Sequence Analysis - RNA-Seq 2

RNA-Seq

- Identify sequence of RNA molecules
- “Unbiased” - possible to sequence any molecule in sample
- Molecules sequenced in proportion to relative abundance in sample
- Most often used for gene abundance estimation

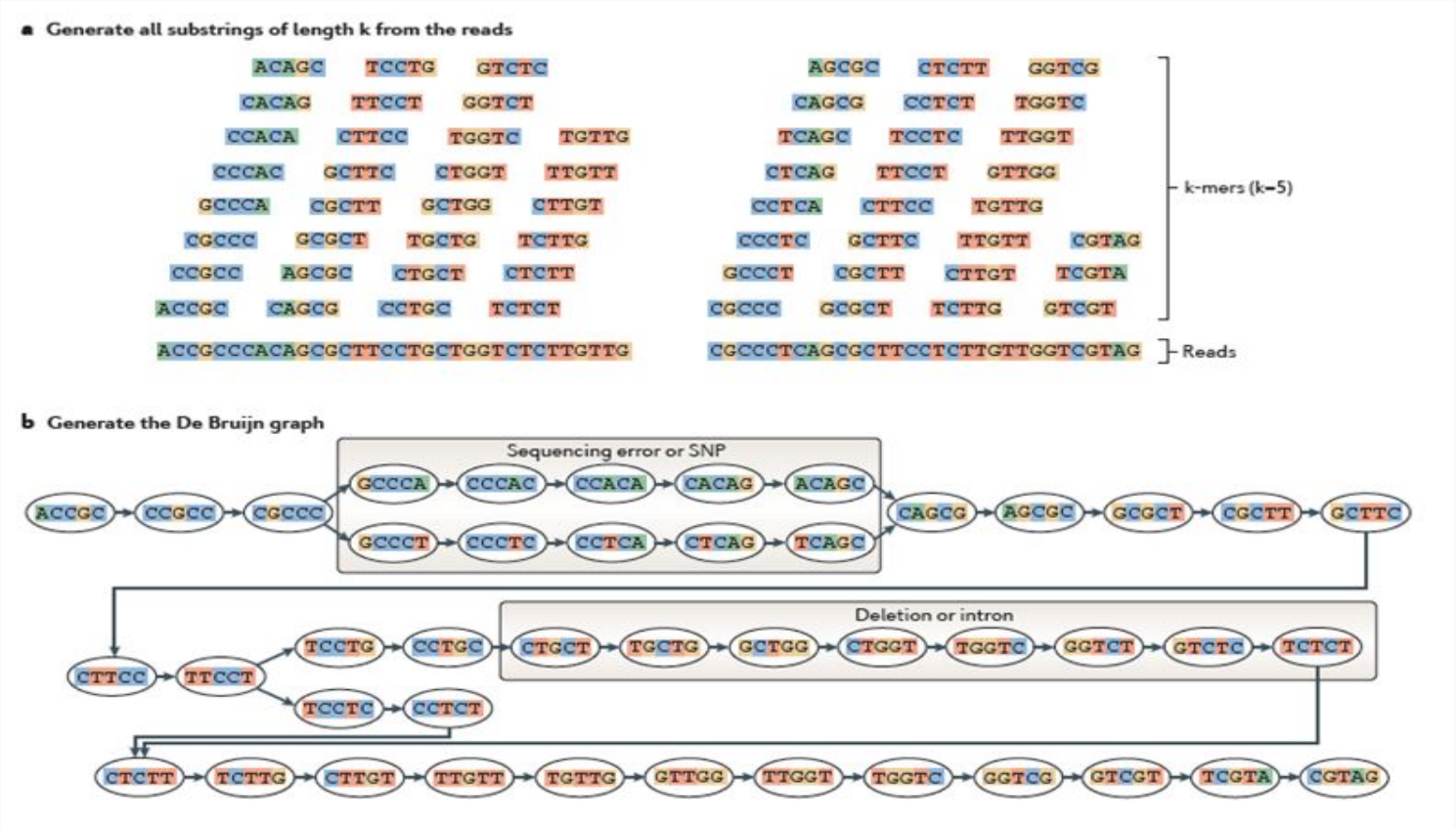
Common Types of RNA-Seq Analyses

- Sequence based
  - Transcriptome reconstruction
  - Splicing analysis
  - Gene fusion discovery
  - Coding variants
- Abundance based
  - Differential expression
  - Allele-specific expression (with genotyping)

