RNASEQ PROJECT DESIGN

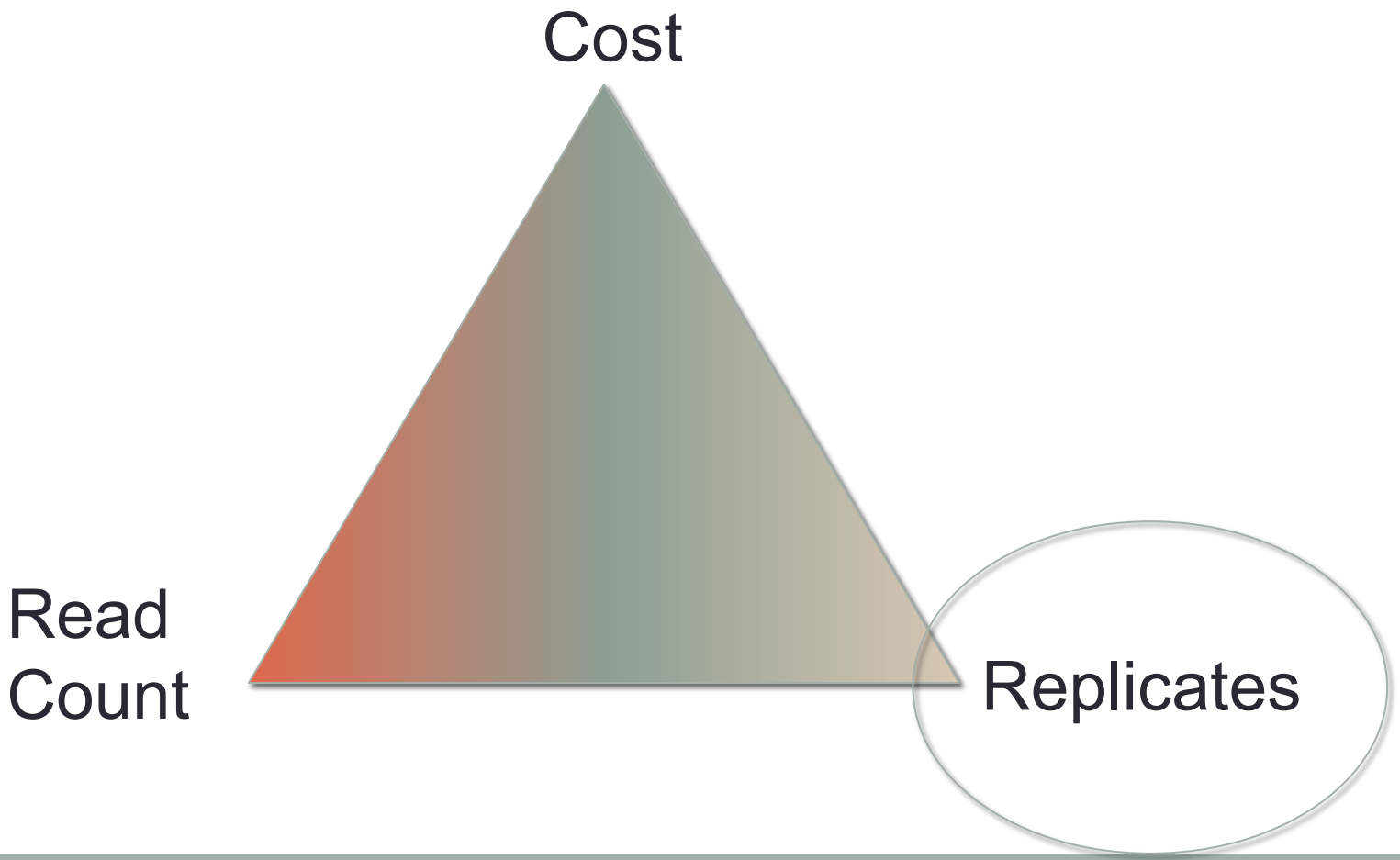
Experimental Design

Assembly in Non-Model Organisms

And other (hopefully useful) Stuff

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

Major Considerations for Project Design

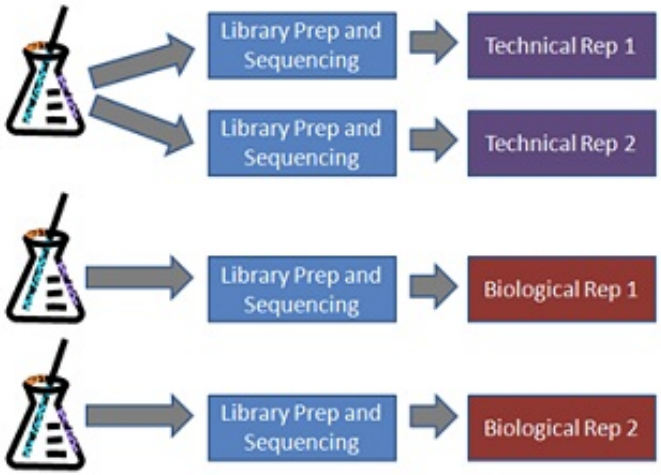


Pro Tip: Who is your resident statistician? Buy them a coffee and make friends.

