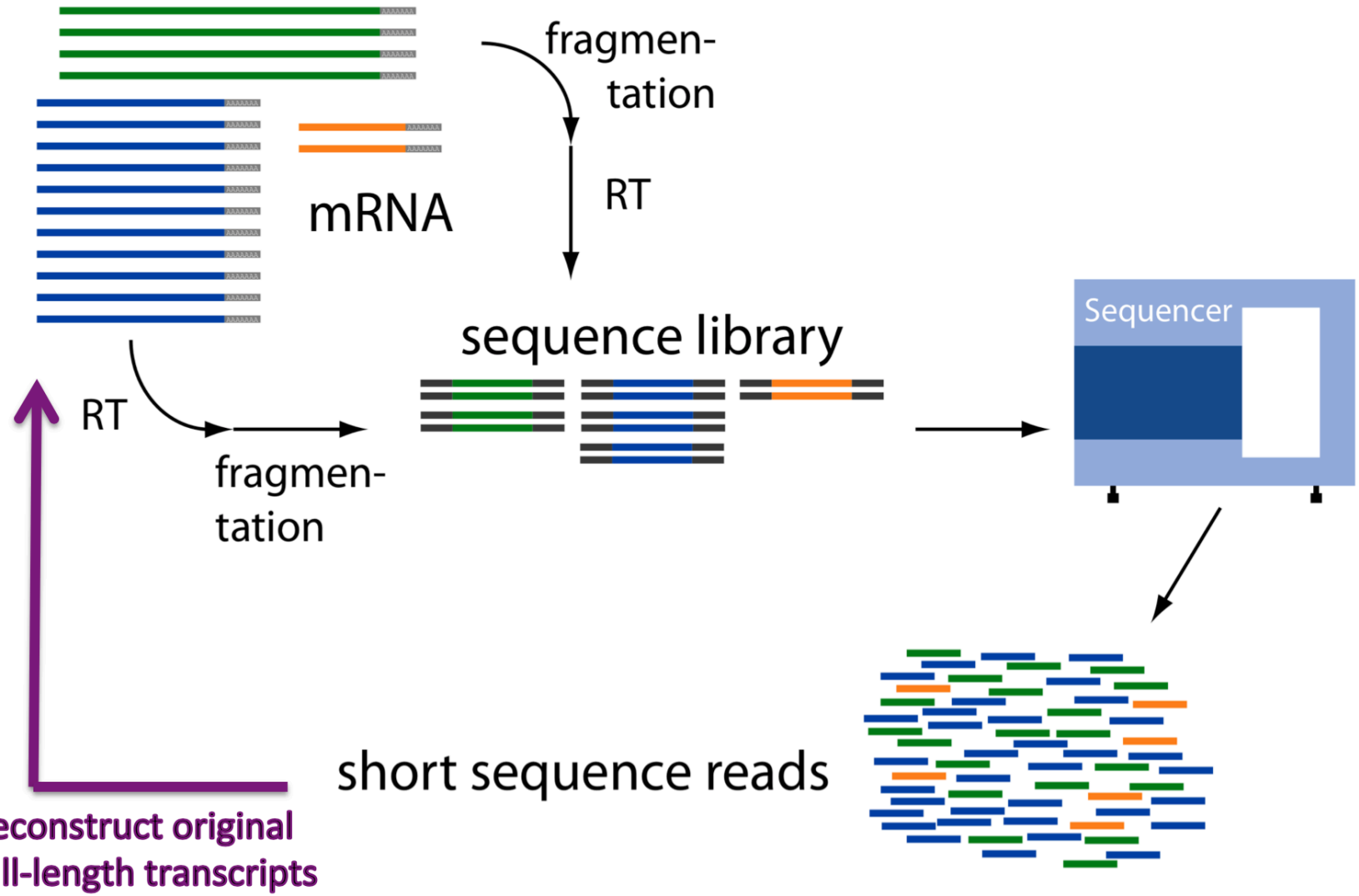
BI694

Bioinformatics & Phylogenetics

Winter Semester 2017

Read Mapping and DE Analysis Review

Overview of RNA-Seq



From: http://www2.fml.tuebingen.mpg.de/raetsch/members/research/transcriptomics.html

Transcript Reconstruction from RNA-Seq Reads

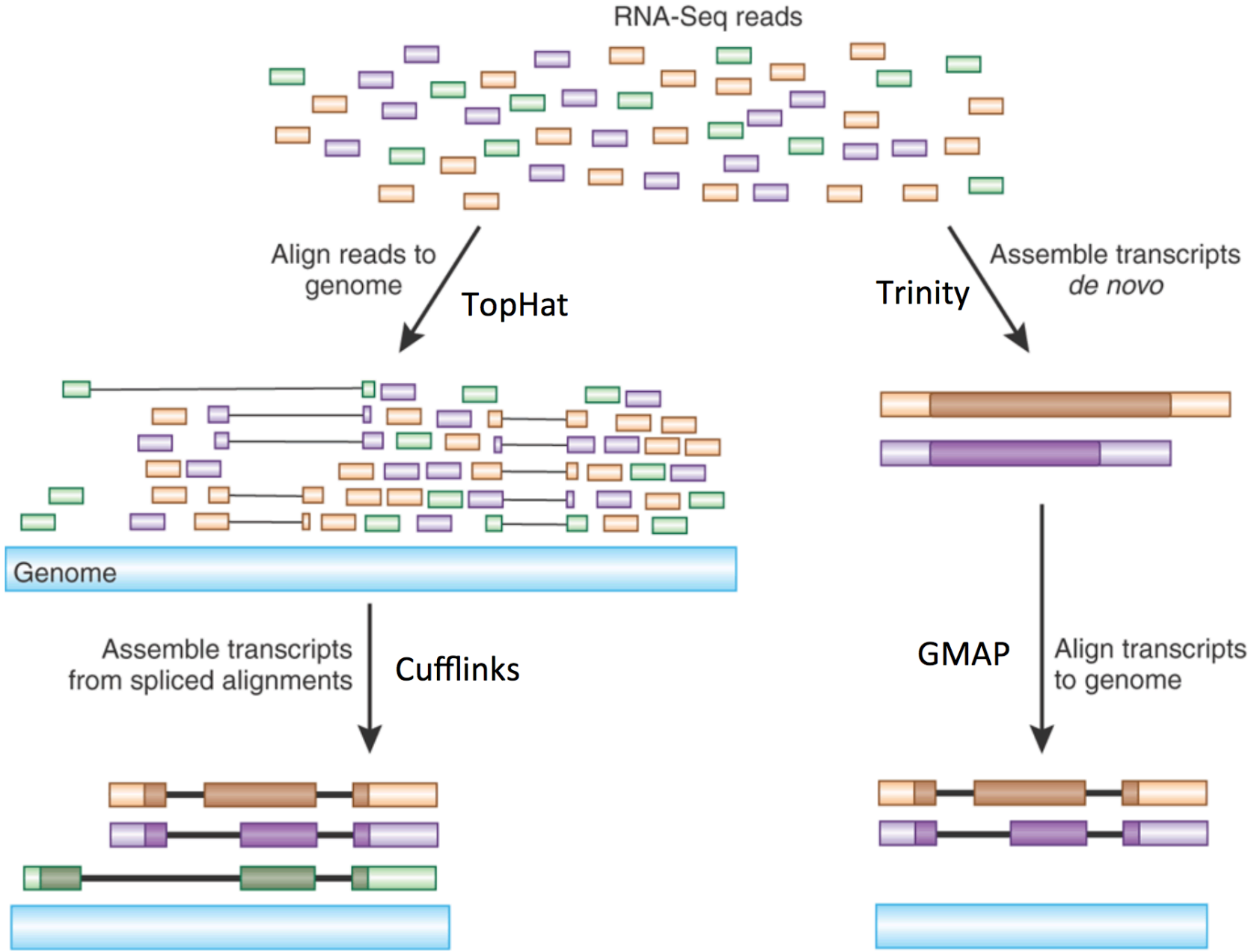
Advancing RNA-Seq analysis

Brian J Haas & Michael C Zody

Nature Biotech, 2010

New methods for analyzing RNA-Seq data enable *de novo* reconstruction of the transcriptome.

