

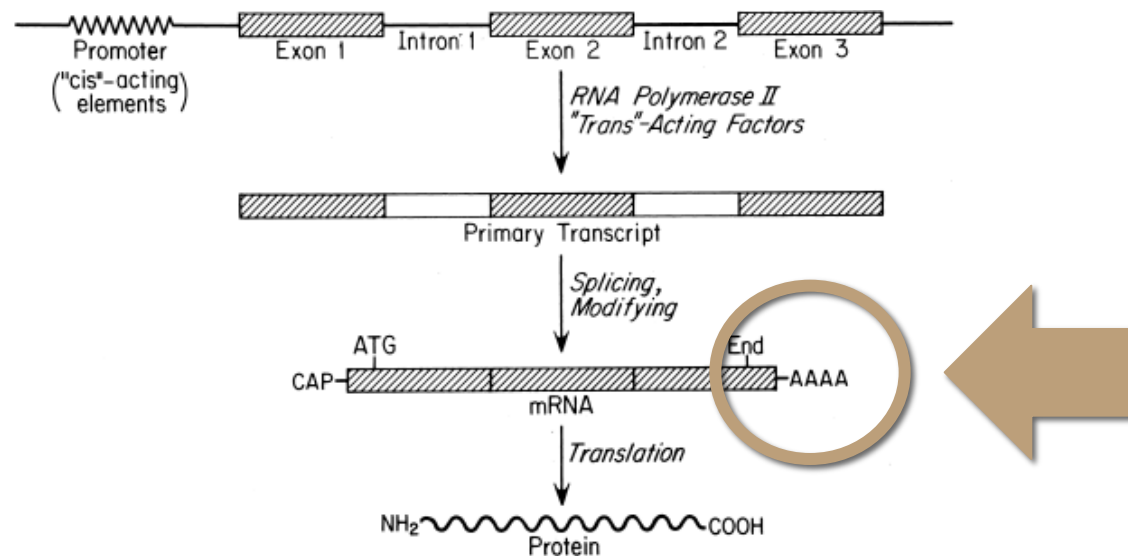
# BASICS OF RNASEQ

---

# Some background

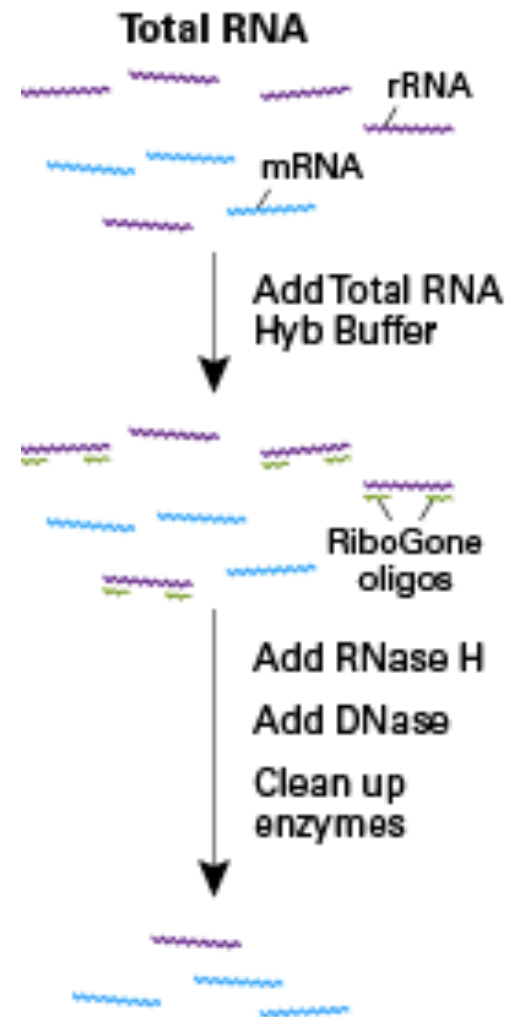
# Targeting mRNA for sequencing

- To target mRNA
  - **Poly-A enrichment** - purify the poly-A containing mRNA molecules using poly-T oligo attached magnetic beads
  - Only works for eukaryotes



# Remove rRNA

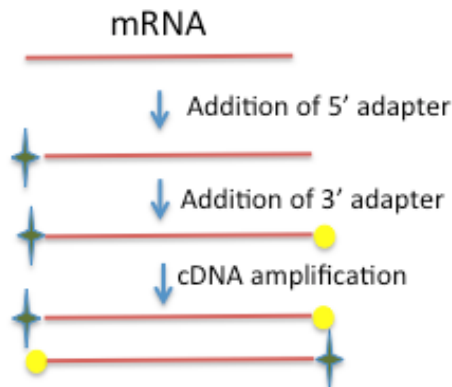
- Goal is to subtract rRNA, thus enriching for mRNA
- Hybridization/bead capture procedure that selectively binds target sequences using biotinylated capture probes
- This leaves other types of RNA, including non-coding types



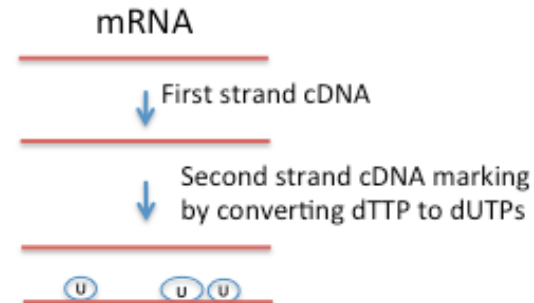
# Advantages of Strand Specific Sequencing

- Different protocols take different approaches, but all result in sequencing the original strand only (not the RC)
  - Ligation-based
  - dUTP-based
- Good for assembly and mapping
- Differentiate overlapping genes, pseudogenes, antisense transcripts
- Identifying the transcribed strand for non-coding RNAs

