

Lecture 14. RNAseq

Michael Schatz

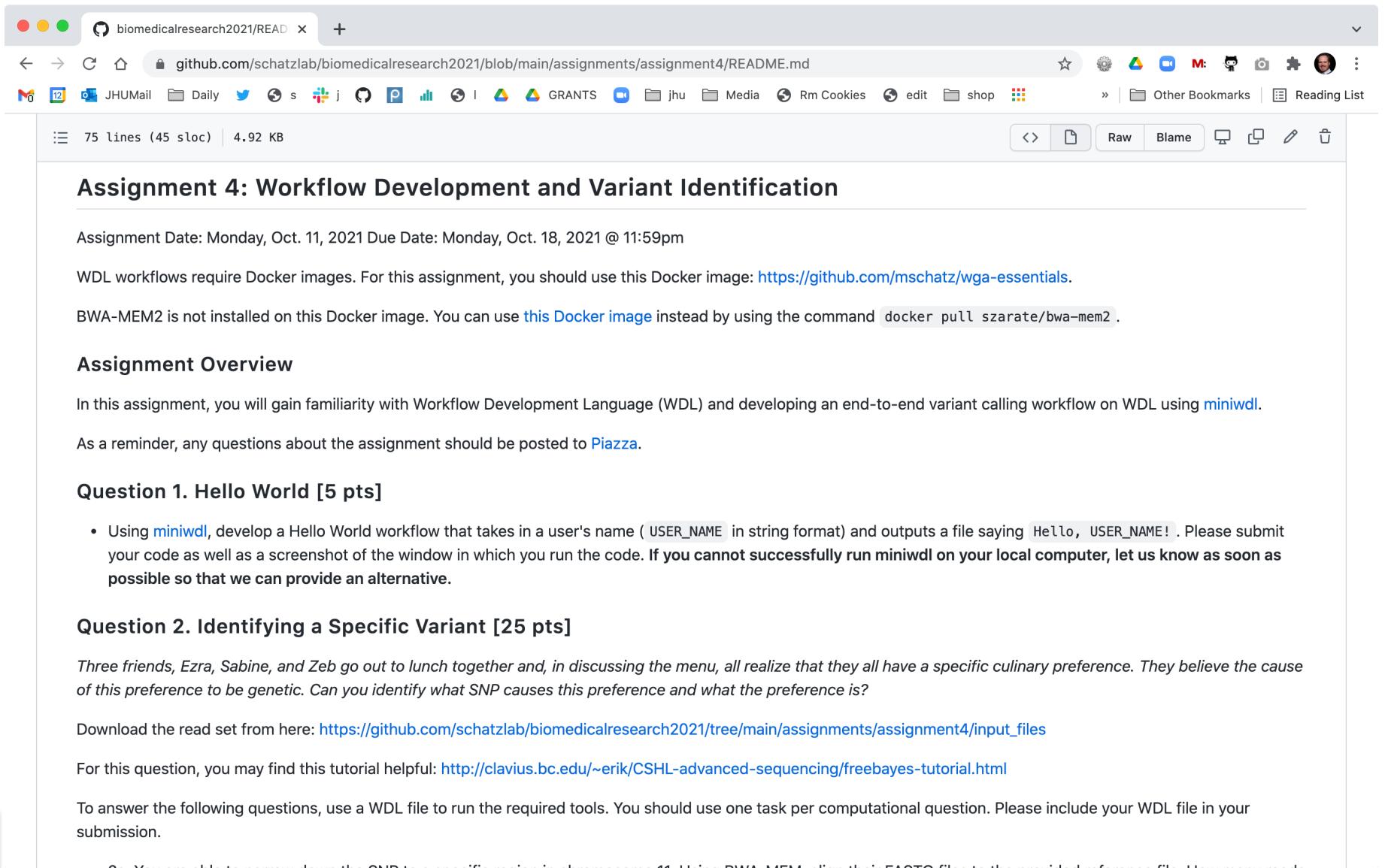
October 18, 2021

Advanced Biomedical Research



Assignment 4: WDLs

Due Oct 18 @ 11:59pm



A screenshot of a web browser window displaying a GitHub README file. The title bar shows the URL github.com/schatzlab/biomedicalresearch2021/blob/main/assignments/assignment4/README.md. The page content is as follows:

Assignment 4: Workflow Development and Variant Identification

Assignment Date: Monday, Oct. 11, 2021 Due Date: Monday, Oct. 18, 2021 @ 11:59pm

WDL workflows require Docker images. For this assignment, you should use this Docker image: <https://github.com/mschatz/wga-essentials>.

BWA-MEM2 is not installed on this Docker image. You can use [this Docker image](#) instead by using the command `docker pull szarate/bwa-mem2`.

Assignment Overview

In this assignment, you will gain familiarity with Workflow Development Language (WDL) and developing an end-to-end variant calling workflow on WDL using [miniwdl](#).

As a reminder, any questions about the assignment should be posted to [Piazza](#).

Question 1. Hello World [5 pts]

- Using [miniwdl](#), develop a Hello World workflow that takes in a user's name (`USER_NAME` in string format) and outputs a file saying `Hello, USER_NAME!`. Please submit your code as well as a screenshot of the window in which you run the code. If you cannot successfully run miniwdl on your local computer, let us know as soon as possible so that we can provide an alternative.

Question 2. Identifying a Specific Variant [25 pts]

Three friends, Ezra, Sabine, and Zeb go out to lunch together and, in discussing the menu, all realize that they all have a specific culinary preference. They believe the cause of this preference to be genetic. Can you identify what SNP causes this preference and what the preference is?

Download the read set from here: https://github.com/schatzlab/biomedicalresearch2021/tree/main/assignments/assignment4/input_files

For this question, you may find this tutorial helpful: <http://clavius.bc.edu/~erik/CSHL-advanced-sequencing/freebayes-tutorial.html>

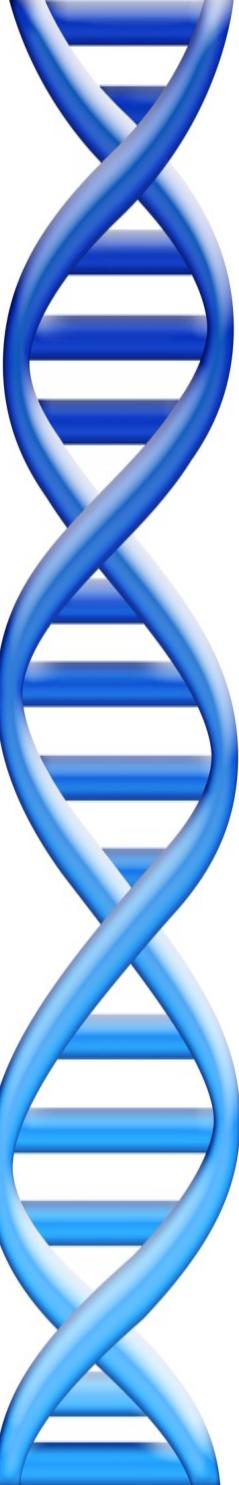
To answer the following questions, use a WDL file to run the required tools. You should use one task per computational question. Please include your WDL file in your submission.

<https://github.com/schatzlab/biomedicalresearch2021>

Goal: Genome Annotations

aatgcatgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcggctatgctaatt
gcatgcggctatgcaaggctggatccgatgactatgctaagctggatccgatgacaatgcattgcggctatgctaatt
aatgaatggtcttggattttaccttggaaatgtctaagctggatccgatgacaatgcattgcggctatgctaatt
tggtcttggattttaccttggaaatgtctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcg
gctatgctaattgcattgcggctatgcaaggctggatccgatgactatgctaagctgcggctatgctaattgcattgcg
gctatgctaagctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatcc
gctatgctaattgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcgg
atgctaattgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgacaatgcattgcgg
gctatgctaagctggatccgatgacaatgcattgcggctatgctaagctggatccgatgacaatgcattgcgg
atgactatgctaagctgcggctatgctaattgcattgcggctatgctaagctcatgcggctatgctaagctgg
gcatgcggctatgctaaggctggatccgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagct
ggatccgatgactatgctaagctgcggctatgctaattgcattgcggctatgctaagctcggtatgctaatt
gtcttggattttaccttggaaatgtctaagctggatccgatgacaatgcattgcggctatgctaatt
gatttaccttggaaatgtctaattgcattgcggctatgctaagctggatgcattgcggctatgctaagctgg
cgatgacaatgcattgcggctatgctaattgcattgcggctatgctaagctggatccgatgactatgctaagctgc
gctatgctaattgcattgcggctatgctaagctcatgcgg

Gene!



