

# BIOS226

# Visualisation and Dimension Reduction Analysis

**Rob Morris**

**r.morris10@liverpool.ac.uk**

# Learning Outcomes

- Explain why visualisation is essential in high-dimensional biological data.
- Interpret heatmaps, clustering dendograms, and PCA plots.
- Explain what dimensionality reduction achieves conceptually.
- Describe what principal components represent.
- Identify potential batch effects or outliers from PCA.
- Understand why QC and normalisation must precede visualisation.

# What have we done so far...

- Identified where we can gather nucleotide sequencing data online
- Navigated Galaxy for it's use in sequencing analysis
- QC for sequencing read files (.fastq) - FastQC
- Mapped genomic sequences to reference genome - QC and important stats
- Recognised the importance of normalisation in sequencing analysis
- All while identifying similarities and differences between genomic and transcriptomic sequencing data - technical and biological

# After Alignment, the Paths Diverge

- Up to now: FASTQ → Quality Control → Alignment → BAM file
- Now DNA and RNA sequencing differ:

DNA-seq	RNA-seq
<i>Measures sequence variation</i>	<i>Measures gene expression</i>
Identify variants (SNPs/indels)	Count reads per gene
Create variant (VCF) file	Create gene count matrix
Filter for quality & depth	Normalise for library size
Compare genetic differences	Compare expression patterns

Although the biological questions now differ, the statistical challenge is the same:

***High-dimensional data requiring interpretation.***

# Dimensionality in our data

- The number of variables describing each observation
- Each variable is a new axis (dimension) to the data
- Variables = genes (transcriptomics) or SNPs (genomics)
- Humans cannot visualize more than 3 dimensions (3D)....

# What is High-Dimensional Data?

- Dimension = Variable
- Each gene (or SNP) = a variable (= dimension)
- RNA-seq: 20,000 genes = 20,000-dimensional space
- Humans cannot visualise >3 dimensions...
- We need methods that can summarise high-dimensional data without losing biological meaning

# Before we visualise data, we must make sure:

- ✓ Data aligned
- ✓ Low-quality samples removed
- ✓ Low-count genes filtered
- ✓ Data normalised