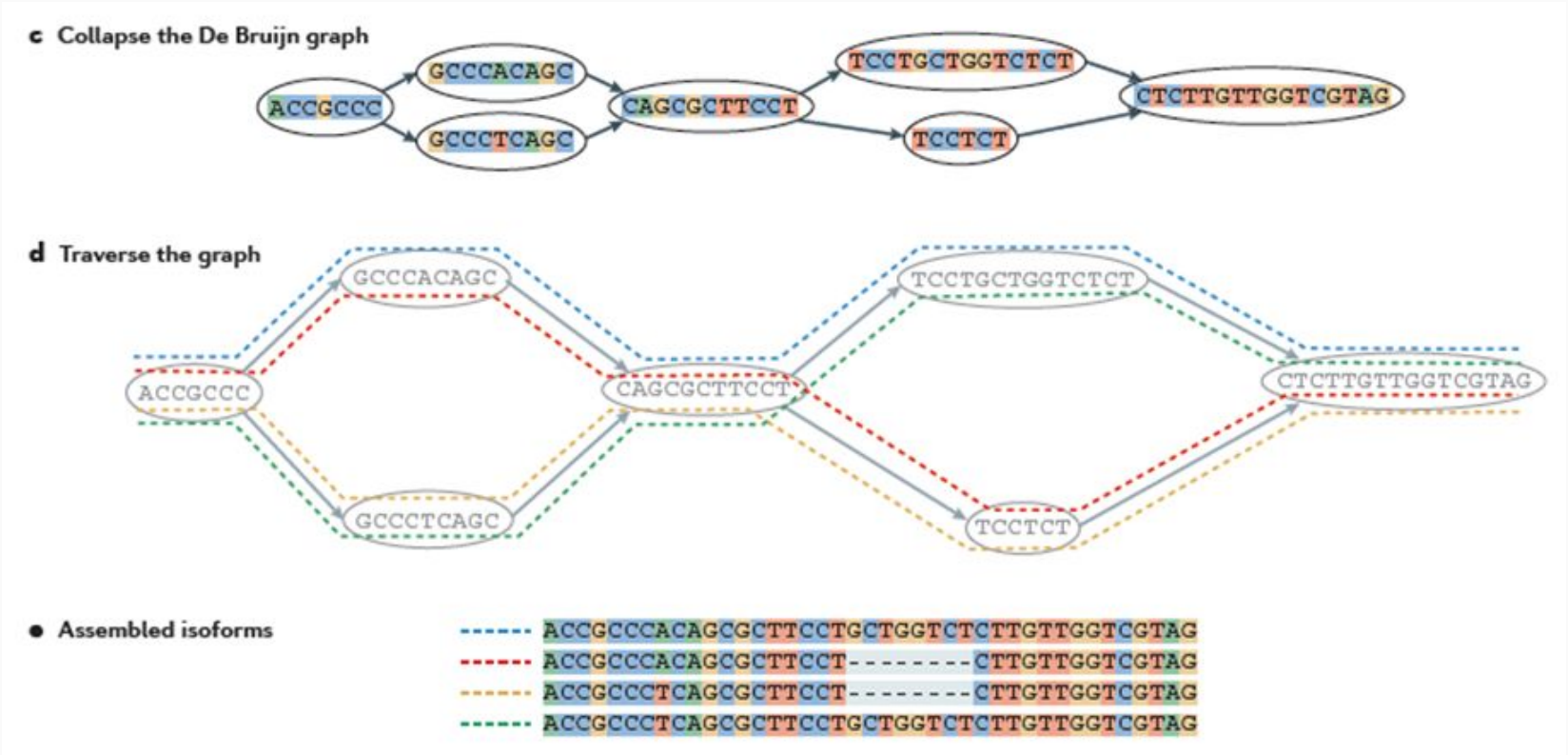
Transcriptome reconstruction

- Given RNA fragments, recover original transcript
- Two approaches:
  - *De novo* - no reference transcriptome available
  - *Reference or genome guided* - reference available
- Very challenging with short reads!

De novo transcriptome reconstruction

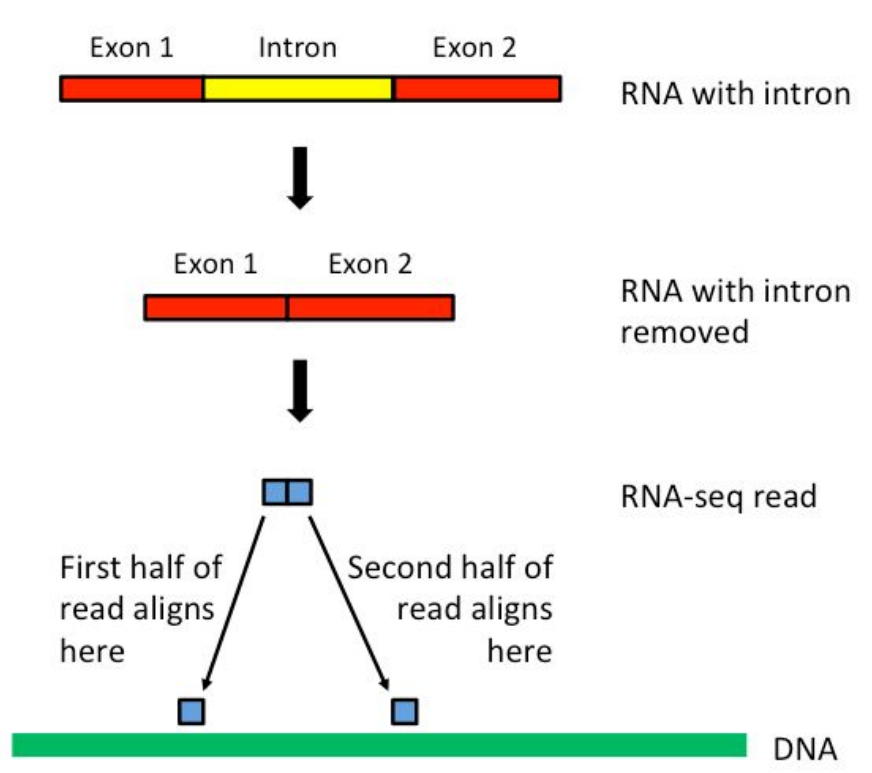


De novo transcriptome reconstruction



Concept: Spliced Alignment

- Coding sequences interrupted by introns
- mRNA molecules have introns excised
- Some reads span *splice junctions*
- Spliced alignment aligns *junction* or *spliced reads*



