The Elizabeth H. and James S. McDonnell III
McDonnell Genome Institute
at Washington University

Washington University in St. Louis
School of Medicine

# PMBIO Module 06

## **RNAseq.** RNA-sequence analysis

Malachi Griffith, Obi Griffith, Zachary Skidmore, Huiming Xia
Introduction to bioinformatics for DNA and RNA sequence analysis (IBDR01)
29 October – 2 November, 2018
Glasgow

# Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)

This is a human-readable summary of (and not a substitute for) the license. Disclaimer.

## You are free to:


**Share** — copy and redistribute the material in any medium or format

**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.

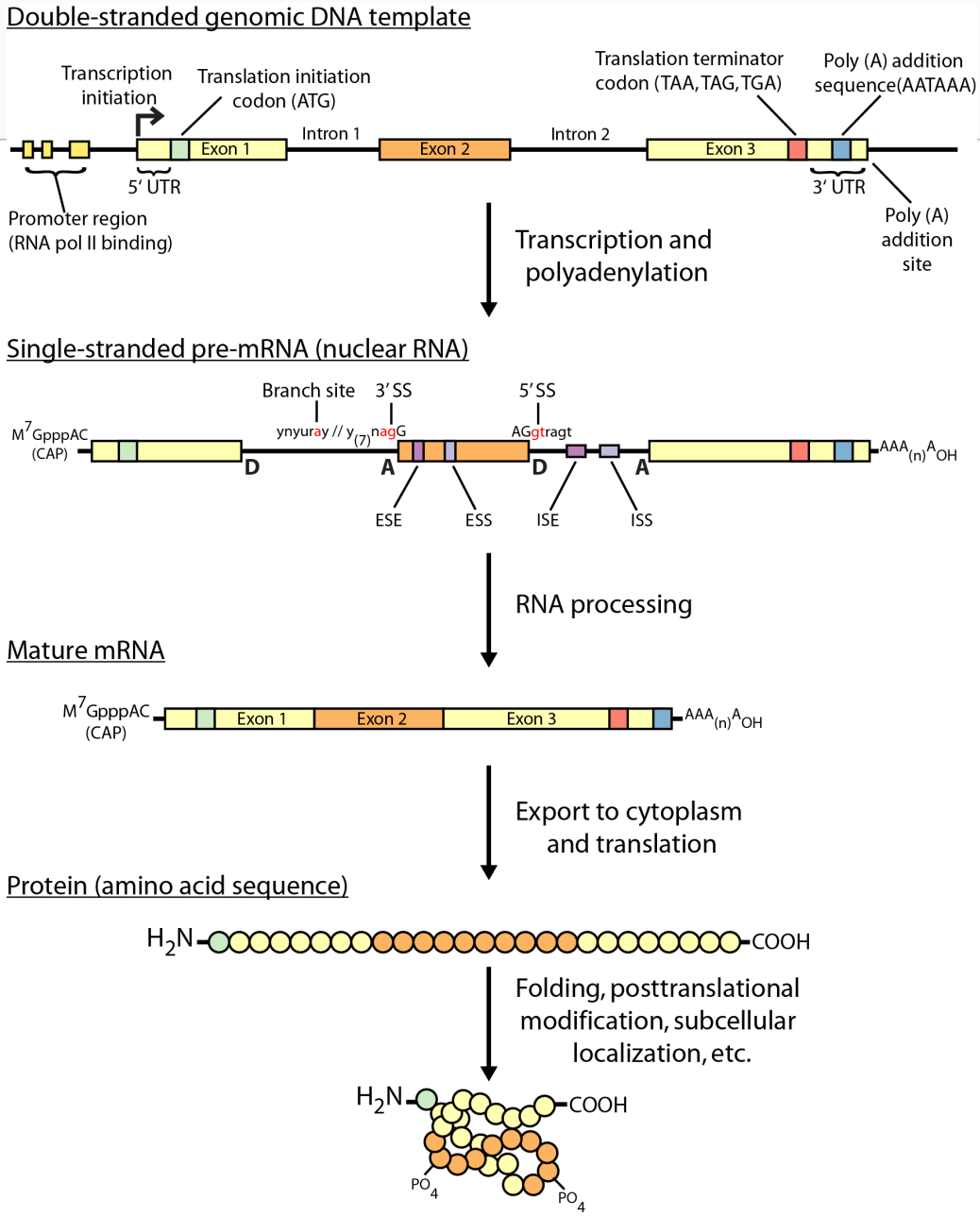
**No additional restrictions** — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.

