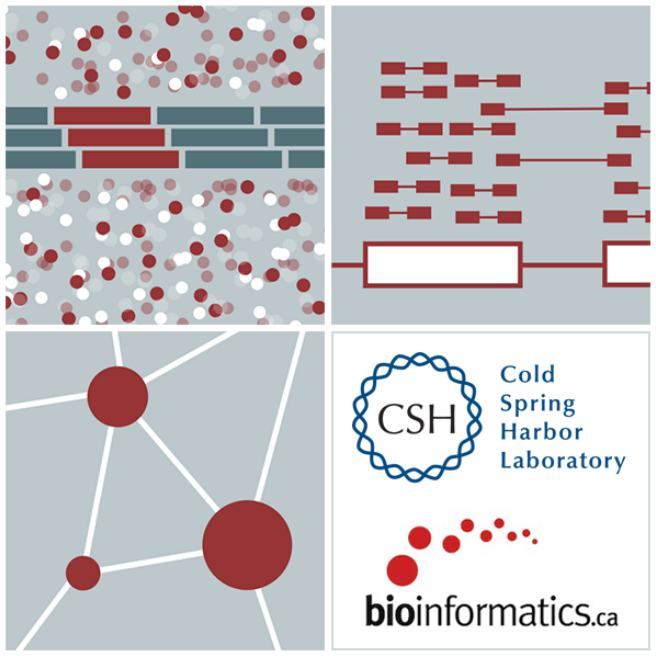
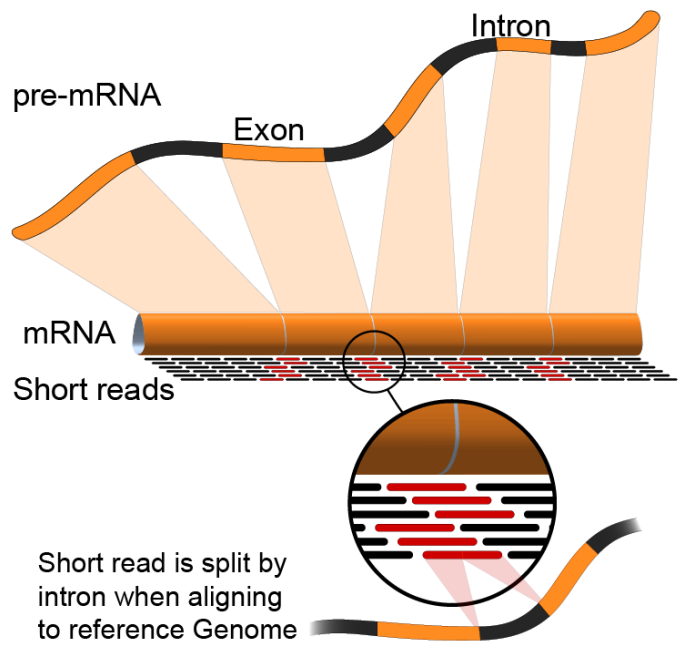
Cold Spring Harbor Laboratory

Introduction to RNA sequencing (lecture)

Felicia Gomez, Charlz Jerold, Obi Griffith, Malachi Griffith,
My Hoang, Mariam Khanfar, Chris Miller, Kartik Singhal, Jennie Yao

Advanced Sequencing Technologies & Bioinformatics Analysis November 10-21, 2025





Washington University in St. Louis
School of Medicine

Introduction to course – philosophy and goals

Do "the bioinformatics" for someone, and you help them for a day.
Teach someone to do bioinformatics, and you help them for a lifetime.

- Ancient Chinese proverb

- Course goals
  - Learn concepts and develop skills for sequence analysis
  - Build the foundation for tackling your own analysis challenges
  - Learn to think like a bioinformatician
  - Have fun

Introductions to Bioinformatics instructors (WashU team)

Malachi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

Obi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

Chris Miller
Associate Professor of Medicine

Felicia Gomez
Assistant Professor of Medicine

Charlz Jerold
Bioinformatics
Research Analyst

My Hoang
Graduate Student

Mariam Khanfar
Graduate Student

Kartik Singhal
Graduate Student

Jennie Yao
Graduate Student

Washington University in St. Louis
School of Medicine

GRIFFITHLAB

griffithlab.org
rnabio.org genviz.org pmbio.org
civicdb.org pvactools.org dgidb.org
regtools.org vatools.org

Introductions to Bioinformatics instructors (WashU team)

Beard

Malachi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

Obi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

Chris Miller
Associate Professor of Medicine

Felicia Gomez
Assistant Professor of Medicine

Charlz Jerold
Bioinformatics
Research Analyst

My Hoang
Graduate Student

Mariam Khanfar
Graduate Student

Kartik Singhal
Graduate Student

Jennie Yao
Graduate Student

Washington University in St. Louis
School of Medicine

GRIFFITHLAB

griffithlab.org
rnabio.org genviz.org pmbio.org
civicdb.org pvactools.org dgidb.org
regtools.org vatools.org

Introductions to Bioinformatics instructors (WashU team)

Malachi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

Obi Griffith
Professor of Medicine
Professor of Genetics
Assistant Director, MGI

No Beard

Chris Miller
Associate Professor of Medicine

Felicia Gomez
Assistant Professor of Medicine

Charlz Jerold
Bioinformatics
Research Analyst

My Hoang
Graduate Student

Mariam Khanfar
Graduate Student

Kartik Singhal
Graduate Student

Jennie Yao
Graduate Student

Washington University in St. Louis
School of Medicine

GRIFFITHLAB

griffithlab.org
rnabio.org genviz.org pmbio.org
civicdb.org pvactools.org dgidb.org
regtools.org vatools.org

Student poll (respond in slack)

Not counting the pre-requisites and materials for this course:

- Do you consider yourself a bioinformatician? Computational biologist?
- Are you familiar with linux/command line?
  - Intermediate?
  - Expert?
- Do you sometimes write code?
- Are you familiar with R?
  - Intermediate?
  - Expert?
- Do you use git/github?
- What organism do you work with? (Put an animal emoji in slack)
- Are you interested in bulk RNAseq, scRNAseq, or both?