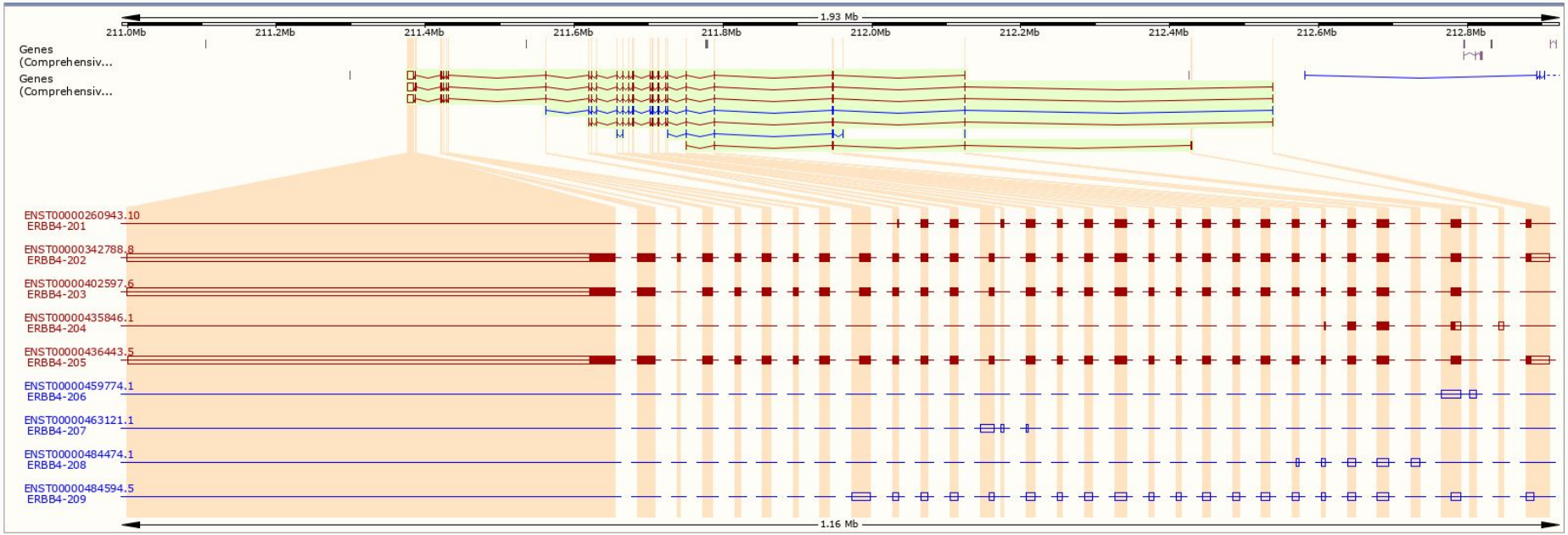
https://discoveringthegenome.org/discovering-genome/rna-sequencing-up-close-data/spliced-alignment

Concept: Spliced Alignment

- Splicing-aware programs:
  - STAR, TopHat
- Splicing *unaware* programs:
  - bwa, bowtie, most others
- *De novo* - use only genome sequence
- *Transcriptome guided* - use known splice junctions to guide alignment

