# RNASEQ PROJECT DESIGN

Experimental Design

Assembly in Non-Model Organisms

And other (hopefully useful) Stuff

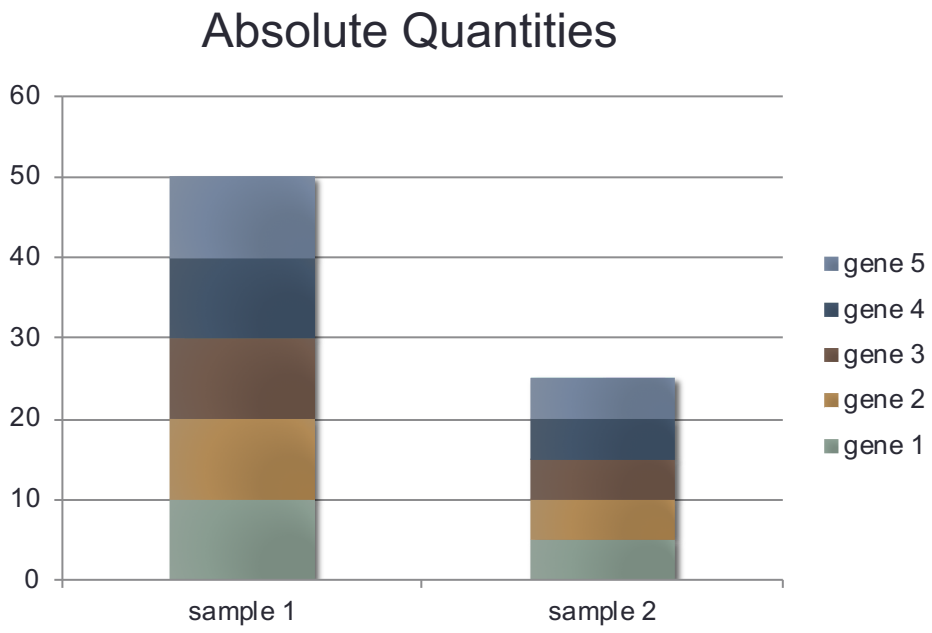
Meg Staton
mstaton1@utk.edu
University of Tennessee
Knoxville, TN