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

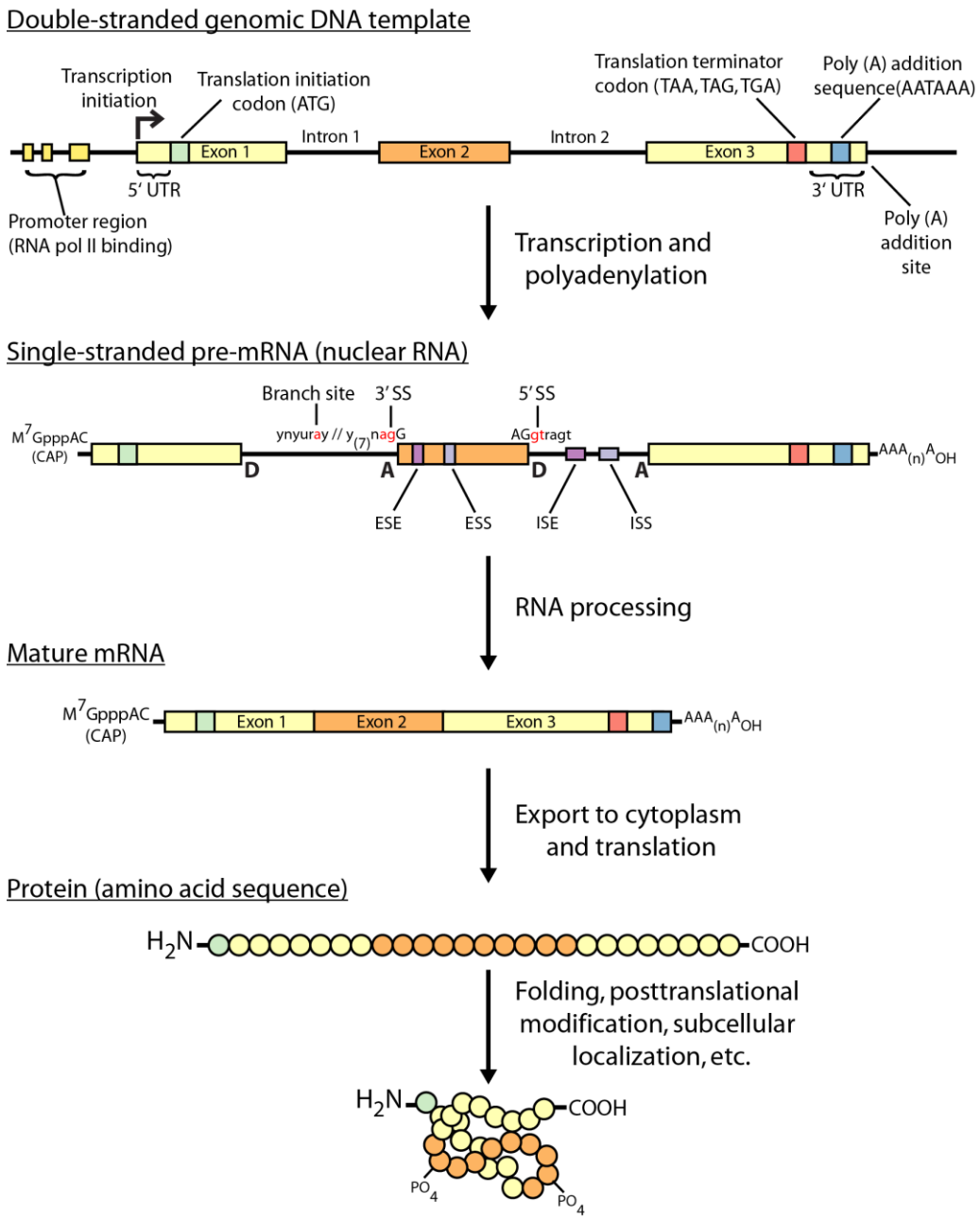
Learning objectives of module 1

- 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 workflows
  - Common technical questions related to RNA-seq analysis
  - Introduction to the RNA-seq hands on tutorial

