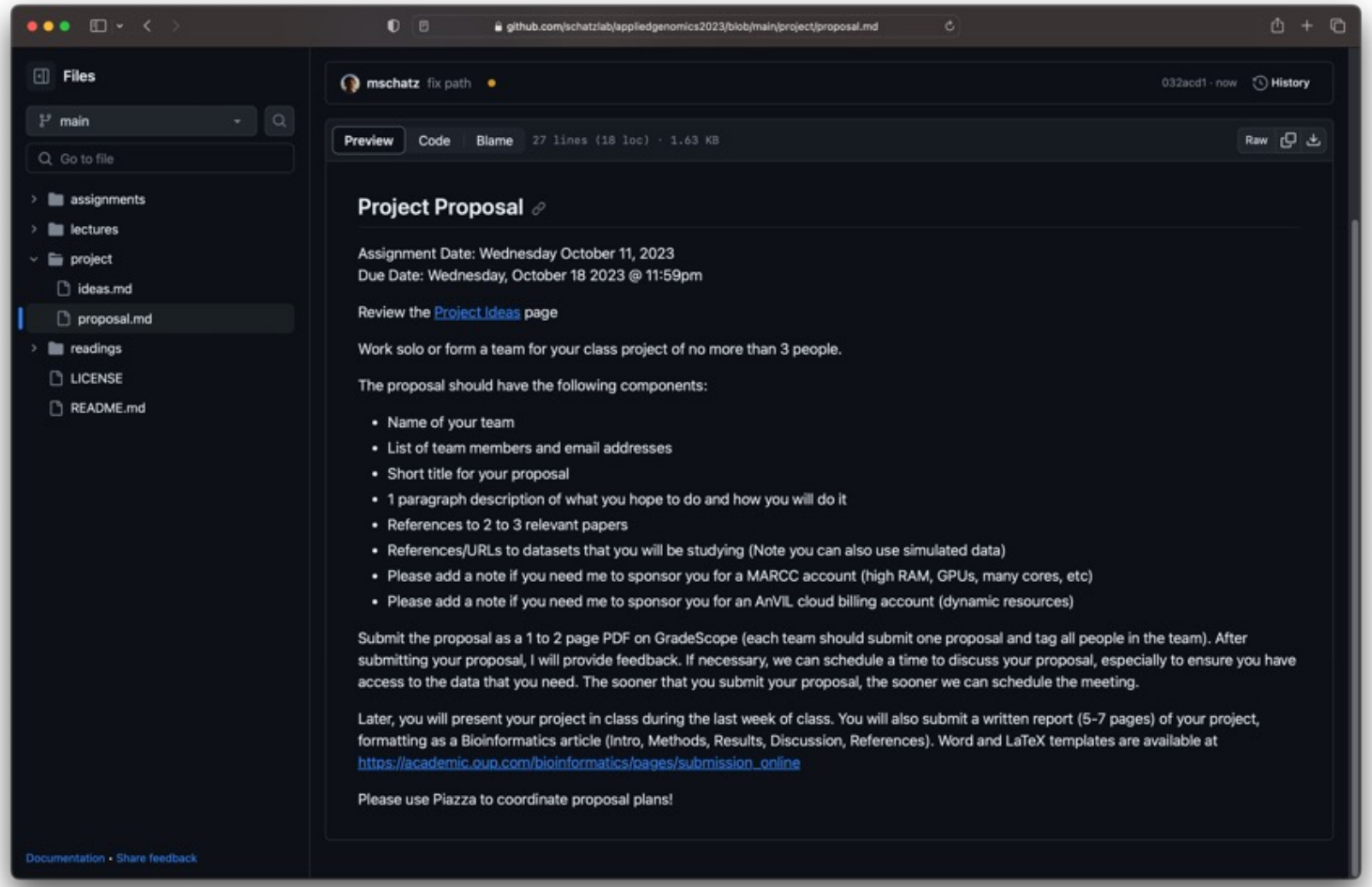
RNAseq (pt2)

Michael Schatz

October 25, 2023

Lecture 17. Applied Comparative Genomics

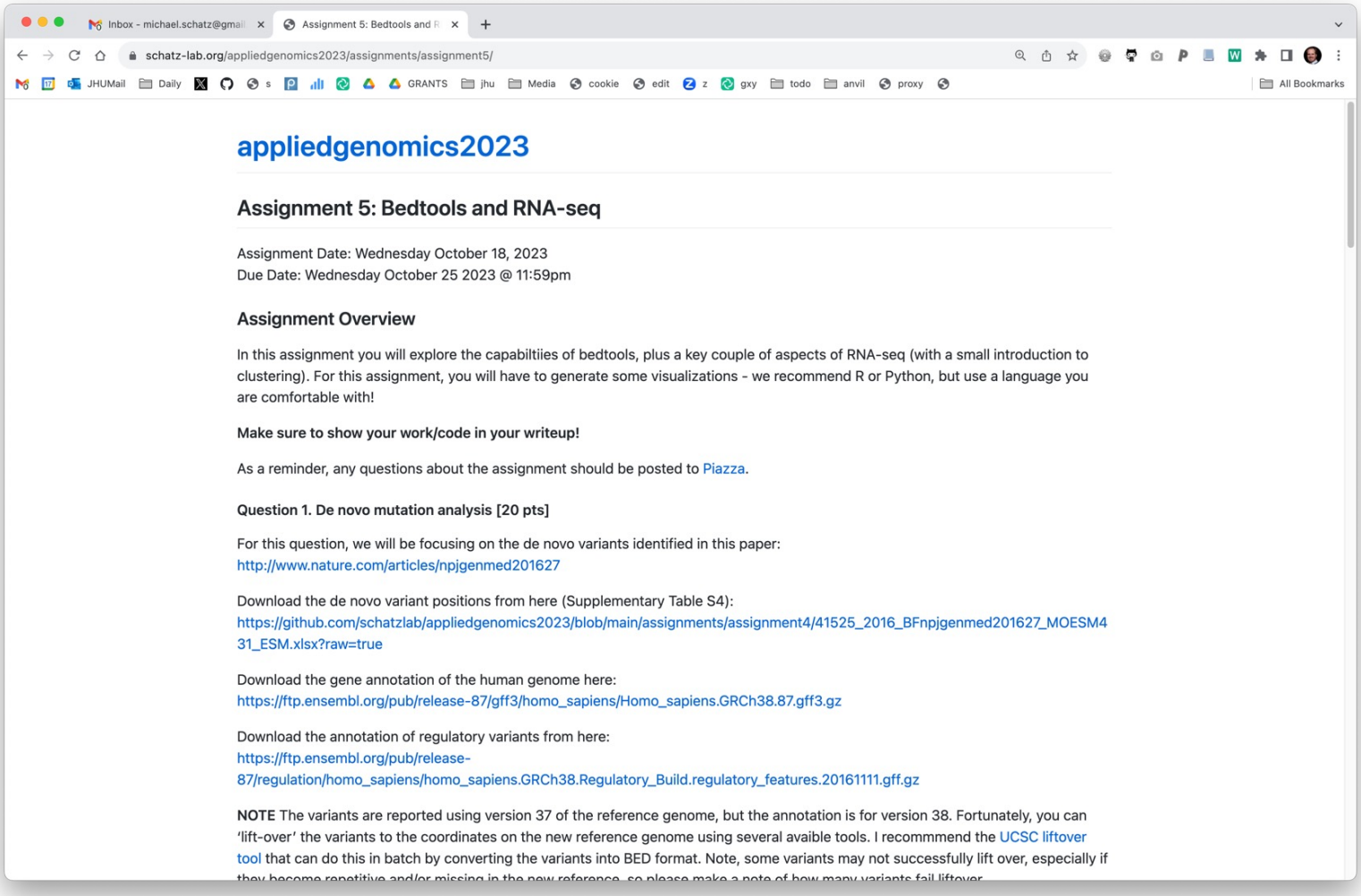
Project Proposal
Due Wednesday Oct 18 by 11:59pm



https://github.com/schatzlab/appliedgenomics2023/blob/main/project/proposal.md

Check Piazza for questions!

