JOHNS HOPKINS
SCHOOL of MEDICINE

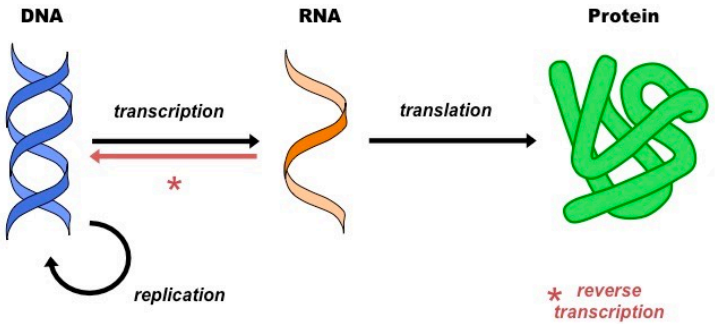
Quantitative Molecular Neurogenomics 2023
Module 6 - RNA-Seq Overview

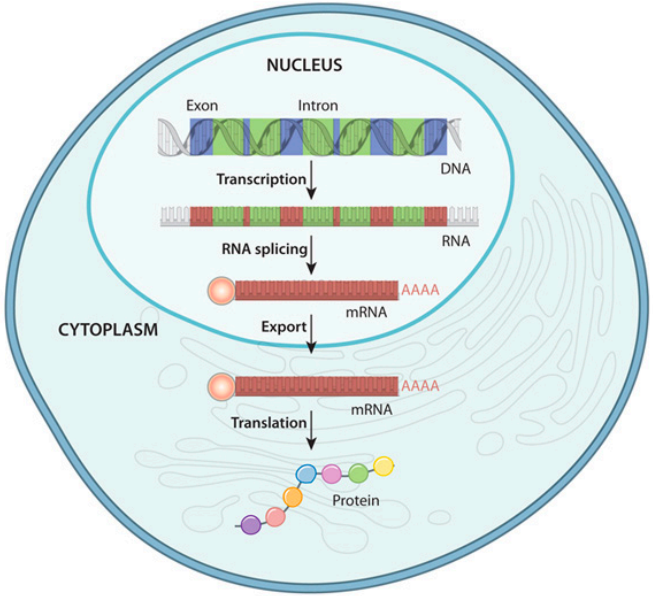
ME.440.825
Loyal Goff - Course Director
Genevieve Stein-O'Brien - Instructor
Richard Sriworarat - TA
Kyla Woyshner - TA
Jeanette Johnson - TA

October 9th, 2023

Gene Expression

- Information about cell type/state can be inferred from knowing what genes are ***expressed*** and when
- Differences in gene expression can highlight important changes and/or differentiating features
- mRNA quantification is a proxy measurement for (most) gene expression





Why RNA?

- RNA->Protein rates are not captured
- Many more tools exist to manipulate/engineer/modify nucleic acids
- While not a ***perfect*** proxy, mRNA levels are more easily and accurately measurable *en masse* than proteins

Measuring Gene Expression

- qRT-PCR
- Microarrays
- RNA-Seq

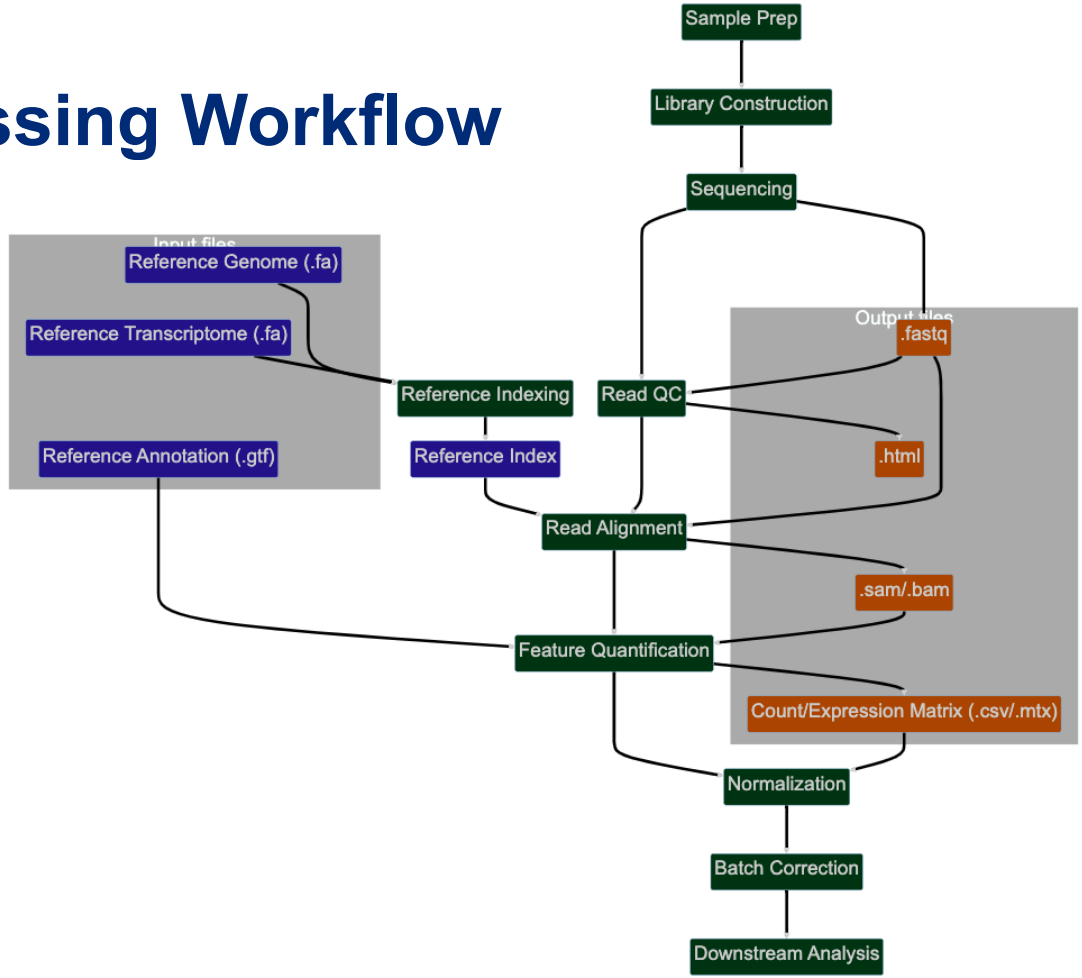
Why RNA-Seq?

