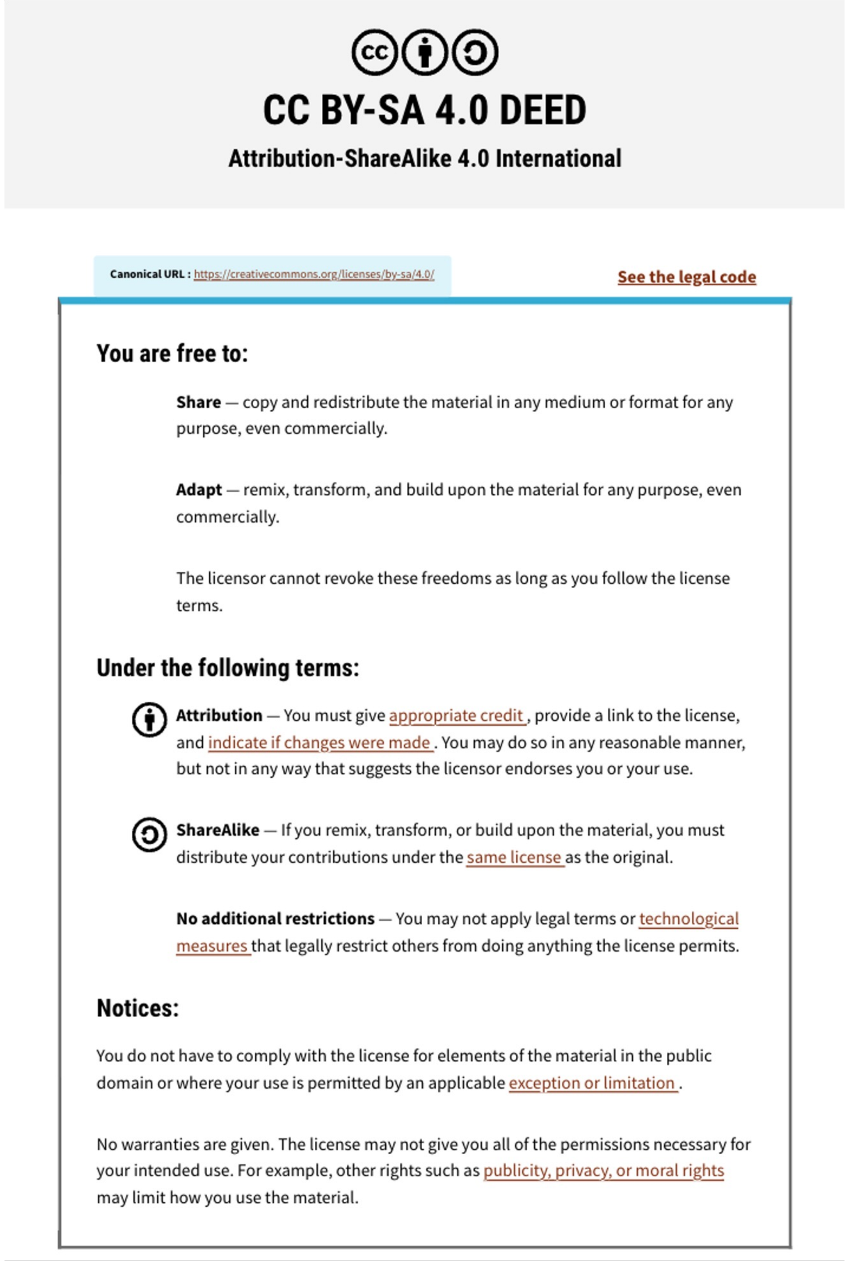
bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io



CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Canonical URL : https://creativecommons.org/licenses/by-sa/4.0/

See the legal code

You are free to:

Share — copy and redistribute the material in any medium or format for any purpose, even commercially.

Adapt — remix, transform, and build upon the material for any purpose, even commercially.

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

Attribution — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation .

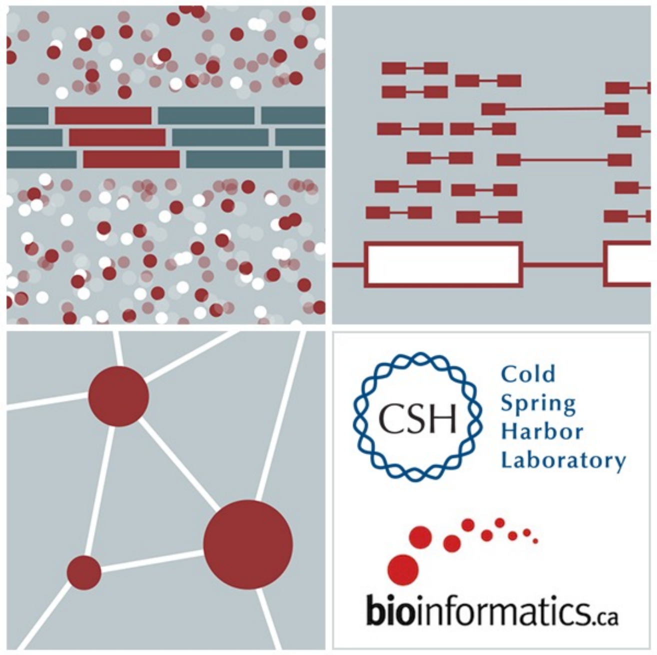
No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material.

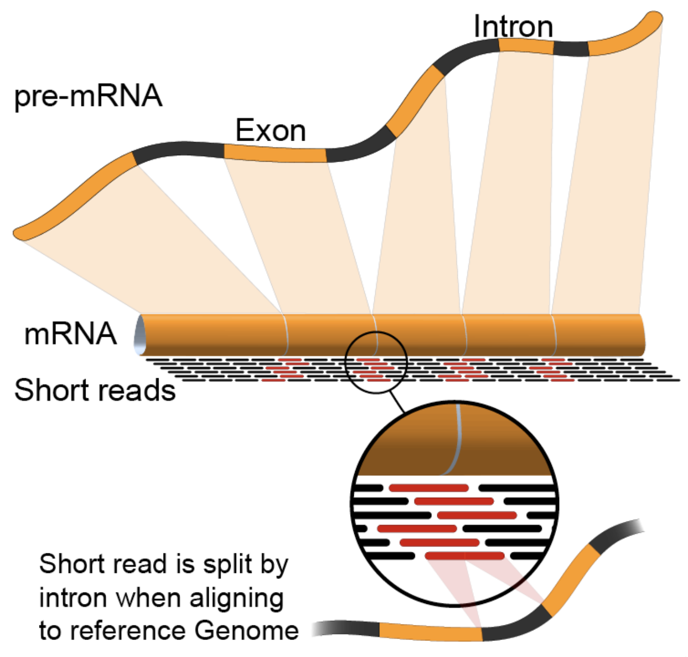
Introduction to RNA sequencing (lecture)