Outline

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

Basic Local Alignment Search Tool

- Rapidly compare a sequence Q to a database to find all sequences in the database with a score above some cutoff S .
 - Which protein is most similar to a newly sequenced one?
 - Where does this sequence of DNA originate?
- Speed achieved by using a procedure that typically finds “most” matches with scores $> S$.
 - Tradeoff between sensitivity and specificity/speed
 - Sensitivity – ability to find all related sequences
 - Specificity – ability to reject unrelated sequences

Very Similar Sequences

Query: HBA_HUMAN Hemoglobin alpha subunit

Sbjct: HBB_HUMAN Hemoglobin beta subunit

Score = 114 bits (285), Expect = 1e-26

Identities = 61/145 (42%), Positives = 86/145 (59%), Gaps = 8/145 (5%)

Query 2 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-----DLSHGSAQV 55

L+P +K+ V A WGKV + E G EAL R+ + +P T+ +F F D G+ +V

Sbjct 3 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAMGNPKV 60

Query 56 KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRDPVNFKLLSHCLLVTLAHLPA 115

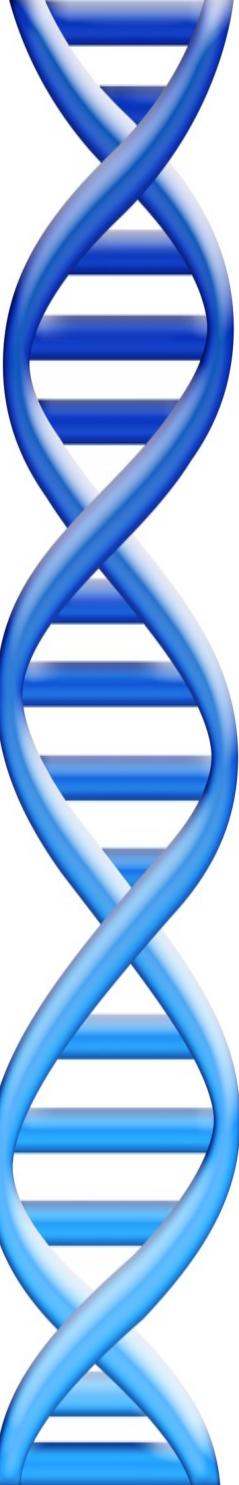
K HGKKV A ++ +AH+D++ + LS+LH KL VDP NF+LL + L+ LA H

Sbjct 61 KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120

Query 116 EFTP AVHASLDKFLASVSTVLTSKY 140

EFTP V A+ K +A V+ L KY

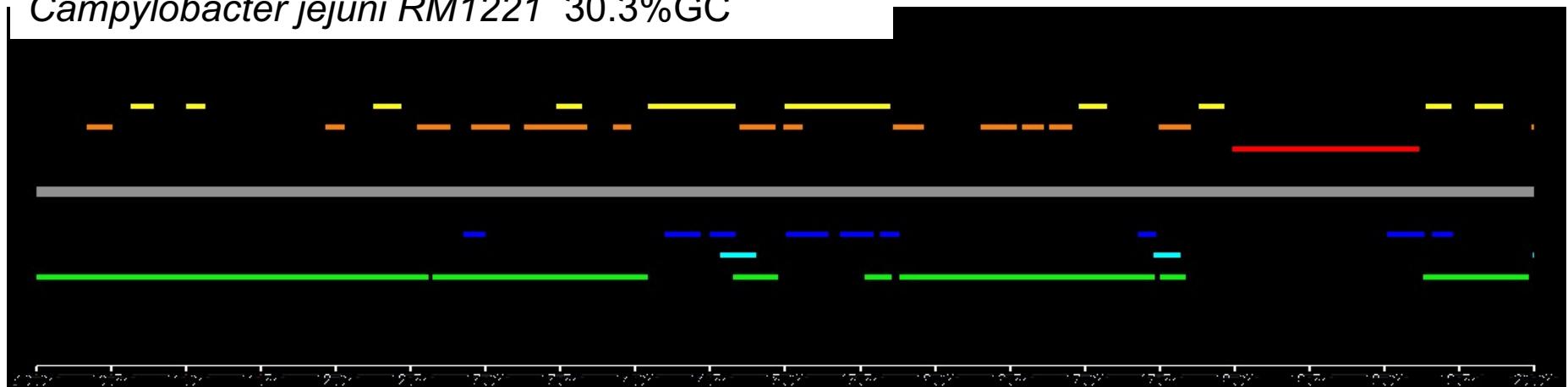
Sbjct 121 EFTPPVQAAYQKVVAGVANALAHKY 145



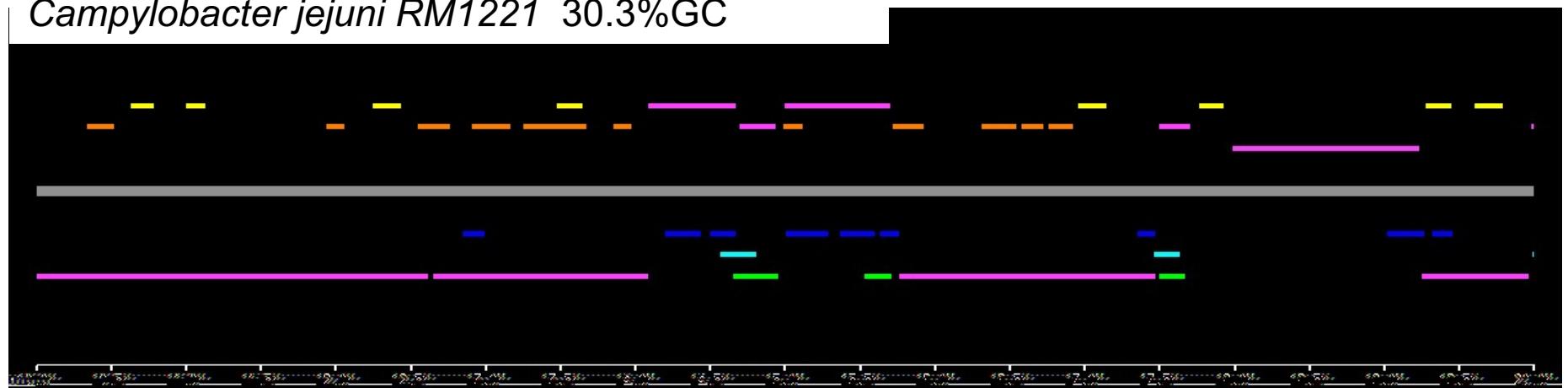
Outline

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

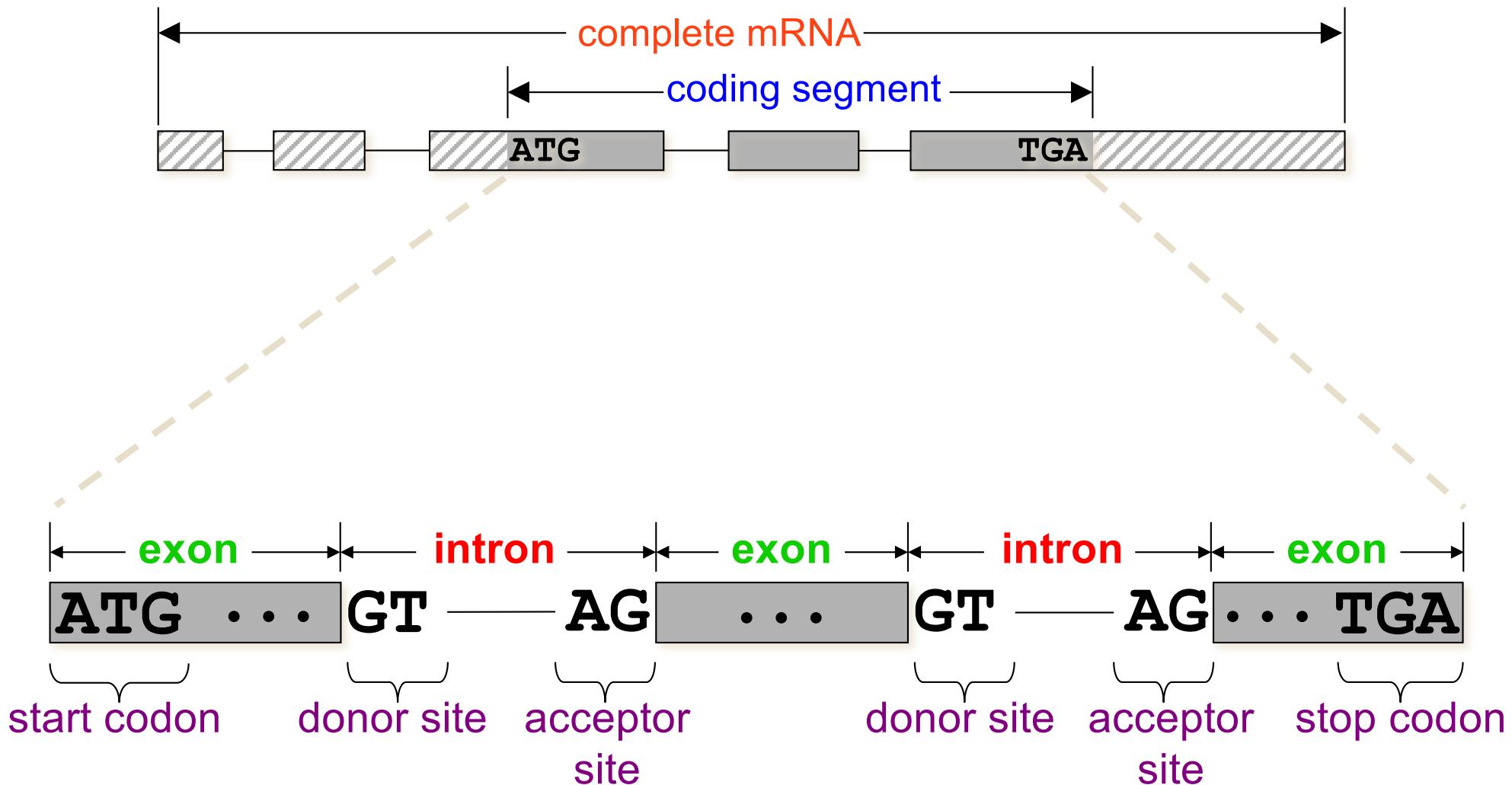
Campylobacter jejuni RM1221 30.3%GC



Campylobacter jejuni RM1221 30.3%GC



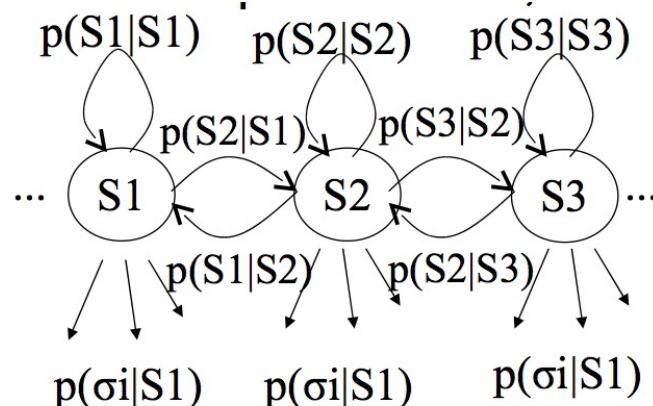
Eukaryotic Gene Syntax



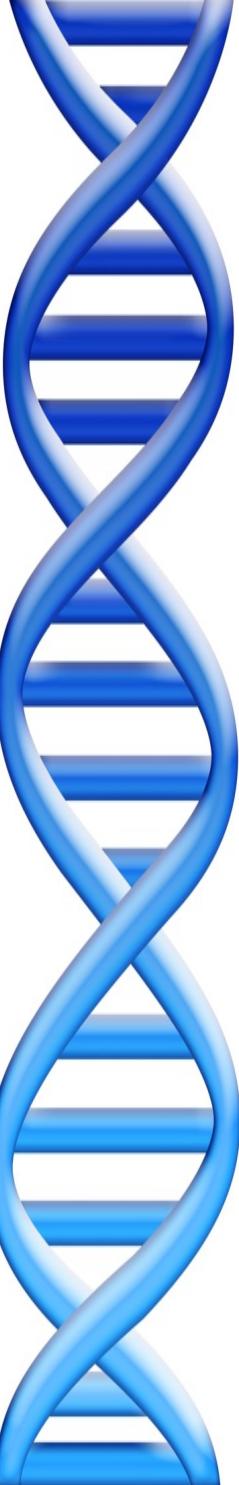
Regions of the gene outside of the CDS are called **UTR's** (*untranslated regions*), and are mostly ignored by gene finders, though they are important for regulatory functions.