# Let's look at a gene count matrix (RNA-seq) for humans...

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH				
25857	CDT2A	0	1.9587226	5.67377	0	0	0	0	0	0	4.252129	4.3026201	0	3.7456663	0	0	0	0	0	6.5381994	0	0	0	6.1249034	0	0	0	0	0	0	1.0527152						
25858	HSFY2	0	0	4.6867641	0	5.6880829	0	1.6441556	0.8651785	0	0	7.6410328	6.1548777	0	5.6626599	0	0	0	0	0	2.3618664	0	0	0	4.1856252	0	0	0	1.2205459	0	0	0	1.7127299	1.4806359			
25859	HSFY1	0	0	4.6867641	0	5.6880829	0	1.6441556	0.8651785	0	0	7.6410328	6.1548777	0	5.6626599	0	0	0	0	0	2.3618664	0	0	0	4.1856252	0	0	0	1.2205459	0	0	0	1.7127299	1.4806359			
25860	TTTY9A	0	0	1.4822387	0	0	0	0	0	0	0	5.2137642	2.5067871	0	0	0	0	0	0	0	0	0	0	1.429177	3.2628057	0	0	0	0	0	0	0	0				
25861	TTTY9B	0	0	1.4822387	0	0	0	0	0	0	0	5.2137642	2.5067871	0	0	0	0	0	0	0	0	0	0	1.429177	3.2628057	0	0	0	0	0	0	0	0				
25862	TTTY14	0.4785538	0.6948184	0.7803334	0	0	0	0	0.6925283	0	0	6.8870359	6.0567876	2.4057833	0	4.9371679	6.6853471	0	0	0	0	6.5381984	0	3.2142334	4.1508764	6.6072866	5.1454293	5.9281794	0	0	1.0763018	1.2362881	0	0	0	1.0527152	
25863	CD24	7.34227	3.1094266	5.6378387	5.4286783	6.5209731	7.3474707	4.2796413	5.5164973	7.6483608	4.3470367	11.729369	9.6928647	8.5174124	9.2212763	7.9476143	13.036211	10.748531	10.889345	10.670318	4.6288012	9.7622212	10.301516	9.7066942	10.116609	10.314818	3.9403626	4.7753271	3.7041808	3.7632125	7.9627351	7.2361641	5.1738492				
25864	BCORP1	0	0	2.8636149	1.5596277	0	0	0.7727721	1.875814	0	0	7.7659612	4.6098523	0	5.6626599	7.0522058	2.8063566	0	0	0	0	7.5304158	5.4506456	0	2.2404093	1.429177	7.9678364	6.6194776	0	0	1.2362881	0	0	0.4618749	0	0.7792097	
25865	TXLNGY	5.6027329	9.0345345	8.5801633	6.948545	8.3707148	0.4357692	9.1271172	8.7057481	1.9948877	8.2402581	8.4638494	9.0312561	9.1400613	7.8102617	9.0530016	10.147313	9.9835826	0	9.6375114	8.7186647	3.3917938	10.408366	9.0799383	9.5815859	9.8471155	10.003728	5.987129	8.6232954	9.2040979	0	7.5744695	0.2507398	9.2480535			
25866	KDM5D	7.7559585	12.03049	10.958803	9.7356718	10.752315	1.2693487	10.781193	10.73027	1.1888125	10.037957	9.5338549	10.150707	8.7368213	9.4173659	9.0167747	9.327512	10.263439	7.1391192	9.2781286	8.5265087	5.2848383	9.309416	9.1806199	10.658261	10.462656	10.98715	1.4823374	11.017826	11.062795	0.9371732	10.135796	1.626667	11.202436			
25867	TTTY10	0	0	0.6211922	0	0	0	0	0	0	0	7.9351496	5.4313477	0	5.2570847	0	0	0	0	0	6.8973147	0	5.9315664	2.2404093	0	5.4591821	0	0	0	0	0	0	0				
25868	EIF1AY	5.7204162	9.0174127	9.3419135	6.861812	9.0571166	0.4357692	9.5196255	9.5852186	0.5125304	9.9669498	7.2711882	7.3656535	5.593699	5.2570847	6.5591753	6.7670302	6.1946349	0	7.2850829	5.3363564	0	8.3396151	7.444824	8.0686871	7.9174161	9.1921607	2.4075227	9.5434514	10.210324	0.5433835	7.7312768	0.6502602	10.320436			
25869	RPS4Y2	0	1.2890012	5.8091636	0	1.2563557	0	0.9150902	0	0	0	0	4.6098523	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