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!
- Different genes have different variances and are potentially subject to different errors and biases.



"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?

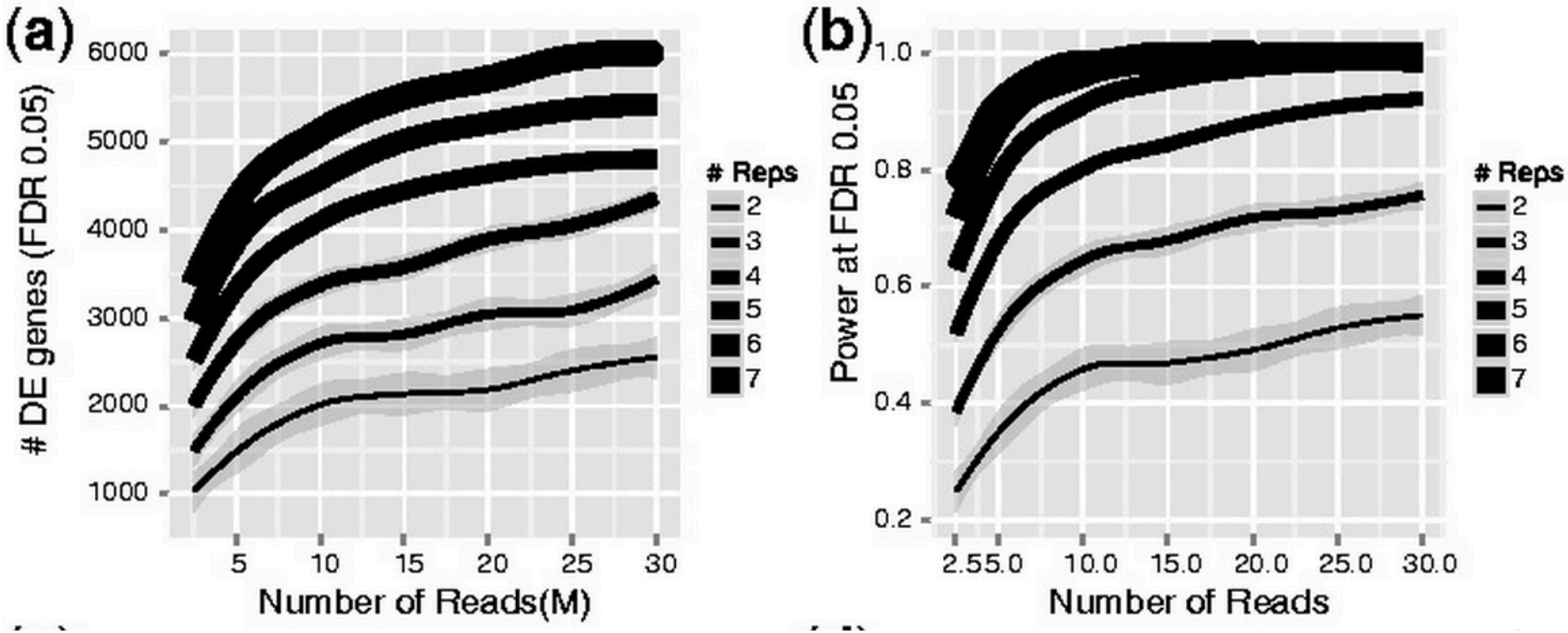
- beyond a depth of 10 million reads, replicates provide more statistical power than depth for detecting differential gene expression

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.

- At least 6 according to Schurch et al., 2016
- Others say many more!