HMMs for Gene Finding

- Similar to Markov models used for prokaryotic gene finding, but system may transition between multiple models called states (gene/non-gene, intergenic/exon/intron)
- Observers can see the emitted symbols of an HMM (i.e., nucleotides) but have no ability to know which state the HMM is currently in.
 - But we can *infer* the most likely hidden states of an HMM based on the given sequence of emitted symbols.



AAAGCATGCATTAACGTGAGCACAAATAGATTACA



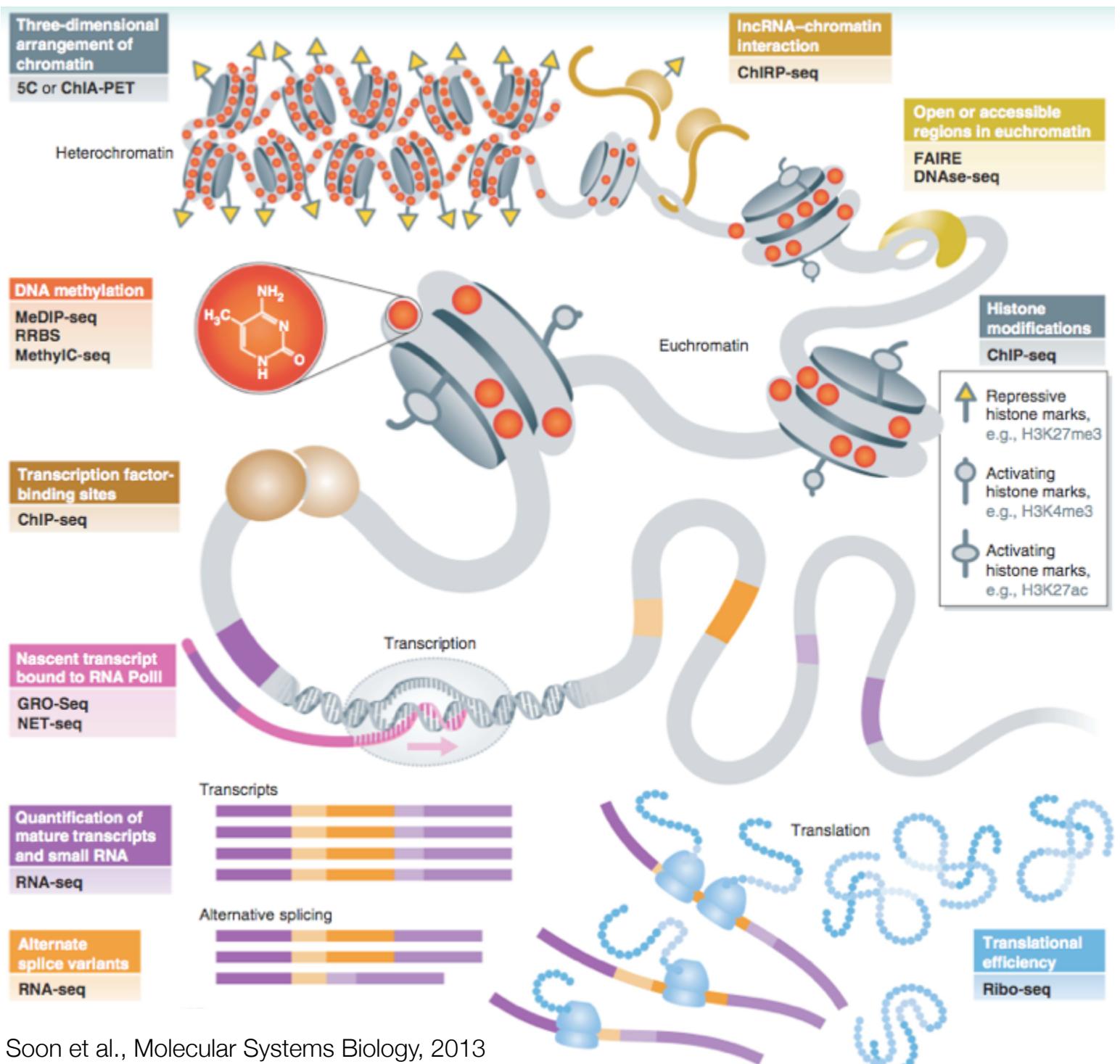
Outline

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

Sequencing Assays

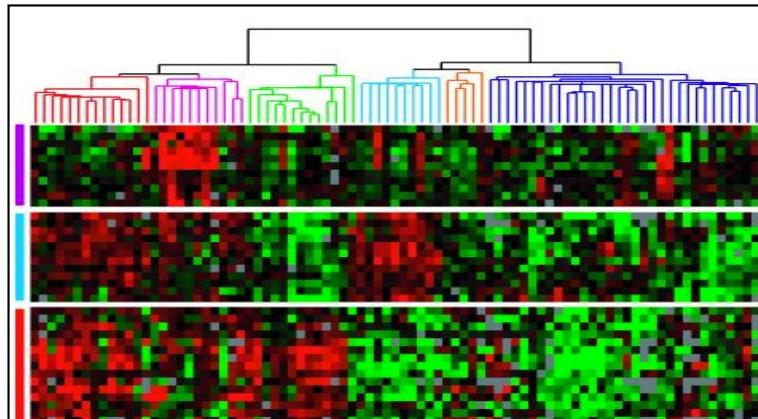
The *Seq List (in chronological order)

1. Gregory E. Crawford et al., “Genome-wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS),” *Genome Research* 16, no. 1 (January 1, 2006): 123–131, doi:10.1101/gr.4074106.
2. David S. Johnson et al., “Genome-Wide Mapping of in Vivo Protein-DNA Interactions,” *Science* 316, no. 5830 (June 8, 2007): 1497–1502, doi:10.1126/science.1141319.
3. Tarjei S. Mikkelsen et al., “Genome-wide Maps of Chromatin State in Pluripotent and Lineage-committed Cells,” *Nature* 448, no. 7153 (August 2, 2007): 553–560, doi:10.1038/nature06008.
4. Thomas A. Down et al., “A Bayesian Deconvolution Strategy for Immunoprecipitation-based DNA Methylome Analysis,” *Nature Biotechnology* 26, no. 7 (July 2008): 779–785, doi:10.1038/nbt1414.
5. Ali Mortazavi et al., “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq,” *Nature Methods* 5, no. 7 (July 2008): 621–628, doi:10.1038/nmeth.1226.
6. Nathan A. Baird et al., “Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers,” *PLoS ONE* 3, no. 10 (October 13, 2008): e3376, doi:10.1371/journal.pone.0003376.
7. Leighton J. Core, Joshua J. Waterfall, and John T. Lis, “Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters,” *Science* 322, no. 5909 (December 19, 2008): 1845–1848, doi:10.1126/science.1162228.
8. Chao Xie and Martti T. Tammi, “CNV-seq, a New Method to Detect Copy Number Variation Using High-throughput Sequencing,” *BMC Bioinformatics* 10, no. 1 (March 6, 2009): 80, doi:10.1186/1471-2105-10-80.
9. Jay R. Hesselberth et al., “Global Mapping of protein-DNA Interactions in Vivo by Digital Genomic Footprinting,” *Nature Methods* 6, no. 4 (April 2009): 283–289, doi:10.1038/nmeth.1313.
10. Nicholas T. Ingolia et al., “Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling,” *Science* 324, no. 5924 (April 10, 2009): 218–223, doi:10.1126/science.1168978.
11. Alayne L. Brunner et al., “Distinct DNA Methylation Patterns Characterize Differentiated Human Embryonic Stem Cells and Developing Human Fetal Liver,” *Genome Research* 19, no. 6 (June 1, 2009): 1044–1056, doi:10.1101/gr.088773.108.
12. Mayumi Oda et al., “High-resolution Genome-wide Cytosine Methylation Profiling with Simultaneous Copy Number Analysis and Optimization for Limited Cell Numbers,” *Nucleic Acids Research* 37, no. 12 (July 1, 2009): 3829–3839, doi:10.1093/nar/gkp260.
13. Zachary D. Smith et al., “High-throughput Bisulfite Sequencing in Mammalian Genomes,” *Methods* 48, no. 3 (July 2009): 226–232, doi:10.1016/j.ymeth.2009.05.003.
14. Andrew M. Smith et al., “Quantitative Phenotyping via Deep Barcode Sequencing,” *Genome Research* (July 21, 2009).