Malachi Griffith, Obi Griffith, Isabel Risch,
Nicolas Ho, Melisa Acun, Varinder Verma, Mobin Khoramjoo

RNA-seq Analysis 2025. July 7-9, 2025

bioinformatics.ca




Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

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

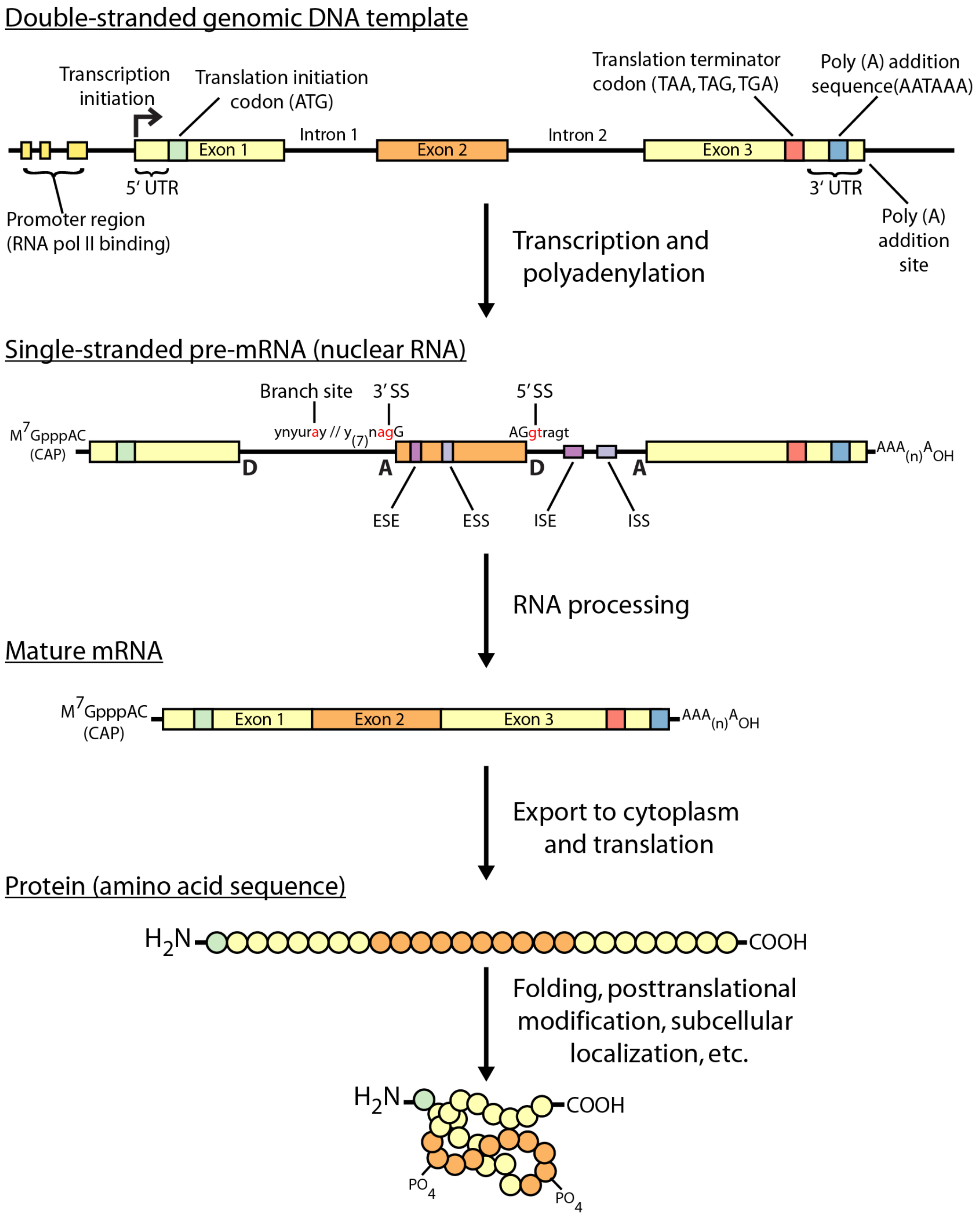
Learning objectives of module 1

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

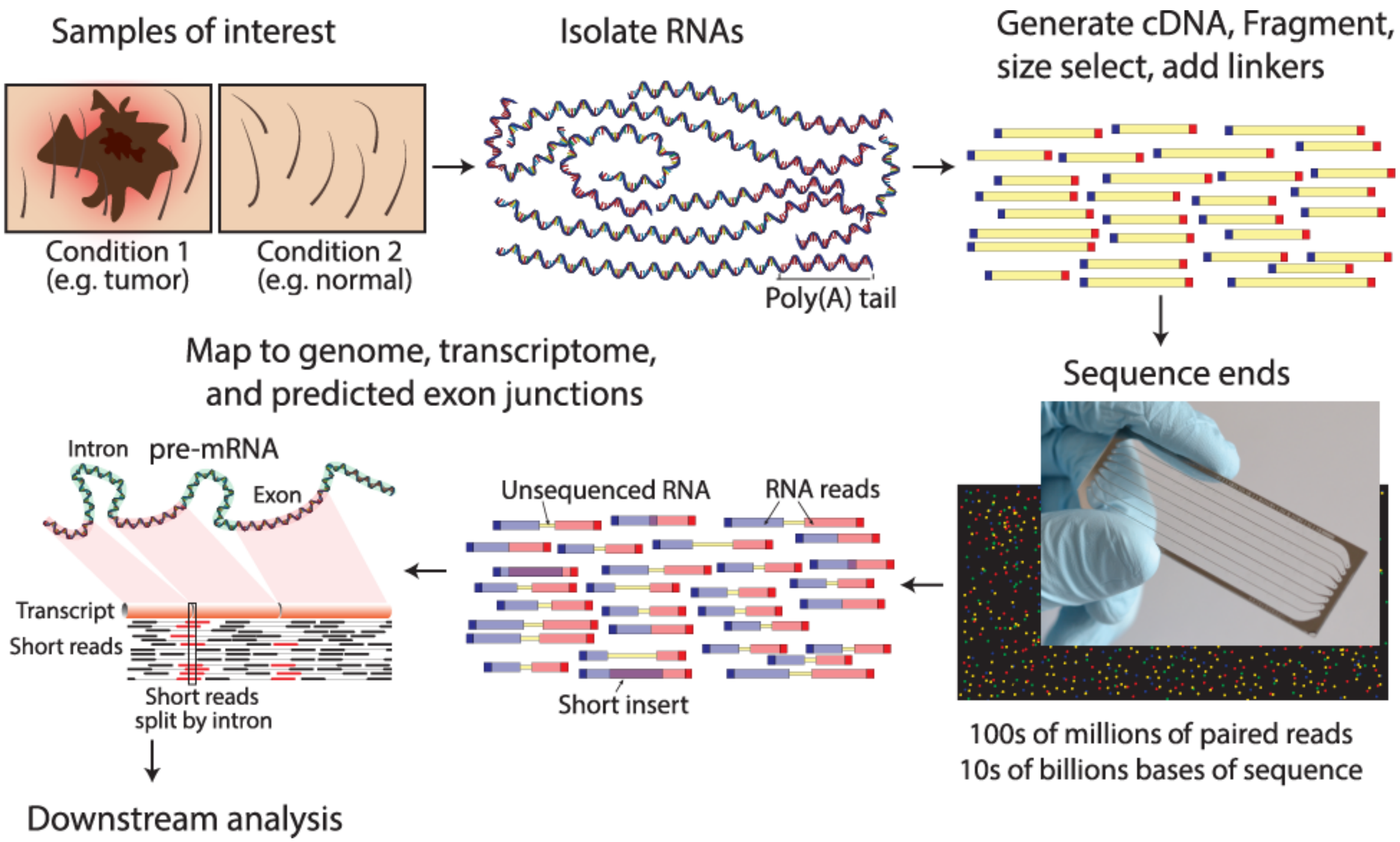
Gene expression

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?


Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
M^7GpppAC (CAP)
ynyuray // $y_{(7)}$nagG
AGgtragt
D
A
D
A
$AAA_{(n)}A_{OH}$
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M^7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
$AAA_{(n)}A_{OH}$
Export to cytoplasm and translation
Protein (amino acid sequence)
H_2N
COOH
Folding, posttranslational modification, subcellular localization, etc.
H_2N
COOH
PO_4
PO_4

RNA sequencing



Samples of interest

Condition 1
(e.g. tumor)

Condition 2
(e.g. normal)

