BIOL 343
Applied Bioinformatics I

Differential expression

Dr. Nic Wheeler

Counting converts mapped reads to expression values

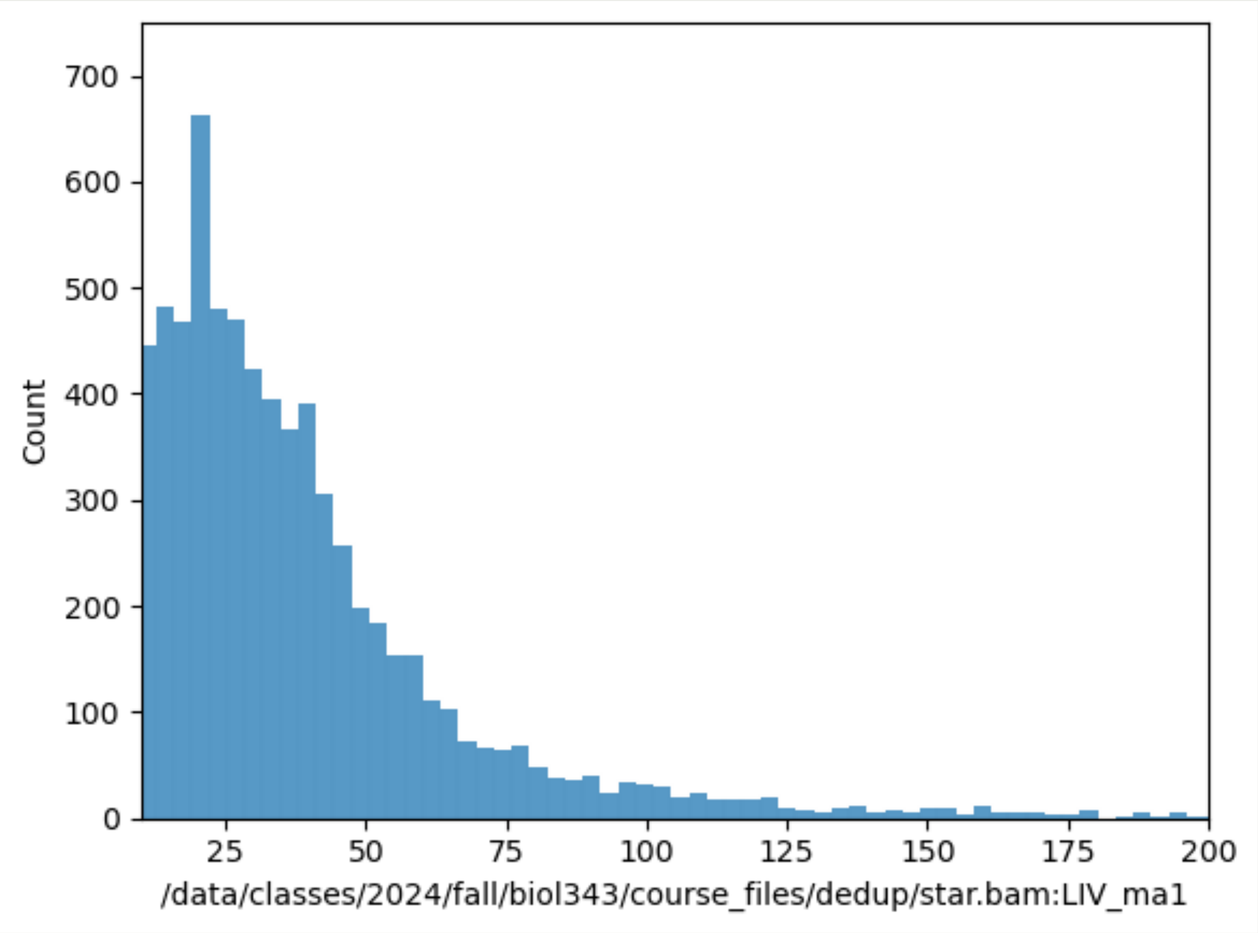
Expression values will be used by statistical models to identify DEGs

- Recall the goal of our RNA-seq experiments…
  - Treatment vs Control
  - Mutant vs Wild type
  - ***Identify differentially expressed genes (DEGs)***
- DEGs will be identified using statistical tests comparing ***expression values*** of transcripts/genes
- Expression values will be calculated based on the number of reads that ***align/map*** to a given genomic locus
- There are different ways to count reads

