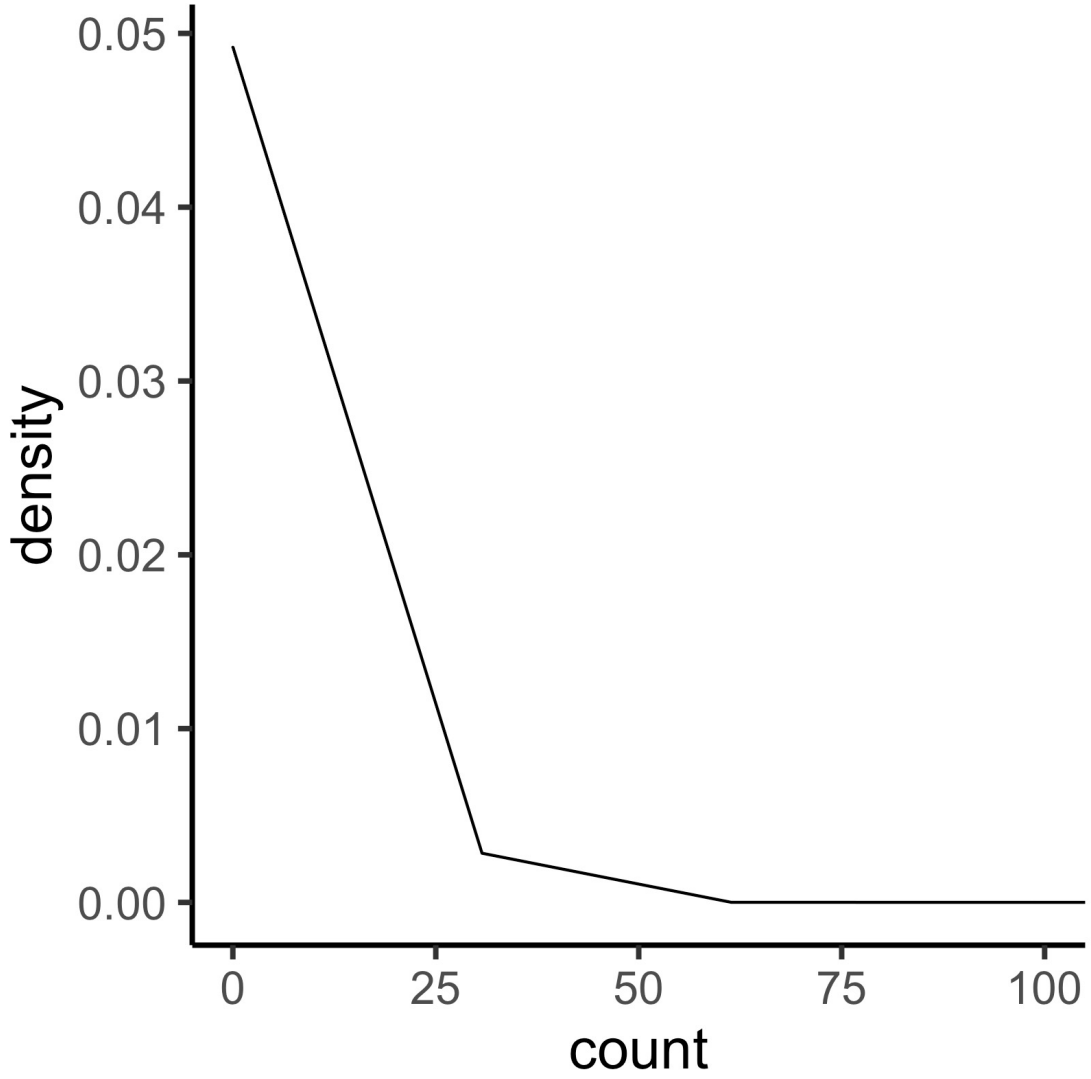
Clustering to Check for Effects and Differential Expression