- Cost-effective
- Unbiased
  - do not need to know anything about RNA content or sequence *a priori*
- Sensitive

What questions can be addressed with RNA-Seq?

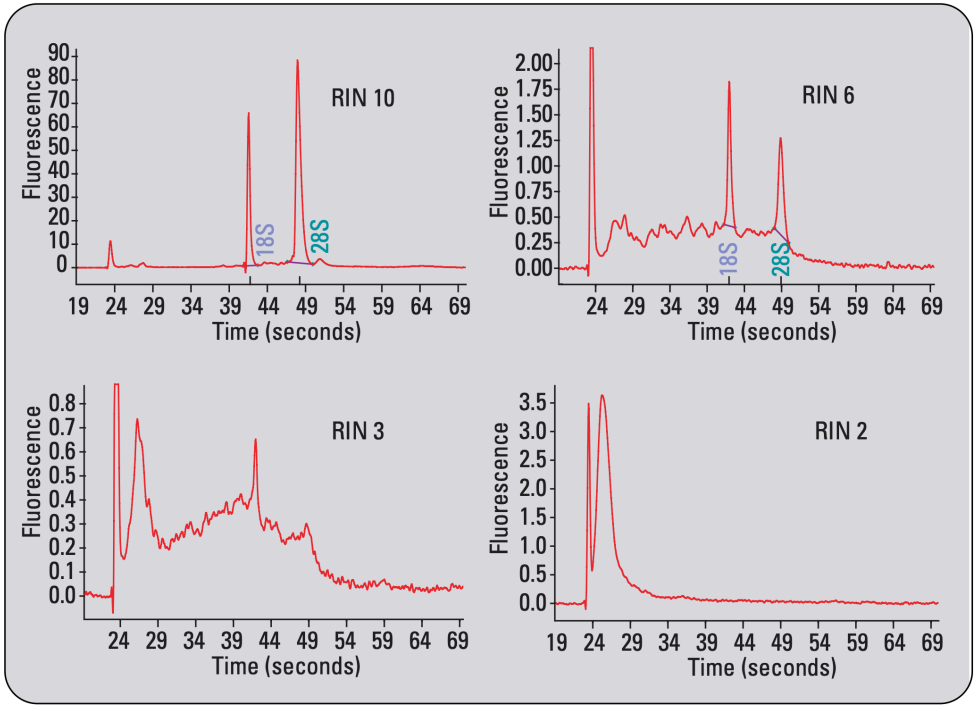
- Transcriptome assembly
- Novel transcript/isoform discovery
- ***Estimating*** gene expression
- Differential analysis
- Allele-specific expression
- Post-transcriptional modifications

RNA-seq Preprocessing Workflow

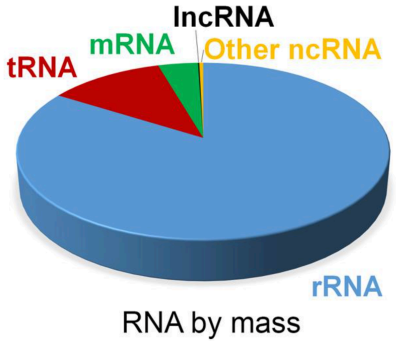


Sample Preparation

- Total RNA Extraction
  - Critical first step to obtain high-quality RNA
  - Purity and quality of extracted RNA are essential
- High-resolution gel or electropherogram essential to assess total RNA integrity
  - Quality measured by proxy of 18S & 28S rRNA integrity and ratio.
- Tissue/context and isolation method can affect RNA quality and yield