Assignment 5
Due: Wednesday Oct 25, 2023 by 11:59pm



schatz-lab.org/appliedgenomics2023/assignments/assignment5/

appliedgenomics2023

Assignment 5: Bedtools and RNA-seq

Assignment Date: Wednesday October 18, 2023
Due Date: Wednesday October 25 2023 @ 11:59pm

Assignment Overview

In this assignment you will explore the capabiltiies of bedtools, plus a key couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. De novo mutation analysis [20 pts]

For this question, we will be focusing on the de novo variants identified in this paper:
http://www.nature.com/articles/npjgenmed201627

Download the de novo variant positions from here (Supplementary Table S4):
https://github.com/schatzlab/appliedgenomics2023/blob/main/assignments/assignment4/41525_2016_BFnpjgenmed201627_MOESM431_ESM.xlsx?raw=true

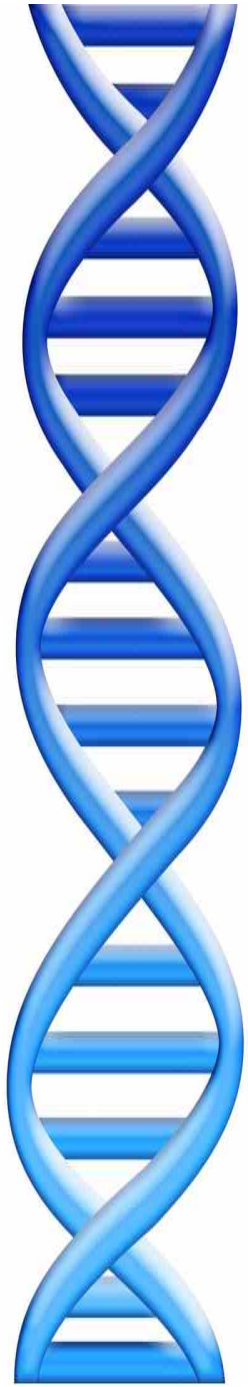
Download the gene annotation of the human genome here:
https://ftp.ensembl.org/pub/release-87/gff3/homo_sapiens/Homo_sapiens.GRCh38.87.gff3.gz

Download the annotation of regulatory variants from here:
https://ftp.ensembl.org/pub/release-87/regulation/homo_sapiens/homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20161111.gff.gz

NOTE The variants are reported using version 37 of the reference genome, but the annotation is for version 38. Fortunately, you can 'lift-over' the variants to the coordinates on the new reference genome using several avaible tools. I recommmend the UCSC liftover tool that can do this in batch by converting the variants into BED format. Note, some variants may not successfully lift over, especially if they become repetitive and/or missing in the new reference, so please make a note of how many variants fail liftover

https://schatz-lab.org/appliedgenomics2023/assignments/assignment5/

Check Piazza for questions!



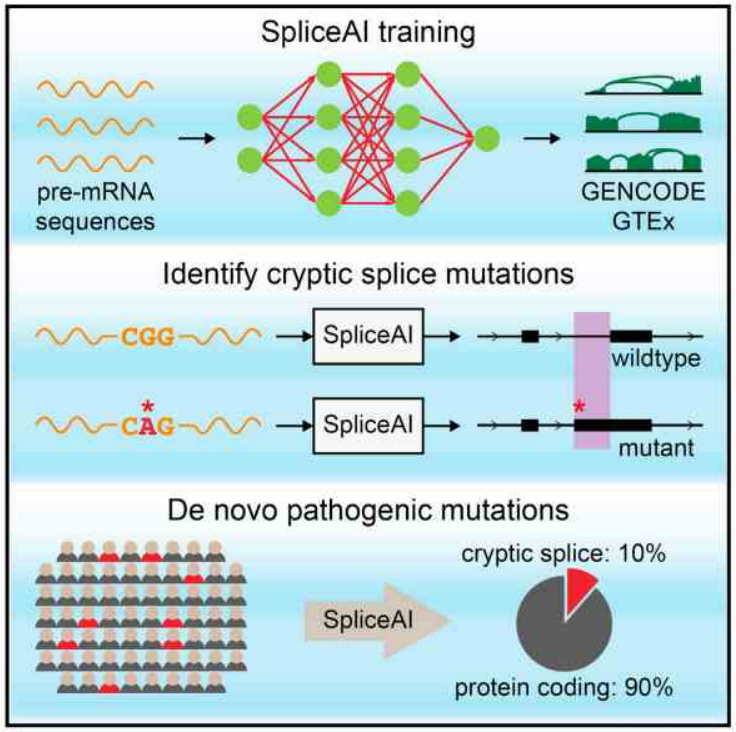
Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Cell

Predicting Splicing from Primary Sequence with Deep Learning

Graphical Abstract



Highlights

- SpliceAI, a 32-layer deep neural network, predicts splicing from a pre-mRNA sequence
- 75% of predicted cryptic splice variants validate on RNA-seq
- Cryptic splicing may yield ~10% of pathogenic variants in neurodevelopmental disorders
- Cryptic splice variants frequently give rise to alternative splicing

Authors

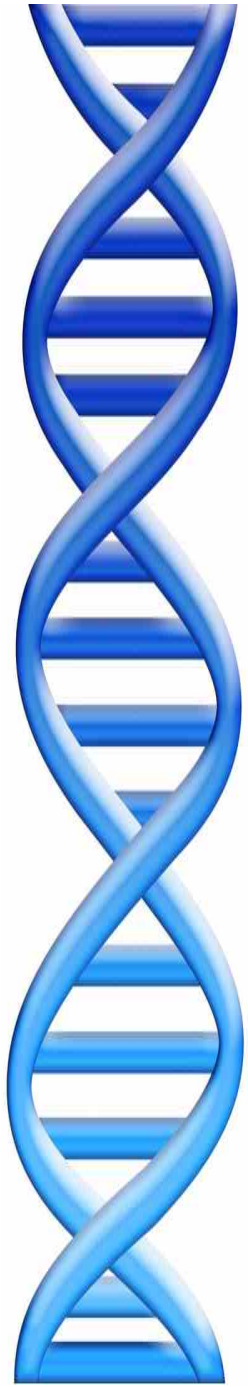
Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, ..., Serafim Batzoglou, Stephan J. Sanders, Kyle Kai-How Farh

Correspondence

kfarh@illumina.com

In Brief

A deep neural network precisely models mRNA splicing from a genomic sequence and accurately predicts noncoding cryptic splice mutations in patients with rare genetic diseases.

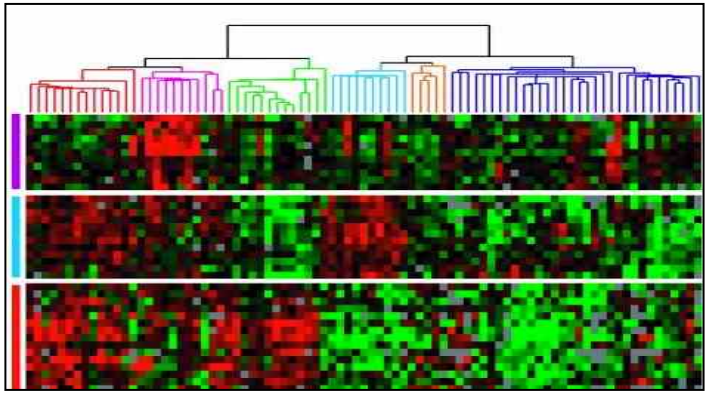


Outline

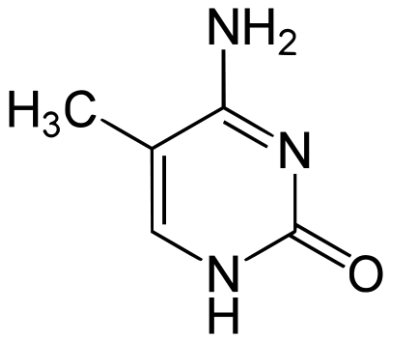
1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

