

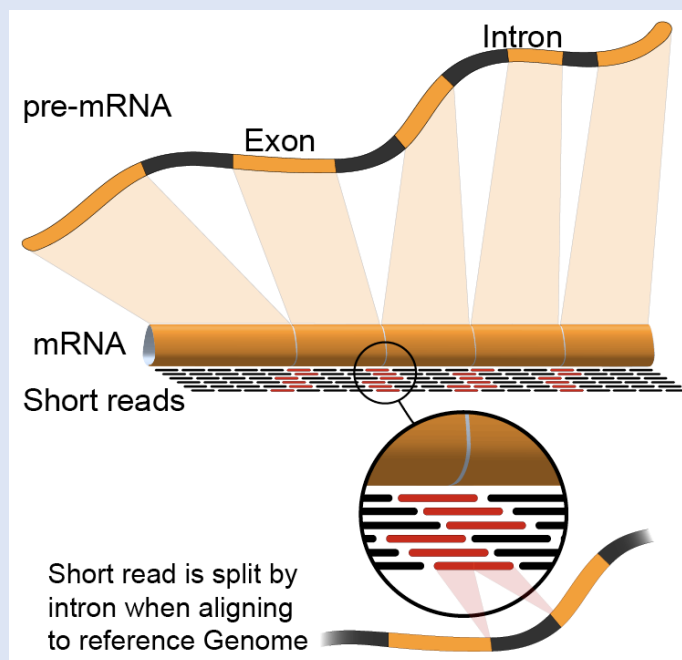
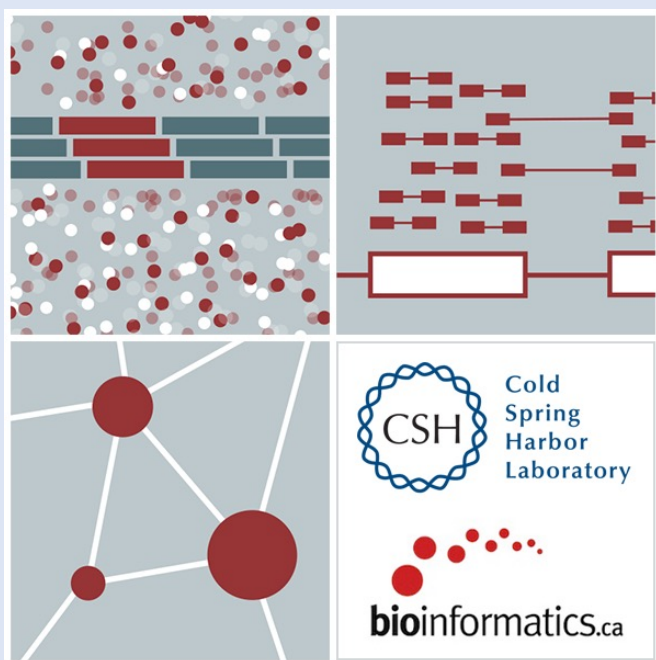


Cold
Spring
Harbor
Laboratory

Introduction to RNA sequencing (lecture)

Arpad Danos, Felicia Gomez, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal

Advanced Sequencing Technologies & Bioinformatics Analysis November 5-19, 2023



 Washington University in St. Louis
SCHOOL OF MEDICINE

Learning objectives of the course

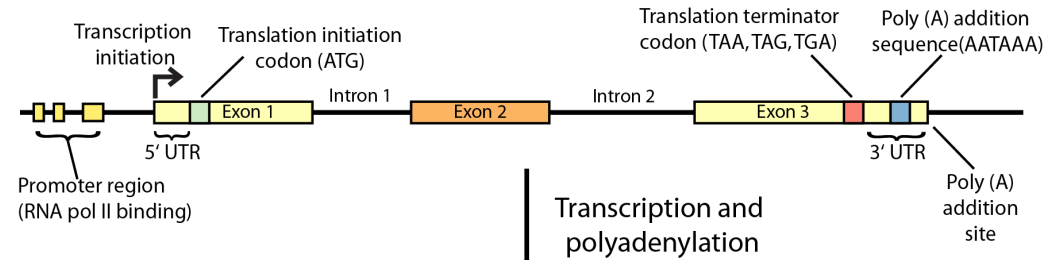
- **Module 1: Introduction to RNA Sequencing**
- Module 2: Alignment and Visualization
- Module 3: Expression and Differential Expression
- Module 4: Alignment Free Expression Estimation
- Tutorials
 - Provide a working example of an RNA-seq analysis pipeline
 - Run in a ‘reasonable’ amount of time with modest computer resources
 - Self contained, self explanatory, portable

Learning objectives of module 1

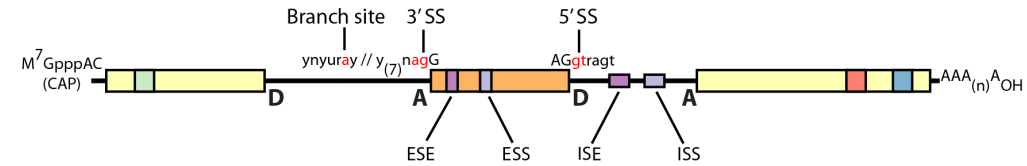
- Introduction to the theory and practice of RNA sequencing (RNA-seq) analysis
 - Background molecular biology
 - Challenges specific to RNA-seq
 - General goals and themes of RNA-seq analysis work flows
 - Common technical questions related to RNA-seq analysis
 - Introduction to the RNA-seq hands on tutorial