## Ligation- based method



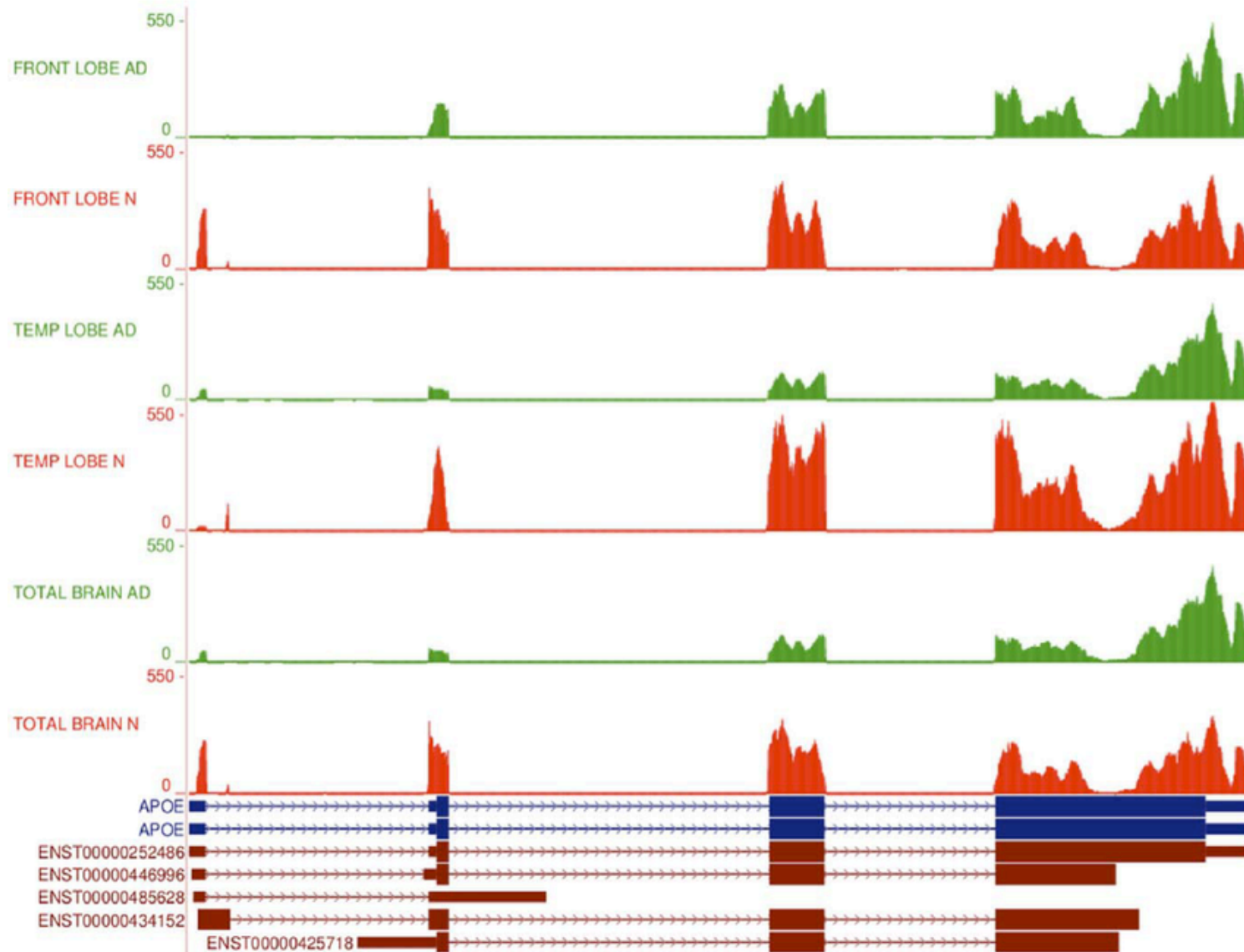
## dUTP second strand based method



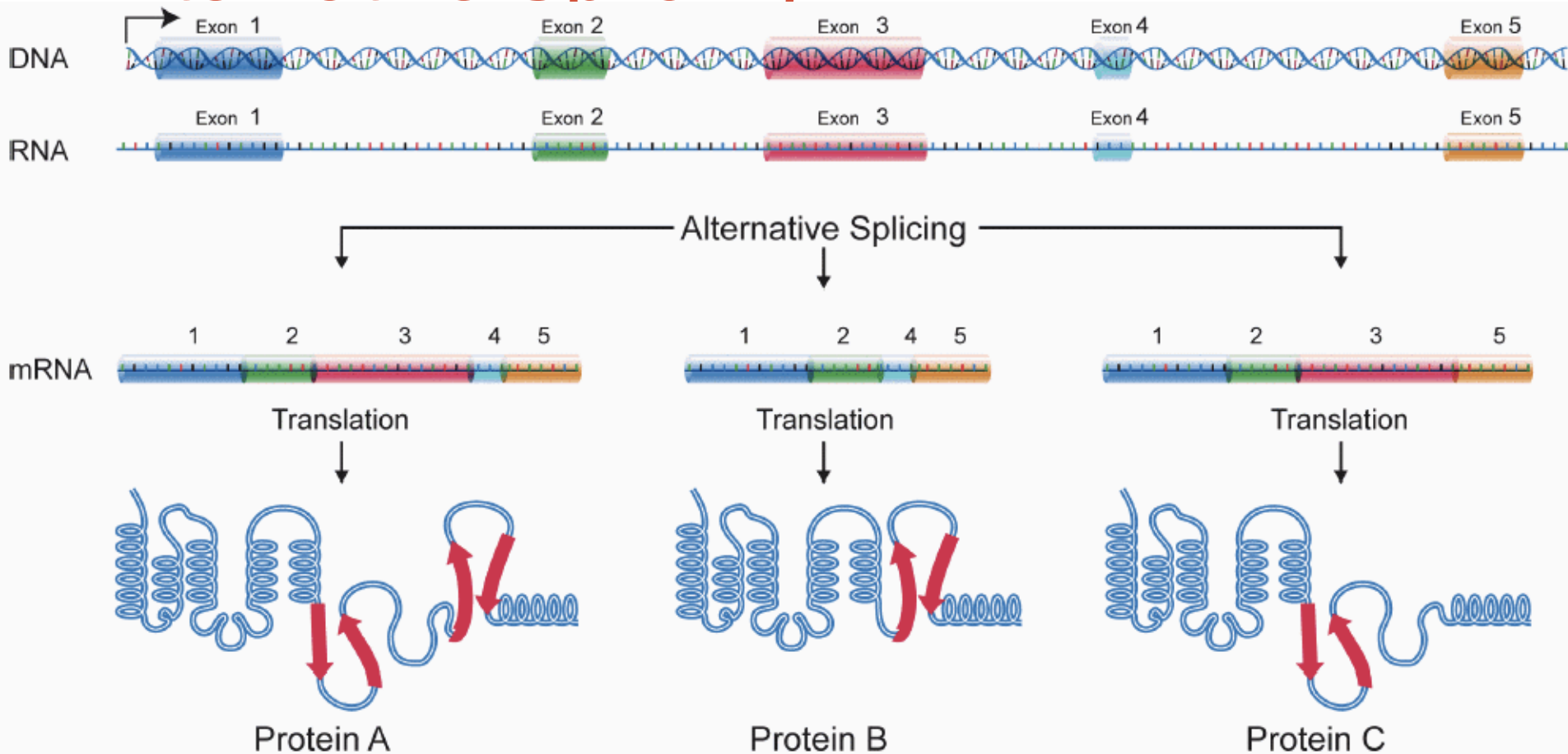
# Experimental Goals for mRNA Seq

- Catalog of genes
- Gene expression levels
- Differential gene expression levels
- All of the above for alleles and splice variants
- Annotating the genes in a reference genome
- Variant (Genetic marker) discovery
- Post-transcriptional modifications, RNA-editing