25870	PRORY	0	0	4.1588137	1.9714669	2.2895284	0	0	0	0	0	4.252125	5.0782984	0	0	0	0	0	0	0	4.7691386	2.5230838	0	0	0	0	0	0	0	0	0	0	0				
25871	RBMY2EP	0	0.2705013	3.209698	0	0.8744816	0	0	0	0	0	4.252125	0	0	5.2570847	0	4.642738	0	0	0	5.3363564	0	0	0	0	0	0	0	0	0	0	0	0				
25872	RBMY1B	0	0.4982037	6.3558135	0	1.0780169	0	0	0	0	2.7138428	7.5042518	6.2467211	0	5.6626599	0	0	0	0	0	6.7289017	0	0	0	5.7321159	4.8118249	6.1249034	3.0851684	0	0	0	6.8799295	0	0	0	0	
25873	RBMY1A1	0	0.4982037	5.8137718	0	0.6374397	0	0	0	0	2.4458683	6.9930802	5.7776982	0	5.2570847	0	0	0	0	0	5.7424391	0	0	0	5.3260974	4.2944176	5.4591821	3.0851684	0	0	0	6.3637836	0	0	0	0	
25874	RBMY1D	0	0.4982037	6.3558135	0	1.0780169	0	0	0	0	2.7138428	6.7725741	6.2467211	0	5.6626599	0	0	0	0	0	6.7289017	0	0	0	5.7321159	4.8118249	6.1249034	3.0851684	0	0	0	6.8799295	0	0	0	0	
25875	RBMY1E	0	0.4982037	5.881184	0	1.0780169	0	0	0	0	2.5273892	6.1914943	5.8380026	0	5.2570847	0	0	0	0	0	6.3138908	0	0	0	5.1096864	4.2944176	5.7167404	3.0851684	0	0	0	6.3637836	0	0	0	0	
25876	TTTY13	0	0	0	0	0	0	0	0	0	6.5121764	6.0567876	0	0	0	0	0	0	0	3.8210995	0	2.3618664	1.517354	0	0	0	0	0	0	0	0	0	0				
25877	PRY2	0	2.8276129	0.2567423	0	0	0	0	0	0	0	7.5742627	5.7147627	0	5.2570847	0	3.6993646	0	0	0	0	0	6.0589675	0	0	0	0	0	0	0	0	0	0	0	0		
25878	PRY	0	2.8276129	0.2567423	0	0	0	0	0	0	0	7.5742627	5.7147627	0	5.2570847	0	3.6993646	0	0	0	0	0	6.0589675	0	0	0	0	0	0	0	0	0	0	0	0		
25879	LOC101929	0	0.9236509	0.2567423	0	0	0	0	0	0	0	7.9874065	8.0145457	0	5.6626599	1.0687664	3.6993646	0	4.7250249	0	8.4741862	3.3917938	0	0	0	0	0	0	0	0	0	0	0	0	0.8111924	0	0
25880	TTTY6	0	0	0	0	0	0	0	0	0	0	6.5121764	1.7406102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
25881	TTTY6B	0	0	0	0	0	0	0	0	0	0	6.5121764	1.7406102	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
25882	RBMY1F	2037	5.898641	0	1.2563557	0	0	0	0	0	2.8084925	7.5042518	5.7776982	0	5.2570847	0	0	0	0	0	0	6.0589675	0	0	0	0	0	0	0	0	0	0	0	0			
25883	RBMY1J	2037	5.898641	0	1.2563557	0	0	0	0	0	2.8084925	7.5042518	5.648956	0	5.2570847	0	0	0	0	0	0	6.0589675	0	0	0	0	0	0	0	0	0	0	0	0			
25884	TTTY5	0	0	0	0	0	0	0	0	0	0	6.8870359	3.911748	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
25885	RBMY2FP	1.2037	5.1526385	0	0	0	0	0	0	0	0	1.8754344	0	3.3793948	0	3.7456663	0	2.8063566	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
25886	LOC1006521	0	1.8565344	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
25887	TTTY17B	0	0	0	0	0	0	0	0	0	0	3.6677586	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
25888	TTTY17C	0	0	0	0	0	0	0	0	0	0	3.6677586	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
25889	TTTY17A	0	0	0	0	0	0	0	0	0	0	3.6677586	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
25890	TTTY4C	8184	1.8120245	0	0	0	0	0	0	0	0	9.181625	7.8809285	5.8958871	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
25891	TTTY4B	8184	1.8120245	0	0	0	0	0	0	0	0	9.181625	7.8809285	5.8958871	0	0																					