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

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

Pooling

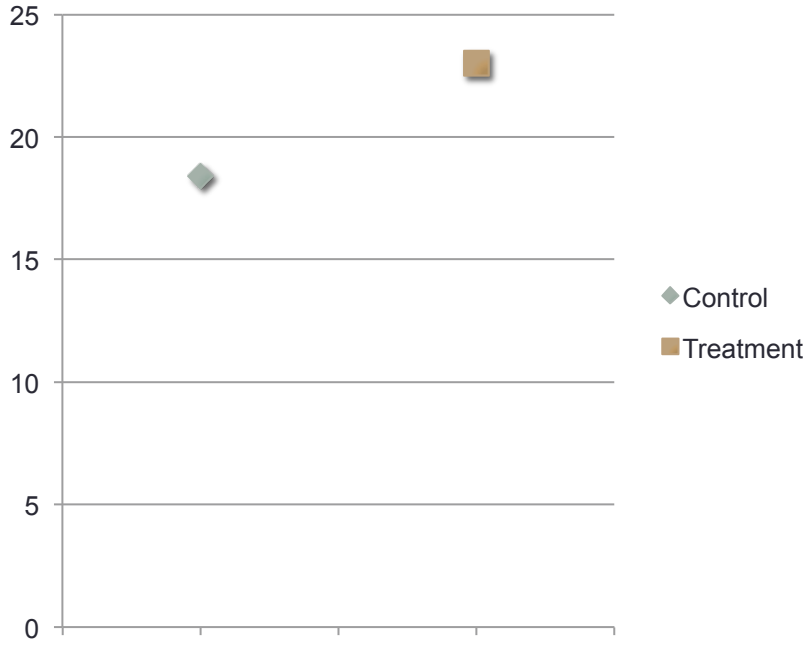
Does pooling my samples count as biological replicates?