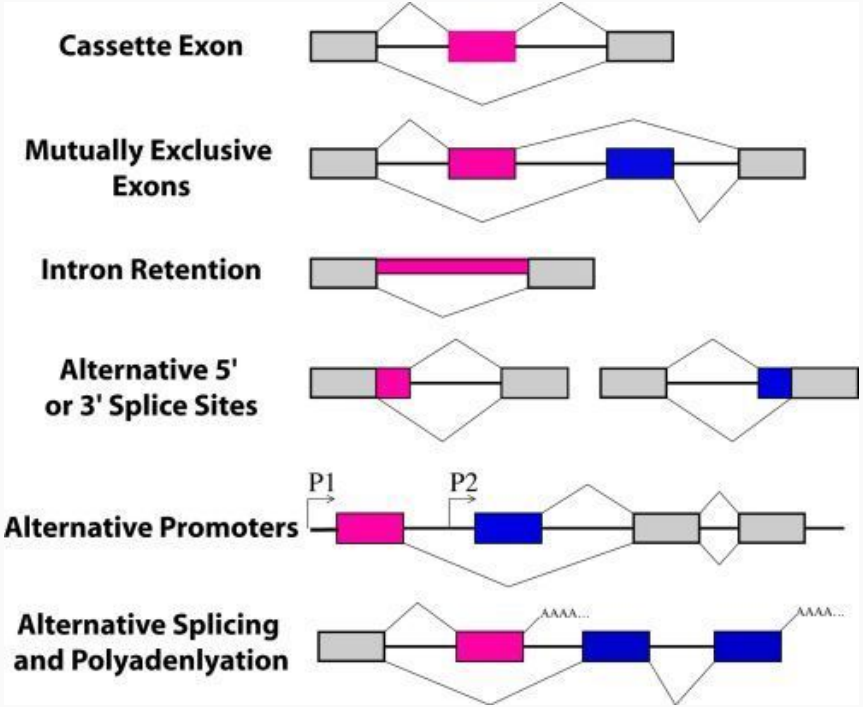
Concept: mRNA Isoforms

- **Isoform**: pattern of exons/transcribed sequence



Alternative Splicing (AS) Analysis

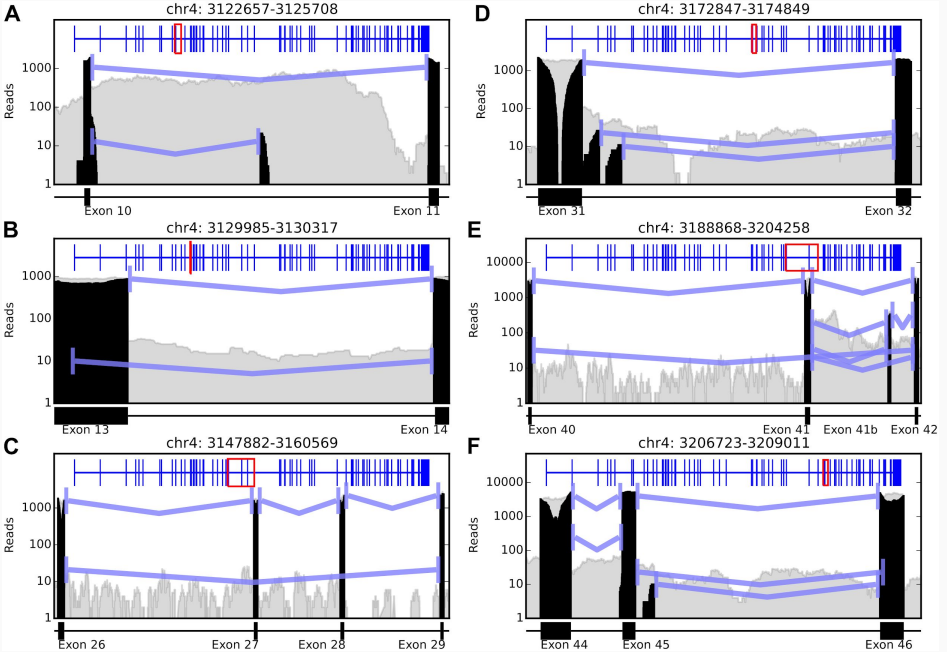
- Detect and/or quantify isoforms
- Different AS types
- Examine pattern of exons in spliced reads
- Methods:
  - Whippet
  - MISO
  - rMATS
  - IRFinder, and many others