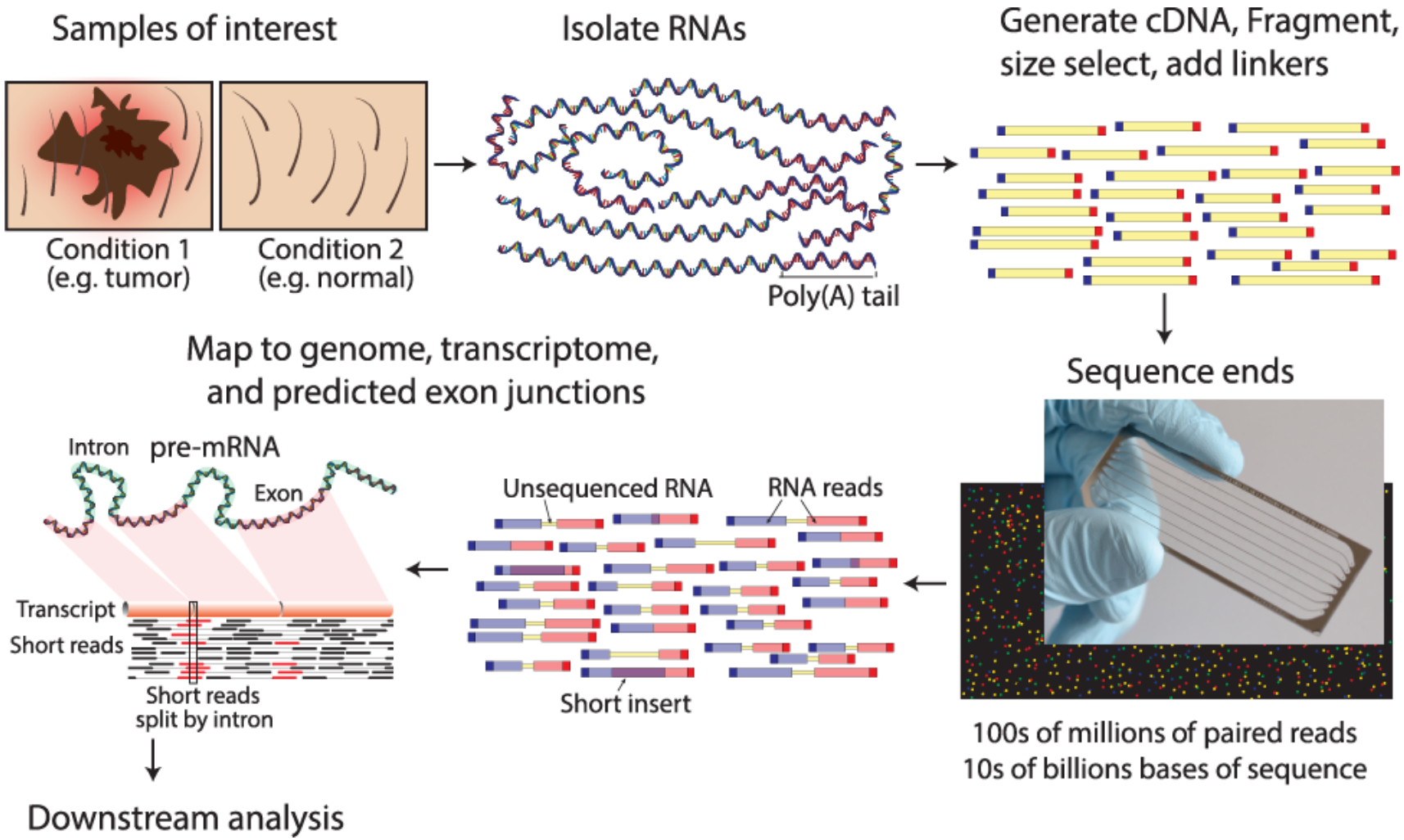
Gene expression (brief review of central dogma)

Thinking about the molecular biology here, what is actually being sequenced in an RNA-seq experiment?

Does it differ depending on the sequencing platform? Or for bulk vs single cell sequencing?



RNA sequencing overview



MPS (NGS) Platforms: Illumina is currently dominant

Production-scale sequencers

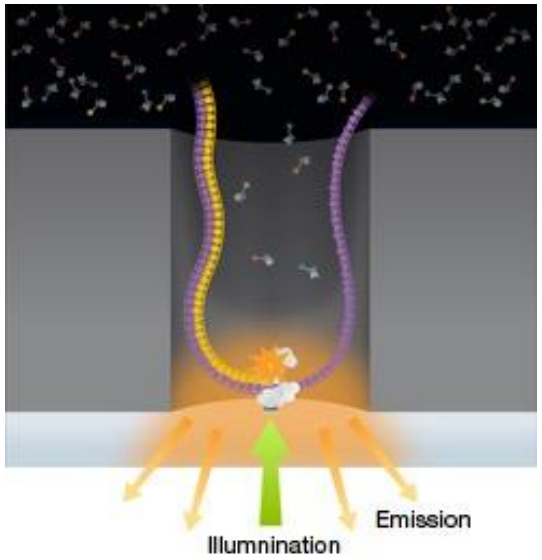
Key specifications	 NextSeq 1000 and 2000 Systems	 NovaSeq 6000 System	 NovaSeq X Series
Max output per flow cell	540 Gb[a]	3 Tb[b]	8 Tb[c]
Run time (range)[d]	~8–44 hr	~13–44 hr	~17–48 hr
Max reads per run (single reads)	1.8B[a]	10B (single flow cell)[b] 20B (dual flow cells)	26B (single flow cell)[c] 52B (dual flow cells)[c,e]
Max read length	2 × 300 bp	2 × 250 bp	2 × 150 bp

- Higher accuracy, range of capacity and throughput
- Slightly longer read lengths on some platforms

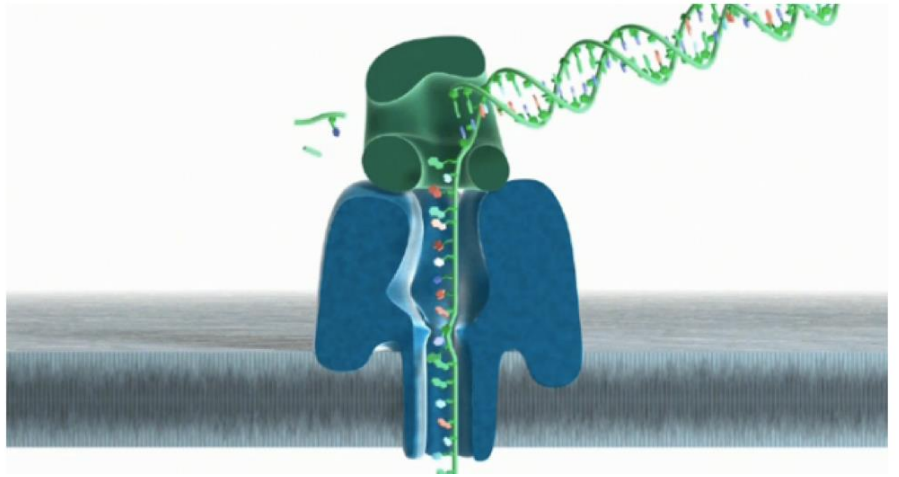
Next-next (3rd) generation sequencing platforms

Defining Characteristics: Long reads (10-100 kb) from single molecules.