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

Experiment

Values from control replicates:
10, 13, 16, 20, 23, 23, 24
Average: 18.4

Values from treatment replicates:
16, 19, 22, 24, 25, 27, 28
Average: 23

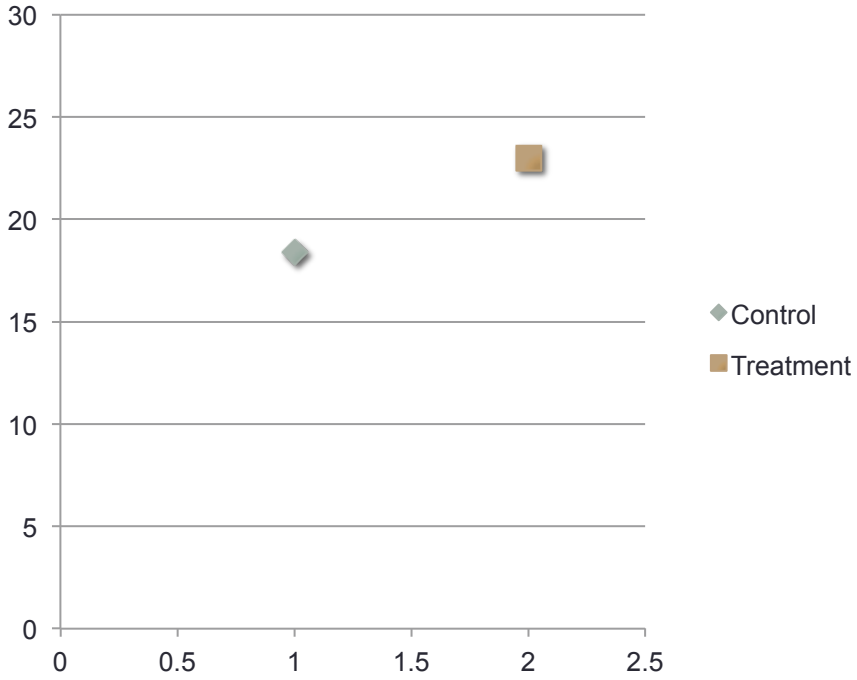


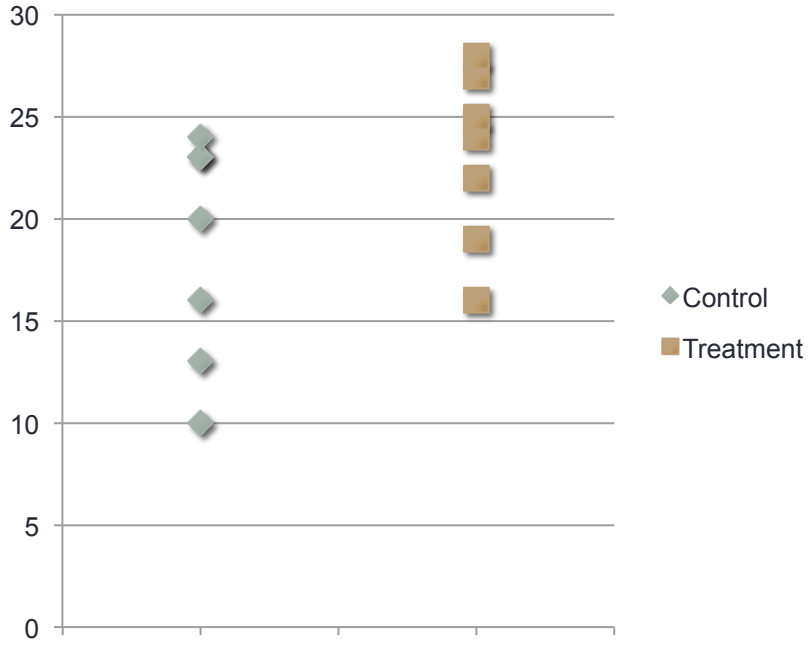
Pooling

Does pooling my samples count as biological replicates?

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

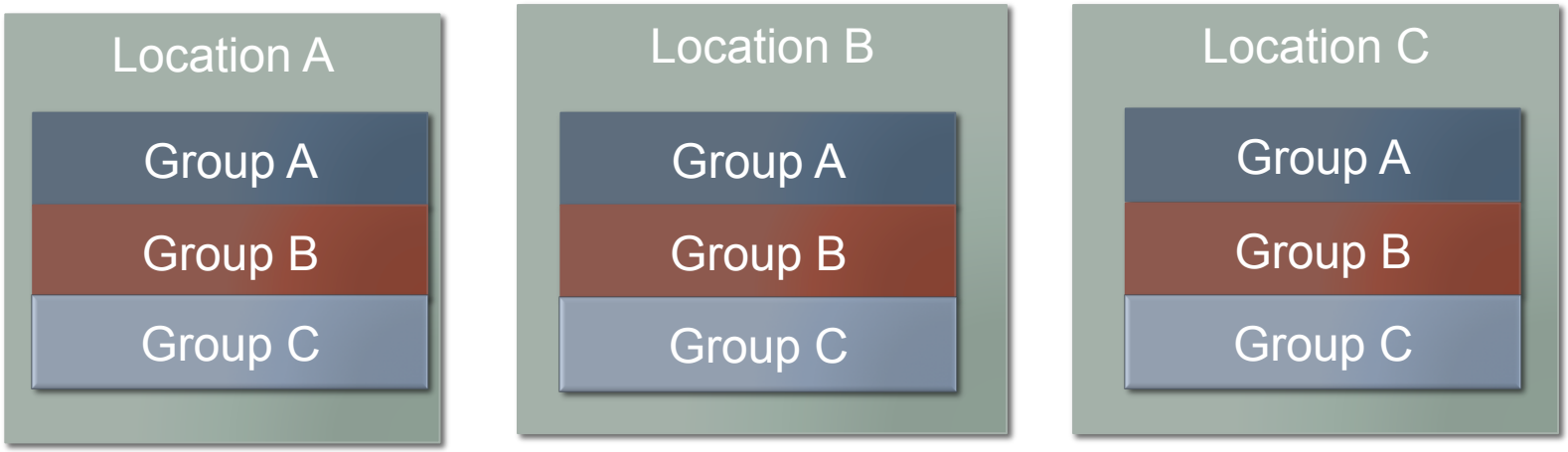
Perform a t-test – this is NOT statistically significant.