Isolate RNAs

Poly(A) tail

Generate cDNA, Fragment, size select, add linkers

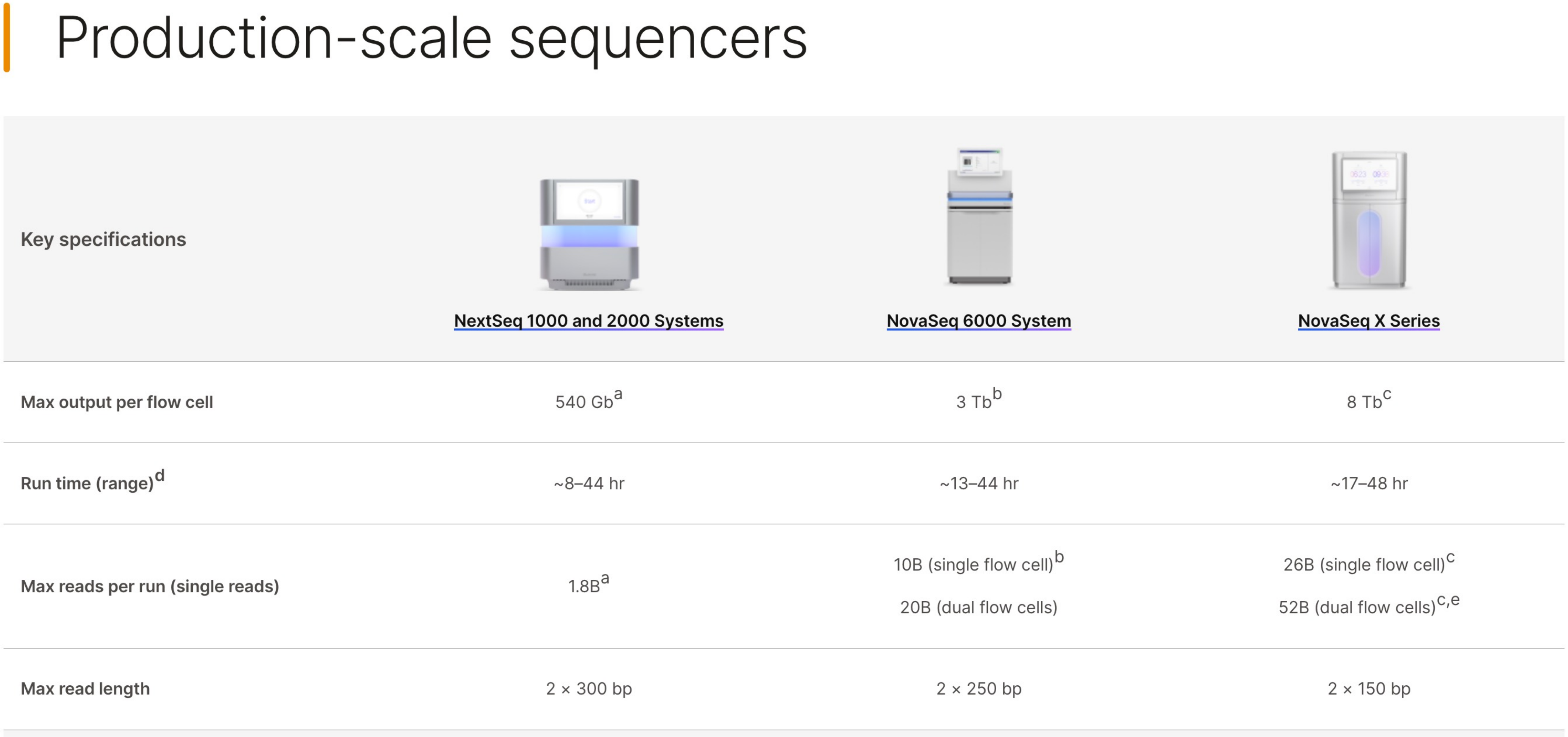
Sequence ends

100s of millions of paired reads
10s of billions bases of sequence

Map to genome, transcriptome, and predicted exon junctions

Downstream analysis

MPS (NGS) Platforms: Illumina is currently dominant



Production-scale sequencers

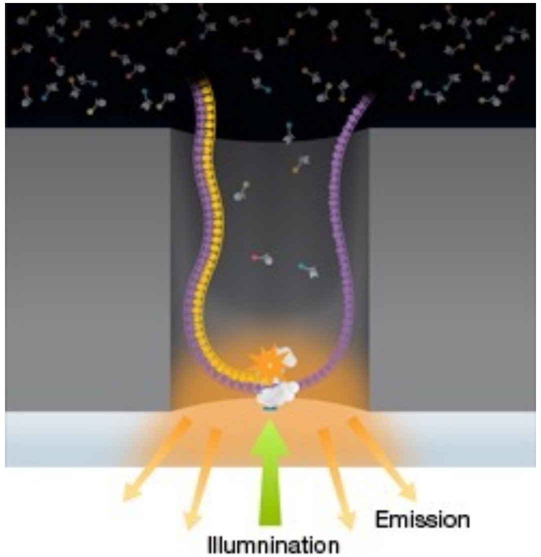
Key specifications	NextSeq 1000 and 2000 Systems	NovaSeq 6000 System	NovaSeq X Series
Max output per flow cell	540 Gb[a]	3 Tb[b]	8 Tb[c]
Run time (range)[d]	~8–44 hr	~13–44 hr	~17–48 hr
Max reads per run (single reads)	1.8B[a]	10B (single flow cell)[b] 20B (dual flow cells)	26B (single flow cell)[c]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