# Genome Annotation



# 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.

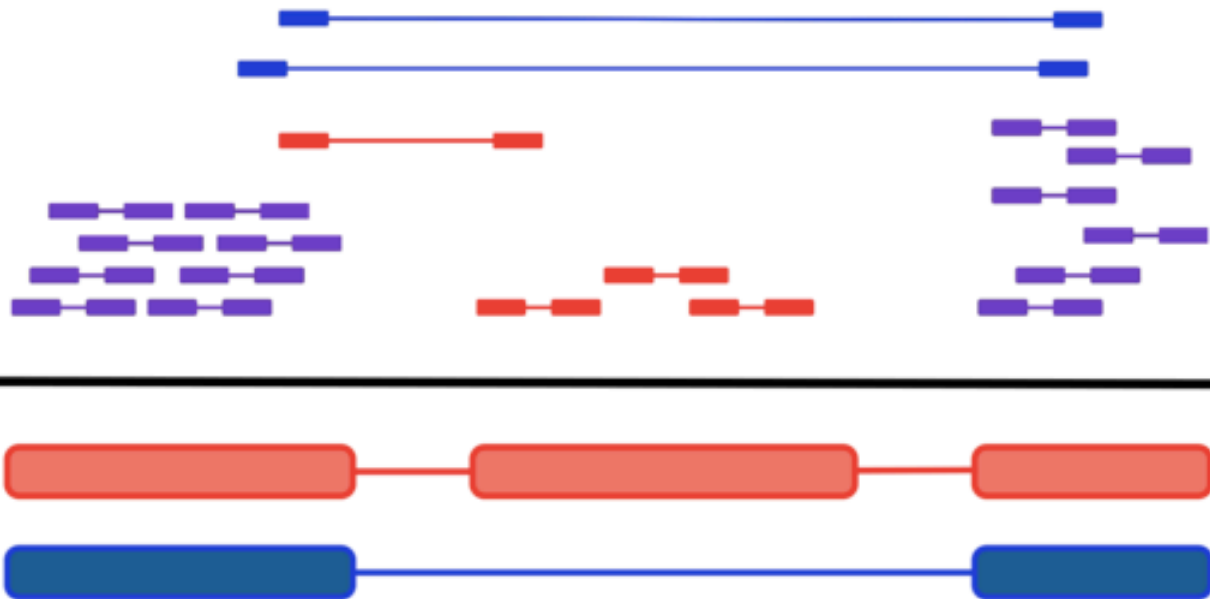
Indicate isoform B.