Blocking

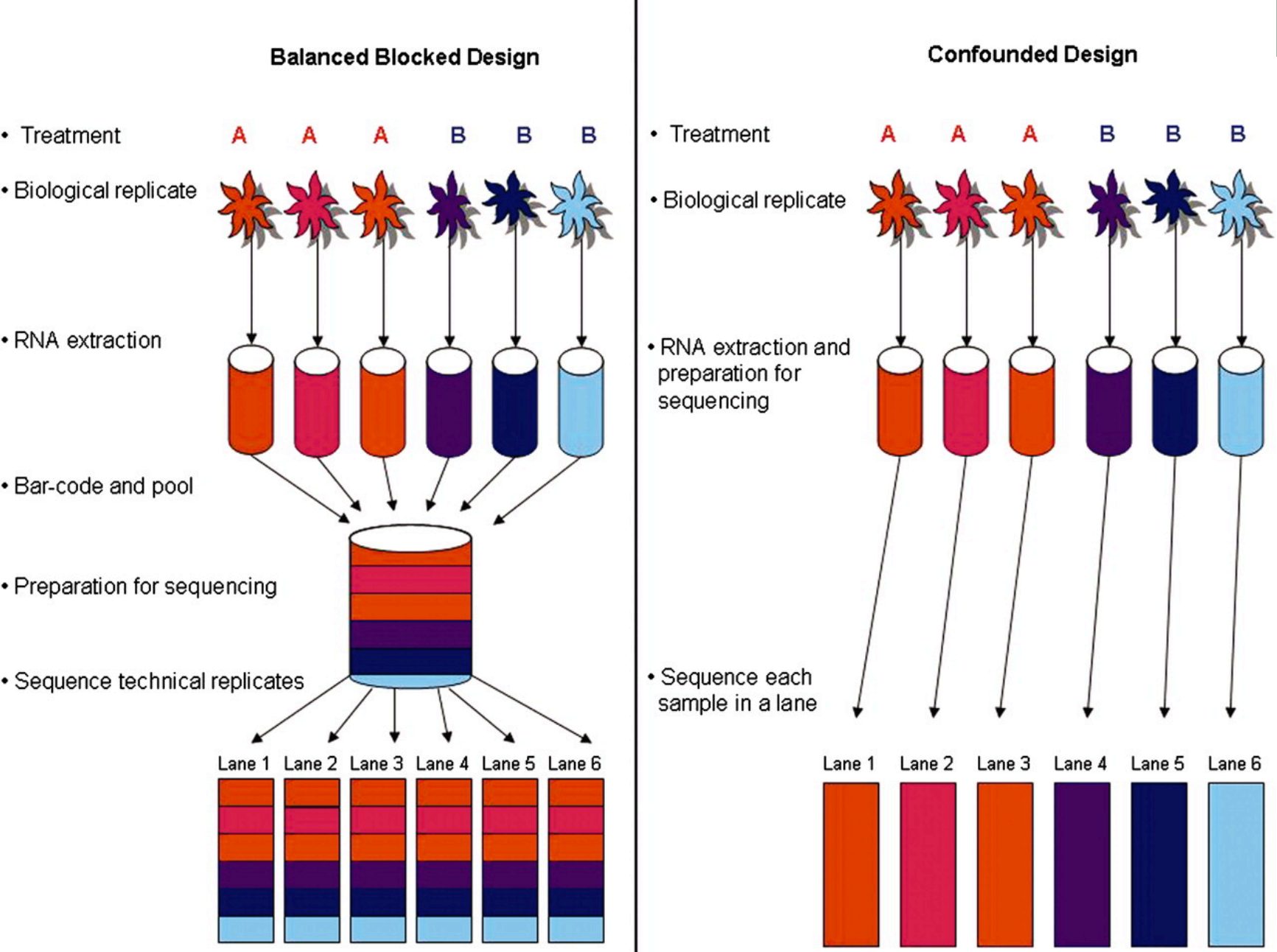
- Complete Randomized Block Design
- Randomize - assigning individuals at random to treatments in an experiment
- Blocking - Experimental units are grouped into homogeneous clusters in an attempt to improve the comparison of treatments