Pacific Biosciences: watching a polymerase synthesize DNA/cDNA in real time



Oxford Nanopore: Translocating DNA/RNA through a nanopore with electrode-based detection



The promise: Long reads will allow us to accurately sequence and assemble whole human genomes, from scratch, without using the reference genome.

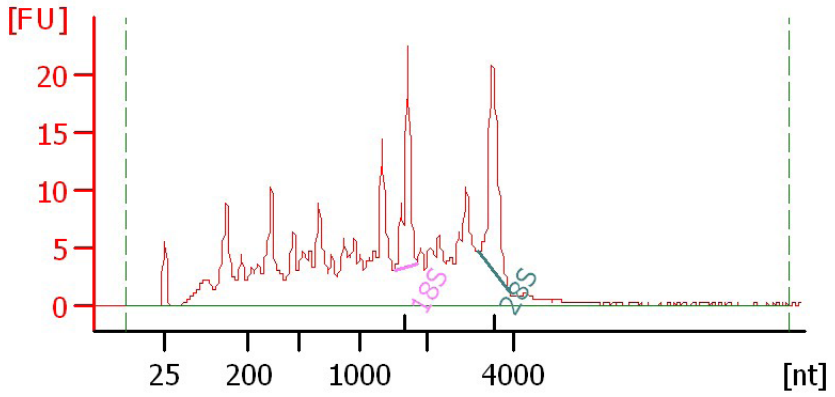
Status: Currently limited by lower throughput, higher base error rate and higher cost. 3rd generation technologies have proven useful, but generally for niche applications so far.

Challenges specific to RNA sequencing

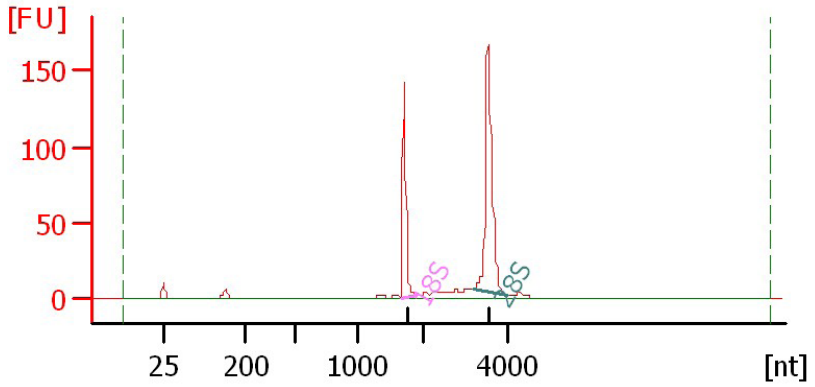
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction 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