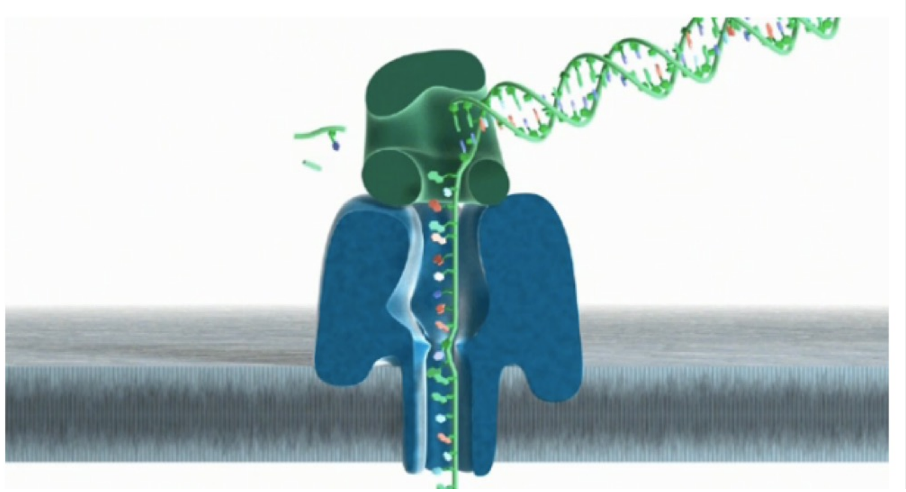
Next-next (3rd) generation sequencing platforms

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



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

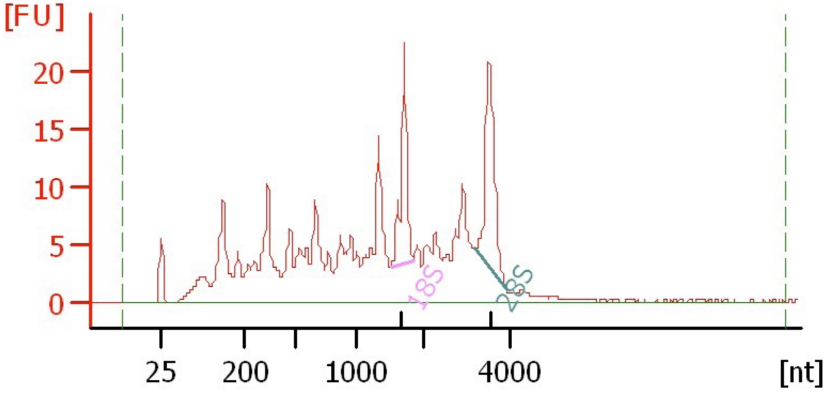
<u>Status</u>: Currently limited by lower throughput, higher base error rate and higher cost. 3rd generation technologies have proven useful, but generally for niche applications so far.

Challenges

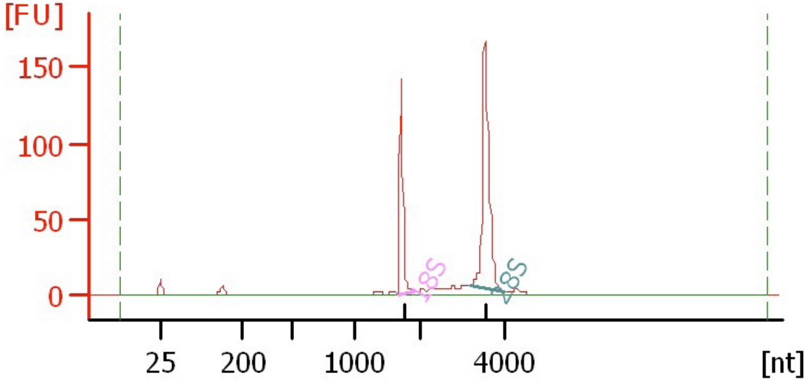
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - $10^5 – 10^7$ orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction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