*-seq in 4 short vignettes

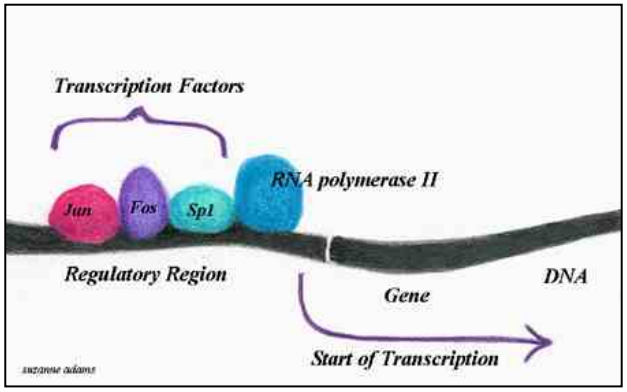
RNA-seq



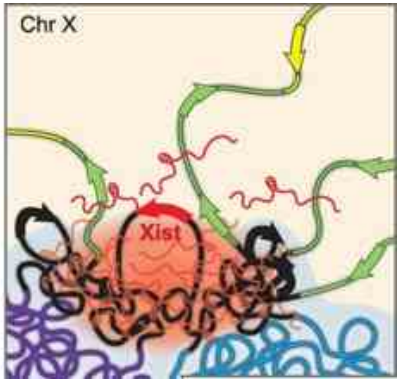
Methyl-seq



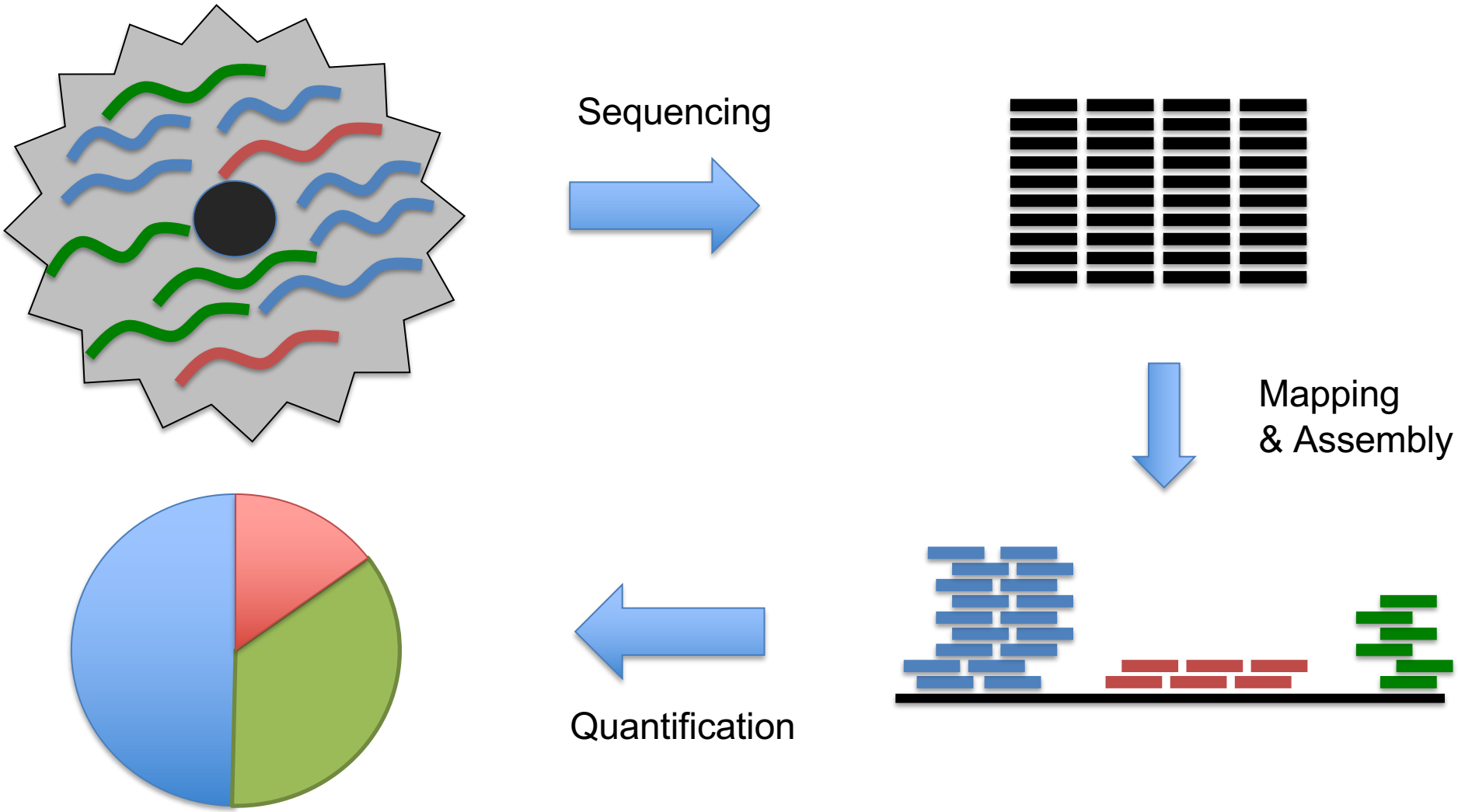
ChIP-seq



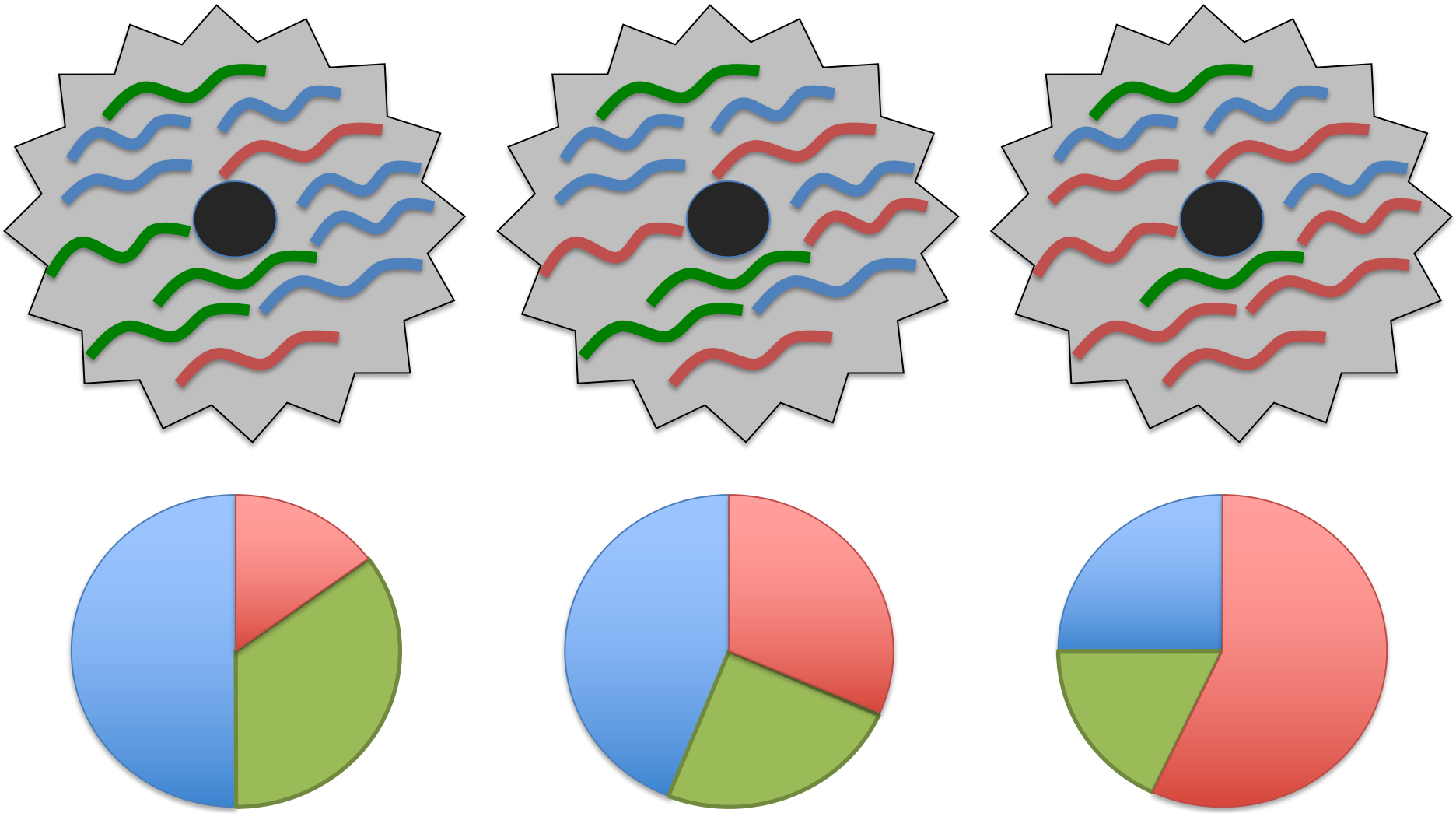
Hi-C



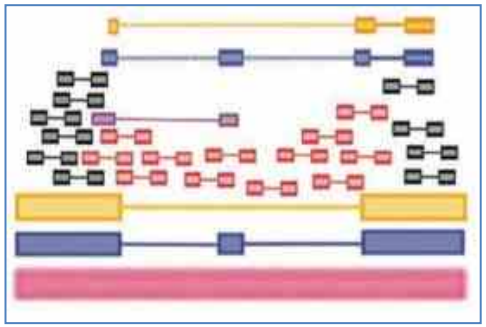
RNA-seq Overview



RNA-seq Overview



RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

RNA-Seq Approaches

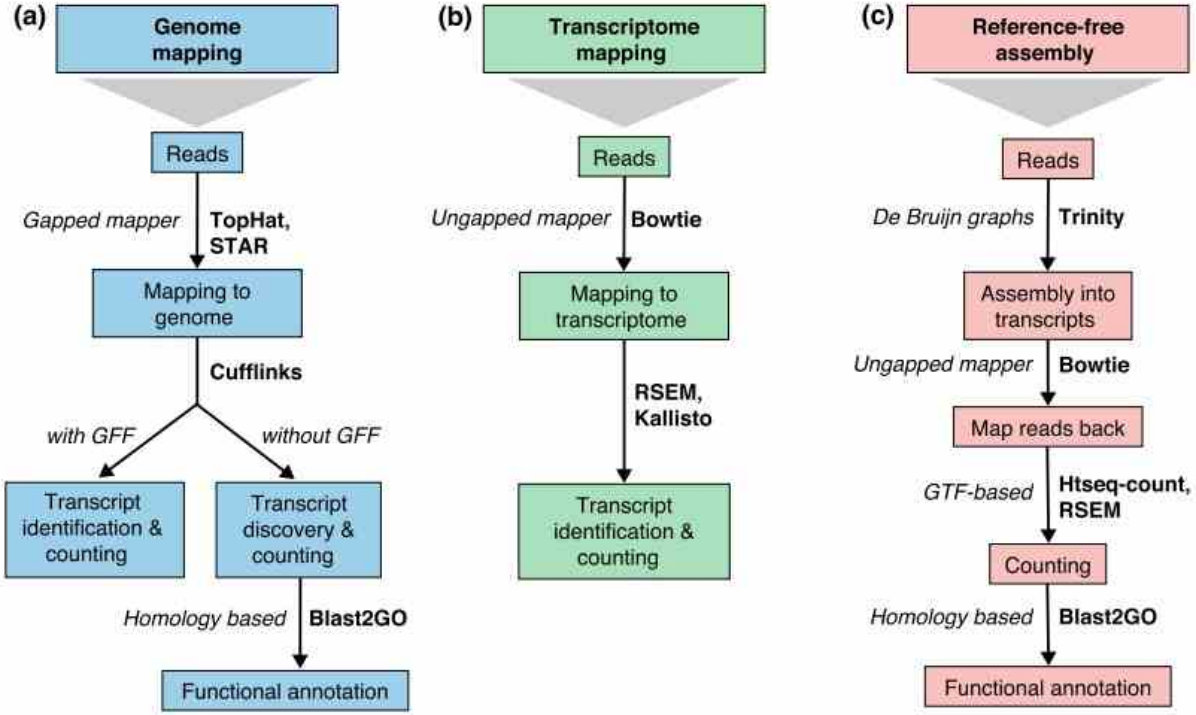


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and re…ovel) transcript discovery and quantification can proceed with or without an ann… no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner. Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads n…o contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further ar…owed by the functional annotation of the novel transcripts as in (**a**). Representative software that can be used a…eated in *bold text*. Abbreviations: *GFF* General Feature Format, *GTF* gene transfer format, *RSEM* RNA-Seq by Expectation Maximization