https://www.ncbi.nlm.nih.gov/books/NBK19730/figure/A1150/

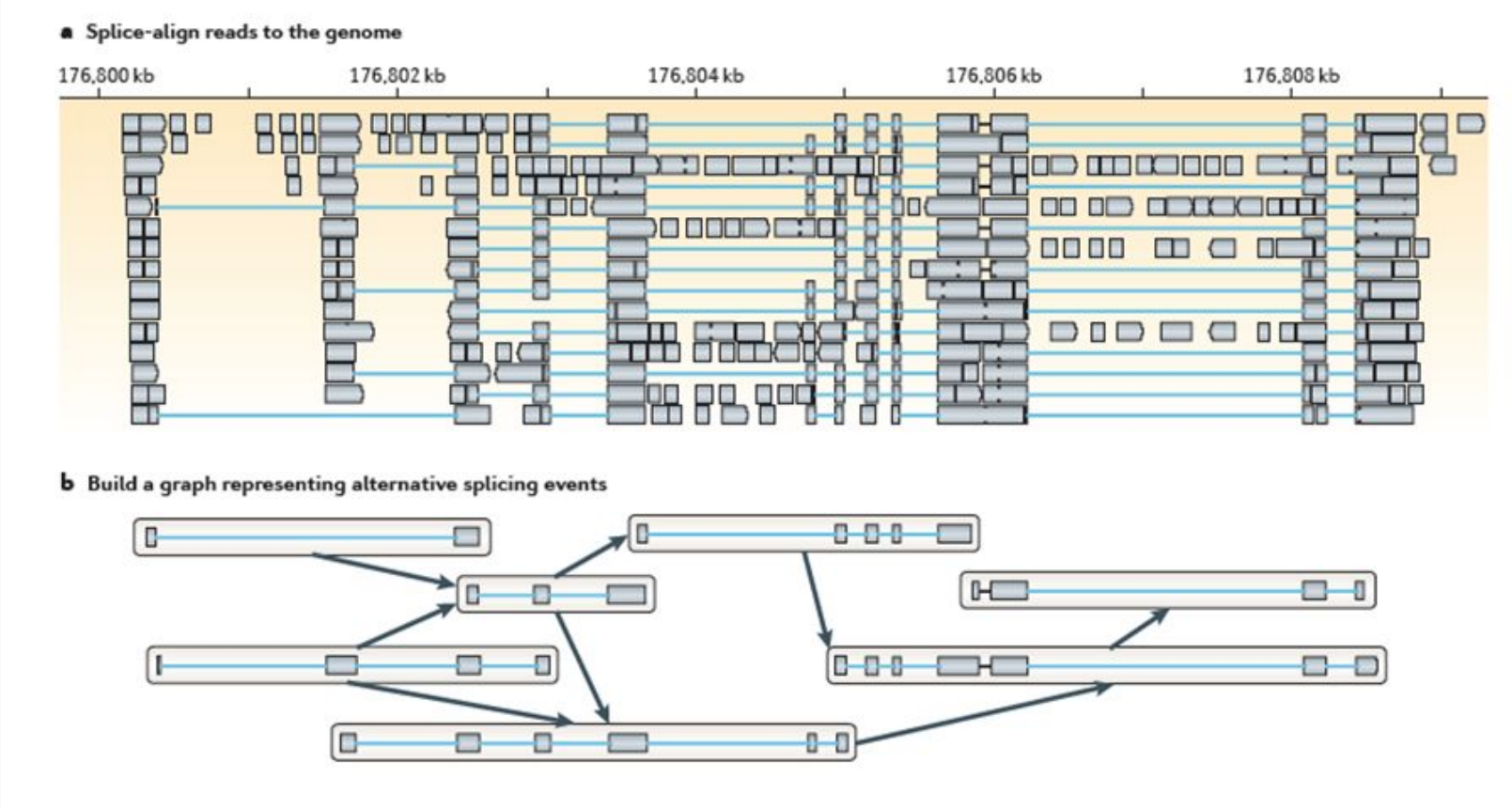
Alternative Splicing (AS) Analysis

- **Read support**: # of reads containing splice junction
- Grey areas → overall aligned read depth
- Black areas → spliced reads
- These have minimum 10 supporting reads per splicing event