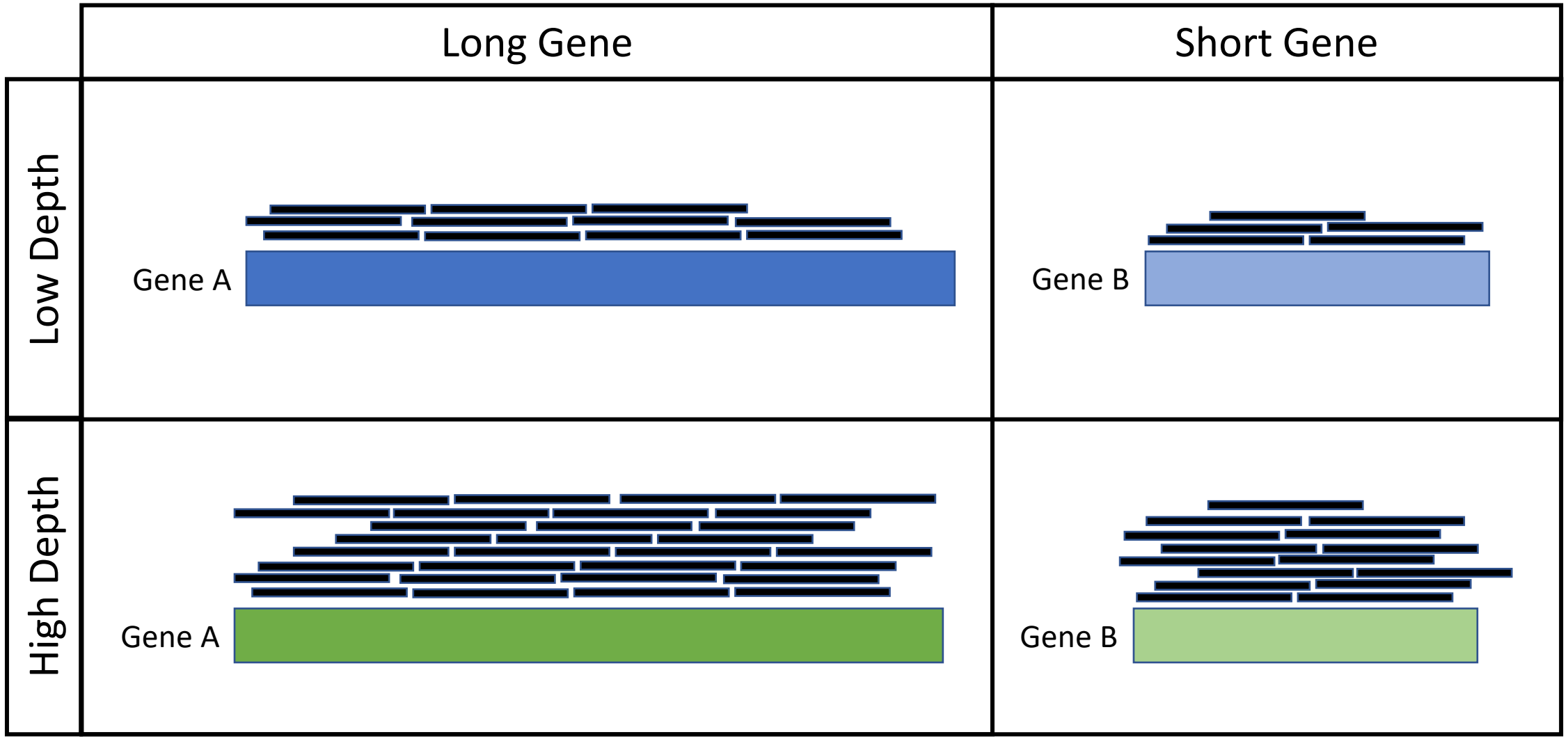
2021-07-21

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



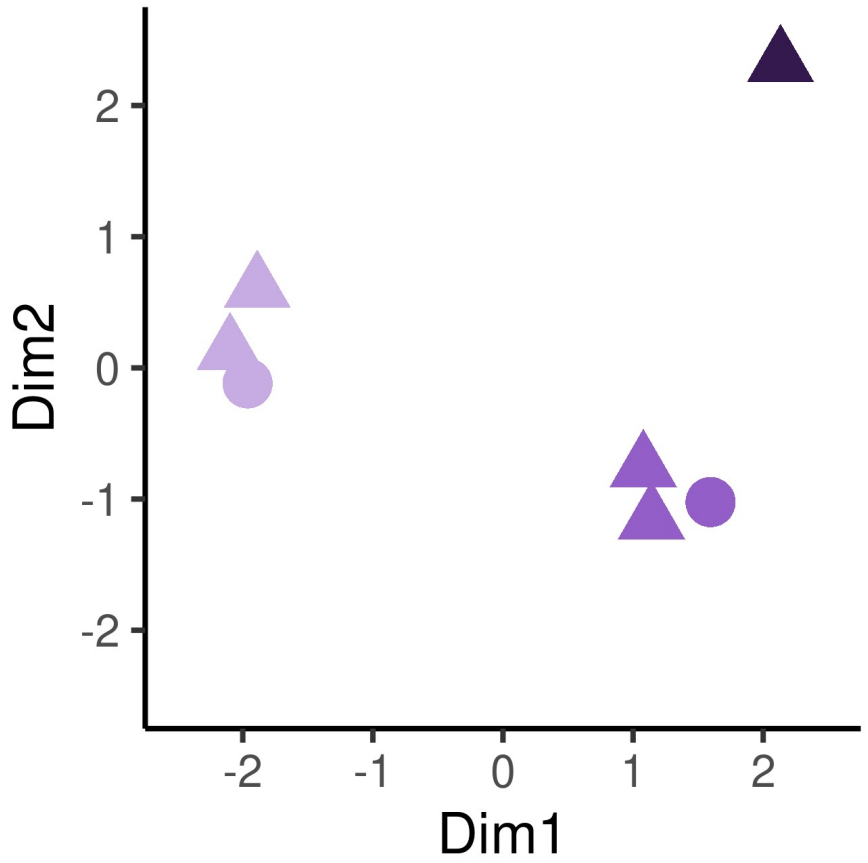
RNA-seq Counts Must be Normalized

Normalization Method	Description	Corrects For:	Use For:
Counts/**F**ragments **P**er **M**illion (**CPM/FPM**)	Counts scaled by the total number of reads in the library	• sequencing depth	• Sample comparison • **NOT** for differential expression testing
Transcripts **P**er kilobase **M**illion (**TPM**)	Counts per length of transcript scaled by number of reads	• sequencing depth • gene length	• Sample comparison • **NOT** for differential expression testing