# Learning objectives of module 06: RNAseq

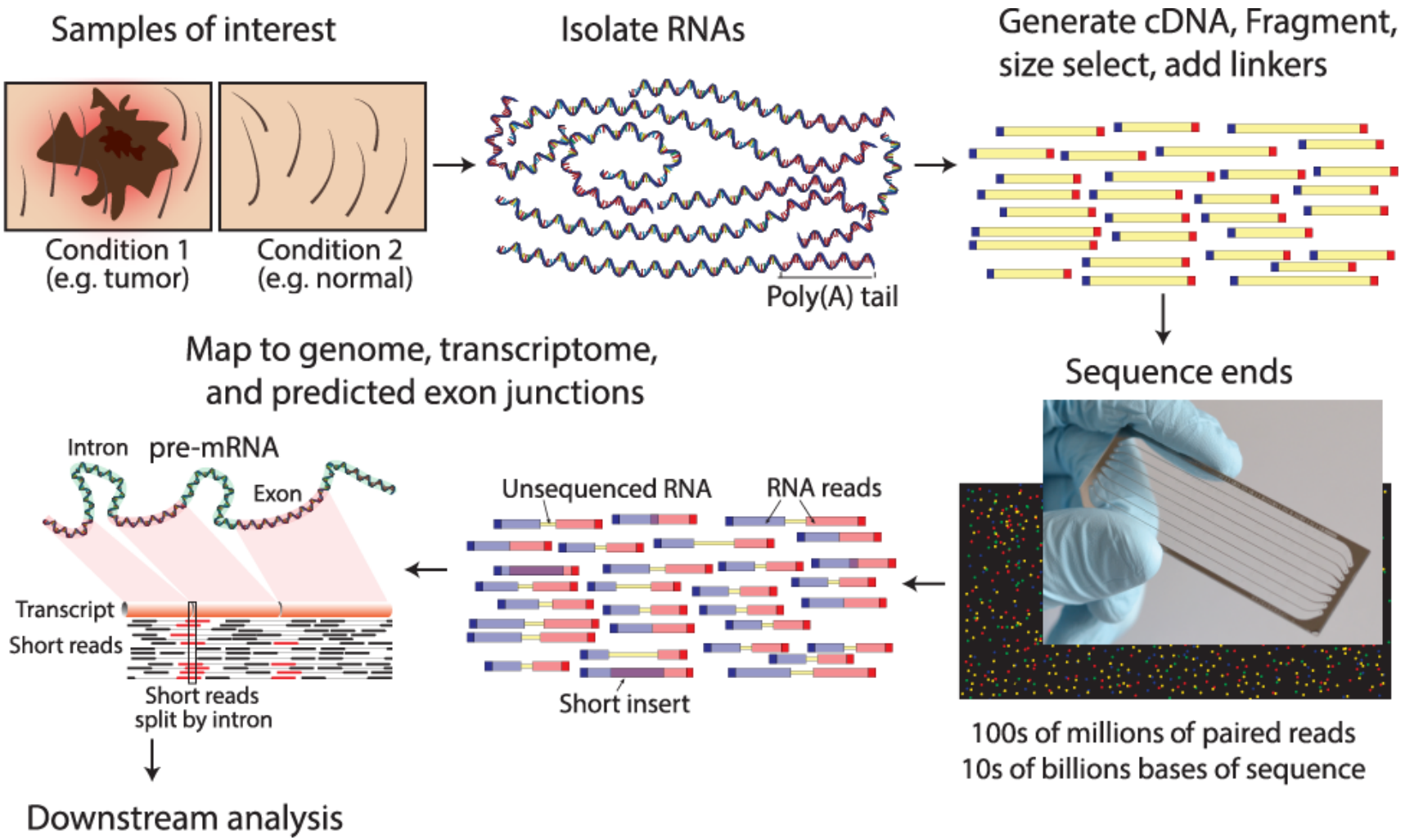
- **Key concepts**: RNA-seq, library strategies, RNA splicing, genes and transcripts, transcription strand, gene/transcript abundance estimation, FPKM and TPM, differential expression (DE) analysis, normalization, batch effects

- Perform alignment of RNA-seq data and a basic QC analysis of the resulting alignments
- Obtain gene and transcript abundance estimates
- Perform a reference-free alignment and abundance estimation and contrast the results with the reference based approach
- Basic differential expression analysis
- Perform an expression outlier analysis
- Run an RNA fusion detection tool and parse the results for interesting candidates
- Assess expression of specific gene markers of relevance (e.g. HER2, ER, PR)
- Assess the expression of specific variants identified in previous sections.
- Identify an example of allele specific expression

# Gene expression



# RNA sequencing

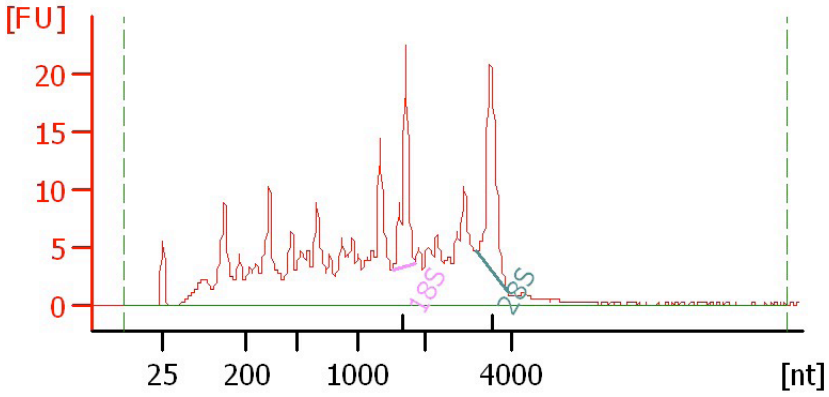


# Challenges

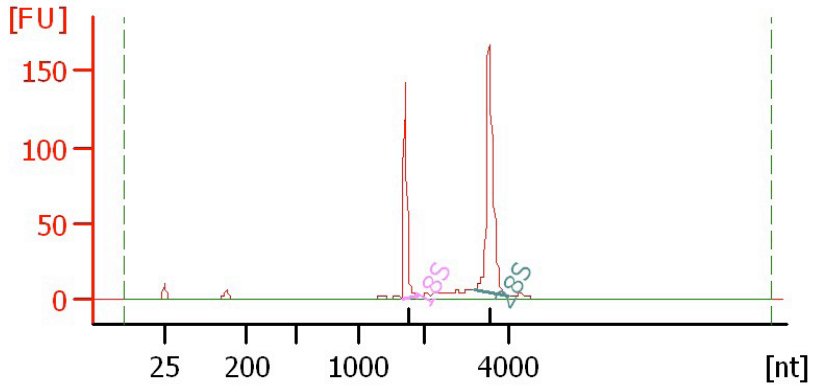
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - $10^5$ – $10^7$ orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction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