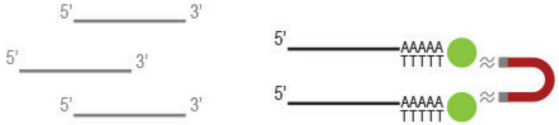


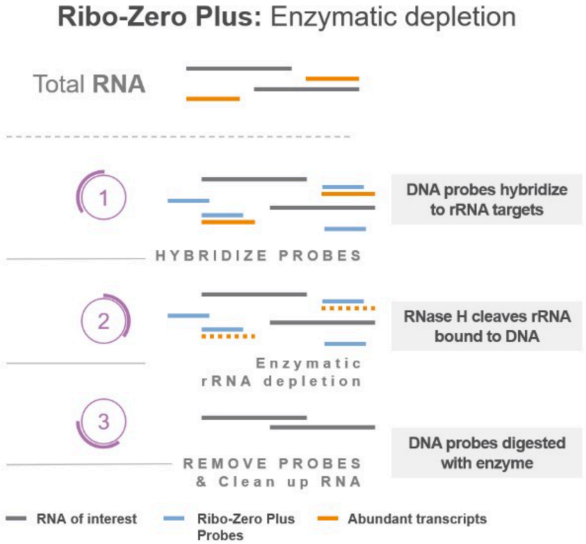
Poly(A) enrichment vs rRNA depletion



- Our target molecule is (usually) mRNAs and/or other PolII transcripts

- Positive selection:
  - polyA-selection
- Negative selection:
  - rRNA depletion

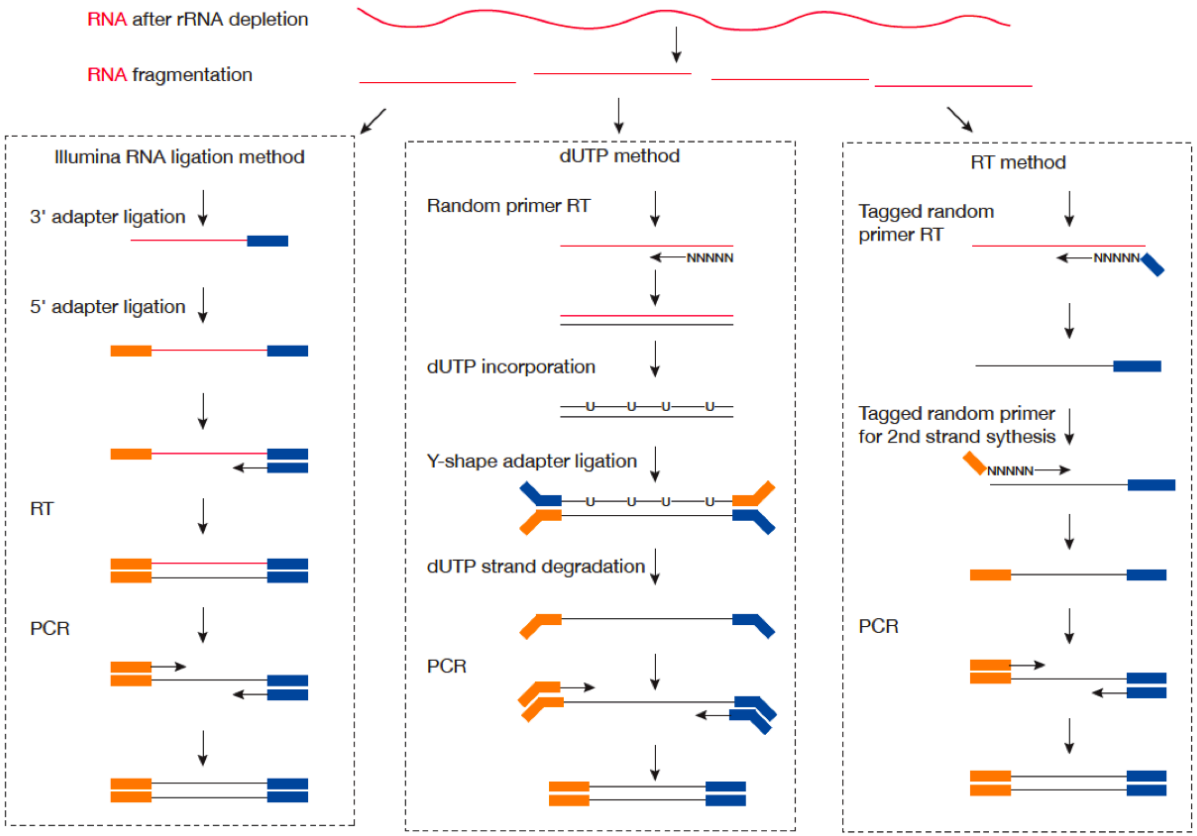




Library Preparation - Total RNA

- Useful for ‘unbiased’ assessment of RNA-species including non-polyA transcripts (e.g. lncRNAs)

