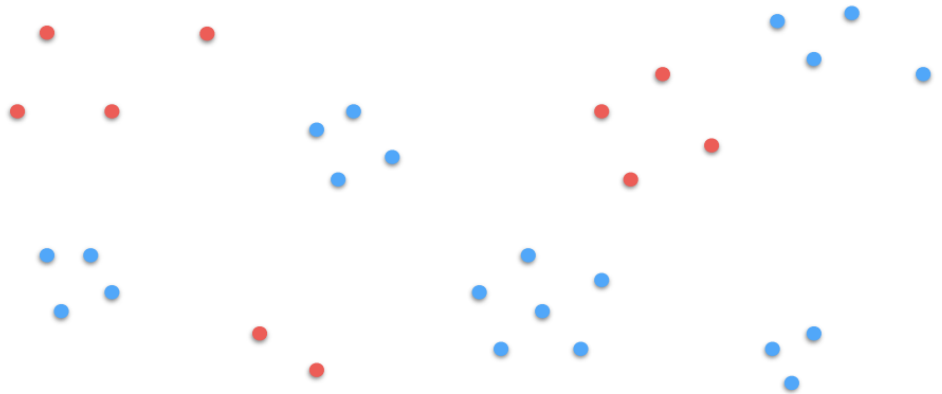


Given: (1) Collection of RNA-Seq fragments
(2) A set of known (or assembled) transcript sequences

Estimate: The relative abundance of each transcript

First, consider this non-Biological example

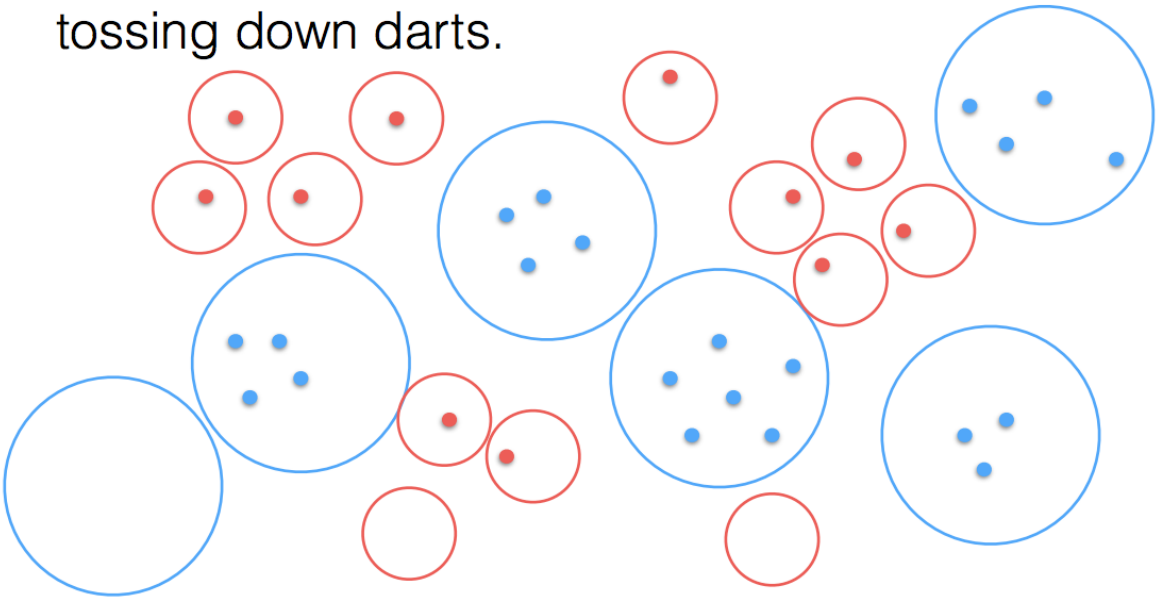
Imagine I have two colors of circle, red and blue. I want to estimate the **fraction of circles** that are red and blue. I'll *sample* from them by tossing down darts.



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.

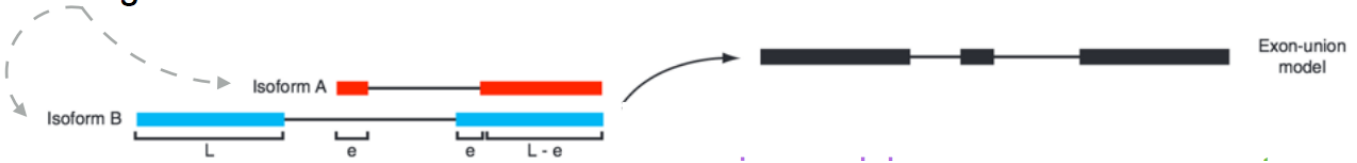


You're missing a **crucial piece of information!**

The areas!

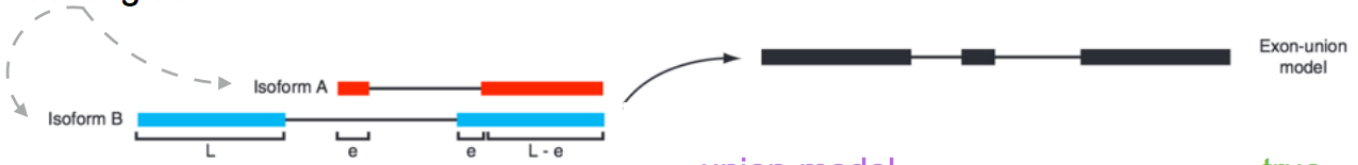
Resolving multi-mapping is fundamental to quantification

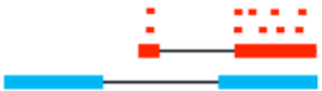
Isoform A is half
as long as isoform B



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.