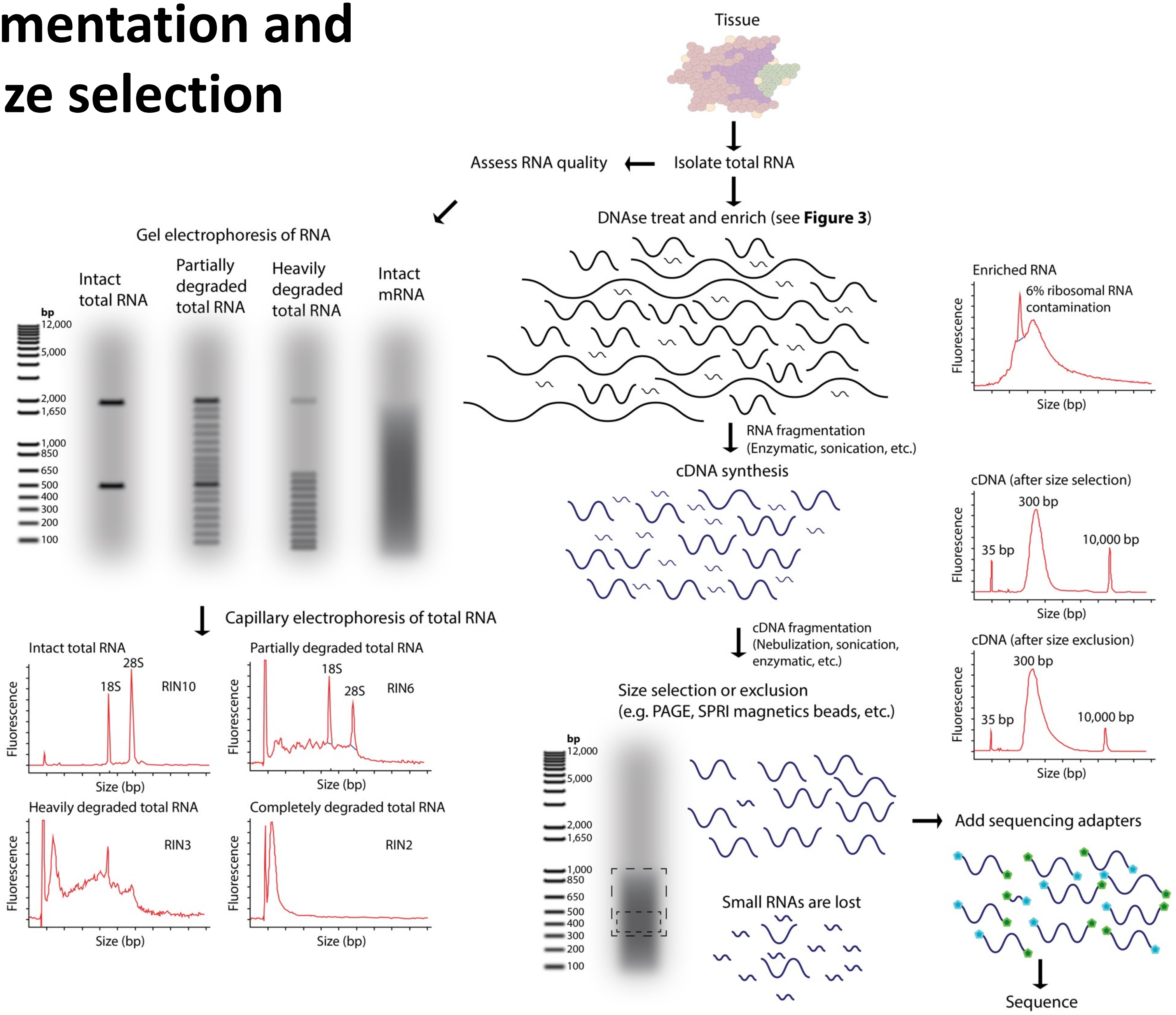
RIN = 10

There are many RNA-seq library construction strategies

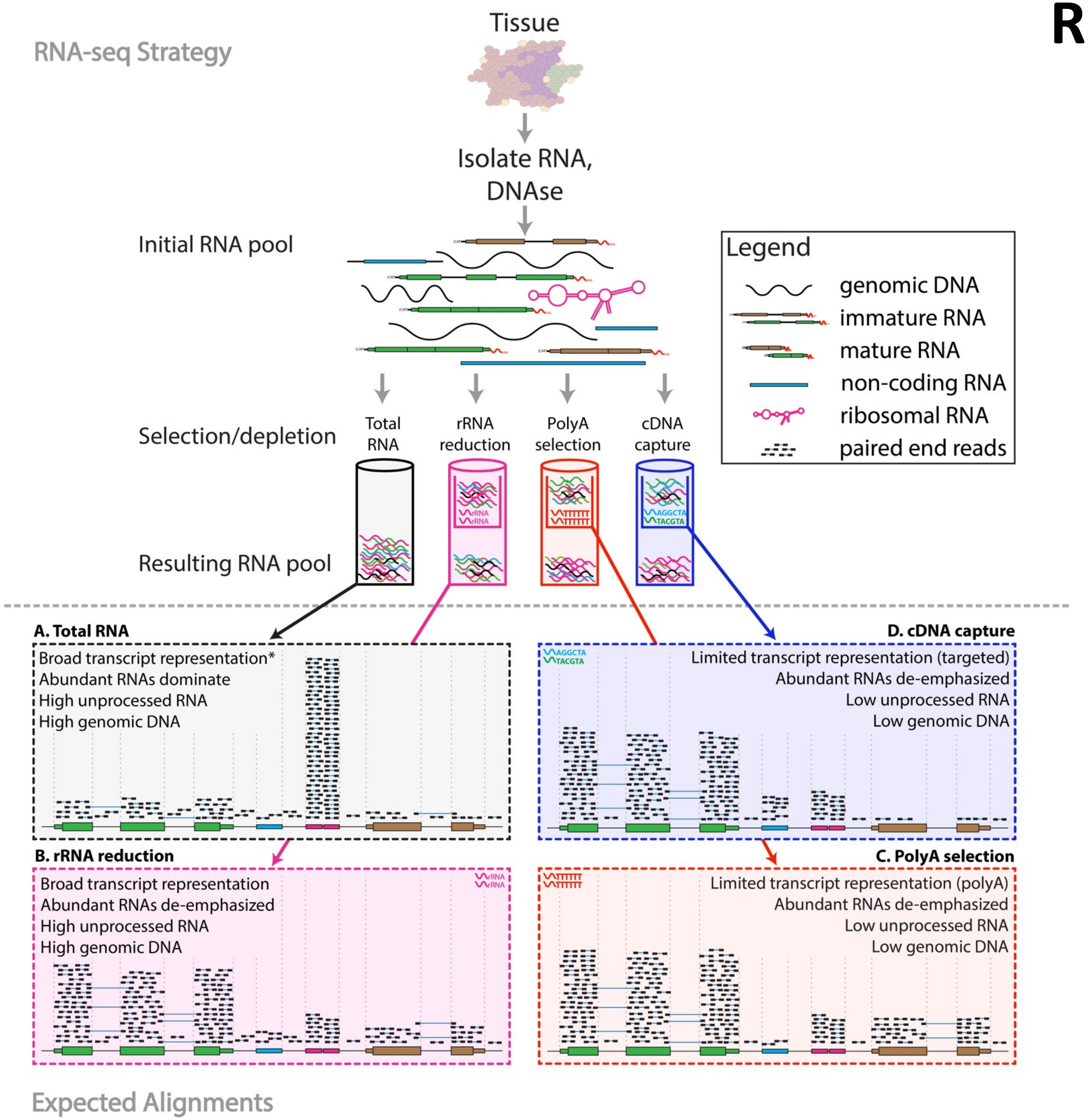
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library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