# Visualisation of data and Quality Control

Because:

- We detect outliers
- We identify batch effects
- We verify experimental grouping
- We assess data quality
- We avoid false conclusions

# Distance Between Samples

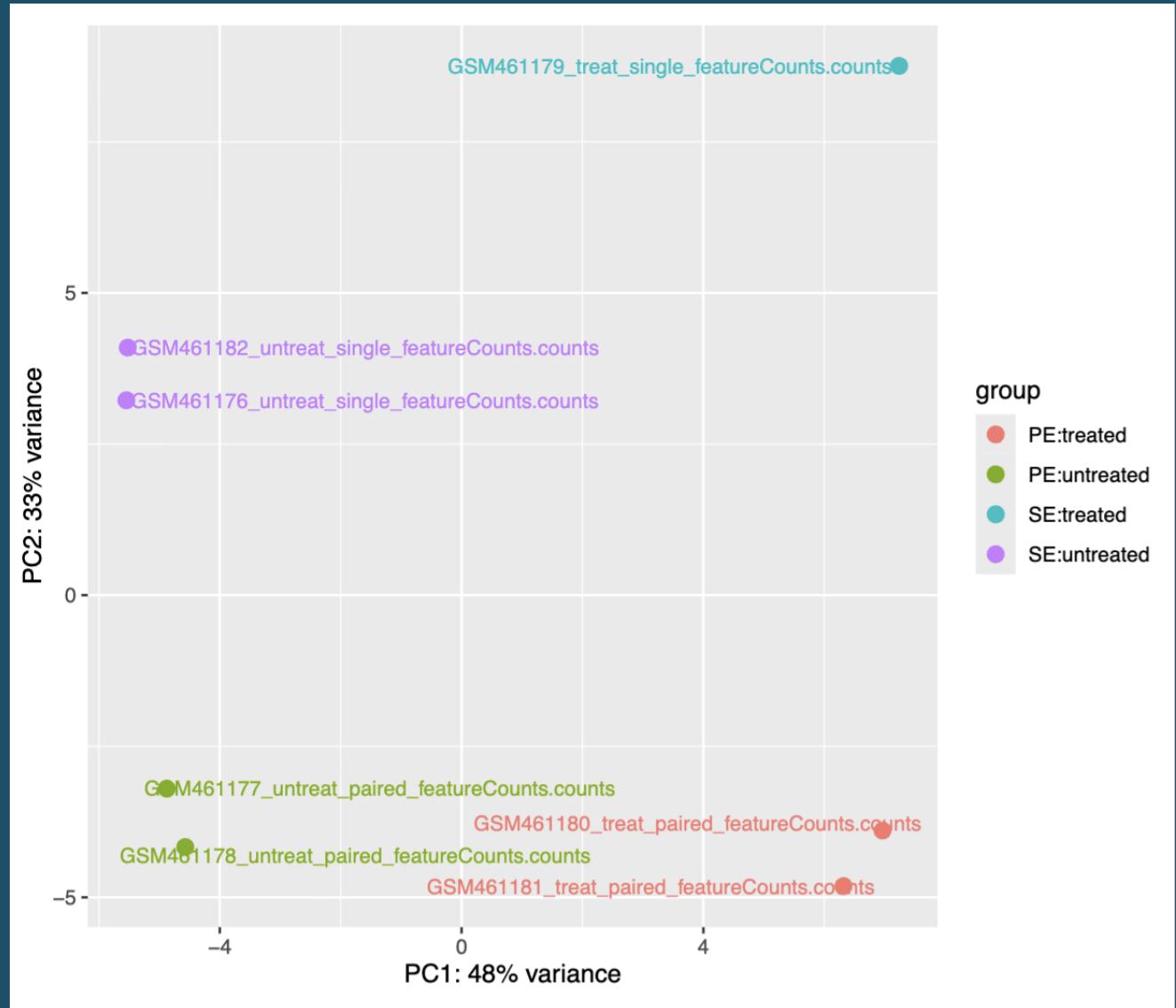
- Each sample = vector of gene expression values
  - We can calculate distance between samples
  - Similar samples cluster together
- 
- Hierarchical clustering
  - Dendograms

# Dimensionality Reduction - The Issue

- If each sample is 20,000-dimensional:
- How do we represent it in 2D?
- **Dimensionality reduction**
  - A mathematical method that reduces many variables into fewer composite variables.