Transcript Reconstruction from RNA-Seq Reads



De novo transcriptome assembly

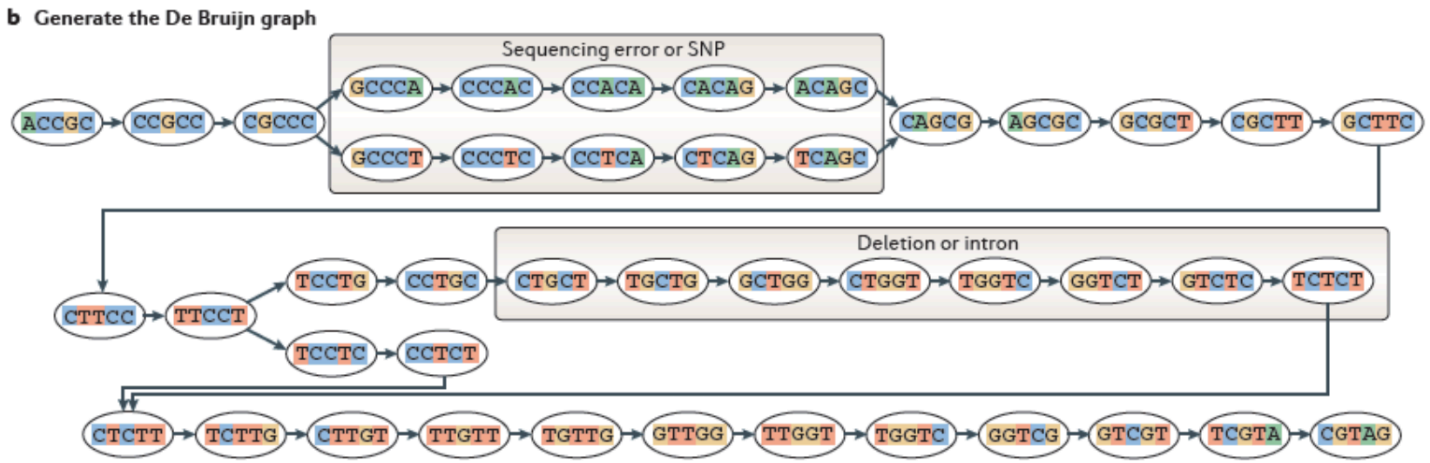
No genome required

Empower studies of non-model organisms

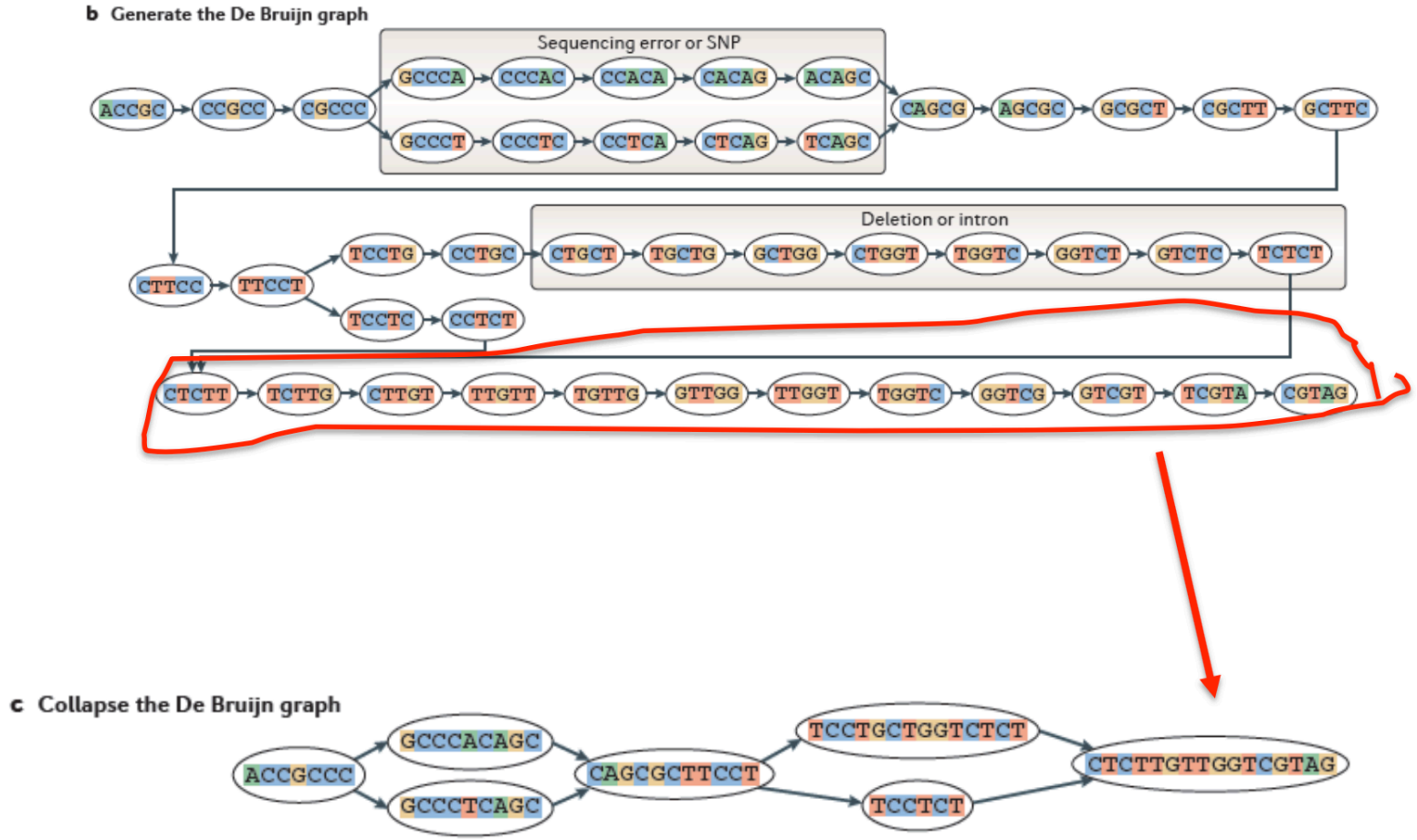
- expressed gene content
- transcript abundance
- differential expression

The General Approach to *De novo* RNA-Seq Assembly Using De Bruijn Graphs

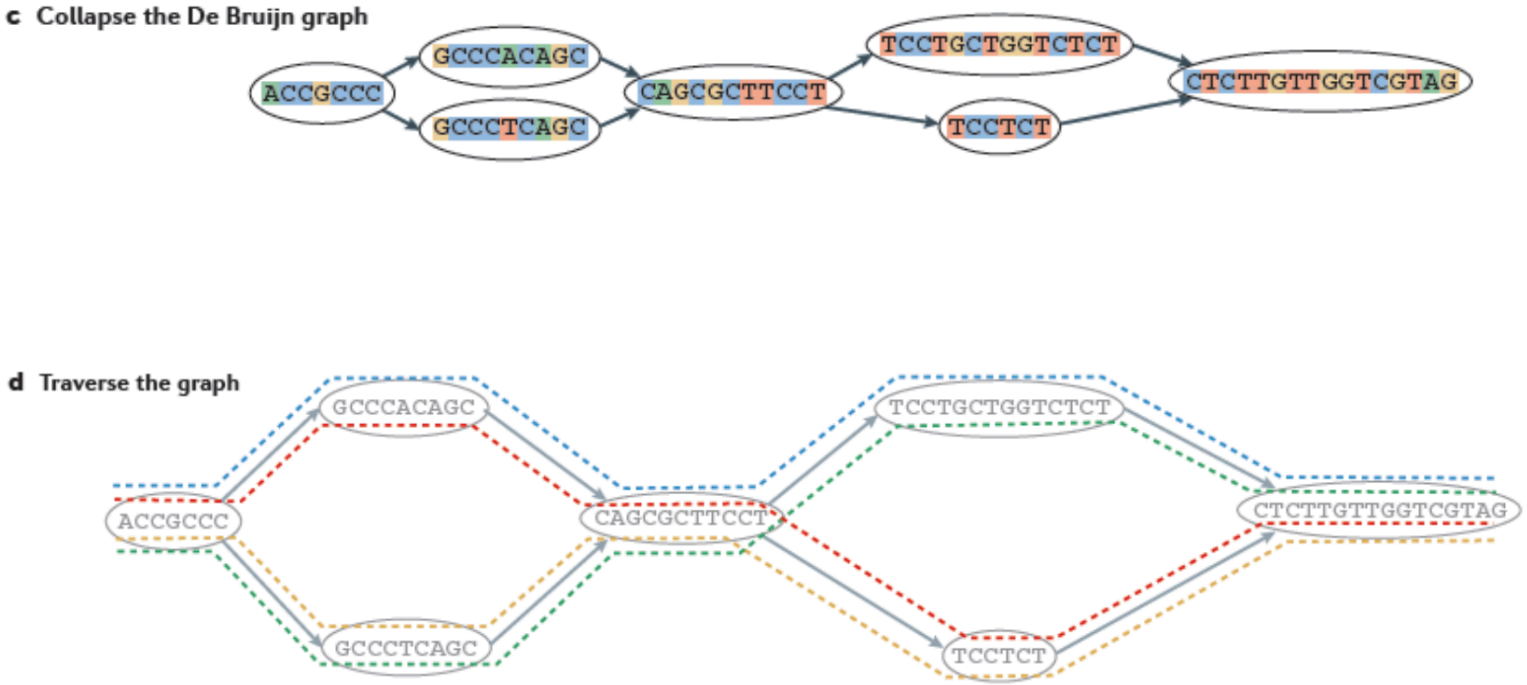
Sequence Assembly via De Bruijn Graphs



From Martin & Wang, Nat. Rev. Genet. 2011



From Martin & Wang, Nat. Rev. Genet. 2011



From Martin & Wang, Nat. Rev. Genet. 2011

Contrasting Genome and Transcriptome Assembly

Genome Assembly

- Uniform coverage
- Single contig per locus
- Double-stranded

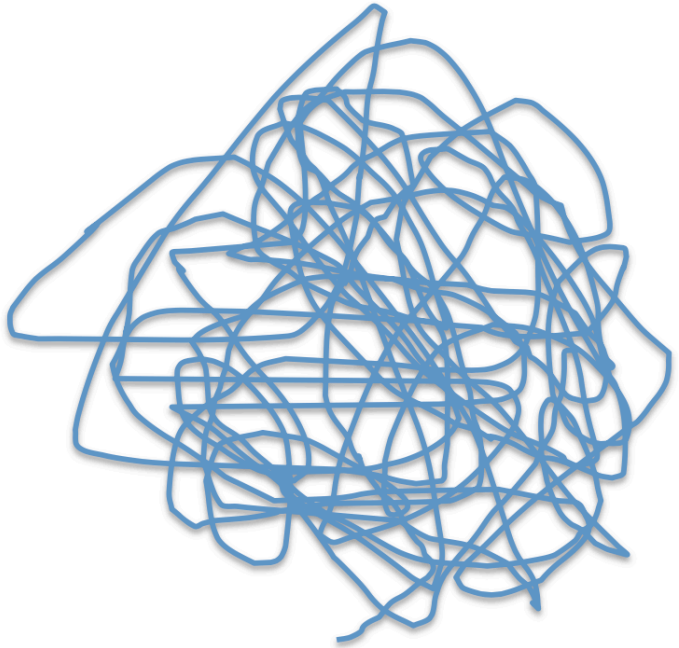
Transcriptome Assembly

- Exponentially distributed coverage levels
- Multiple contigs per locus (alt splicing)
- Strand-specific

Trinity Aggregates Isolated Transcript Graphs

Genome Assembly

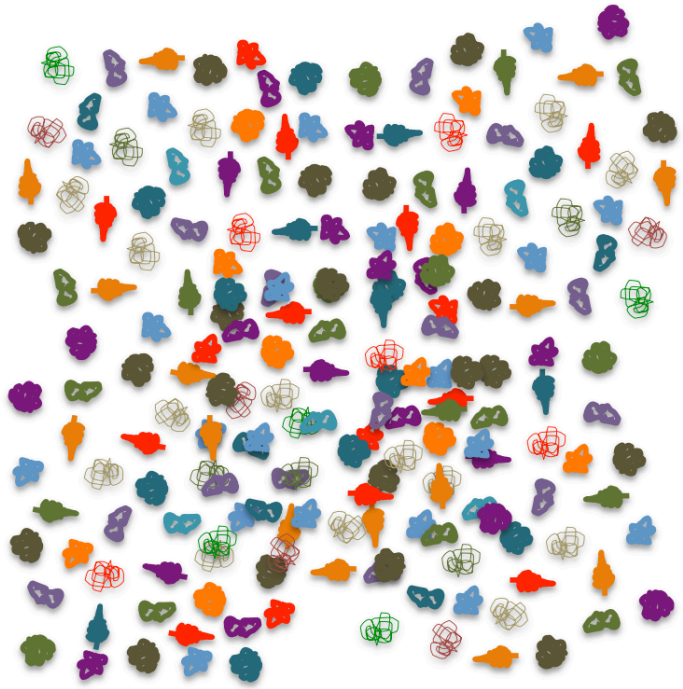
Single Massive Graph



Entire chromosomes represented.

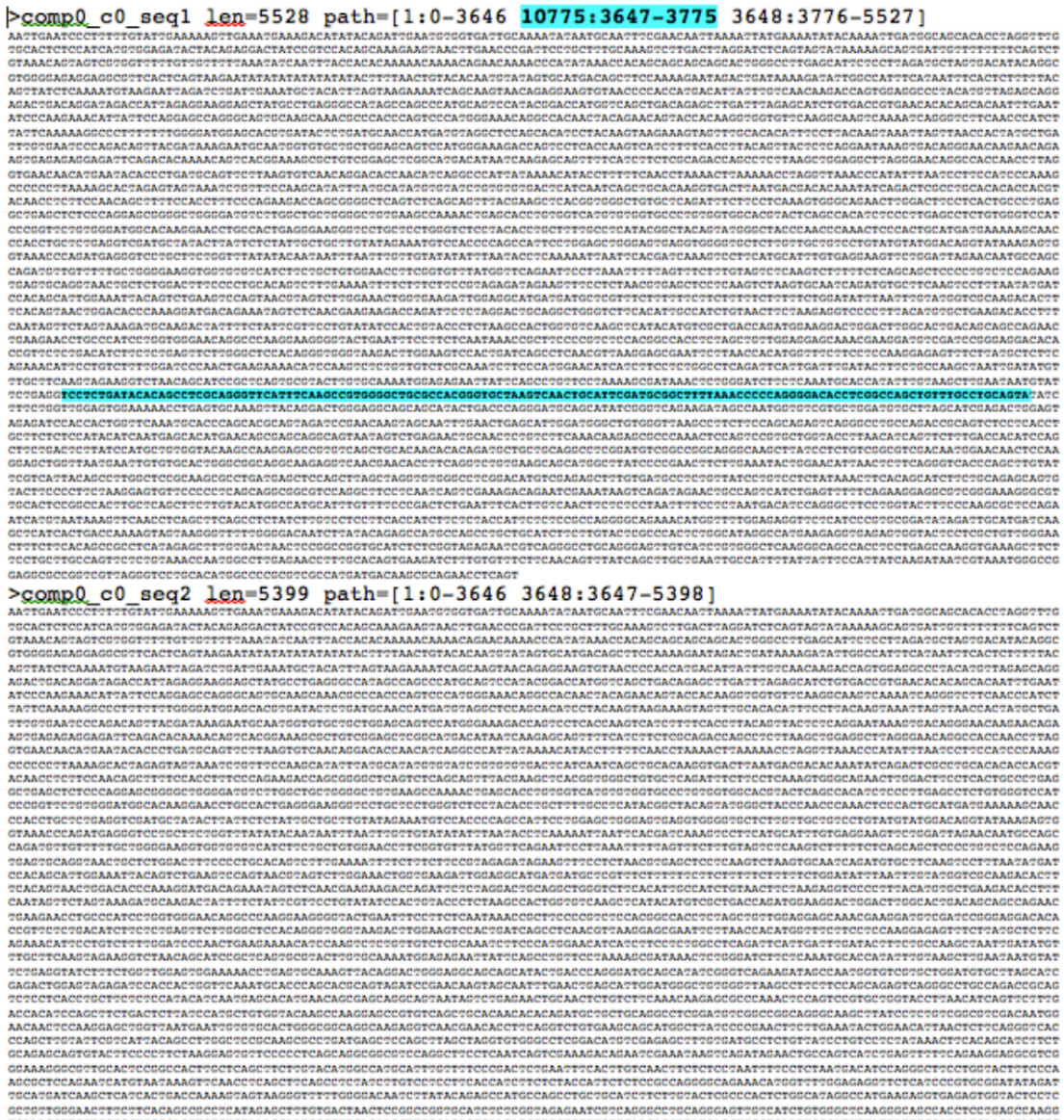
Trinity Transcriptome Assembly

Many Thousands of Small Graphs



Ideally, one graph per expressed gene.

Trinity output: A multi-fasta file



Abundance Estimation
(Aka. Computing Expression Values)

How abundant is a transcript?

GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT Transcript

GGCGTCTATATCT

GGCGTCTATATCTCG

TATCTCGGCTCTAGG

TATCTCAGCTCTAGGCC Mapped reads

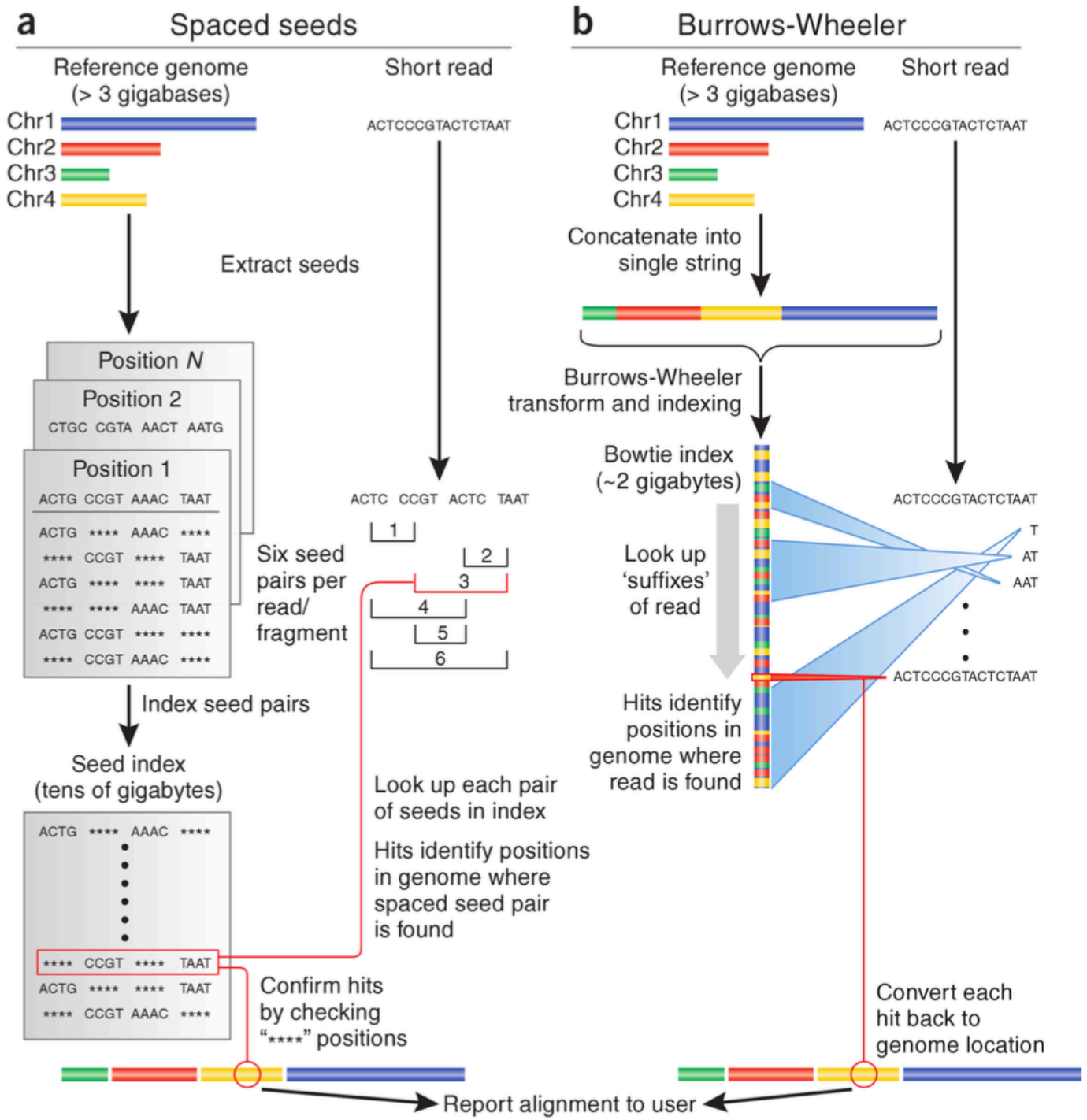
TATCTCAGCTCTAGGCCCTCA

CTCGGCTCTAGGCCCTCATTTT

GGCTCTAGGCCCTCATTTTTT

CTCTAGGCCCTCATTTTTT

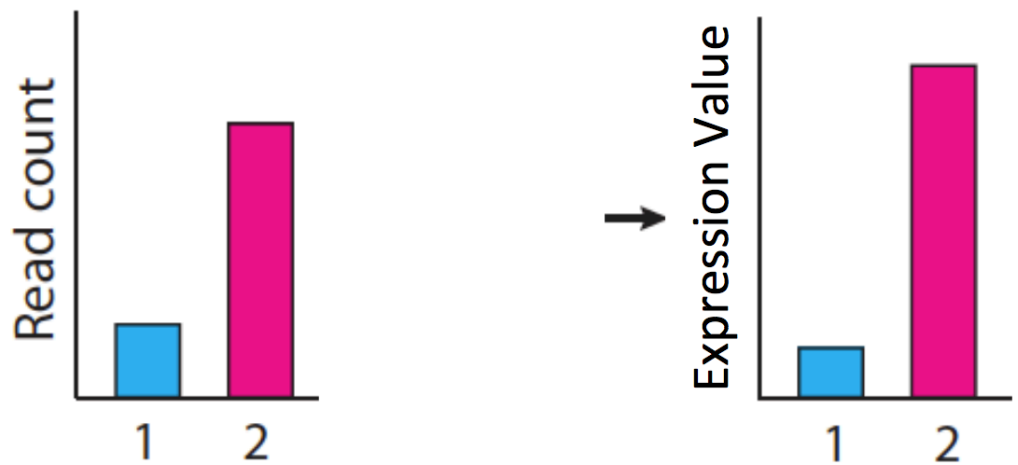
CTAGGCCCTCATTTTTT



What is a Burrows Wheeler Transform?

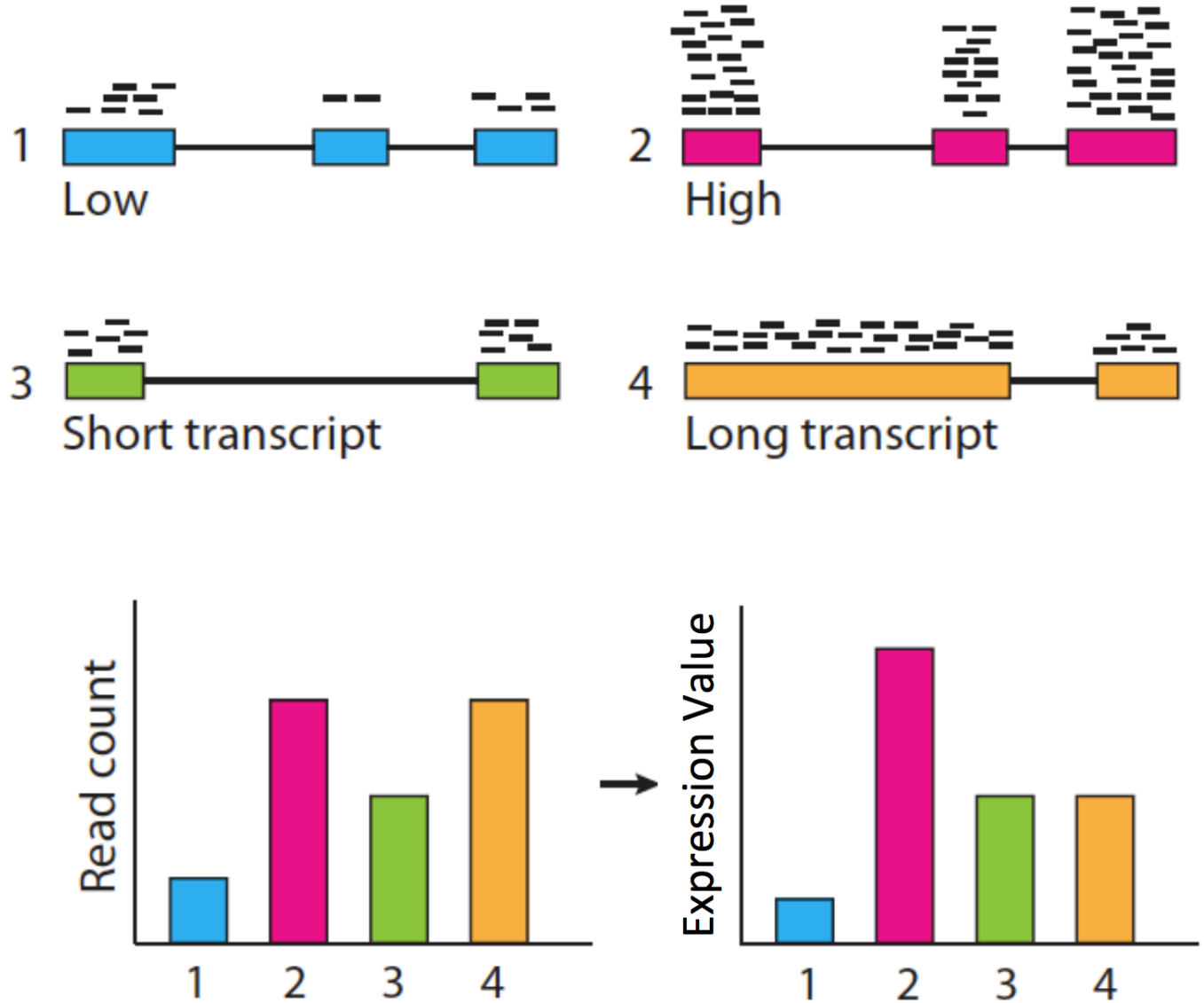
Trapnell & Salzberg (2009) Nat Biotechnol

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

Normalized Expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.

- Reported as: Number of RNA-Seq **F**ragments
  Per **K**ilobase of transcript
  per total **M**illion fragments mapped

FPKM

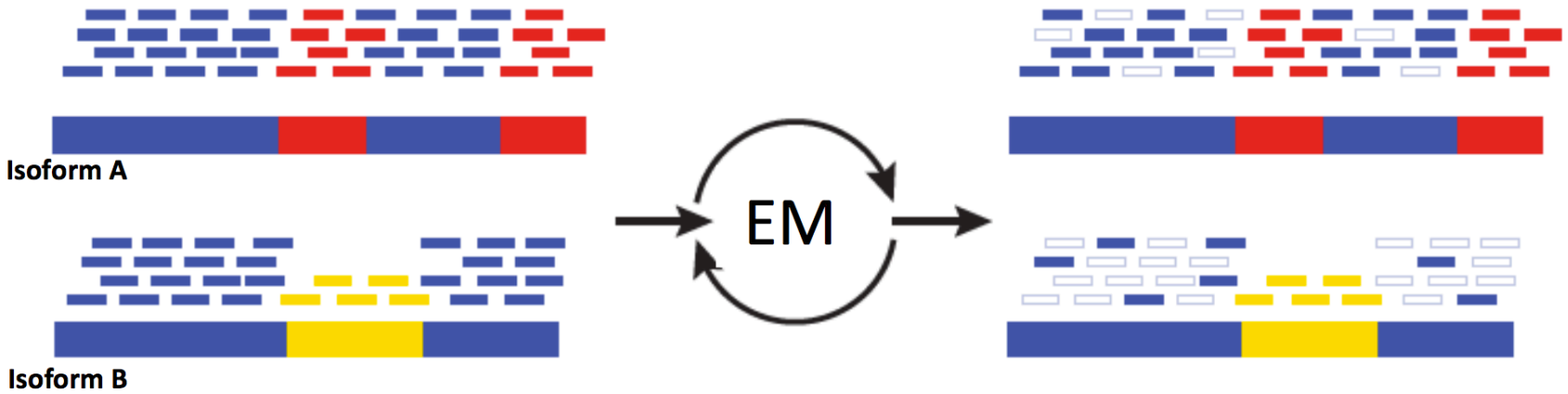
Multiply-mapped Reads Confound Abundance Estimation

Isoform A

Isoform B

Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

Multiply-mapped Reads Confound Abundance Estimation



Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

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

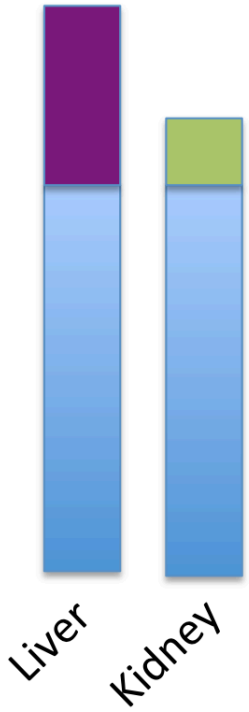
Performed by:
RSEM, eXpress, kallisto, salmon, …

New fast alignment-free methods now available! eg. Kallisto

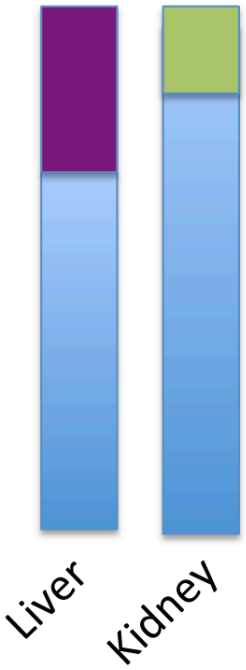
Differential Expression Analysis Using RNA-Seq

Why cross-sample normalization is important

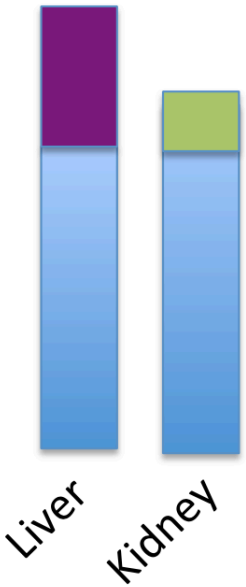
Absolute RNA quantities per cell



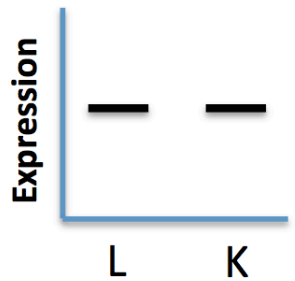
Measured relative abundance via RNA-Seq

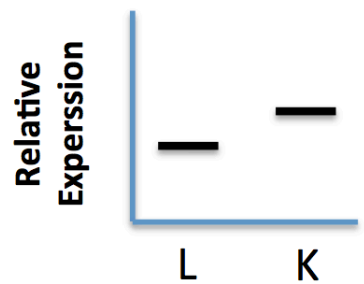


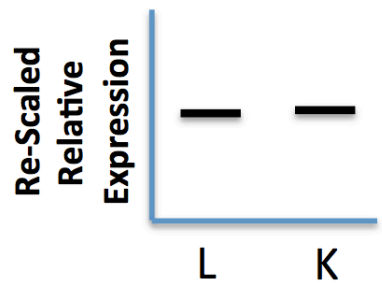
Cross-sample normalized (rescaled) relative abundance



eg. Some housekeeping gene's expression level:







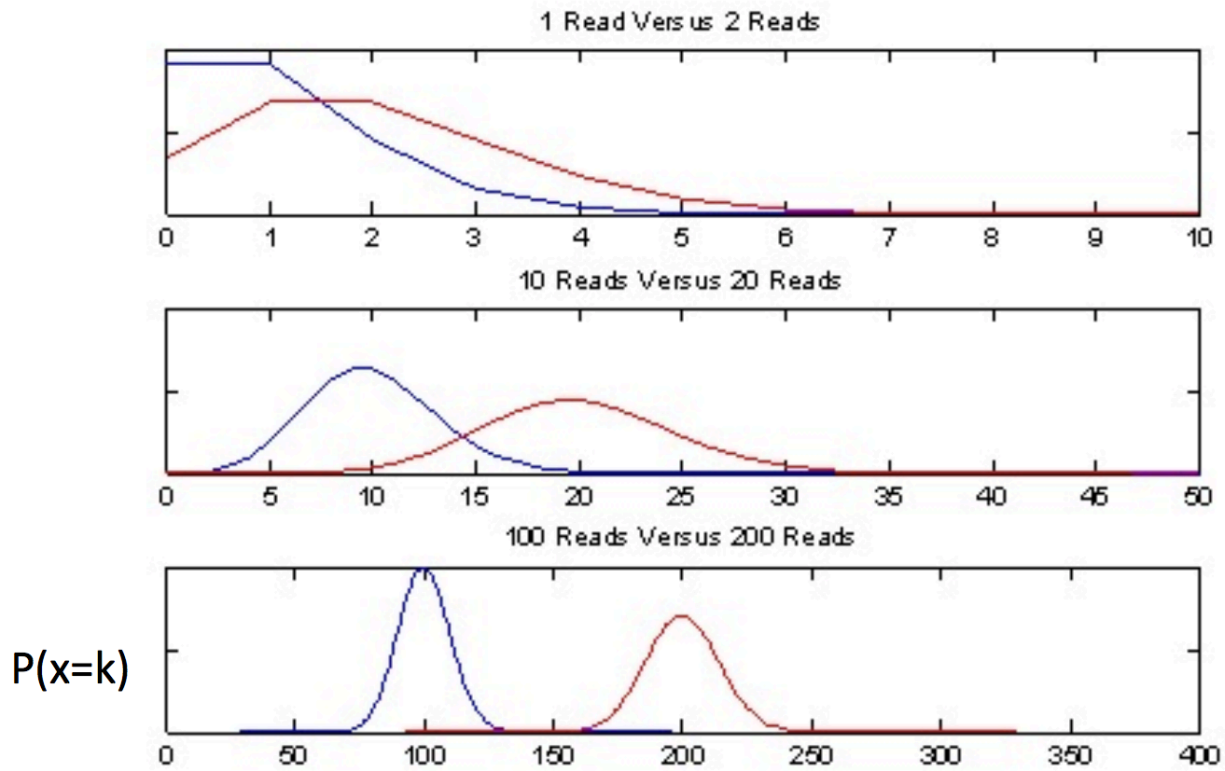
Diff. Expression Analysis Involves

- Counting reads
- Statistical significance testing

	Sample_A	Sample_B	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Beware of concluding fold change from small numbers of counts