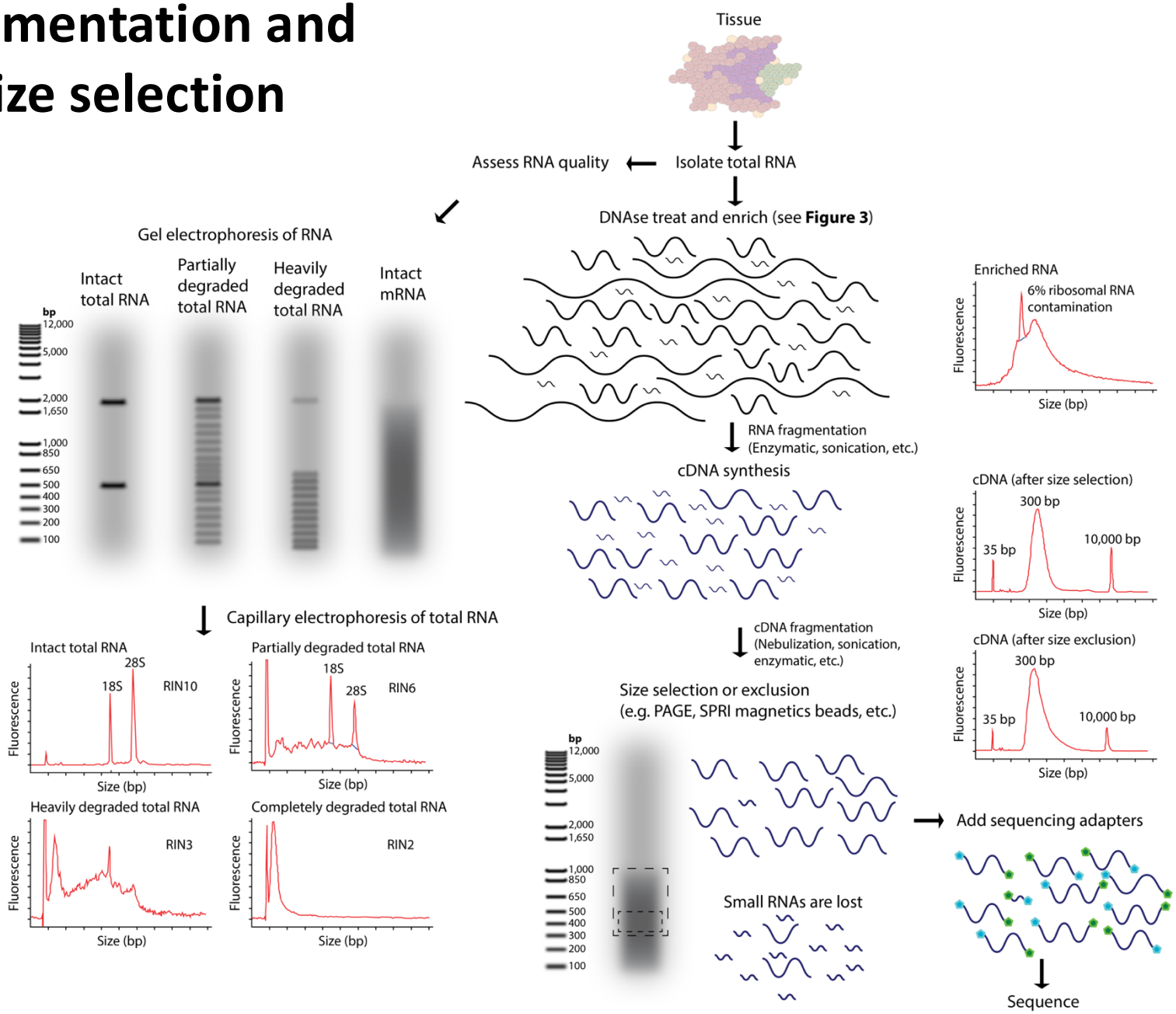
RIN = 10

There are many RNA-seq library construction strategies

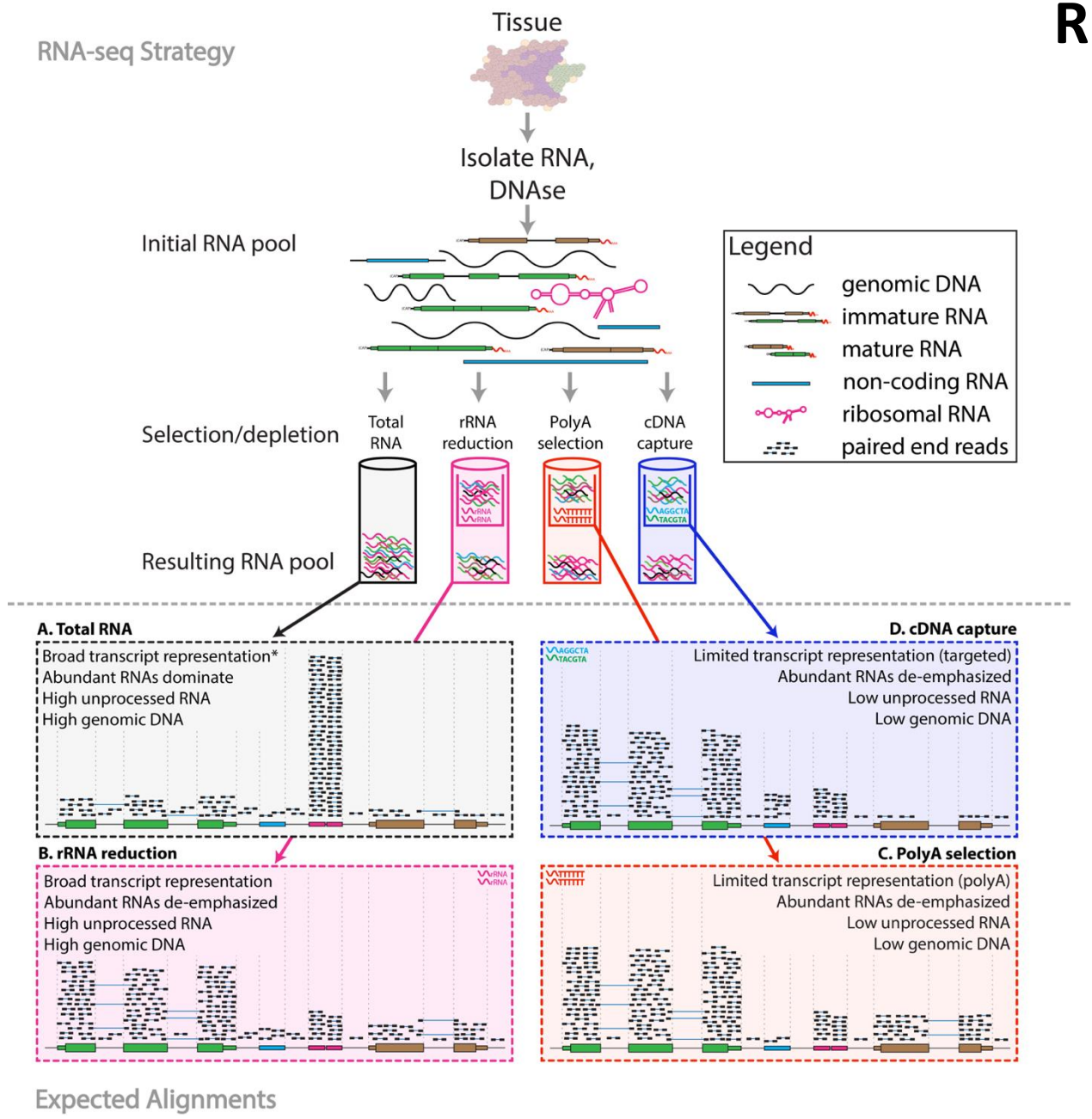
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Library normalization?
  - Exome captured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

- These details can affect analysis strategy
  - Especially comparisons between libraries