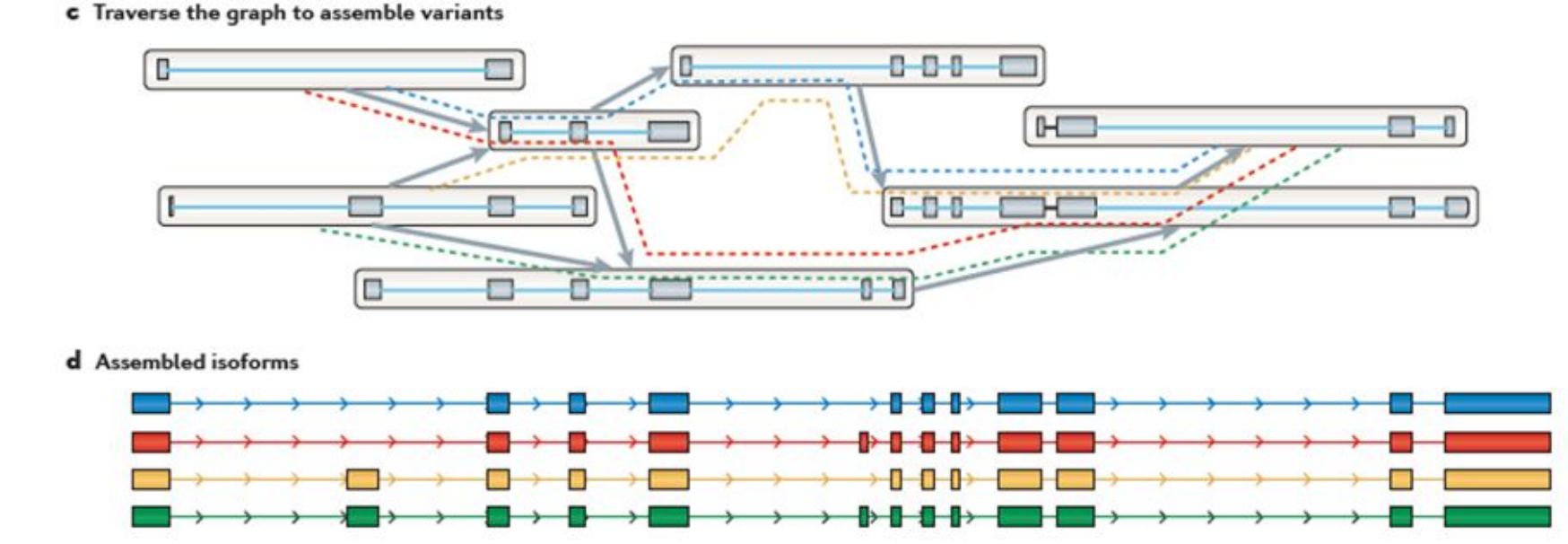


https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141298

Reference guided reconstruction

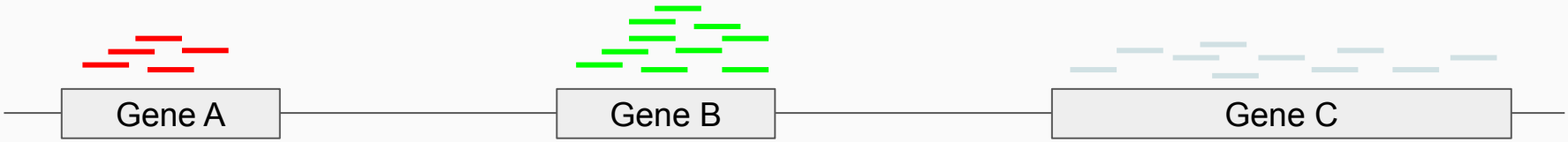


Reference guided reconstruction



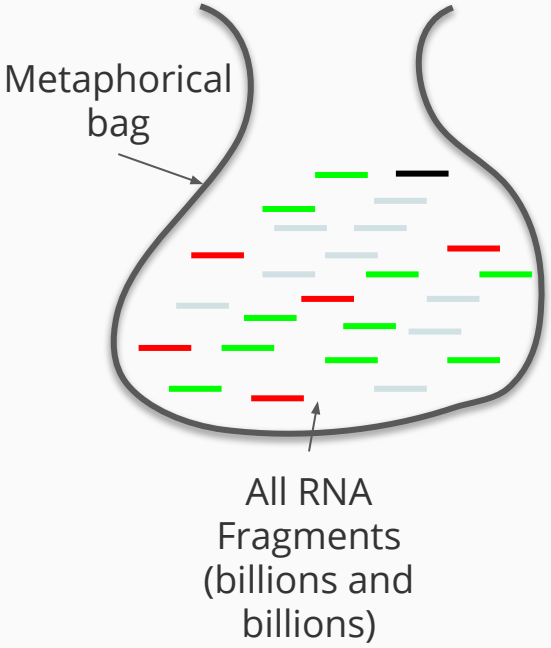
Gene Expression (RNA Abundance)

- Measure relative abundance of RNA species
- # reads mapping to a gene is proportional to # of molecules transcribed
- Example:
  - Gene A = 5, Gene B = 10, Gene C = 10 reads
  - Gene A is about half as abundant Gene B
  - Gene A and Gene C have about the same abundance



Bag of Fragments Analogy

- Reads are samples drawn from the distribution of all RNA fragments

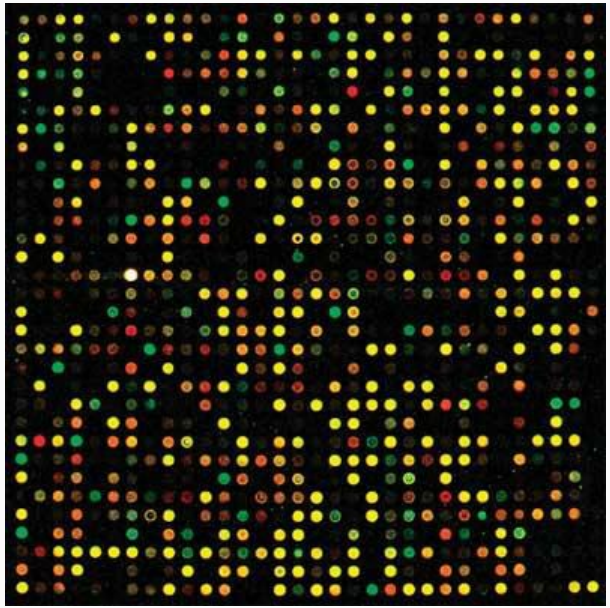


- Drawn in proportion to frequency
- High abundance transcripts drawn frequently
- Low abundance transcripts might not be drawn at all (black read)
- More reads sequenced → more chance to draw low abundance transcripts
- **Absence of evidence is not evidence of absence!**