Gene expression

Double-stranded genomic DNA template

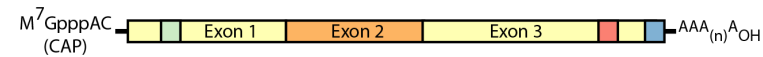


Single-stranded pre-mRNA (nuclear RNA)



RNA processing

Mature mRNA

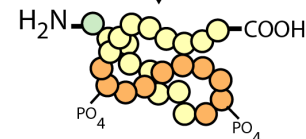


Export to cytoplasm and translation

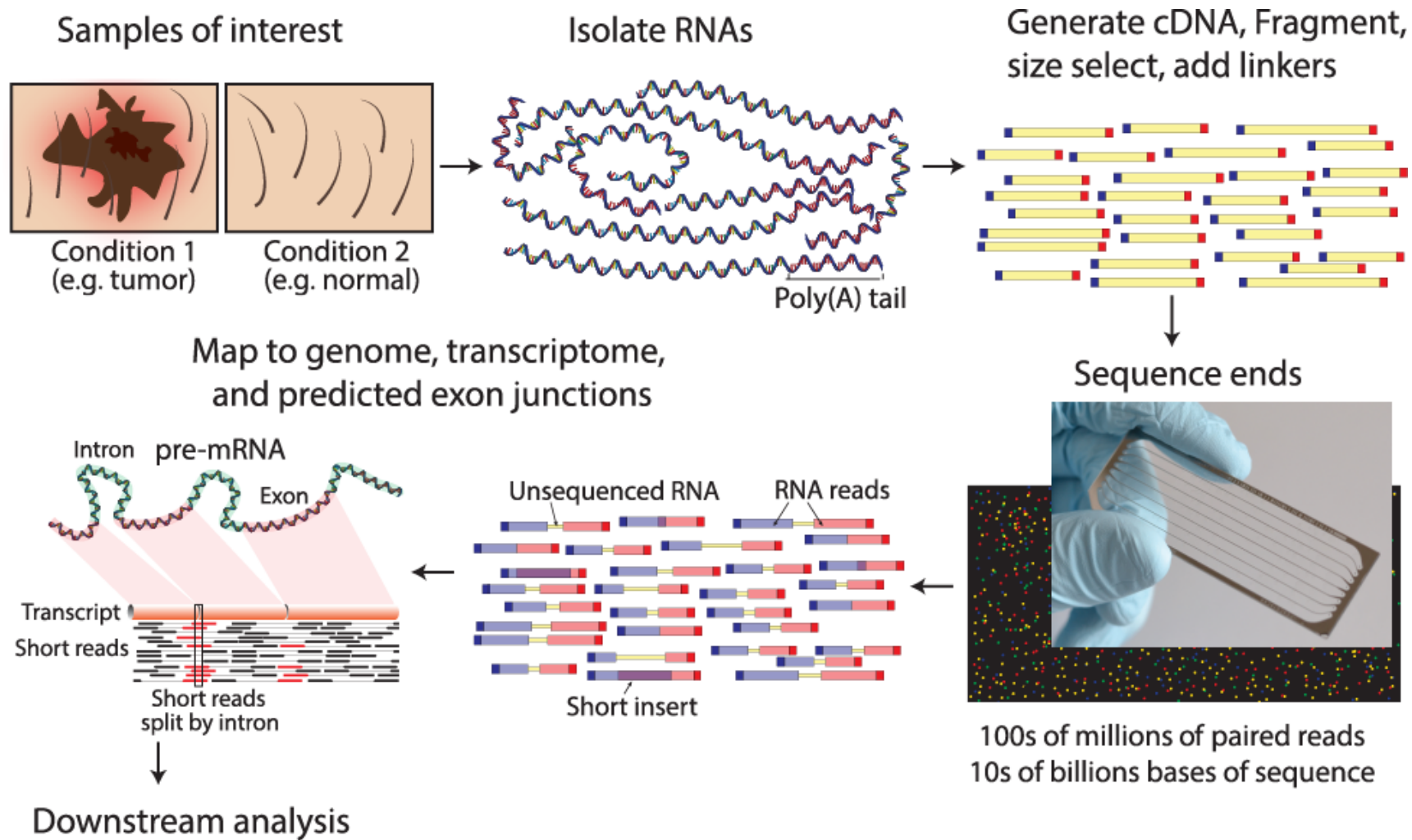
Protein (amino acid sequence)



Folding, posttranslational modification, subcellular localization, etc.