Differential expression performs statistics on expression values

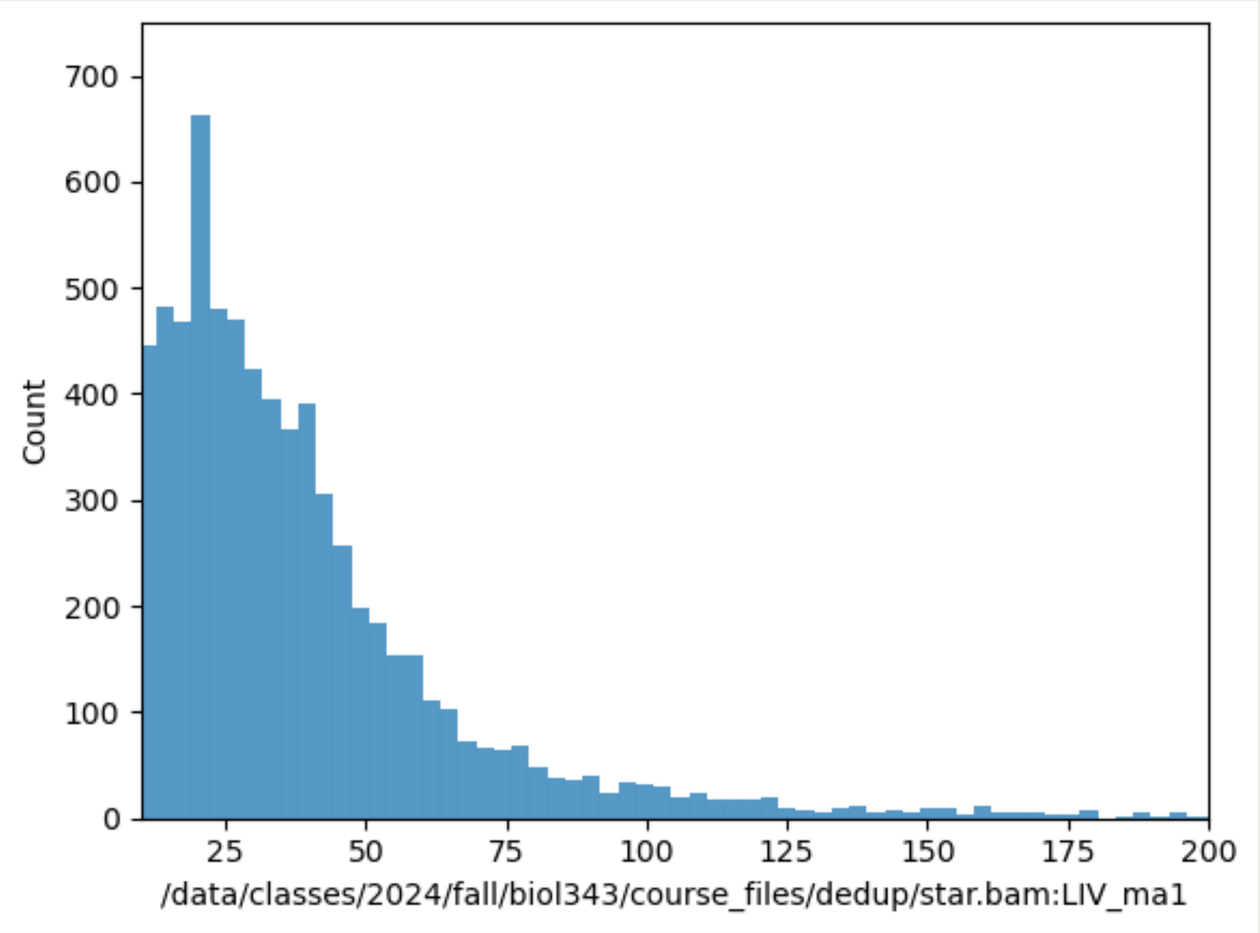
Idiosyncrasies of a count matrix

- Most experiments have 3 replicates
- Very high variation in gene expression values even between replicates
  - If using a t-test, will never find sig. diffs because of high standard deviation
- Many genes
- Skewed, non-normal distribution
  - ANOVA and t-test assume normality