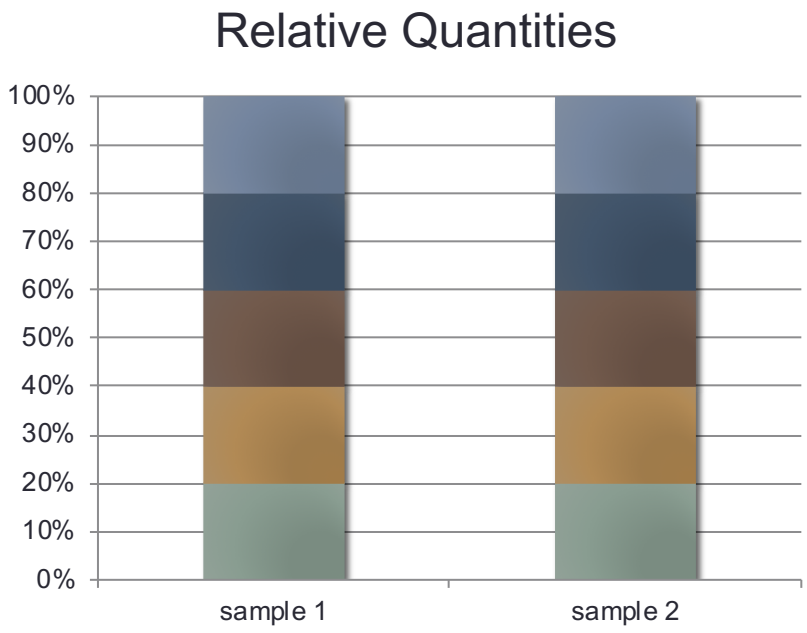
# Gene Expression Quantification

- This is probably the most common experimental goal, lets start with that…

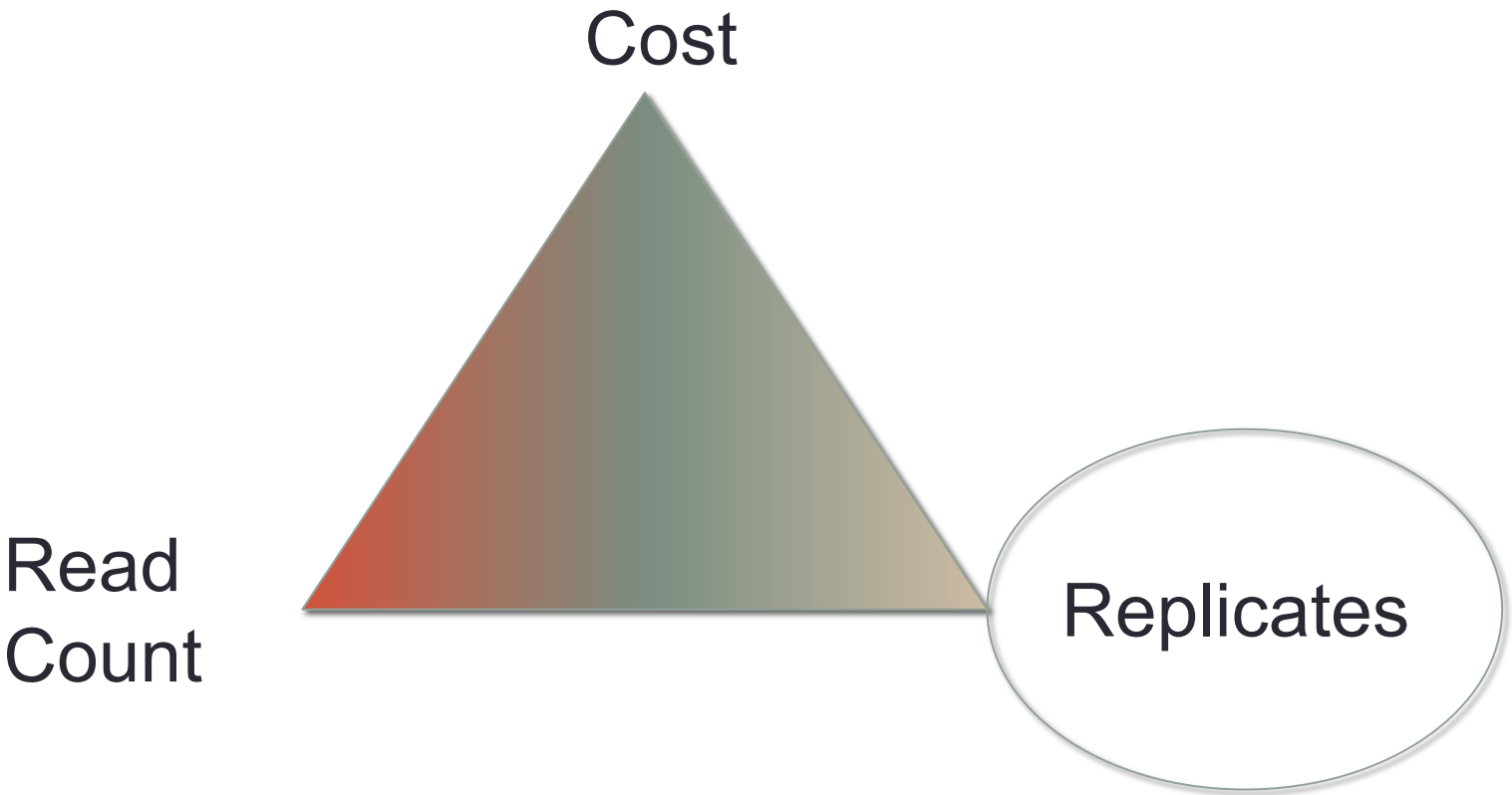
# Limitations

RNASeq gives you relative abundance only



# Major Considerations for Project Design



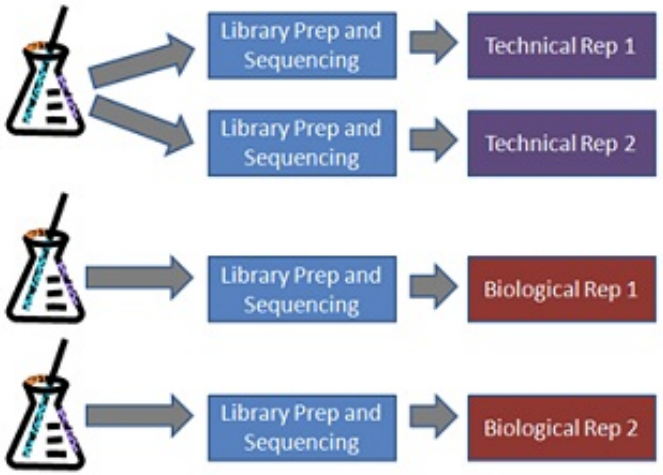

**EXPERIMENTAL DESIGN**

Who is your resident statistician and/or bioinformatician? Buy them a coffee and make friends. **Preferably before starting the experiment!**

# Replicates

- Biological Replicates – independent biological sample, processed separately and barcoded
- Technical Replicates – independent library construction or sequencing of the same biological sample

- Technical reproducibility is very good for RNASeq
- Biological variation is much greater!




"Thinking About RNA Seq Experimental Design for Measuring Differential Gene Expression: The Basics"
http://gkno2.tumblr.com/post/24629975632/thinking-about-rna-seq-experimental-design-for

Marioni, J.C., et al. (2008) RNA-seq: An assessment of technical reproducibility and

# Replicates – How many?

- beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression

Liu Y, Zhou J, White KP. **RNA-seq differential expression studies: more sequence or more replication?** Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.

- Very difficult to publish with 1 rep
- Publications still coming out with 3 reps

# Replicates – How many?

- The ultimate test – 48 replicates. What were the results?

**How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?**

Nicholas J. Schurch,[1,6] Pietá Schofield,[1,2,6] Marek Gierliński,[1,2,6] Christian Cole,[1,6] Alexander Sherstnev,[1,6] Vijender Singh,[2] Nicola Wrobel,[3] Karim Gharbi,[3] Gordon G. Simpson,[4] Tom Owen-Hughes,[2] Mark Blaxter,[3] and Geoffrey J. Barton[1,2,5]