Blocking

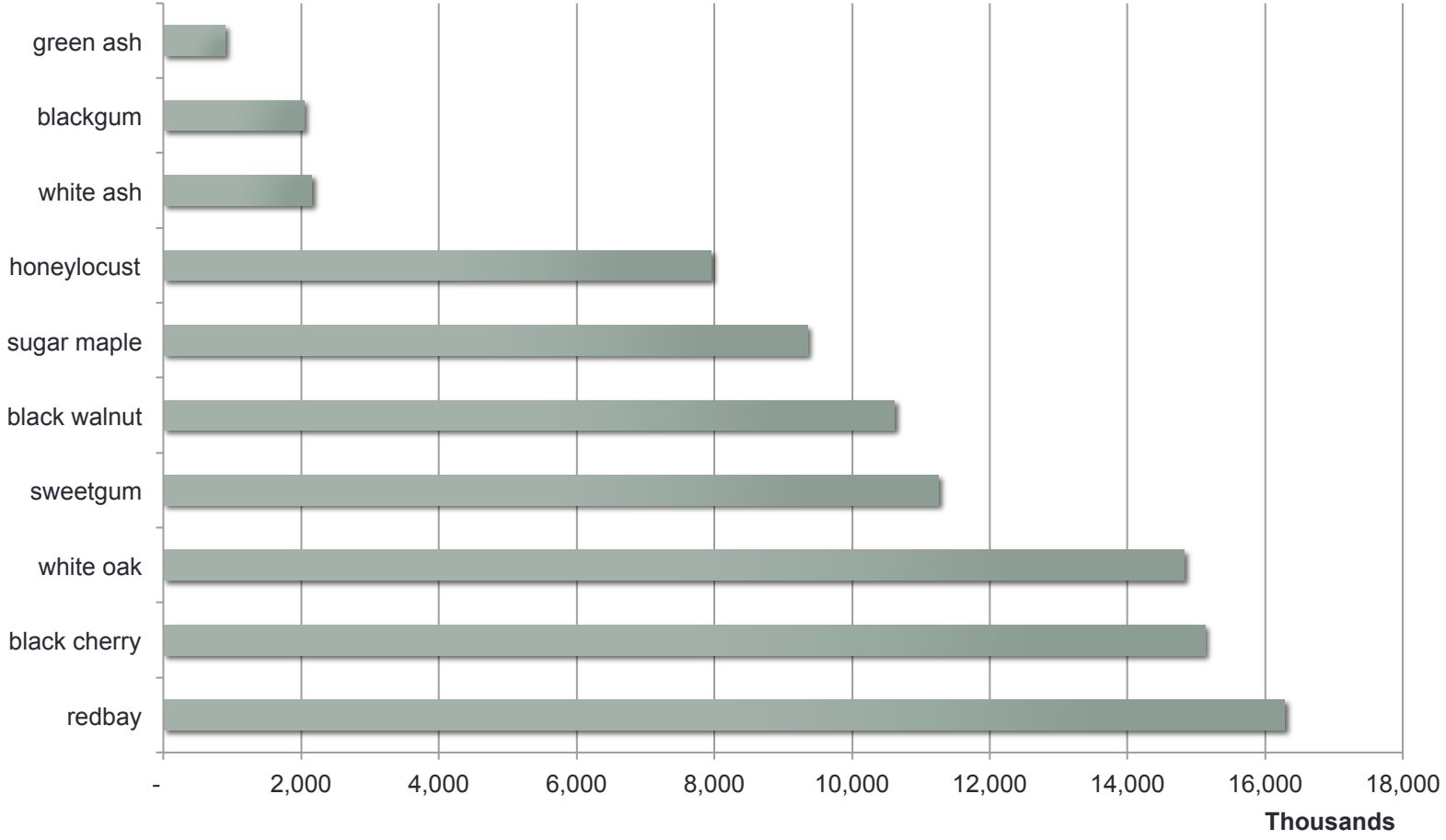
- Need to remember statistical design during laboratory procedures as well
- Lane effects
  - systematically bad sequencing cycles and errors in base calling

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



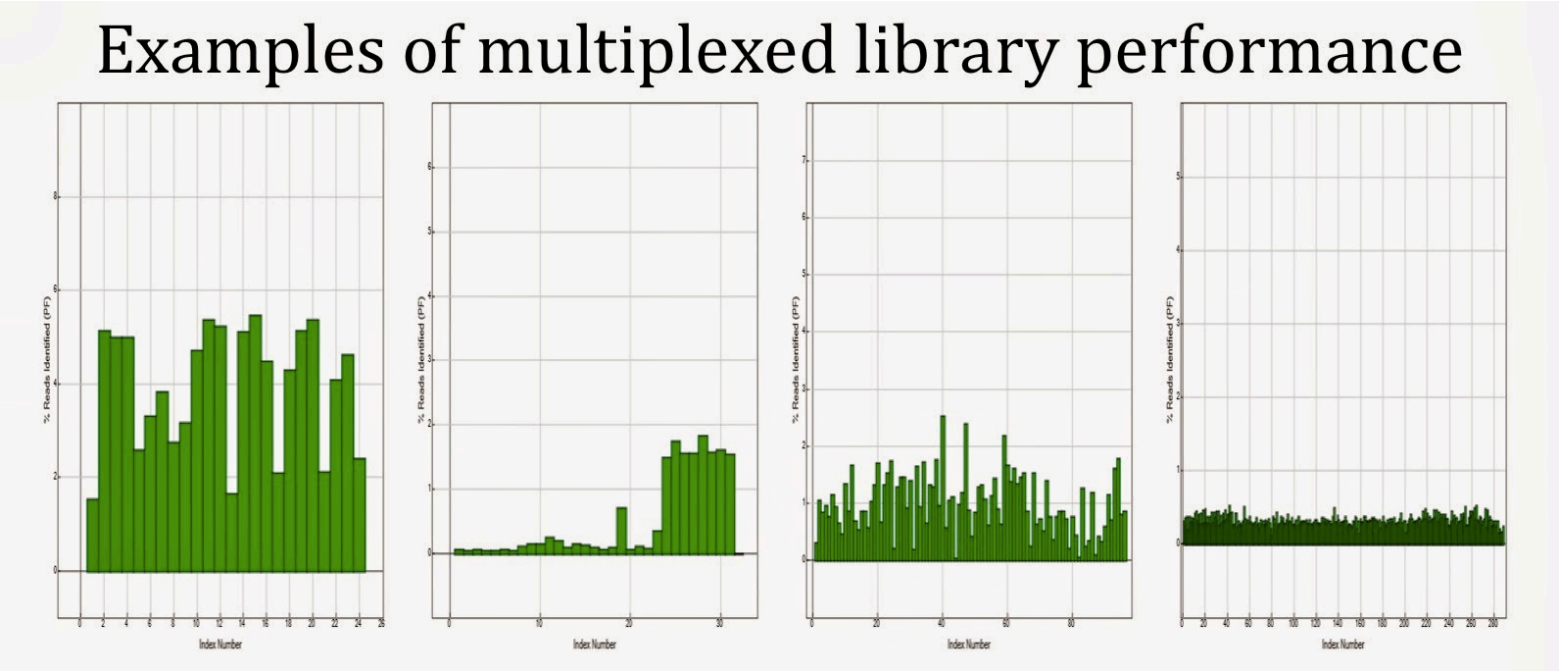
Auer and Doge Statistical Design and Analysis of RNA Sequencing Data Genetics June 2010 vol. 185 no. 2 405-416