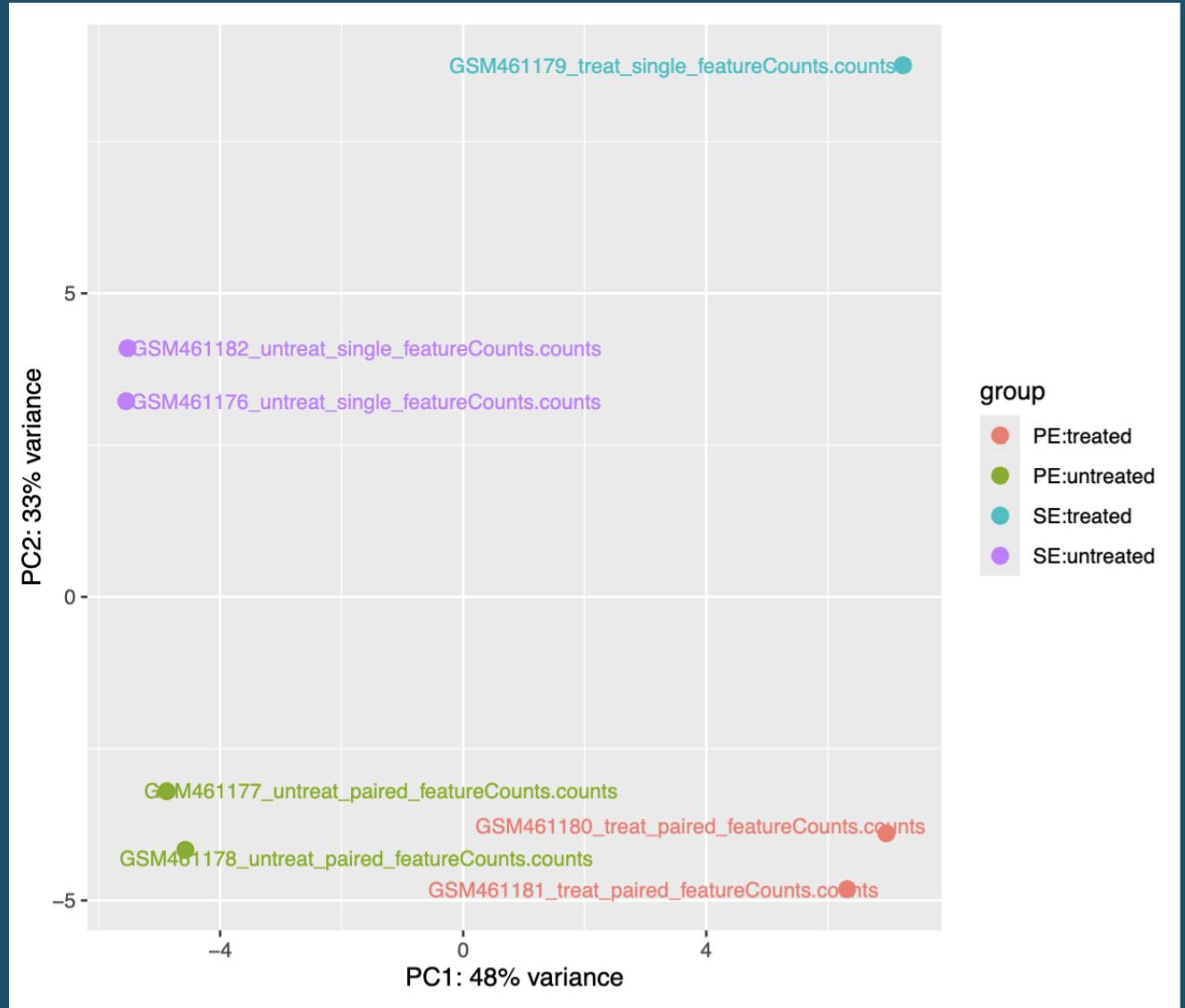
# Principal Component Analysis (PCA)

- Finds directions of maximum variance between samples
- PC1 explains the greatest variance
- PC2 explains the next greatest, orthogonal to PC1
- **Principal components are weighted combinations of all genes**