RNA sequencing

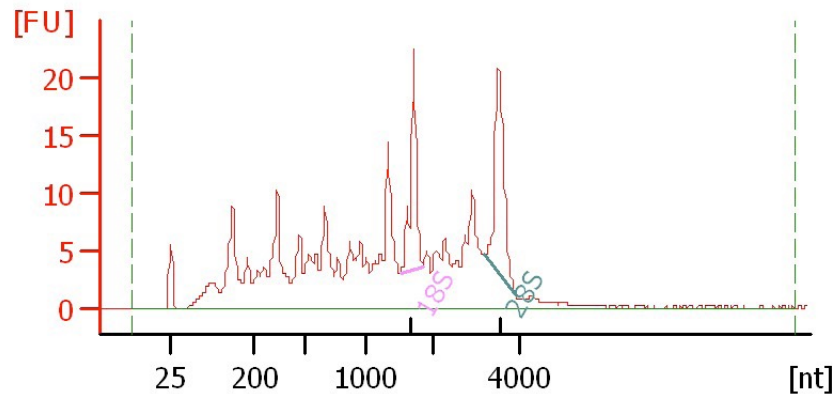


Challenges

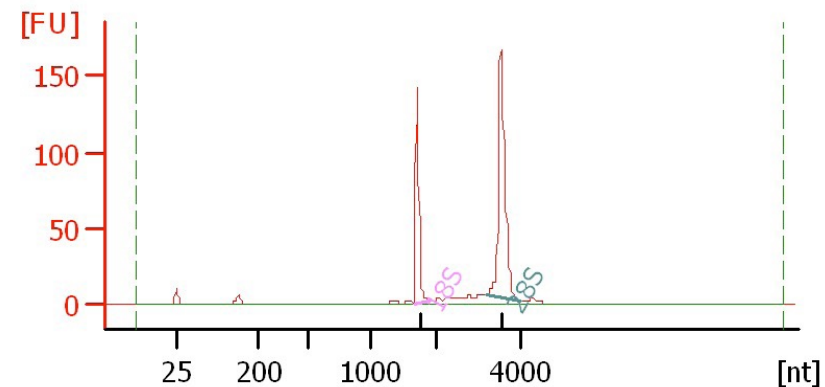
- Sample
 - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
 - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
 - $10^5 - 10^7$ orders of magnitude
 - Since RNA sequencing works by random sampling, a small fraction of highly expressed genes may consume the majority of reads
 - Ribosomal and mitochondrial genes
- RNAs come in a wide range of sizes
 - Small RNAs must be captured separately
 - PolyA selection of large RNAs may result in 3' end bias
- RNA is fragile compared to DNA (easily degraded)

Agilent example / interpretation

- <https://goo.gl/uC5a3C>
- ‘RIN’ = RNA integrity number
 - 0 (bad) to 10 (good)