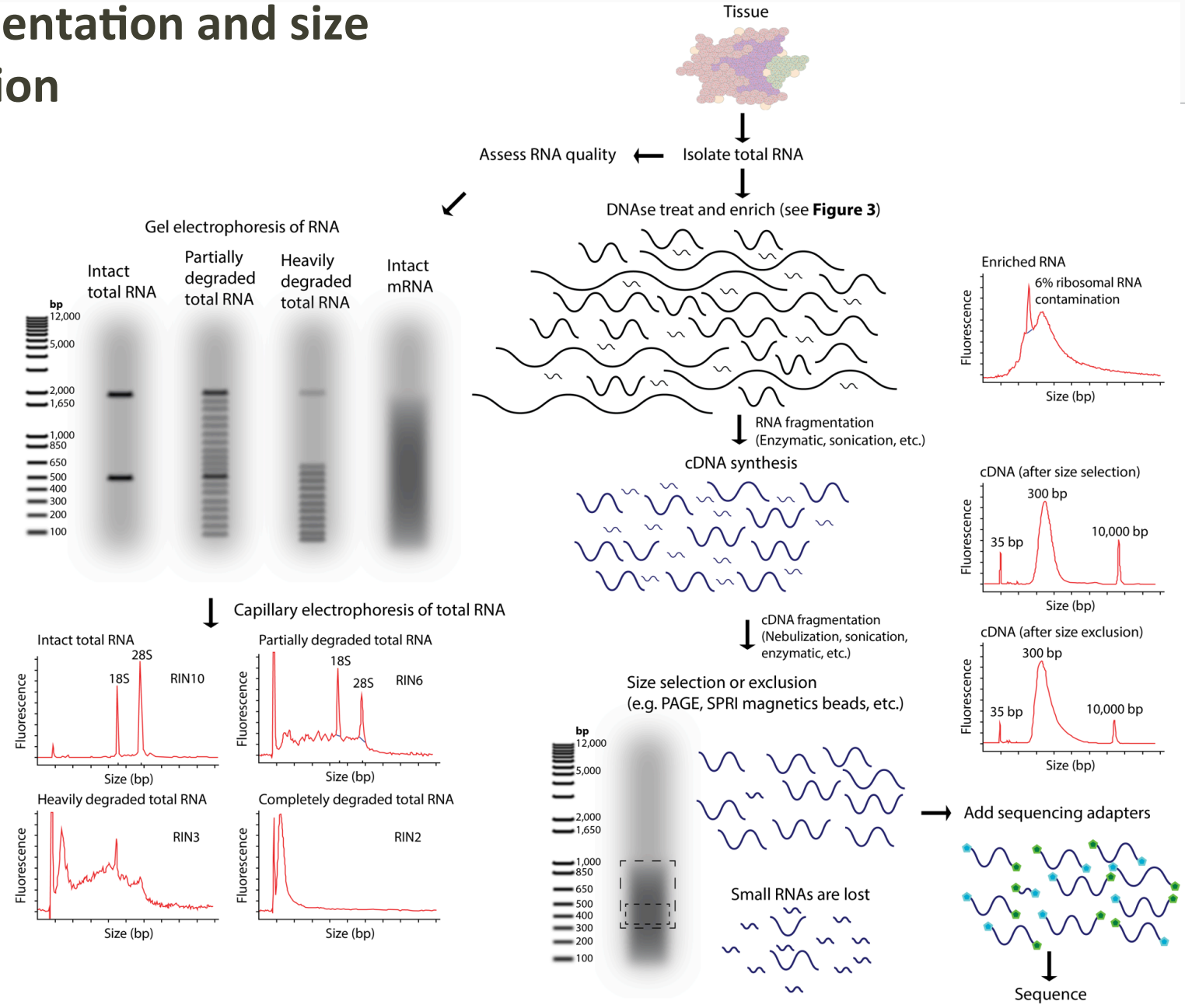

RIN = 10

# There are many RNA-seq library construction strategies

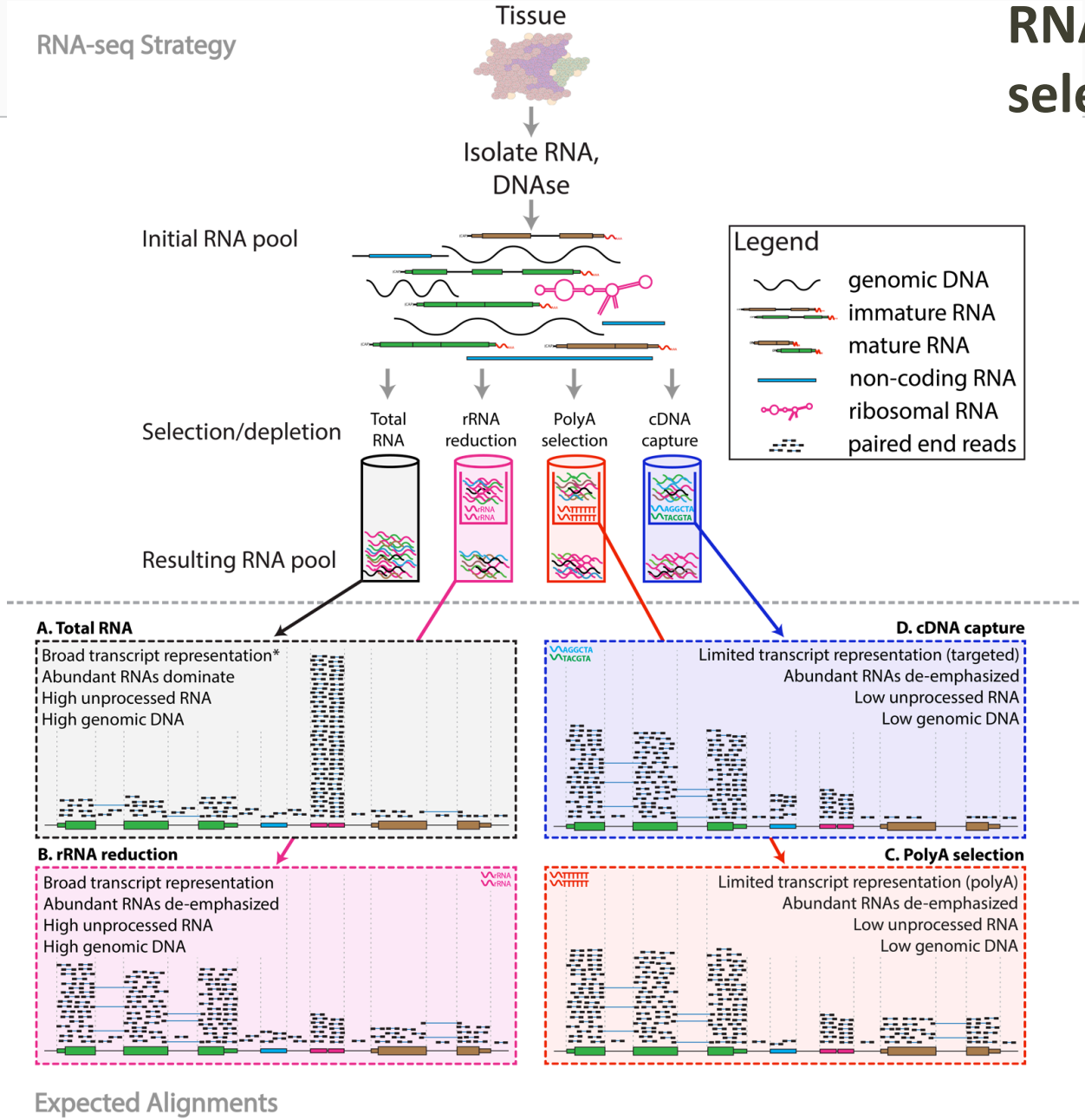
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library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