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

Fragmentation and size selection

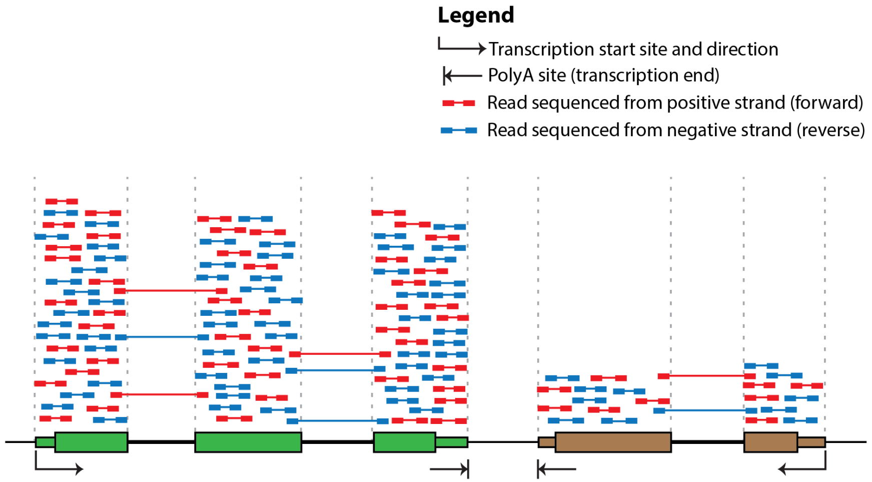

Tissue
Isolate total RNA
Assess RNA quality
DNAse treat and enrich (see **Figure 3**)
Gel electrophoresis of RNA
Intact total RNA
Partially degraded total RNA
Heavily degraded total RNA
Intact mRNA
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100
Enriched RNA
6% ribosomal RNA contamination
Fluorescence
Size (bp)
RNA fragmentation (Enzymatic, sonication, etc.)
cDNA synthesis
cDNA (after size selection)
300 bp
35 bp
10,000 bp
Fluorescence
Size (bp)
Capillary electrophoresis of total RNA
Intact total RNA
28S
18S
RIN10
Fluorescence
Size (bp)
Partially degraded total RNA
18S
28S
RIN6
Fluorescence
Size (bp)
Heavily degraded total RNA
RIN3
Fluorescence
Size (bp)
Completely degraded total RNA
RIN2
Fluorescence
Size (bp)
cDNA fragmentation (Nebulization, sonication, enzymatic, etc.)
cDNA (after size exclusion)
300 bp
35 bp
10,000 bp
Fluorescence
Size (bp)
Size selection or exclusion (e.g. PAGE, SPRI magnetics beads, etc.)
bp
12,000
5,000
2,000
1,650
1,000
850
650
500
400
300
200
100
Small RNAs are lost
Add sequencing adapters
Sequence

RNA sequence enrichment (selection/depletion)

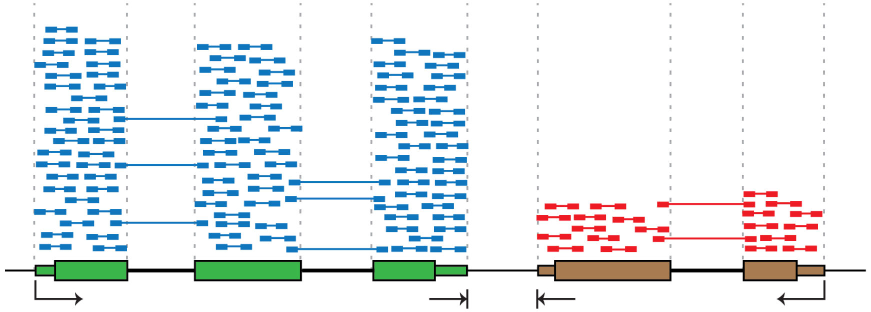


Stranded vs. unstranded

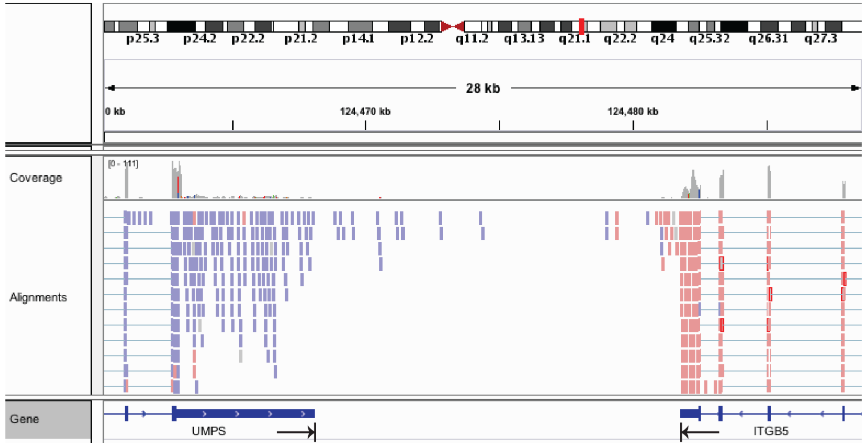
A. Depiction of cDNA fragments from an unstranded library


Legend
Transcription start site and direction
PolyA site (transcription end)
Read sequenced from positive strand (forward)
Read sequenced from negative strand (reverse)

B. Depiction of cDNA fragments from an stranded library



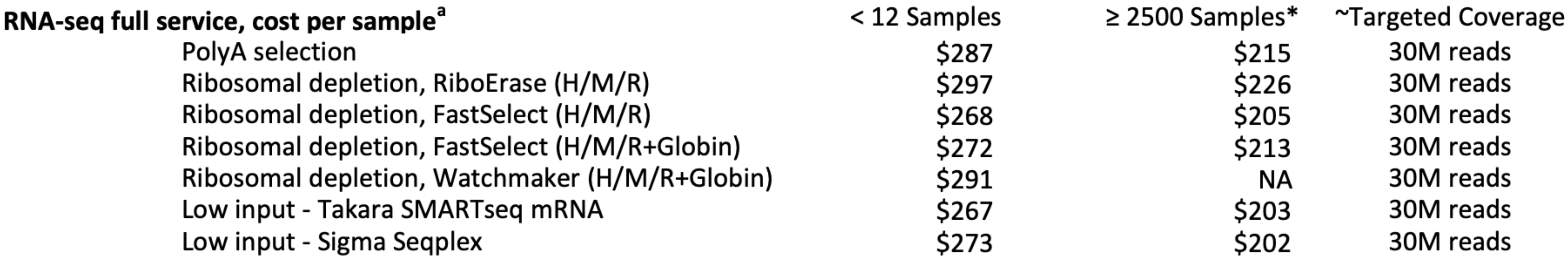
C. Viewing strand of aligned reads in IGV



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?)

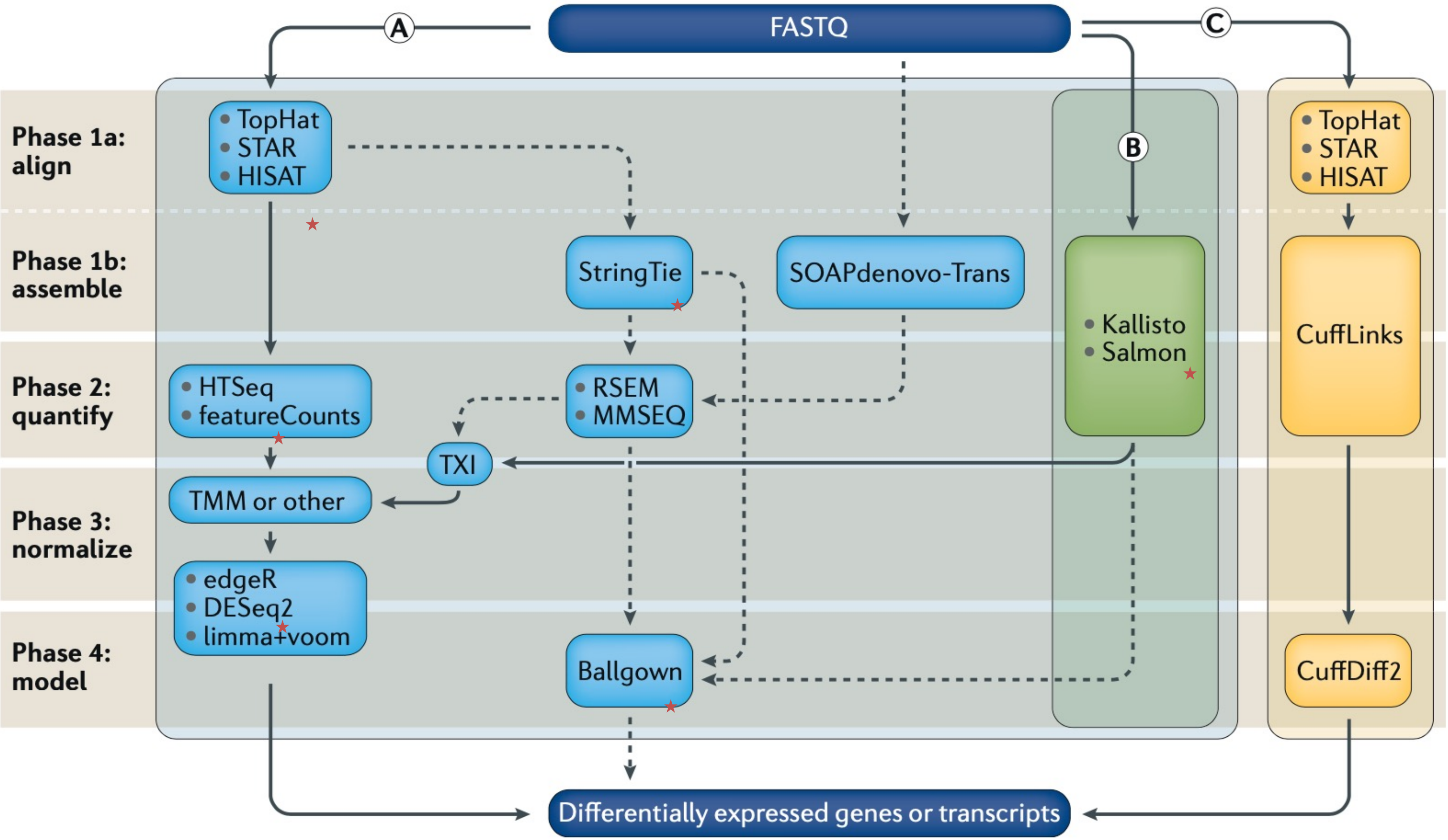
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

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. ‘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



RNA sequencing: the teenage years

★Covered in rnabio.org

Discussion of bulk vs single cell RNA-seq

Image from 10x genomics

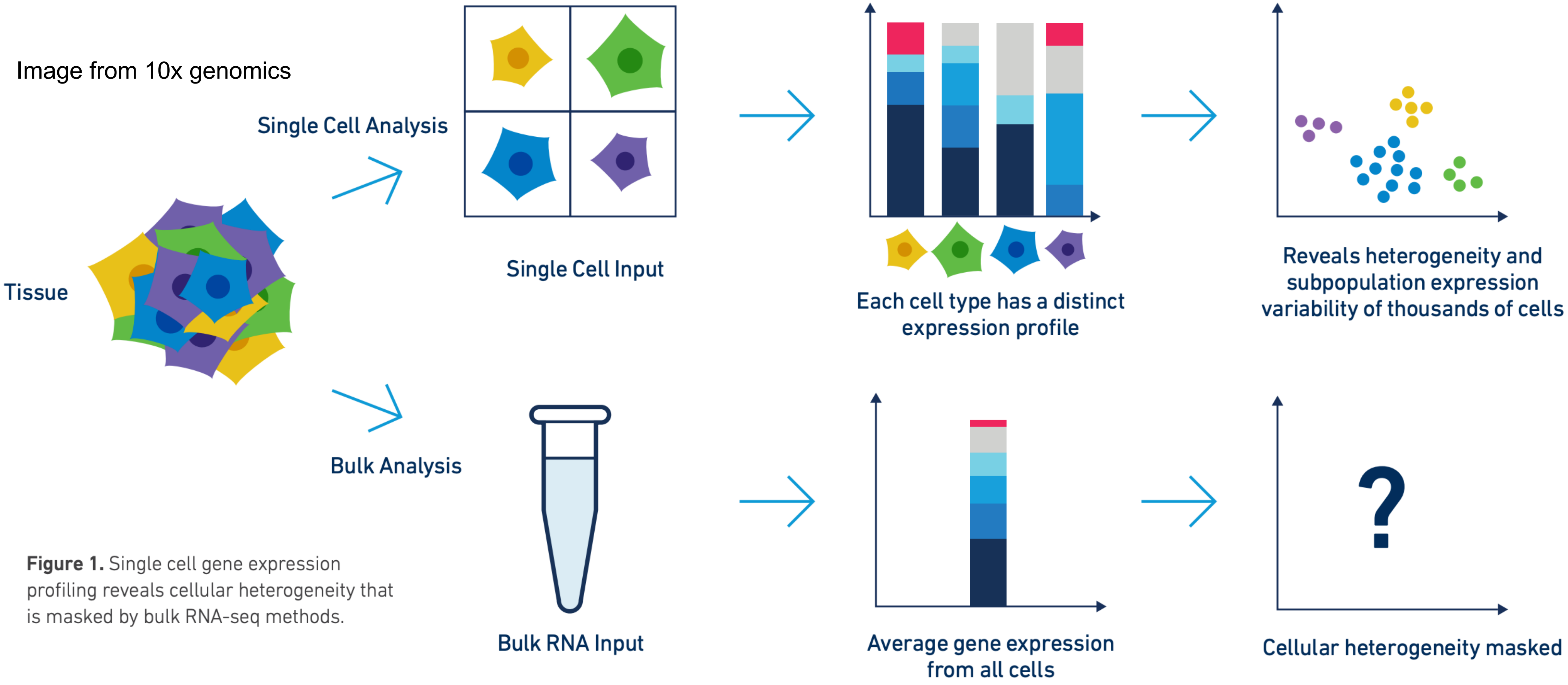

Single Cell Analysis
Tissue
Single Cell Input
Each cell type has a distinct expression profile
Reveals heterogeneity and subpopulation expression variability of thousands of cells
Bulk Analysis
Bulk RNA Input
Average gene expression from all cells
?
Cellular heterogeneity masked

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.

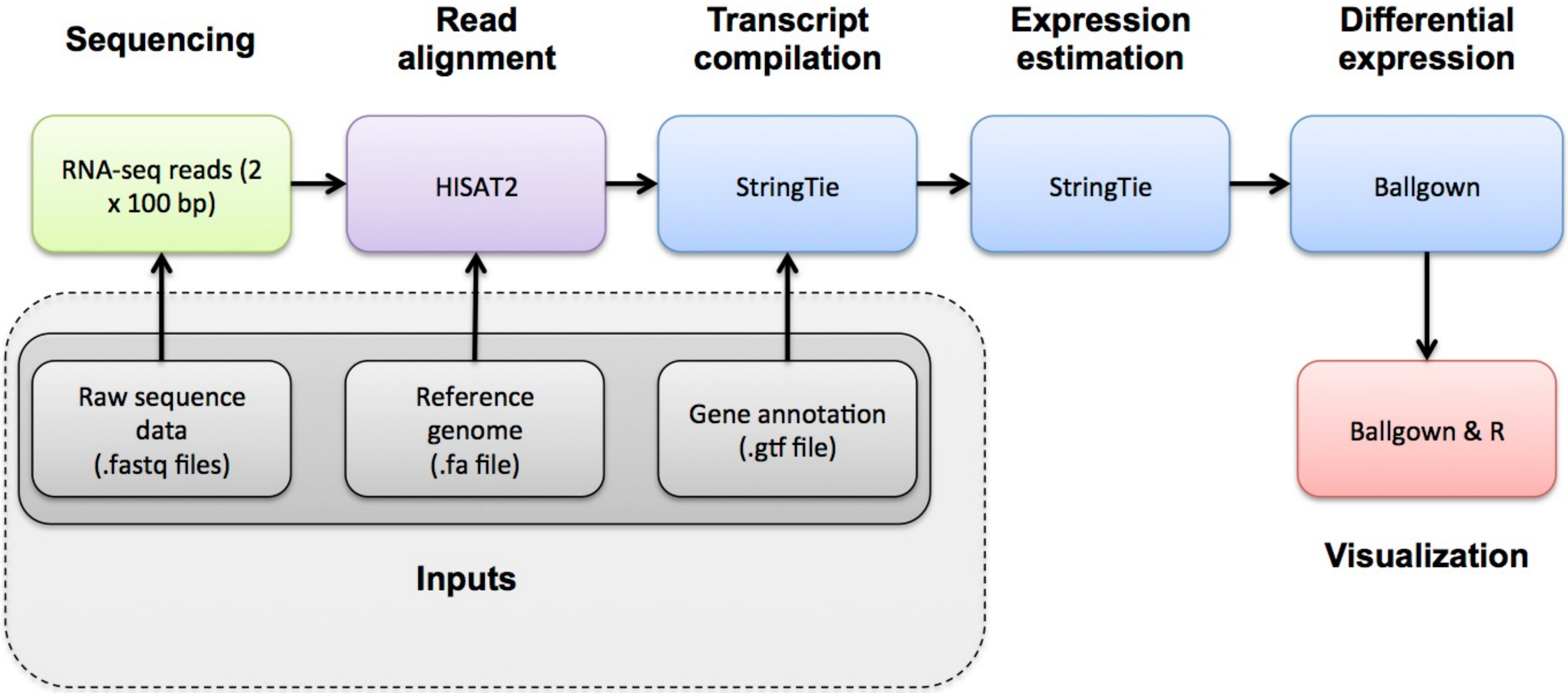
Common questions (and answers)

- 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



Module 1

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:

C3G Canadian Centre for Computational Genomics

HPC4Health

OICR Ontario Institute for Cancer Research

Ontario Genomics

GenomeCanada