Poisson distributions for counts based on **2-fold** expression differences



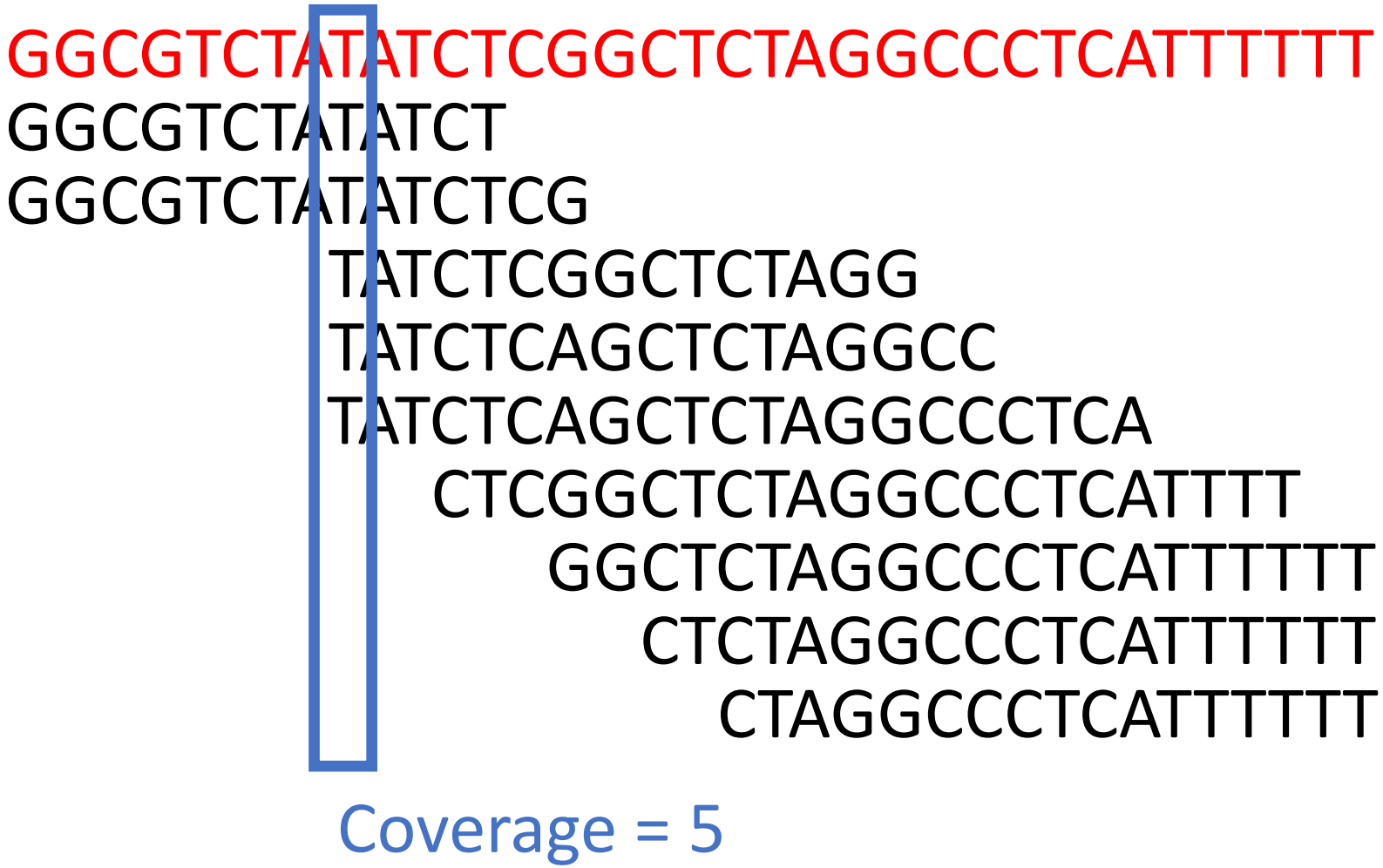
No confidence in 2-fold difference. Likely observed by chance.

High confidence in 2-fold difference. Unlikely observed by chance.

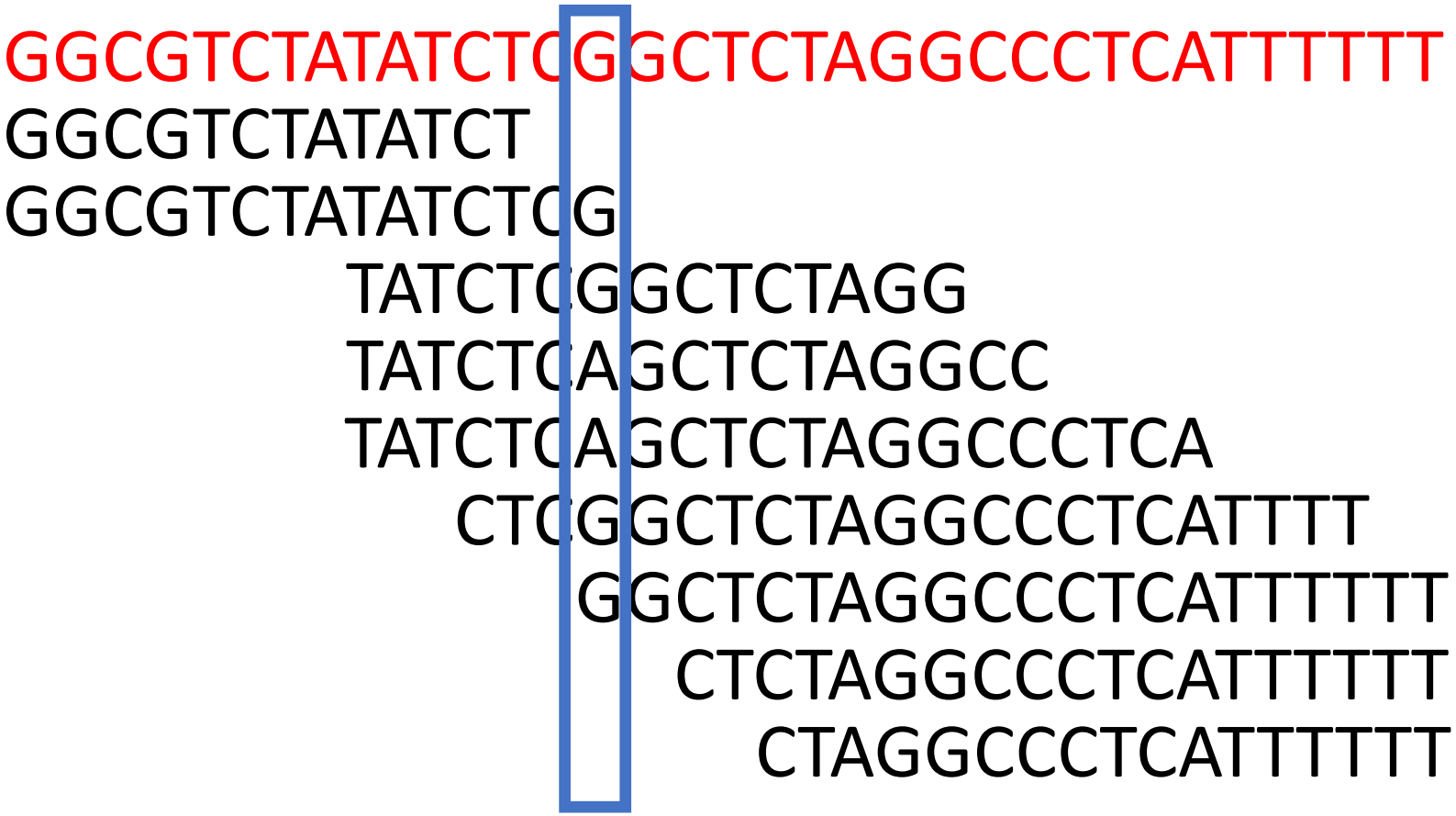
From: http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for