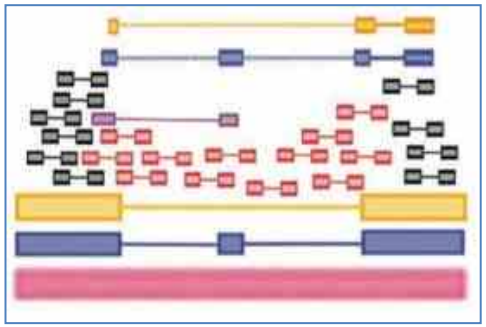
Which approach should we use?

It depends….

A survey of best practices for RNA-seq data analysis

Conesa et al (2016) Genome Biology. doi 10.1186/s13059-016-0881-8

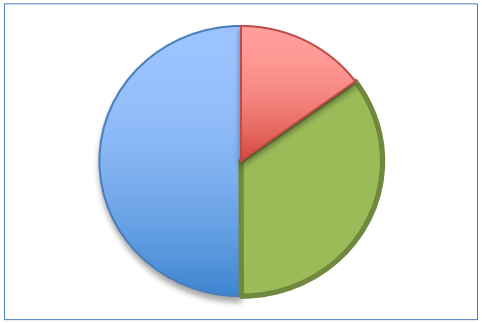
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

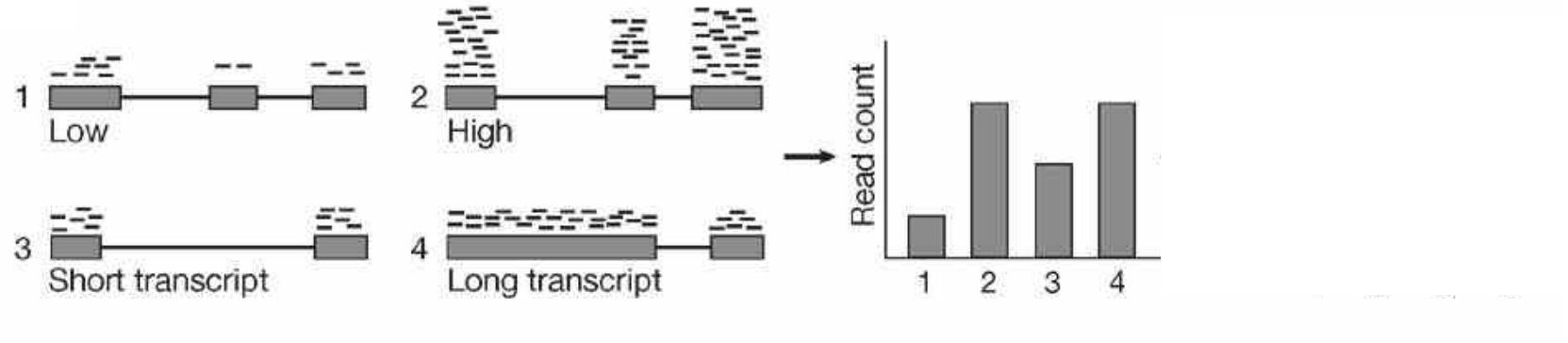
Solution: Use a spliced aligner, and assemble isoforms

TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

RPKM, FPKM, TPM

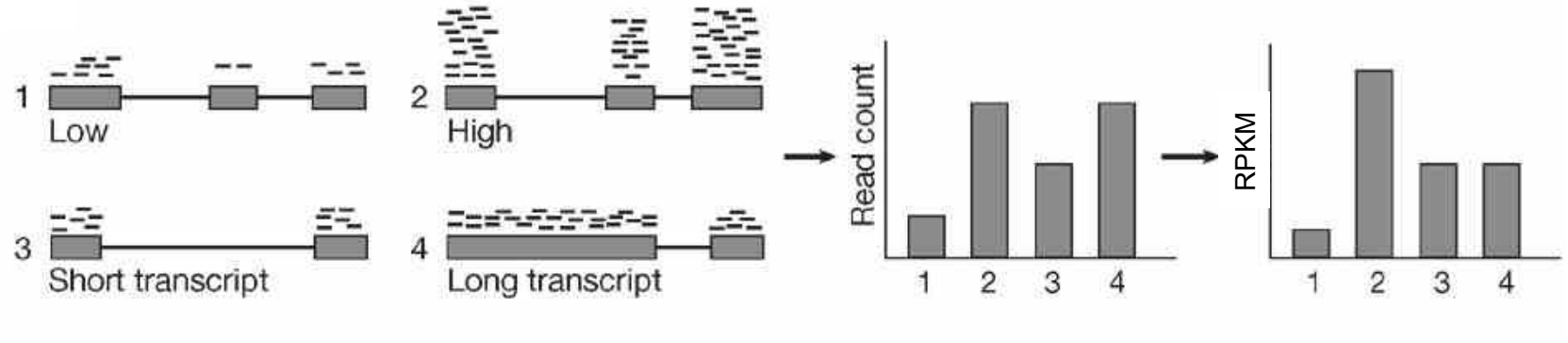


Counting Reads that align to a gene DOESN'T work!

- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)

RPKM, FPKM, TPM



Counting Reads that align to a gene DOESN'T work!

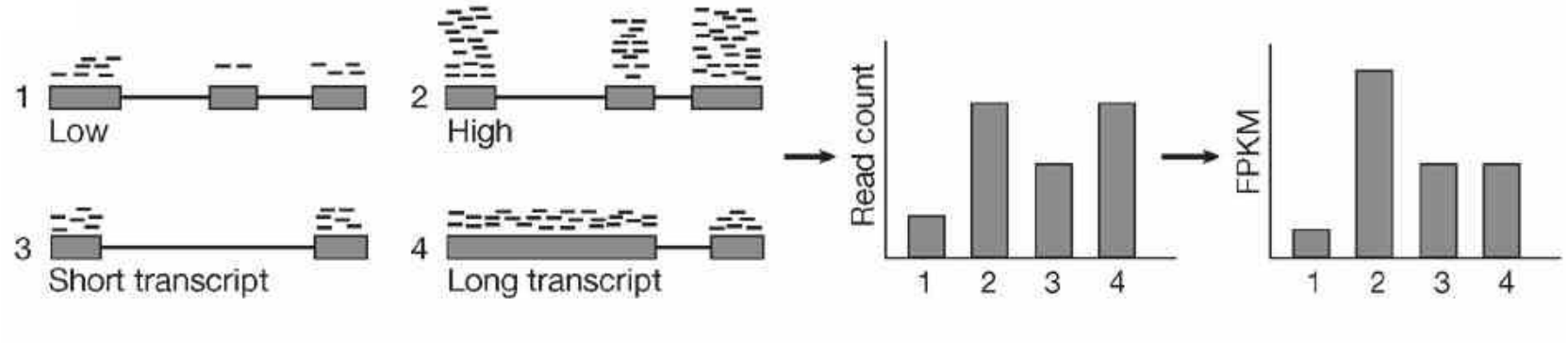
- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)

(Count reads aligned to gene) / (length of gene in kilobases) / (# millions of read mapped)

=> Wait a second, reads in a pair arent independent!