RIN = 6.0



RIN = 10

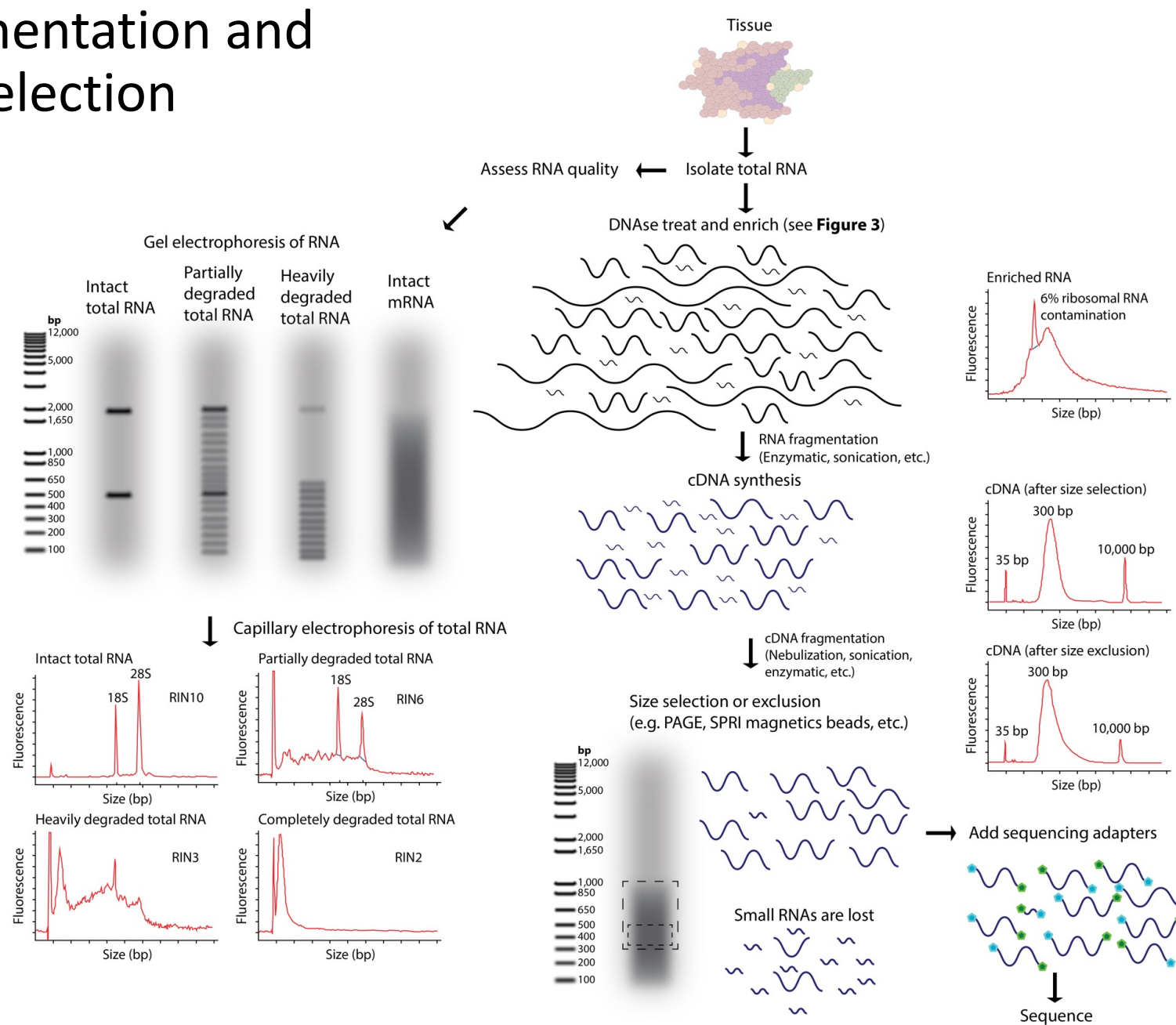
Design considerations

- Standards, Guidelines and Best Practices for RNA-seq
 - The ENCODE Consortium
 - Download from the Course Wiki
 - Meta data to supply, replicates, sequencing depth, control experiments, reporting standards, etc.
- <https://goo.gl/6LePBW>
- Several additional initiatives are underway to develop standards and best practices that cover many of these concepts. These include: the Sequencing Quality Control (SEQC) consortium, the Roadmap Epigenomics Mapping Consortium (REMC), and the Beta Cell Biology Consortium (BCBC).

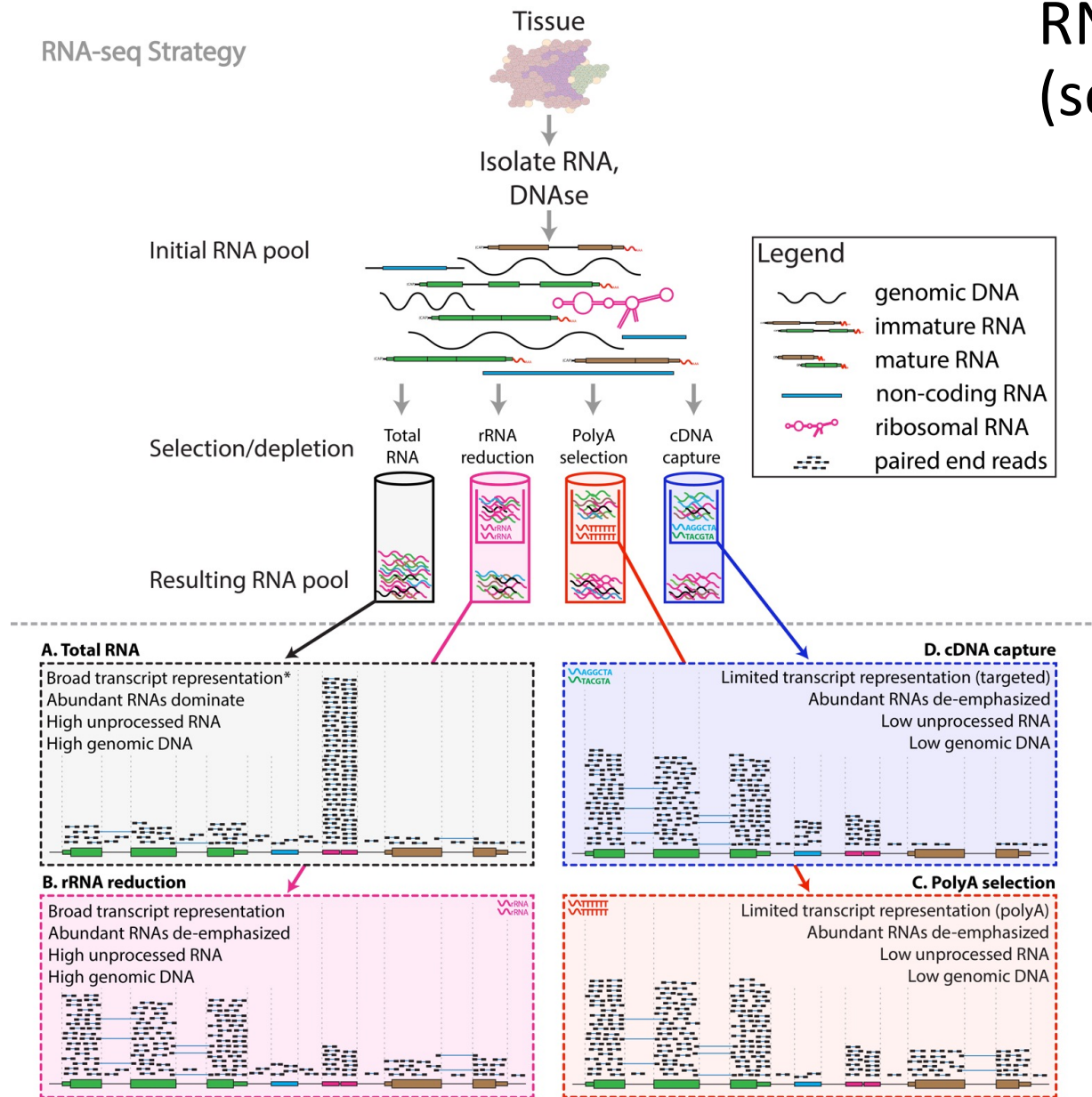
There are many RNA-seq library construction strategies

- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
 - Small RNAs (microRNAs) vs. large RNAs?
 - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library normalization?
- These details can affect analysis strategy
 - Especially comparisons between libraries

Fragmentation and size selection



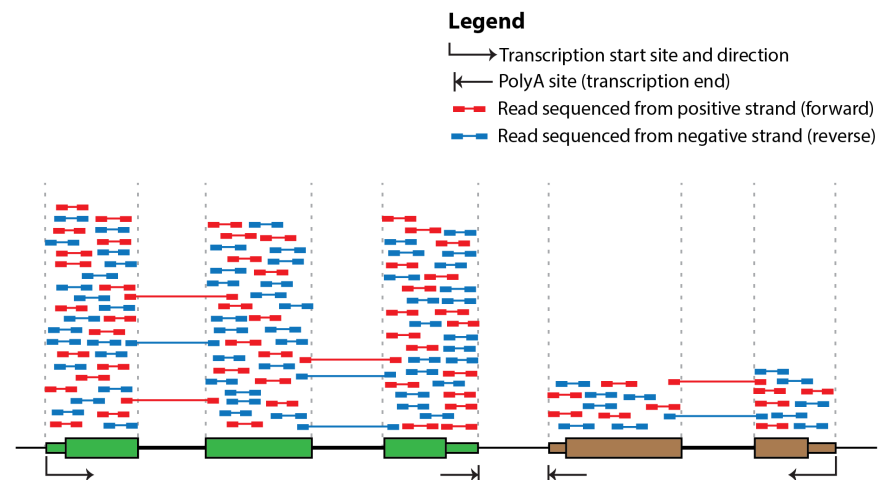
RNA sequence enrichment (selection/depletion)



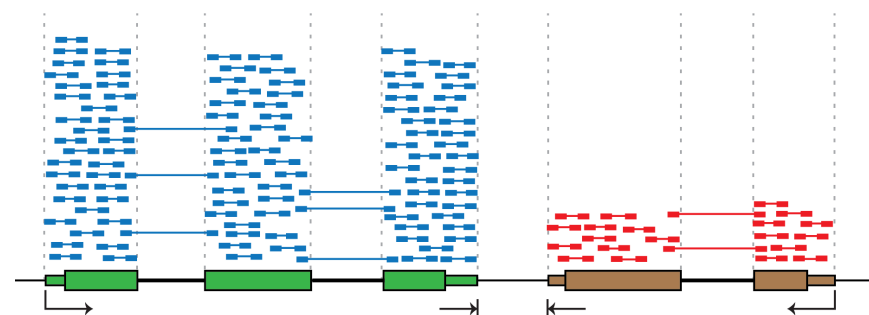
Expected Alignments

Stranded vs. unstranded

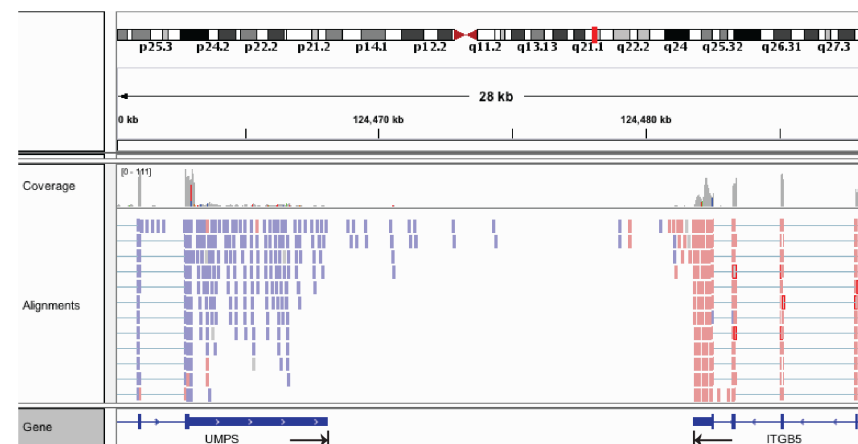
A. Depiction of cDNA fragments from an unstranded library