Reads/**F**ragments **P**er **K**ilobase of exon per **M**illion reads (**RPKM/FPKM**)	Same as TPM, but per exon instead of per transcript	• 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	• sequencing depth • RNA composition	• **NOT** Sample comparison • differential expression testing
edgeR **T**rimmed **M**ean of **M** values (**TMM**)	uses a weighted trimmed mean of the log expression ratios between samples	• sequencing depth • gene length • RNA composition	• Sample comparison • differential expression testing

Why do we cluster the data?

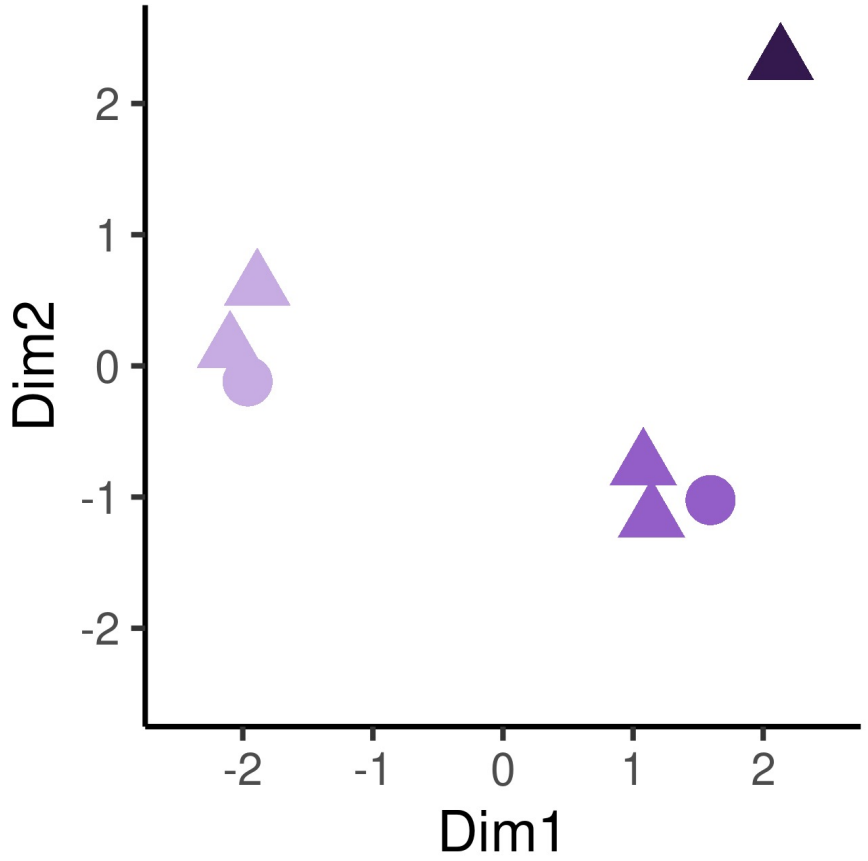
Why do we cluster the data?

