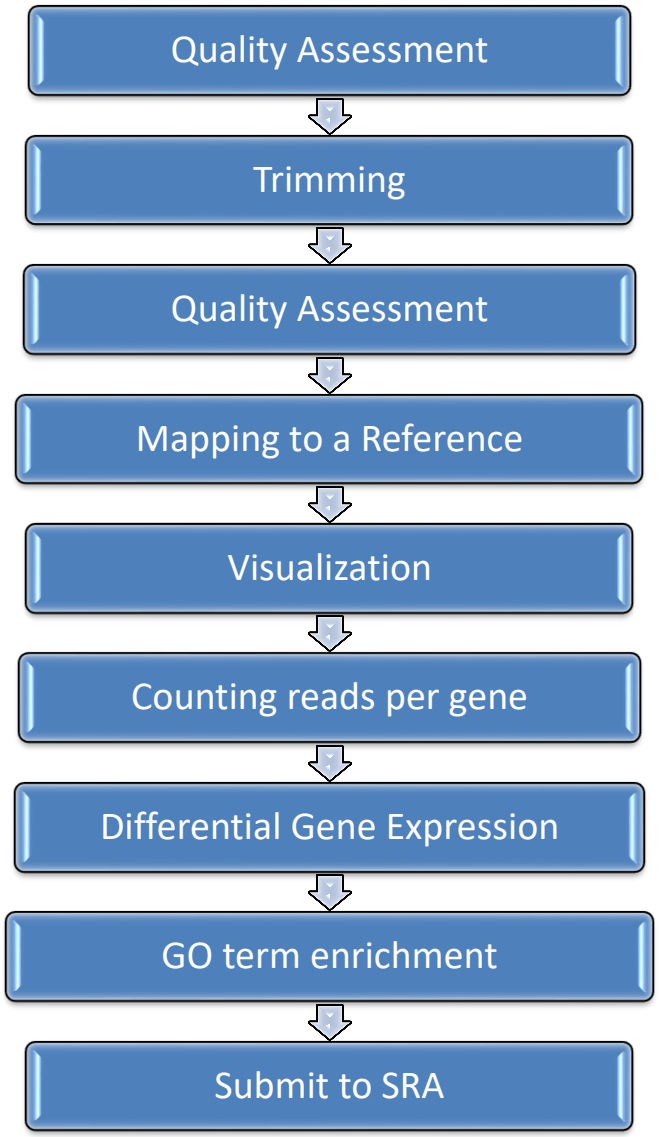
RNASeq Data Analysis Pipeline



FastQC

Babraham Bioinformatics

Skewer

MultiQC

STAR

igv Integrative Genomics Viewer

HTSeq

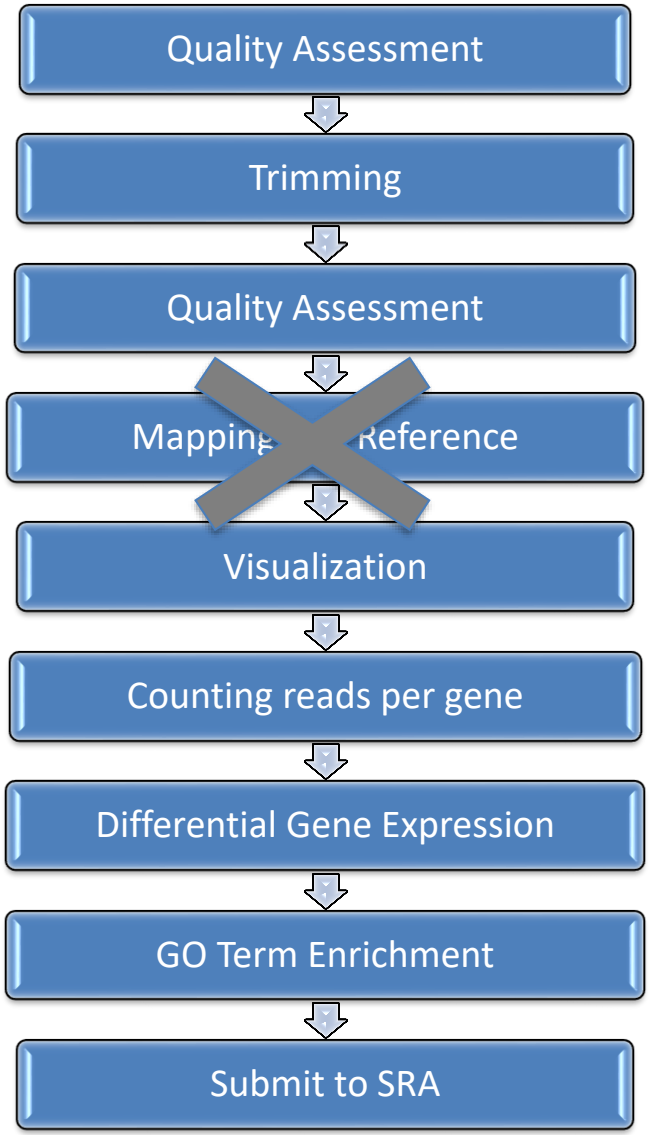
DESeq2

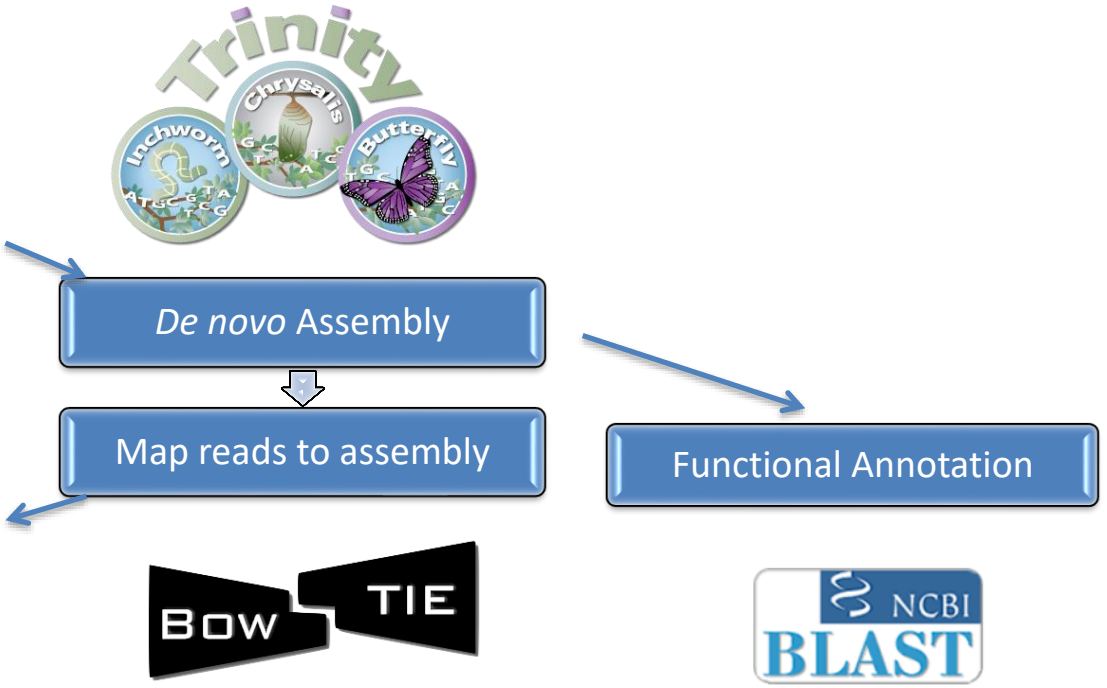
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

GENEONTOLOGY
Unifying Biology

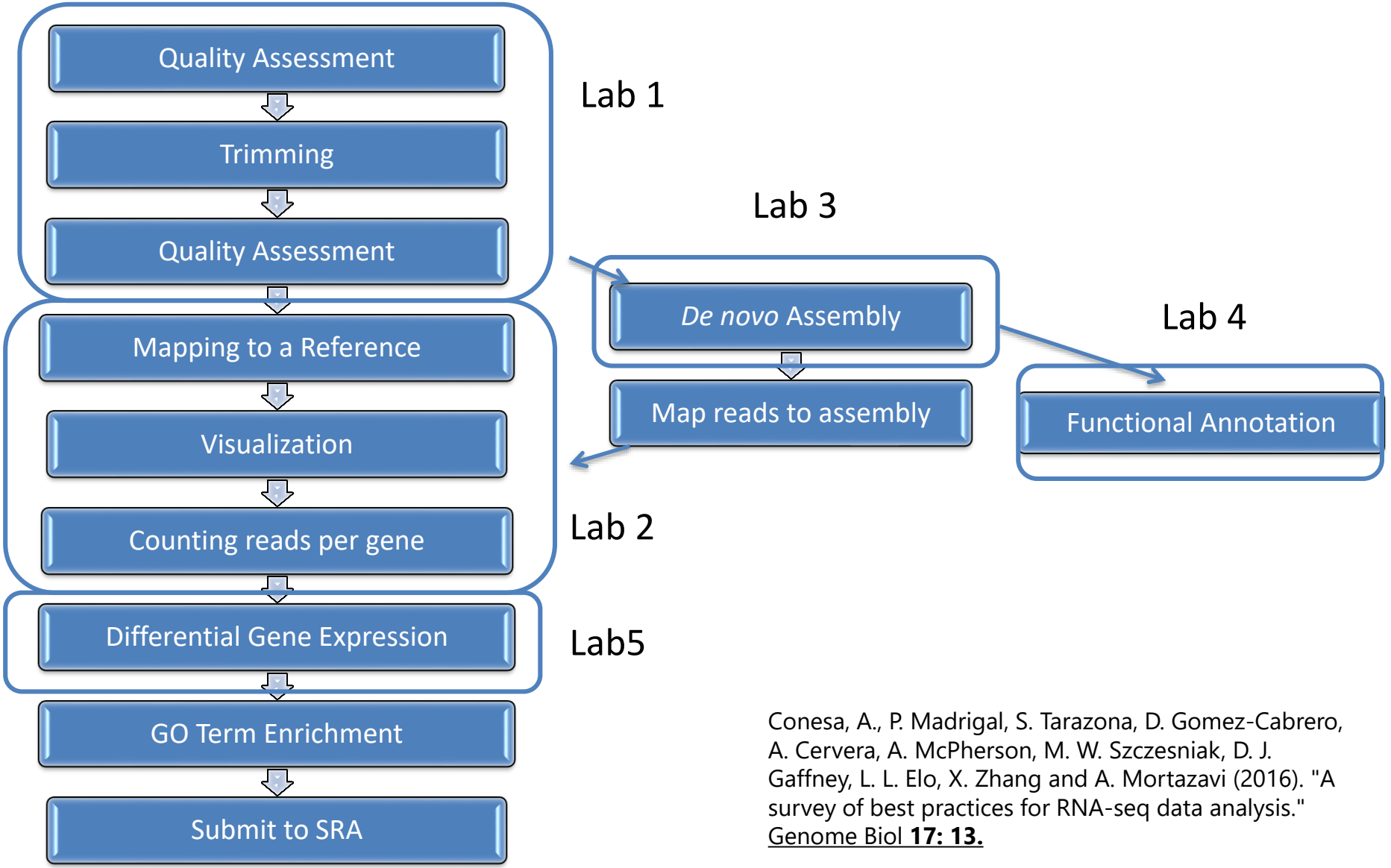
NCBI

What if you don't have a reference?





Bolger, M. E., B. Arsova and B. Usadel (2017). "Plant genome and transcriptome annotations: from misconceptions to simple solutions." Brief Bioinform.



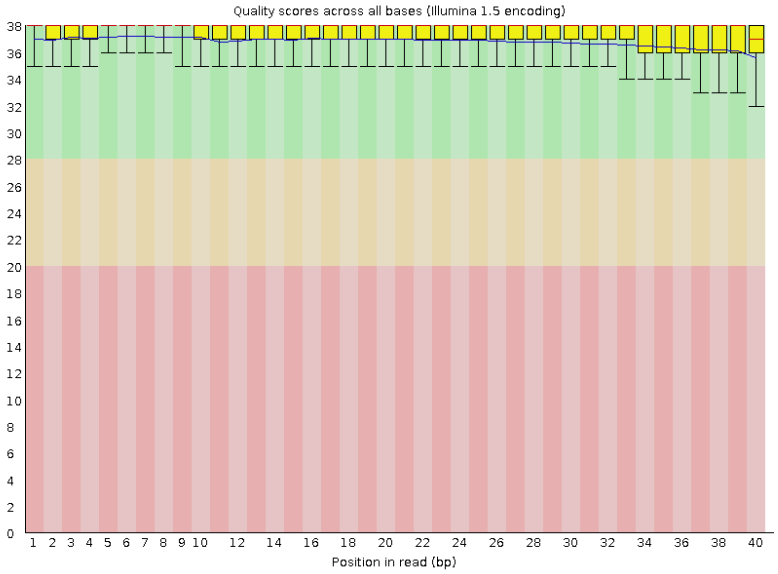
Conesa, A., P. Madrigal, S. Tarazona, D. Gomez-Cabrero, A. Cervera, A. McPherson, M. W. Szczesniak, D. J. Gaffney, L. L. Elo, X. Zhang and A. Mortazavi (2016). "A survey of best practices for RNA-seq data analysis." Genome Biol **17: 13.**

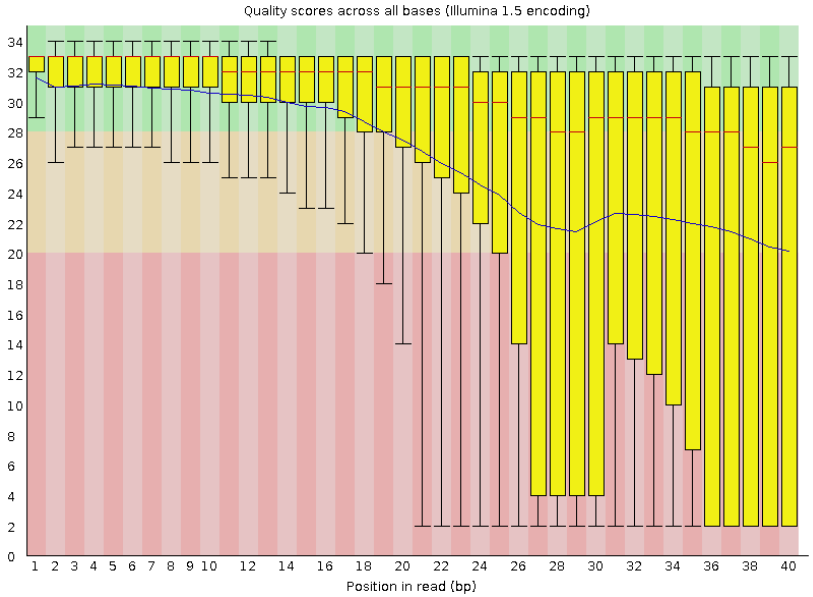
Quality Control

- Is my data of sufficient quality?
- The instrument assigns a confidence value to each base. Are the bases high quality overall?
- Does the complexity look normal?

FastQC

Babraham Bioinformatics





http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimming

- Get rid of the bad data, keep the good data
- Adapter trimming
  - Cut adapter and other Illumina-specific sequences from the read
- Quality trimming
  - Trim off low quality bases
  - Drop a read entirely if is too low quality or too short

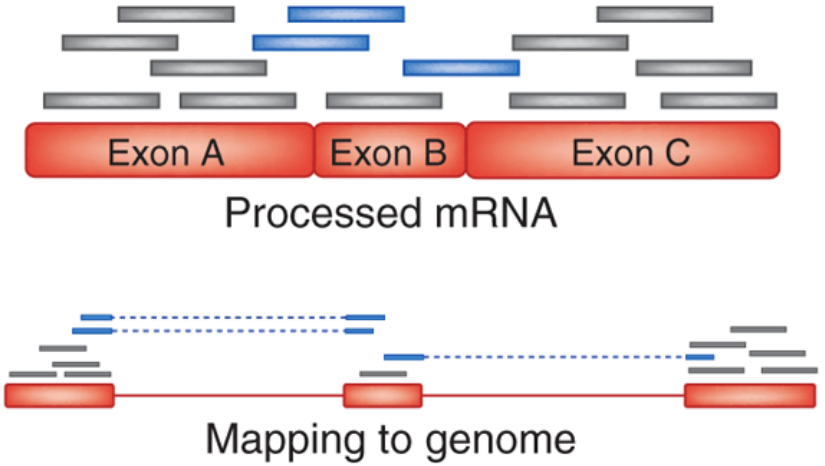
Skewer

Jiang, H., R. Lei, S. W. Ding and S. Zhu (2014). "Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads." BMC Bioinformatics **15: 182.**

Newest research:
Gentle trimming is better.

Mapping to the Reference

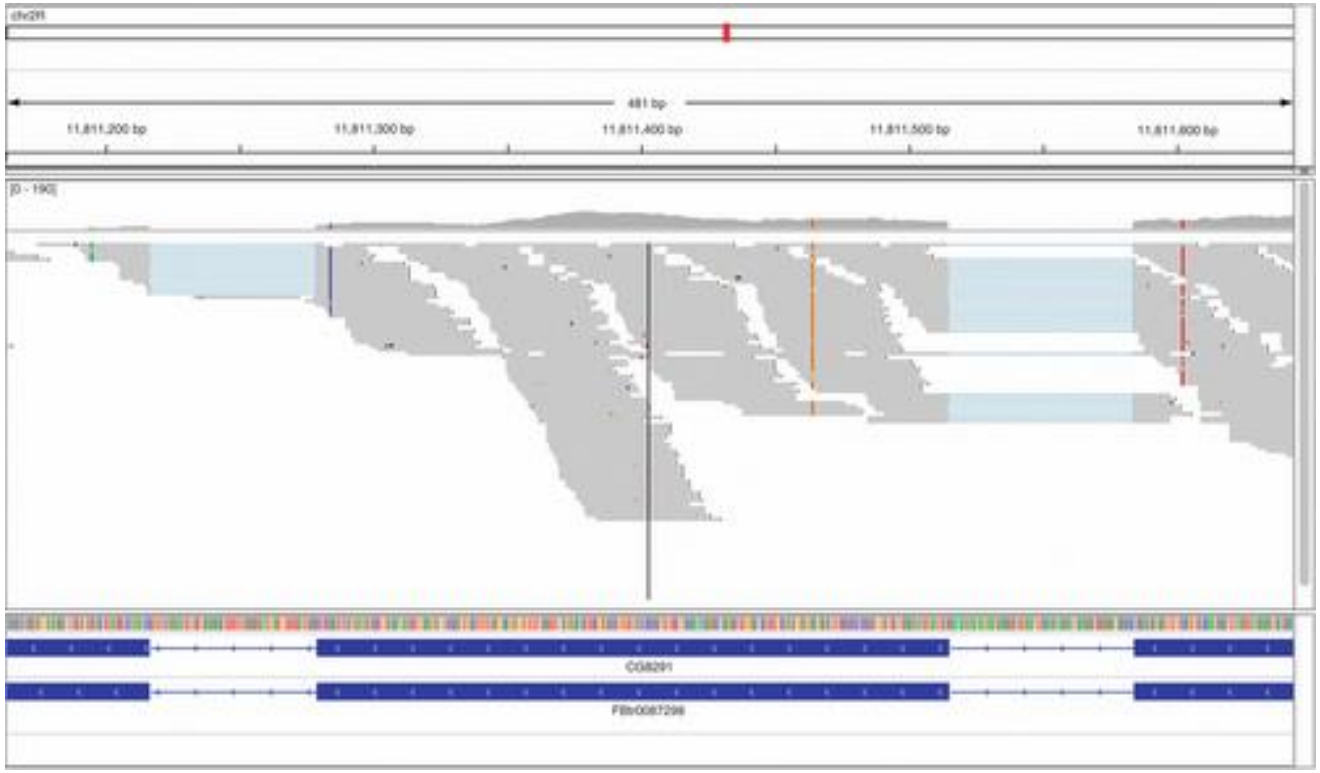
- Mapping RNA to a eukaryotic genome is more complicated than mapping DNA
  - Introns
  - Alternative splicing
- Use a mapping software designed for spliced RNASeq
  - The software will use a file (gff3) to know where the genes are located
  - If this is not available, some mapping software can infer gene structures (This is good for identifying novel genes and isoforms)