Aligned  
Fragments

Genome

Isoform A

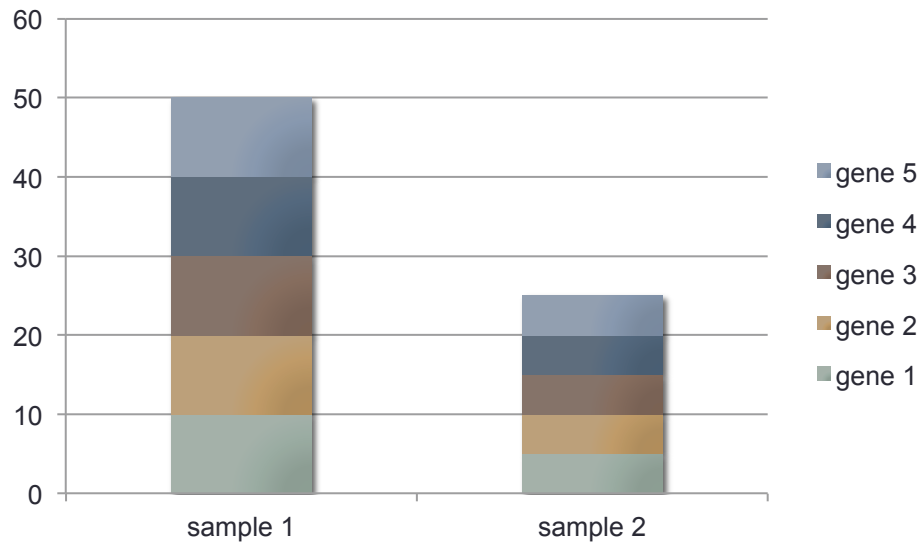
Isoform B



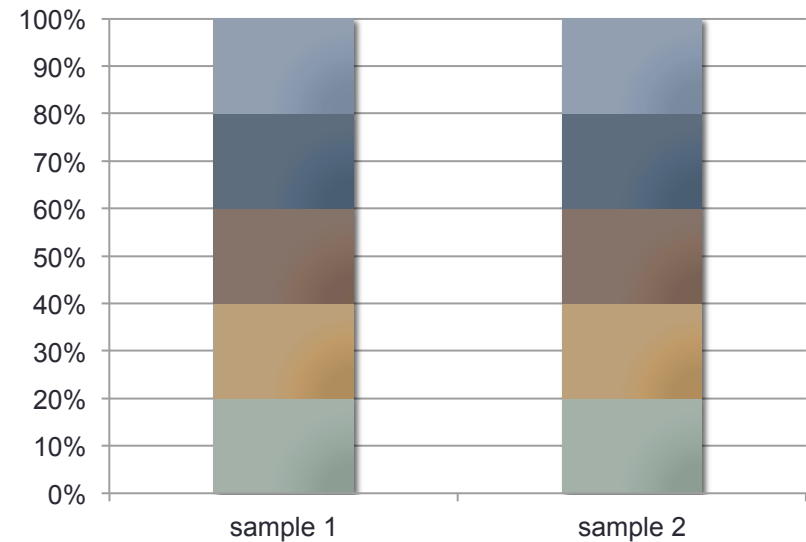
# Limitations

RNASeq gives you relative abundance only

Absolute Quantities



Relative Quantities



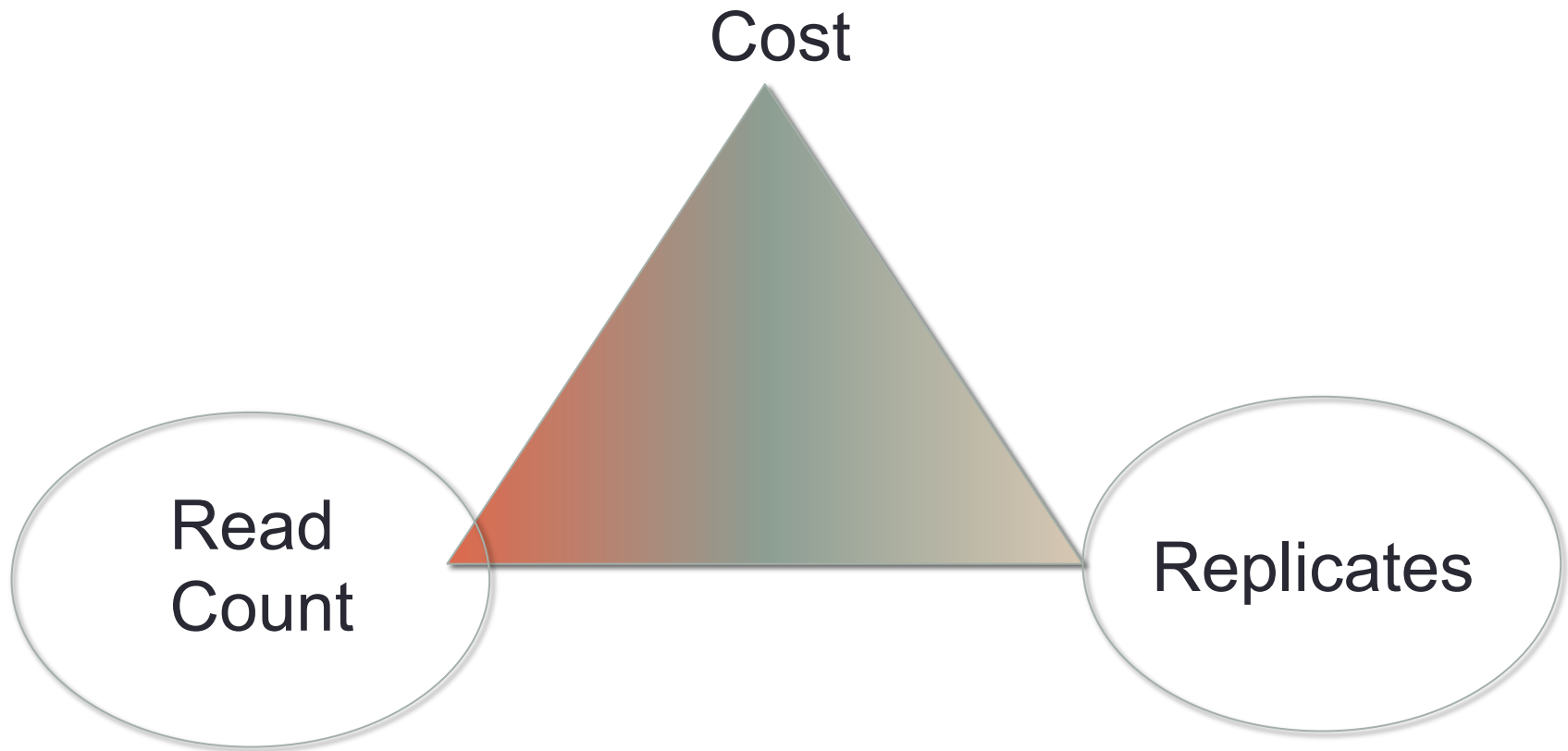
# Limitations

- Reverse transcription, PCR and fragmentation steps can introduce biases
  - depletion of reads at both 5' and 3' ends
    - Difficult to identify the true start and end of novel transcripts
    - May underestimate expression level of short genes
  - GC bias, length bias
- PCR-free preps are available

# RNASEQ PROJECT DESIGN

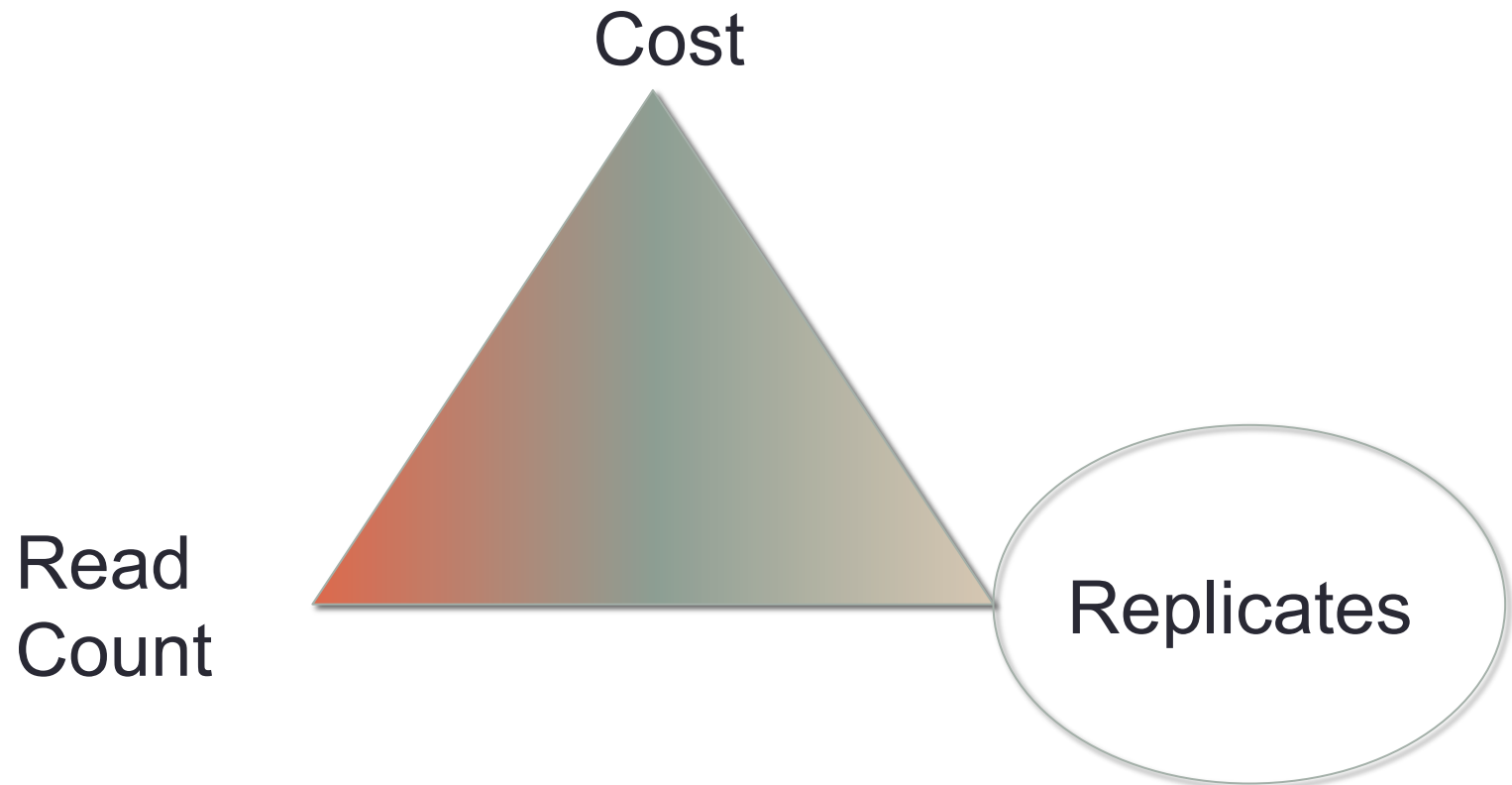
---

# Major Considerations for DEG Project Design



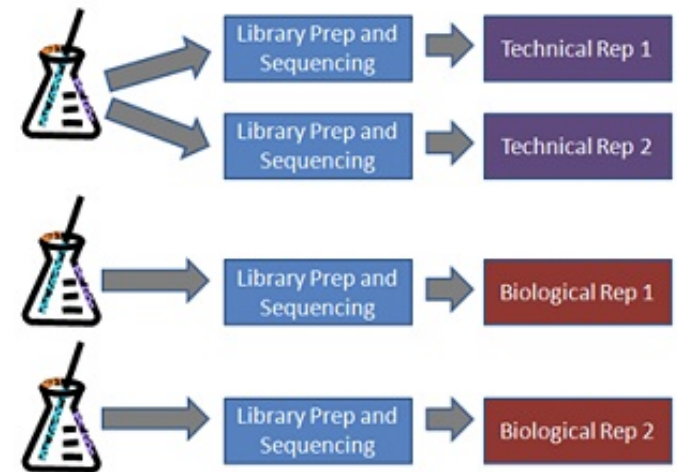
Who is your resident statistician and/or bioinformatician? Buy them a coffee and make friends. **Preferably before starting the experiment!**

# Major Considerations for DEG Project Design



# 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 comparison with gene expression arrays. *Genome Research* 18: 1500-1517

