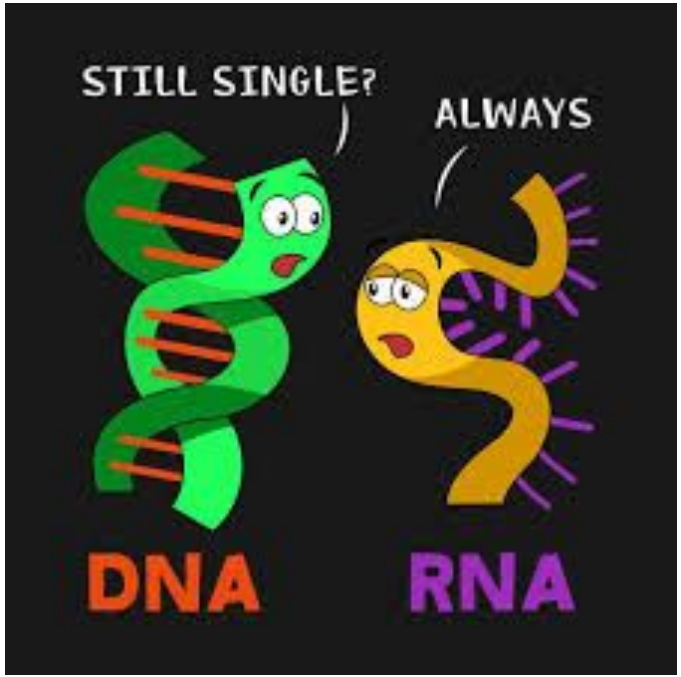
BCB 716 - Sequence Analysis



- I still haven't made progress on Problem Sets #1 and #2.
- I will announce some changes on Thursday

RNA Sequence Analysis

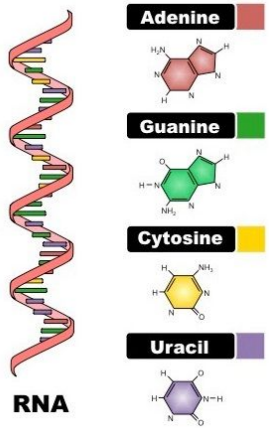
Ribonucleic Acid (RNA)

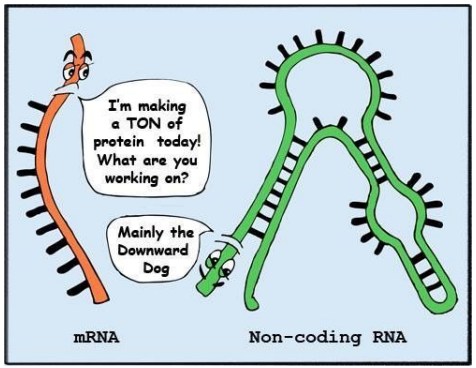
Like DNA, RNA is a long polymer consisting of nucleotides.

- RNA, however, is *single-stranded*
- The strand has a 5'-end (with a phosphate group) and a 3'-end (with a hydroxyl group).
- It is composed of ribonucleotides
  Adenine (A), Cytosine (C), *Uracil (U)*, and Guanine (G).

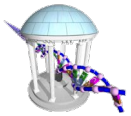
To enhance stability RNA typically folds and bonds with itself.

Often RNA will be annotated using it's complementary DNA (cDNA) sequence with U → T

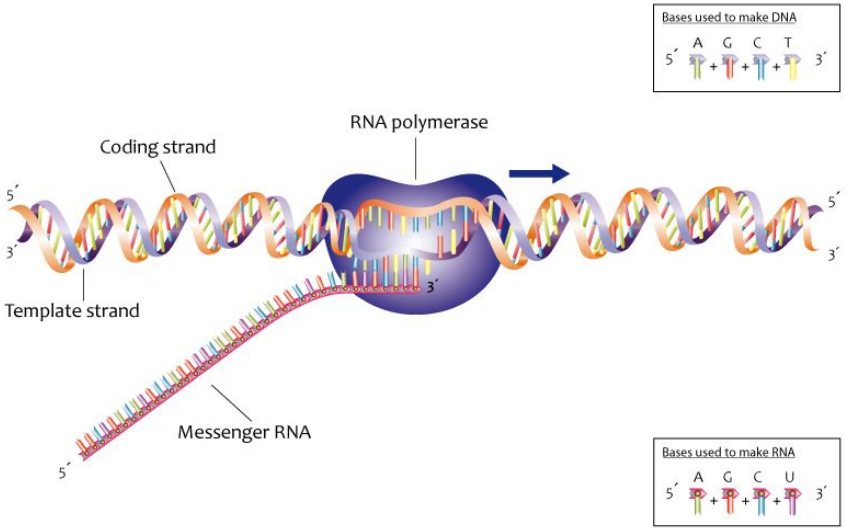




RNA is a *functional* sequence

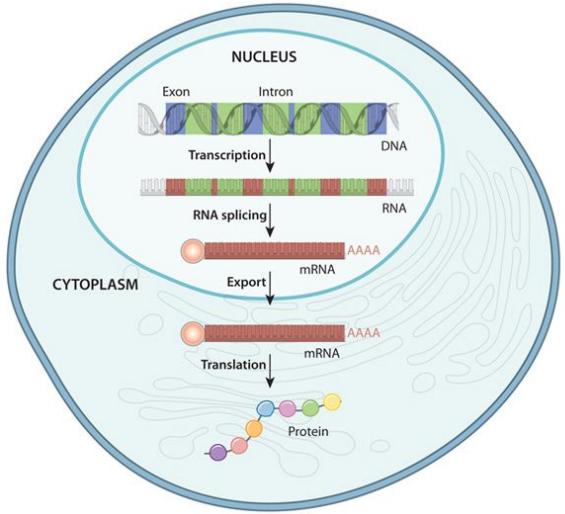


- DNA sequence is the blueprint for all cellular processes
- RNA sequence is functional byproduct of reading the DNA blueprint a.k.a. transcribing DNA
- RNA is often an intermediary between a gene and a functional Protein
- RNA can be directly functional
- RNA expression varies
  - Cell type
  - Cell state
  - In response to environmental insults
  - Genotype state

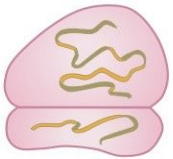


Major types of RNA

- Ribosomal (rRNA)
  - Responsible for protein synthesis
  - Up to 95% of total RNA in a cell
- Messenger (mRNA)
  - Translated into protein in ribosome
  - 3-4% of total RNA in a cell
  - Has poly-A tails in eukaryotes
- Transfer (tRNA)
  - Bring specific amino acids for protein synthesis
- Micro (miRNA)
  - short (22 bp) non-coding RNA involved in expression regulation
- Others (lncRNA, shRNA, siRNA, snoRNA, etc.)