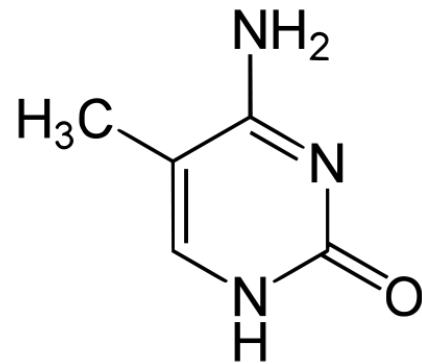


*-seq in 4 short vignettes

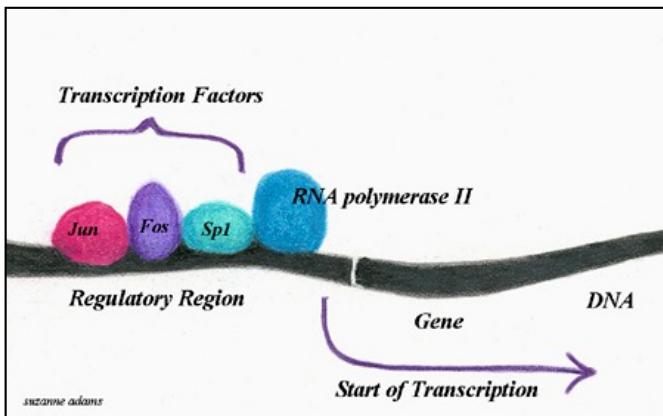
RNA-seq



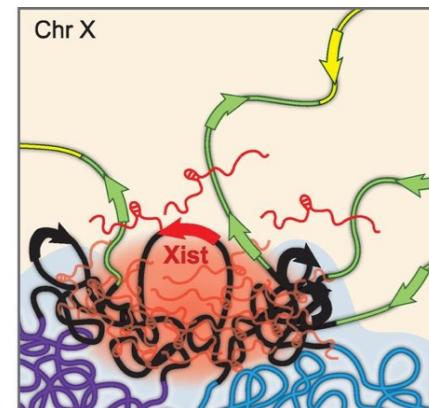
Methyl-seq



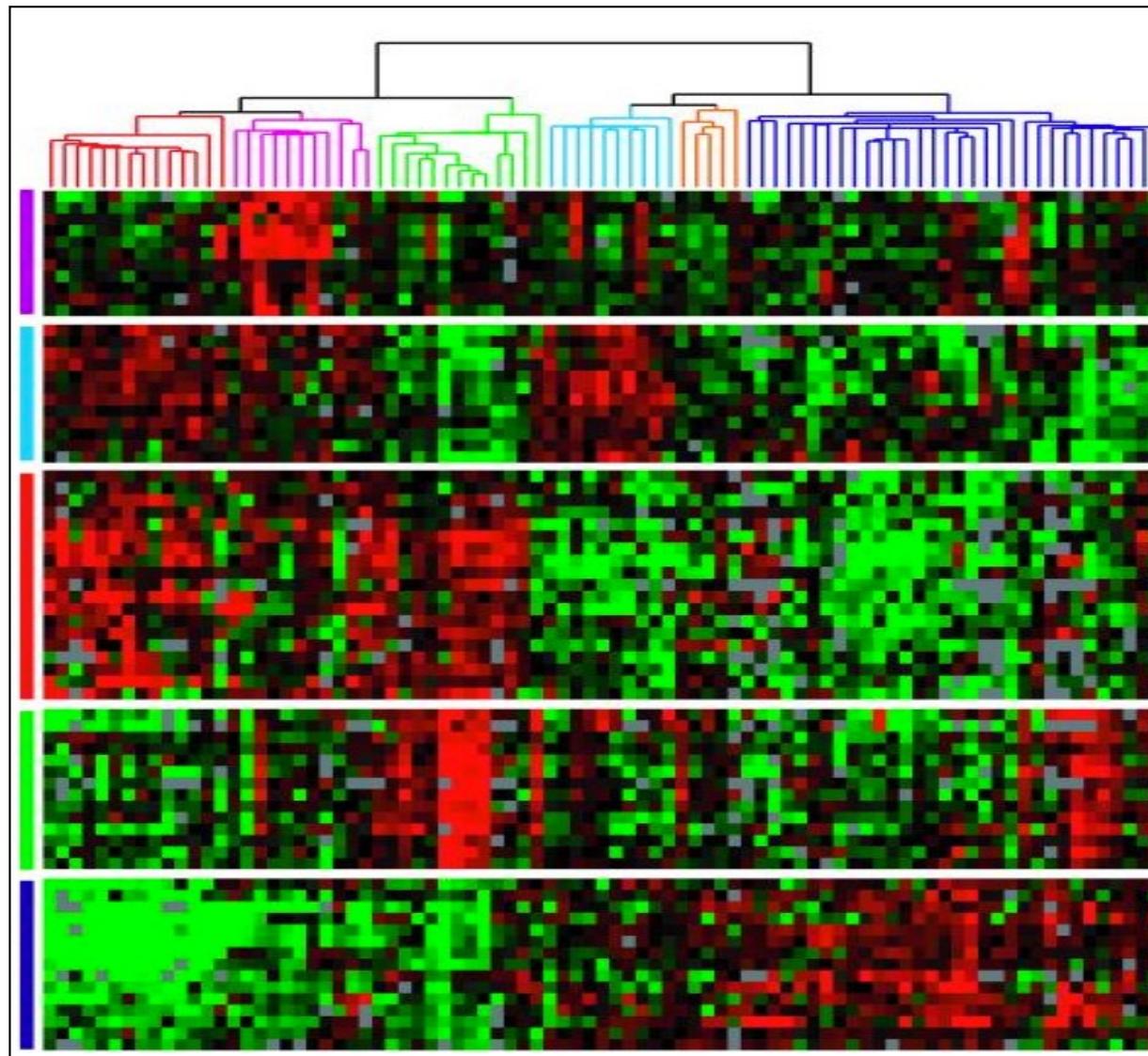
ChIP-seq



Hi-C

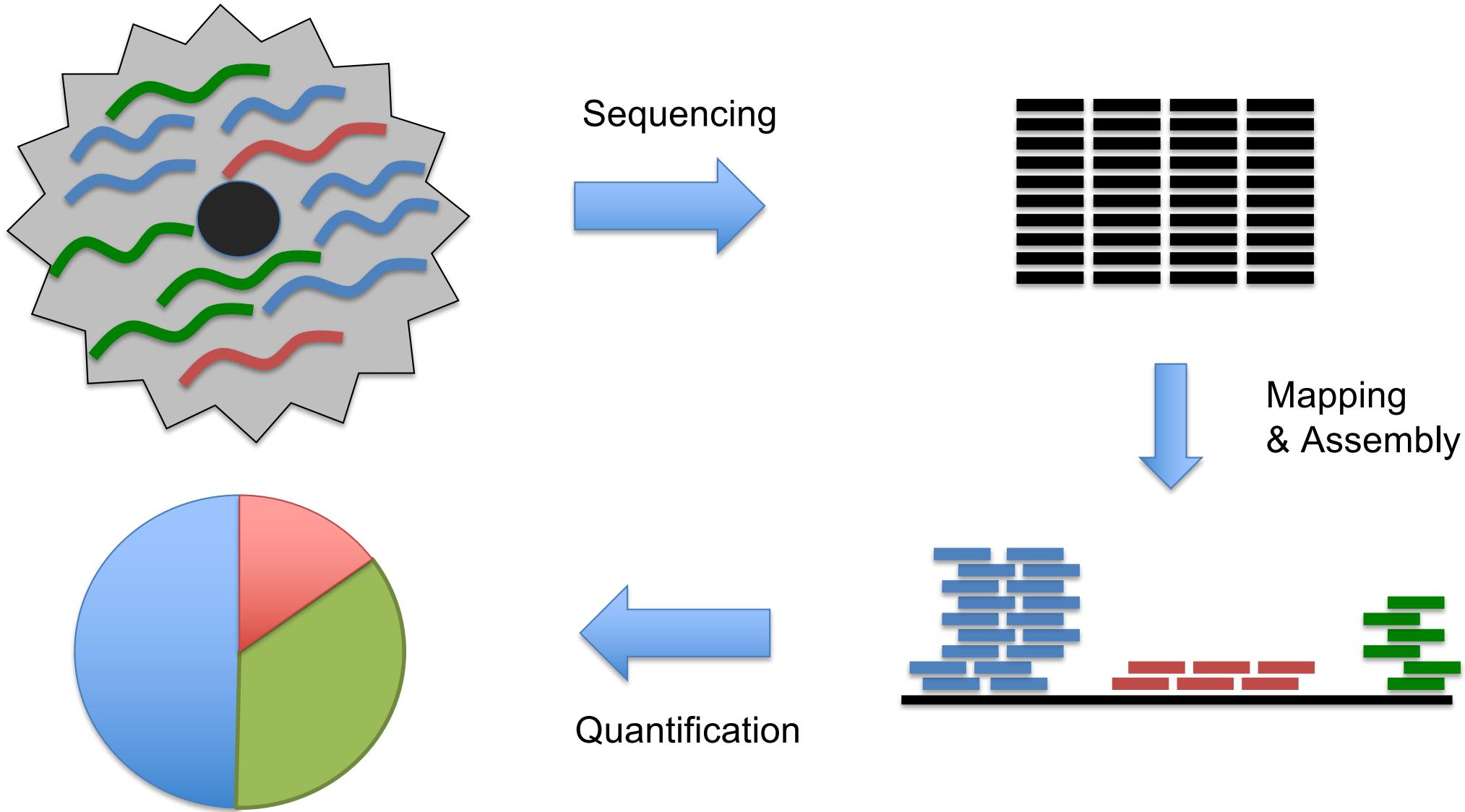


RNA-seq

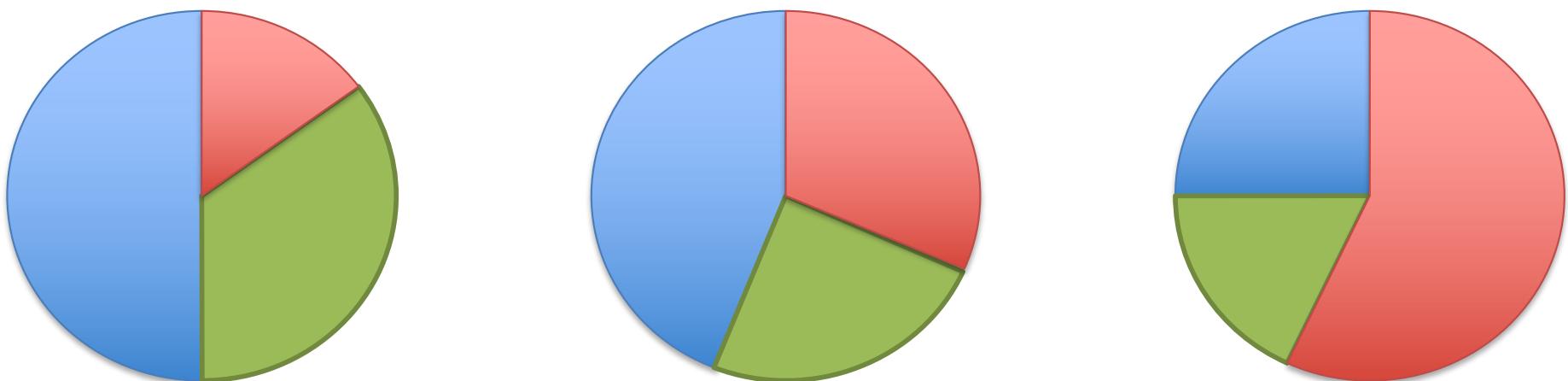
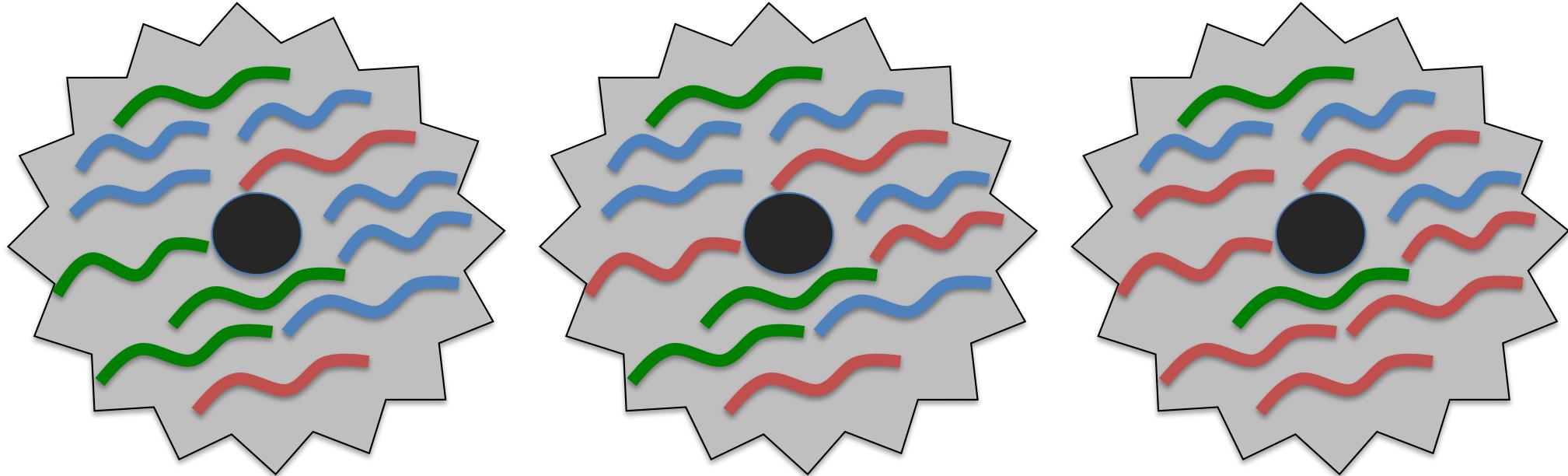


Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.
Sørlie et al (2001) PNAS. 98(19):10869-74.

RNA-seq Overview