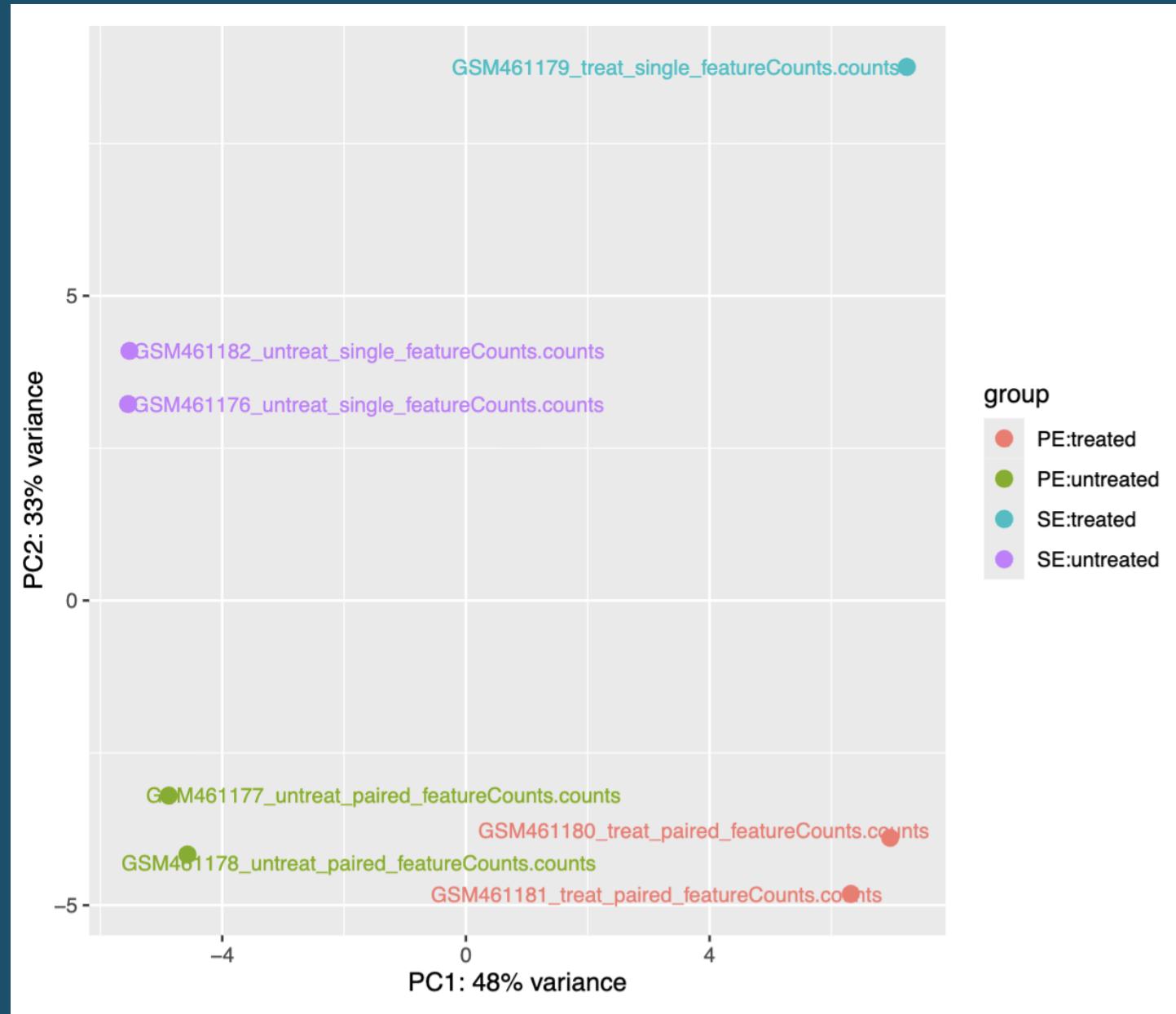
# What PCA Actually Does

- It rotates the coordinate system
- Projects samples into new axes
- Dimensions reduced to 2D plain
- Preserves as much variation as possible
- PCA does not know about treatment groups!



# Interpreting PCA Plots

- Difficult concept to grasp, but mostly comes down to patterns of relatedness
- PC1 separates treated vs control (biological variation)
- PC2 separates batches (technical variation)
- Let's interpret this plot from the upcoming workshop:



# PC1 vs PC2

- **What Does PC1 Represent?**
- Direction of greatest variance
- Largest pattern in dataset
- Could be biology...
- Could be technical....
- But we expect most of the variance to be biological.
- **What Does PC2 Represent?**
- Direction of second greatest variance
- Second largest pattern in dataset
- Orthogonal to PC1 = completely uncorrelated
- We expect most of the variance to be technical.