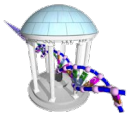



Messenger RNA (mRNA)

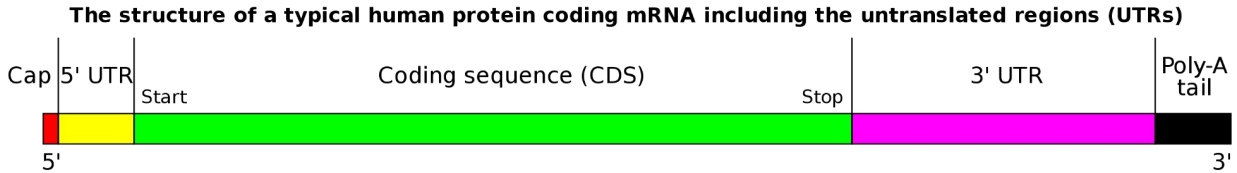

Ribosomal RNA (rRNA)


Transfer RNA (tRNA)

Polyadenylation

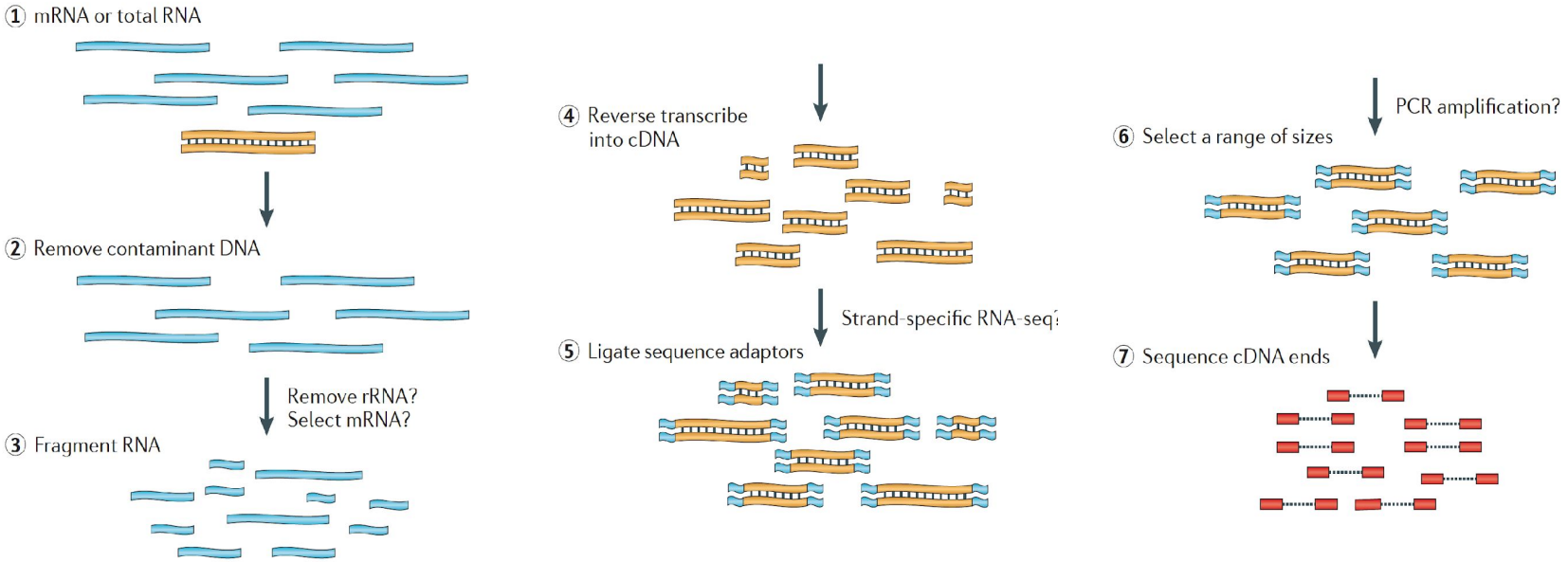


- Polyadenylation is the addition of a poly(A) tail to an RNA transcript, typically a messenger RNA (mRNA)
- The poly(A) tail consists of multiple adenosine monophosphates; in other words, it is a stretch of RNA that has only adenine bases.
- In *eukaryotes*, polyadenylation is part of the process that produces *mature* mRNA for translation. In many bacteria, the poly(A) tail promotes degradation of the mRNA. It, therefore, forms part of the larger process of gene expression.

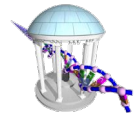


- Various library prep kits select either for or against poly(A)