Benjamin, A. M., M. Nichols, T. W. Burke, G. S. Ginsburg and J. E. Lucas (2014). "Comparing reference-based RNA-Seq mapping methods for non-human primate data." BMC Genomics **15: 570.**

Visualization

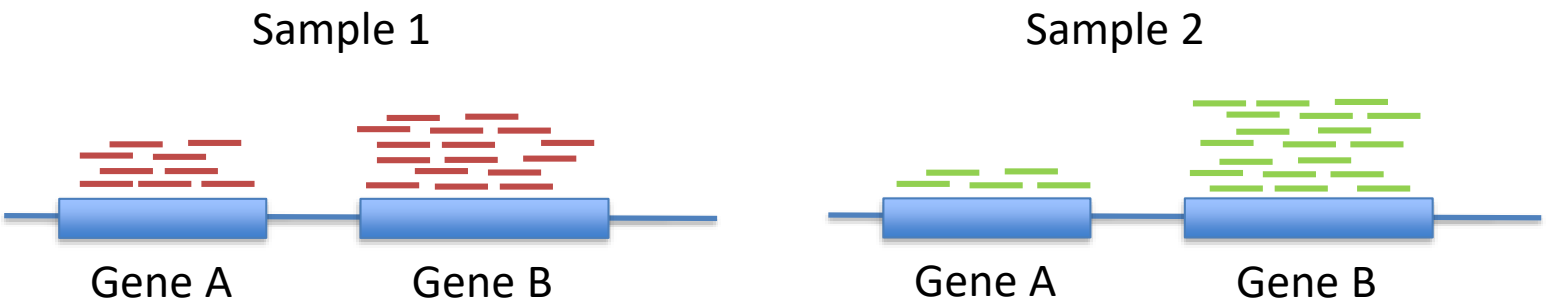
- Look at your data
- The number one most under-appreciate step in data analysis



Integrative Genomics Viewer

Differential expression

- Find genes responding to the conditions
- Biological replicates give power to your results
- Choose an algorithm that suits the data
  - RNASeq, replicates



Generate sequence counts
for all genes in genome

Schurch, N. J., P. Schofield, M. Gierlinski, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson, T. Owen-Hughes, M. Blaxter and G. J. Barton (2016). "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" Rna **22(6): 839-851.**

Making data public

- NCBI Short Read Archive (SRA)
  - Stores raw sequence data from "next-generation" sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics.
  - SRA also stores alignment information in the form of read placements on a reference sequence.
- Upload to SRA
  - Make a list of all the things you need to know prior to starting the project, and keep it updated.
  - Most journals require an accession number prior to publication
  - Enhances reproducibility and allows for new discovery by comparing data sets.
  - Overview of submission process:

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=show&f=sra_sub_expl&view=get_started

Upload to SRA