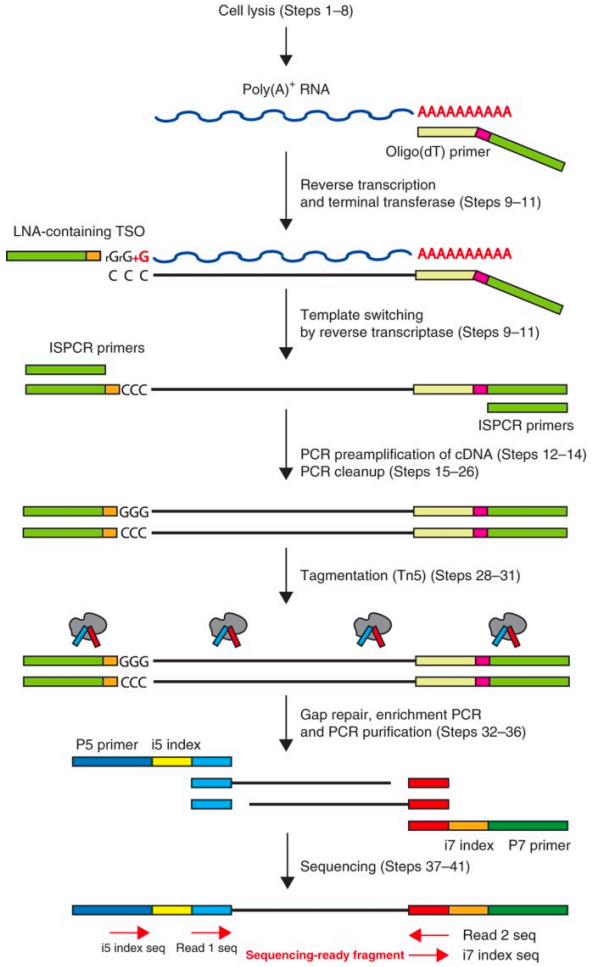
- Pros:
  - Easy and high-efficiency preparations
  - Useful for finding non-PolII transcripts
- Cons:
  - Depletion and/or reduction of rRNA is required beforehand



Library Preparation - mRNA

- Oligo(dT) primer targets poly(A) tail directly for RT
- Reverse transcription converts RNA to complementary DNA (cDNA).
- cDNA is fragmented and adapters are added for sequencing.
- Library amplification via PCR.

- Pros:
  - Does not require rRNA depletion prior
  - Highly enriches for protein coding genes
- Cons:
  - Can prime internally at homopolymer $(A)_n$ sequences
  - Cannot detect RNAs w/o polyA tail

