

Miscellaneous

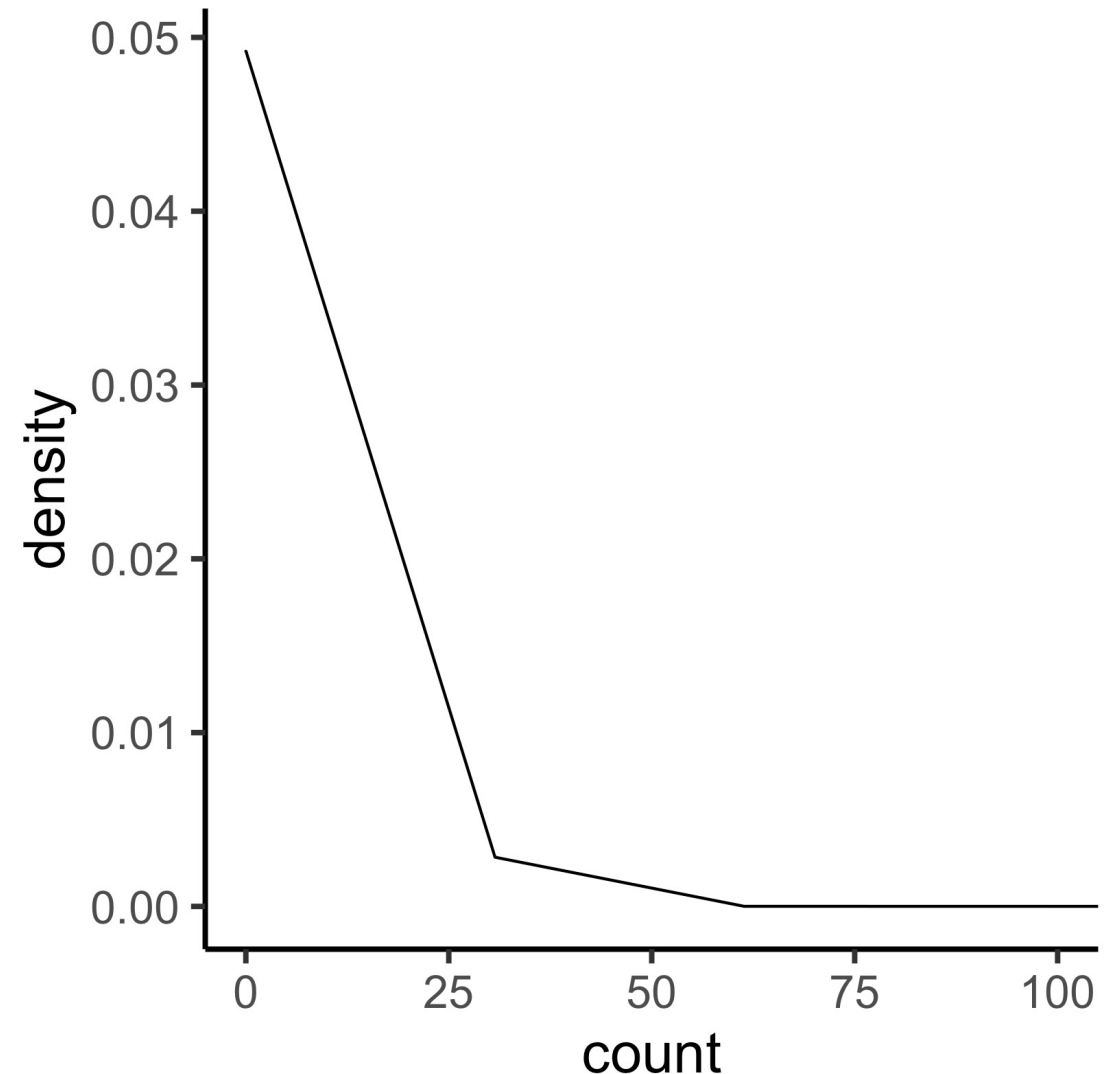
- We're going to take a picture during your exploratory data analysis presentations on Friday, so please wear your BRE t-shirts
- Everyone struggled with the readr/tidyr assignment, so we'll be going over that as a group at some point next week