B. Depiction of cDNA fragments from a stranded library

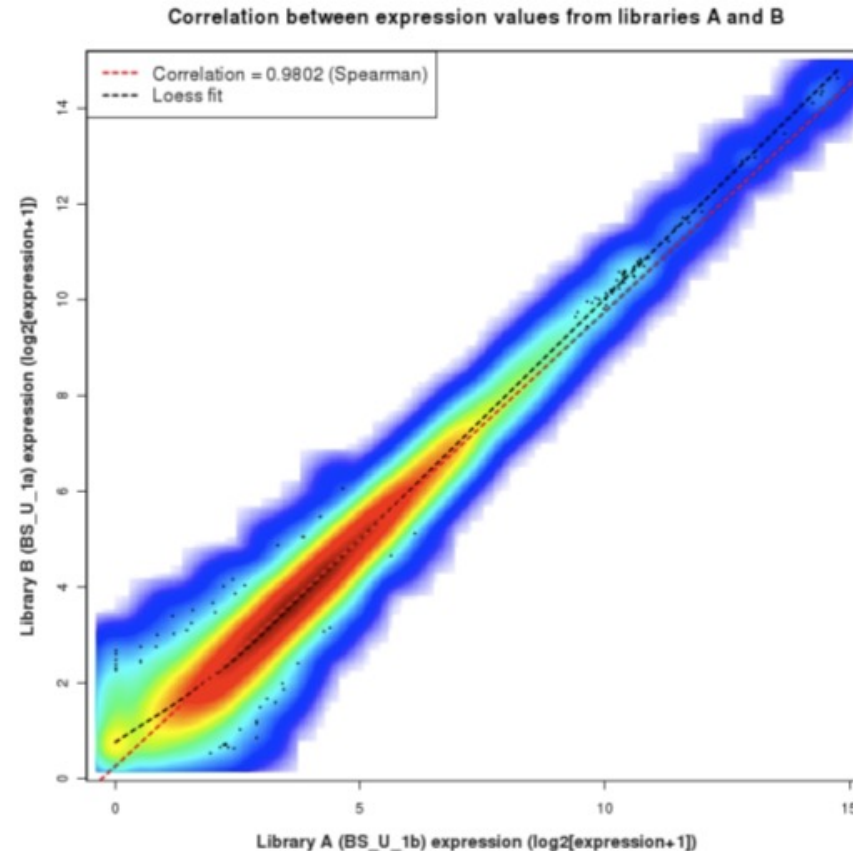


C. Viewing strand of aligned reads in IGV



Replicates

- Technical Replicate
 - Multiple instances of sequence generation
 - Flow Cells, Lanes, Indexes
- Biological Replicate
 - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
 - Some example concerns/challenges:
 - Environmental Factors, Growth Conditions, Time
 - Correlation Coefficient 0.92-0.98



Common analysis goals of RNA-Seq analysis (what can you ask of the data?)

- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
 - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

General themes of RNA-seq workflows

- Each type of RNA-seq analysis has distinct requirements and challenges but also a common theme:

1. Obtain raw data (convert format)

2. Align/assemble reads

3. Process alignment with a tool specific to the goal

- e.g. ‘cufflinks’ for expression analysis, ‘defuse’ for fusion detection, etc.

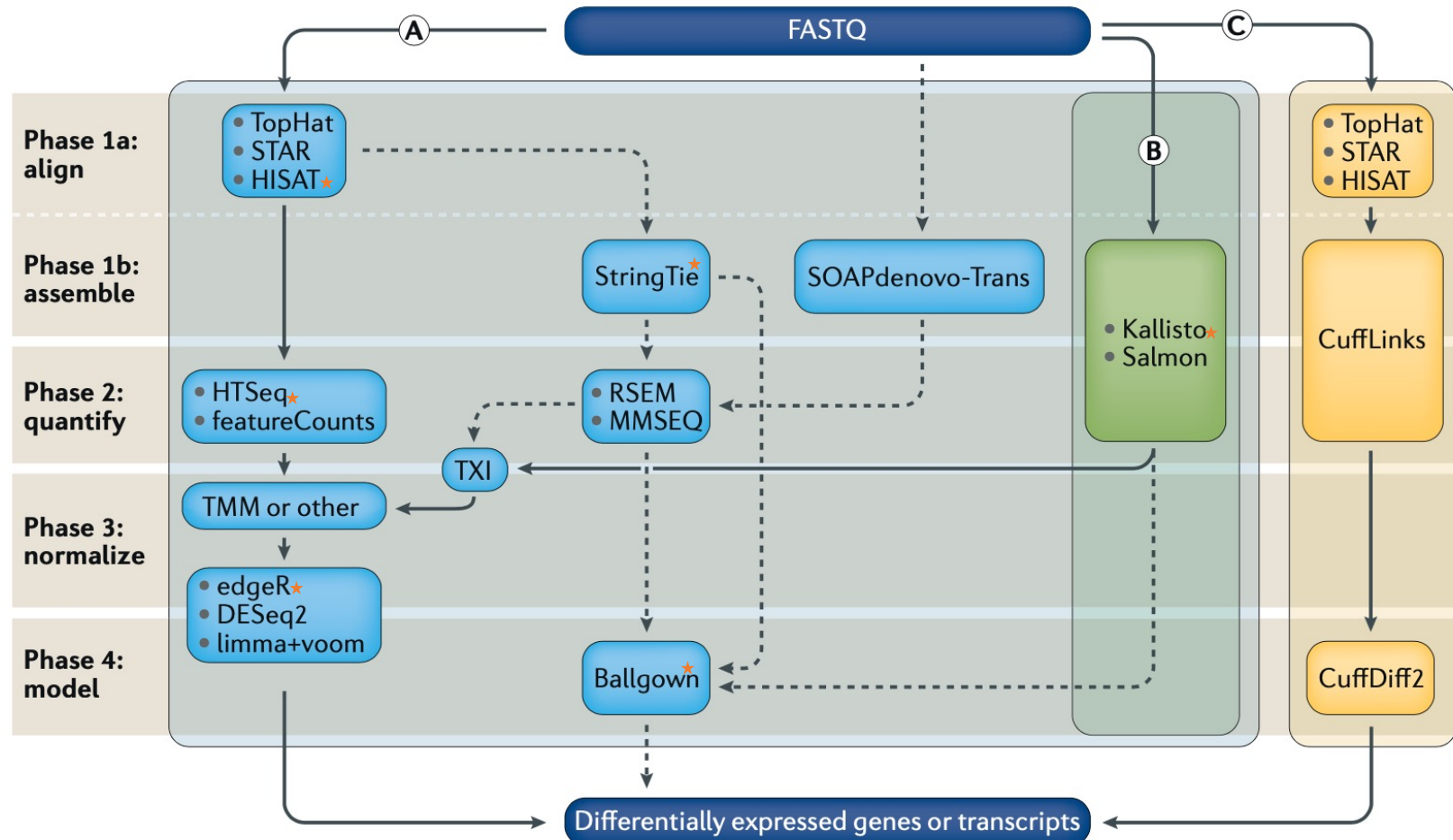
4. Post process

- Import into downstream software (R, Matlab, Cytoscape, Ingenuity, etc.)

5. Summarize and visualize

- Create gene lists, prioritize candidates for validation, etc.

Examples of RNA-seq data analysis workflows for differential gene expression



[Stark et al. 2019](#)

Discussion of bulk vs single cell RNA-seq

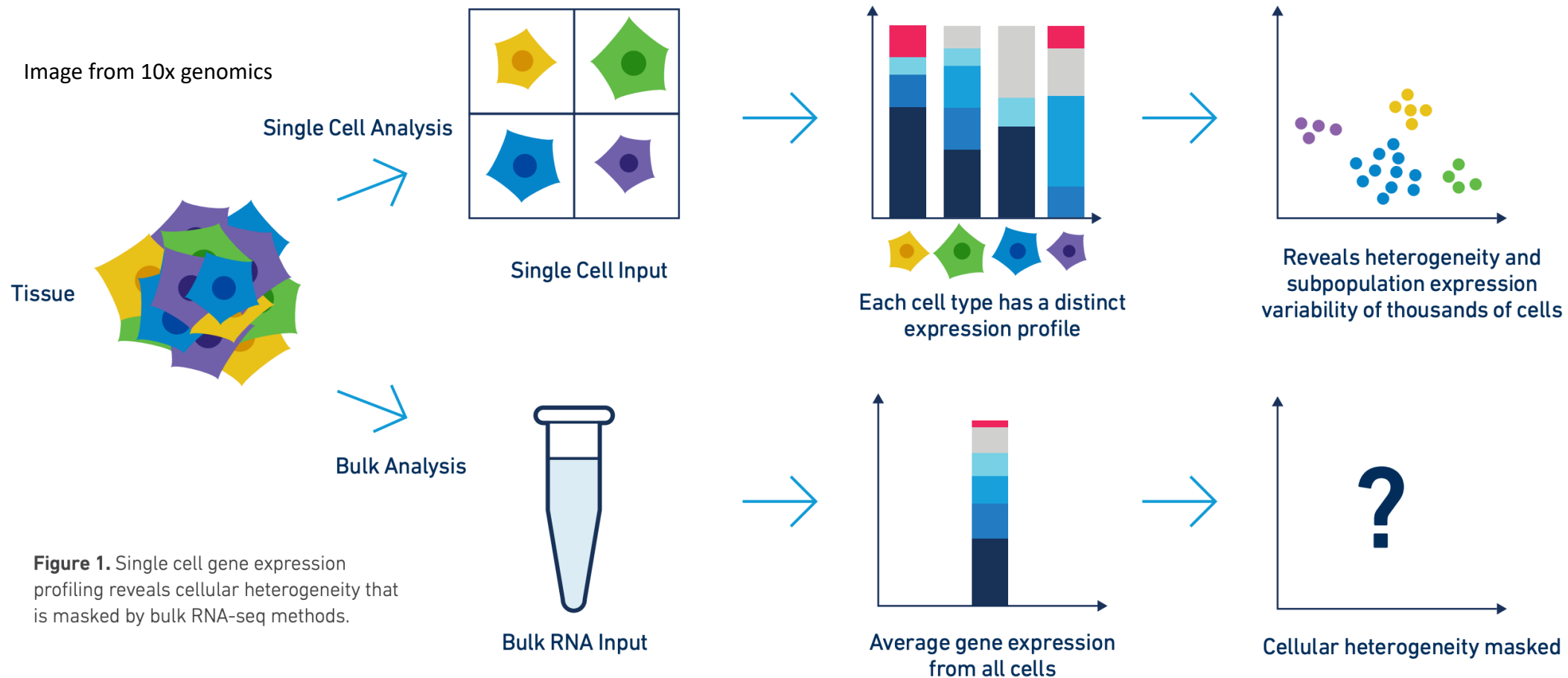


Figure 1. Single cell gene expression profiling reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