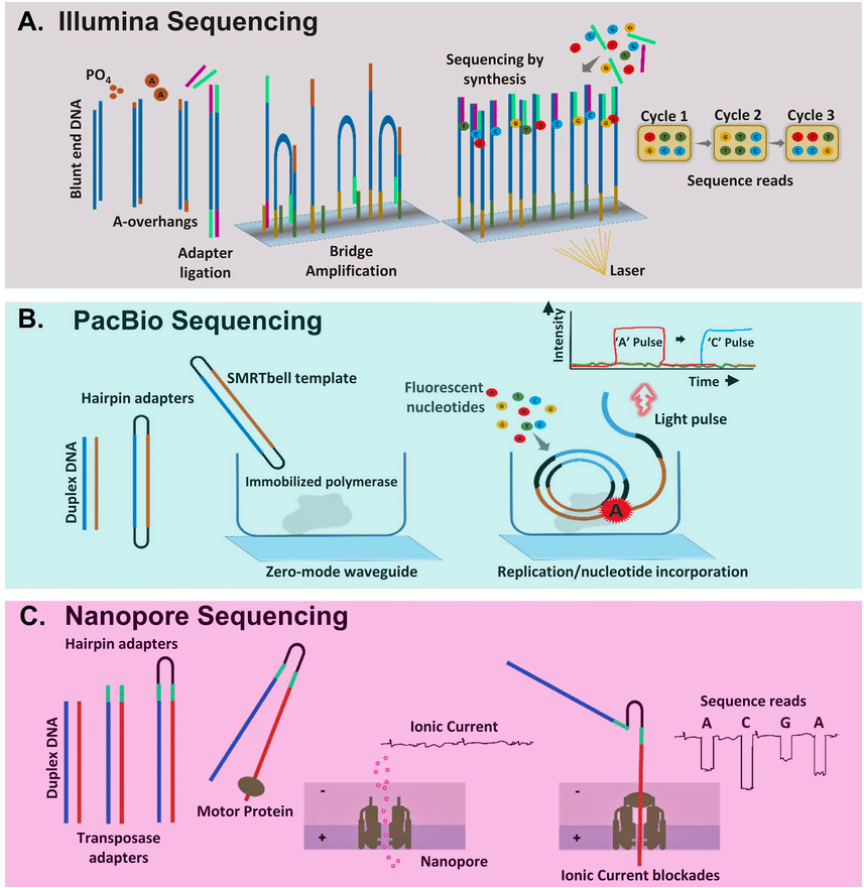


Strand-specific vs Unstranded library prep

- Older prep methods tended to produce ‘unstranded’ libraries
  - Less common these days
- Strand-specific libraries help clarify orientation and quantification
  - Especially for overlapping transcripts (e.g. lncRNAs)

Short read vs Long read sequencing

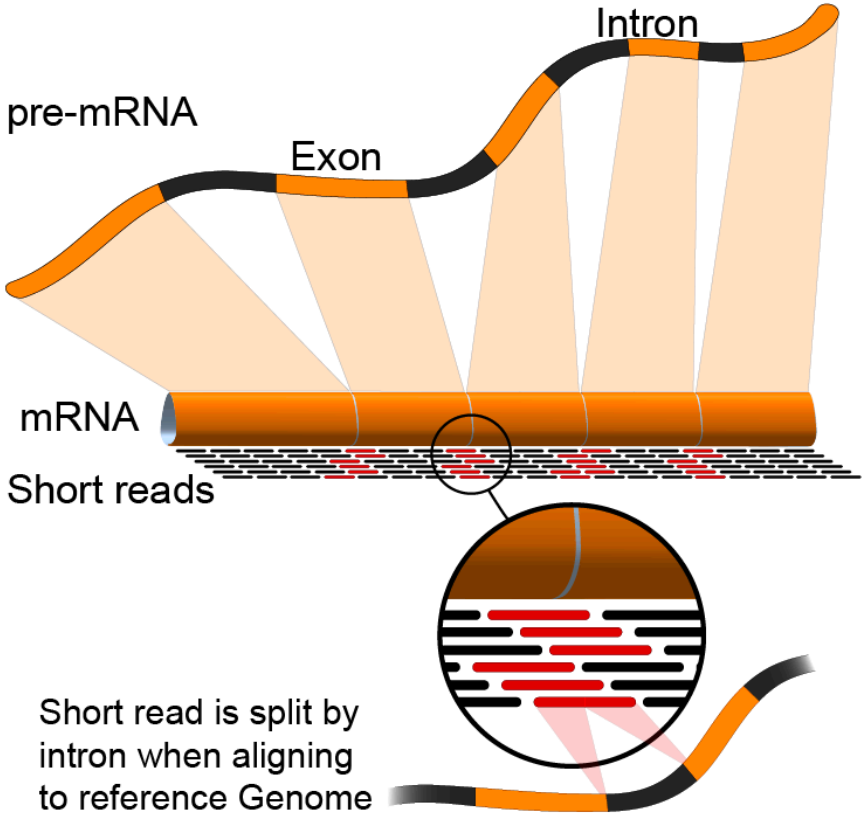
- Often times, library prep is ***very*** similar for both methods.
- Comes down to choice of platform and question
- Illumina short read:
  - Produces between 100-600bp fragments
  - High accuracy (>99%)
  - Billions of fragments per sequencing run
  - Best for variant calling, differential expression, SNP detection.
  - Struggles with repetitive regions, needs a reference for assembly.
- Long read sequencing (e.g. Oxford Nanopore)
  - Several Kb to >2Mb reads, capturing whole transcripts
  - Lower accuracy (~90%)
  - Thousands to millions of reads per run
  - Ideal for de novo assembly, structural variants, splice variants, methylation detection.
  - Can ***directly*** measure RNA sequence



Read QC

Alignment of RNA-seq fragments

- Reads derived from cDNA may contain exon-exon junctions
  - Longer & paired-end reads more useful in this context
- Require use of splice-aware aligner
- Junctions useful for identification & quantification of isoform-level expression