Clustering to Check for Effects and Differential Expression

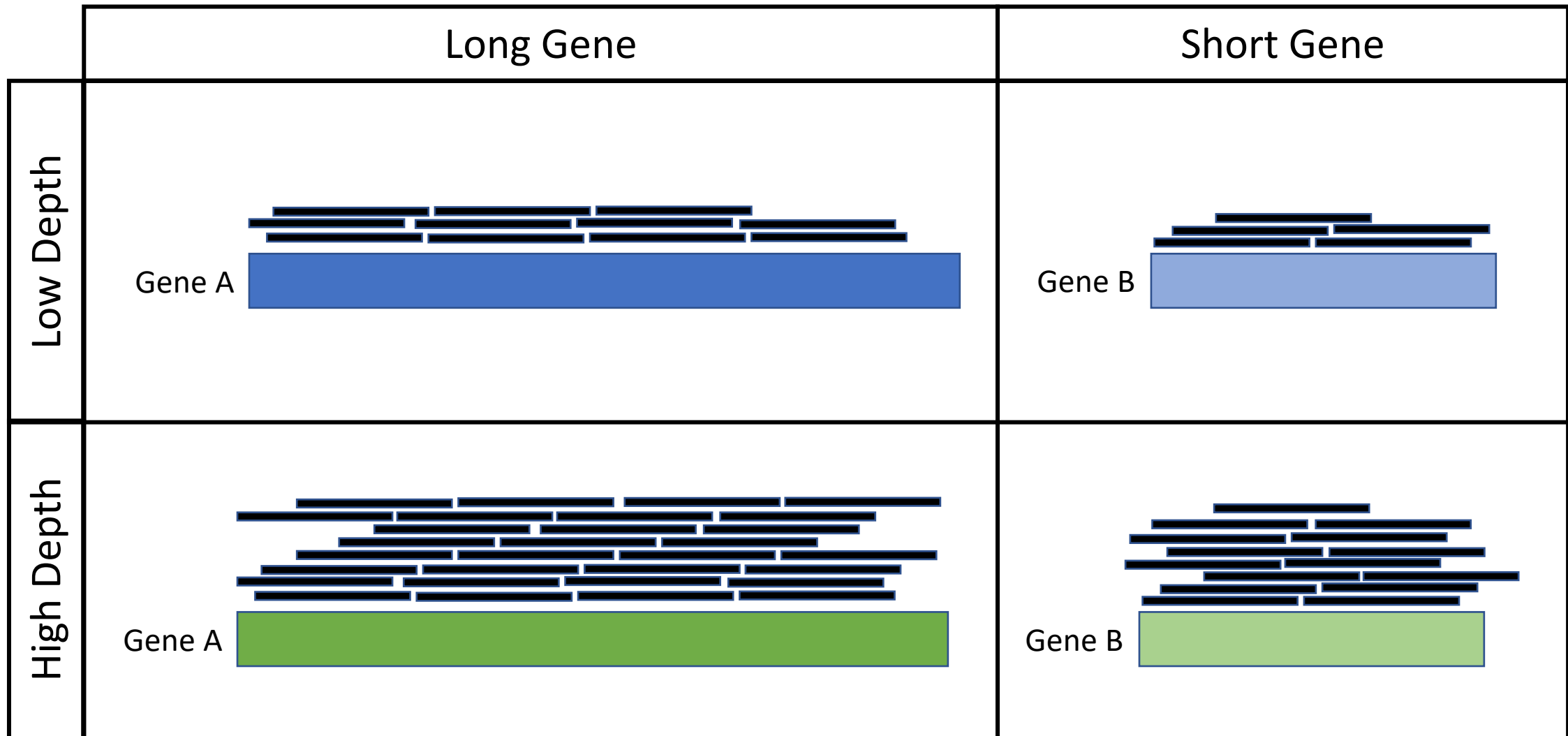
2021-07-21

RNA-seq Counts Must be Normalized

- Start with counts of reads mapping to a gene
- Data is extremely right-skewed
 - True of all sequencing data
 - Formally, this is a negative binomial distribution
- Multiple normalized counts that people use to compensate for it