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

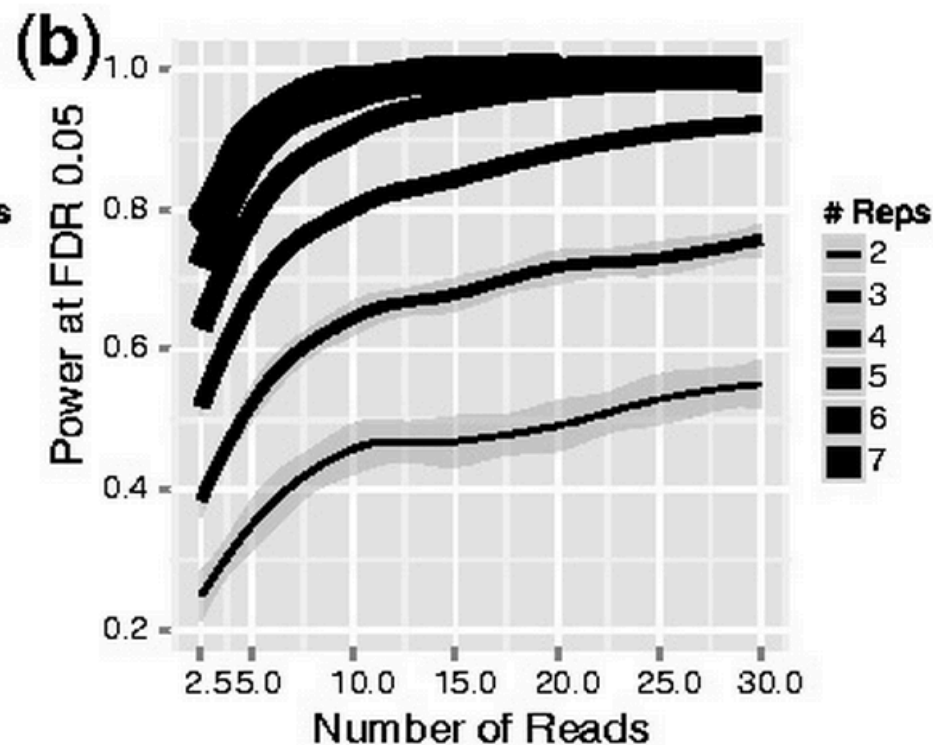
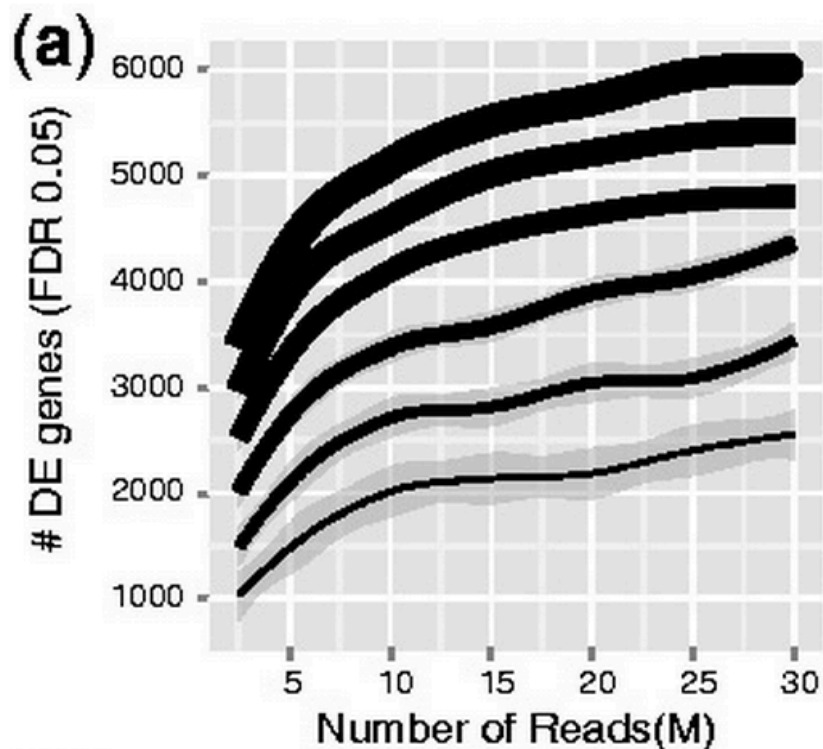
“With three biological replicates, nine of the 11 tools evaluated found only 20%–40% of the significantly differentially expressed (SDE) genes identified with the full set of 42 clean replicates.”

“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 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.”

From our lab: If you have no biological replicates, DESeq(2) probably won't give you any results, you can give EdgeR a try.

# 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

# Lane effects, a cautionary tale

(Lane effects are systematically bad sequencing cycles and errors in base calling)

Original PNAS paper:


Comparison of the transcriptional landscapes between human and mouse tissues

Shin Lin<sup>a,b,1</sup>, Yiing Lin<sup>c,1</sup>, Joseph R. Nery<sup>d</sup>, Mark A. Urich<sup>d</sup>, Alessandra Breschi<sup>e,f</sup>, Carrie A. Davis<sup>g</sup>,

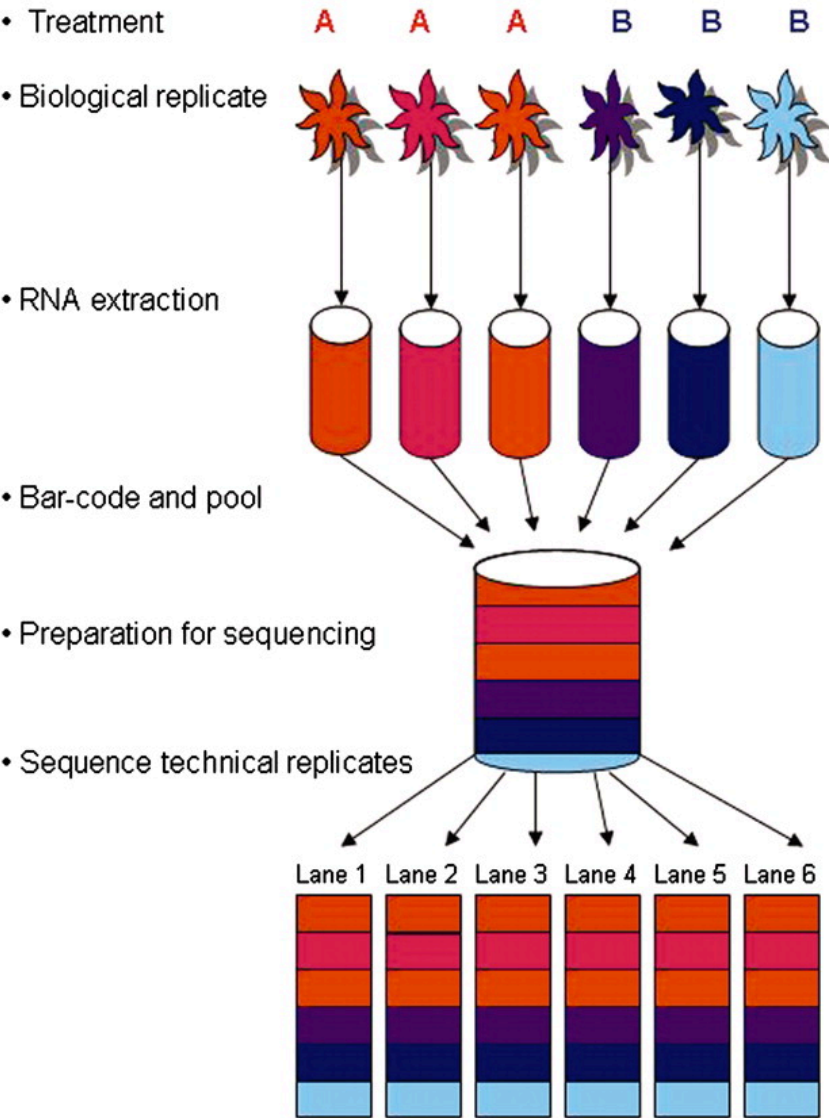
Reanalysis pointing out flawed statistical design and questioning results:

“Here we show that the Mouse ENCODE gene expression data were collected using a flawed study design, which confounded sequencing batch (namely, the assignment of samples to sequencing flowcells and lanes) with species.