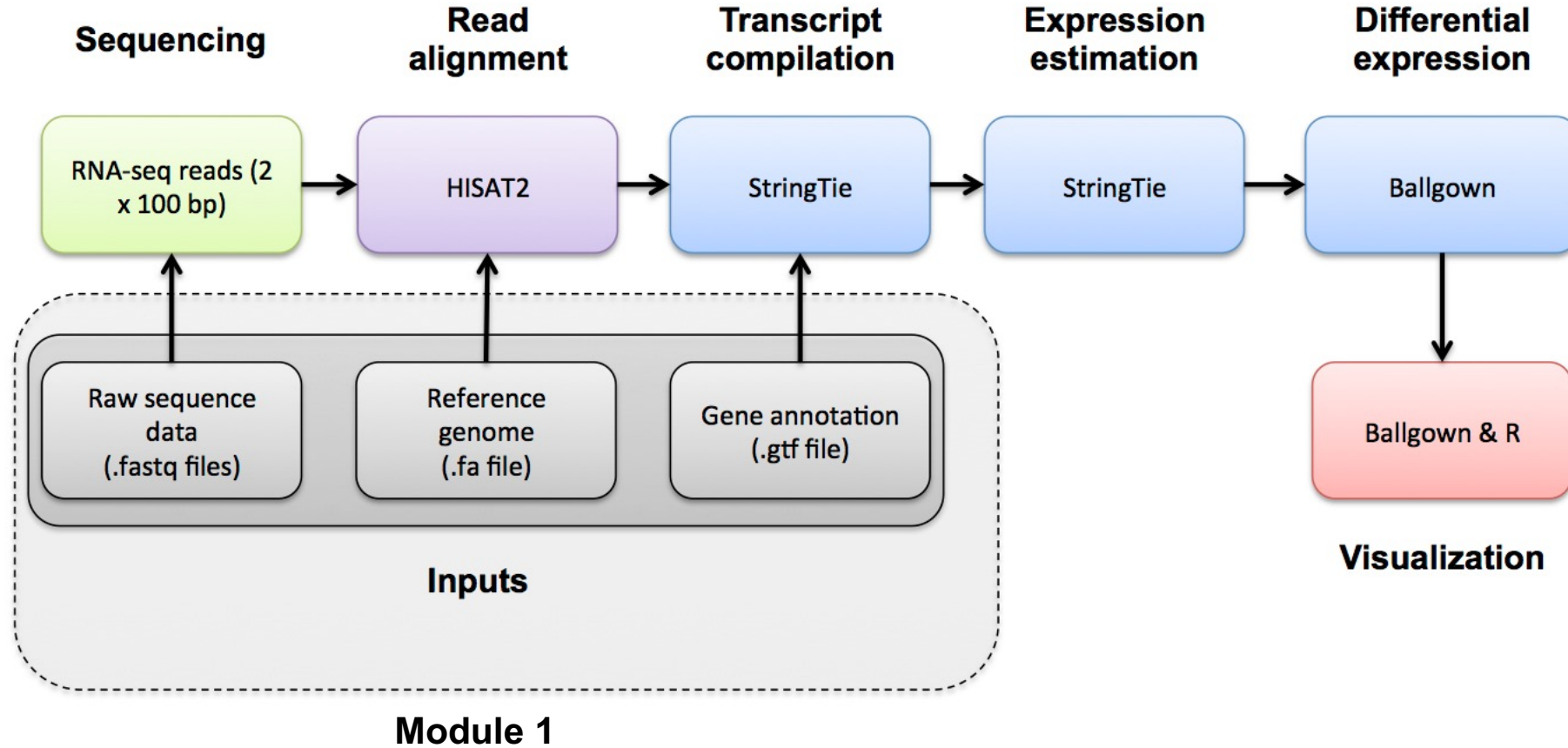
Factors to compare: Cost, complexity of library prep, complexity of analysis, qualitative and quantitative differences in richness of information obtained.

Common questions (and answers)

- [Supplementary Table 7](#)
- Malachi Griffith*, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, Obi L. Griffith*. 2015. Informatics for RNA-seq: A web resource for analysis on the cloud. 11(8):e1004393. 2015.
 - <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393>

Introduction to tutorial (Module 1)

HISAT2/StringTie/Ballgown RNA-seq Pipeline



We are on a Coffee Break & Networking
Session