RPKM, FPKM, TPM



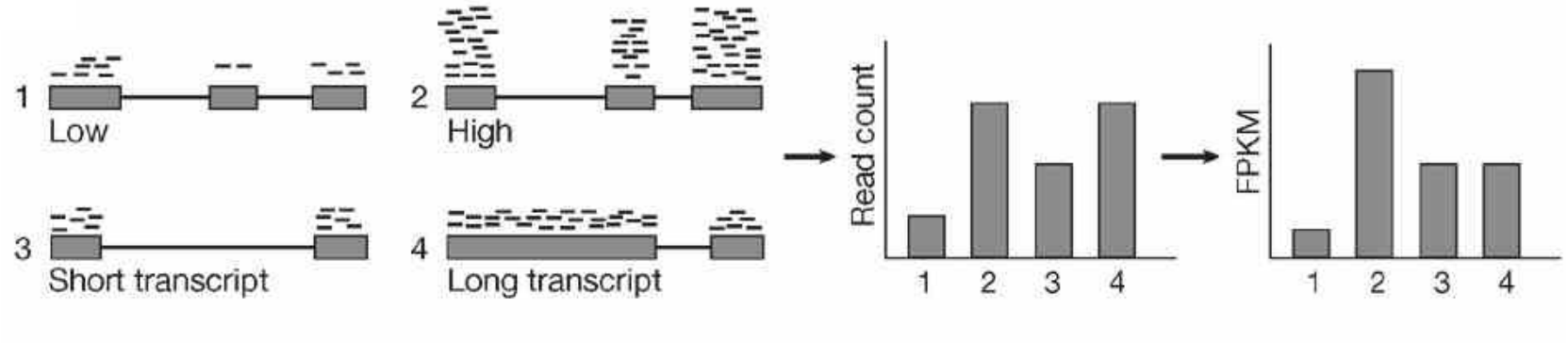
Counting Reads that align to a gene DOESN'T work!
- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)
=> Wait a second, reads in a pair arent independent!

2. FPKM: Fragments Per Kilobase of Exon Per Million Reads Mapped (Trapnell et al, 2010)
$\Rightarrow$ Does a much better job with short exons & short genes by boosting coverage

$\Rightarrow$ Wait a second, FPKM depends on the average transcript length!

RPKM, FPKM, TPM



Counting Reads that align to a gene DOESN'T work!

- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

1. RPKM: Reads Per Kilobase of Exon Per Million Reads Mapped (Mortazavi et al, 2008)

=> Wait a second, reads in a pair arent independent!

2. FPKM: Fragments Per Kilobase of Exon Per Million Reads Mapped (Trapnell et al, 2010)

=> Wait a second, FPKM depends on the average transcript length!

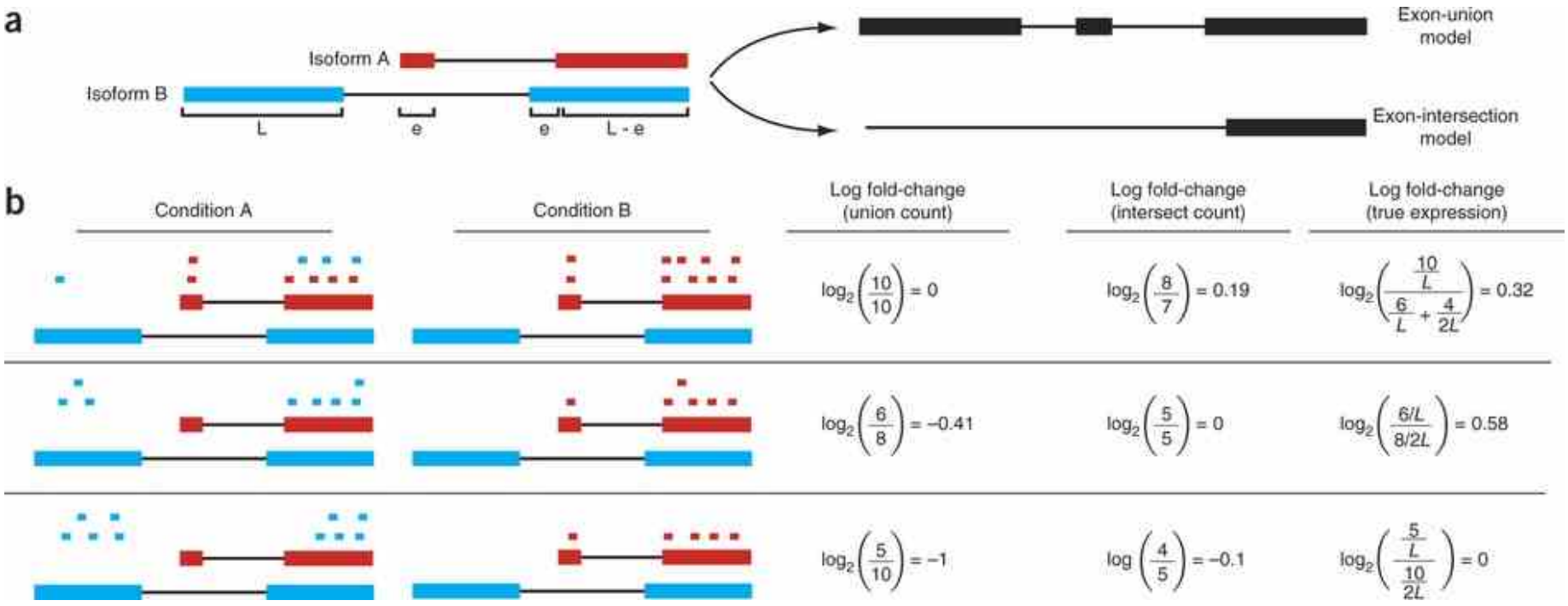
3. TPM: Transcripts Per Million (Li et al, 2011)

$\Rightarrow$ If you were to sequence one million full length transcripts, TPM is the number of transcripts you would have seen of type i, given the abundances of the other transcripts in your sample

=> Recommend you use TPM for all analysis, easy to compute given FPKM

$$\mathrm{TPM}_i = \left(\frac{\mathrm{FPKM}_i}{\sum_j \mathrm{FPKM}_j} \right) \cdot 10^6$$

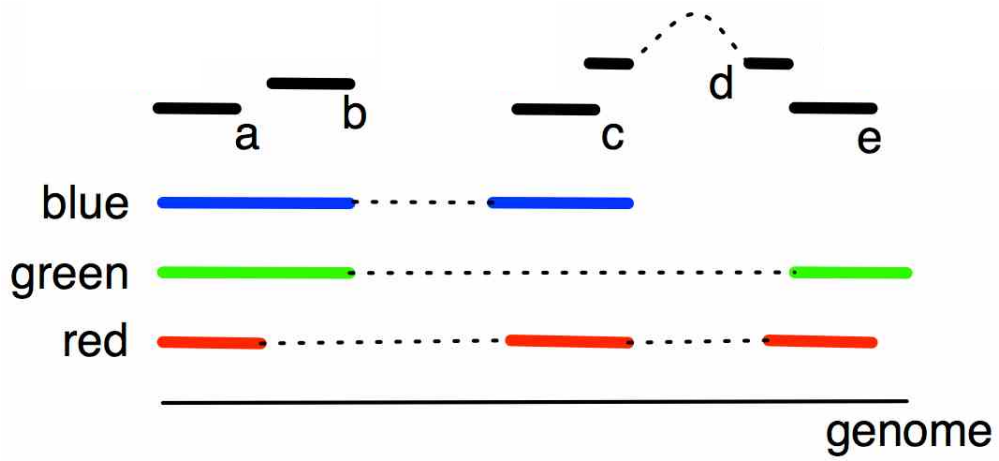
Gene or Isoform Quantification?



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

Differential analysis of gene regulation at transcript resolution with RNA-seq
Trapnell et al (2013) Nature Biotechnology 31, 46–53. doi:10.1038/nbt.2450

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length.
Our initial expectation is all 3 isoforms are equally expressed

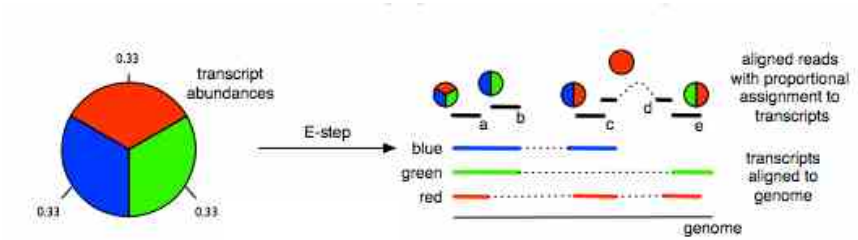
There are five reads (a,b,c,d,e) mapping to the gene.

- Read a maps to all three isoforms
- Read d only to red
- Reads b,c,e map to each of the three pairs of isoforms.

What is the most likely expression level of each isoform?