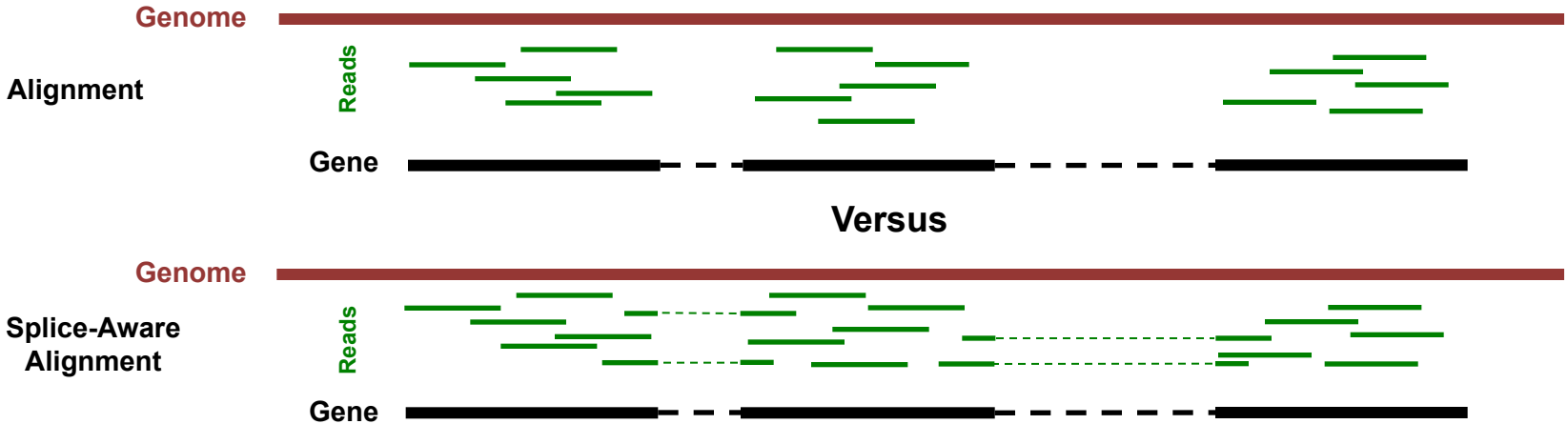
Sequencing RNA



RNA Alignment

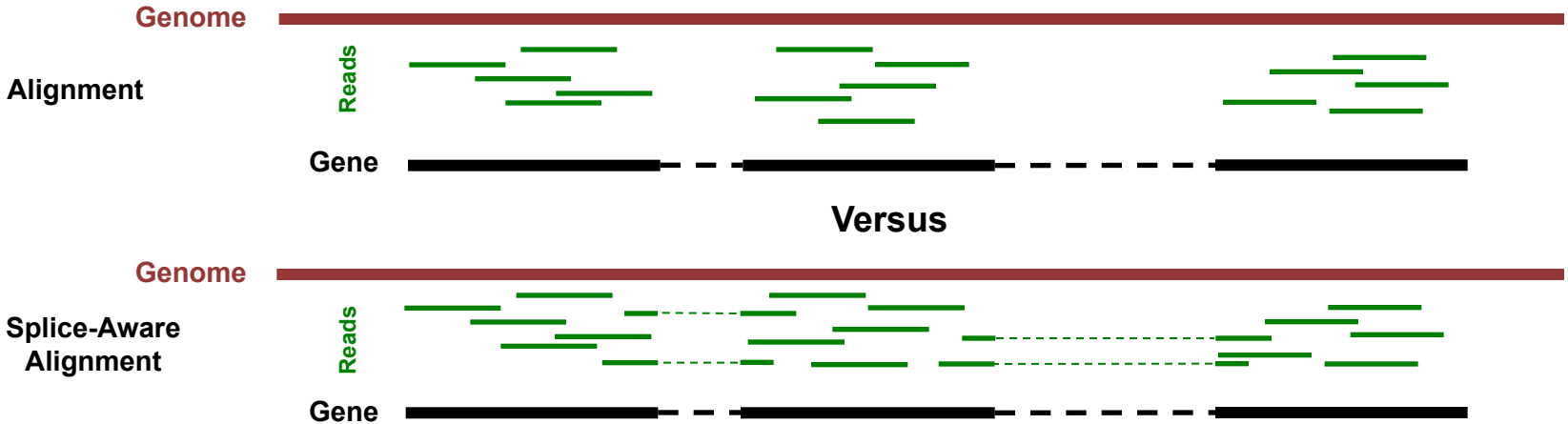


When aligning RNASeq data to a "genome", you will almost always need a "splice-aware" aligner



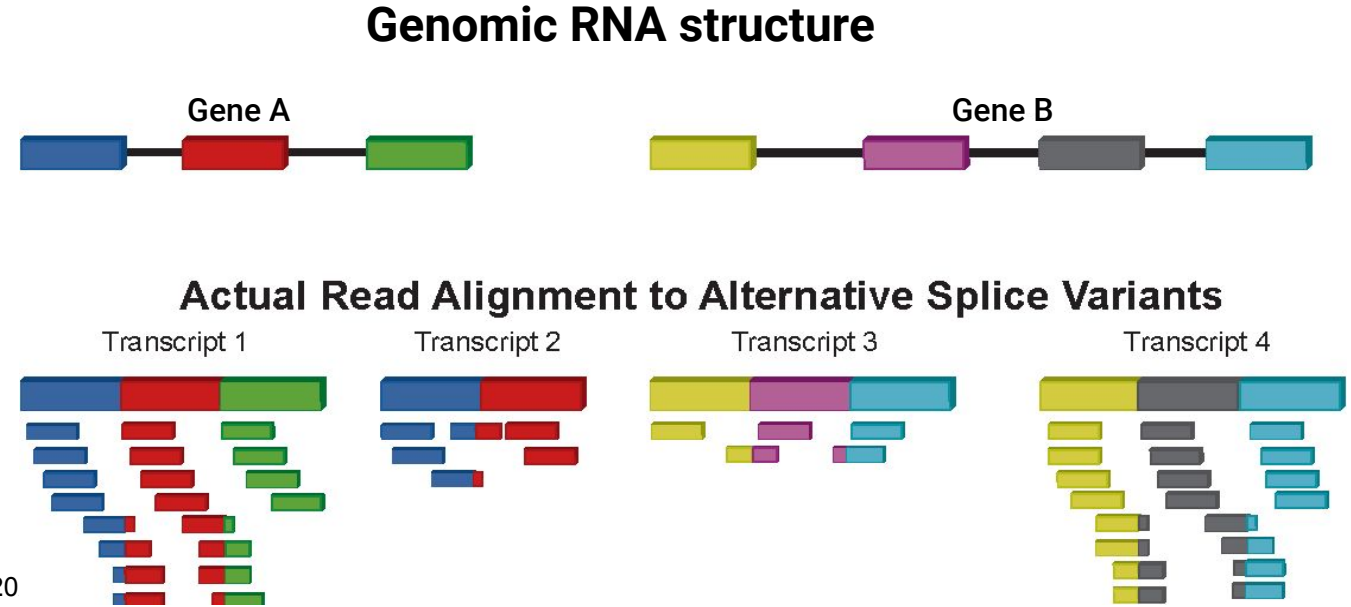
RNA Alignment

When aligning RNASeq data to a "genome", you will almost always need a "splice-aware" aligner (STAR, HiSat2, MapSplice2, GSNAP, etc)

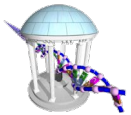


RNA Alignment

A second class of RNA aligners align to "gene models" (a library of transcripts) and assumes all relevant gene models are known. INDEL alignment penalties are adjusted



Considerations when choosing an RNA aligner



Does it deal with reads that map to multiple locations?