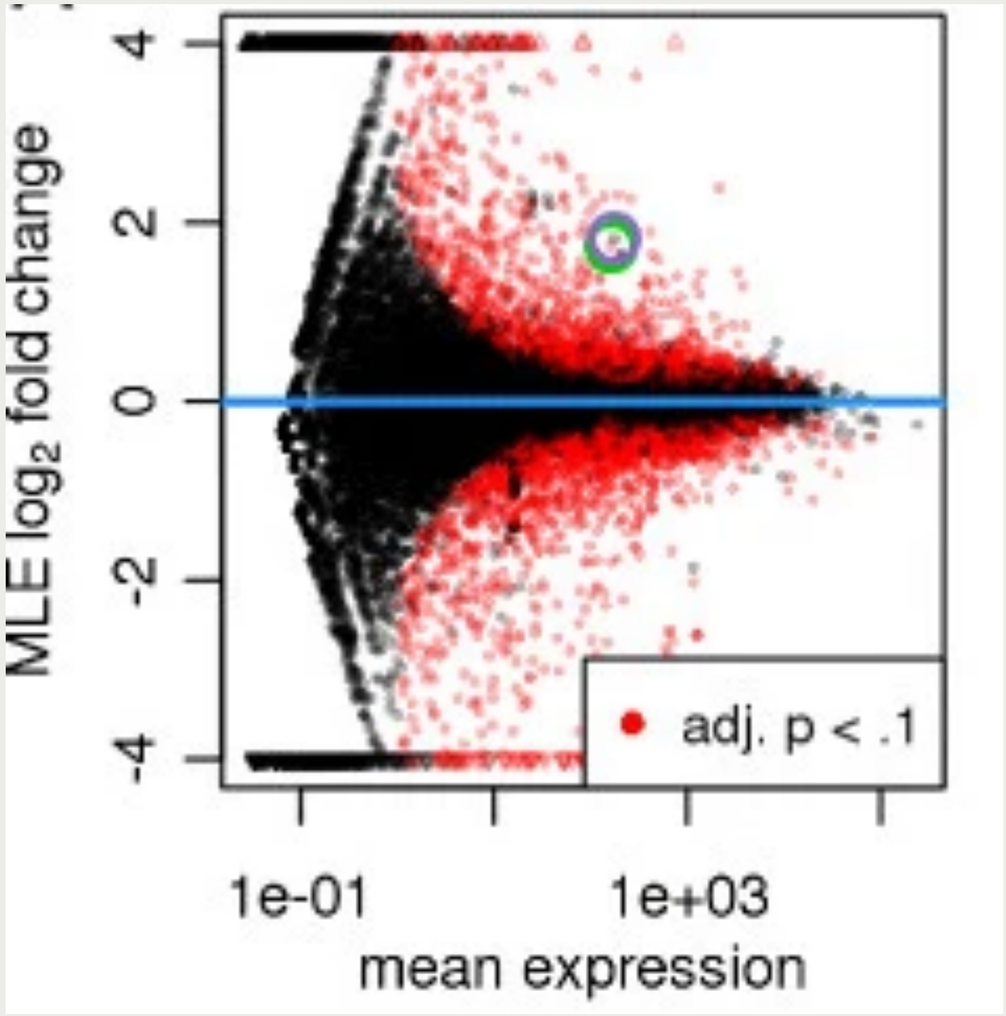
Idiosyncrasies of a count matrix

- Smp_245390
  - INT_ma = 8.67, 8, 1
  - LIV_ma = 269.16, 76.61, 36.19
  - P value and statistical significance:
    - The two-tailed P value equals 0.1666
    - By conventional criteria, this difference is considered to be not statistically significant.
    - SEM = 71.915
- But, these data violate all the assumptions required…