RNA-seq Counts Must be Normalized



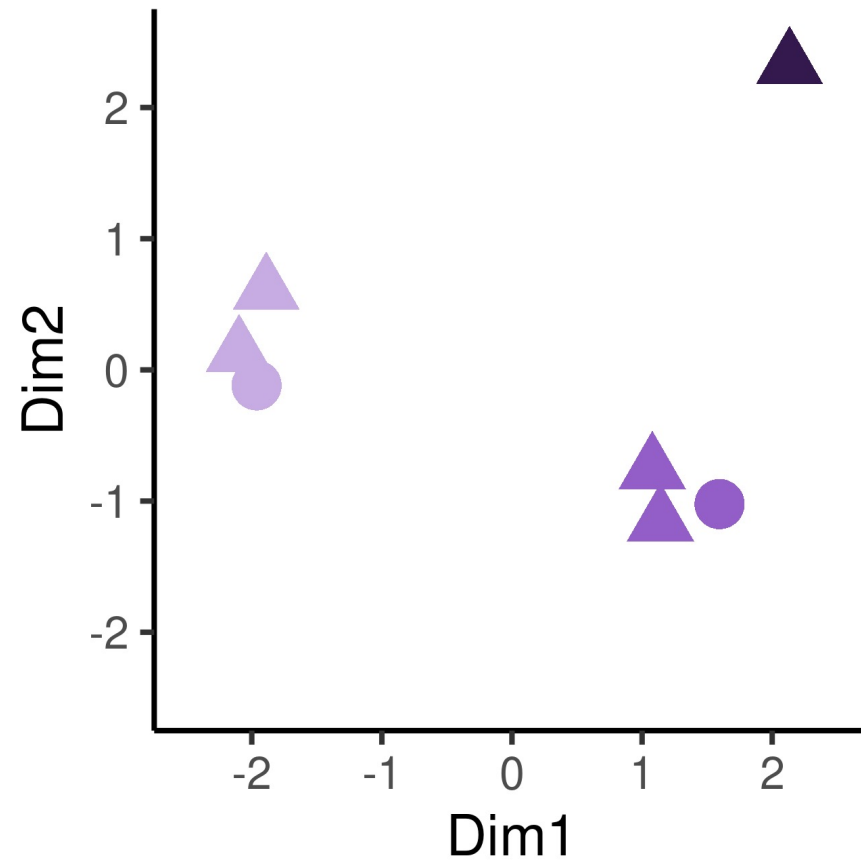
RNA-seq Counts Must be Normalized

Normalization Method	Description	Corrects For:	Use For:
Counts/Fragments Per Million (CPM/FPM)	Counts scaled by the total number of reads in the library	<ul style="list-style-type: none"> sequencing depth 	<ul style="list-style-type: none"> Sample comparison NOT for differential expression testing
Transcripts Per kilobase Million (TPM)	Counts per length of transcript scaled by number of reads	<ul style="list-style-type: none"> sequencing depth gene length 	<ul style="list-style-type: none"> Sample comparison NOT for differential expression testing
Reads/Fragments Per Kilobase of exon per Million reads (RPKM/FPKM)	Same as TPM, but per exon instead of per transcript	<ul style="list-style-type: none"> sequencing depth gene length 	DO NOT USE because they values are not comparable between samples
DESeq2 median of ratios	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	<ul style="list-style-type: none"> sequencing depth RNA composition 	<ul style="list-style-type: none"> NOT Sample comparison differential expression testing
edgeR Trimmed Mean of M values (TMM)	uses a weighted trimmed mean of the log expression ratios between samples	<ul style="list-style-type: none"> sequencing depth gene length RNA composition 	<ul style="list-style-type: none"> Sample comparison differential expression testing

Why do we cluster
the data?