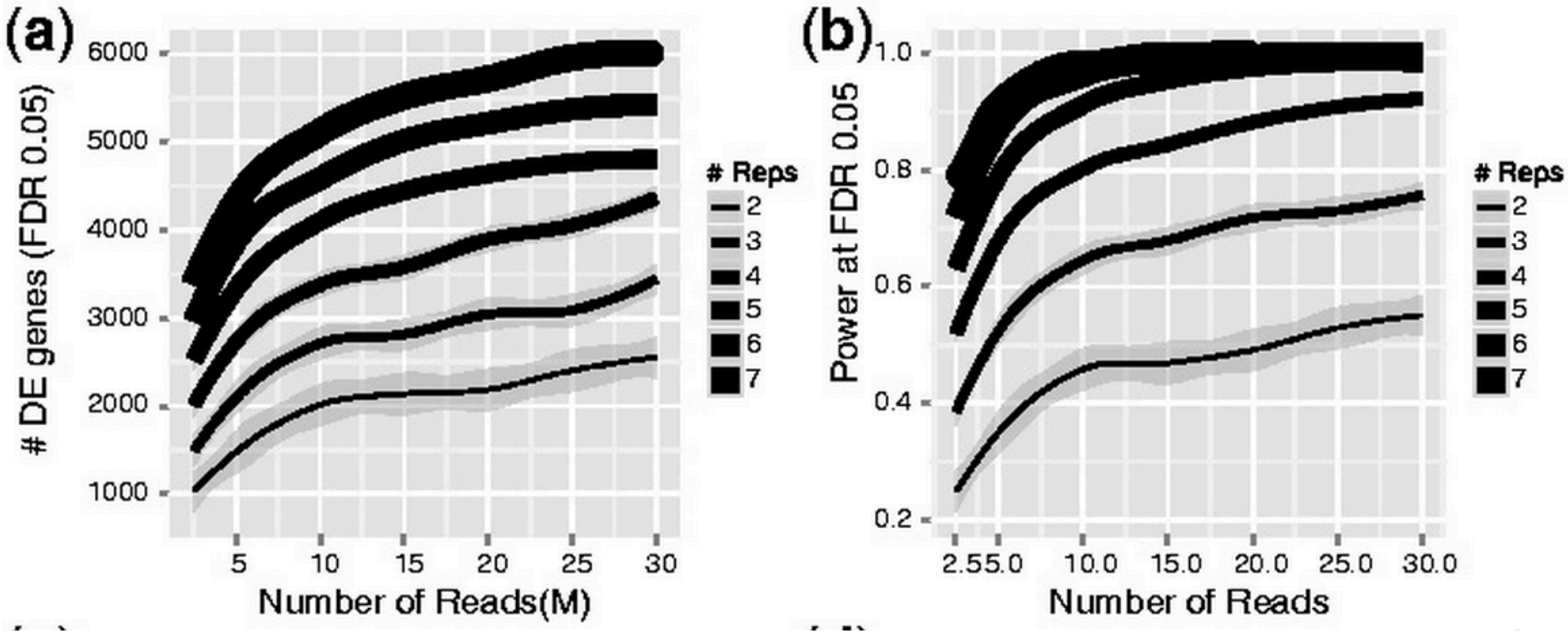
"With three biological replicates, nine of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates."

"these results suggest that at least six biological replicates should be used, rising to at least 12 when it is important to identify SDE genes for all fold changes"

"If fewer than 12 replicates are used, a superior combination of true positive and false positive performances makes edgeR and DESeq2 the leading tools."

# Replicates – How many?

Liu Y, Zhou J, White KP. **RNA-seq differential expression studies: more sequence or more replication?** Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.



# Replicates – Software?

- Both EdgeR and DeSeq will calculate variance from replicates
- Which to use?

- From the horse's mouth:
- "Of course, we like to claim that DESeq is better than edgeR, and for only two or three replicates, I do think so, but for five or more replicates, edgeR's 'moderation' feature really pays off."

  -Simon Anders on SeqAnswers

From Schurch et al 2014:

"For experiments with <12 replicates per condition; use edgeR (exact) or DESeq2.

For experiments with >12 replicates per condition; use DESeq."

http://seqanswers.com/forums/showthread.php?t=10410

# Pooling

Does pooling my samples count as biological replicates?

No. With pooling, you will get an accurate mean, but not an accurate measure of variability across individuals.

Literature is mixed on this issue. But it doesn't make solid statistical sense and the downsides are significant:

"the DEGs identified in pooled samples suffered low positive predictive values" - Rajkumar et al, 2015

# Blocking

- Randomized Block Design
- Divide samples (individuals) into blocks in order to control variation between blocks
- Randomize - assigning individuals at random to treatments in an experiment

| | West Virginia | South Carolina |
|---|---|---|
| Early flowering cultivar | 20 | 20 |
| Late flowering cultivar | 20 | 20 |

# Blocking

Lane effects

- systematically bad sequencing cycles and errors in base calling

A cautionary tale

- Original paper:

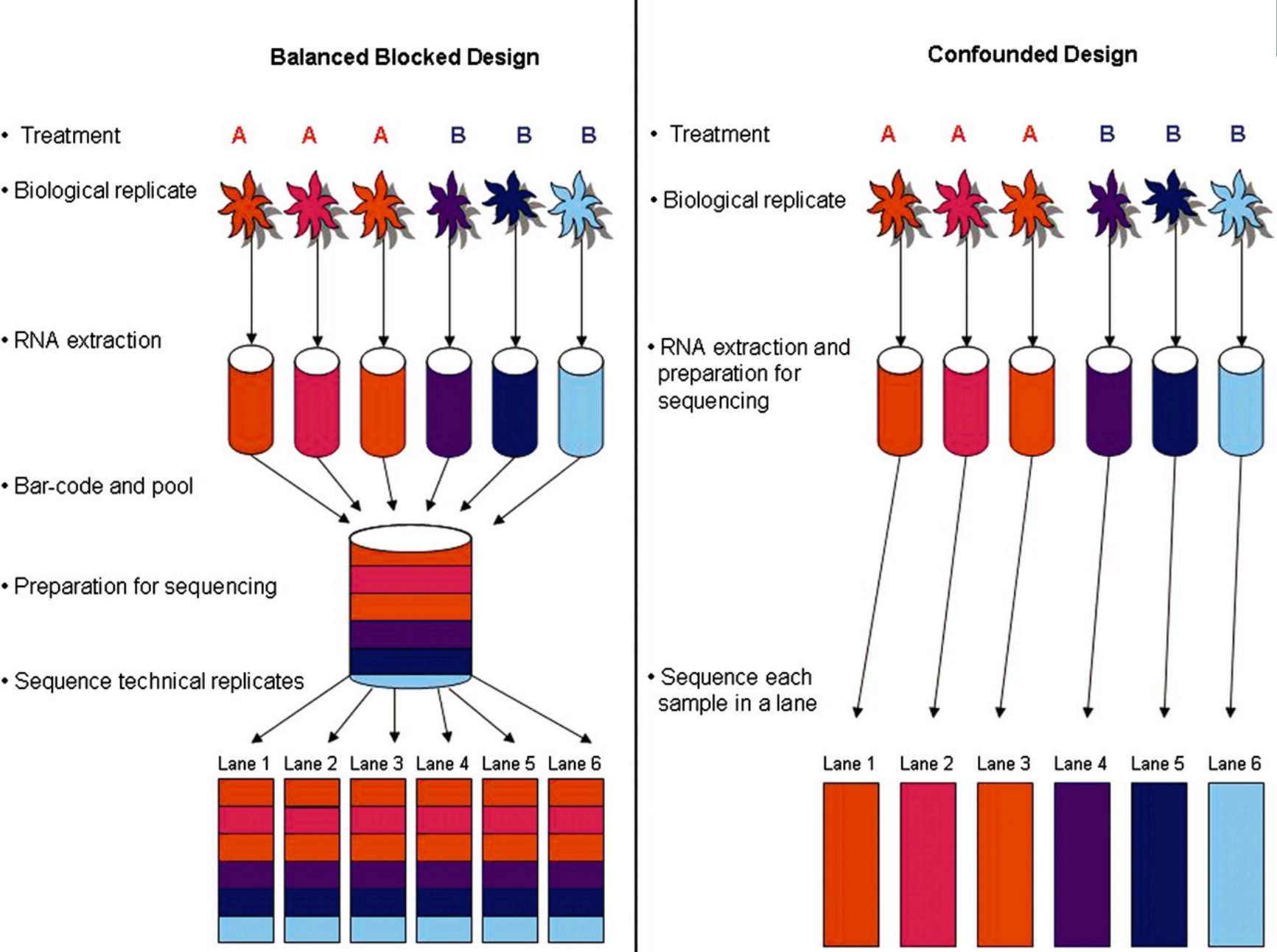
Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin[a,b,1], Yiing Lin[c,1], Joseph R. Nery[d], Mark A. Urich[d], Alessandra Breschi[e,f], Carrie A. Davis[g],

- Reanalysis pointing out flawed statistical design and questioning results

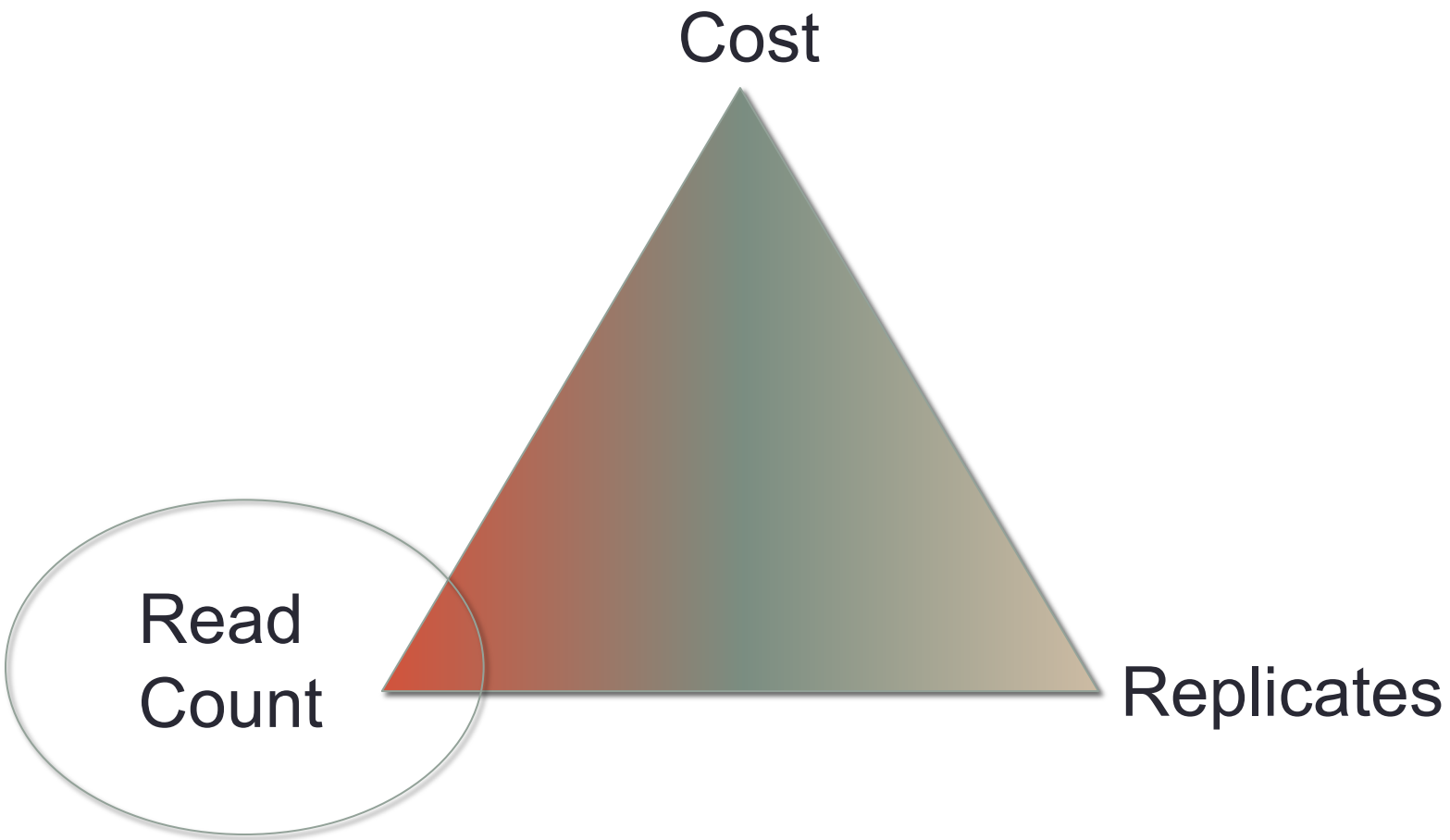
A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