GOOD: Data clusters by effect, not batch

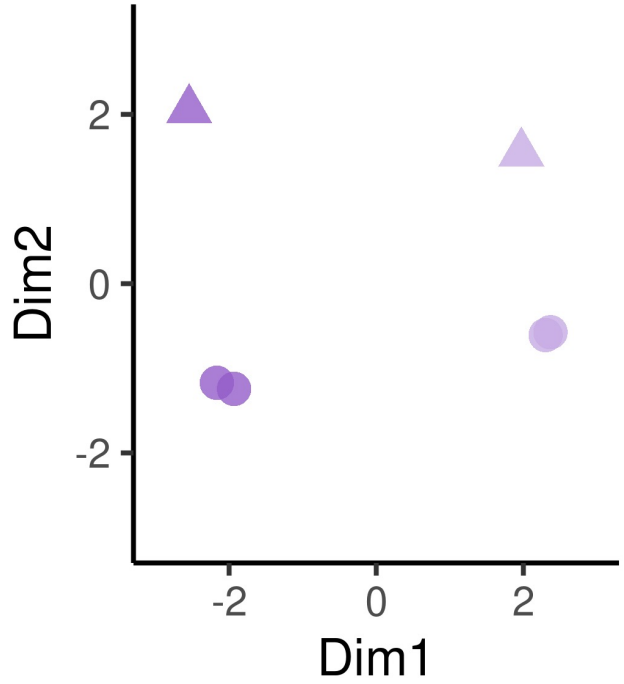


Why do we cluster the data?