Problem: Variation in # of Sequences per Library



Problem: Variation in # of Sequences per Library

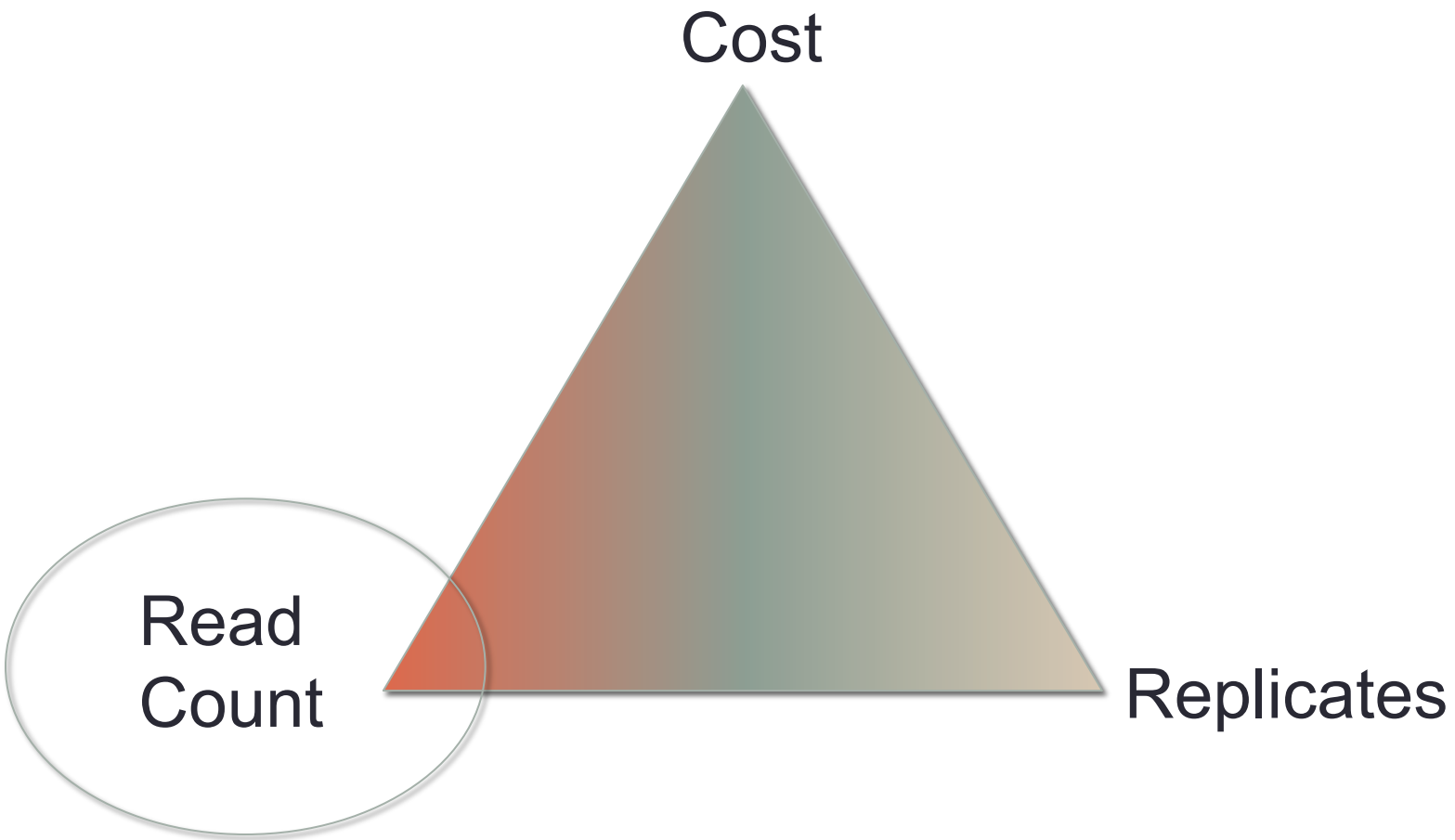
Another example from someone else….