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

# Fragmentation and size selection

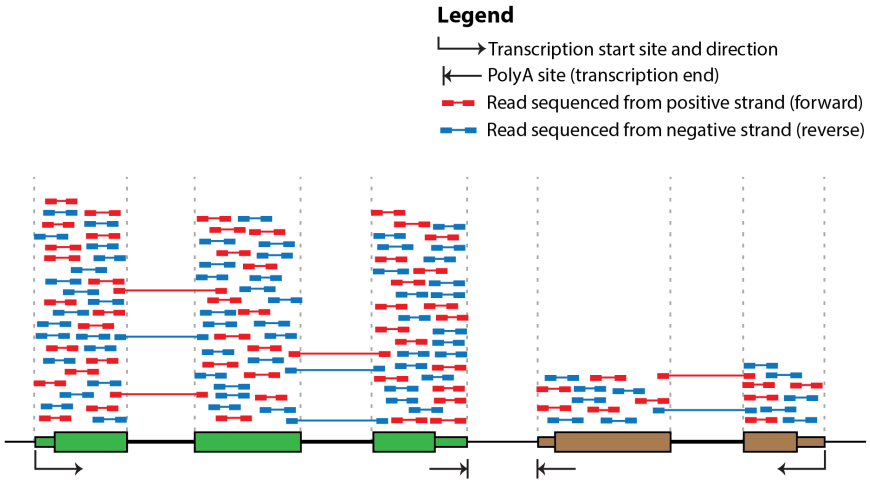


# RNA sequence selection/depletion

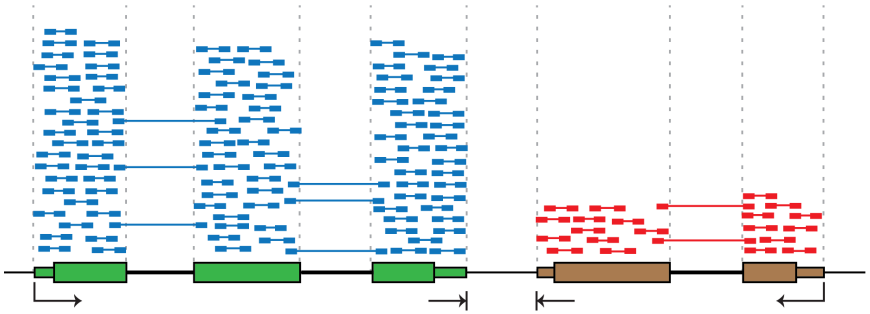


# Stranded vs. unstranded

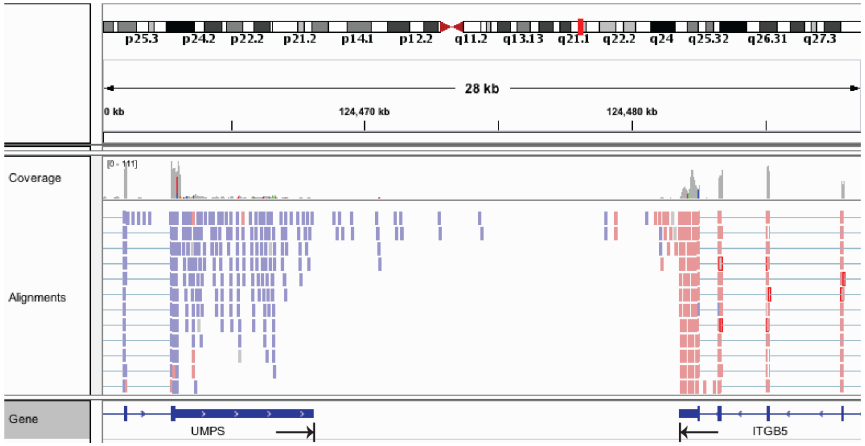
## A. Depiction of cDNA fragments from an unstranded library




## B. Depiction of cDNA fragments from an stranded library

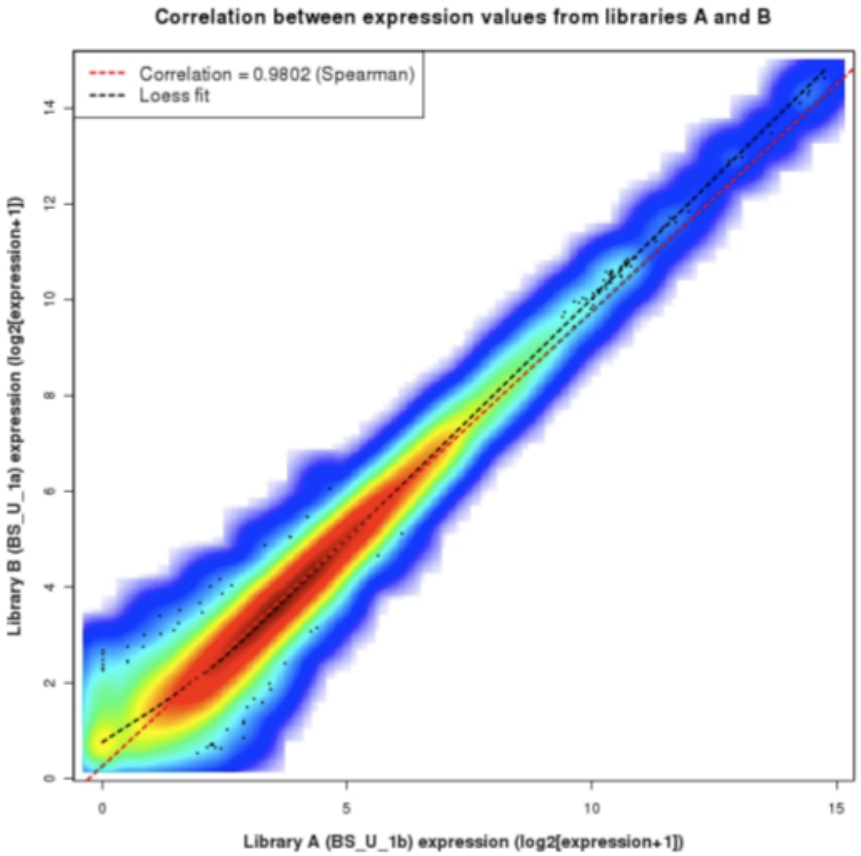


## C. Viewing strand of aligned reads in IGV



# 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

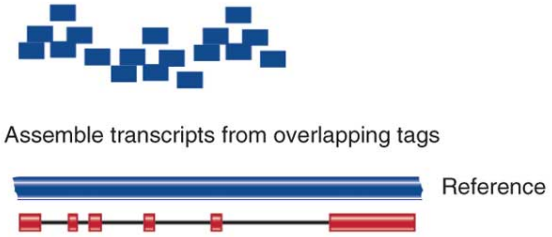


# Common analysis goals of RNA-Seq  analysis (what can you ask of the data?)

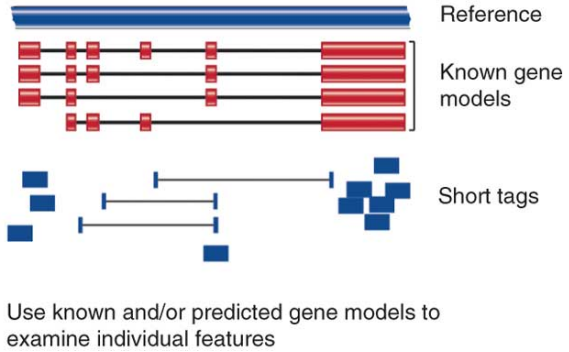
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Allele specific expression
  - Relating to SNPs or mutations
- Mutation discovery
- Fusion detection
- RNA editing

# Three RNA-seq mapping strategies

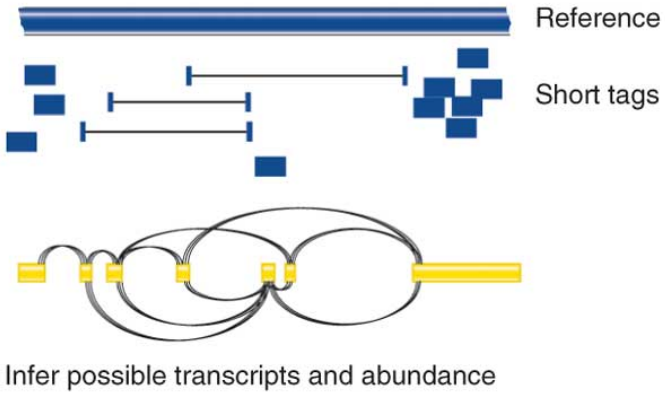
## De novo assembly




## Align to transcriptome

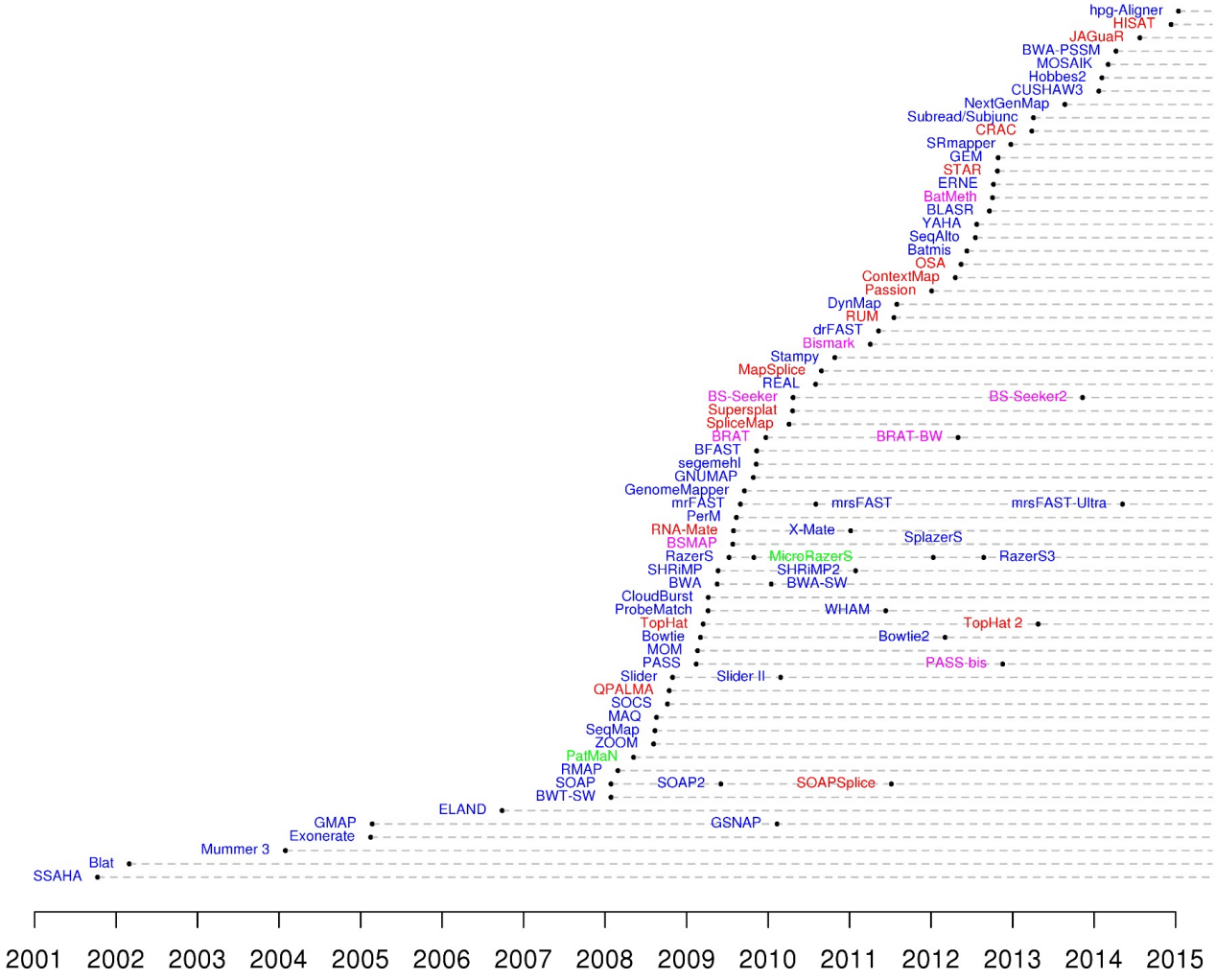



## Align to reference genome




Diagrams from Cloonan & Grimmond, Nature Methods 2010

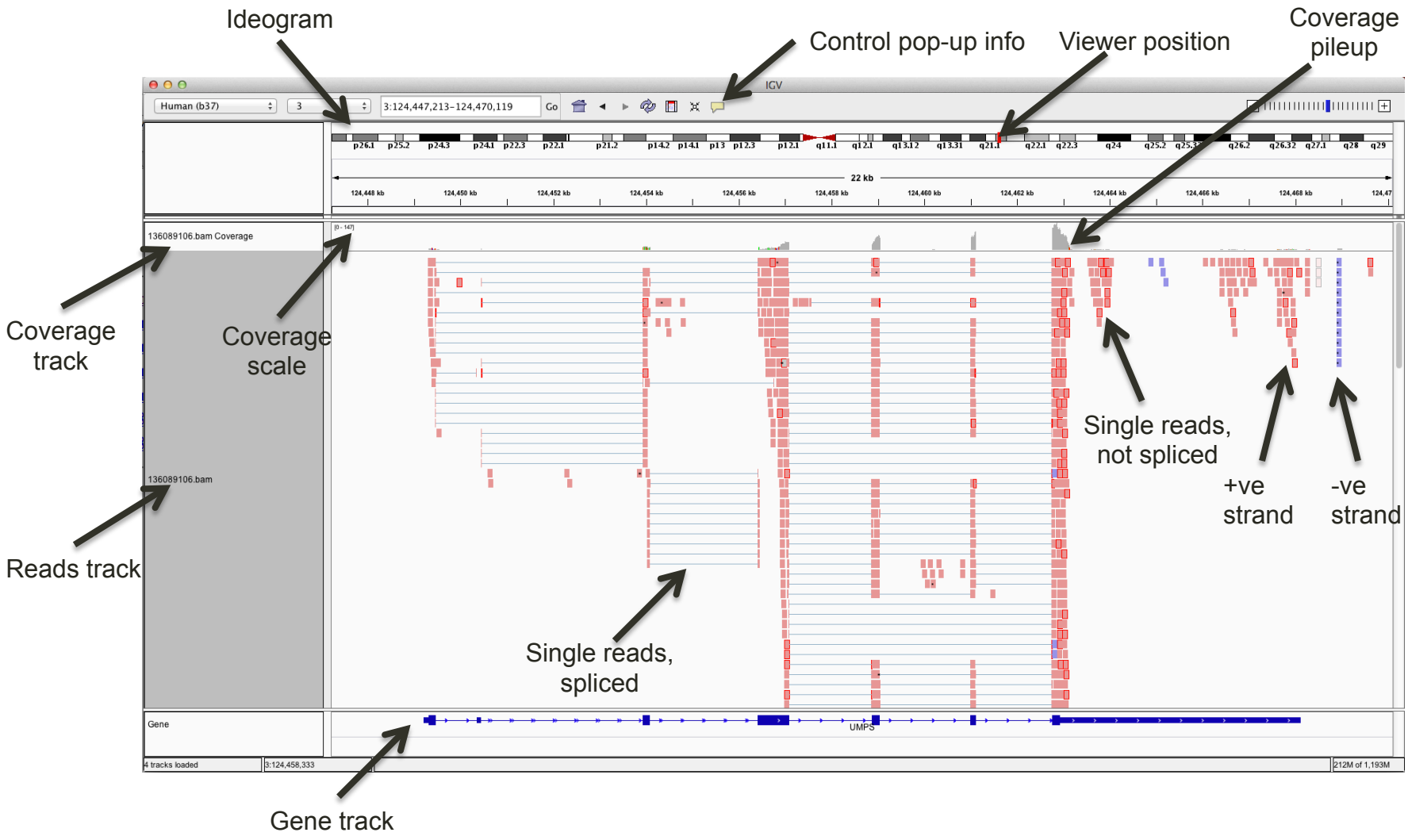
# Which read aligner should I use?



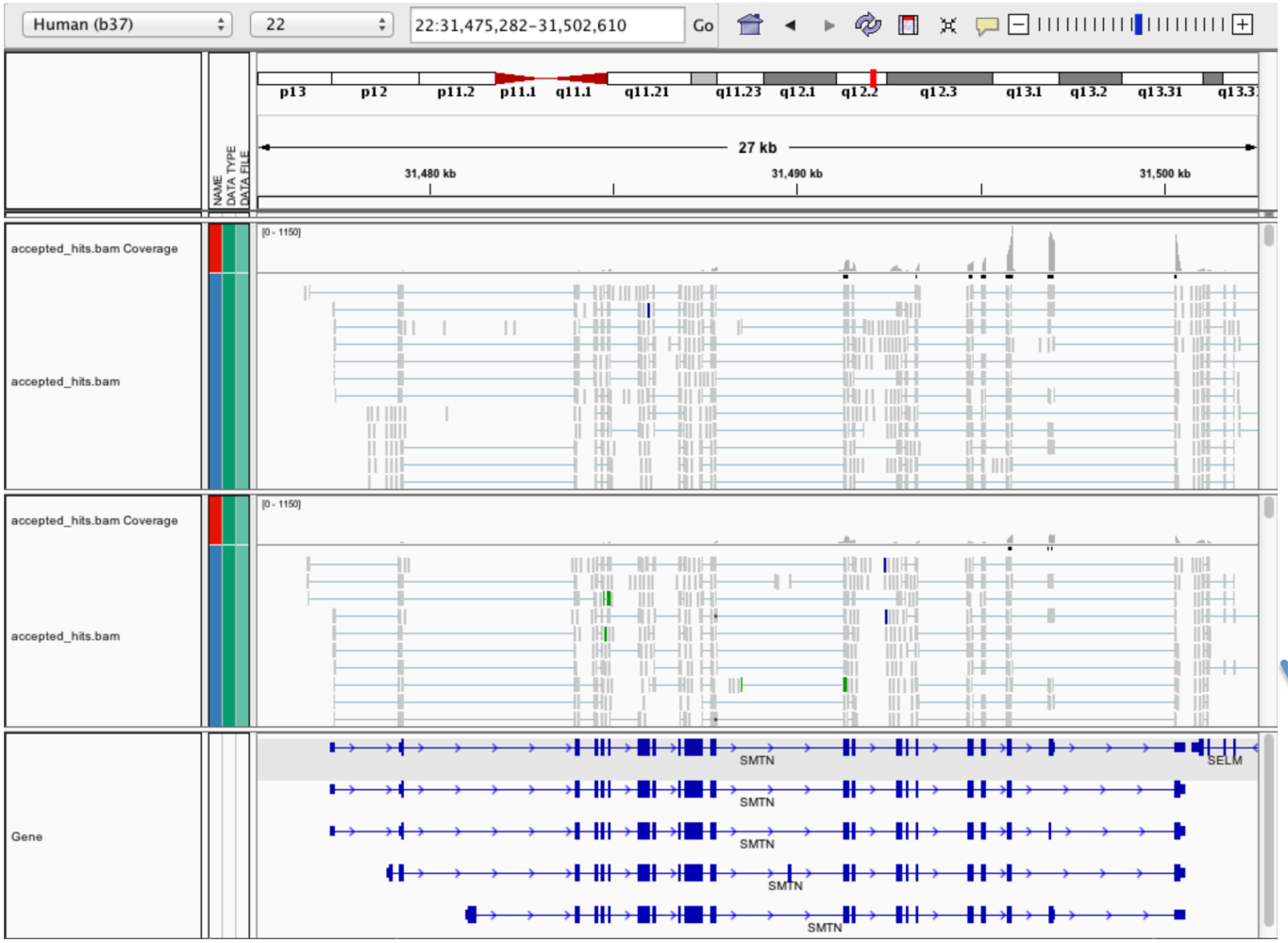

RNA
Bisulfite
DNA
microRNA

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

# Visualization of RNA-seq alignments in IGV browser



# Expression estimation for known genes and transcripts




3' bias

Down-regulated

# What is FPKM (RPKM)

- RPKM: Reads Per Kilobase of transcript per Million mapped reads.
- FPKM: Fragments Per Kilobase of transcript per Million mapped reads.
- In RNA-Seq, the relative expression of a transcript is proportional to the number of cDNA fragments that originate from it. However:
  - The number of fragments is also biased towards larger genes
  - The total number of fragments is related to total library depth
- FPKM (RPKM) attempt to normalize for gene size and library depth

- FPKM (RPKM) = (10^9 * C) / (N * L)
  - C = number of mappable reads/fragments for a gene/transcript/exon/etc
  - N = total number of mappable reads/fragments in the library
  - L = number of base pairs in the gene/transcript/exon/etc

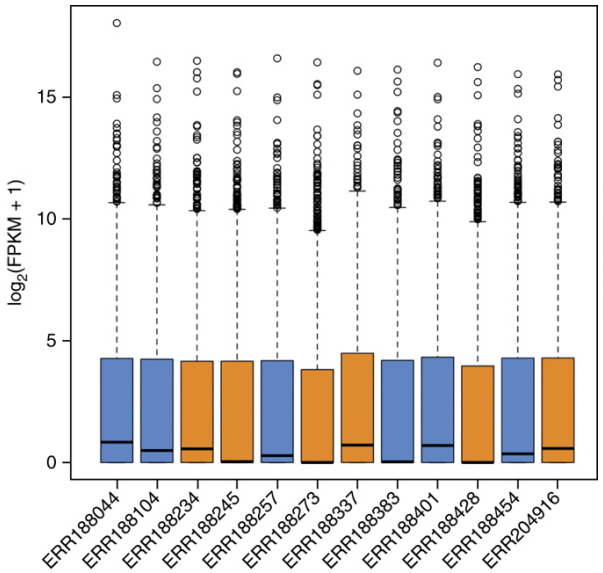
- http://www.biostars.org/p/11378/
- http://www.biostars.org/p/68126/

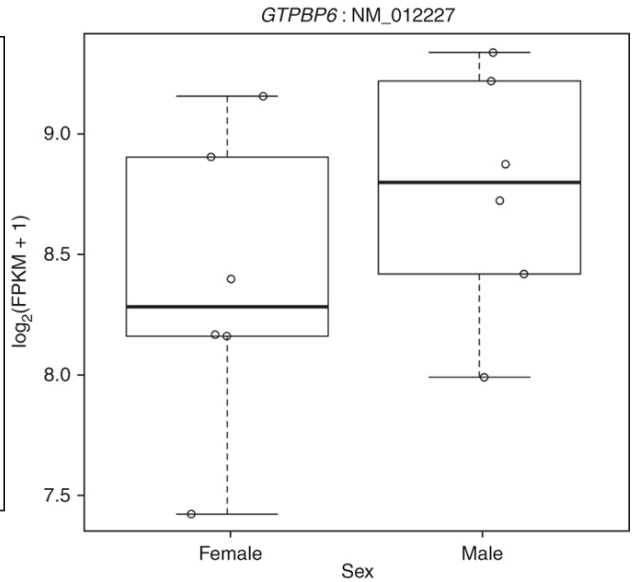
# How do FPKM and TPM differ?

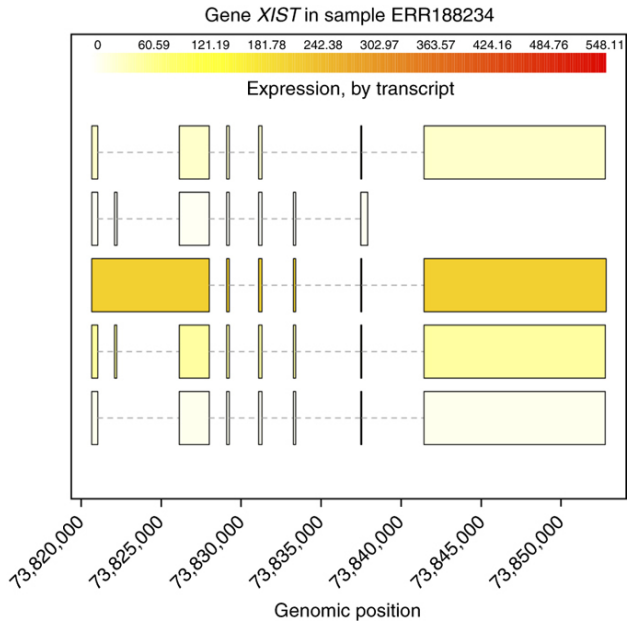
- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:
  - FPKM
    - 1) Determine total library/sample fragment count and divide by 1,000,000
      - "per million" scaling factor
    - 2) Divide each gene/transcript fragment count by #1
      - fragments per million, FPM
    - 3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)
  - TPM
    - 1) Divide each gene/transcript fragment count by length of each gene/transcript in kilobases
      - fragments per kilobase, FPK
    - 2) Sum all FPK values for the sample and divide by 1,000,000
      - "per million" scaling factor
    - 3) Divide #1 by #2 (TPM)

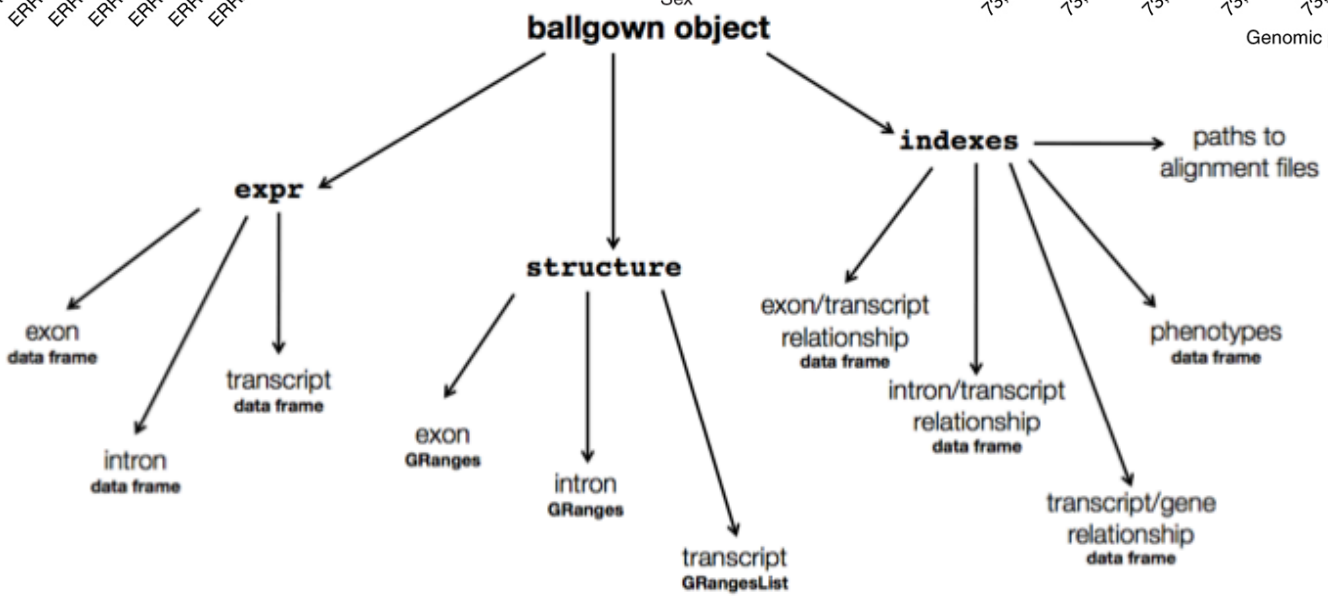
- The sum of all TPMs in each sample is the same. Easier to compare across samples!
- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

# Ballgown for Visualization with R



# Learning objectives of module 4

- Alignment free estimation of transcript abundance

- Introduction to k-mers
- Alignment free tools
  - Sailfish, RNA-Skim, Kallisto, Salmon
- Abundance estimation and differential expression analysis with Kallisto and Sleuth

# What is a k-mer?

- A fixed sized (***K***) sequence

- A string of length ***N*** contains ***N-K+1*** k-mers

1-mer

| |
|---|
| A |
| C |
| G |
| T |

2-mer

| | | | |
|---|---|---|---|
| AA | AC | AG | AT |
| CA | CC | CG | CT |
| GA | GC | GG | GT |
| TA | TC | TG | TT |

ATTCGACAGTAGCCATGACTGG

...

- One can build *K*-mer index to represent a string

| 7-mer | iD | N |
|---|---|---|
| ATTCGAC | 1 | 1 |
| TTCGACA | 2 | 1 |
| TCGACAG | 3 | 1 |
| ... | | |

Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms Rob Patro, Stephen M. Mount, and Carl Kingsford. *Manuscript Submitted* (2013) http://www.cs.cmu.edu/~ckingsf/class/02714-f13/Lec05-sailfish.pdf

https://www.slideshare.net/duruofei/cmsc702-project-final-presentation

# Basic concept of alignment free approaches for transcript abundance

1. Obtain reference transcript sequences (e.g. Ensembl, Refseq, or GENCODE)
2. Build a **k-mer index** of all of the k-mers occurring in each transcript sequence
   - Store each k-mer and its position within the transcript. “hashing”
3. Parse all RNA-seq reads and count how many times each k-mer occurs within each read
   - Model relationship between RNA-seq read k-mers and the transcript k-mer index.
   - What transcript is the most likely source for each read?
   - Called “pseudoalignment” , “quasi-mapping”, etc.
4. Handle sequencing errors, isoforms, ambiguity, and determine abundance estimates
   - Transcriptome de Bruijn graphs, likelihood function, expectation maximization, etc.

# Advantages/disadvantages of alignment free approaches

- Advantages
  - Very fast and efficient
    - Similar accuracy to alignment based approach but with much, much shorter run time.
  - Do not need a reference genome, only a reference transcriptome

- Disadvantages
  - You don’t get a proper BAM file
  - Information in reads with sequence errors may be ignored
  - Limited potential for transcript discovery, variant calling, fusion detection, etc.

# Common alignment free tools

- Sailfish
  - “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.” 2014
  - https://www.ncbi.nlm.nih.gov/pubmed/24752080
- RNA-Skim
  - “RNA-Skim: a rapid method for RNA-Seq quantification at transcript level.” 2014
  - https://www.ncbi.nlm.nih.gov/pubmed/24931995
- Kallisto
  - “Near-optimal probabilistic RNA-seq quantification.” 2016
  - https://www.ncbi.nlm.nih.gov/pubmed/27043002
- Salmon
  - “Salmon provides fast and bias-aware quantification of transcript expression.” 2017
  - https://www.ncbi.nlm.nih.gov/pubmed/28263959

# Which is best?

- Somewhat controversial …
- https://liorpachter.wordpress.com/2017/08/02/how-not-to-perform-a-differential-expression-analysis-or-science/

- Various sources suggest that Salmon, Kallisto, and Sailfish results are quite comparable
- Usability, documentation, and supporting downstream tools could be used to decide

# HISAT2/StringTie/Ballgown RNA-seq Pipeline