# Biological Signal vs Batch Effect

Biological variation:

- Treatment
- Genotype
- Disease state

Technical variation:

- Sequencing run
- Library prep
- Date

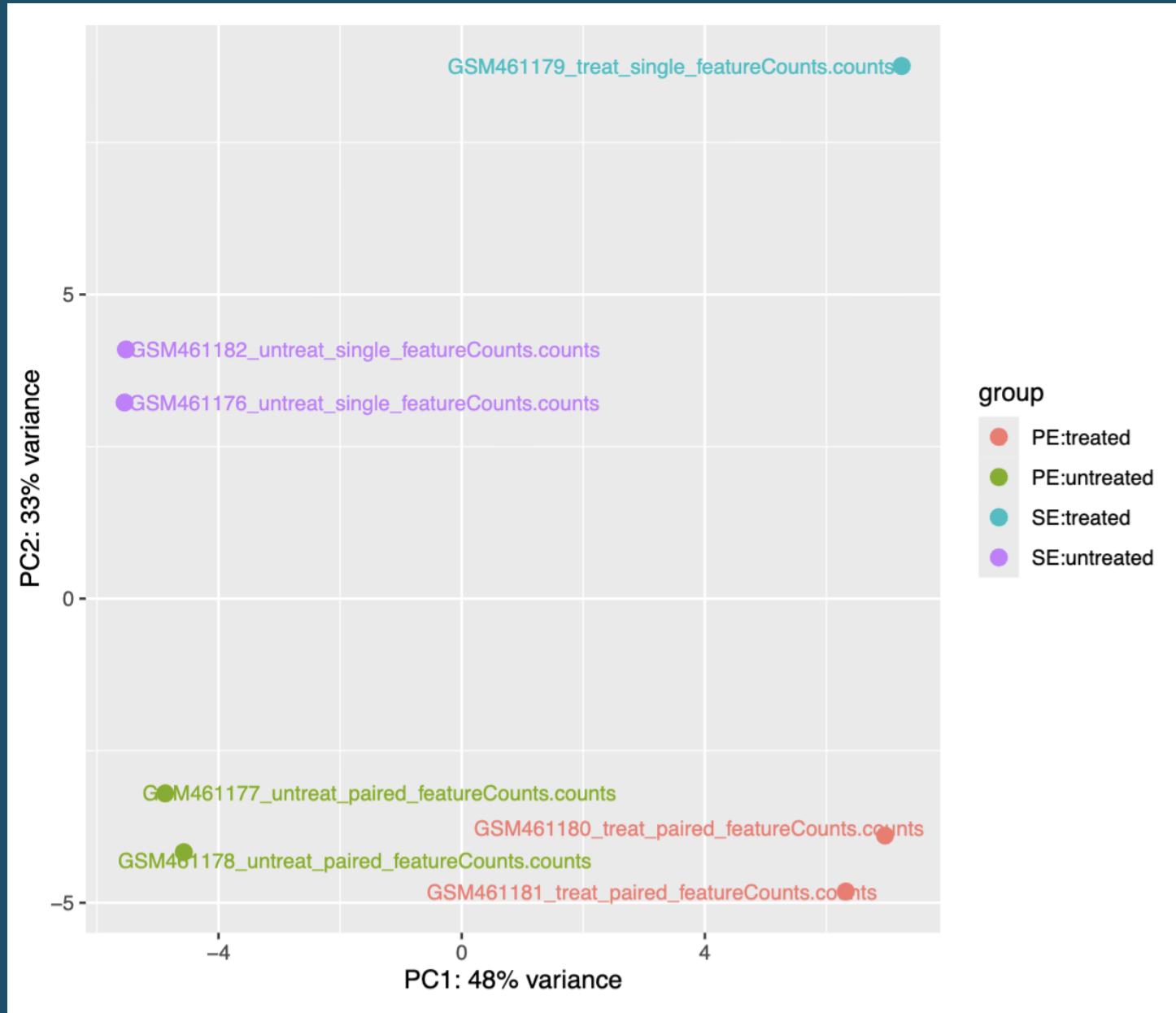
*In well-controlled experiments testing distinct biological conditions, we would expect the largest source of variation to reflect the biological differences being measured. The next largest sources of variation often arise from technical factors, such as differences in sample preparation or sequencing runs.*

*However, if technical variation exceeds biological variation, this may indicate the presence of batch effects or other confounding factors.*

