



Cold  
Spring  
Harbor  
Laboratory

# Advanced Sequencing Technologies & Applications

<http://meetings.cshl.edu/courses.html>



Cold  
Spring  
Harbor  
Laboratory

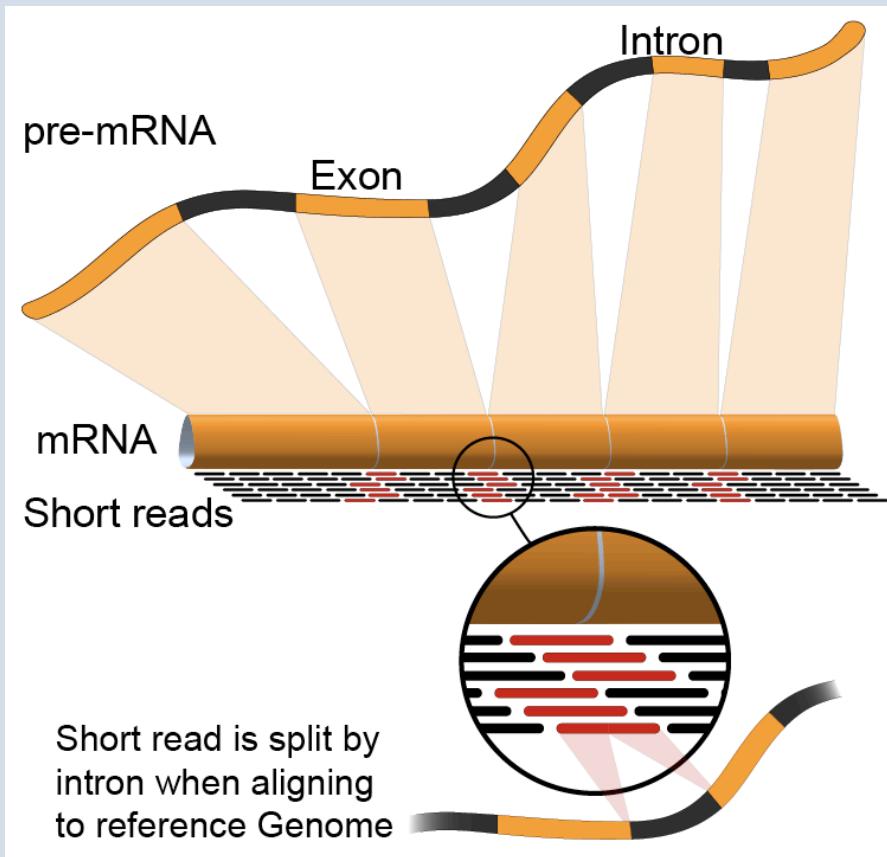
RNA-Seq Module 1

## Introduction to RNA Sequencing (lecture)

Malachi Griffith, Obi Griffith, Jason Walker

Advanced Sequencing Technologies & Applications

November 10 - 22, 2015



# Learning objectives of the course

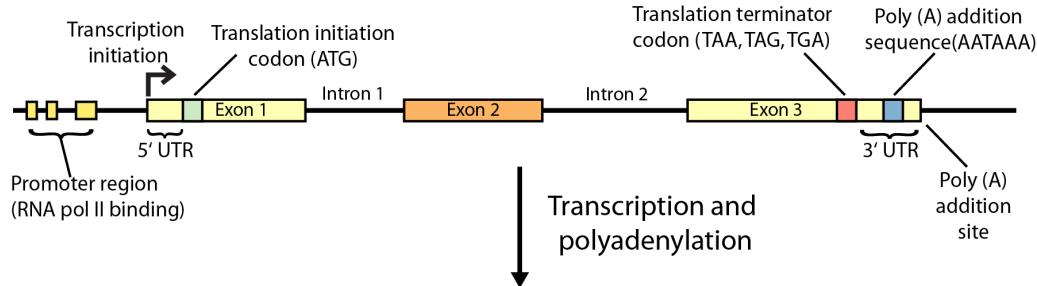
- **Module 1: Introduction to RNA Sequencing**
- Module 2: Alignment and Visualization
- Module 3: Expression and Differential Expression
- Module 4: Isoform Discovery and Alternative Expression
- Tutorials
  - Provide a working example of an RNA-seq analysis pipeline
  - Run in a ‘reasonable’ amount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

# Learning objectives of module 1

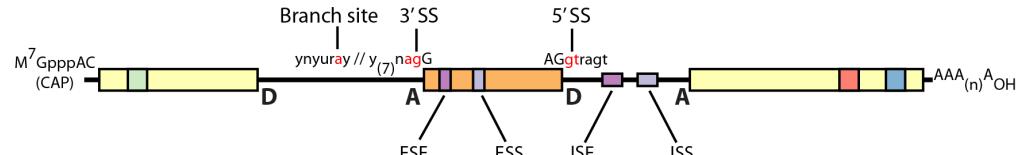
- Introduction to the theory and practice of RNA sequencing (RNA-seq) analysis
  - Rationale for sequencing RNA
  - Challenges specific to RNA-seq
  - General goals and themes of RNA-seq analysis work flows
  - Common technical questions related to RNA-seq analysis
  - Getting help outside of this course
  - Introduction to the RNA-seq hands on tutorial