Yoav Gilad, Orna Mizrahi-Man




Auer and Doge Statistical Design and Analysis of RNA Sequencing Data Genetics June 2010 vol. 185 no. 2 405-416

# Major Considerations for Project Design



# Read Count - How to Decide?

- Standards, Guidelines and Best Practices for RNA-Seq
- V1.0 (June 2011)
- The ENCODE Consortium
- What are you trying to do?
  - Compare two mRNA samples for differential expression (30M PE per sample)
  - Discover novel elements, perform more precise quantification, especially of lowly expressed transcripts (100-200M PE per sample)

http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

# Read Count - How to Decide?

- “As low as one million reads can provide the same sequencing accuracy in transcript abundance (r=0.99) as >30 million reads for highly-expressed genes in all six species”
- Caveat: This only applies to the 50% most highly expressed genes
  - Lei R, Ye K, Gu Z, Sun X. (2014) Diminishing returns in next-generation sequencing (NGS) transcriptome data. Gene S0378-1119(14)01386-9.

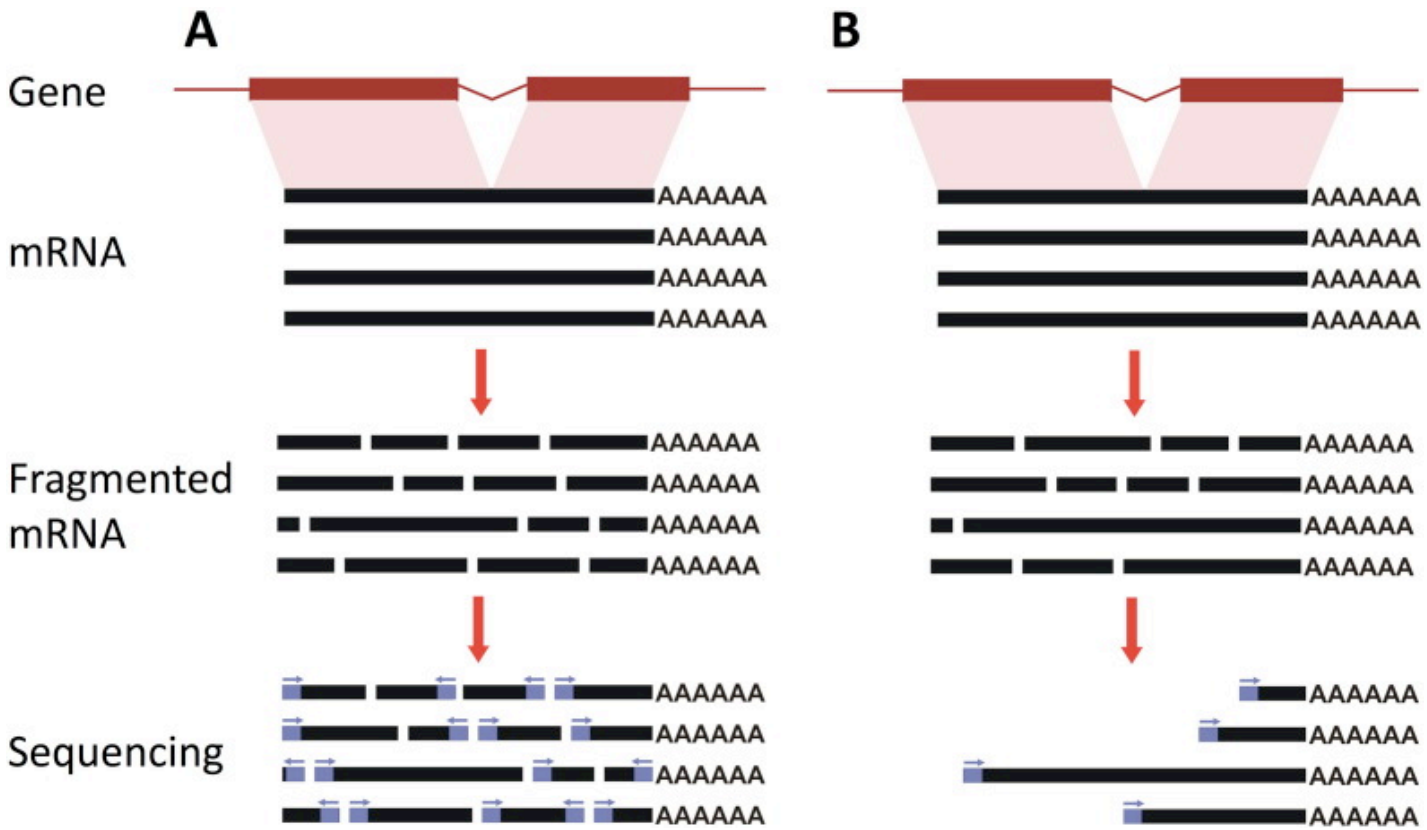
- Beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression
  - Liu Y, Zhou J, White KP. RNA-seq differential expression studies: more sequence or more replication? Bioinformatics. 2014;30(3):301-304. doi:10.1093/bioinformatics/btt688.