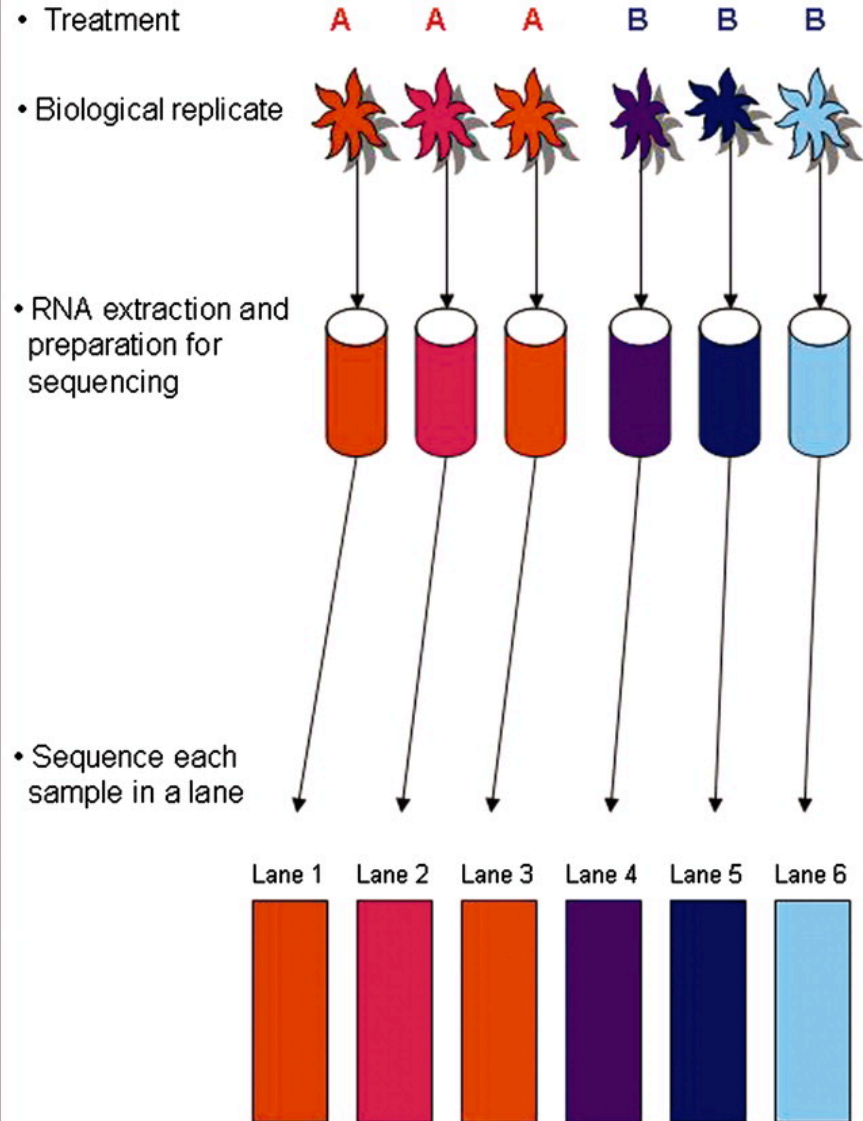
A reanalysis of mouse ENCODE comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]

 Yoav Gilad, Orna Mizrahi-Man

## Balanced Blocked Design



## Confounded Design



# Major Considerations for Project Design

Cost



Read  
Count

Replicates



# 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)

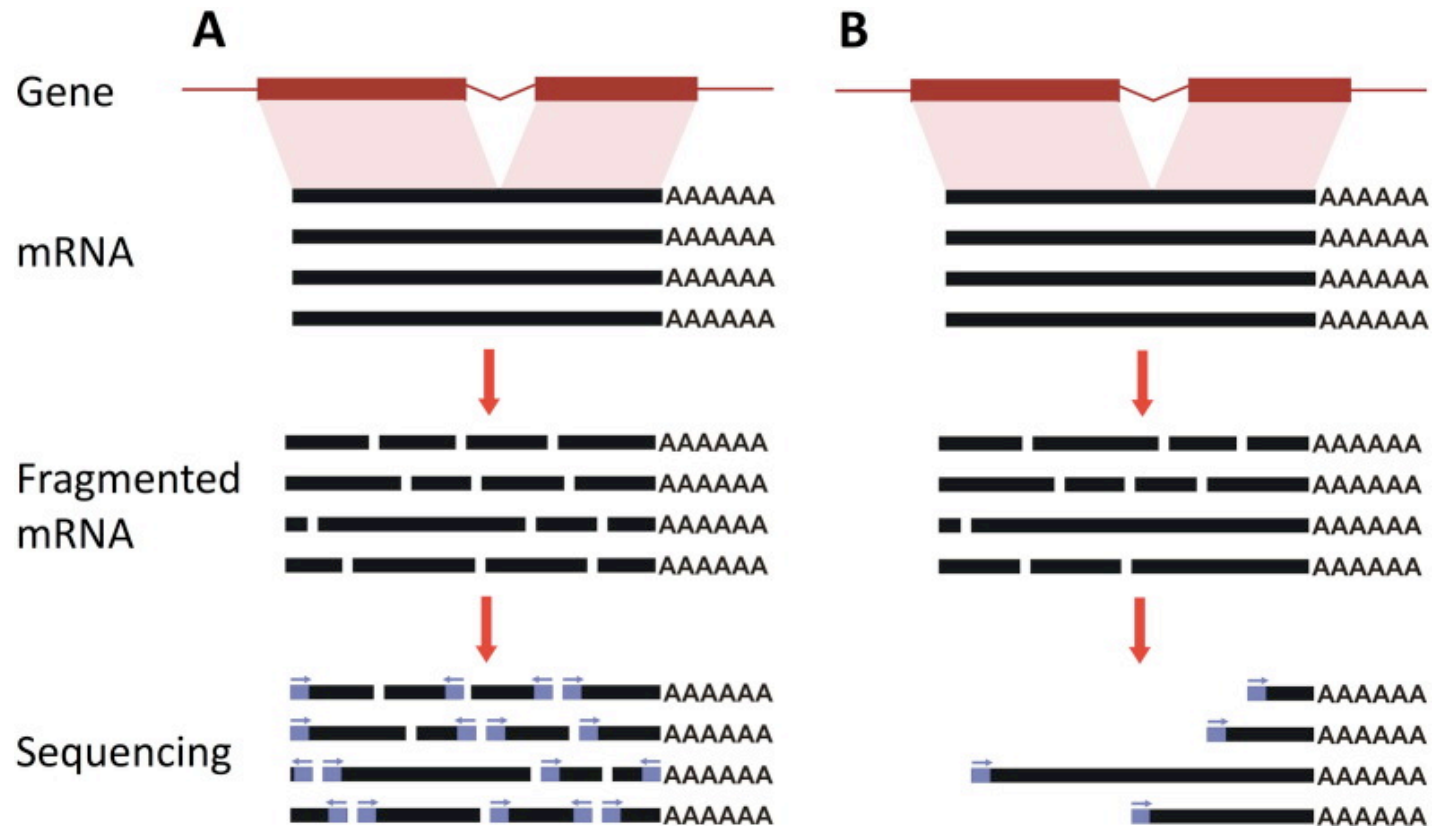
# Read Count - How to Decide?