Many genes are similar and share sequence

How does it handle paired-end or only single-end read data?

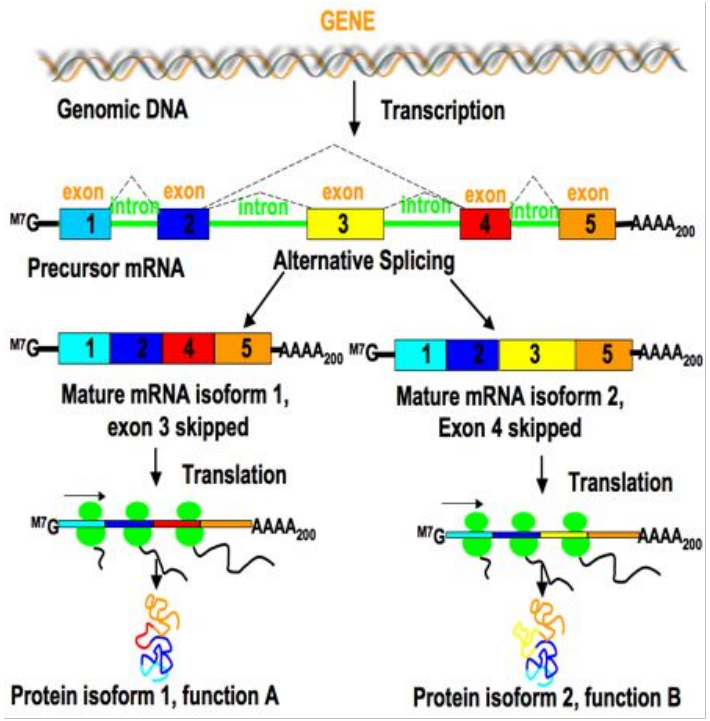
How many mismatches will it allow between the genome and the reads?

What assumptions does it make about the genome?

Once Aligned, then what?

Transcriptome Assembly

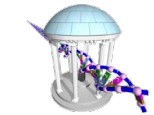
- Genes can be complicated
- Multiple forms of the same gene
- Transcripts and Transcriptome
- In the transcript assembly process reads are processed/grouped into sets of estimated transcripts
- Two approaches
  - reference based
  - de novo