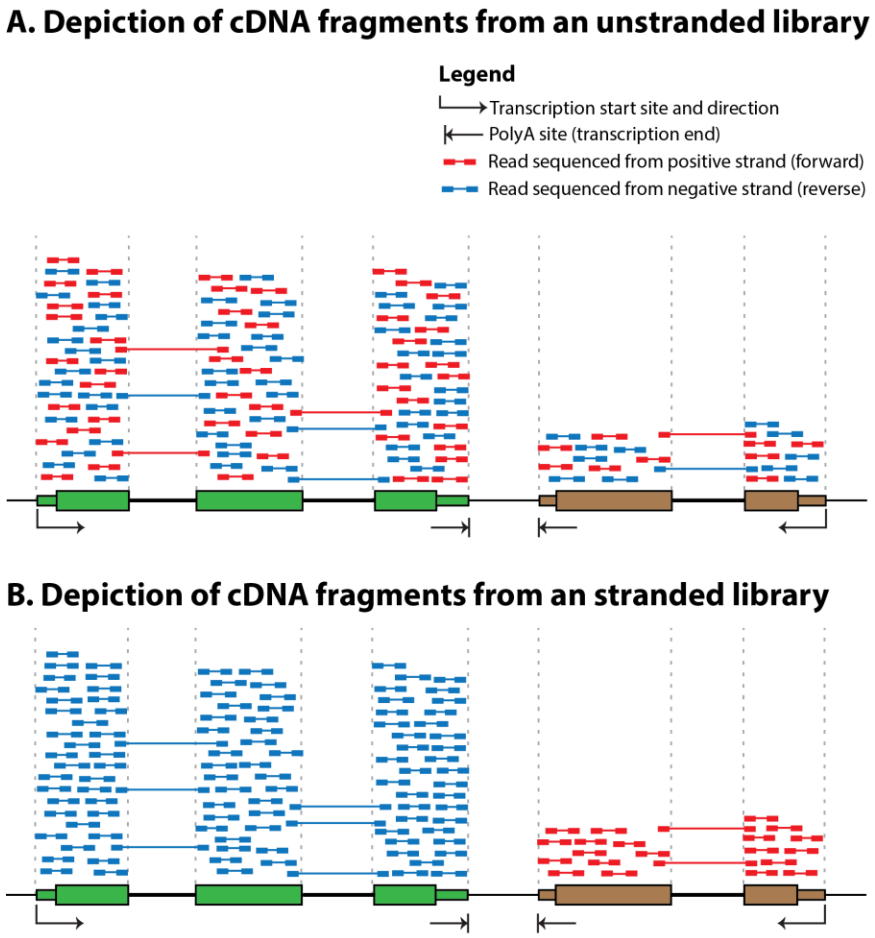
Fragmentation and size selection

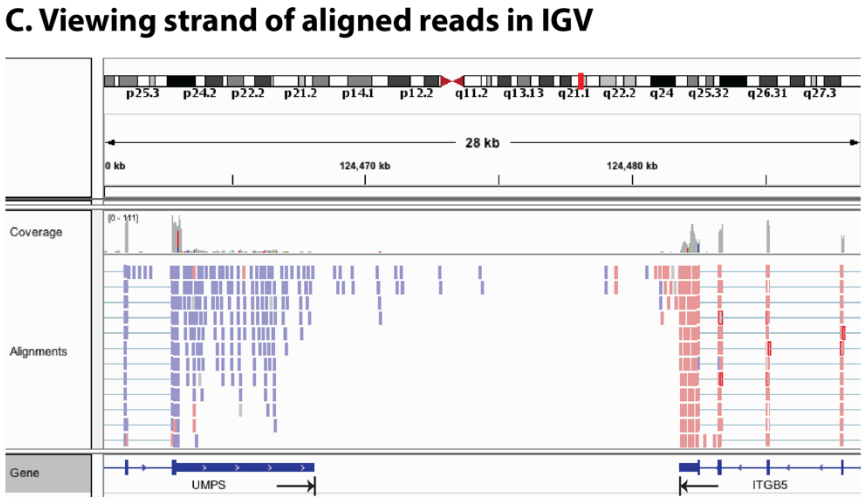


RNA sequence enrichment (selection/depletion)



Stranded vs. unstranded





https://rnabio.org/module-09-appendix/0009/12/01/StrandSettings/
(detailed discussion and cheat sheet)

Ordering RNA-seq data, “coverage”, and cost?

RNA-seq full service, cost per sample[a]	< 12 Samples	≥ 2500 Samples*	~Targeted Coverage
PolyA selection	$287	$215	30M reads
Ribosomal depletion, RiboErase (H/M/R)	$297	$226	30M reads
Ribosomal depletion, FastSelect (H/M/R)	$268	$205	30M reads
Ribosomal depletion, FastSelect (H/M/R+Globin)	$272	$213	30M reads
Ribosomal depletion, Watchmaker (H/M/R+Globin)	$291	NA	30M reads
Low input - Takara SMARTseq mRNA	$267	$203	30M reads
Low input - Sigma Seqplex	$273	$202	30M reads

- An example menu from a sequencing core facility (circa 2024)
- Options primarily relate to method of enrichment and input amounts
- “Coverage” is a non-intuitive concept in bulk-RNAseq.
  - 30M reads is sufficient for gene abundance estimation (increase for other applications)