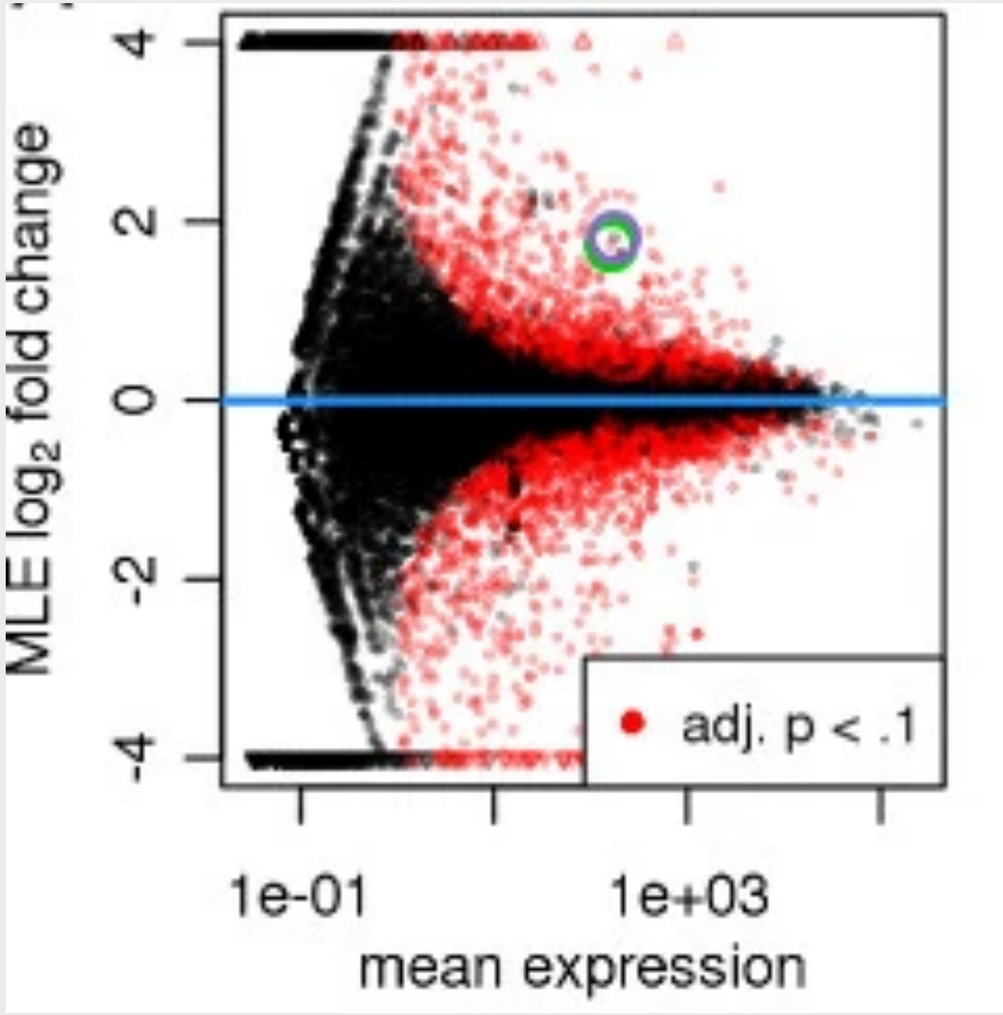
What we want to measure: fold change

- Fold change is the ratio of two measurements
- For example: mean expression of 10 vs mean expression of 5 = two-fold change
- Problem: genes with low expression values are more likely to experience a fold change based just on randomness (noisy fold changes for genes with low counts)