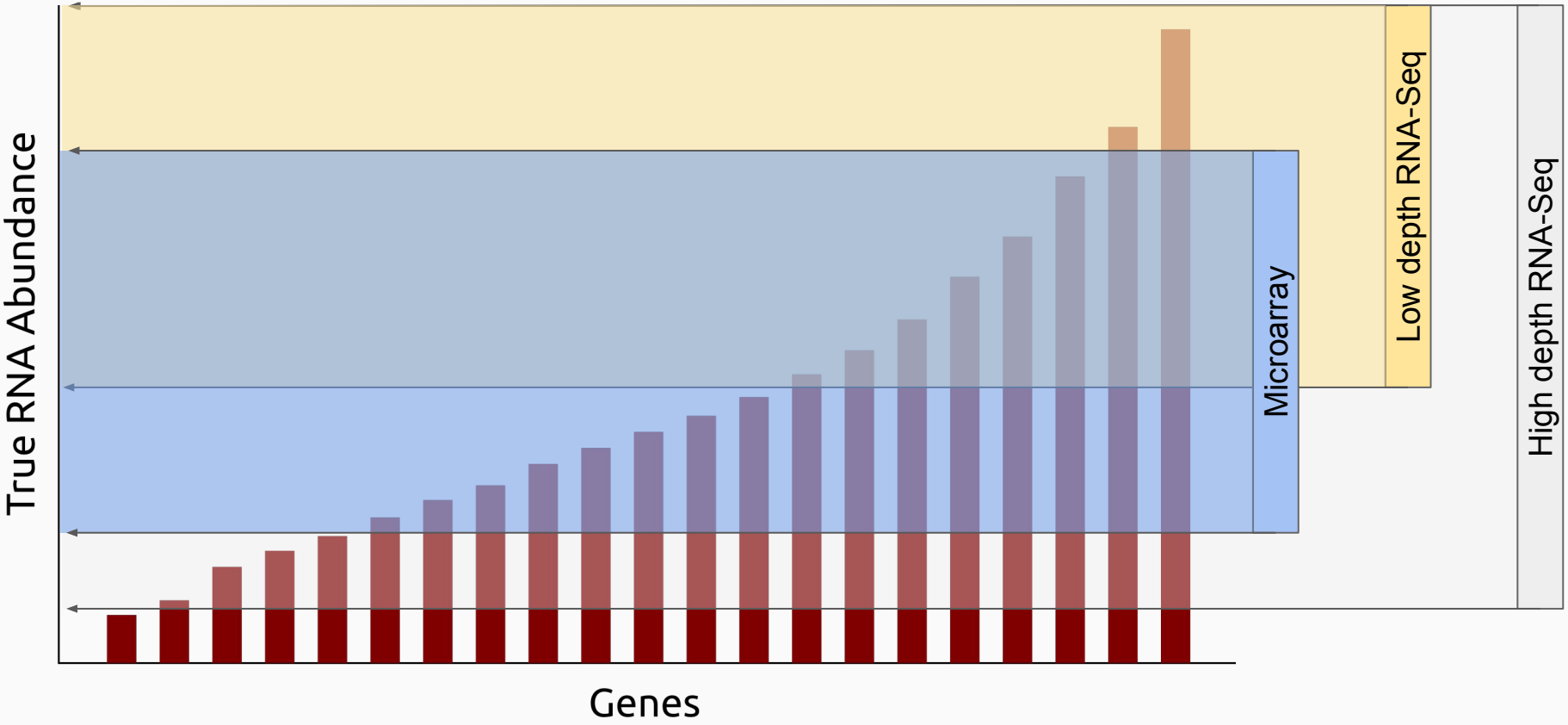
RNA-Seq vs Microarray

RNA-Seq compared to microarray:

- “Unbiased” - de novo sequences
- Larger dynamic range
- Information rich
- More complex data
- Much larger data
- More expensive



Dynamic Range - Detection Limits



Gene Expression Analysis Strategies

- Align+count
  - Explicit alignment against genome
  - Count reads aligned to known loci (e.g. genes)
- Quantify
  - “All-in-one” approach
  - Quasi-alignment (i.e. “good enough” alignment) against transcriptome only
  - Statistical model estimates abundance

Abundance Estimation: Align and Count

1. Align against reference genome
2. Compare alignments with annotation
3. Count # of reads within desired features (e.g. exons, coding sequence)
4. Sum to transcript or gene level
5. Read counts are estimates of abundance

Concept: Multimapping Reads

- Paralogs, repetitive sequence may transcribe identical RNA
- **Multimapping read**: read that maps equally well to multiple loci
- Can cause abundance estimation bias
- Mitigate by:
  - Filtering out multimappers
  - Limit reads to aligning to a maximum # of loci (e.g. 1, 10)
  - Multimap resolution methods (e.g. mmr, ORMAN)

Abundance Estimation (Quantification)

- Quasi- (or pseudo-)alignment: map reads to sequences without explicit alignment
- Must build reference transcriptome
- Statistical model performs abundance inference
- Handles multimapping reads implicitly
- Similar accuracy, faster than align+count

Expression Analysis Strategies Summary

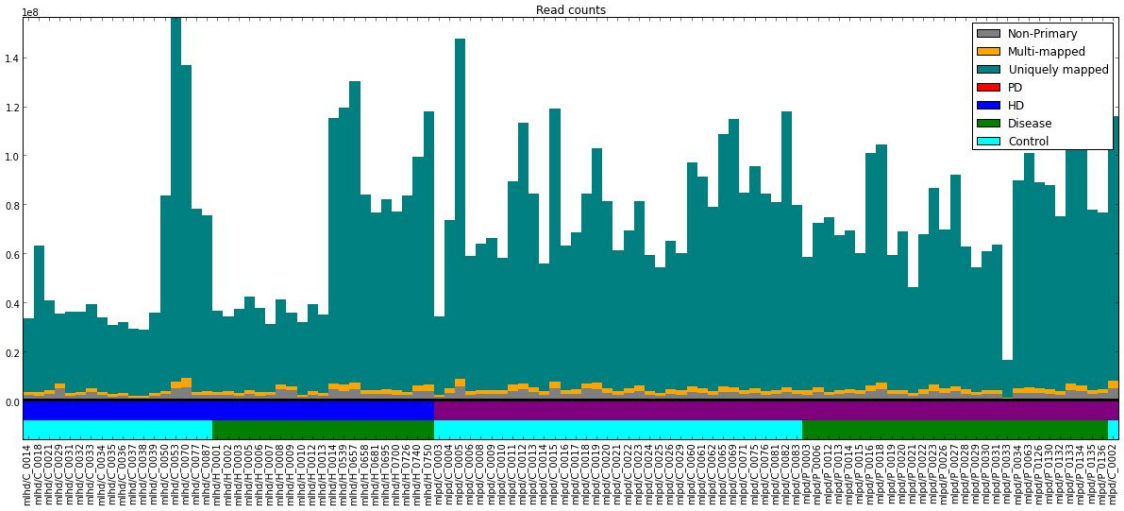
- Align+count:
  - Spliced alignment methods: STAR, TopHat
  - Counting methods: htseq, featurecounts, VERSE
  - Advantages: flexible, accurate, whole genome
  - Disadvantages: slow, many parameters to choose
- Quantify
  - Methods: salmon, kallisto
  - Advantage: very fast, accurate, handles multimaps
  - Disadvantage: transcriptome only