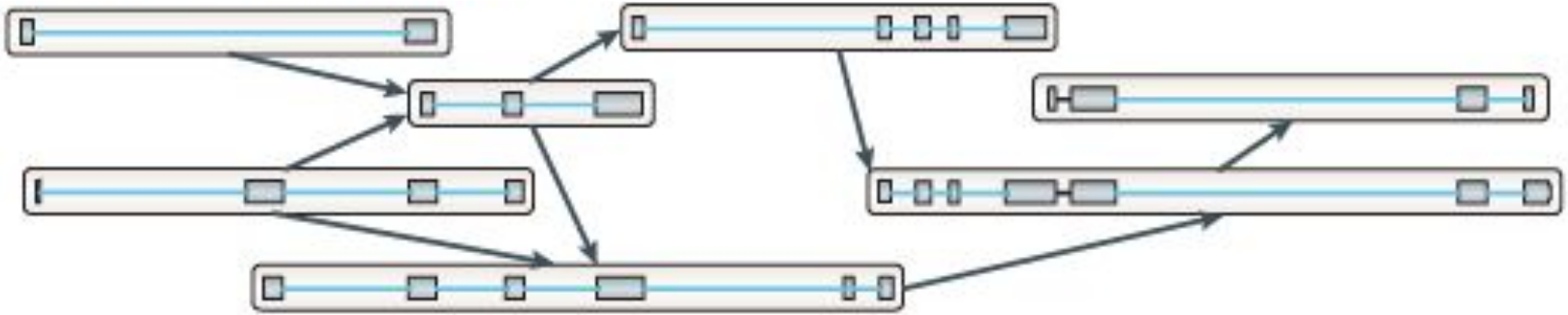
Reference-based Transcriptome Assembly

Cartoon of an alignment result from a "splice-aware" aligner

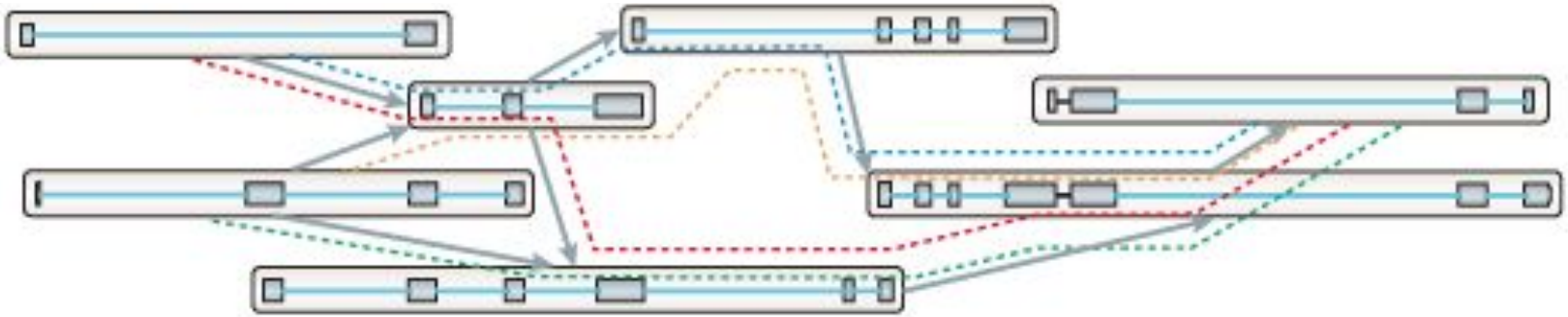
Reference-based Transcriptome Assembly



A. Build graph representing alternative splicing events

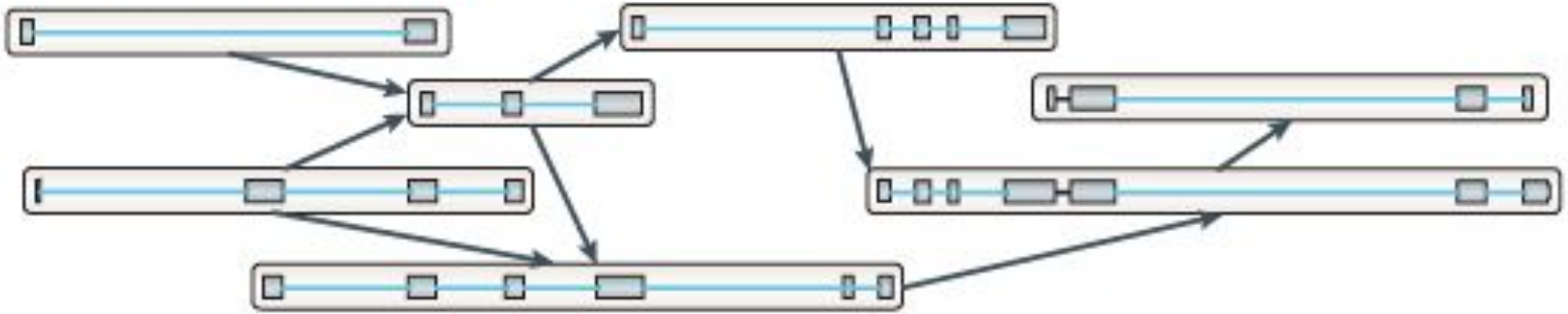


B. Traverse the graph to assemble variants

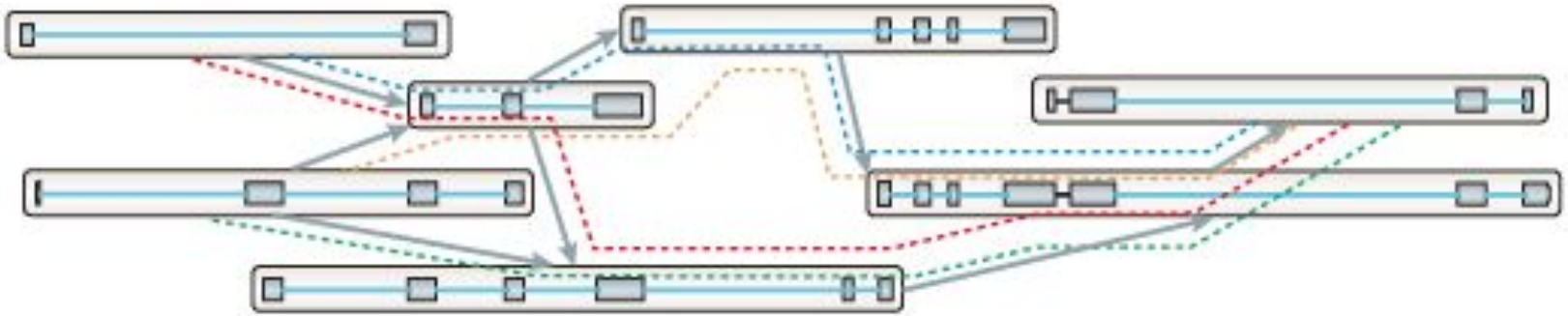


Reference-based Transcriptome Assembly

A. Build graph representing alternative splicing events

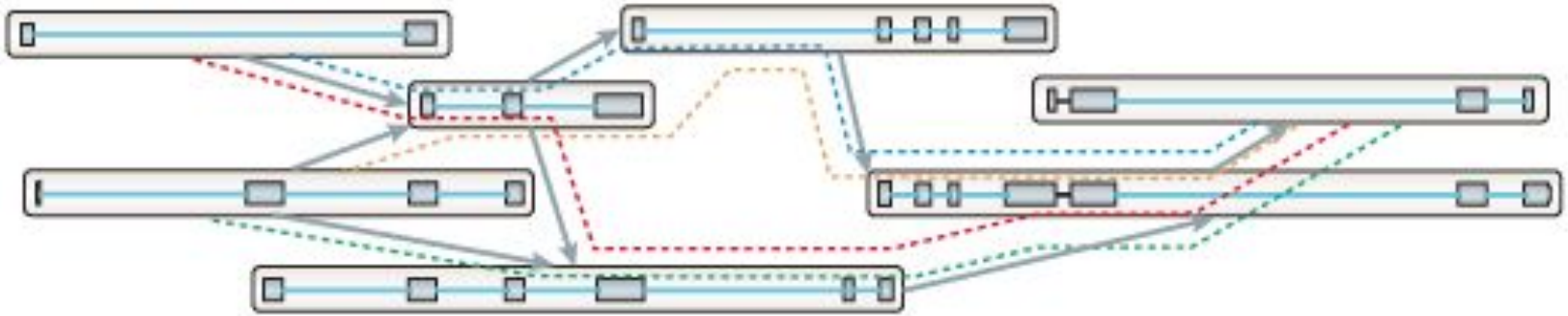


B. Traverse the graph to assemble variants

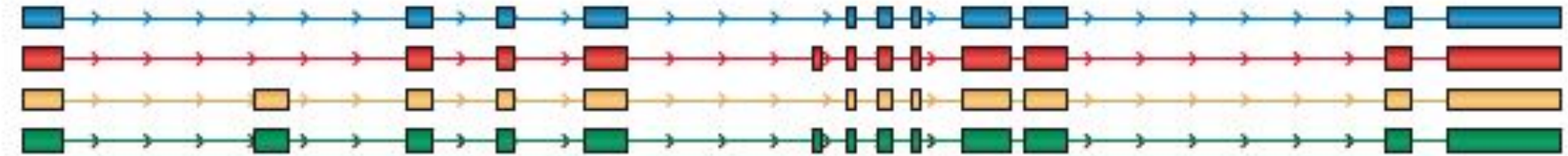


Reference-based Transcriptome Assembly

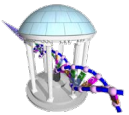
B. Traverse the graph to assemble transcript variants