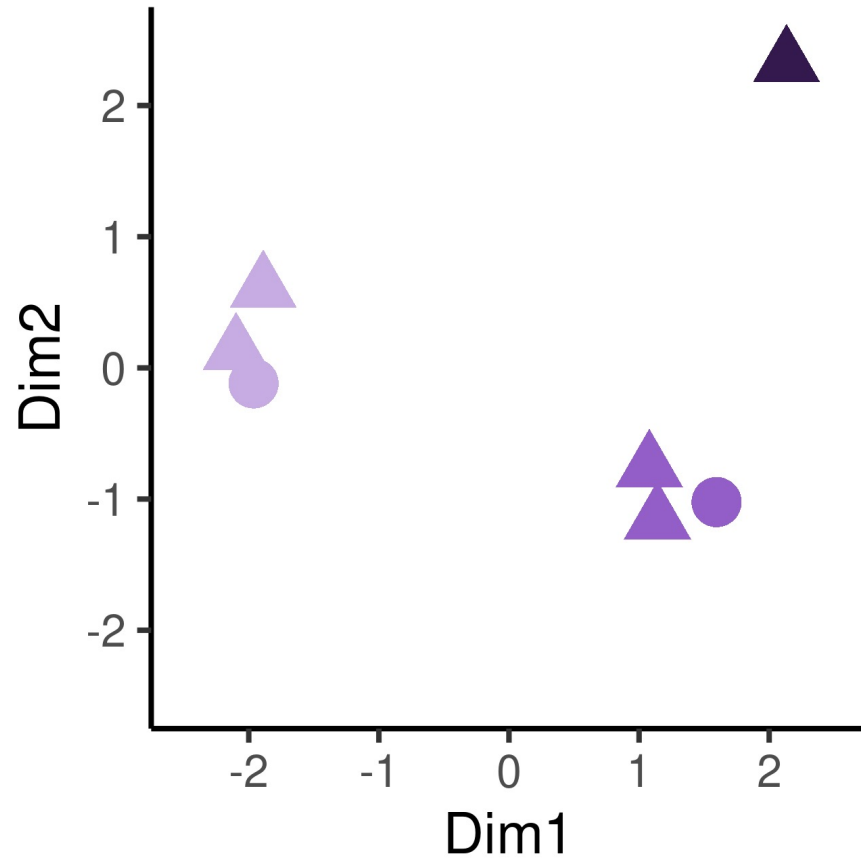
Why do we cluster the data?

GOOD: Data clusters by effect, not batch

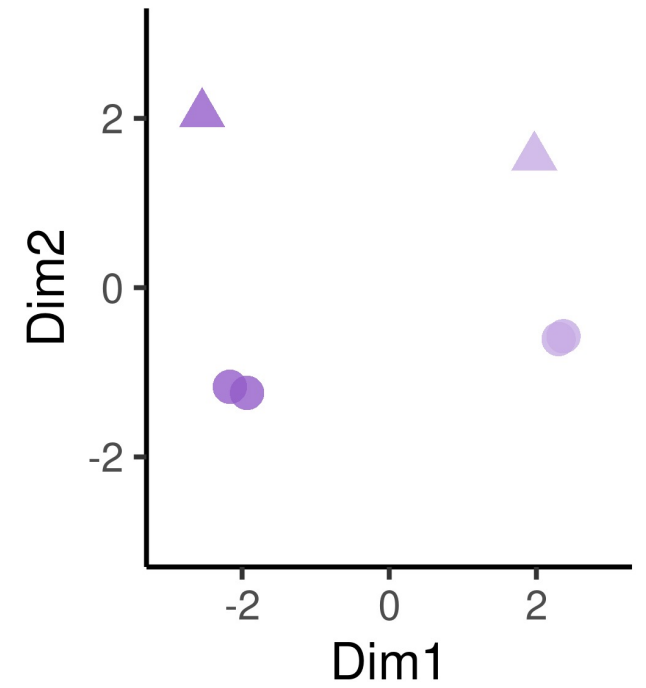


Why do we cluster the data?