Differential Expression (DE)

- Start with expression matrix of genes x sample counts/estimates
- Which genes have counts associated with variable of interest, e.g. case vs control?
- Each gene will have
  - significance (e.g. p-value)
  - Effect size (e.g. log2 fold change)

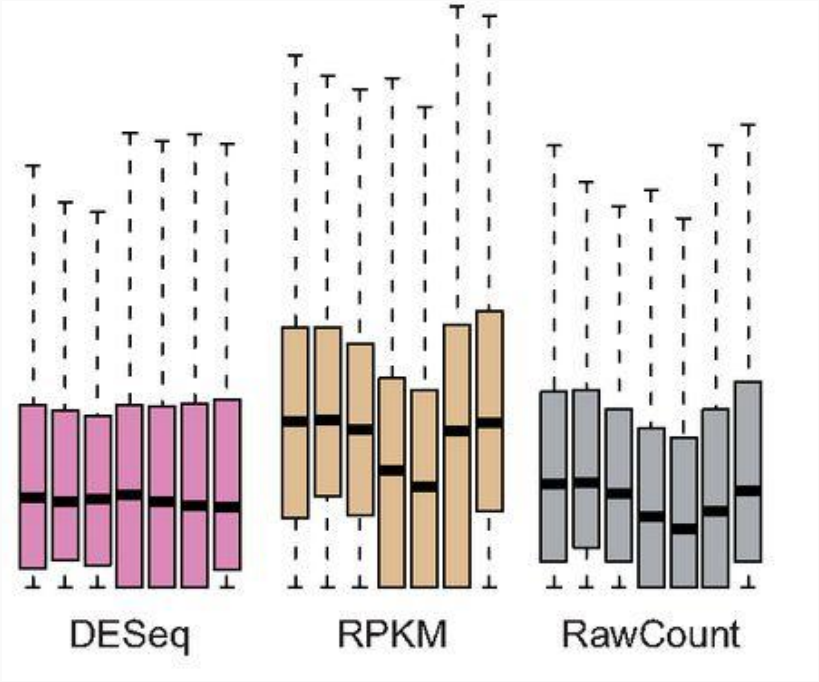
Concept: Count Normalization

- # of reads differ between libraries
- Counts proportional to library size
- Must be *normalized* for samples to be comparable
- Un-normalized counts called *raw counts*



Count Normalization Strategies

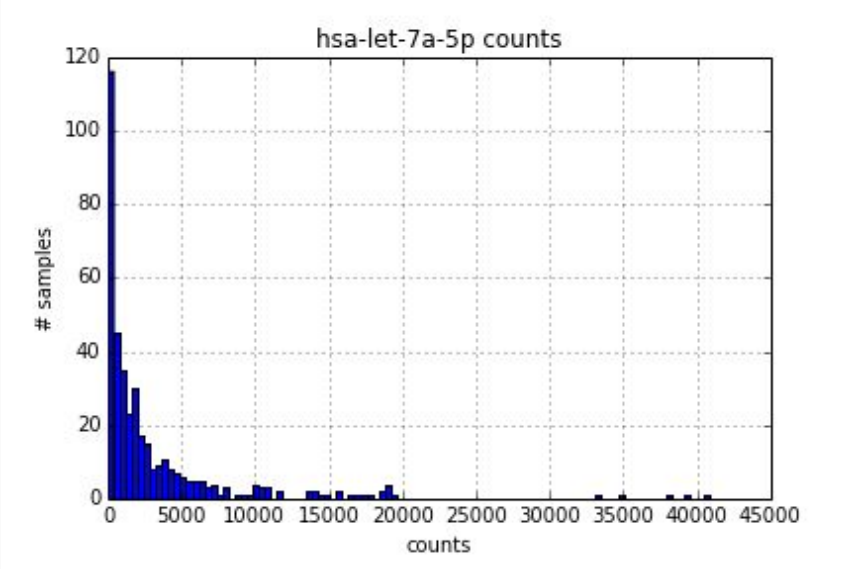
- DESeq2 - median of geometric mean count ratio
- FPKM/RPKM - fragments (reads) per kilobase per million reads
  - Divide each gene count by length of gene*10^6
- Others proposed, these two most common



Dillies, M.-A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* **14,** 671–683 (2013). http://bib.oxfordjournals.org/content/14/6/671.full

Modeling Count Data

- Count data are not normally distributed:
  - Non-negative integers
  - Mean-Variance dependence
  - Long upper tail
- Modeled as Negative Binomial Distributed
- Negative Binomial Regression Utilized for DE



Differential Expression Methods

- Current state of the art:
  - DESeq2 (Negative Binomial Regression)
  - edgeR (Negative Binomial Regression)
  - Count transformation + limma (linear regression)
- Deprecated:
  - Cufflinks (Negative Binomial Regression)
- These methods perform normalization