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



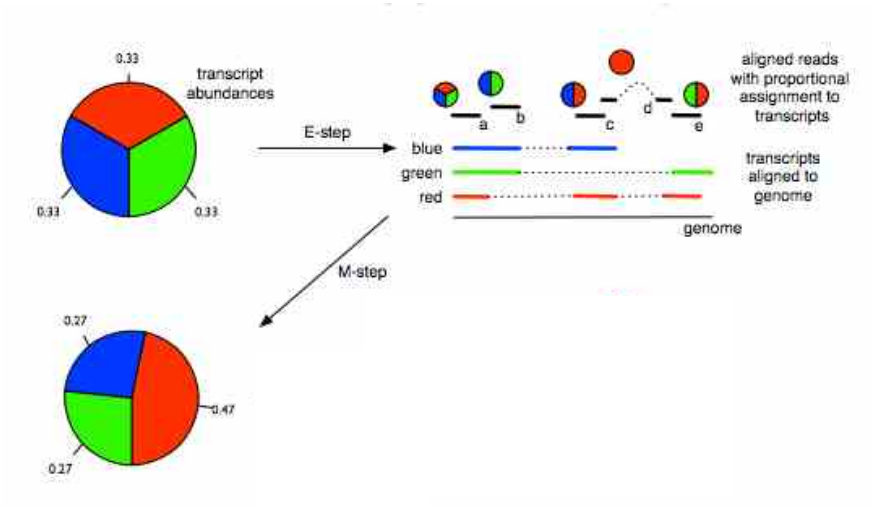
The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:

red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)

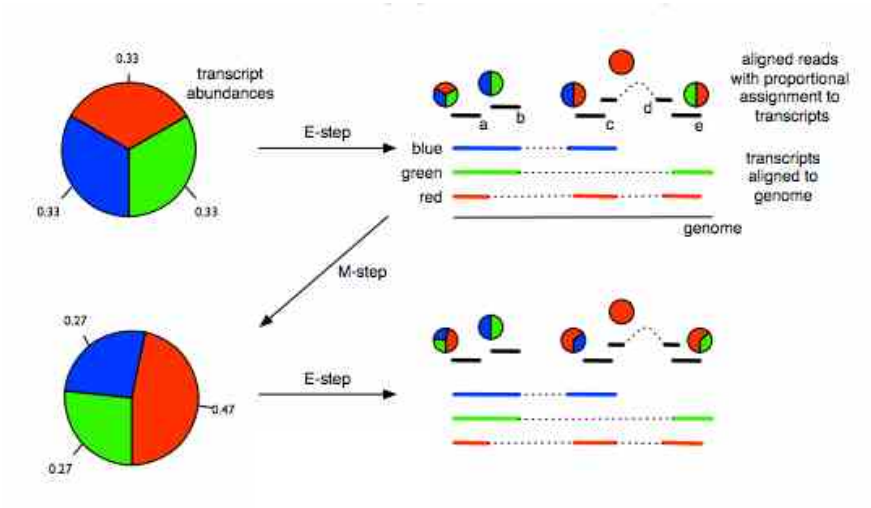
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

Models for transcript quantification from RNA-seq

Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

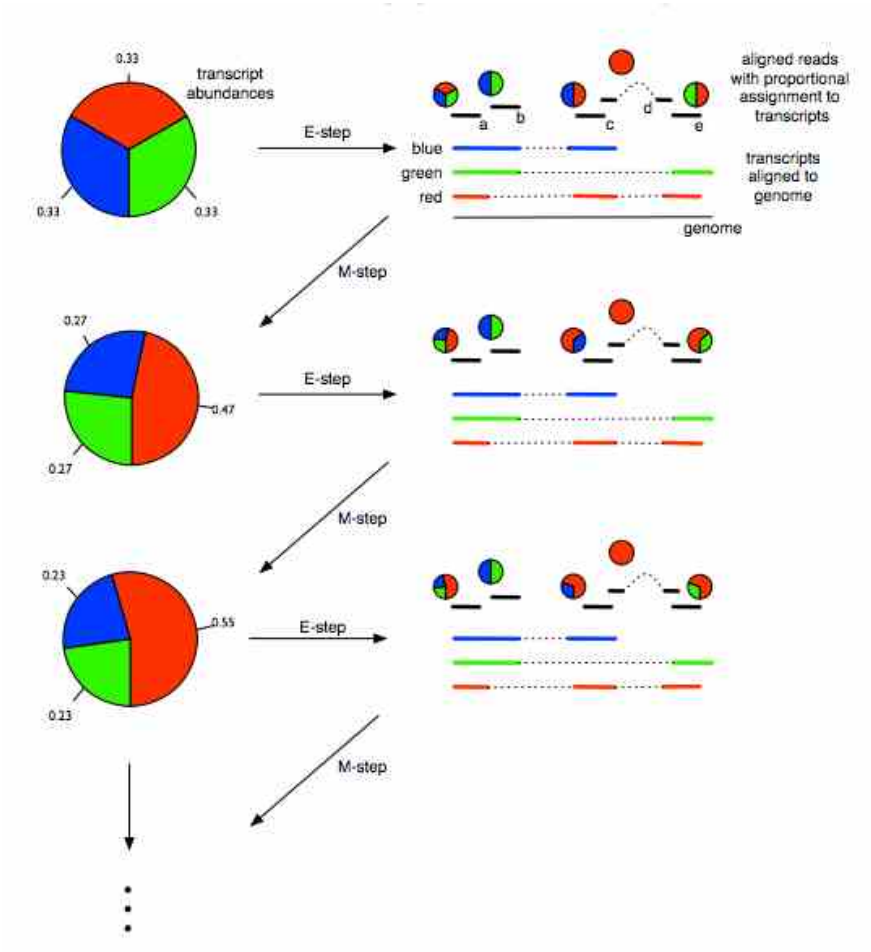
During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:
red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

Repeat until convergence!

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Multi-mapping? Isoform ambiguity? Expectation Maximization to the Rescue



The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

There are five reads (a,b,c,d,e) mapping to the gene. Read a maps to all three isoforms, read d only to red, and the other three (reads b,c,e) to each of the three pairs of isoforms.

During the expectation (E) step reads are proportionately assigned to transcripts according to the (current) isoform abundances (RGB): a=(.33,.33,.33), b=(0,.5,.5), c=(.5,.5), d=(1,0,0), e=(.5,.5,0)

Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:

red: 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)

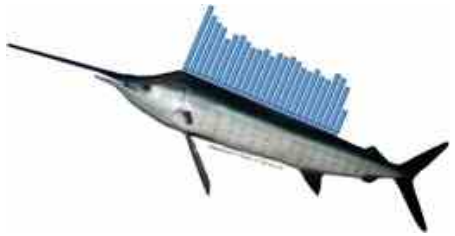
blue: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

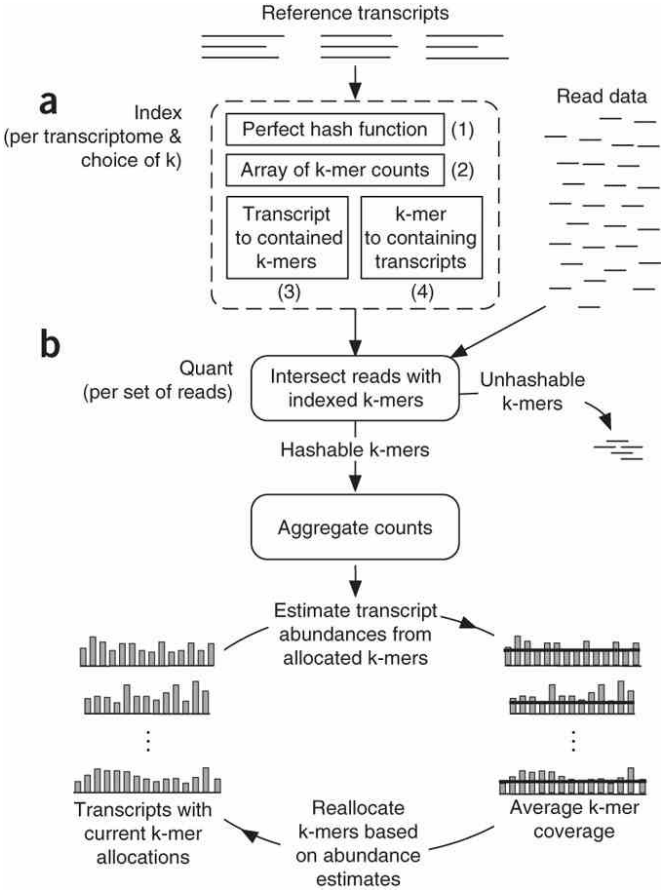
green: 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)

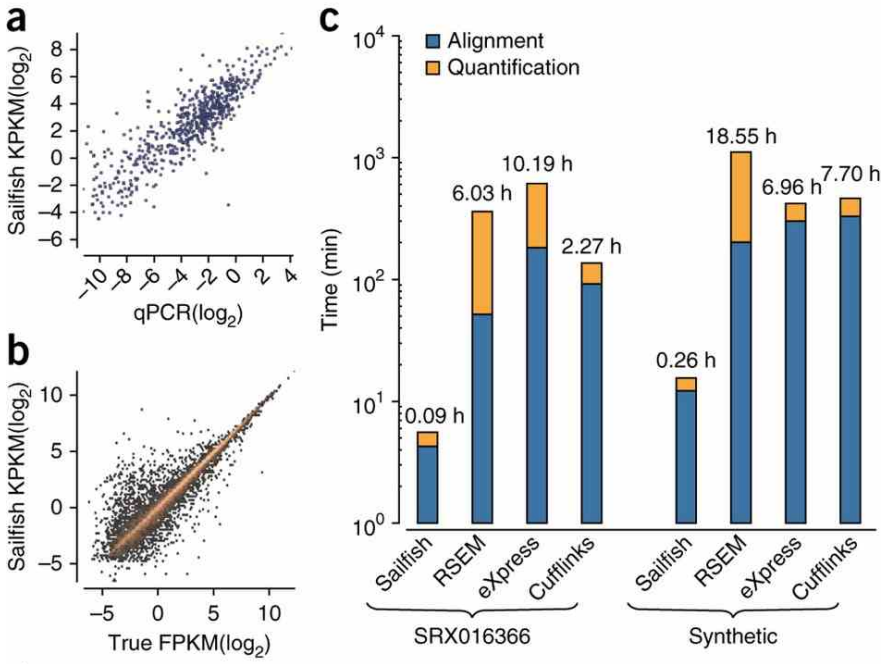
Repeat until convergence!