GOOD: Data clusters by effect, not batch

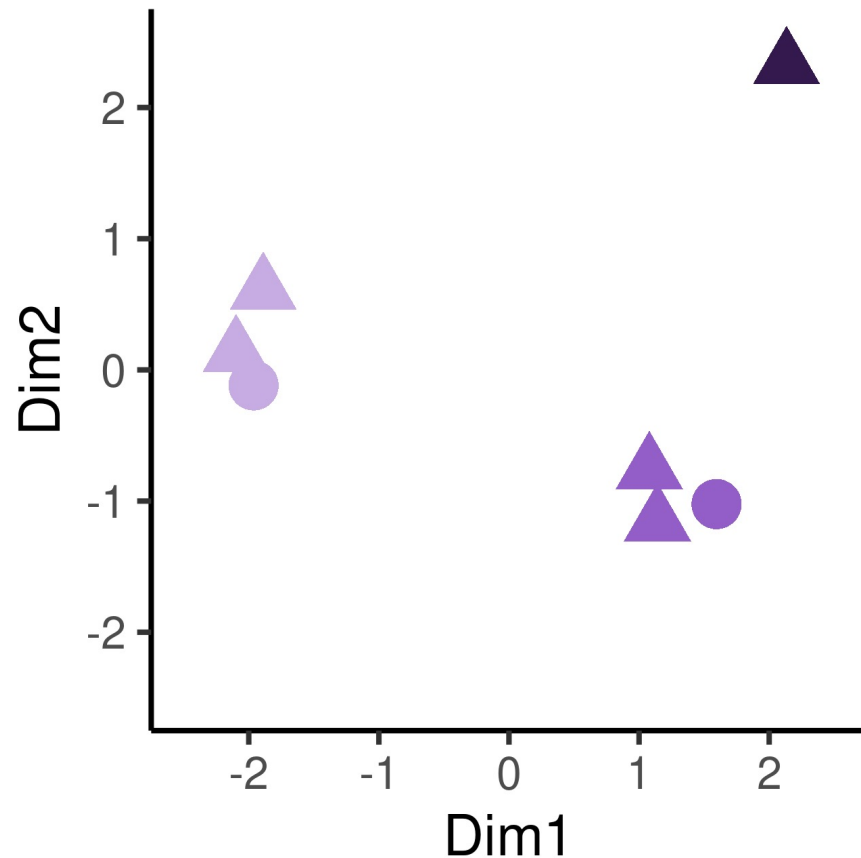


BAD: Data clusters by effect, then batch

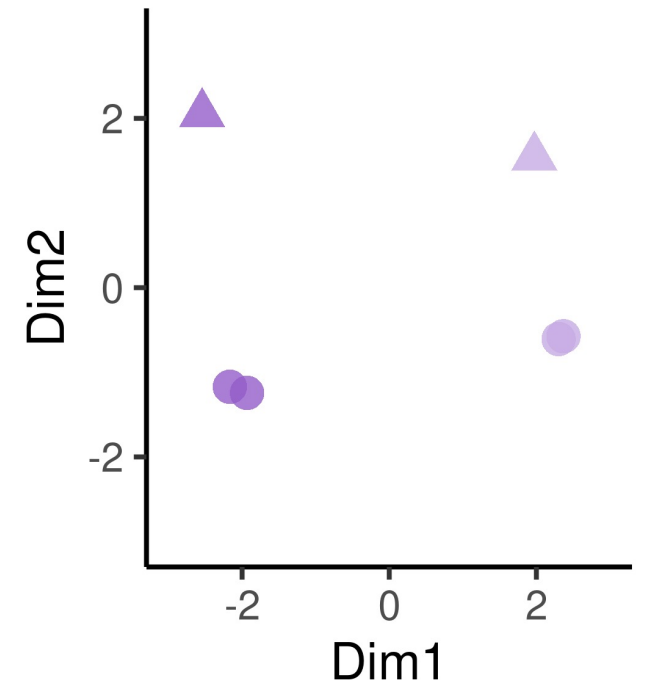


Why do we cluster the data?