- 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

# 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

# LAB

- A standard protocol for differential gene expression.

# mRNA Data Analysis Pipeline

Quality Assessment

FastQC



Babraham Bioinformatics

Trimming

Trimmomatic

Quality Assessment

FastQC



Babraham Bioinformatics

Mapping to a Reference

STAR

Visualization



Integrative  
Genomics  
Viewer

Counting reads per gene

HTSeq



Bioconductor  
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Differential Gene Expression

DESeq2

GO Term Enrichment

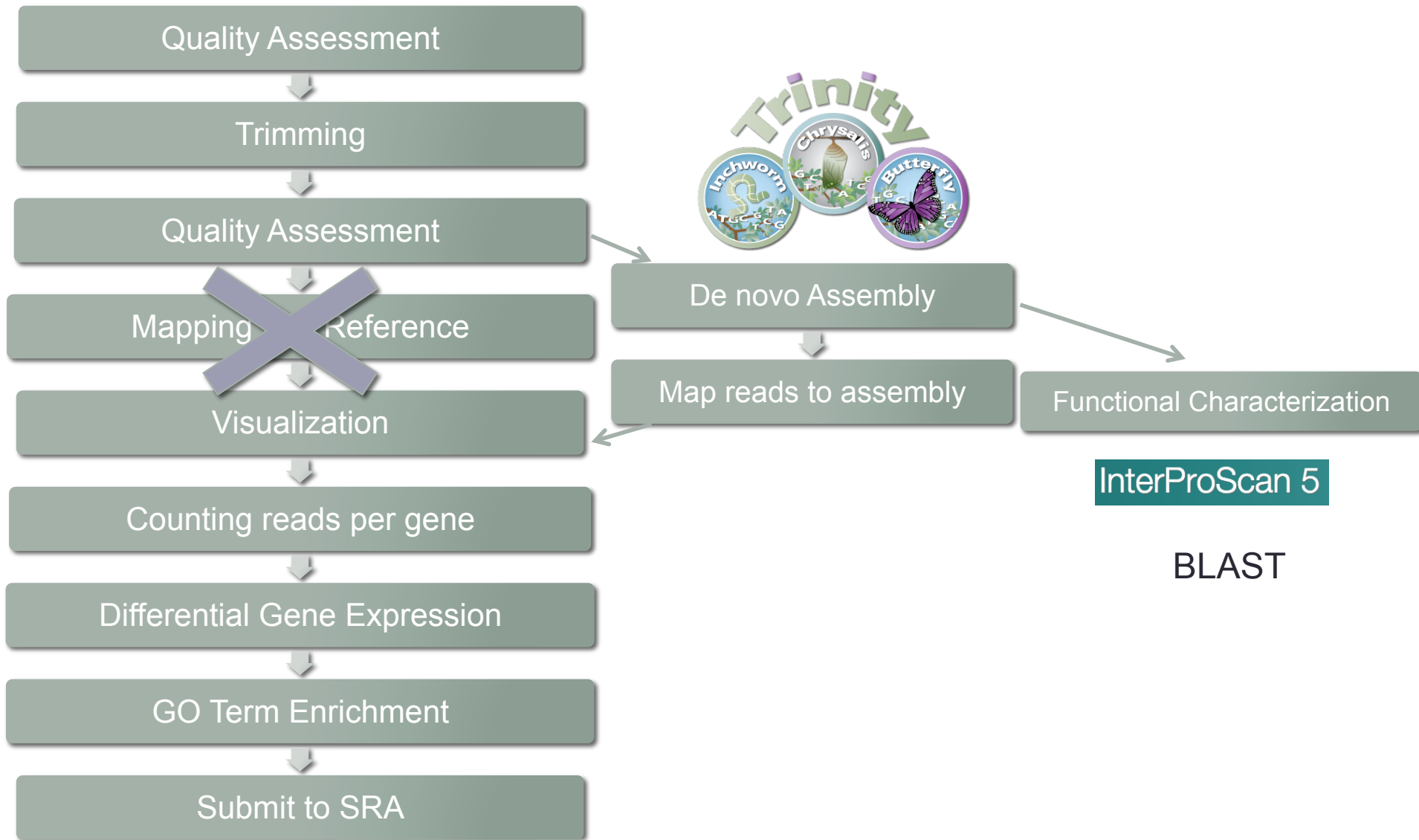


BINGO

Submit to SRA



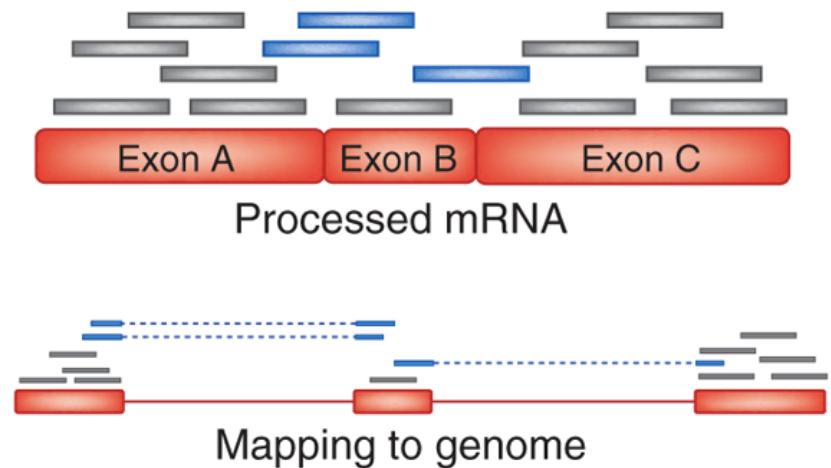
# What if you don't have a reference?





# Mapping to the Reference

- Mapping RNA to a eukaryotic genome is more complicated than mapping DNA
  - Introns
  - Alternative splicing
- Usually, you want to use a mapping software designed for RNASeq
  - The software will use a file (gff3) to know where the genes are located
  - Many RNASeq mapping software packages will also infer gene structures (This is good for identifying novel genes and isoforms)



# GFF- Generic Feature Format

- GFF was the original file format
- Represent genomic features on a sequence
  - gene on a chromosome
- But it did not cover all the use cases needed. Eventually different groups chose to extend it in their own custom ways, and multiple new formats then became common, confusing everyone.