C. Assembled isoforms



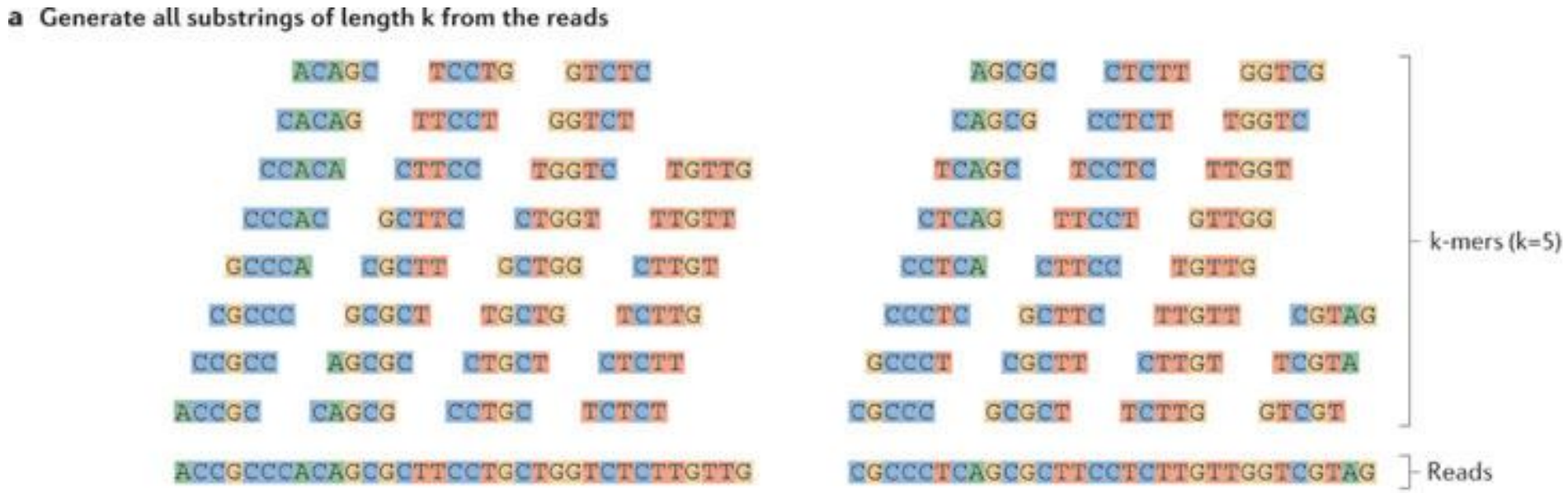
De novo Transcript assembly



Used when very little information is available for the genome

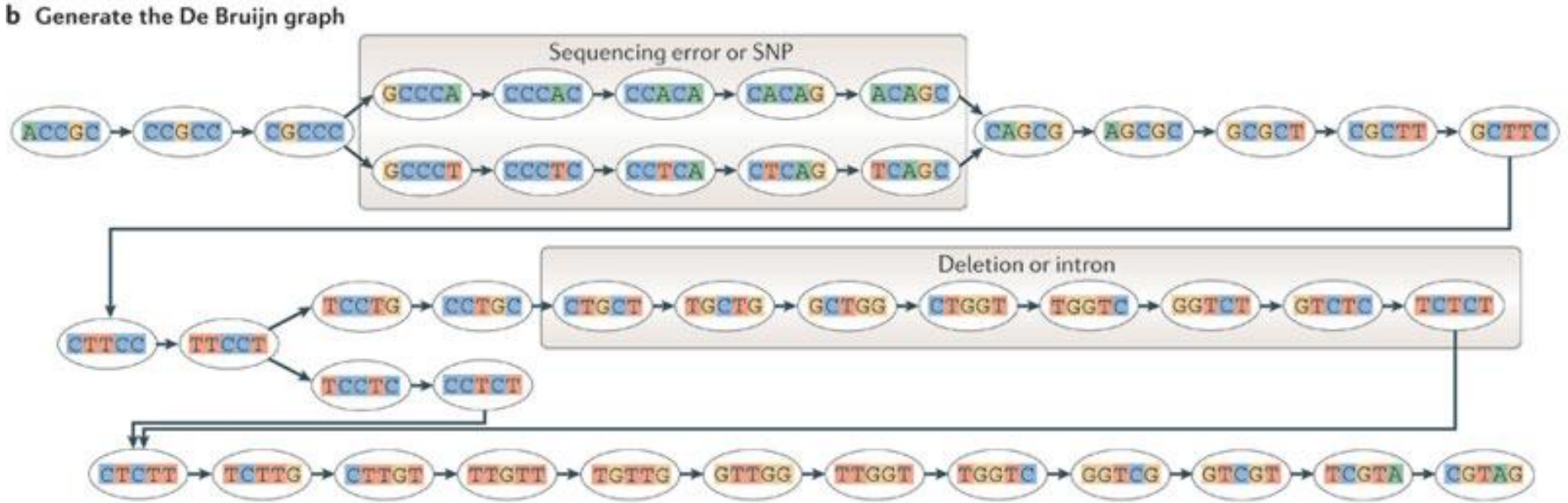
- **Often a first step in putting together information about an unknown genome**
- **Amount of RNA reads needed for a good de novo assembly is higher than for a reference-based assembly**
- **Can be used for genome annotation, once the genome is assembled**
- **Trinity, SPAdes, and TransABySS, are examples of well-regarded transcriptome assemblers**