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

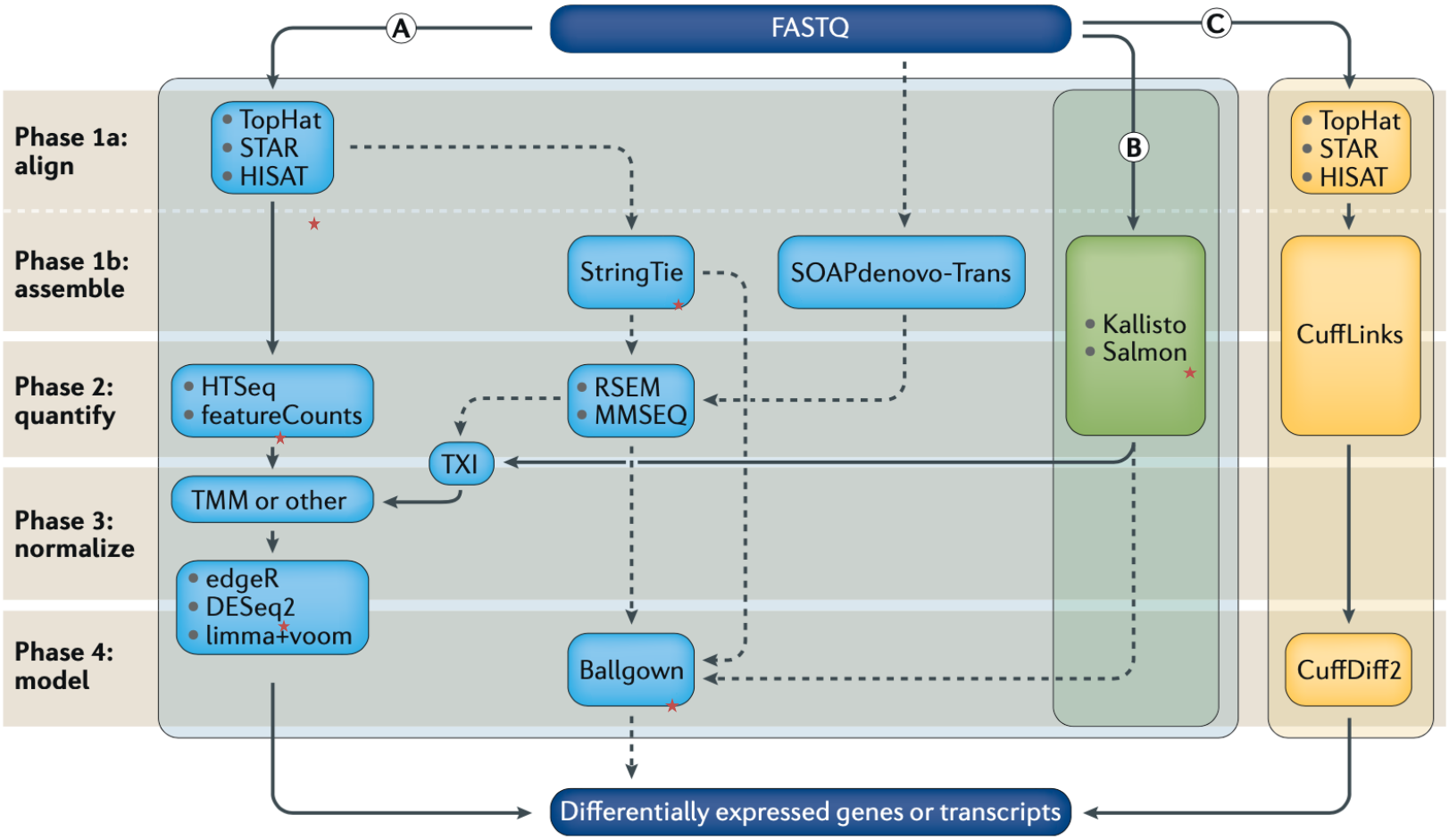
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

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. ‘stringtie’ for expression analysis, ‘star-fusion’ for fusion detection, etc.
4. Post process
   - Import into downstream software (R, python, WebGestalt, etc.)
5. Summarize and visualize
   - Create gene lists, prioritize candidates for validation, etc.

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



[RNA sequencing: the teenage years]