# Gene expression

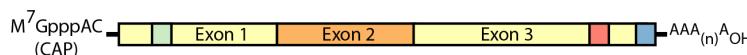
## Double-stranded genomic DNA template



## Single-stranded pre-mRNA (nuclear RNA)



## Mature mRNA

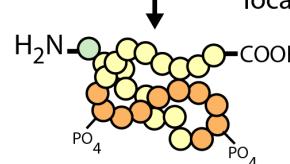


## RNA processing

## Protein (amino acid sequence)

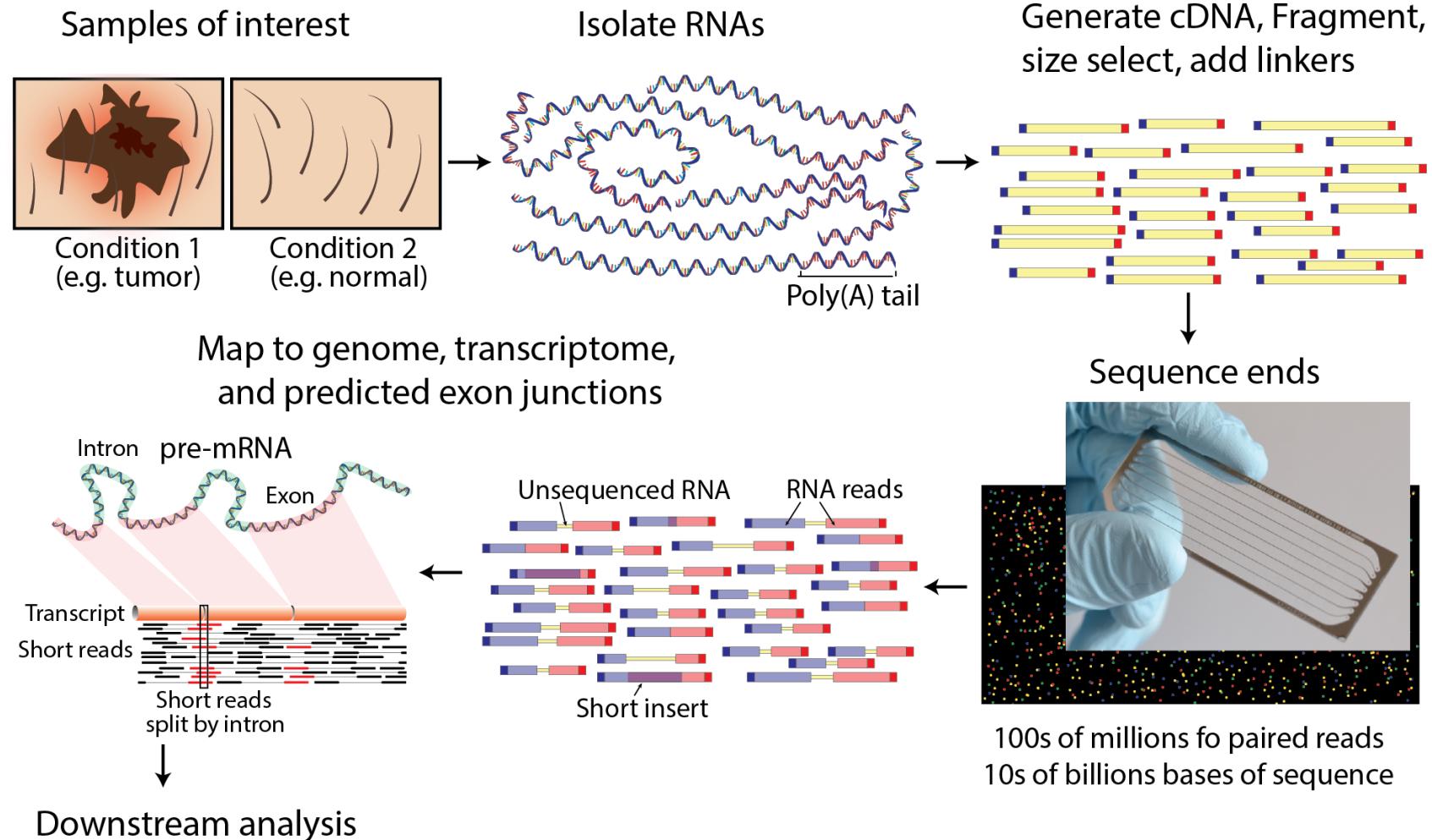


## Export to cytoplasm and translation



## Folding, posttranslational modification, subcellular localization, etc.

# RNA sequencing



# Why sequence RNA (versus DNA)?

- Functional studies
  - Genome may be constant but an experimental condition has a pronounced effect on gene expression
    - e.g. Drug treated vs. untreated cell line
    - e.g. Wild type versus knock out mice
- Predicting transcript sequence from genome sequence is difficult
  - Gene annotation is revolutionized by RNA-seq
- Some molecular features can only be observed at the RNA level
  - Alternative isoforms, fusion transcripts, RNA editing

# Why sequence RNA (versus DNA)?

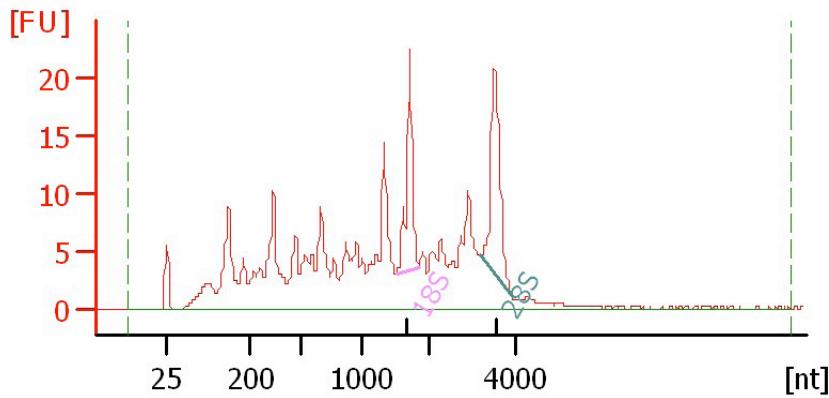
- Interpreting mutations that do not have an obvious effect on protein sequence
  - ‘Regulatory’ mutations that affect what mRNA isoform is expressed and how much
- Prioritizing protein coding somatic mutations (often heterozygous)
  - If the gene is not expressed, a mutation in that gene would be less interesting
  - If the gene is expressed but only from the wild type allele, this might suggest loss-of-function (haploinsufficiency)
  - If the mutant allele itself is expressed, this might suggest a candidate drug target

# Challenges

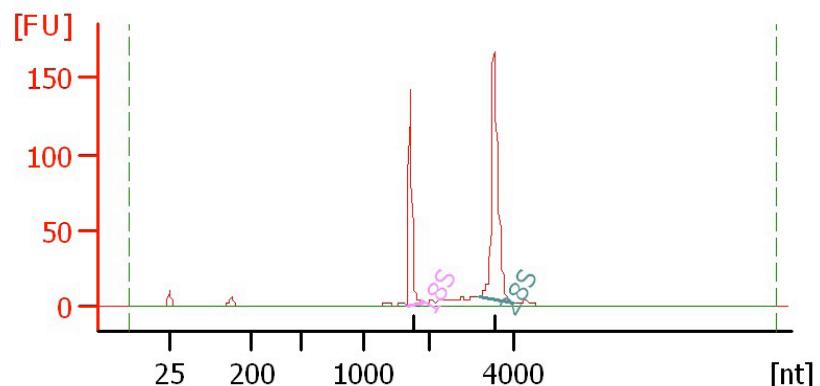
- Sample
  - Purity?, quantity?, quality?
- RNAs consist of small exons that may be separated by large introns
  - Mapping reads to genome is challenging
- The relative abundance of RNAs vary wildly
  - $10^5 - 10^7$  orders of magnitude
  - Since RNA sequencing works by random sampling, a small fraction 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://github.com/griffithlab/rnaseq\\_tutorial/wiki/Resources/Agilent\\_Trace\\_Examples.pdf](https://github.com/griffithlab/rnaseq_tutorial/wiki/Resources/Agilent_Trace_Examples.pdf)
- ‘RIN’ = RNA integrity number
  - 0 (bad) to 10 (good)



RIN = 6.0



RIN = 10

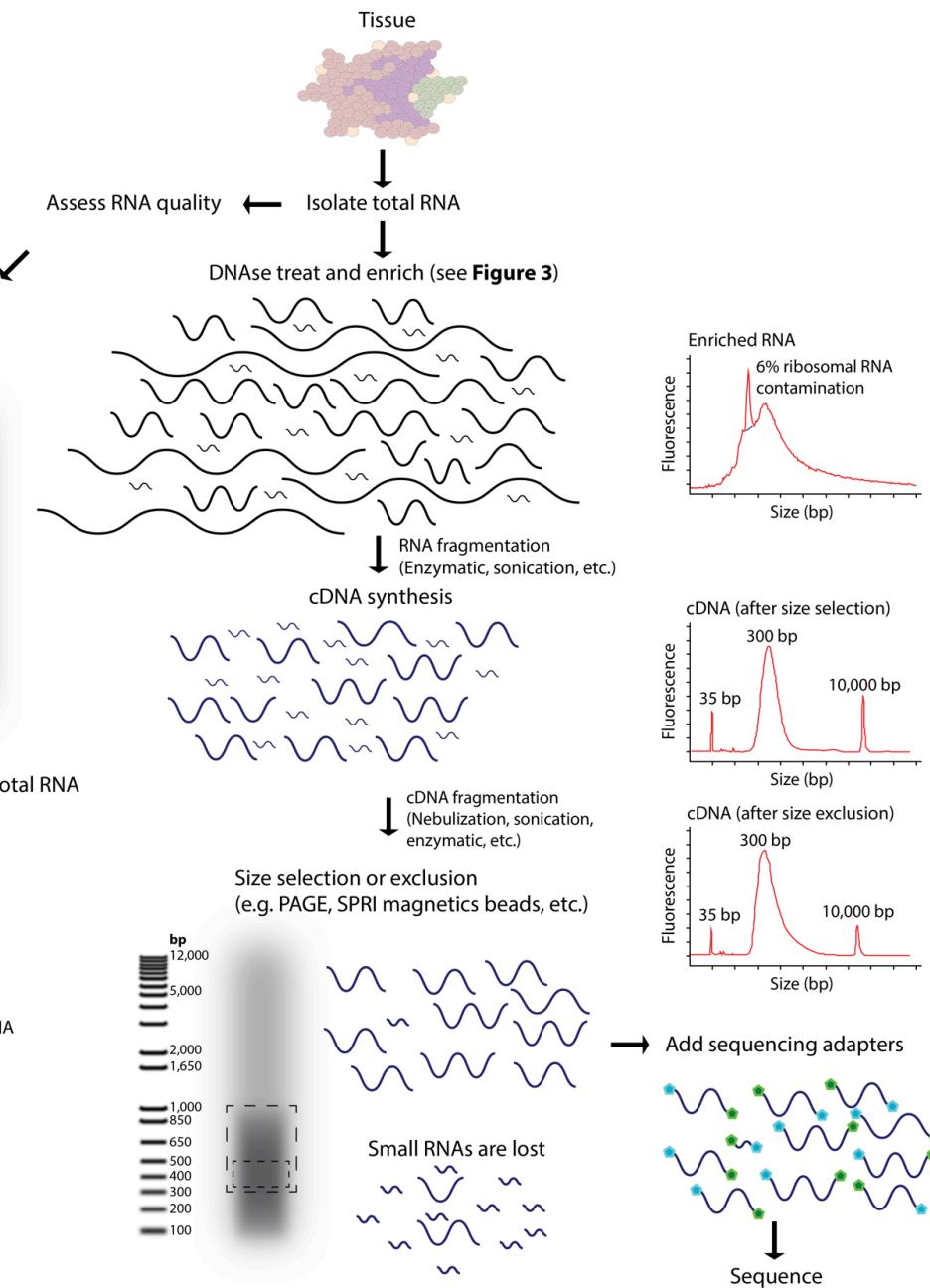
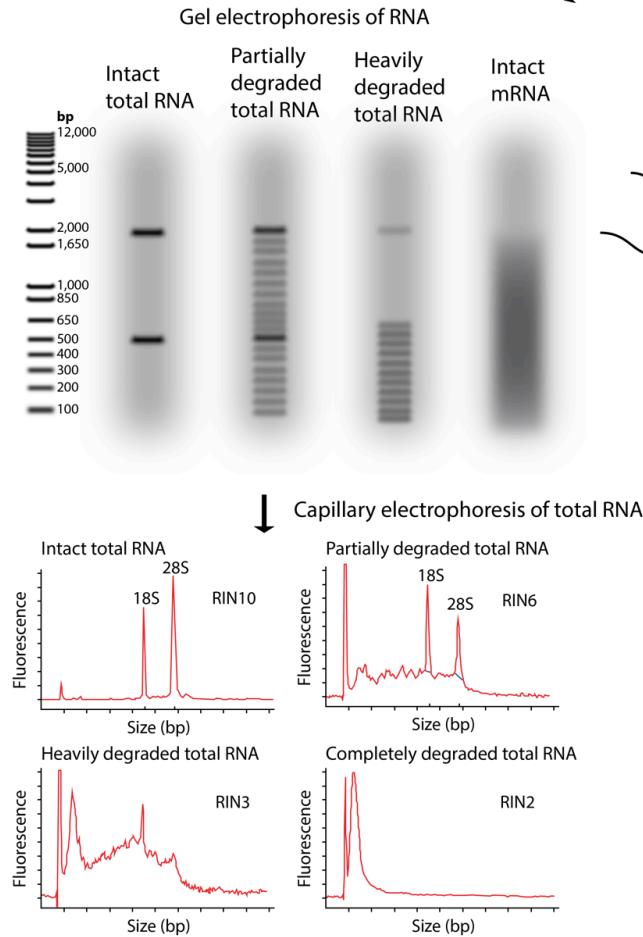
# Design considerations

- 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://github.com/griffithlab/rnaseq\\_tutorial/wiki/Resources/ENCODE\\_RNAseq\\_standards\\_v1.0.pdf](https://github.com/griffithlab/rnaseq_tutorial/wiki/Resources/ENCODE_RNAseq_standards_v1.0.pdf)

# There are many RNA-seq library construction strategies

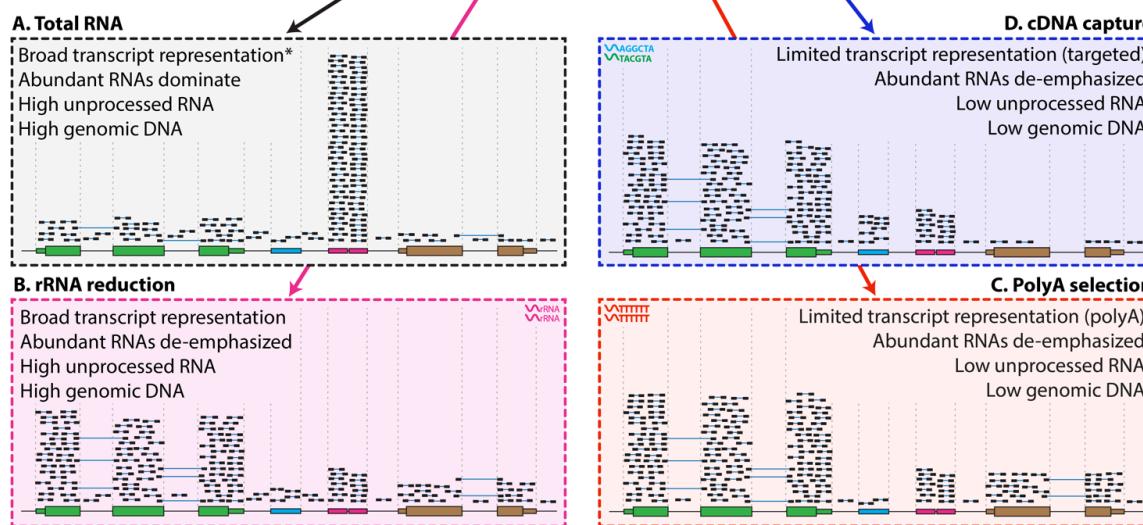
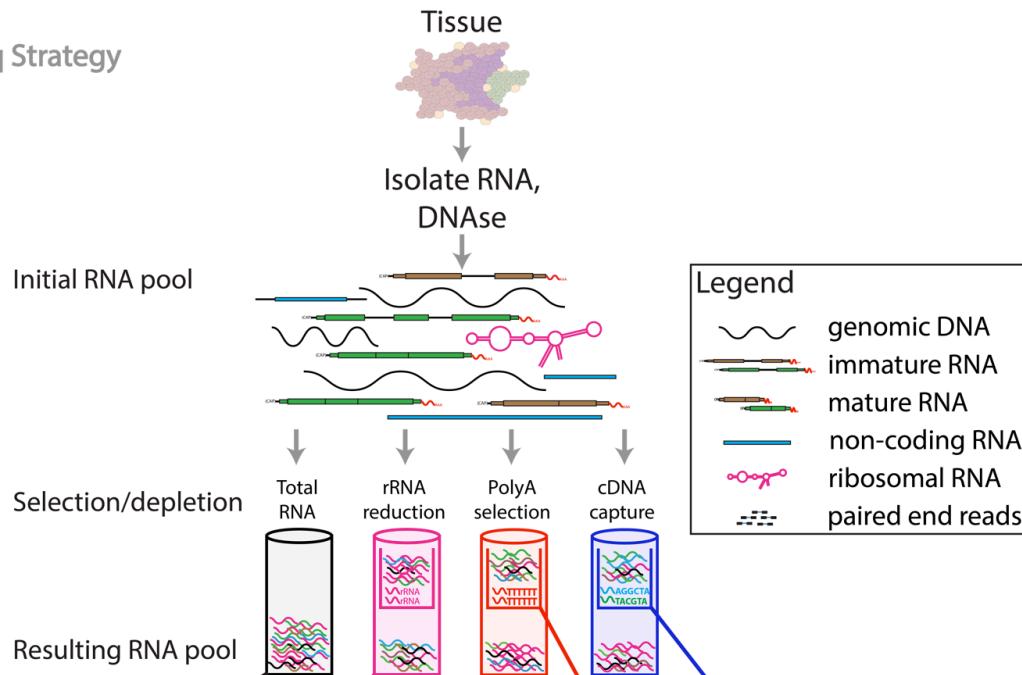
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library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

# Fragmentation and size selection



# RNA sequence selection/depletion

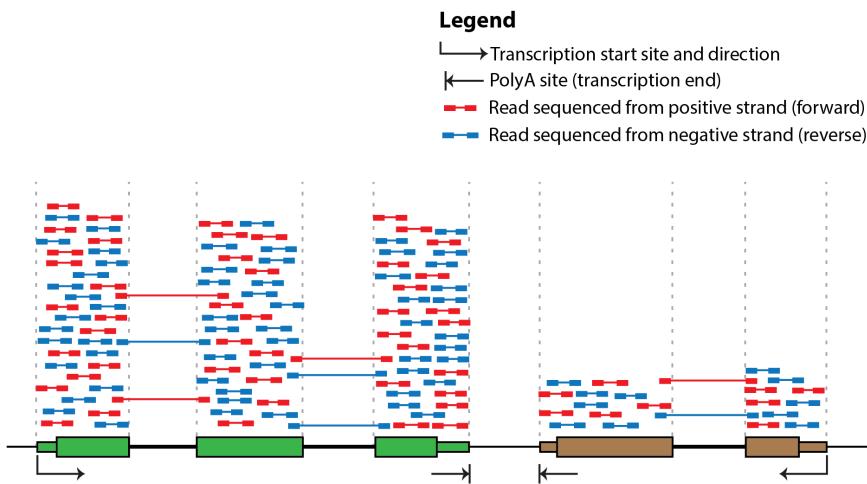
## RNA-seq Strategy



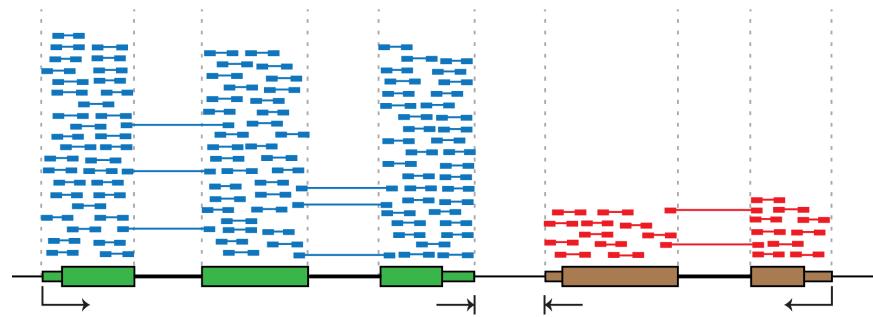
Expected Alignments

# Stranded vs. unstranded

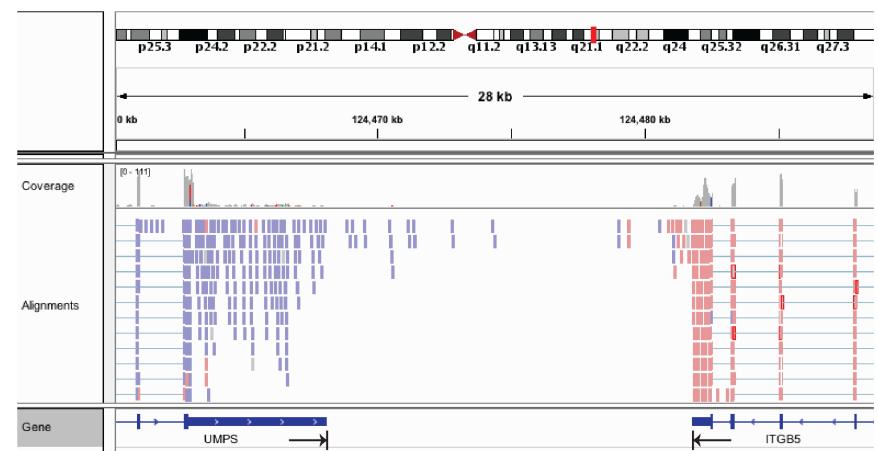
## A. Depiction of cDNA fragments from an unstranded library



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

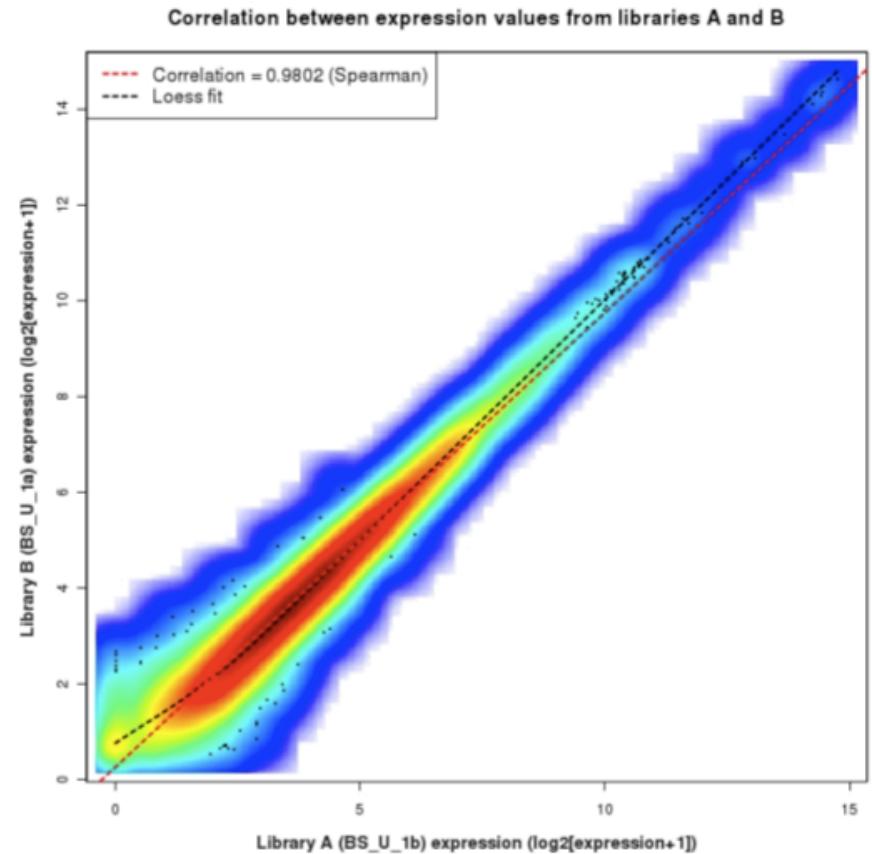


## C. Viewing strand of aligned reads in IGV



# 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.

# BioStar exercise

- Go to the BioStar website:
  - <http://www.biostars.org/>
  - If you do not already have an OpenID (e.g. Google, Yahoo, etc.)
  - Login -> ‘get one’
- Login and set up your user profile
- Tasks:
  - Find a question that seems useful and ‘vote it up’
  - Answer a question [optional]
  - Search for a topic area of interest and ask a question that has not already been asked [optional]

# Common questions: Should I remove duplicates for RNA-seq?

- Maybe... more complicated question than for DNA
- Concern.
  - Duplicates may correspond to biased PCR amplification of particular fragments
  - For highly expressed, short genes, duplicates are expected even if there is no amplification bias
  - Removing them may reduce the dynamic range of expression estimates
- If you do remove them, assess duplicates at the level of paired-end reads (fragments) not single end reads

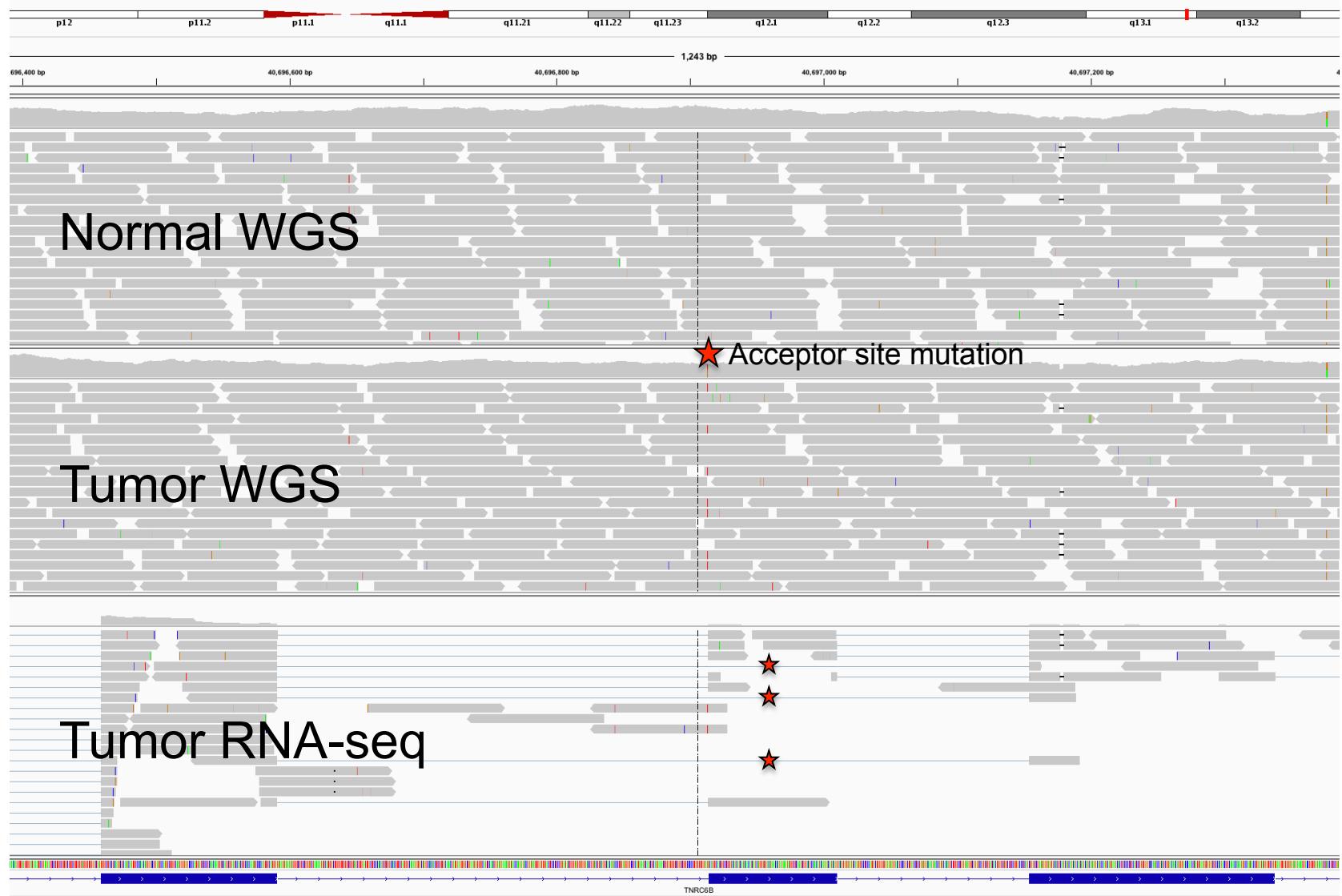
# Common questions: How much library depth is needed for RNA-seq?

- Depends on a number of factors:
  - Question being asked of the data. Gene expression? Alternative expression? Mutation calling?
  - Tissue type, RNA preparation, quality of input RNA, library construction method, etc.
  - Sequencing type: read length, paired vs. unpaired, etc.
  - Computational approach and resources
- Identify publications with similar goals
- Pilot experiment
- Good news: 1-2 lanes of recent Illumina HiSeq data should be enough for most purposes

# Common questions: What mapping strategy should I use for RNA-seq?

- Depends on read length
- < 50 bp reads
  - Use aligner like BWA and a genome + junction database
  - Junction database needs to be tailored to read length
    - Or you can use a standard junction database for all read lengths and an aligner that allows substring alignments for the junctions only (e.g. BLAST ... slow).
    - Assembly strategy may also work (e.g. Trans-ABySS)
- > 50 bp reads
  - Spliced aligner such as Bowtie/TopHat

# Visualization of spliced alignment of RNA-seq data

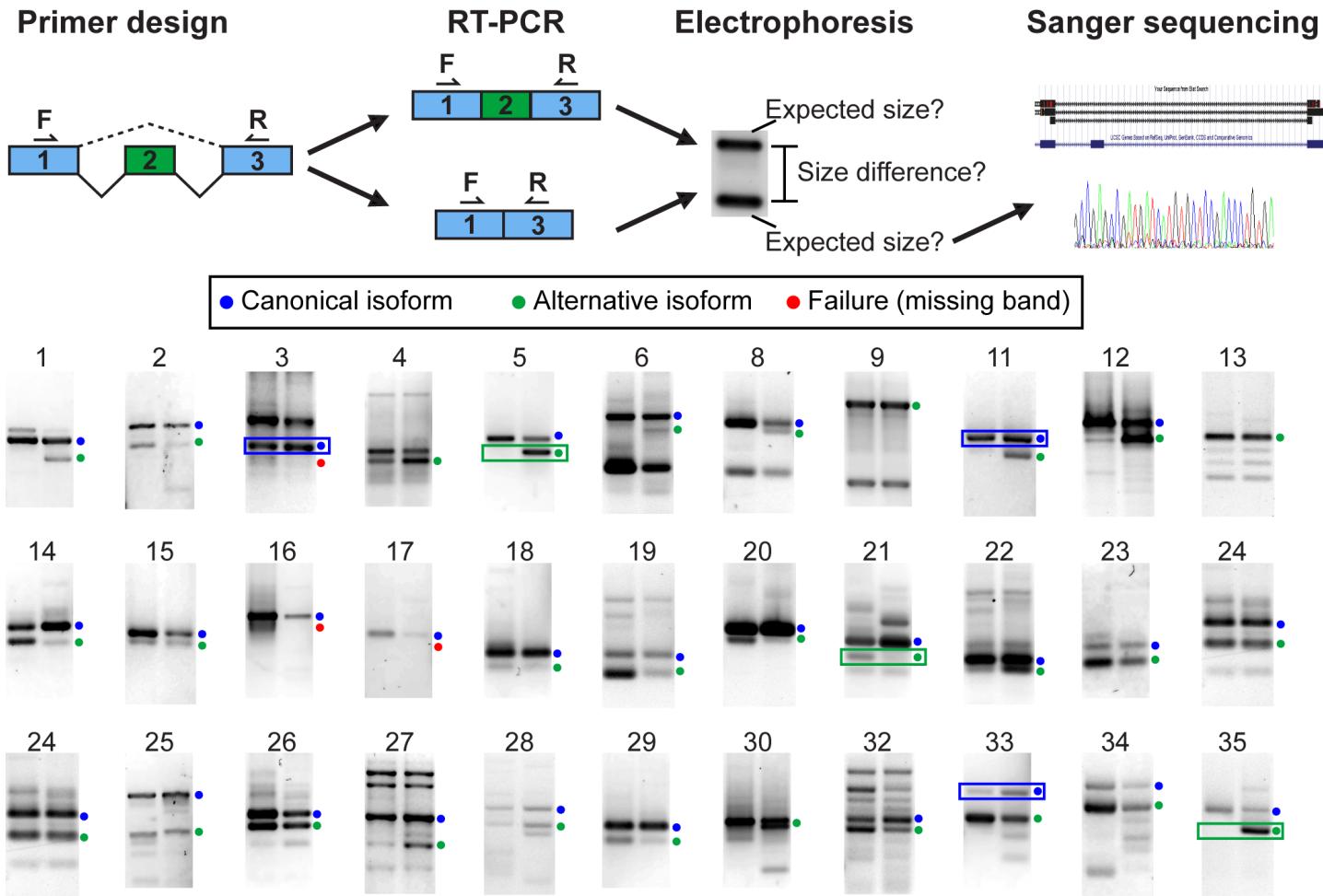


IGV screenshot

# Common questions: how reliable are expression predictions from RNA-seq?

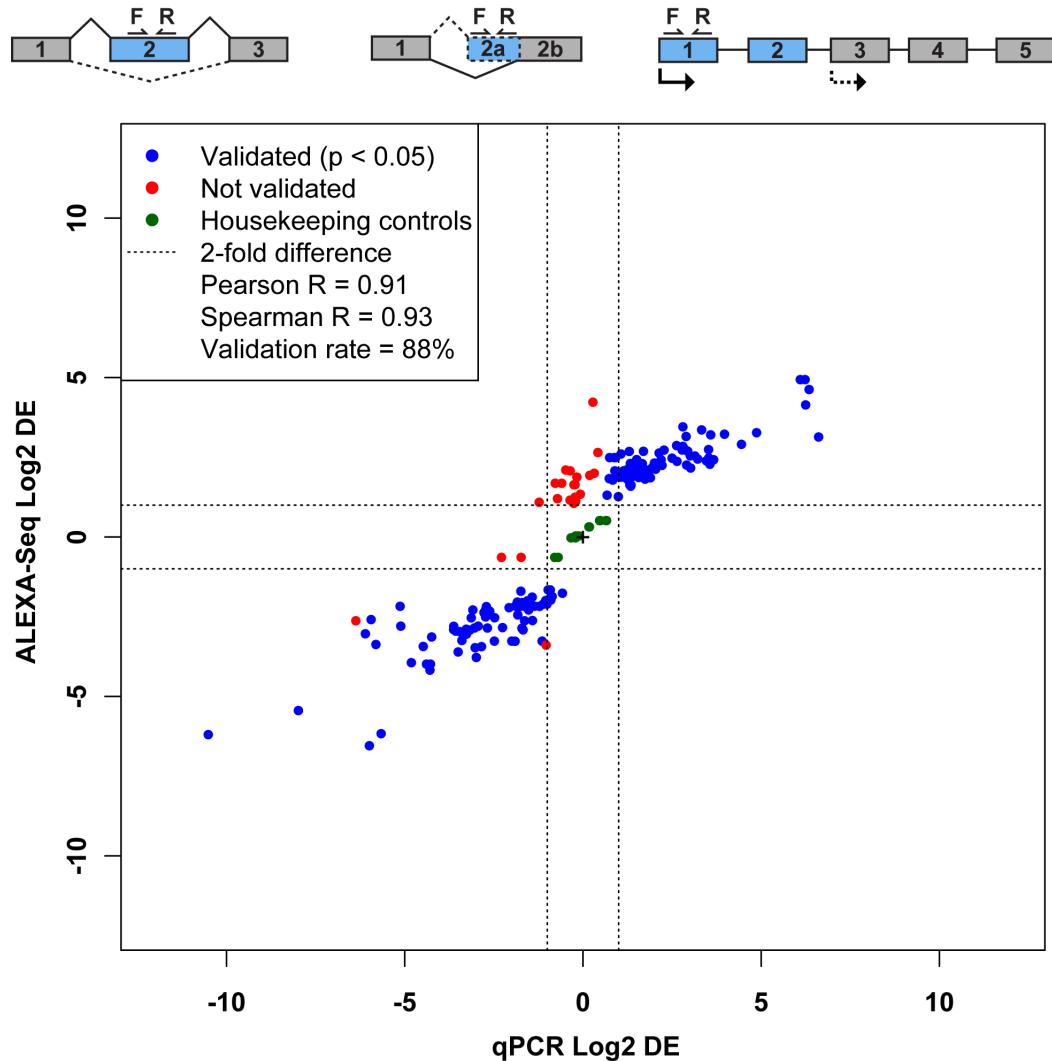
- Are novel exon-exon junctions real?
  - What proportion validate by RT-PCR and Sanger sequencing?
- Are differential/alternative expression changes observed between tissues accurate?
  - How well do DE values correlate with qPCR?
- 384 validations
  - qPCR, RT-PCR, Sanger sequencing
- See ALEXA-Seq publication for details:
  - Also includes comparison to microarrays
  - Griffith et al. *Alternative expression analysis by RNA sequencing*. Nature Methods. 2010 Oct;7(10):843-847.