Steps of a De novo assembly

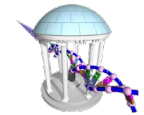


Steps of a De novo assembly

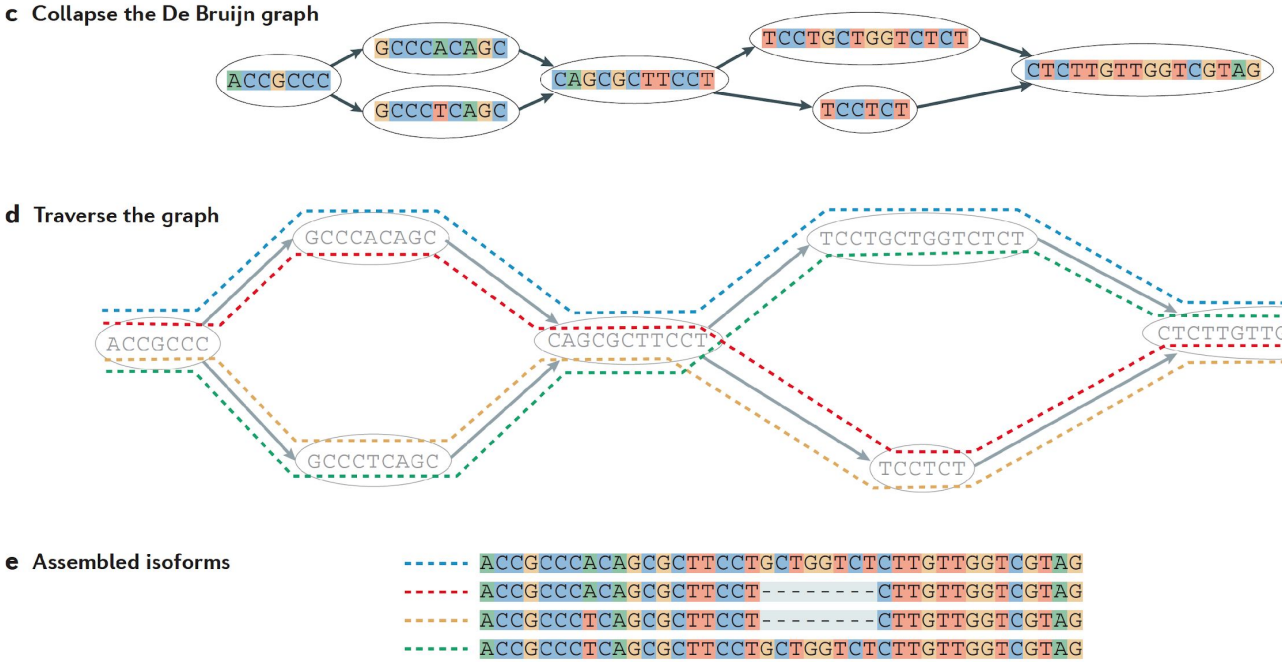
Adds a directed edge between a k-mer whose k-1 suffix matches the k-1 prefix of a second k-mer



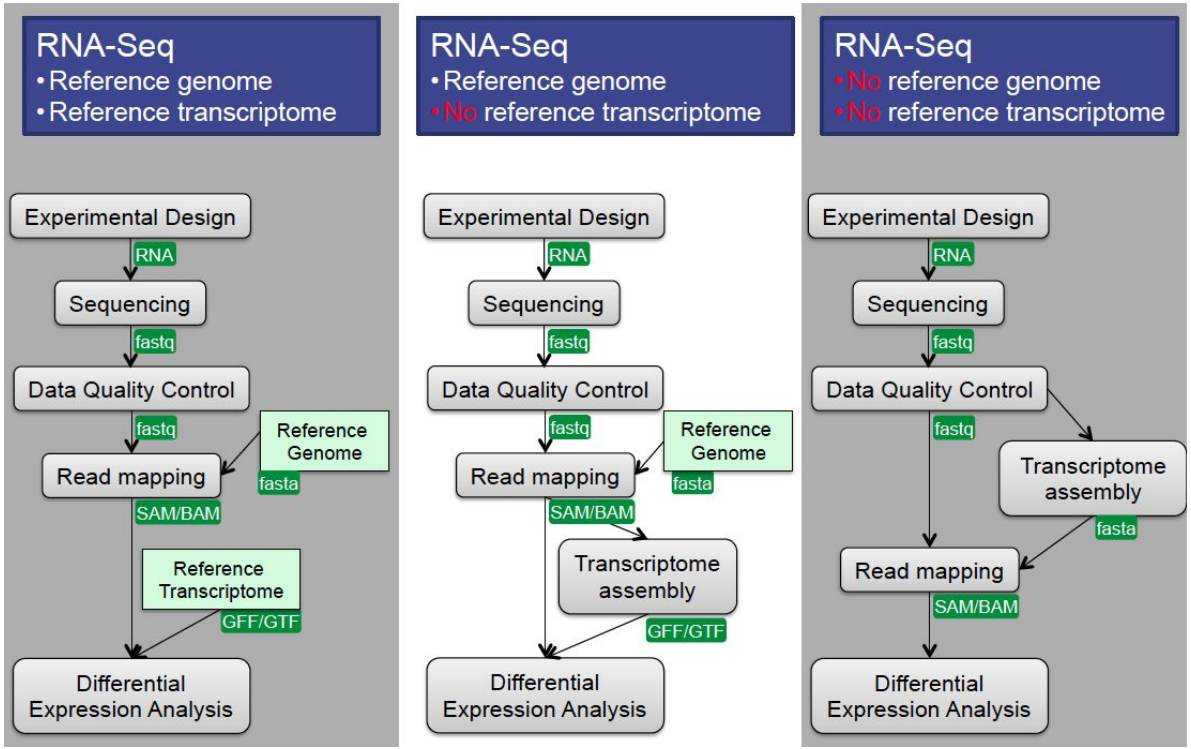
Steps of a De novo assembly



Simplify by merging nodes connected by single edges. The enumerate all paths in the graph