Mapping can tell us something about structural variation…

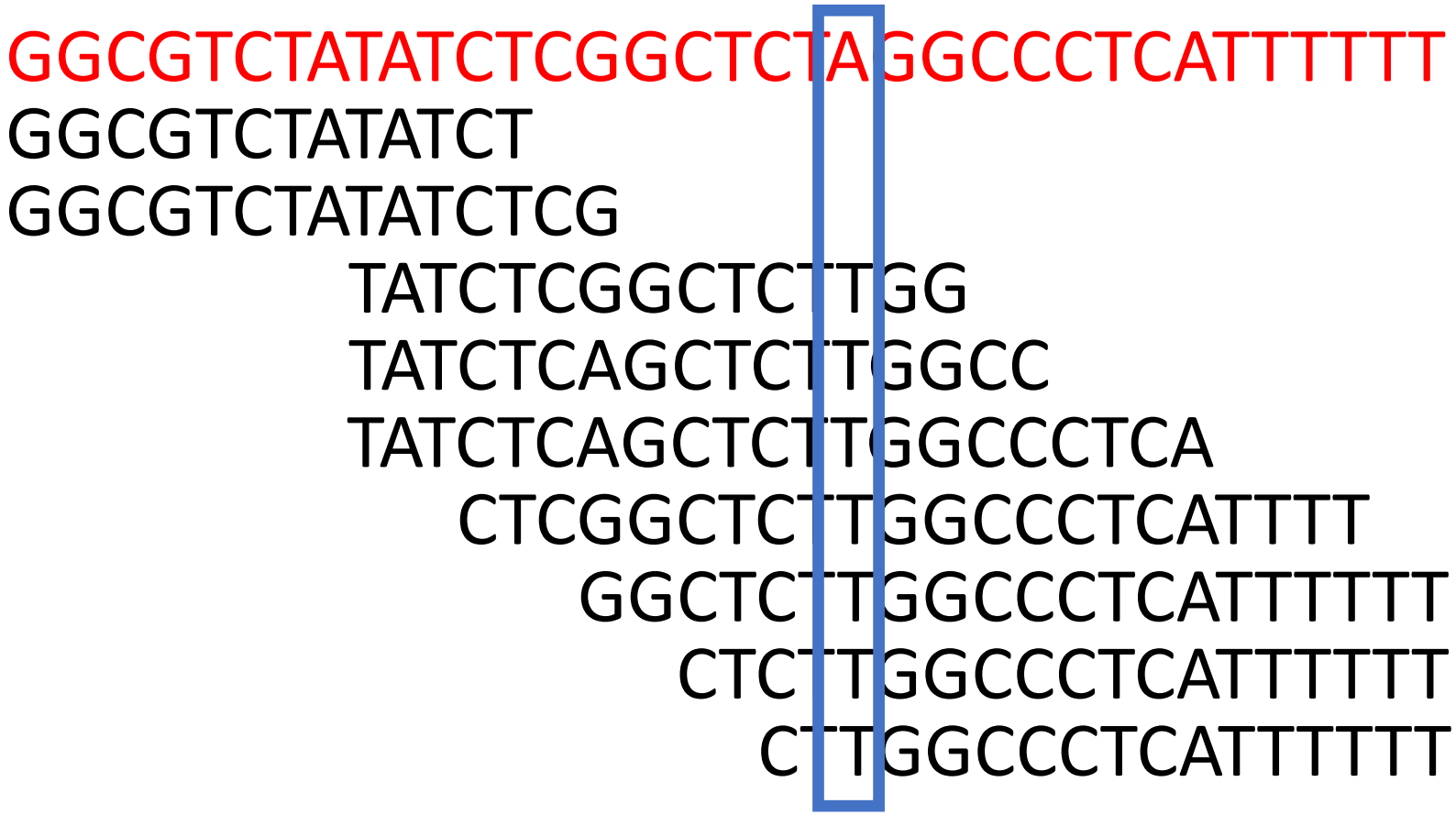
SNP Calling


GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCT
GGCGTCTATATCTCG
TATCTCGGCTCTAGG
TATCTCAGCTCTAGGCC
TATCTCAGCTCTAGGCCCTCA
CTCGGCTCTAGGCCCTCATTTT
GGCTCTAGGCCCTCATTTTTT
CTCTAGGCCCTCATTTTTT
CTAGGCCCTCATTTTTT
Coverage = 5

SNP Calling


GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCT
GGCGTCTATATCTCG
TATCTCGGCTCTAGG
TATCTCAGCTCTAGGCC
TATCTCAGCTCTAGGCCCTCA
CTCGGCTCTAGGCCCTCATTTT
GGCTCTAGGCCCTCATTTTTT
CTCTAGGCCCTCATTTTTT
CTAGGCCCTCATTTTTT

SNP Calling


GGCGTCTATATCTCGGCTCTAGGCCCTCATTTTTT
GGCGTCTATATCT
GGCGTCTATATCTCG
TATCTCGGCTCTTGG
TATCTCAGCTCTTGGCC
TATCTCAGCTCTTGGCCCTCA
CTCGGCTCTTGGCCCTCATTTT
GGCTCTTGGCCCTCATTTTTT
CTCTTGGCCCTCATTTTTT
CTTGGCCCTCATTTTTT

SNP Calling

Basic Procedure:

- Map read to reference
- Read out differences

Considerations:

- Reads a re short but genomes/transcriptomes long and complex
- Is the mapping position unique? Think about abundance est…

- Reads contain errors and errors are not random
- Base quality score? Error model?