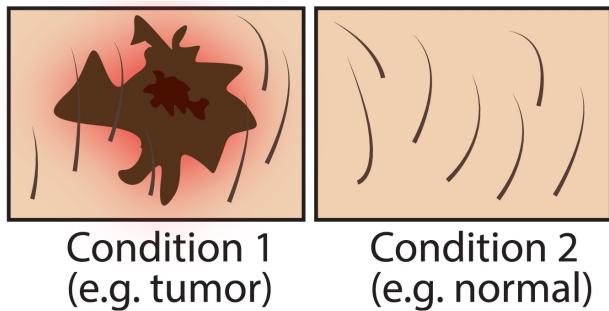


RNA-seq Overview

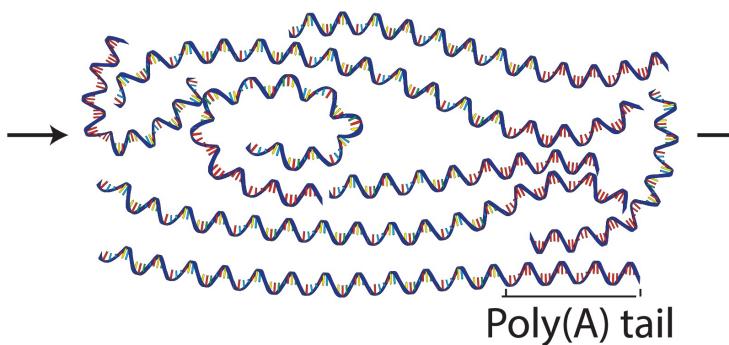


RNA-seq Overview

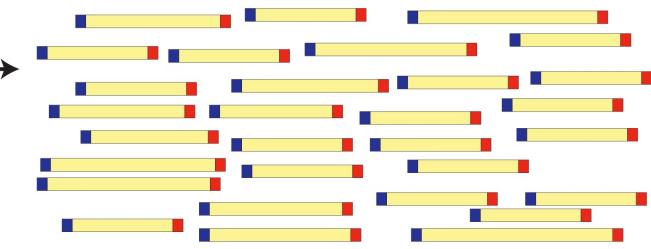
Samples of interest



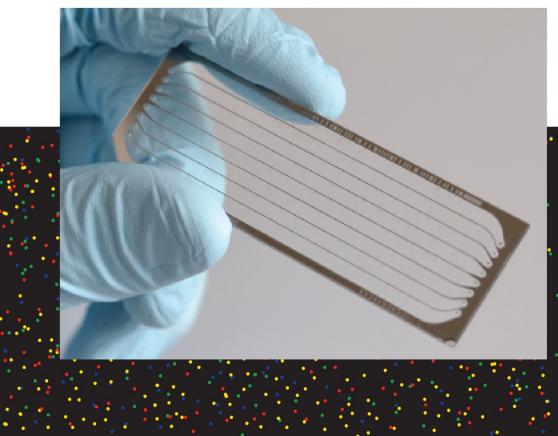
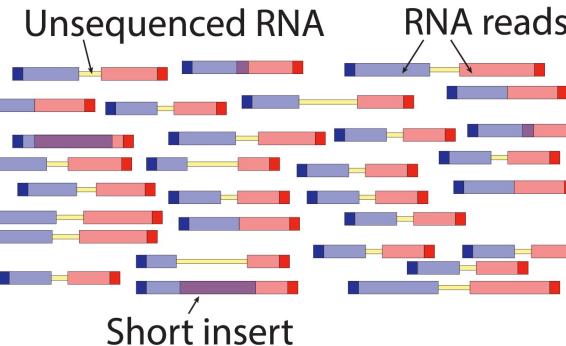
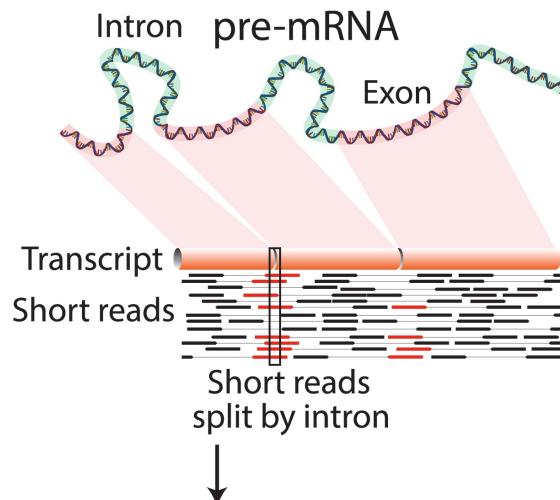
Isolate RNAs



Generate cDNA, fragment, size select, add linkers



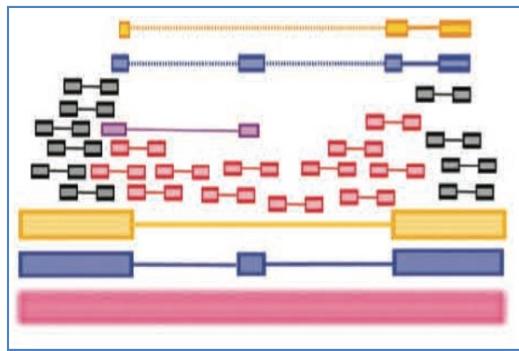
Map to genome, transcriptome, and predicted exon junctions