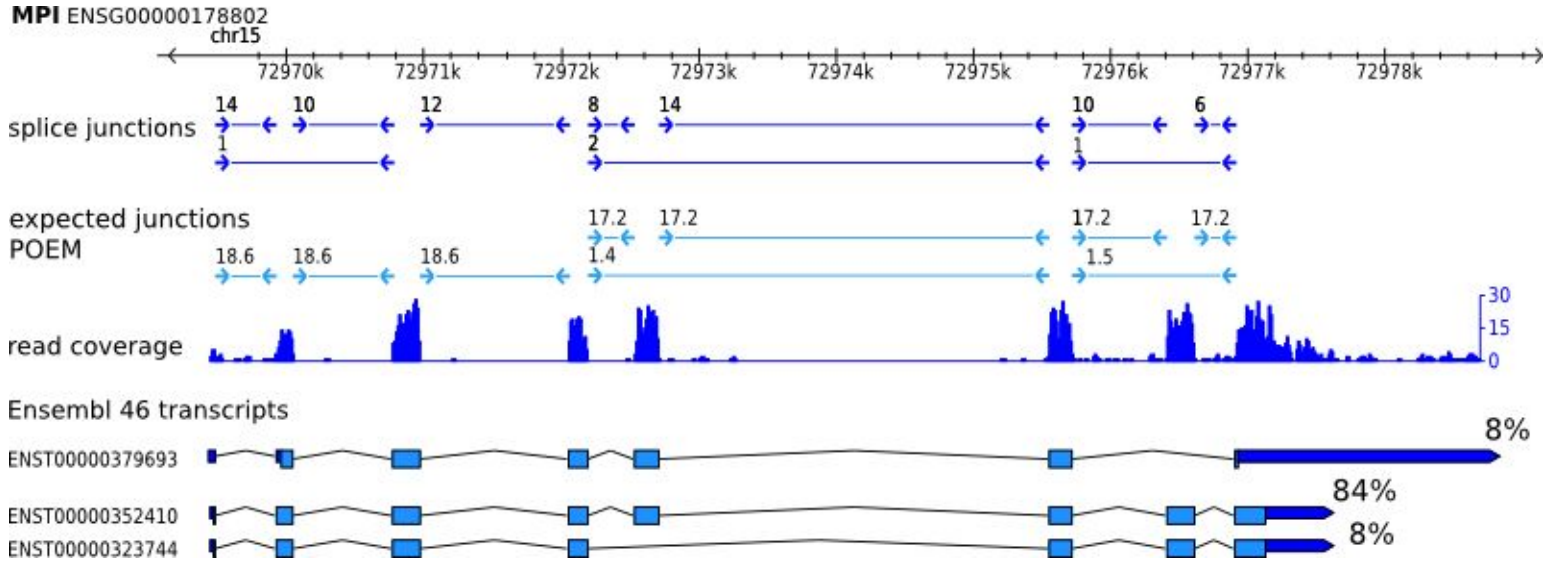


Typical RNA Processing Pipelines

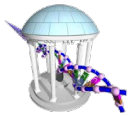


RNA Expression Analysis

Relative transcript abundances. Given N transcripts estimate how many of each best approximate the observed read coverage.



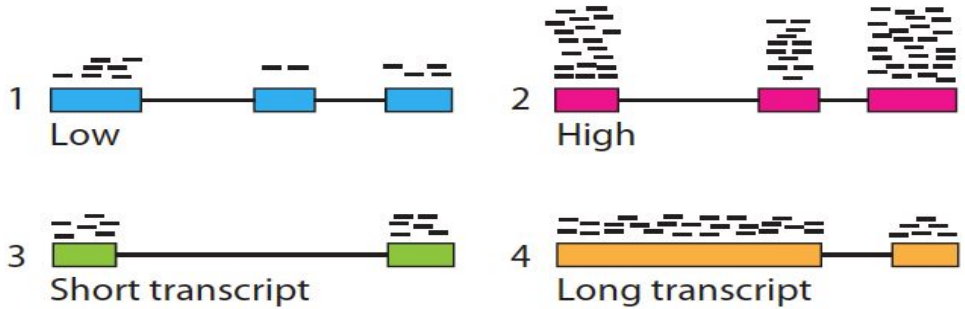
Read counts to transcript expression estimates

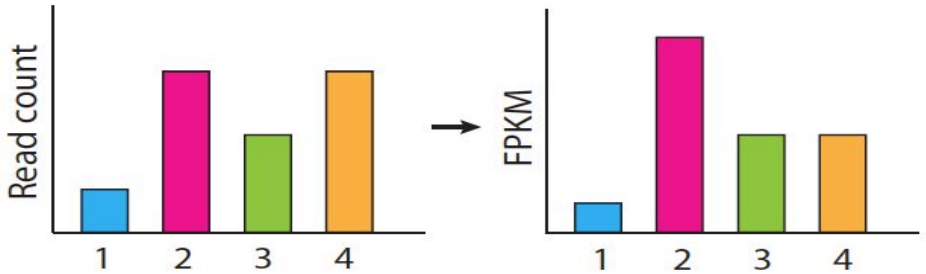


Small transcripts with a few reads might be expressed more than a larger transcript with more reads.

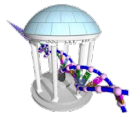
Need to add a normalization of read counts to accommodate for transcript sizes.

Convert reads per transcript to Fragments Per Kilobase of transcript per Million mapped reads.

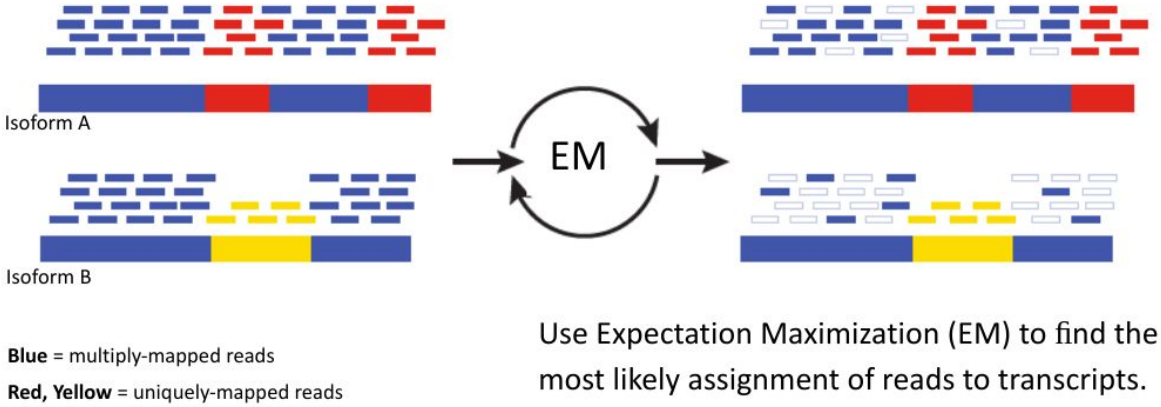




Expectation Maximization algorithms

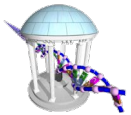


Expectation Maximization (EM) to find the most likely assignment of reads to the transcripts.



- Cufflinks and Cuffdiff (Tuxedo)
- RSEM
- eXpress
- Salmon/kallisto (non-alignment based)

Non-alignment based RNA quantification



Assume we have two things:

A model of the transcriptome

Counts of k-mer frequencies from our sequenced RNAseq reads

Make a table of k-mers from the transcriptome

	ACAGC	TCCTG	AGCGC	CTCTT	GGTCC	GCGCT	TTCCT	Abundance
$GeneAt_1$	1	0	0	0	1	0	1	8
$GeneBt_1$	0	1	0``	0	0	1	1	3
$GeneCt_1$	1	0	1	1	0	1	0	4
$GeneCt_2$	0	0	1	0	0	1	0	1
counts	12	3	5	4	8	8	11	

Next Time

RNAseq pipelines