★ Covered in rnabio.org

Discussion of bulk vs single cell RNA-seq

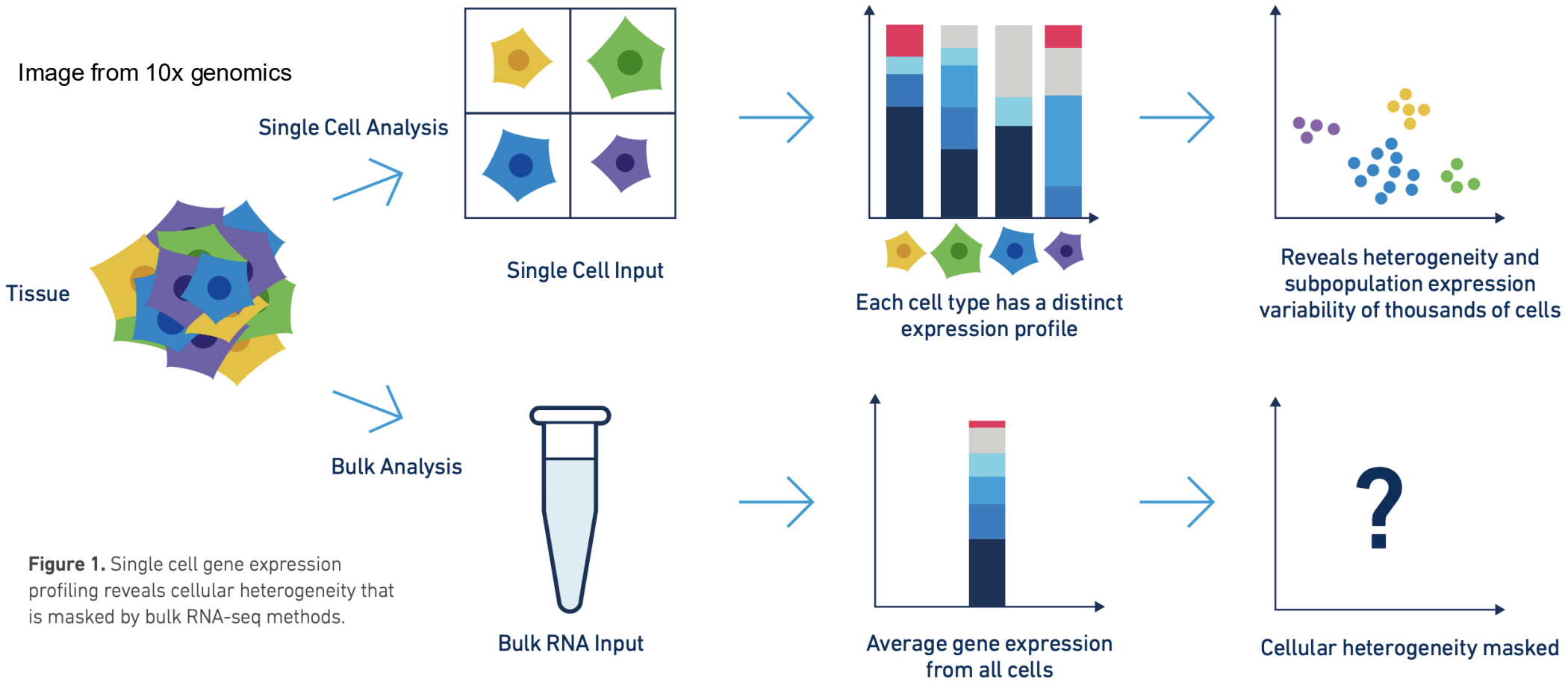
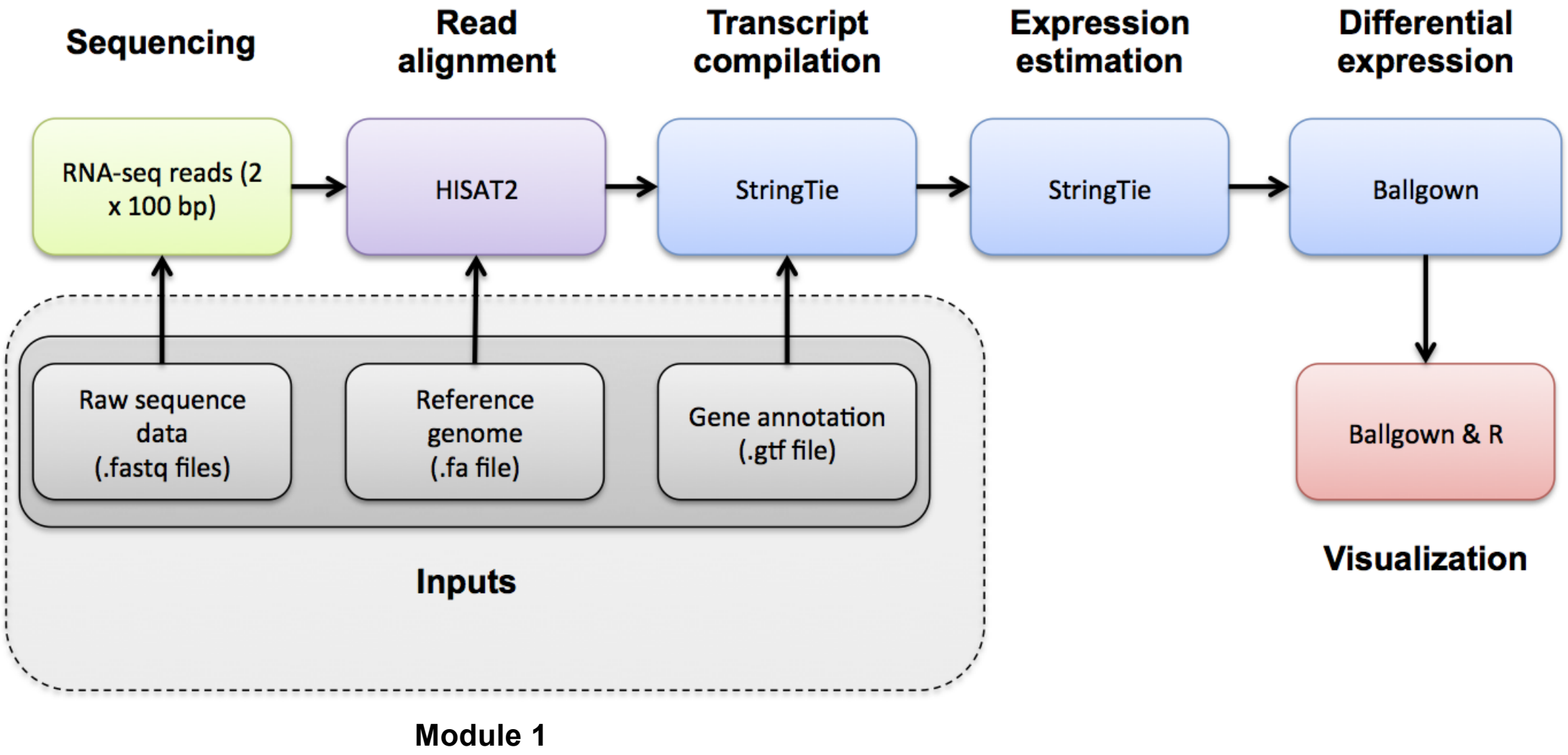


Figure 1. Single cell gene expression profiling reveals cellular heterogeneity that is masked by bulk RNA-seq methods.

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

Introduction to tutorial (Module 1)

HISAT2/StringTie/Ballgown RNA-seq Pipeline



Bioinformatics troubleshooting cheat sheet

- ❑ Check your inputs!
- ❑ Mix of incompatible reference genomes used (see this tutorial)
- ❑ Mix of incompatible gene/transcript identifiers
- ❑ Reference sequence names (e.g. “1” vs “chr1”)
- ❑ 1-based vs 0-based coordinates (see this tutorial)
- ❑ Computational tasks fail due to resource limitations (memory and storage)
- ❑ Dependency hell for bioinformatics tools. Learn to use containers (e.g. docker) or environment managers (e.g. conda)

We are on a Coffee Break & Networking Session

Workshop Sponsors: