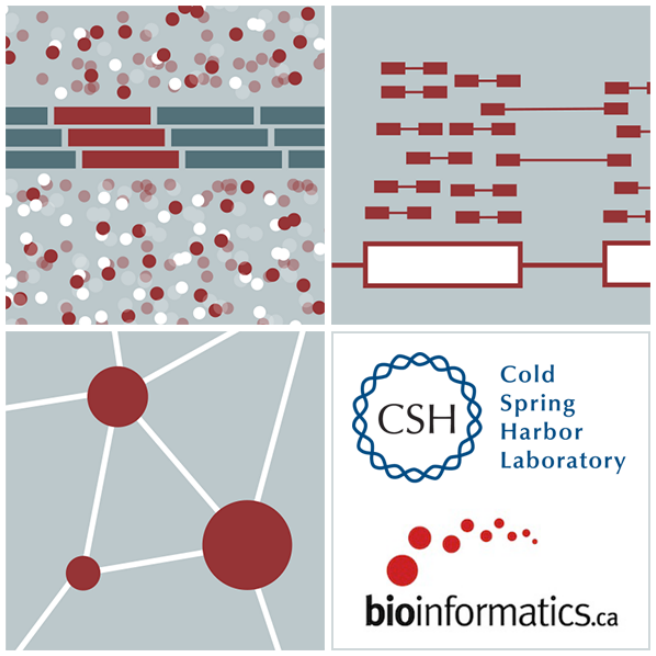
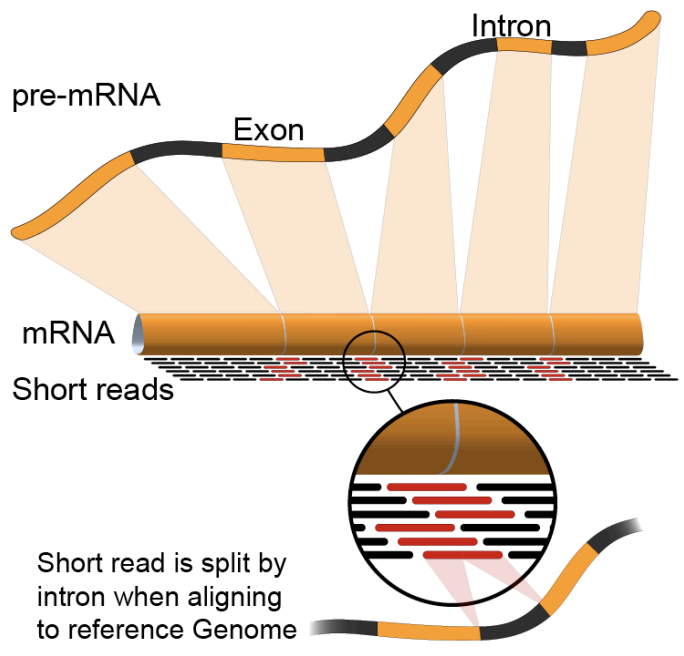
Introduction to RNA sequencing)

Kelsy Cotto, Felicia Gomez, Obi Griffith, Malachi Griffith, Megan Richters, Huiming Xia

Bfx workshop December 14th, 2020





Washington University in St. Louis
School of Medicine

Learning objectives of the course

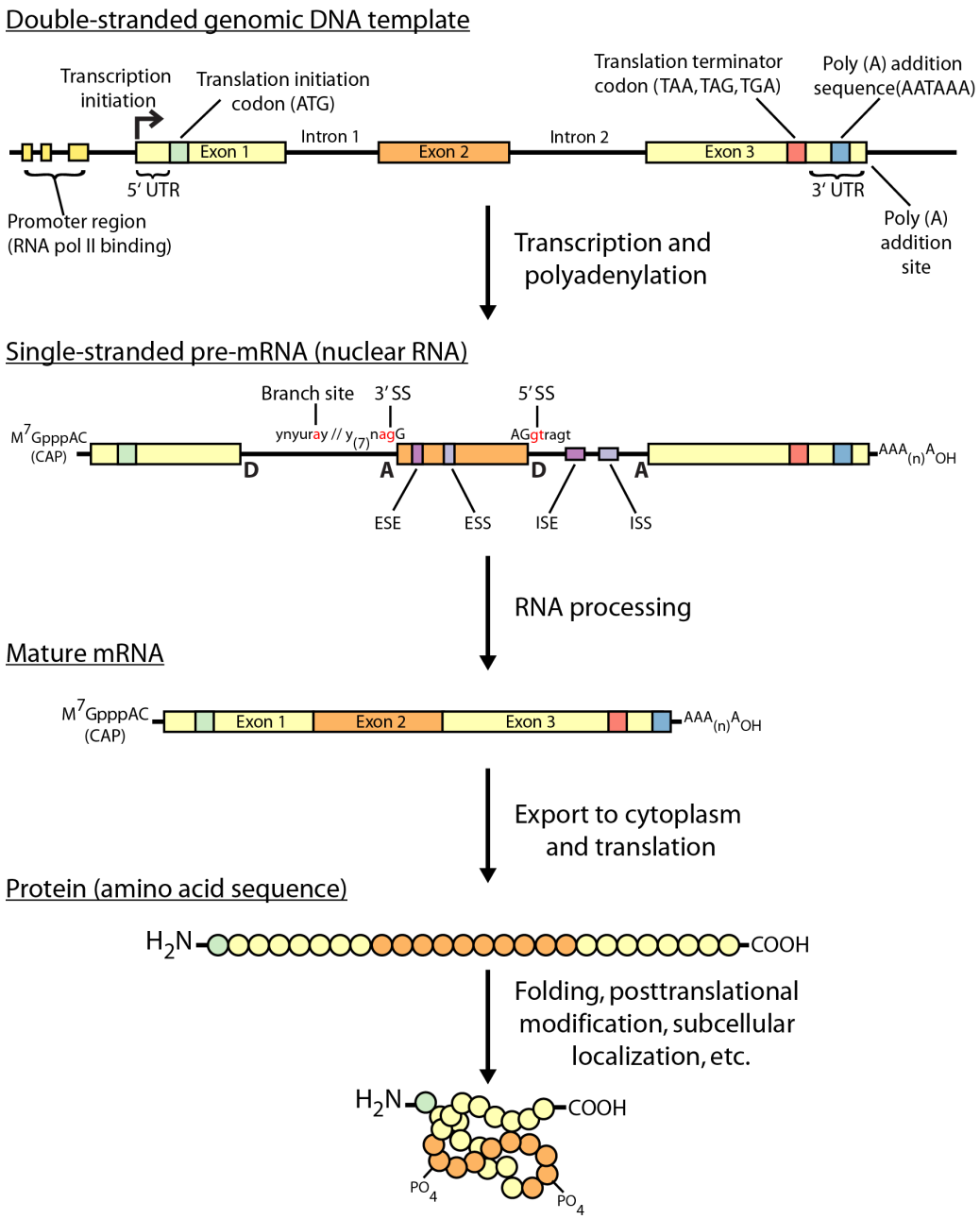
- **Module 1: Introduction to RNA Sequencing**
- Module 2: Alignment and Visualization
- Module 3: Expression and Differential Expression
- Module 4: Alignment Free Expression Estimation
- Module 5: Single Cell RNA-Seq

- 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

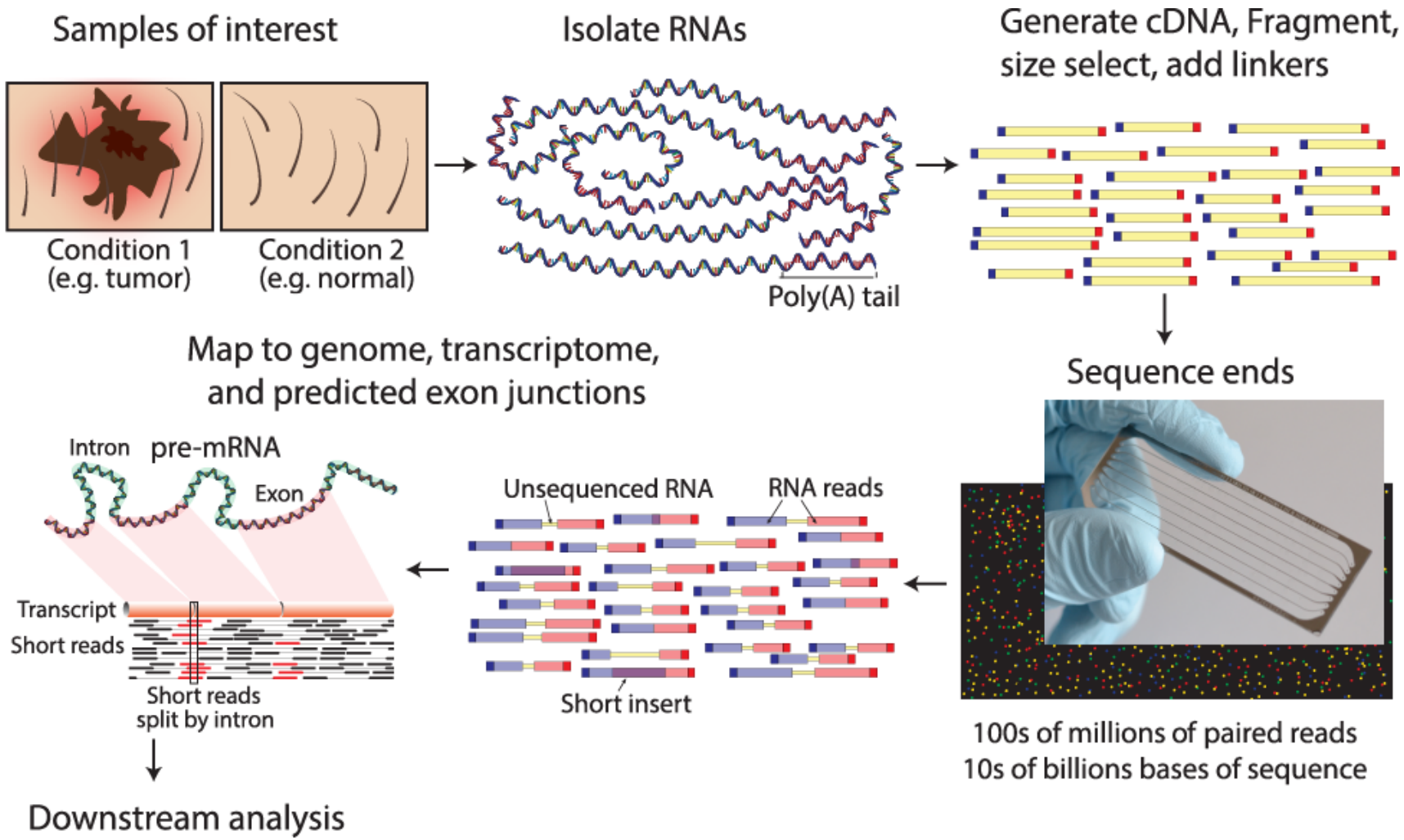
Learning objectives of module 1

- Introduction to the theory and practice of RNA sequencing (RNA-seq) analysis
  - Rationale for sequencing RNA
  - 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



RNA sequencing

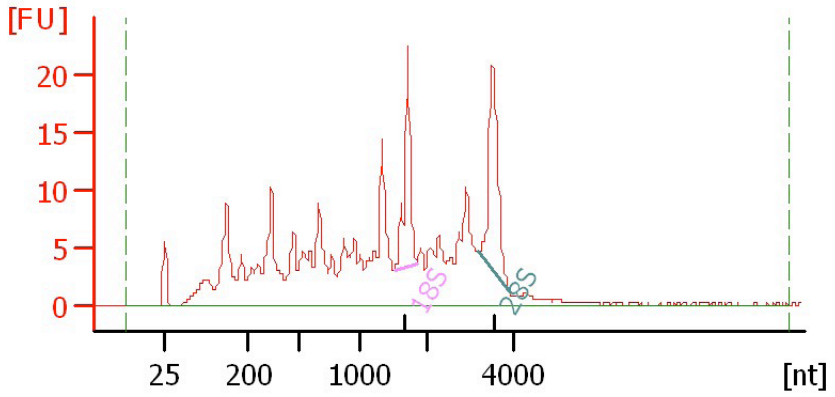


Challenges

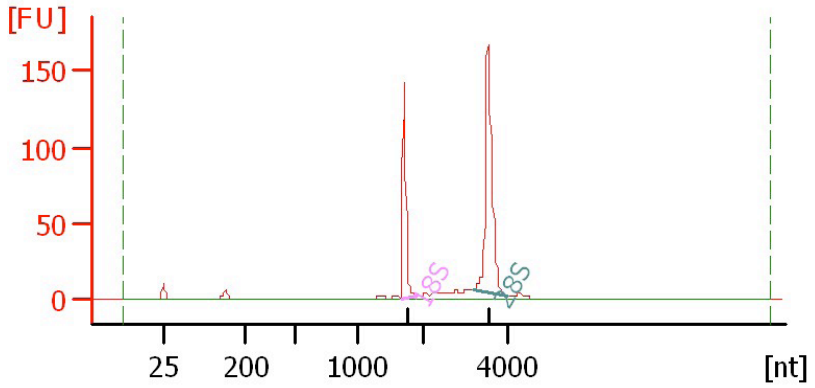
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - 10^5 – 10^7 orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction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

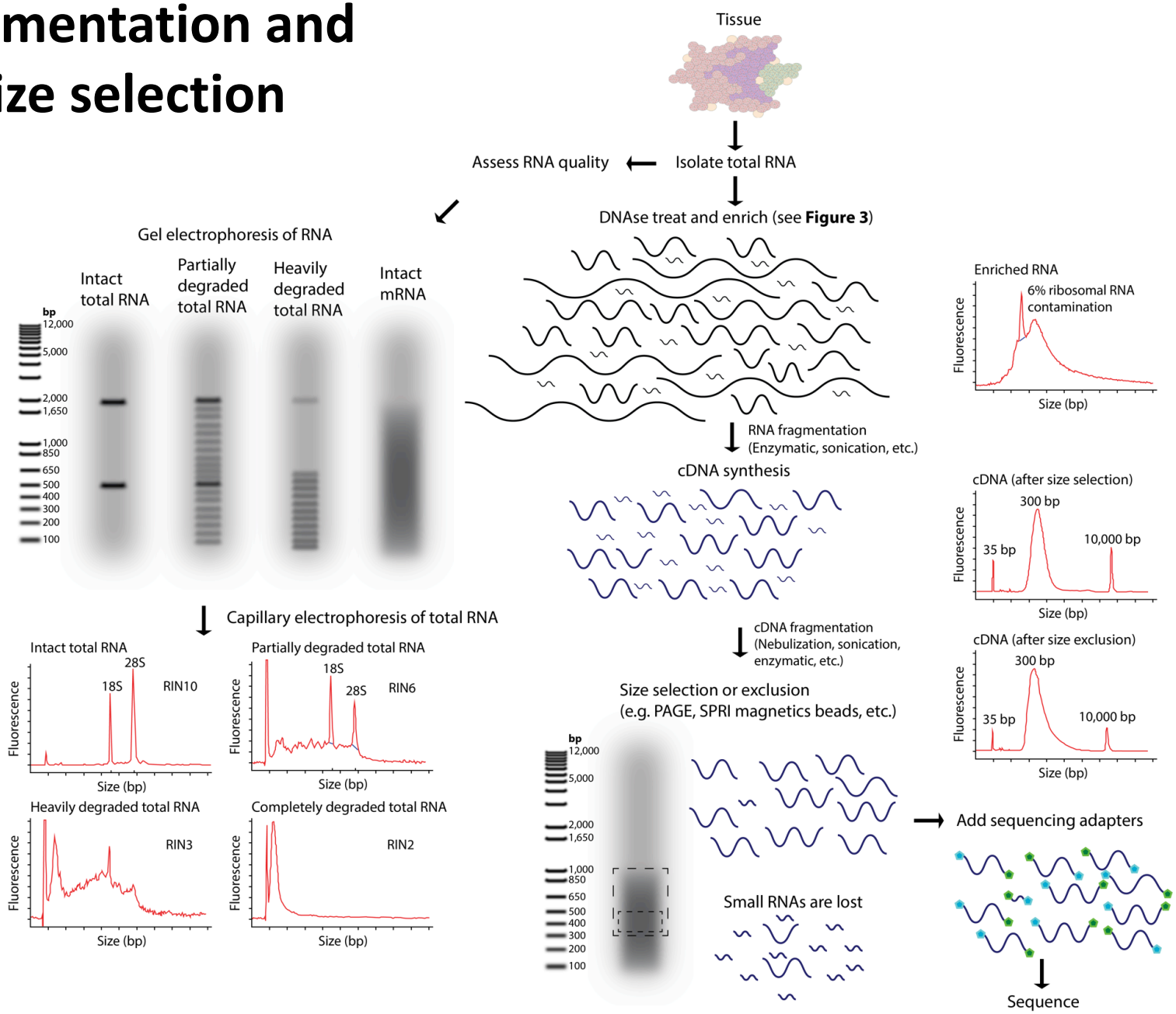
- Several additional initiatives are underway to develop standards and best practices that cover many of these concepts. These include: the Sequencing Quality Control (SEQC) consortium, the Roadmap Epigenomics Mapping Consortium (REMC), and the Beta Cell Biology Consortium (BCBC).

There are many RNA-seq library construction strategies

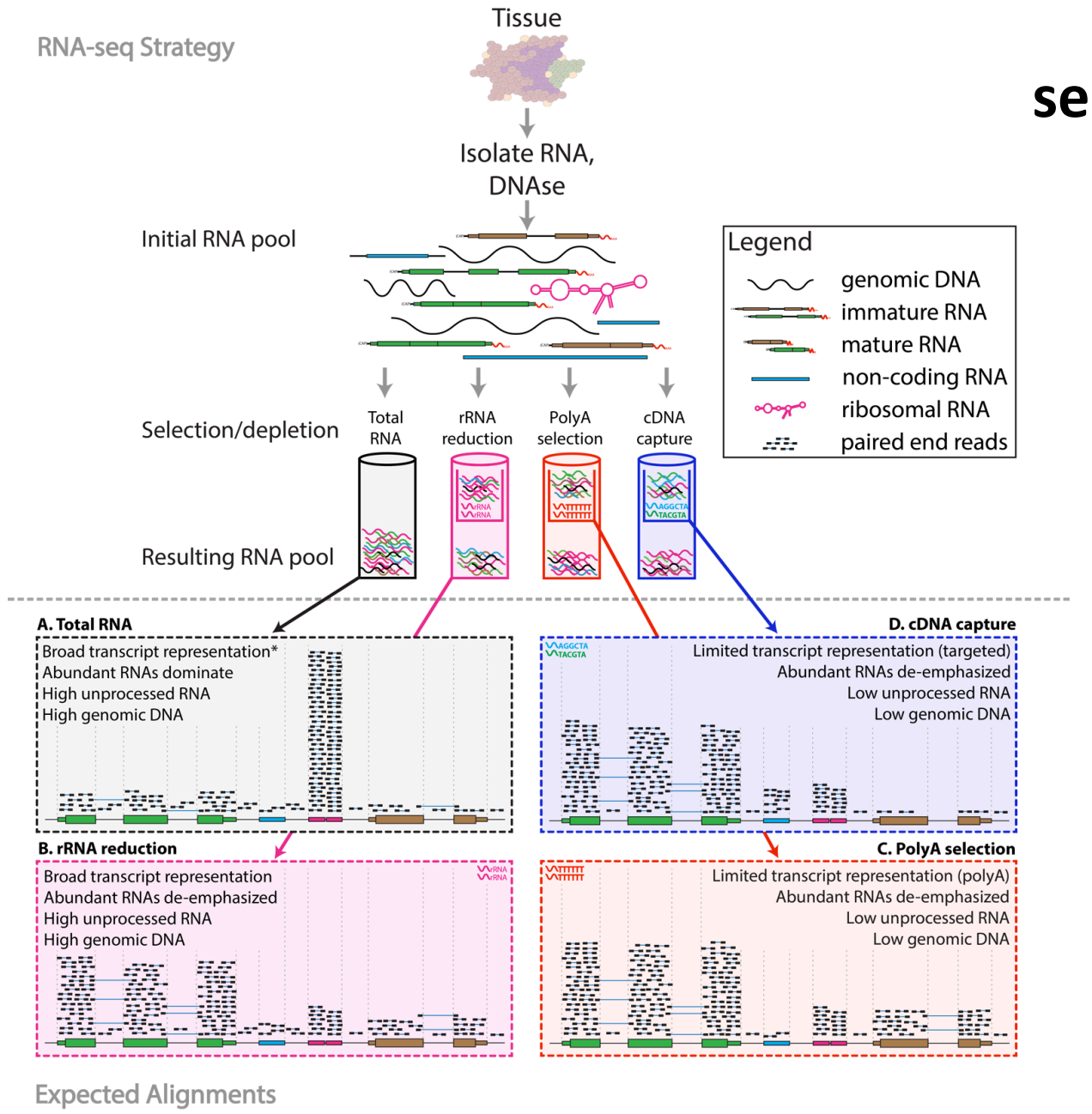
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library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

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

Fragmentation and size selection

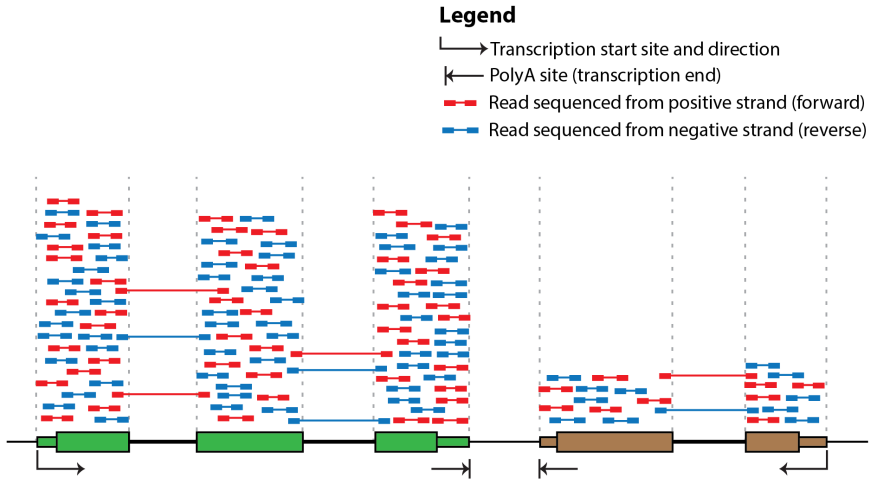


RNA sequence selection/depletion

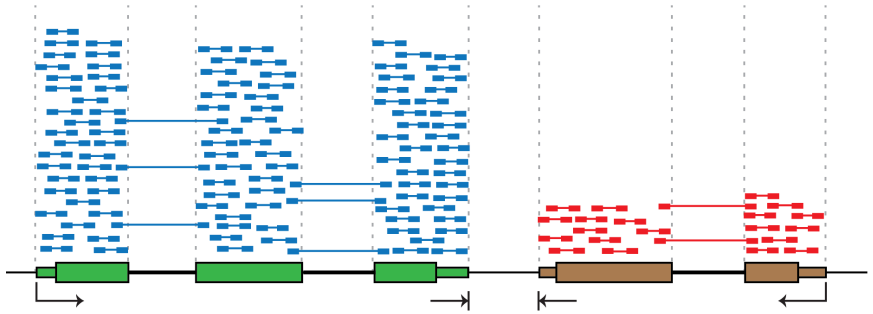


Stranded vs. unstranded

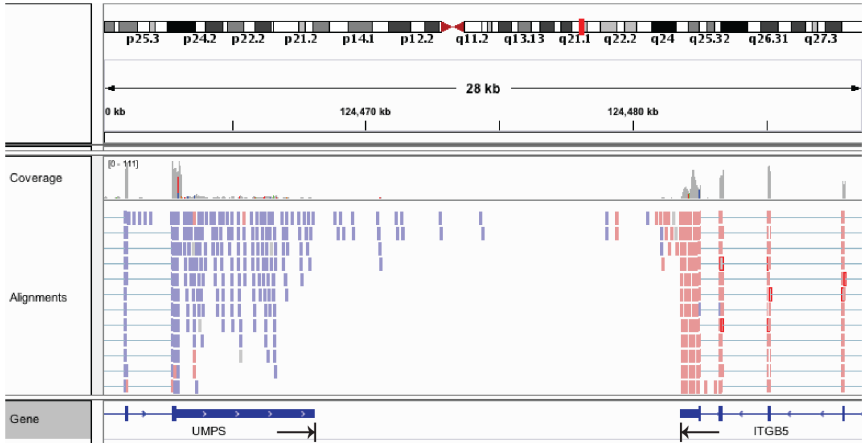
A. Depiction of cDNA fragments from an unstranded library



B. Depiction of cDNA fragments from an stranded library

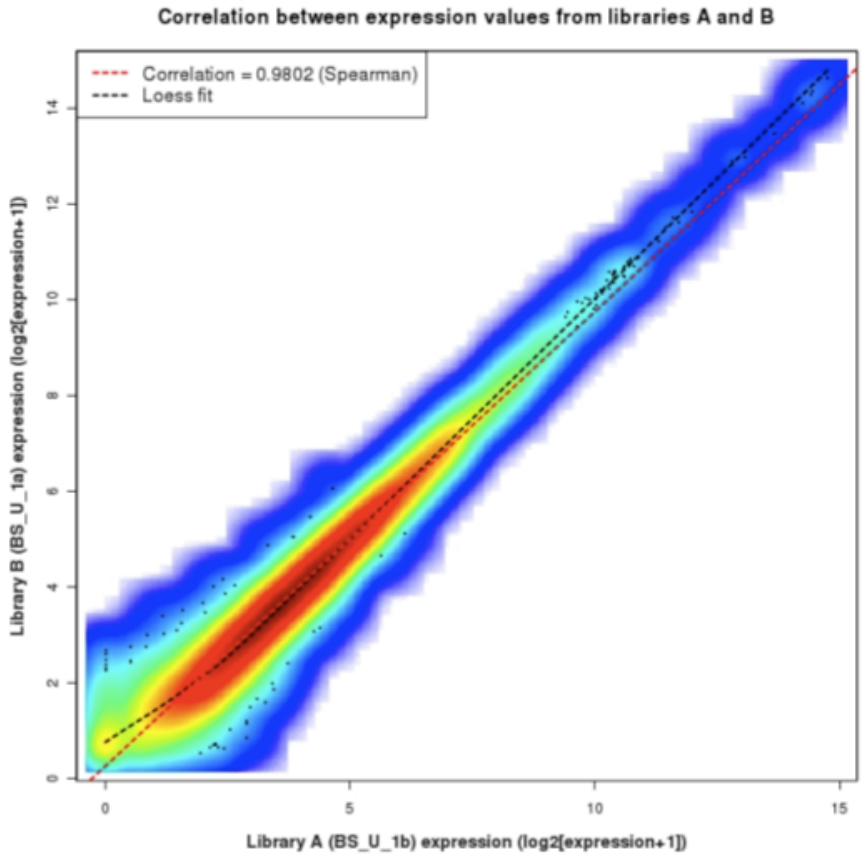


C. Viewing strand of aligned reads in IGV



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?)

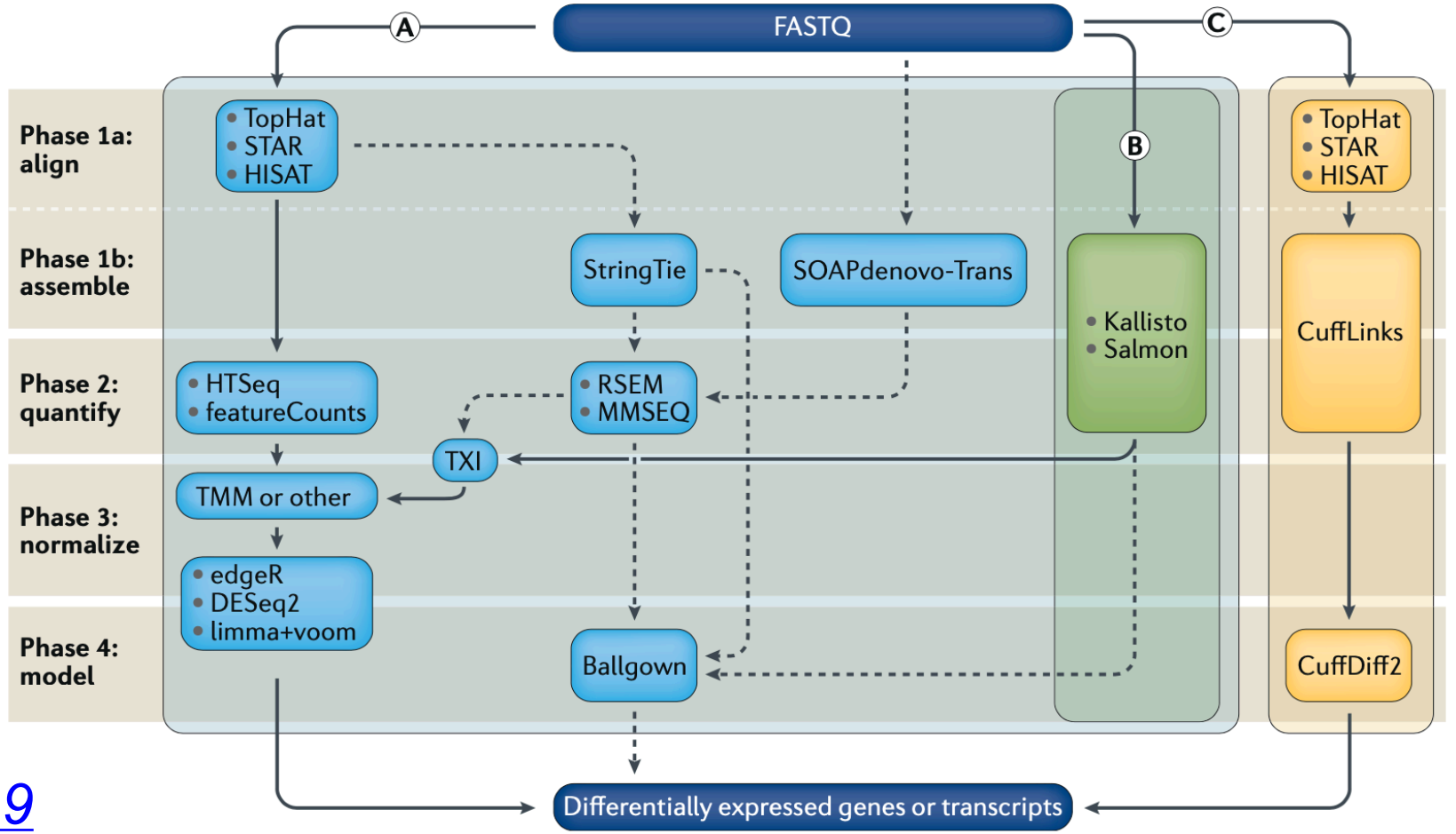
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

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

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

Alignment is central to most genomic research

FASTQ format

Align to genome

Quality control

Find peaks

Quantify transcripts

Find SVs

.
.
.

etc.

Analysis, Interpretation, Discovery

Phase 0

Phase1

Phase 2-100

Alignment - How does it work?

- Alignment is about fitting individual pieces (reads) into the correct part of the puzzle
- The human genome project gave us the picture on the box cover (the reference genome)
- Imperfections in how the pieces fit can indicate changes to a copy of the picture

Reference: AGCCTGAGACCGTAAAAAAGTCAAG

| | | | | | | | | | | | | | | | | | | | | | | |

A read sequence: GAGACCGTAAAAACGTC

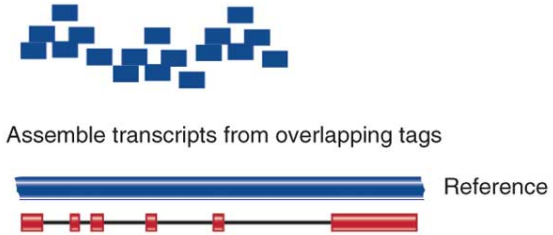
A variant!

RNA-seq alignment challenges

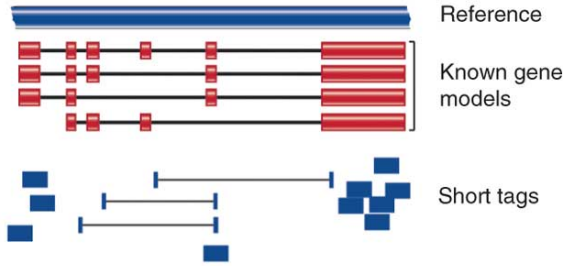
- Computational cost
  - 100's of millions of reads
- Introns!
  - Align to a transcriptome or align to a genome?
    - Spliced vs. unspliced alignments
- Can I just align my data once using one approach and be done with it?
  - Unfortunately, probably not

Three RNA-seq mapping strategies

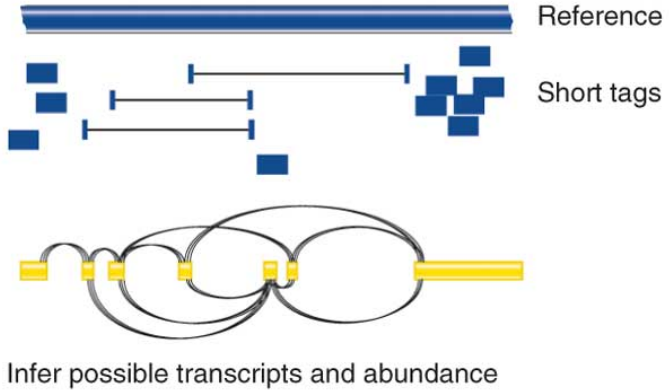
De novo assembly



Align to transcriptome



Align to reference genome



Diagrams from Cloonan & Grimmond, Nature Methods 2010

Which alignment strategy is best?

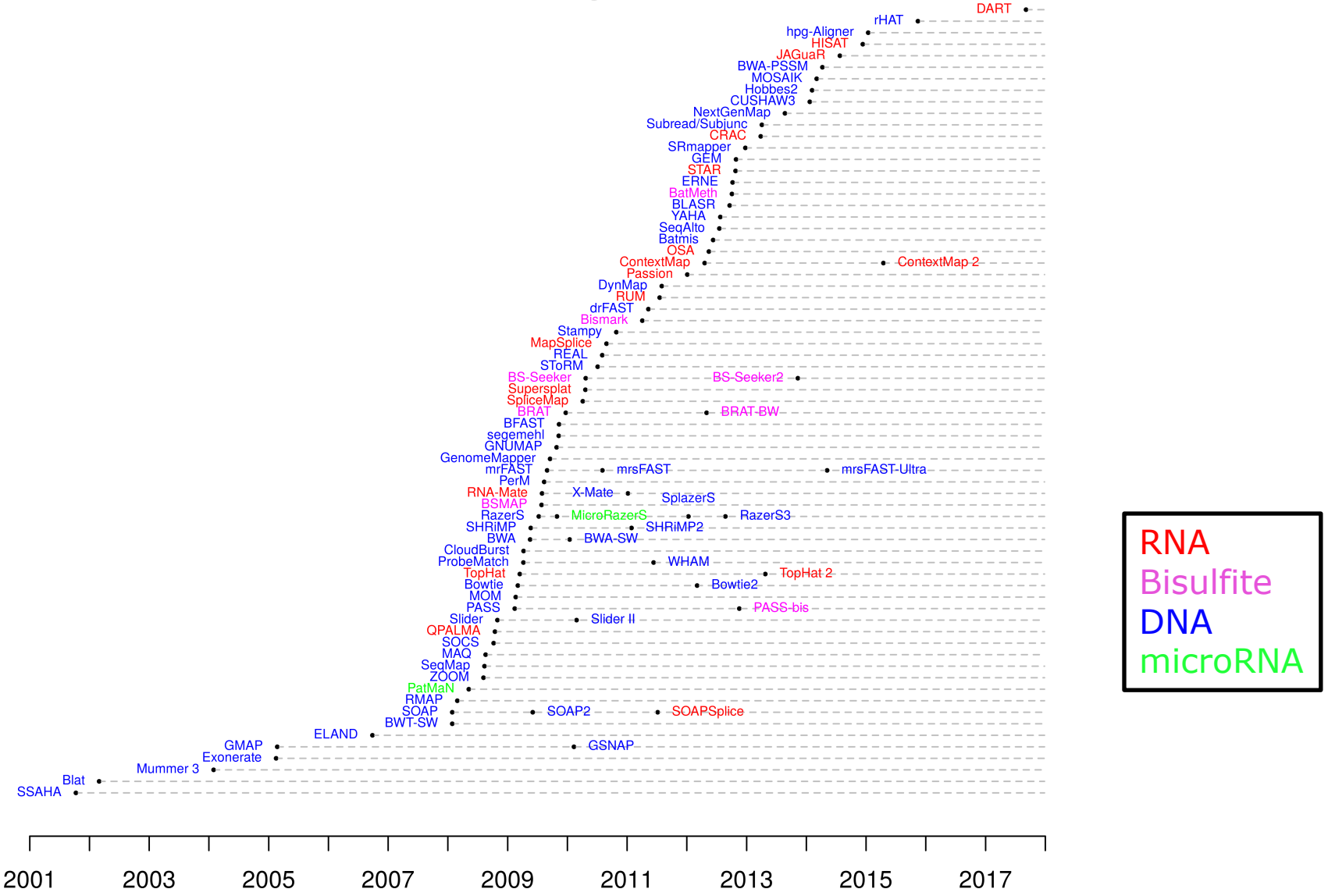
- De novo assembly
  - If a reference genome does not exist for the species being studied
  - If complex polymorphisms/mutations/haplotypes might be missed by comparing to the reference genome

- Align to transcriptome
  - If you have short reads (< 50bp)
  - Relies on known transcripts

- Align to reference genome
  - All other cases
  - Does not rely on known transcripts – allows for discovery

- Each strategy involves different alignment/assembly tools

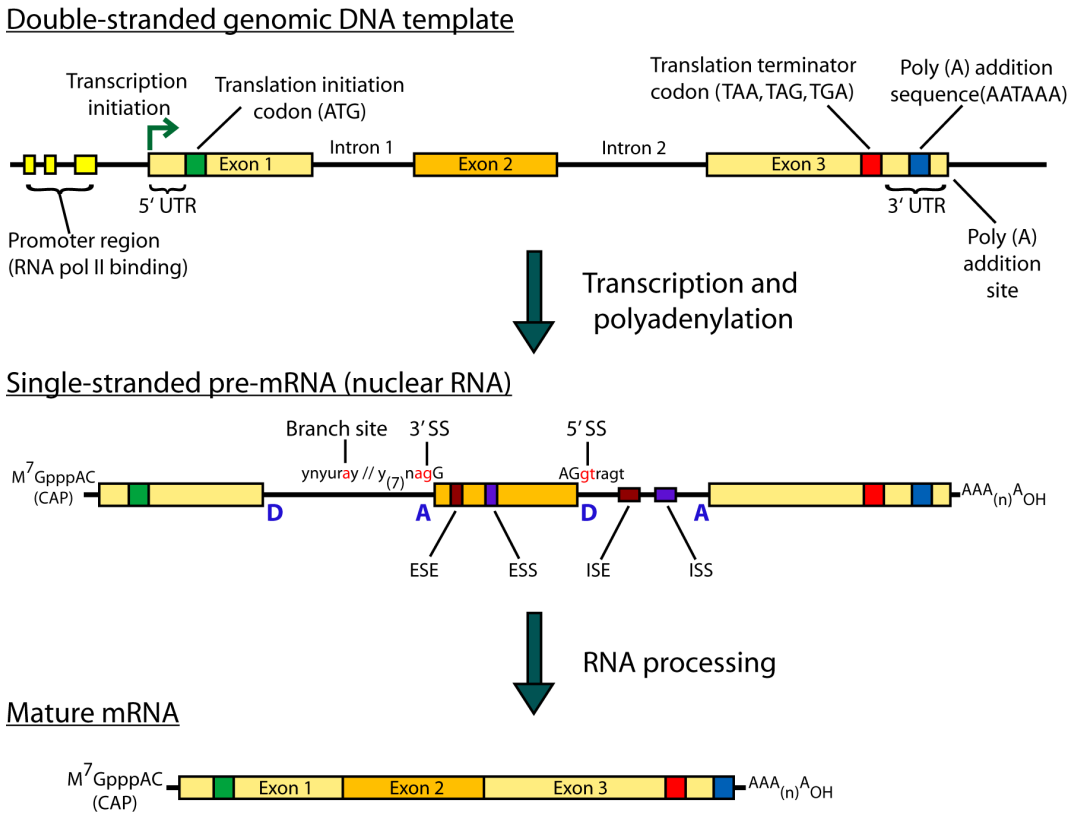
Which read aligner should I use?



RNA
Bisulfite
DNA
microRNA

Should I use a splice-aware or unspliced mapper?

- The fragments being sequenced in RNA-seq represent mRNA - introns are removed
- But we are usually aligning these reads back to the reference genome
- Unless your reads are short (<50bp) you should use a splice-aware aligner
  - HISAT2, STAR, MapSplice, etc.



HISAT/HISAT2

- HISAT is a ‘splice-aware’ RNA-seq read aligner
  - HISAT = **H**ierarchical **I**ndexing for **S**pliced **A**lignments of **T**ranscripts
- Requires a reference genome
- Very fast

- Uses an indexing scheme based on the Burrows-Wheeler transform and the Ferragina-Manzini (FM) index
- Multiple types of indexes for alignment
  - a whole-genome FM index to anchor each alignment
  - numerous local FM indexes for very rapid extensions of these alignments.
  - Whole-genome indices with SNPs and known transcript structures accounted for

Kim et al. 2015. Nat Methods 12:357–360

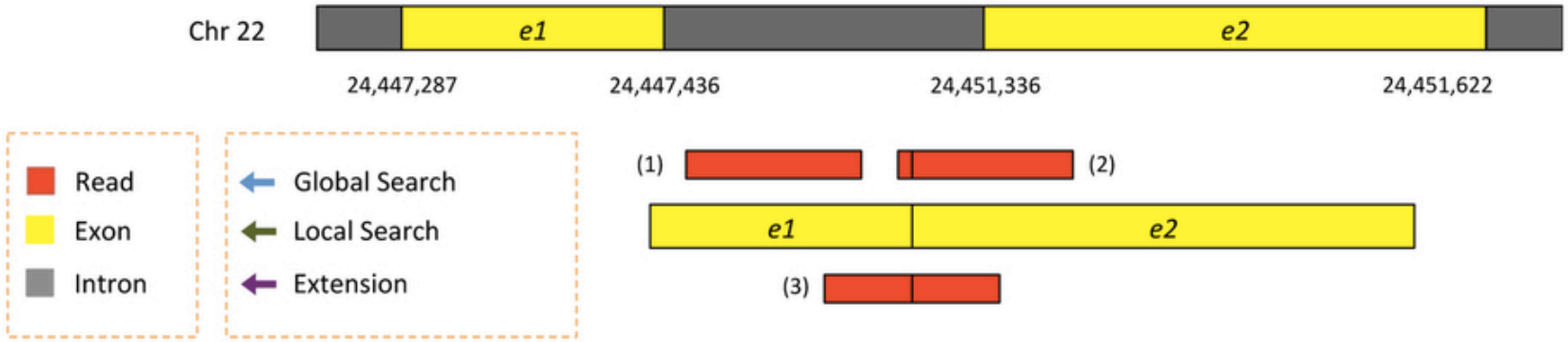
HISAT/HISAT2 algorithm

- Uses a hierarchical indexing algorithm + several adaptive strategies
  - based on the position of a read with respect to splice sites

1) Find candidate locations across the whole genome first
   - mapping part of each read using the global FM index
   - Generally identifies one or a small number of candidates.
2) Do local alignment
   - selects one of ~48,000 local indexes for each candidate
   - uses it to align the remainder of the read.

- For paired reads, each mate is separately aligned
  - If a read fails to align, then the alignments of its mate are used as anchors to map the unaligned mate

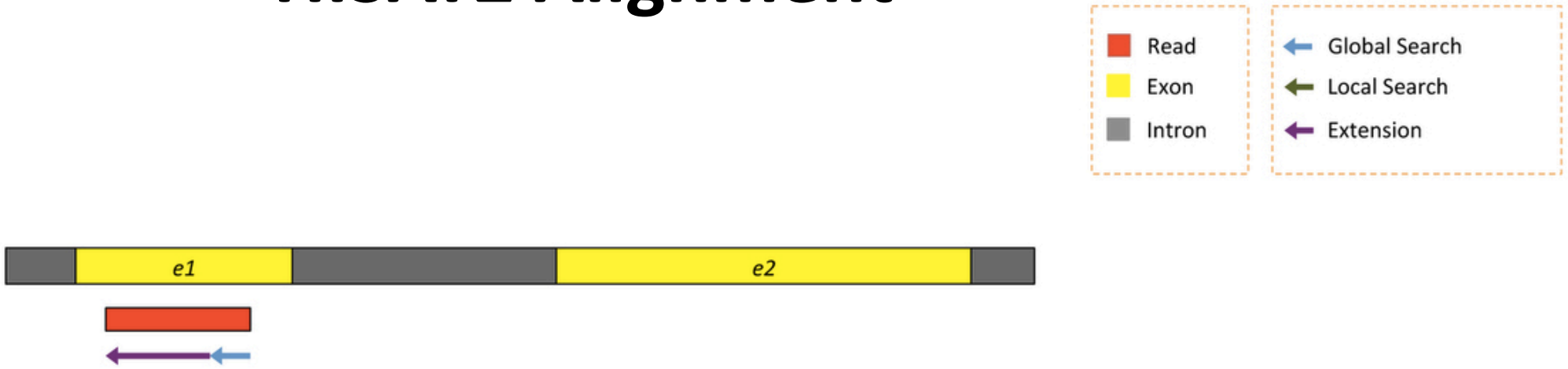
HISAT2 Alignment



- Two exons from chr22
- Three reads

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment

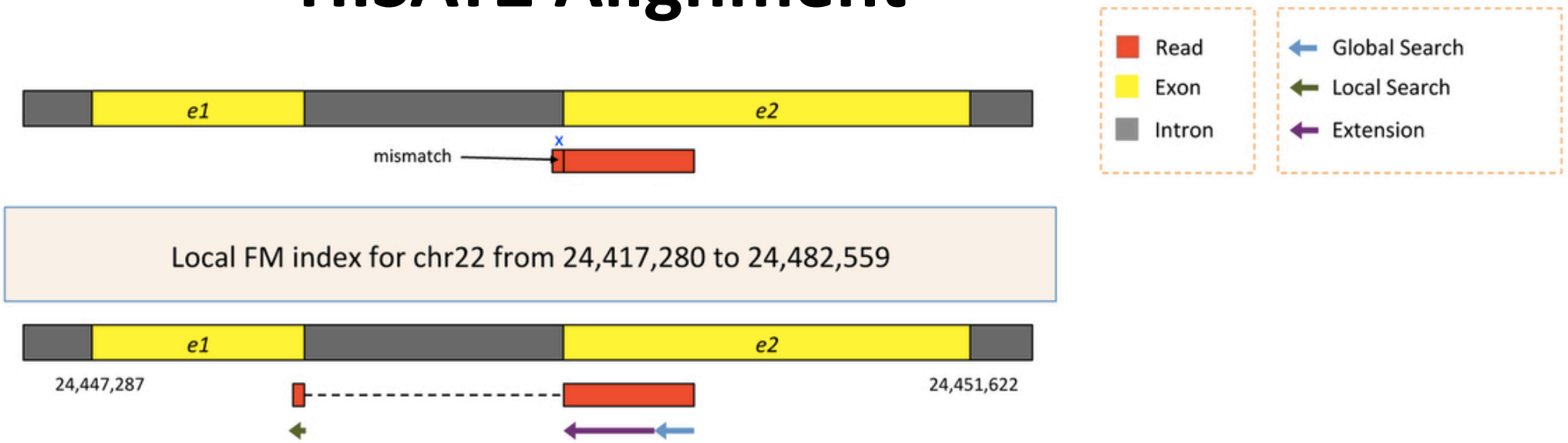


1) Search for read position with global FM index (slower)

2) Once at least 28bp and exactly one location switch to extension mode against reference genome (faster)

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment



1) Search for read position with global FM index (slower)

2) Extend until mismatch at 93bp (faster)

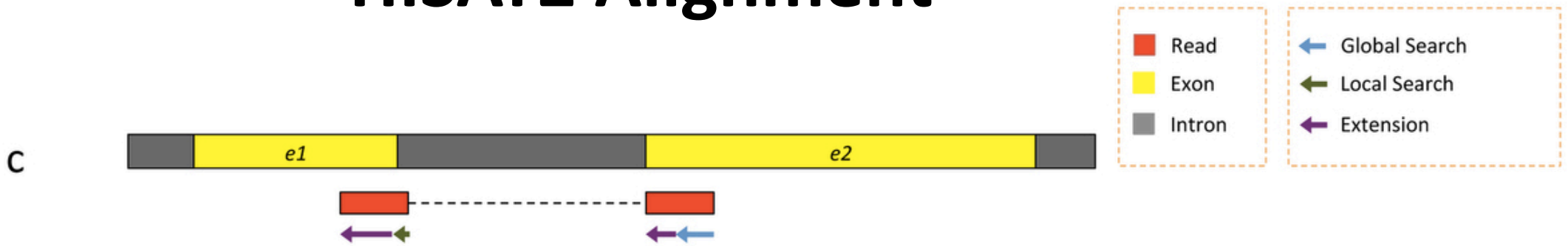
3) Switch to local FM index to align remaining 8bp

- index covers only a small region, so we find just one match

4) Check for compatibility and combine into single spliced alignment

Kim et al. 2015. Nat Methods 12:357–360

HISAT2 Alignment



1) global search until exactly one match of at least 28bp (slower)

2) Extend until mismatch at 51bp (faster)

3) switch to local FM index to align first 8bp of remaining read

    - If too many matches increase prefix size

4) Extend again

5) Check for compatibility and combine into single spliced alignment

Kim et al. 2015. Nat Methods 12:357–360

Should I allow ‘multi-mapped’ reads?

- Depends on the application
- In *DNA* analysis it is common to use a mapper to randomly select alignments from a series of equally good alignments
- In *RNA* analysis this is less common
  - Perhaps disallow multi-mapped reads if you are variant calling
  - Definitely should allow multi-mapped reads for expression analysis with Cufflinks (and StringTie?)
  - Definitely should allow multi-mapped reads for gene fusion discovery

What is the output of HISAT2?

- A SAM/BAM file
  - SAM stands for Sequence Alignment/Map format
  - BAM is the binary version of a SAM file

- Remember, compressed files require special handling compared to plain text files

- How can I convert BAM to SAM?
  - http://www.biostars.org/p/1701/

- Is HISAT2 the only mapper to consider for RNA-seq data?
  - http://www.biostars.org/p/60478/