<http://www.sequenceontology.org/gff3.shtml>



The screenshot shows the Sequence Ontology Project website. The header features the 'SO' logo and the text 'The Sequence Ontology Project'. Below this is a green navigation bar with links: Home, Browser, Wiki, GFF3, GVF, Resources, Software, About, Request A Term, and Site Map. The breadcrumb trail reads 'Home > Resources > GFF3'. The main content area is titled 'Generic Feature Format Version 3 (GFF3)' and includes a 'Summary' section with the author 'Lincoln Stein', the date '26 February 2013', and the version '1.21'. A 'News' section on the right mentions that 'October 2013 GVF was used in the clinical annotation of a whole genome, for precision medicine. Integrating'.

SO The Sequence Ontology Project

Home Browser Wiki GFF3 GVF Resources Software About Request A Term Site Map

Home > Resources > GFF3

## Generic Feature Format Version 3 (GFF3)

### Summary

Author: Lincoln Stein  
Date: 26 February 2013  
Version: 1.21

### News

► **October 2013** GVF was used in the clinical annotation of a whole genome, for precision medicine. Integrating

# GFF3

## Generic Feature Format Version 3

- Gff3 format is an attempt to:
  - add and standardize the most common extensions to gff
  - preserve backward compatibility to gff
- Basics:
  - 9 columns
  - Tab delimited
  - Plain text

Backward compatibility - Maintaining compatibility with earlier models or versions of the same product. A new version of a program is said to be backward compatible if it can use files and data created with an older version of the same program.

Chr1	.	gene	301	2169	.	+	.	ID=SPAC1F7.08;Name=iron%20transport%20multi..
------	---	------	-----	------	---	---	---	---

Column 1: "seqid"

Column 2: "source"

Column 3: "type"

Columns 4 & 5: "start" and "end"

Column 6: "score"

Column 7: "strand"

Column 8: "phase"

Column 9: "attributes"

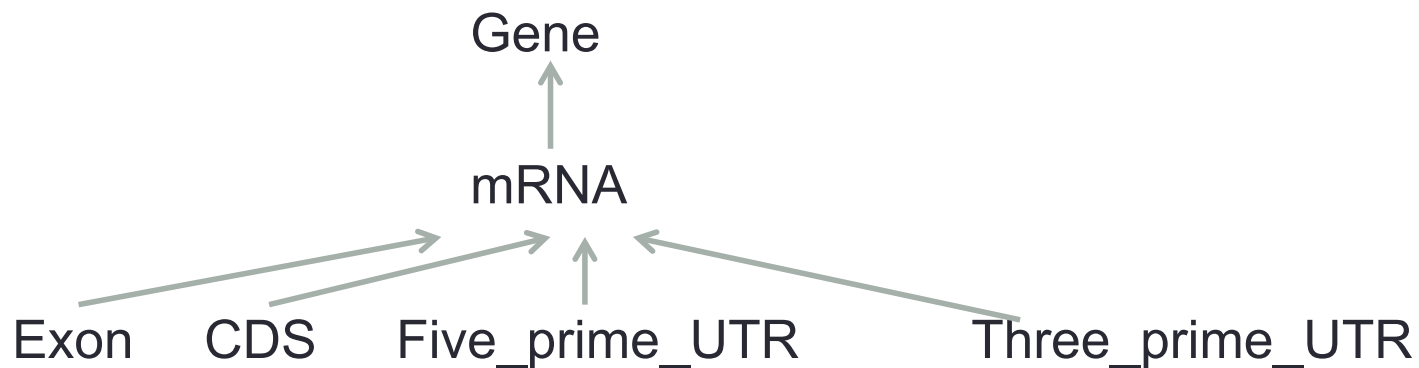
A list of feature attributes in the format tag=value. Multiple tag=value pairs are separated by semicolons

ID= must be unique

genome	.	mRNA	3012169	.	+	.	ID=m.SPAC1F7.08;Parent=SPAC1F7.08;Name=iron...
--------	---	------	---------	---	---	---	--

Parent=

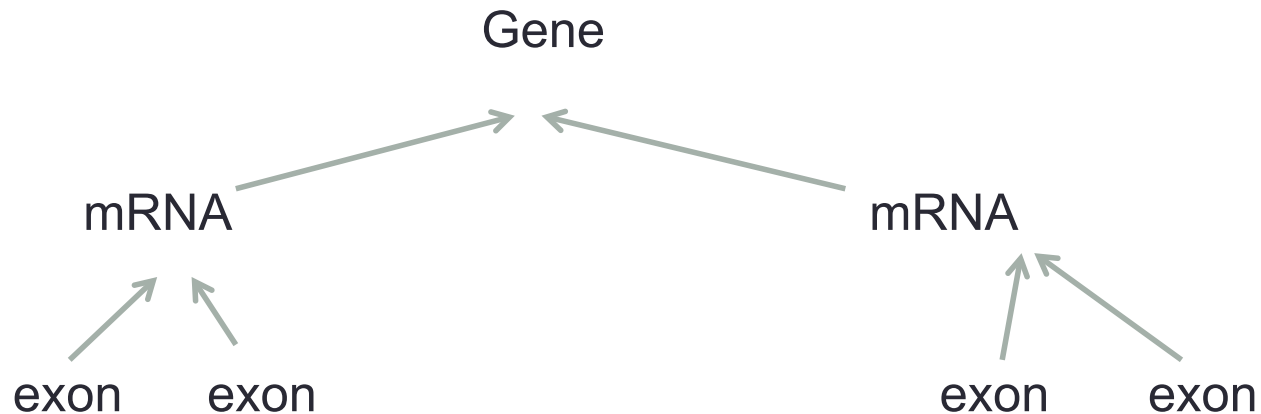
Hierarchy of gene pieces



# GFF3

## Generic Feature Format Version 3

A feature can have many “children”, allowing for isoforms to be represented as well.



# GFF3 – Alternative Isoforms

```
ctg123 example gene          1050 9000 . + . ID=EDEN;Name=EDEN;Note=protein kinase

ctg123 example mRNA          1050 9000 . + . ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1
ctg123 example five_prime_UTR 1050 1200 . + . Parent=EDEN.1
ctg123 example CDS           1201 1500 . + 0 Parent=EDEN.1
ctg123 example CDS           3000 3902 . + 0 Parent=EDEN.1
ctg123 example CDS           5000 5500 . + 0 Parent=EDEN.1
ctg123 example CDS           7000 7608 . + 0 Parent=EDEN.1
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.1

ctg123 example mRNA          1050 9000 . + . ID=EDEN.2;Parent=EDEN;Name=EDEN.2;Index=1
ctg123 example five_prime_UTR 1050 1200 . + . Parent=EDEN.2
ctg123 example CDS           1201 1500 . + 0 Parent=EDEN.2
ctg123 example CDS           5000 5500 . + 0 Parent=EDEN.2
ctg123 example CDS           7000 7608 . + 0 Parent=EDEN.2
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.2

ctg123 example mRNA          1300 9000 . + . ID=EDEN.3;Parent=EDEN;Name=EDEN.3;Index=1
ctg123 example five_prime_UTR 1300 1500 . + . Parent=EDEN.3
ctg123 example five_prime_UTR 3000 3300 . + . Parent=EDEN.3
ctg123 example CDS           3301 3902 . + 0 Parent=EDEN.3
ctg123 example CDS           5000 5500 . + 1 Parent=EDEN.3
ctg123 example CDS           7000 7600 . + 1 Parent=EDEN.3
ctg123 example three_prime_UTR 7601 9000 . + . Parent=EDEN.3
```