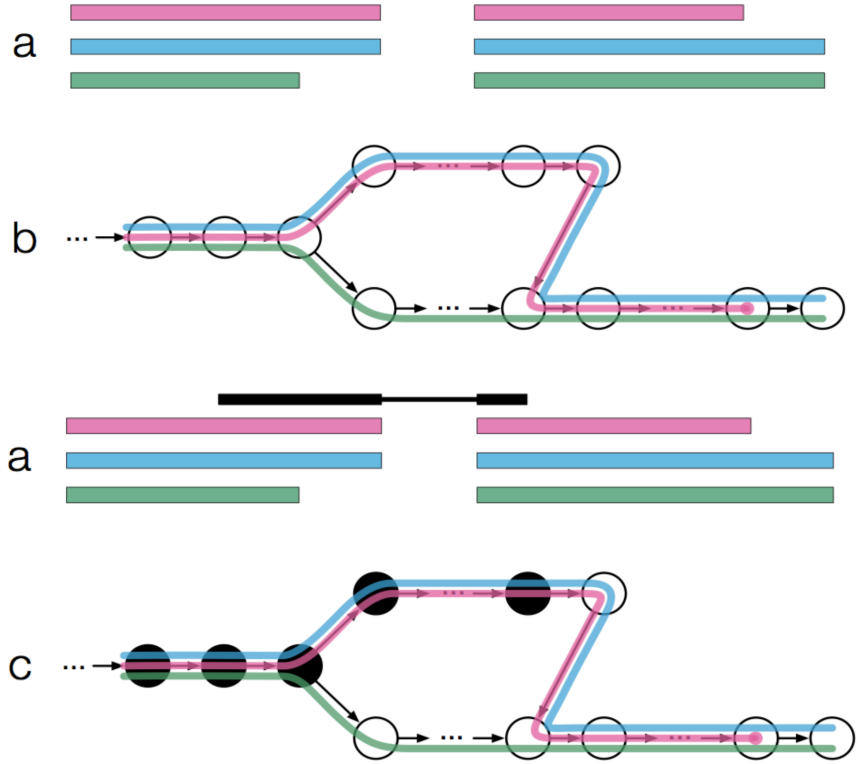
RNA: Alignment vs *Pseudo*alignment

- Alignment
  - Mapping to reads to annotated reference genome
  - Variety of (splice-aware) alignment methods
  - Can be used for discovery of new transcripts
  - Speed depends on method but generally slower and more memory intensive
  - Alignment Tools
    - Tophat2
    - ***HISAT2***
    - STAR

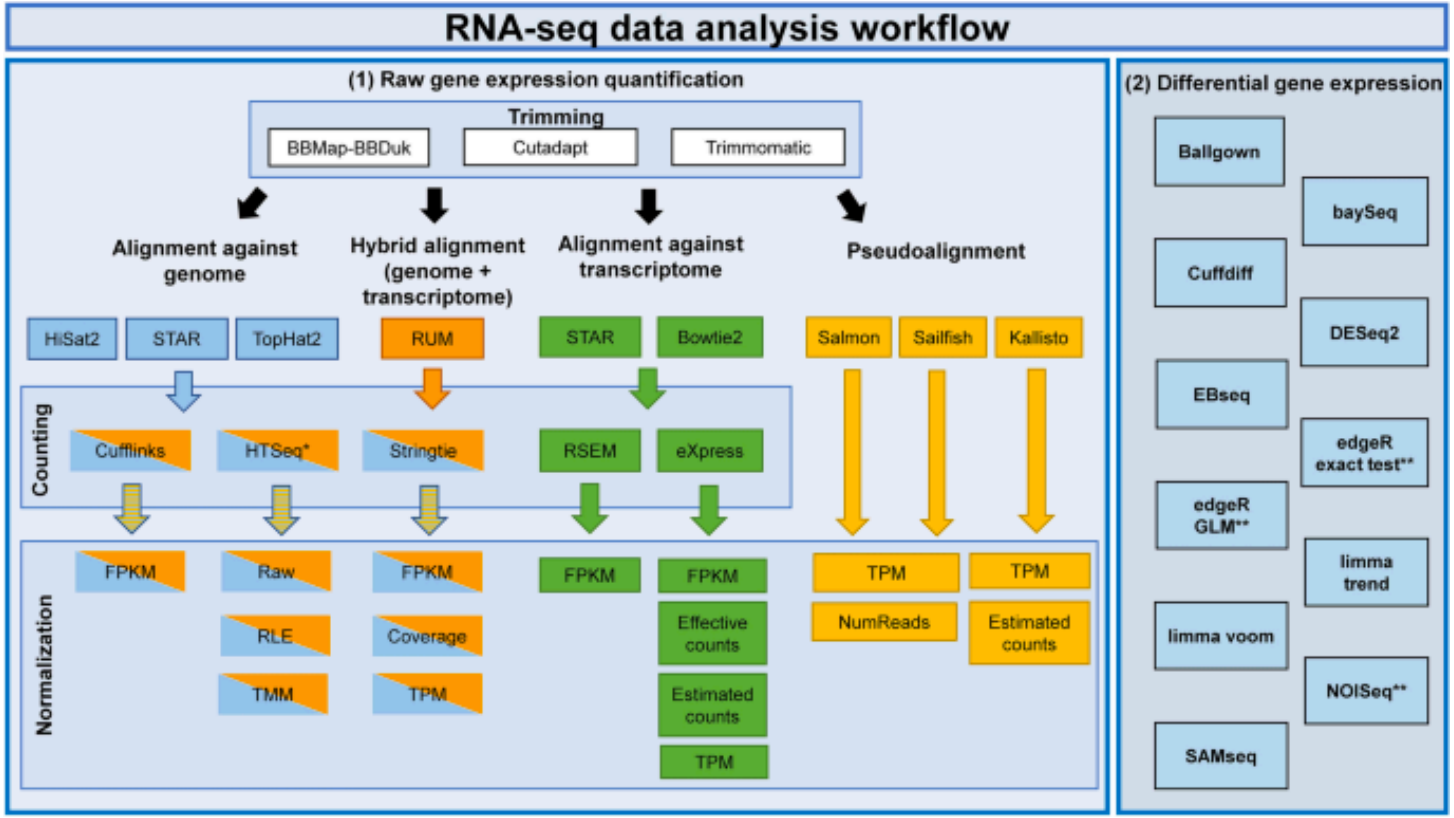
- Pseudoalignment
  - Mapping reads directly to reference transcriptome
  - K-mer-based equivalency class
  - No inference beyond reference transcriptome
  - Very fast
  - Speed allows for bootstrapping estimation of error in quantification
  - Pseudoalignment Tools:
    - ***Kallisto***
    - Salmon

Pseudoalignment of RNA-Seq reads

- Different Goal:
  - *…determine, for each read, not where in each transcript it aligns, but rather which transcripts it is compatible with…*
  - Useful for ‘counting-based’ problems
- Hashed index of Kmers that are ‘compatible’ with each reference transcript isoform
- Intersection of all k-compatibility classes for a read identifies the ‘compatible isoforms’ for a given read.



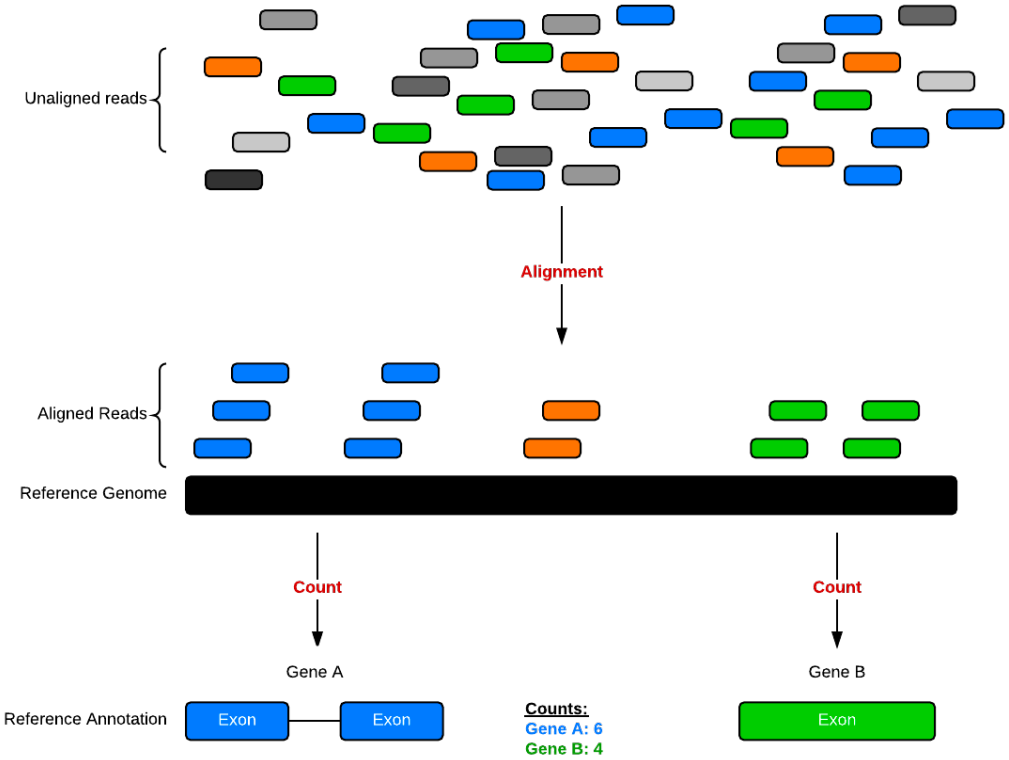
RNA: Alignment vs *Pseudo*alignment



https://www.nature.com/articles/s41598-020-76881-x

Feature Quantification

- ***Estimation*** of the abundance of transcripts or genes in RNA-Seq data.
- ***Objective:*** Translate read alignments into meaningful measures of gene or transcript expression.
- Provides the foundational data for differential expression analysis, co-expression networks, and other downstream analyses.



Counting Methods

- Transcript-based
  - Counts reads uniquely aligned to individual transcripts (isoforms)
    - Kallisto or Salmon
  - Can be aggregated to gene-level post hoc

- Gene-based
  - Aggregates counts for all transcripts of a gene
    - featureCounts or HTSeq
  - Cannot be disambiguated to isoform-level

- The choice between transcript-based and gene-based quantification often depends on the research question.
- Transcript-based is more detailed but can be noisier, while gene-based provides a ‘cleaner’ summarized view.

Challenges in Feature Quantification

- Counts ≠ Expression
  - Gene length normalization required
  - Sequencing depth normalization required

- Ambiguous Mapping
  - Reads mapping to multiple regions/genes
  - Reads mapping to multiple isoforms
  - Solutions:
    - Drop multi-mapping reads from quantification
    - Distribute value across mapped features
- Biases
  - GC bias
  - PCR duplicates
  - Overall sequencing depth

Normalization Strategies / Units

- RPKM/FPKM
  - **R**eads/**F**ragments per **K**ilobase of mRNA per **M**illion reads mapped
  - Calculation:

$$\frac{nReads\ mapped/length\ in\ Kb}{Total\ reads\ in\ millions}$$

  - Values are ***not*** comparable across experiments

- TPM
  - **T**ranscripts **P**er **M**illion
  - Calculation:
    - First normalize by RPKM/FPKM
    - Divide by the sum of all normalized read counts and multiply by 1 million.
  - Values are now comparable across experiments

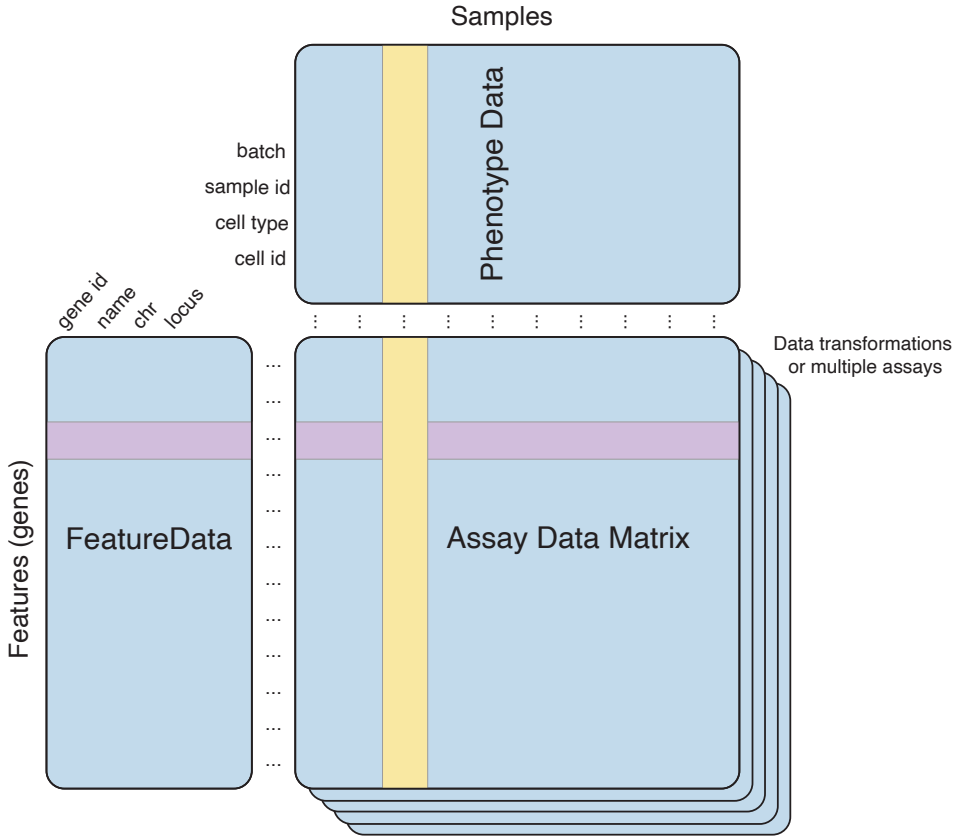
Raw count normalization as part of DE

- Several differential expression tools instead directly model the ***raw read counts*** using a negative binomial distribution.
  - Input should ***not*** be pre-normalized as it is handled internally
  - e.g. DESeq2, edgeR

- These tools use scaling factors derived from the data to adjust for library size differences.
  - For instance, DESeq2 uses the median-of-ratios method.

Aggregation of sample counts to matrix

- Each RNA-Seq sample produces count data for genes/transcripts.
- Aggregating these counts creates a comprehensive expression matrix.
  - Concatenating vectors together provides the foundational matrix for all downstream analyses
- Ensure consistent naming/annotation across samples.
  - Merging/reordering genes is a common error during this step
  - Be sure to validate/check often
- Merge sample metadata and retrieve feature annotation at this stage as well.



Tools and Resources

Neuroscience

Hipposeq: a comprehensive RNA-seq database of gene expression in hippocampal principal neurons

Mark S Cembrowski, Lihua Wang, Ken Sugino, Brenda C Shields, Nelson Spruston

Janelia Research Campus, Howard Hughes Medical Institute, United States

Apr 26, 2016 · https://doi.org/10.7554/eLife.14997

https://elifesciences.org/articles/14997