# Read Count - How to Decide?

- General recommendations:
- If you have to choose between depth and replicates, choose more replicates
- Look at what is being published in your community
- What resources do you already have?
  - Well assembled and annotated genomes – save money by using single ends, shorter reads
  - De novo transcriptome assembly – longer reads, paired ends

http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf

# 3’ RNASEq (3’TagSeq)




Normal RNASeq

3’ RNASeq

# 3’ RNASeq

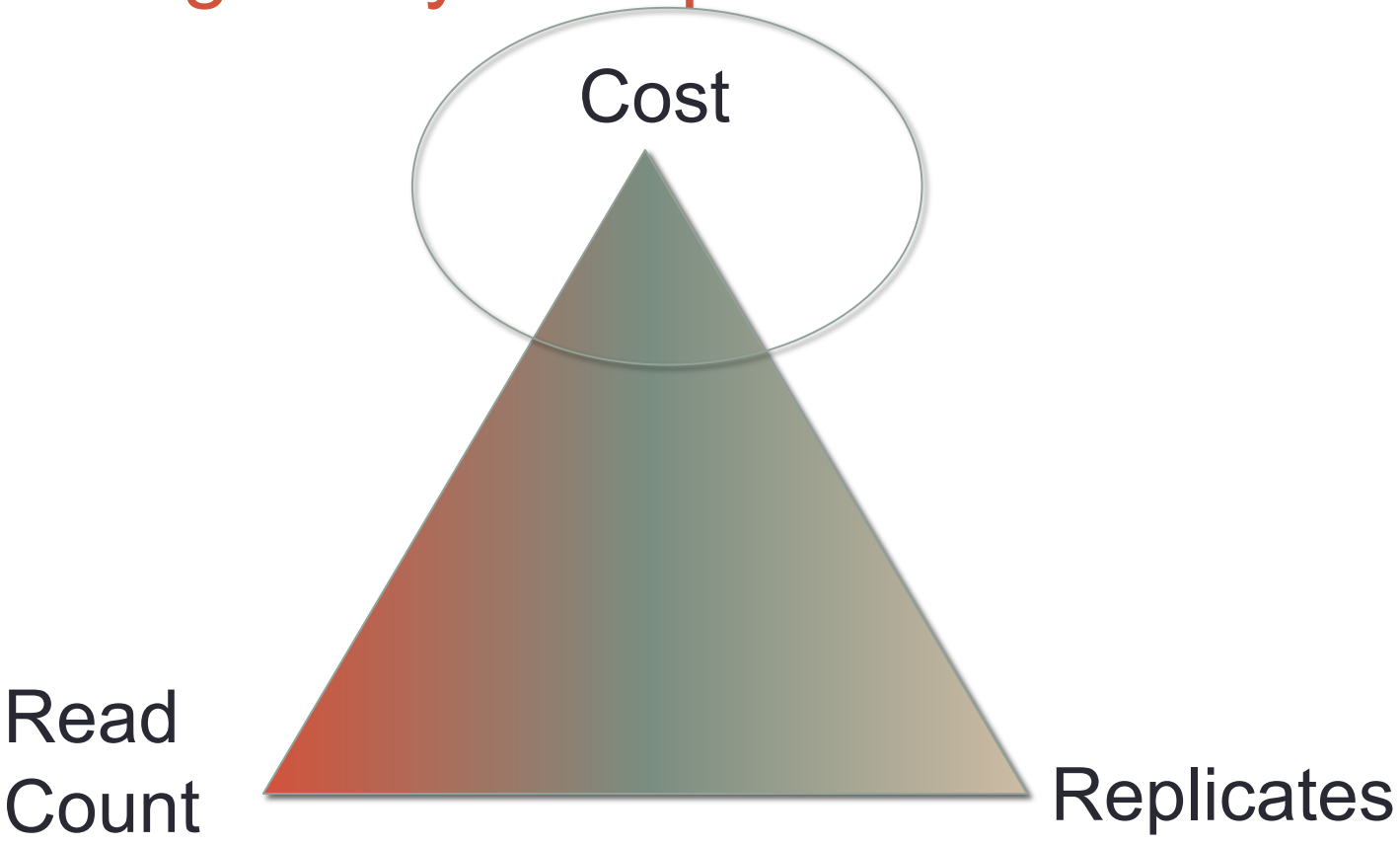
- Advantages
  - Requires fewer reads for same statistical power
  - Easier library prep, costs much less
  - Single read sequencing is sufficient (another cost savings)
- Disadvantages
  - No transcript splicing info
  - Only for eukaryotes
  - Better for organisms with reference genomes:

“when little genomic information is available for the species studied, the standard RNA-seq presents a better cost-benefit compromise, whereas for model species, the 3′ RNA-seq method might more accurately detect differential expression.”

-Tandonnet and Torres, 2017

Traditional *versus* 3′ RNA-seq in a non-model species

# What's right for your experiment?



# Publicly Posted Pricing

UT Genomics Core
https://ceb.utk.edu/dna-sequencing/

UTexas Austin Genomic Sequencing and Analysis Facility
https://wikis.utexas.edu/display/GSAF/Library+Prep+and+NGS+Pricing+Descriptions

Science Exchange
https://www.scienceexchange.com/services/illumina-ngs

Cornell University Institute of Biotechnology
http://www.biotech.cornell.edu/brc/genomics/services/price-list

UC Davis Genome Center
http://dnatech.genomecenter.ucdavis.edu/prices/

# What about other projects?

- Transcriptome assembly
- Splice variants
- Annotating the genes in a reference genome