Downstream analysis

100s of millions of paired reads
10s of billions bases of sequence

RNA-seq Challenges



Challenge I: Eukaryotic genes are spliced

RNA-Seq Approaches

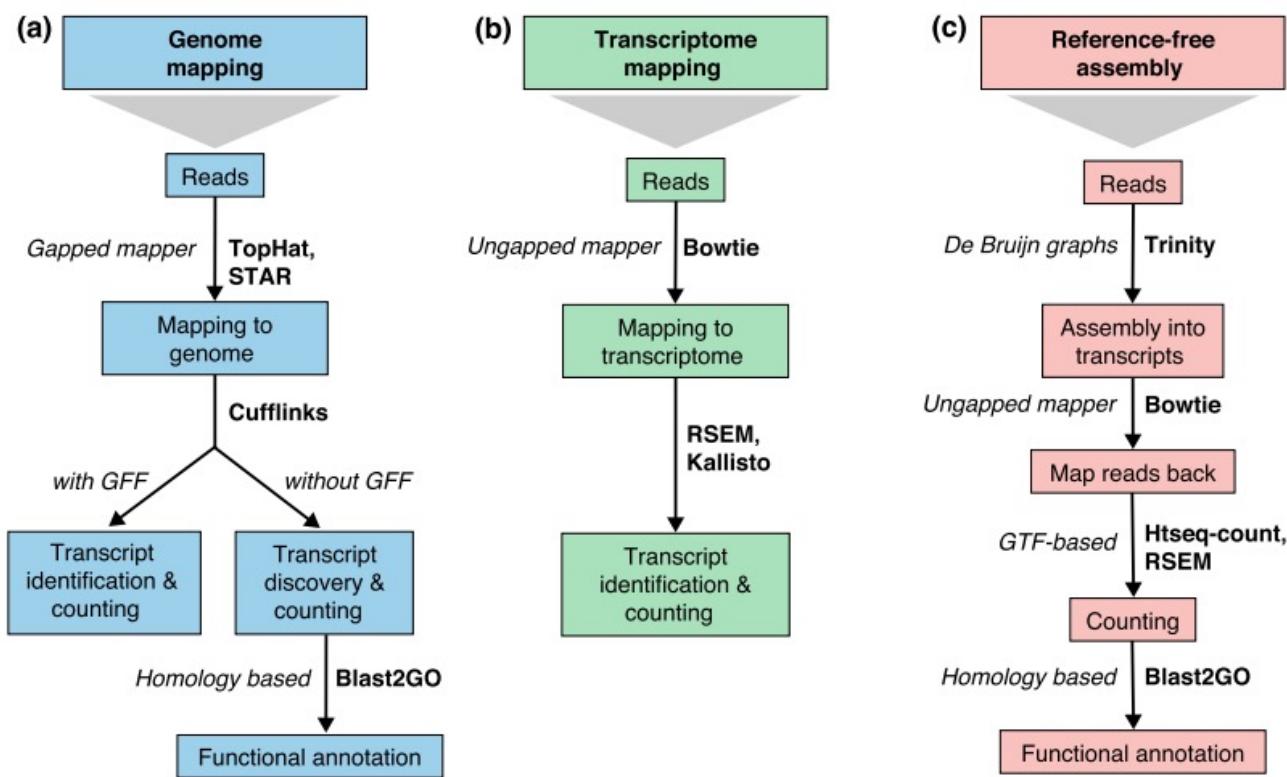


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome with a gapped mapper. Next (novel) transcript discovery and quantification can proceed with or without an annotation file. Novel transcripts are then functionally annotated. **b** If 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 need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analysis proceeds as in **(b)** followed by the functional annotation of the novel transcripts as in **(a)**. Representative software that can be used at each analysis step are indicated in bold text. Abbreviations: GFF General Feature Format, GTF gene transfer format, RSEM RNA-Seq by Expectation Maximization

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 Approaches

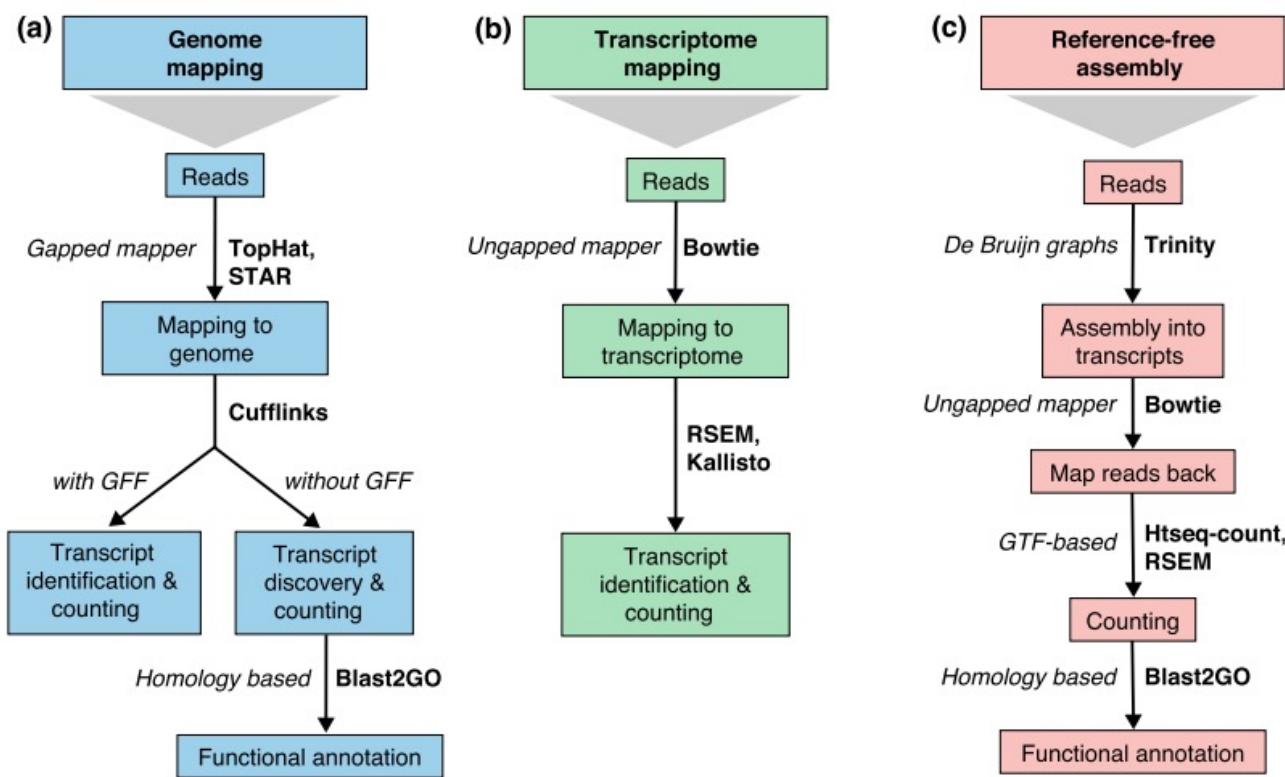


Fig. 2 Read mapping and transcript identification strategies. Three basic strategies for regular RNA-seq analysis. **a** An annotated genome is available and reads are mapped to the genome using a gapped aligner (TopHat, STAR). Transcript identification and quantification can proceed with or without an annotation file (GFF). **b** If no novel transcript discovery is needed, reads can be mapped to the reference transcriptome using an ungapped aligner (Bowtie). Transcript identification and quantification can occur simultaneously. **c** When no genome is available, reads need to be assembled first into contigs or transcripts. For quantification, reads are mapped back to the novel reference transcriptome and further analyzed using a transcriptome assembler (Trinity). This is followed by the functional annotation of the novel transcripts as in **(a)**. Representative software that can be used at each analysis step are indicated in bold text. Abbreviations: GFF General Feature Format, GTF gene transfer format, RSEM RNA-Seq by Expectation Maximization

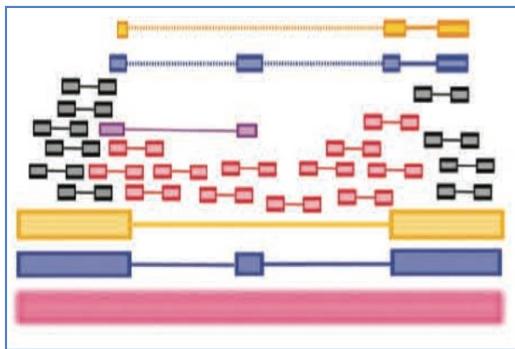
Which approach should we use?

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



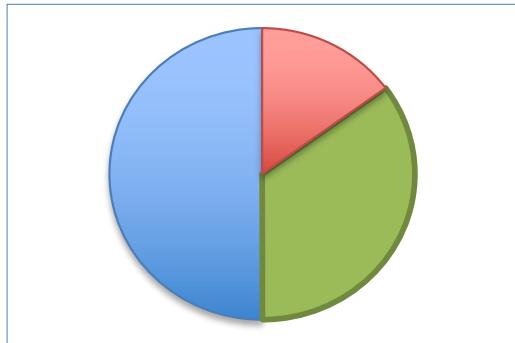
Challenge 1: Eukaryotic genes are spliced

Solution: Use a spliced aligner, and assemble isoforms

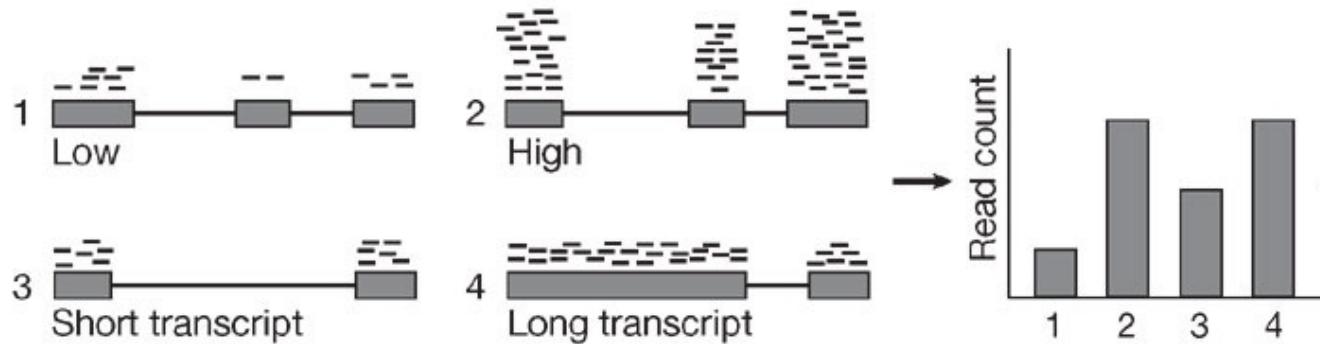
TopHat: discovering spliced junctions with RNA-Seq.