# Validation (qualitative)



33 of 192 assays shown. Overall validation rate = 85%

# Validation (quantitative)



qPCR of 192  
exons identified  
as alternatively  
expressed by  
ALEXA-Seq

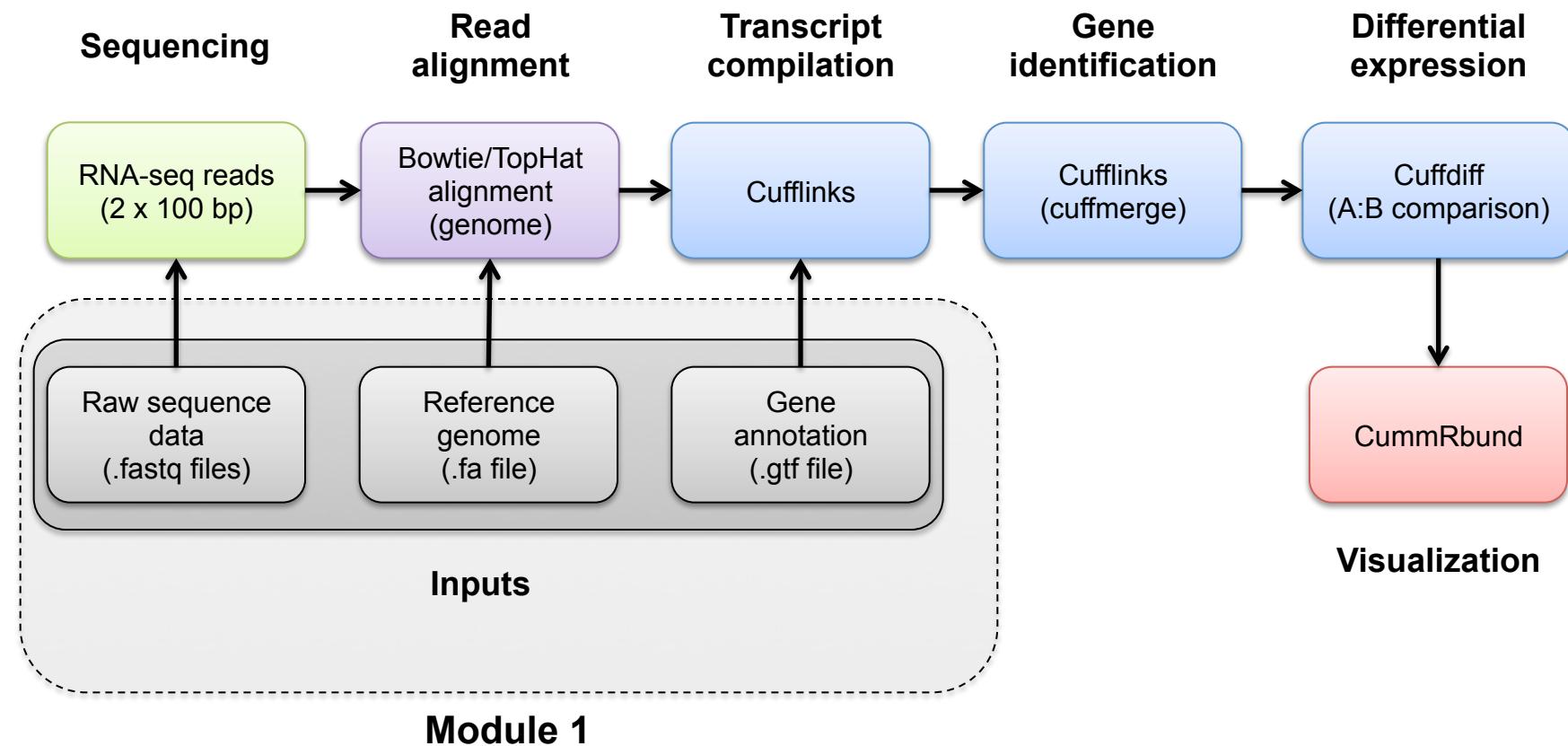
**Validation rate = 88%**

# Common questions: What if I don't have a reference genome for my species?

- Have you considered sequencing the genome of your species?
- If that is not practical or you simply prefer a transcript discovery approach that does not rely on prior knowledge of the genome or transcriptome there are some tools available ...
  - Unfortunately de novo transcriptome assembly is beyond the scope of this workshop
  - The good news is that the skills you learn here will help you figure out how to install and run those tools yourself

# **Introduction to tutorial (Module 1)**

# Bowtie/Tophat/Cufflinks/Cuffdiff RNA-seq Pipeline



# Break