How to assess “abundance”

RPKM — Reads per kilobase per million mapped reads

FPKM — Fragments per kilobase per million mapped reads

Don’t use these measures, TPM measures the “same thing”, but in a better way.

TPM — Transcripts per million

Useful for visualization / assessment etc.

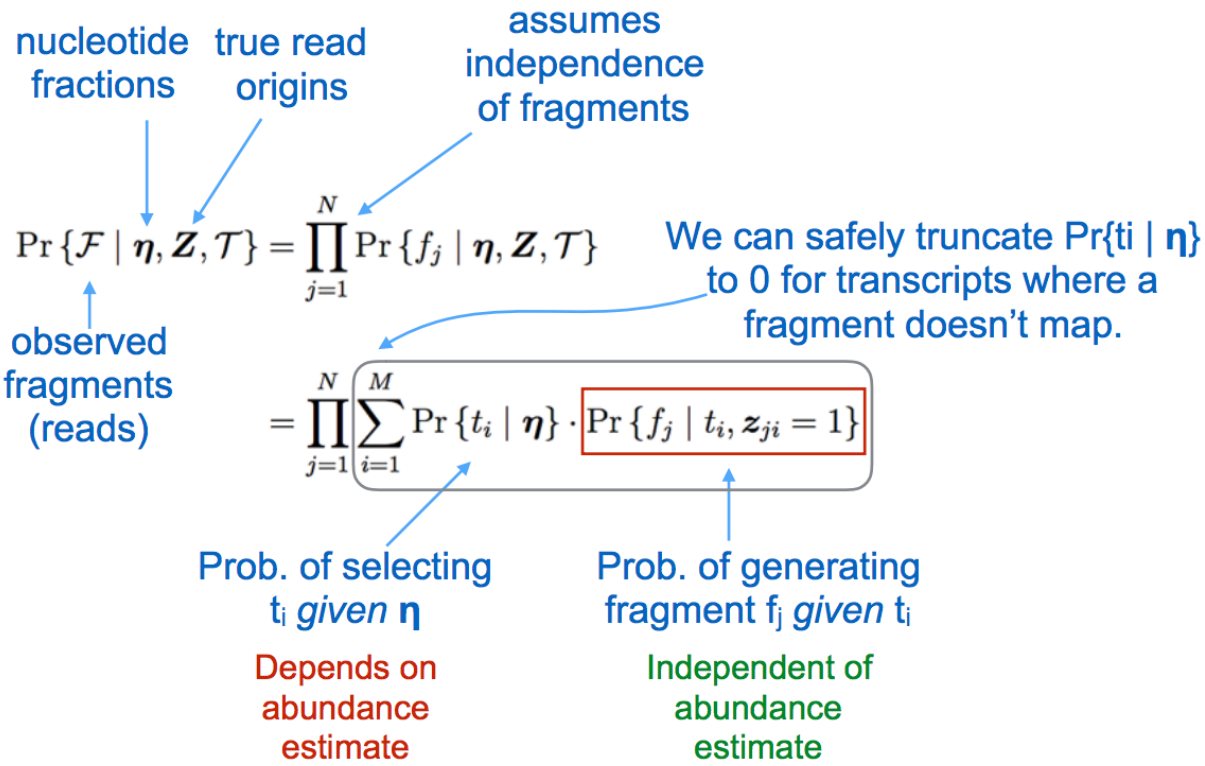
(Estimated) Number of Reads

These are what are used (after normalization) for differential expression. Why can’t we use TPM?

Gene Expression

http://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

A probabilistic view of RNA-Seq quantification



We want to find the values of **η** that ***maximize*** this probability.
We can do this (at least locally) using the EM algorithm.

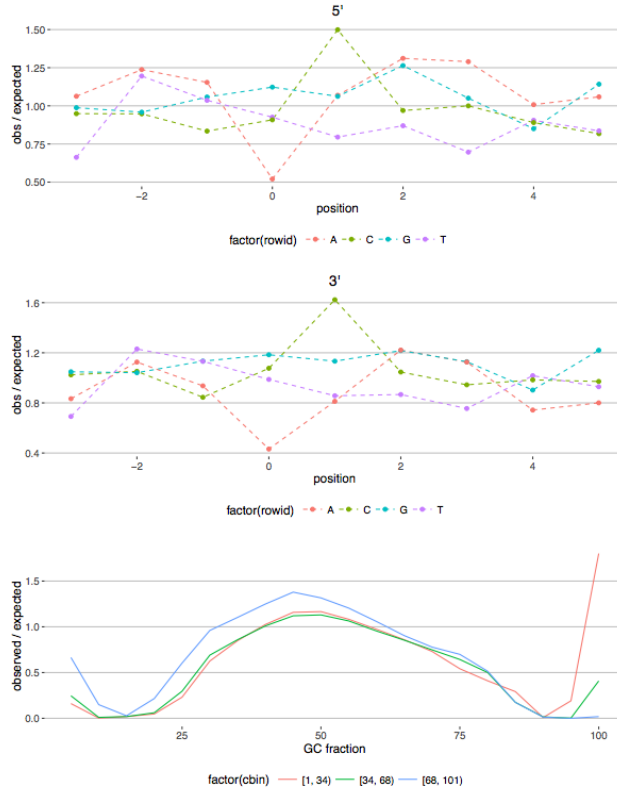
Biases abound in RNA-seq data

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.