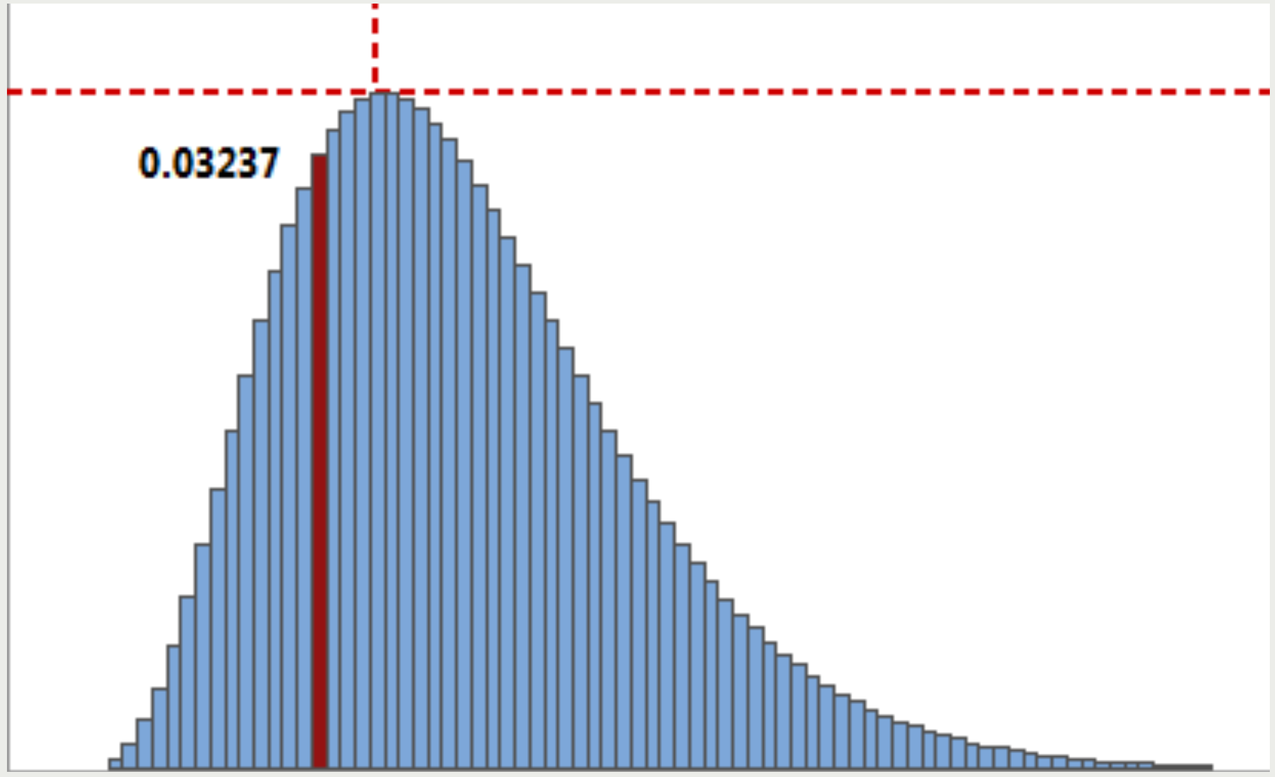
What we want to measure: fold change

- In RNA-seq, we usually use $\log_2$ fold-change (L2FC)
  - L2FC = 1 ⟶ 2 fold-change
  - L2FC = 2 ⟶ 4 fold-change
  - L2FC = -3 ⟶ 8 fold-change
- Easier to visualize a L2FC scale



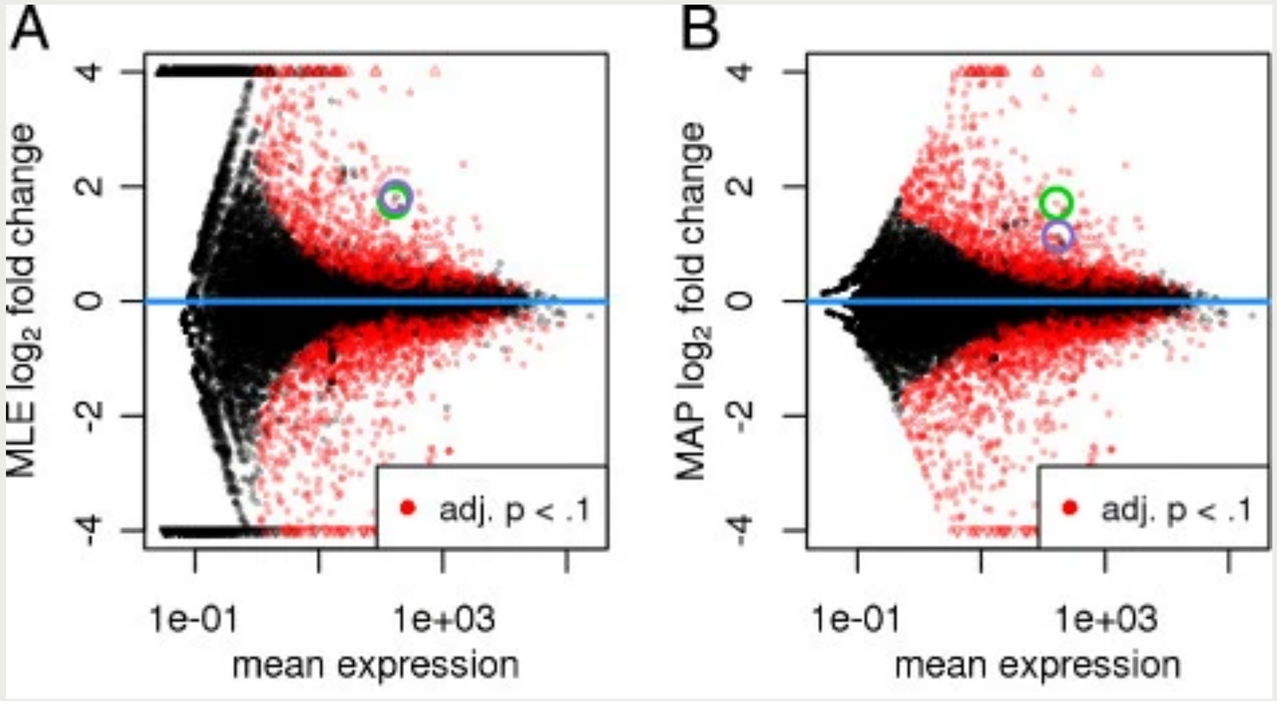
Negative binomial distribution

- Problem: naturally high variation in expression values; most genes have very low expression values
- Solution: model the data differently and share variation across genes
  - Negative binomial distribution
  - Typical for count data, where low counts are more likely to occur than high counts
  - Use “dispersion” instead of standard deviation