Trapnell et al (2009) *Bioinformatics*. 25:0 1105-1111

Challenge 2: Read Count != Transcript abundance



RPKM, FPKM, TPM

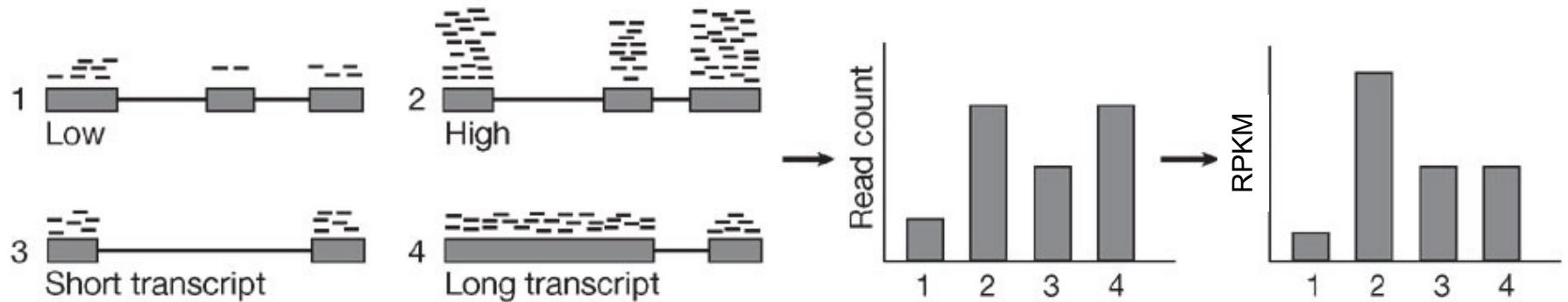


Counting Reads that align to a gene DOESN'T work!

- 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!

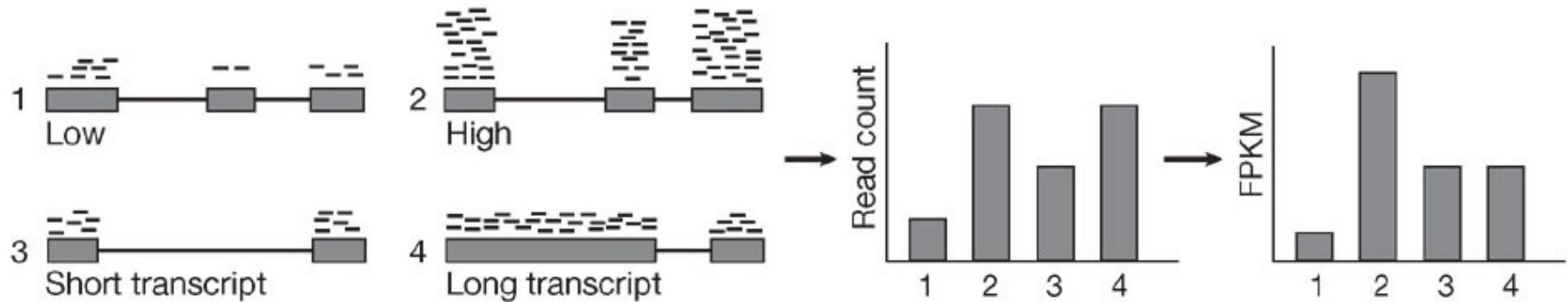
- Overall Coverage: 1M reads in experiment 1 vs 10M reads in experiment 2
- Gene Length: gene 3 is 10kbp, gene 4 is 100kbp

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 aren't 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)

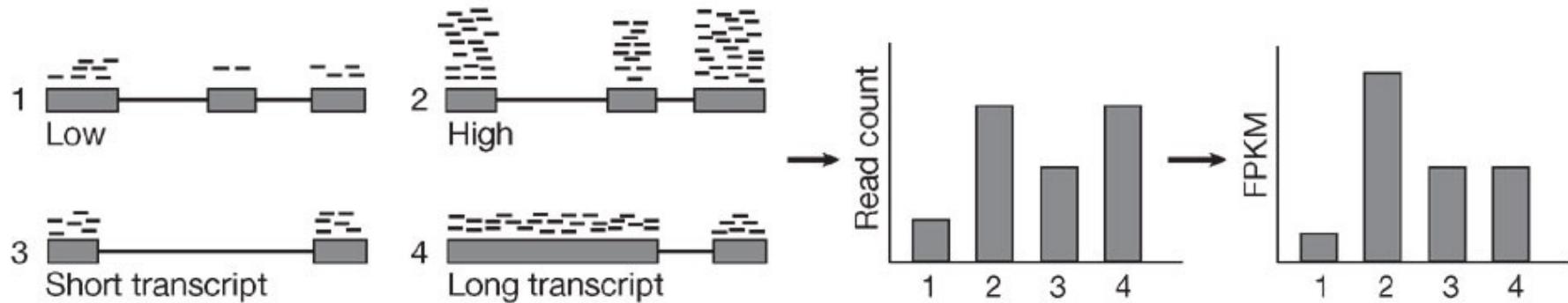
=> Wait a second, reads in a pair aren't independent!

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

⇒ Does a much better job with short exons & short genes by boosting coverage

⇒ 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 aren't 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!

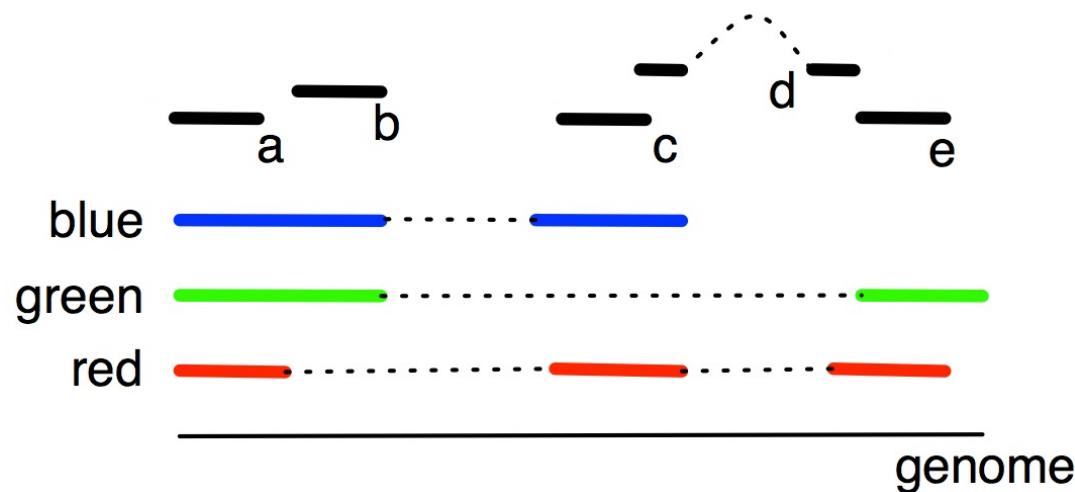
3. TPM: Transcripts Per Million (Li et al, 2011)

⇒ 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

$$TPM_i = \left(\frac{FPKM_i}{\sum_j FPKM_j} \right) \cdot 10^6$$

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



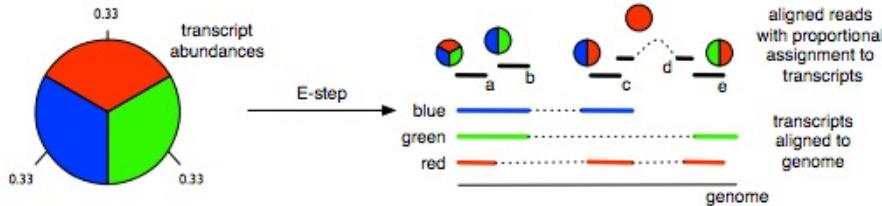
The gene has three isoforms (red, green, blue) of the same length.
Our initial expectation is all 3 isoforms are equally expressed

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?

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

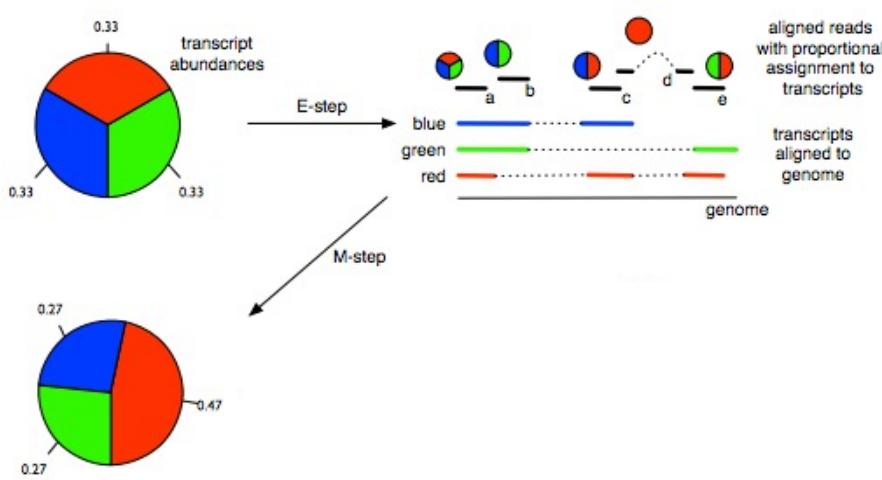


The gene has three isoforms (red, green, blue) of the same length. Initially every isoform is assigned the same abundance (red=1/3, green=1/3, blue=1/3)

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)

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)