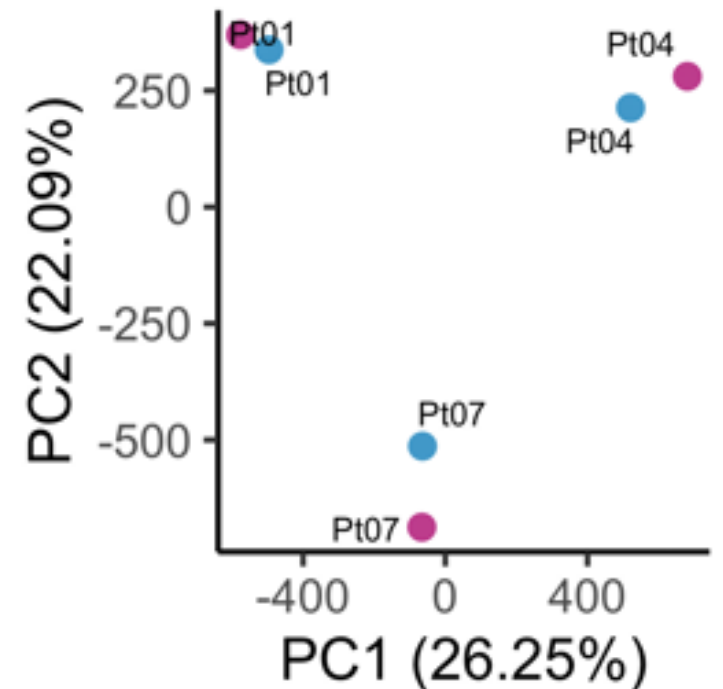
GOOD: Data clusters by effect, not batch



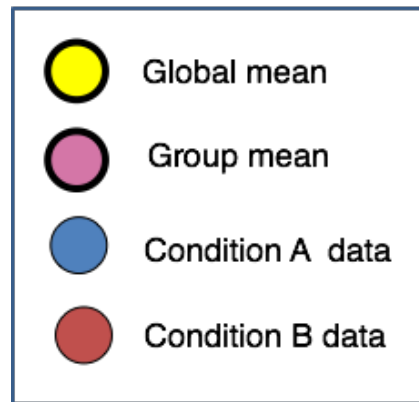
BAD: Data clusters by effect, then batch



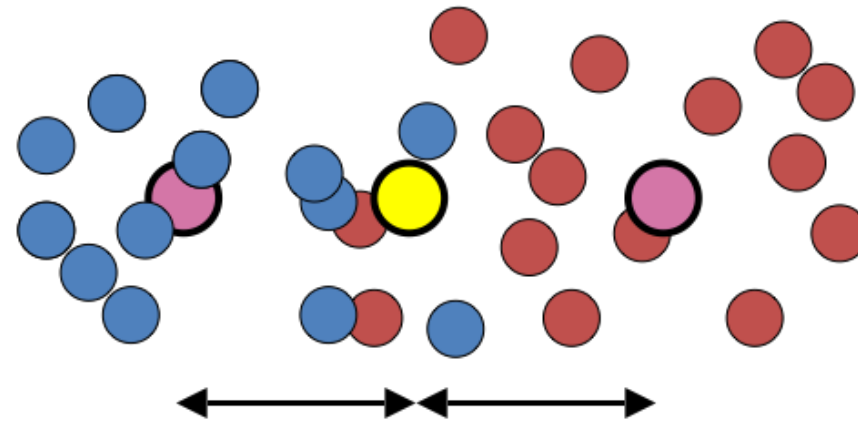
VERY BAD: Data does not show desired effect



How does differential expression work?

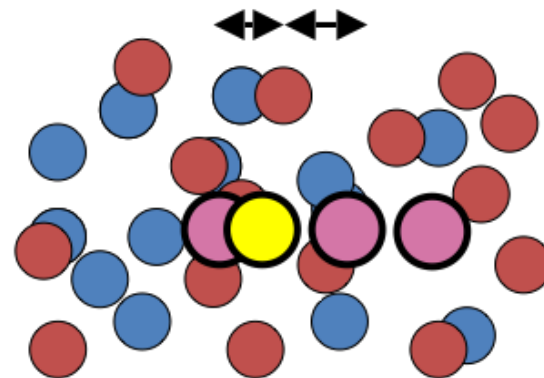


Expression level →



Significant difference

Deviations from global mean



No significant difference