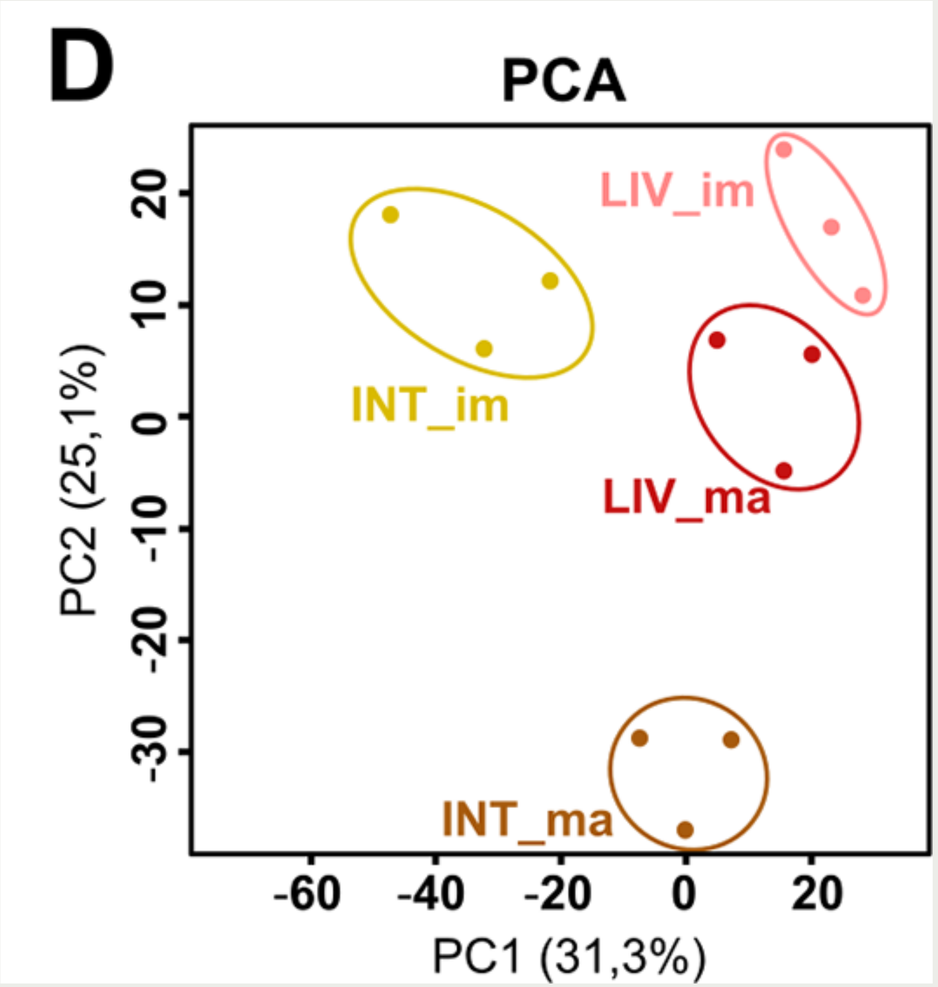
DESeq2

- Statistical package that incorporates negative binomial modeling and shares information (i.e., variation) across genes
- Written in R but ported to Python (PyDESeq2)
- Introduces variance-stabilizing transformations
- Calculates P values that the L2FC != 0
- Requires ignoring genes with very low expression (i.e., average expression < 10)



PCA plots

- PCA plots show the first two PCs
- Type of QC plot to ensure samples cluster as expected (by treatment, tissue type, age, etc.)
- Requires variance-stabilizing transformations to account for high variance at low expression values
- Can indicate contamination or sequencing errors



Volcano plots

- Plot to show statistical significance (p-values) and effect sizes (log2 fold changes)
- Y-axis is -log10 transformed (small p-values lead to higher points)
- Thresholds for p-values (0.05) and log2 fold changes (1/-1 or higher) are shown
- Interesting genes will fall in the top left and top right panels