Models for transcript quantification from RNA-seq
Pachter, L (2011) arXiv. 1104.3889 [q-bio.GN]

Sailfish: Fast & Accurate RNA-seq Quantification





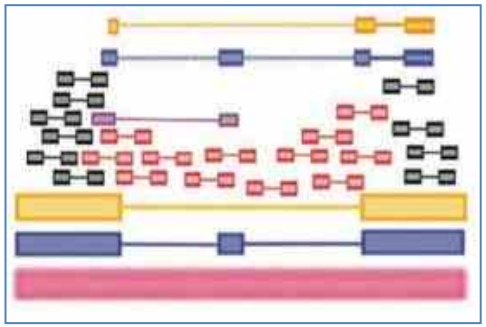


d

	Human brain tissue				Synthetic			
	Sailfish	RSEM	eXpress	Cufflinks	Sailfish	RSEM	eXpress	Cufflinks
Pearson	0.85	0.82	0.85	0.85	0.96	0.96	0.95	0.94
Spearman	0.84	0.80	0.85	0.85	0.76	0.77	0.77	0.75
RMSE	–	–	–	–	6.06	8.90	8.83	10.05
medPE	–	–	–	–	6.50	12.48	14.06	12.42

Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms
Patro et al (2014) Nature Biotechnology 32, 462–464 doi:10.1038/nbt.2862

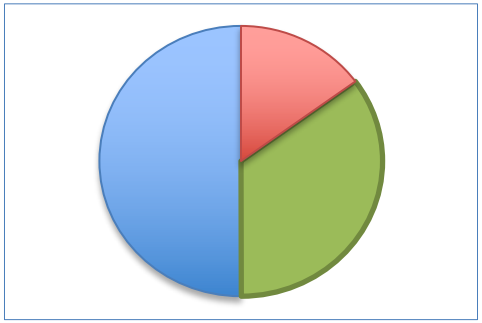
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

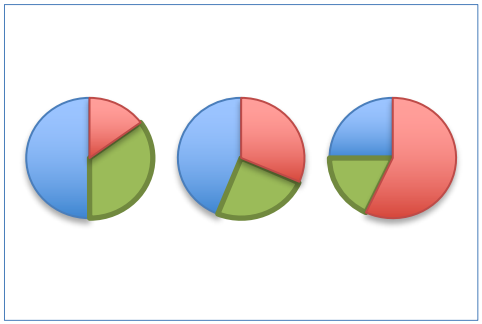
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

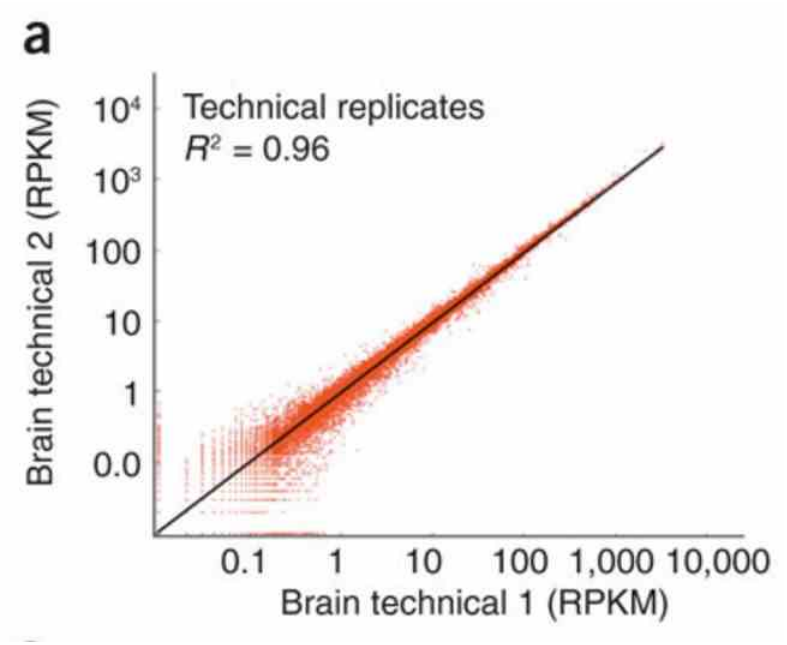
Solution: Infer underlying abundances (e.g. TPM)

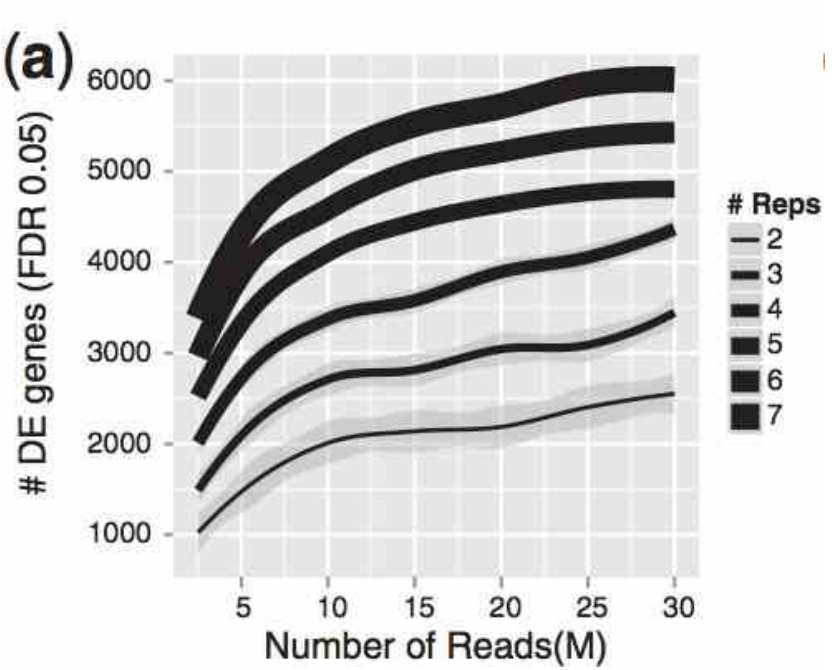
Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515



Challenge 3: Transcript abundances are stochastic

How Many Replicates?



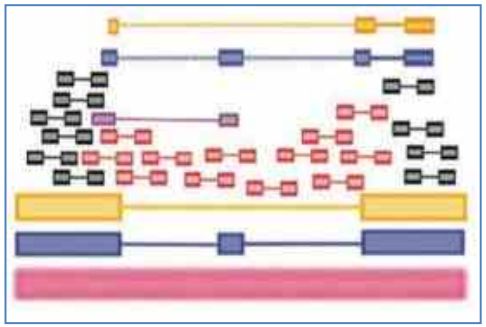


Why don't we have perfect replicates?

Mapping and quantifying mammalian transcriptomes by RNA-Seq
Mortazavi et al (2008) Nature Methods. 5, 62-628

RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) Bioinformatics. doi:10.1093/bioinformatics/btt688

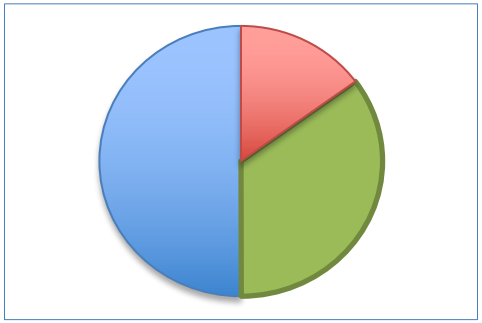
RNA-seq Challenges



Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

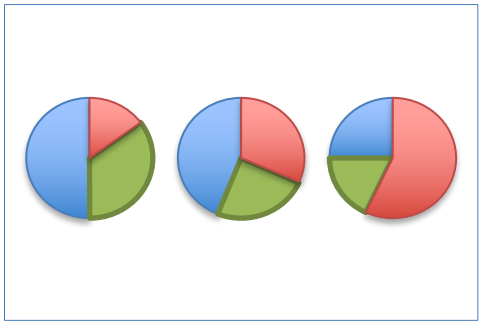
TopHat: discovering spliced junctions with RNA-Seq.
Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111



Challenge 2: Read Count != Transcript abundance

Solution: Infer underlying abundances (e.g. TPM)

Transcript assembly and quantification by RNA-seq
Trapnell et al (2010) *Nat. Biotech.* 25(5): 511-515