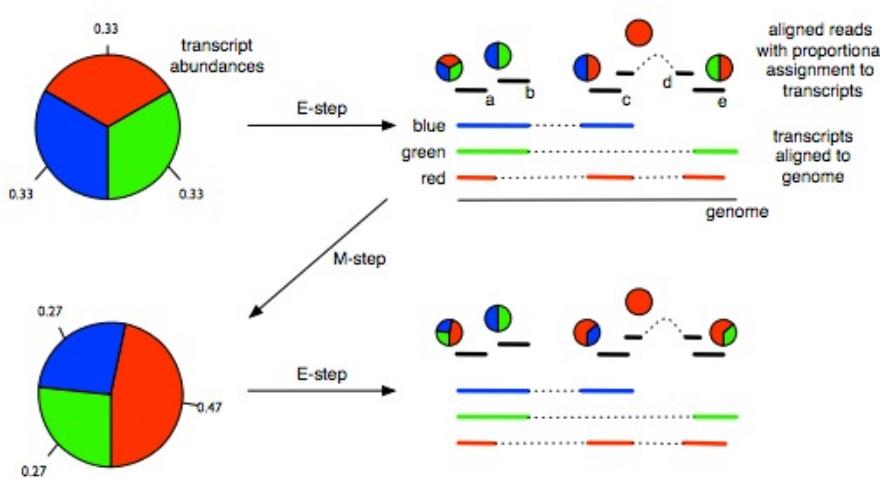
Next, during the maximization (M) step isoform abundances are recalculated from the proportionately assigned read counts:

$$\text{red: } 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)$$

$$\text{blue: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

$$\text{green: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

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:

$$\text{red: } 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)$$

$$\text{blue: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

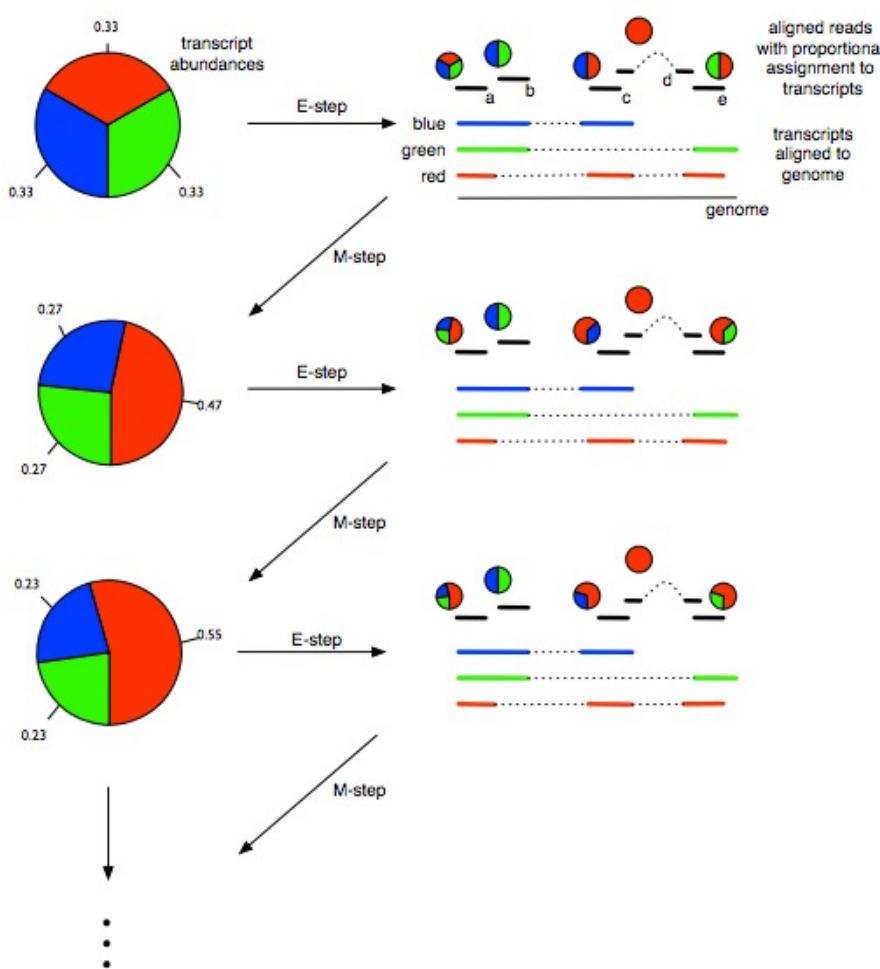
$$\text{green: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

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:

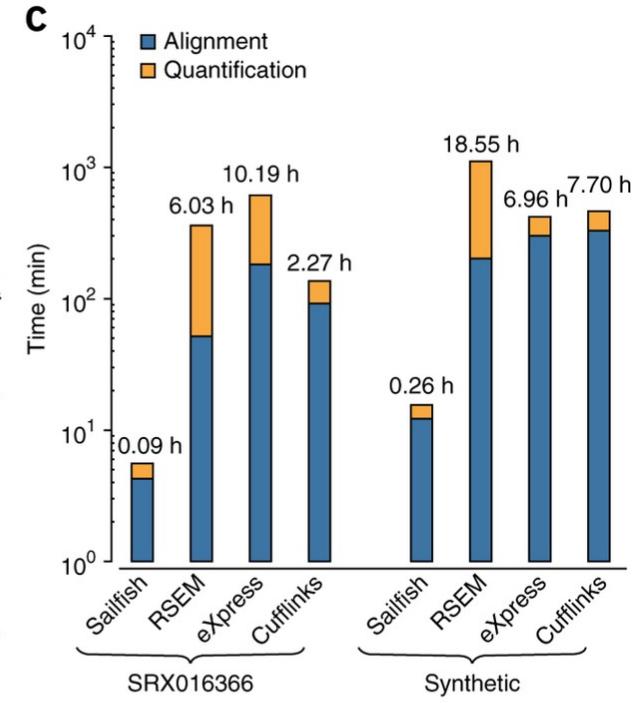
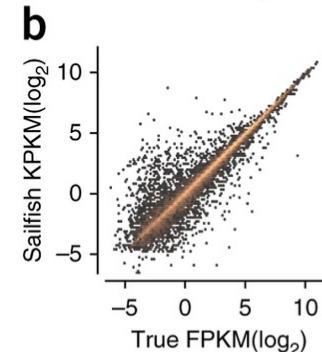
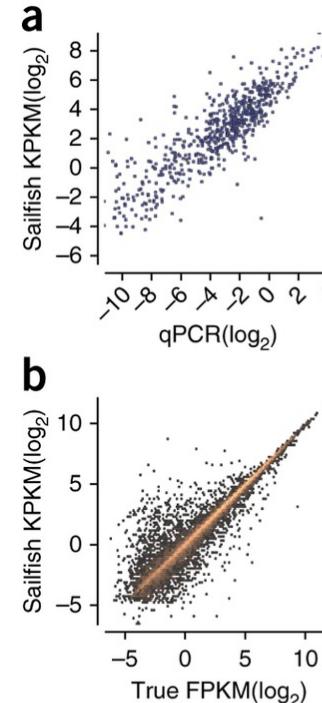
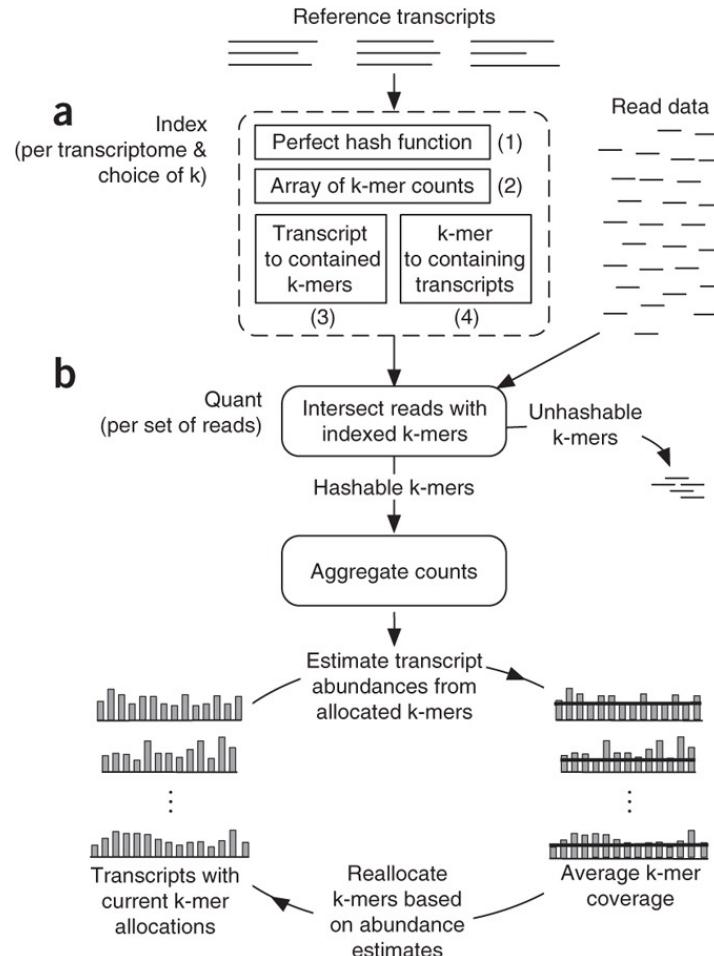
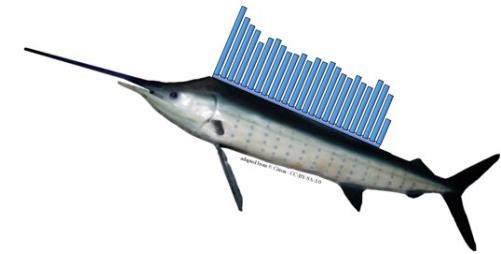
$$\text{red: } 0.47 = (0.33 + 0.5 + 1 + 0.5)/(2.33 + 1.33 + 1.33)$$

$$\text{blue: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

$$\text{green: } 0.27 = (0.33 + 0.5 + 0.5)/(2.33 + 1.33 + 1.33)$$

Repeat until convergence!

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

Annotation Summary

- Three major approaches to annotate a genome
 - I. 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 ☺