GOOD: Data clusters by effect, not batch

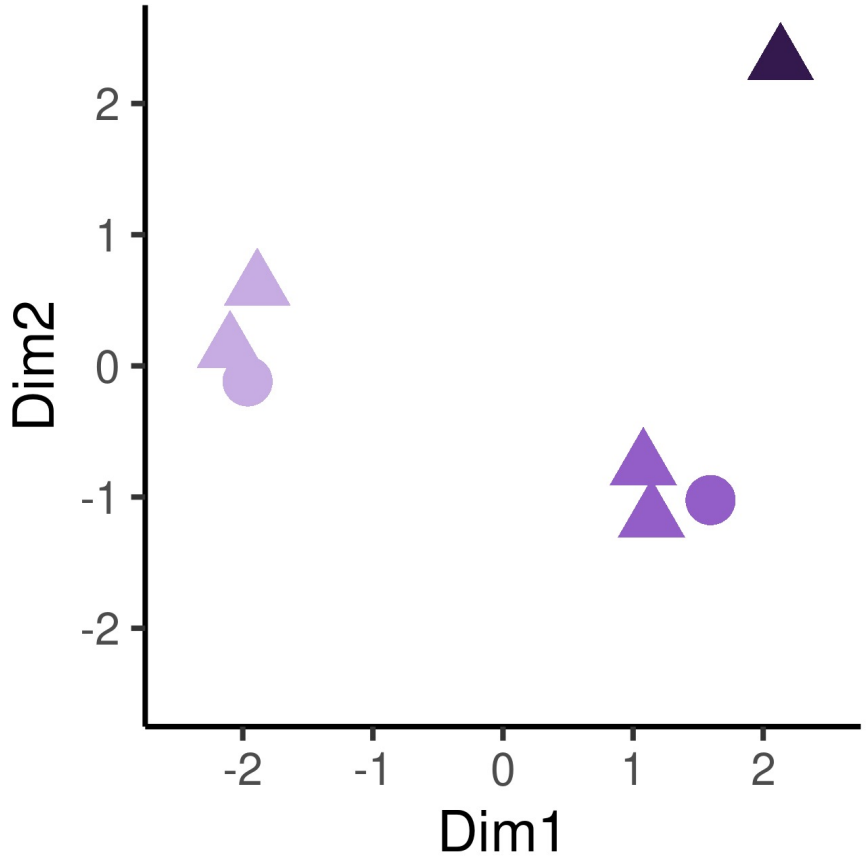


BAD: Data clusters by effect, then batch

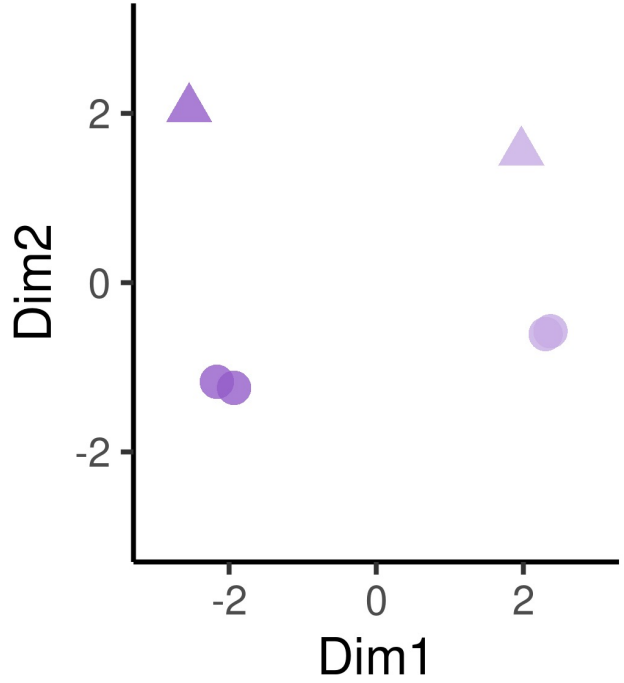


Why do we cluster the data?

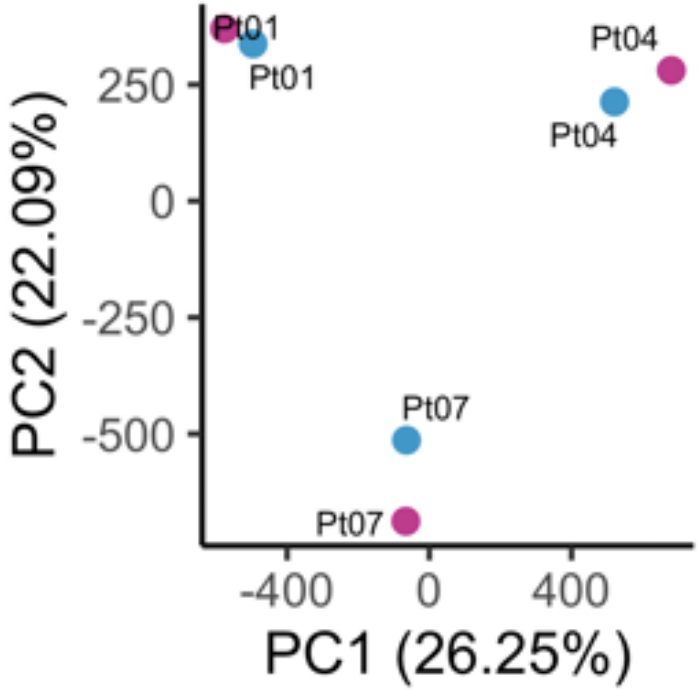
GOOD: Data clusters by effect, not batch



BAD: Data clusters by effect, then batch



VERY BAD: Data does not show desired effect



How does differential expression work?