Conesa et al 2016. **A survey of best practices for RNA-seq data analysis**

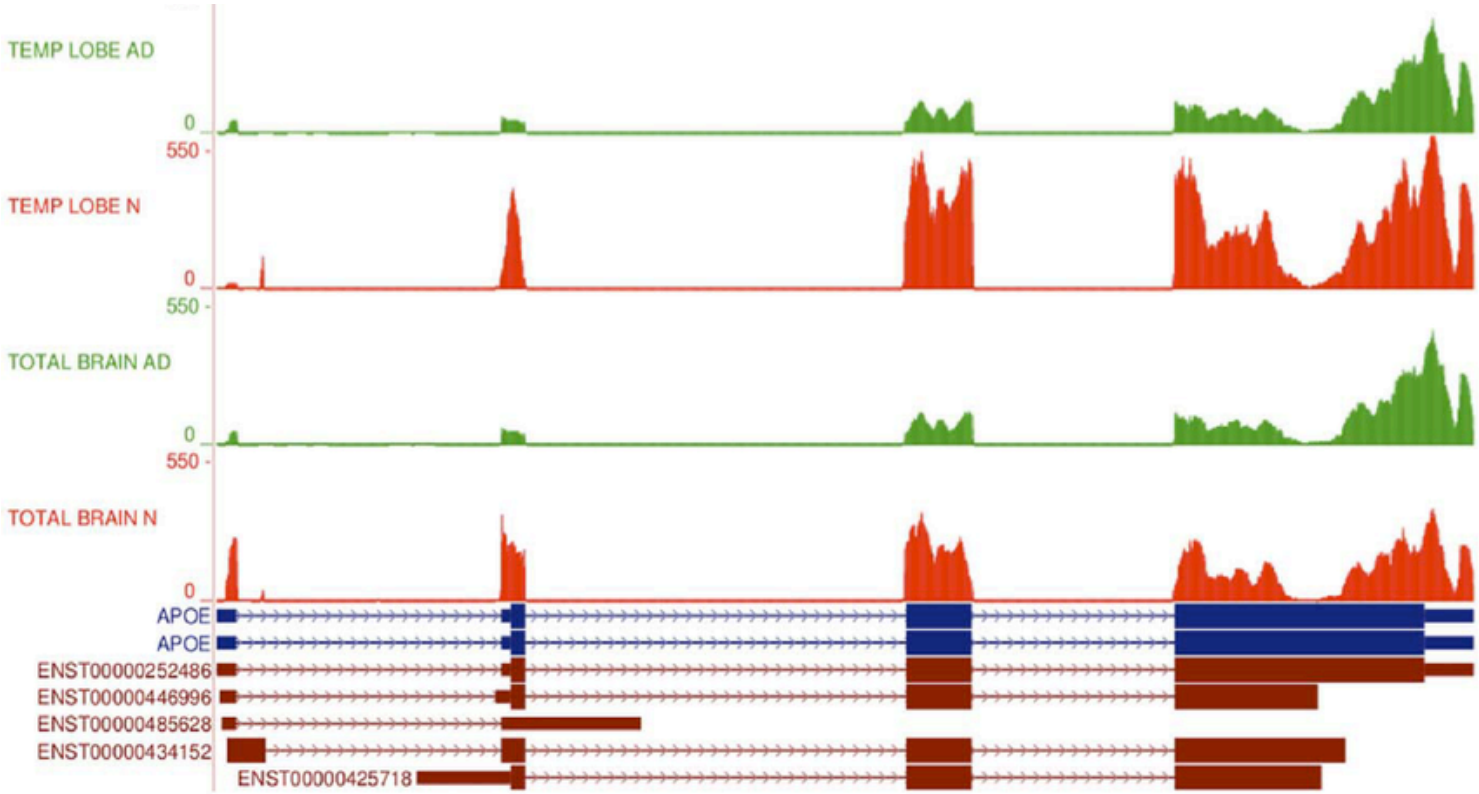
# De novo assembly

- 20-40 million reads (if using short reads)
- Longer read length is better
- Paired end is better
- Consider a long read technology
  - In addition to short read
  - On its own?
- Single individual if possible, many tissues and development stages if possible

MacManes 2018, The Oyster River Protocol: a multi-assembler and kmer approach for de novo transcriptome assembly

MacManes 2016, Establishing evidenced-based best practice for the *de novo* assembly and evaluation of transcriptomes from non-model organisms

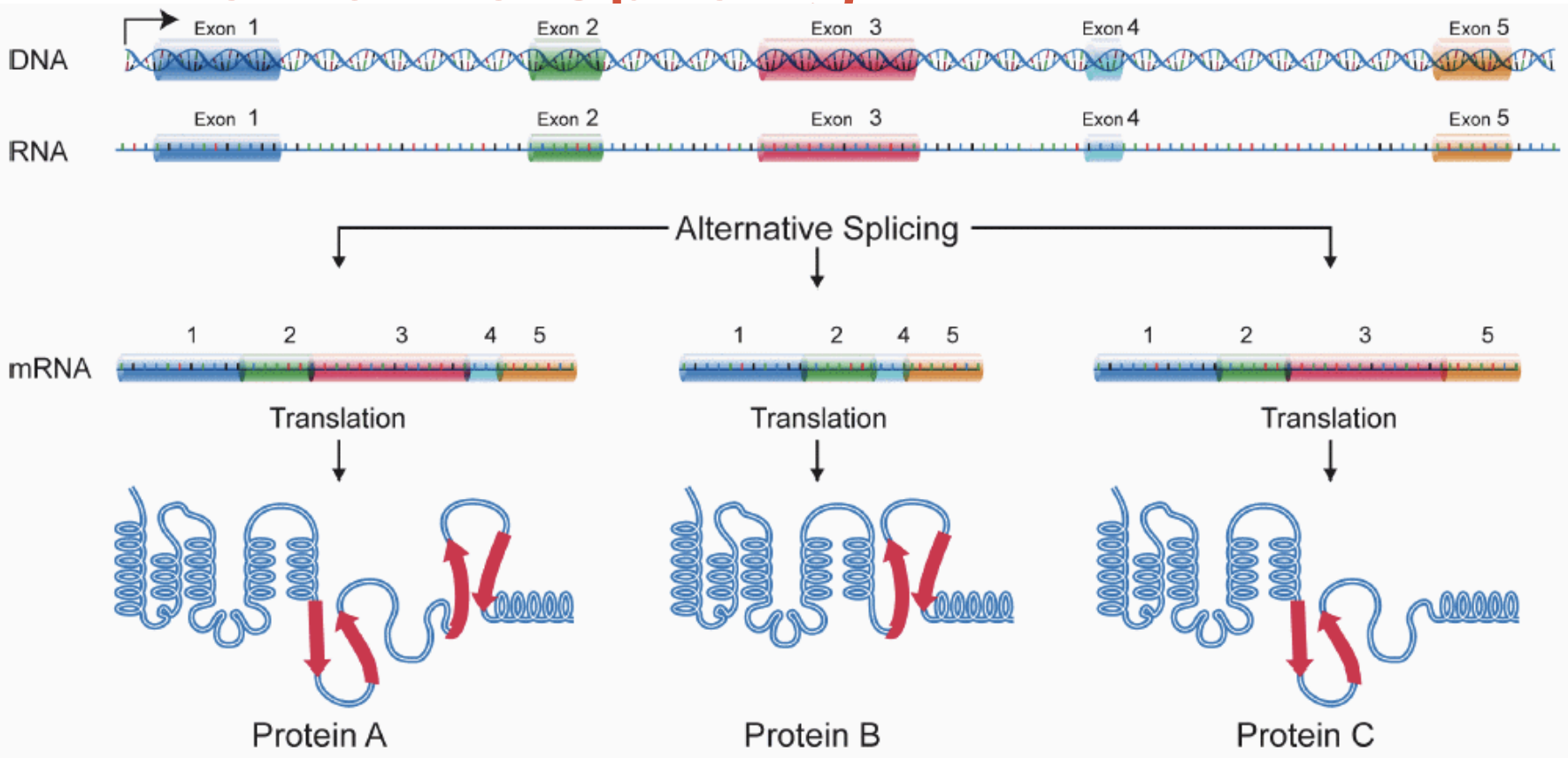
# Genome Annotation




Similar advice to de novo assembly:

- Longer read length is better
- Paired end is better
- Consider a long read technology
- Single individual if possible, many tissues and development stages if possible

# Alternative Splicing



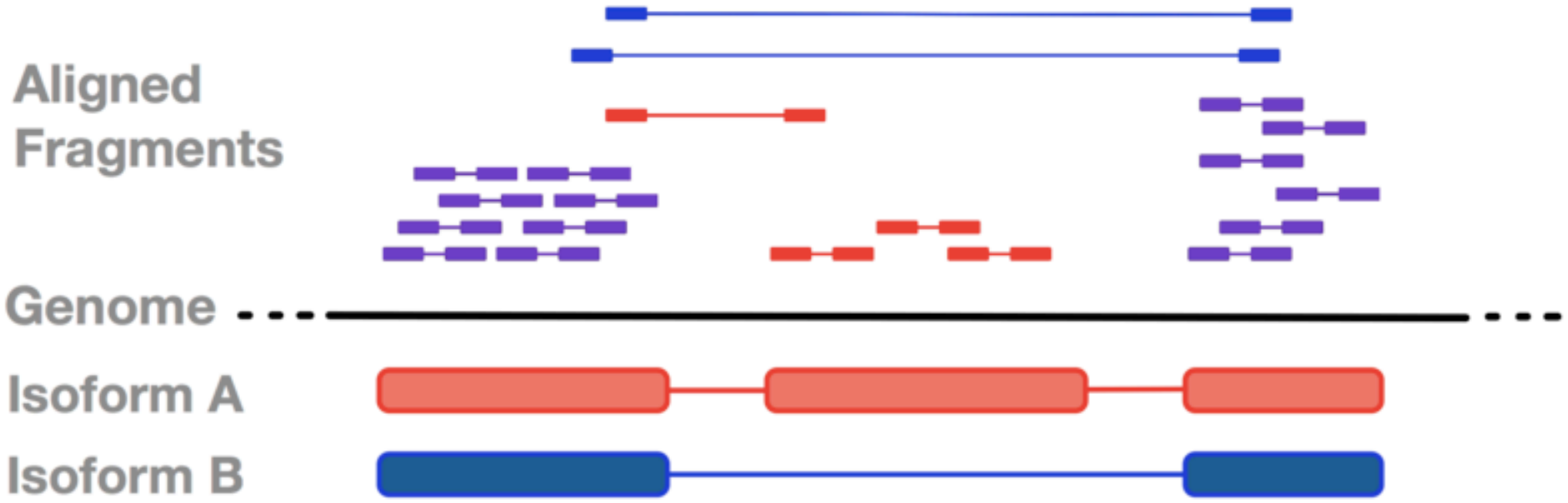

Splice variants are often tissue-specific. In humans, up to 95% of multiexonic genes have multiple splice isoforms.

# Detecting Known Isoform Variants

Ambiguous – No information about isoform.

Indicate isoform A.

Indicate isoform B.




http://cgrlucb.wikispaces.com/Isoform+Deconvolution+and+Unannotated+Species

# Overview

- Project Design
  - Replicates
  - Cost
  - Read Count

May differ based on project goals and how you will use the data.

Read a lot!