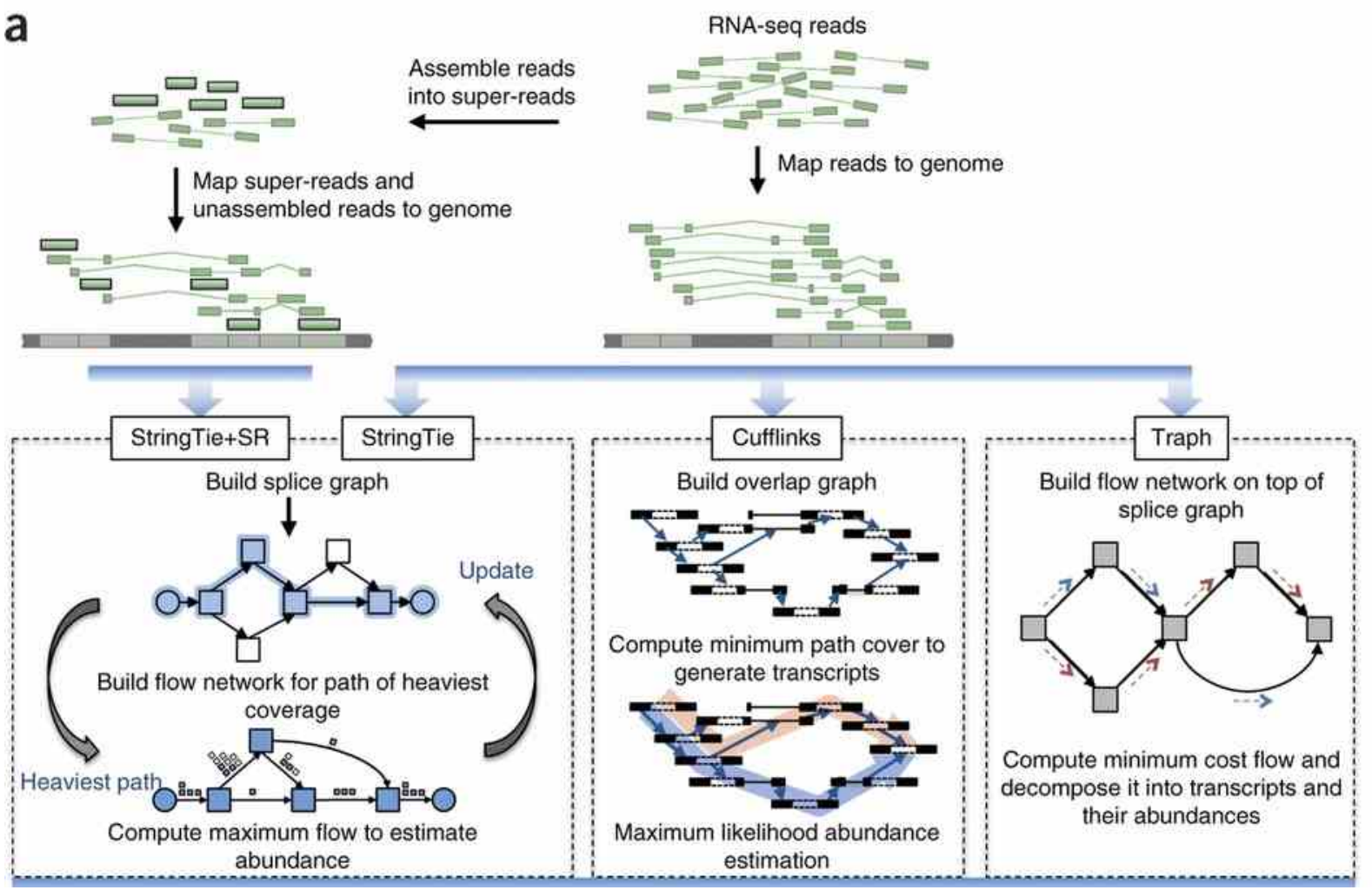


Challenge 3: Transcript abundances are stochastic

Solution: Replicates, replicates, and more replicates

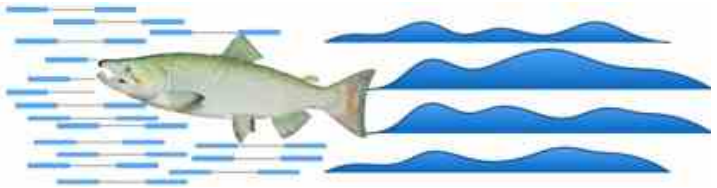
RNA-seq differential expression studies: more sequence or more replication?
Liu et al (2013) *Bioinformatics*. doi:10.1093/bioinformatics/btt688

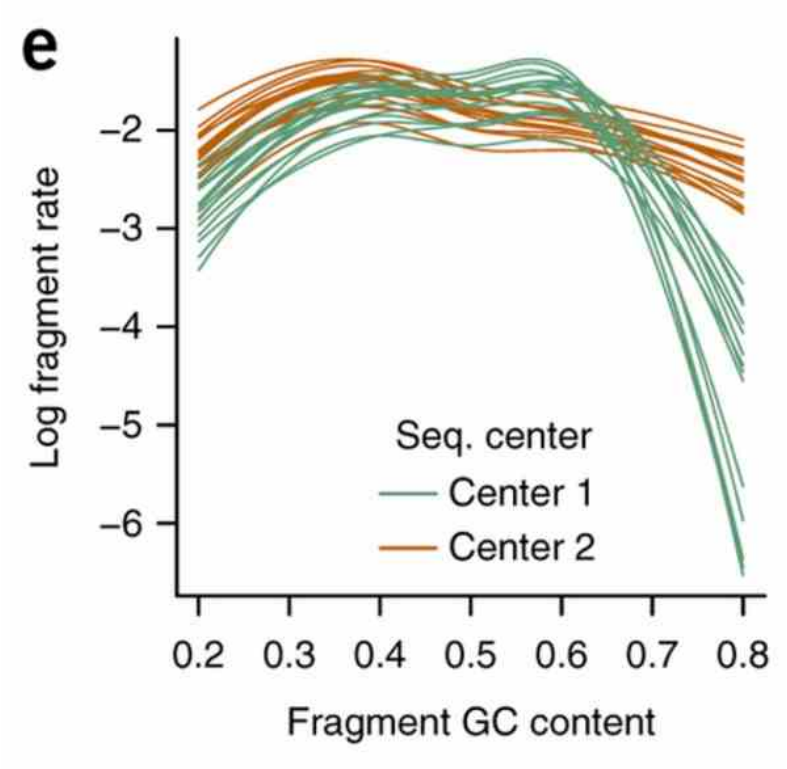
Isoform Quantification Approaches

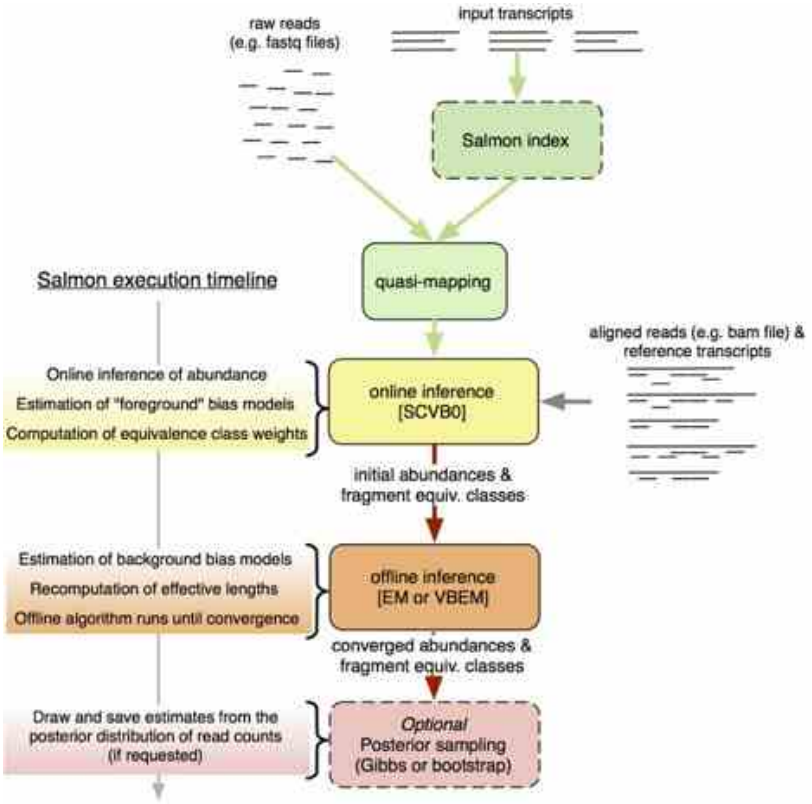


StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
Pertea M, et al. (2015) Nature Biotechnology. doi: 10.1038/nbt.3122.

Salmon: The ultimate RNA-seq Pipeline?







Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation
Love et al (2016) Nature Biotechnology 34, 1287–1291 (2016) doi:10.1038/nbt.3682

Salmon provides fast and bias-aware quantification of transcript expression
Patro et al (2017) Nature Methods (2017) doi:10.1038/nmeth.4197

Annotation Summary

- Three major approaches to annotate a genome
  1. Alignment:
     - Does this sequence align to any other sequences of known function?
     - Great for projecting knowledge from one species to another
  2. Prediction:
     - Does this sequence statistically resemble other known sequences?
     - Potentially most flexible but dependent on good training data
  3. Experimental:
     - Lets test to see if it is transcribed/methylated/bound/etc
     - Strongest but expensive and context dependent

- Many great resources available
  - Learn to love the literature and the databases
  - Standard formats let you rapidly query and cross reference
  - Google is your number one resource ☺