http://core-genomics.blogspot.com/2013/10/how-good-is-your-ngs-multiplexing.html

What's the solution?
Robust library quantification.

Major Considerations for Project Design



Pro Tip: Who is your resident statistician? Buy them a coffee and make friends.

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?

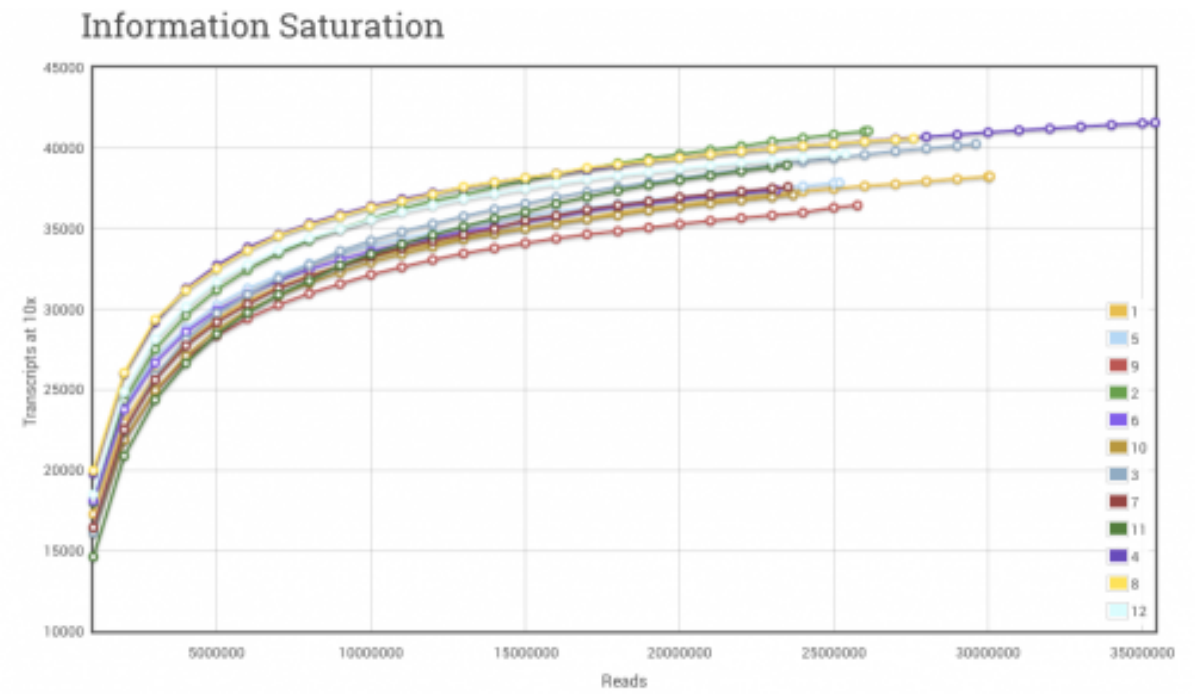
- If you are working in human, mouse or yeast, the work has been done.

- If not…
- If you have to choose between depth and replicates, choose more replicates
- What is being published in your community?
- What resources do you already have?
  - Well assembled and annotated genomes – single ends, shorter reads
  - De novo – longer reads, paired ends

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

How to know if you've sampled everything?

- New Discovery Rate – Explore with a Saturation curve
  - How many new genes are being discovered with each additional slice of data?



https://cofactorgenomics.com/1-graph-will-give-you-new-perspective-sequencing-experiment/