*How would you distinguish between the two?*

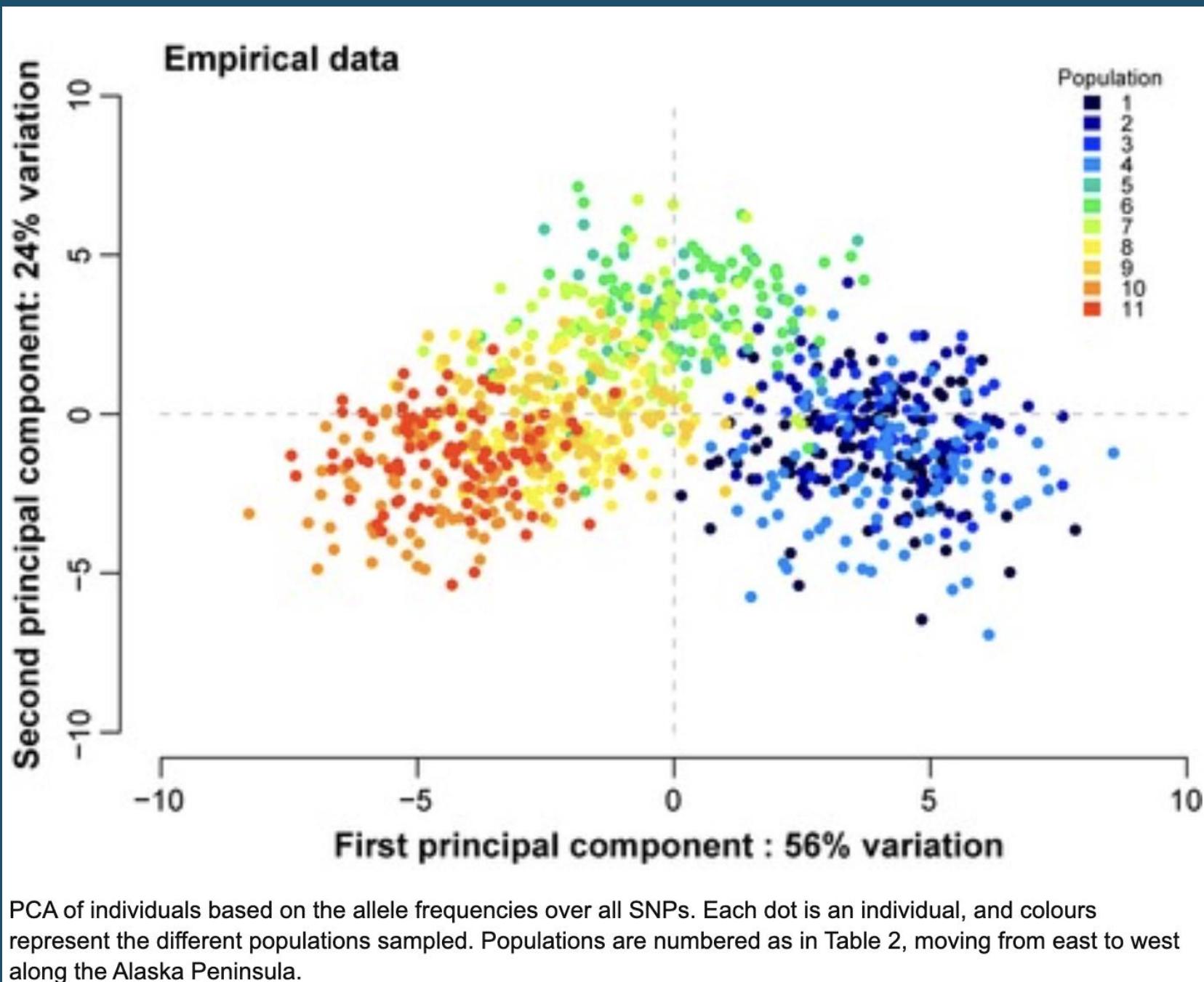
# Outliers in PCA

- Sample far from cluster
- May indicate technical failure
- May indicate real biology
- Important:
- Do not automatically remove outliers.
- Investigate.



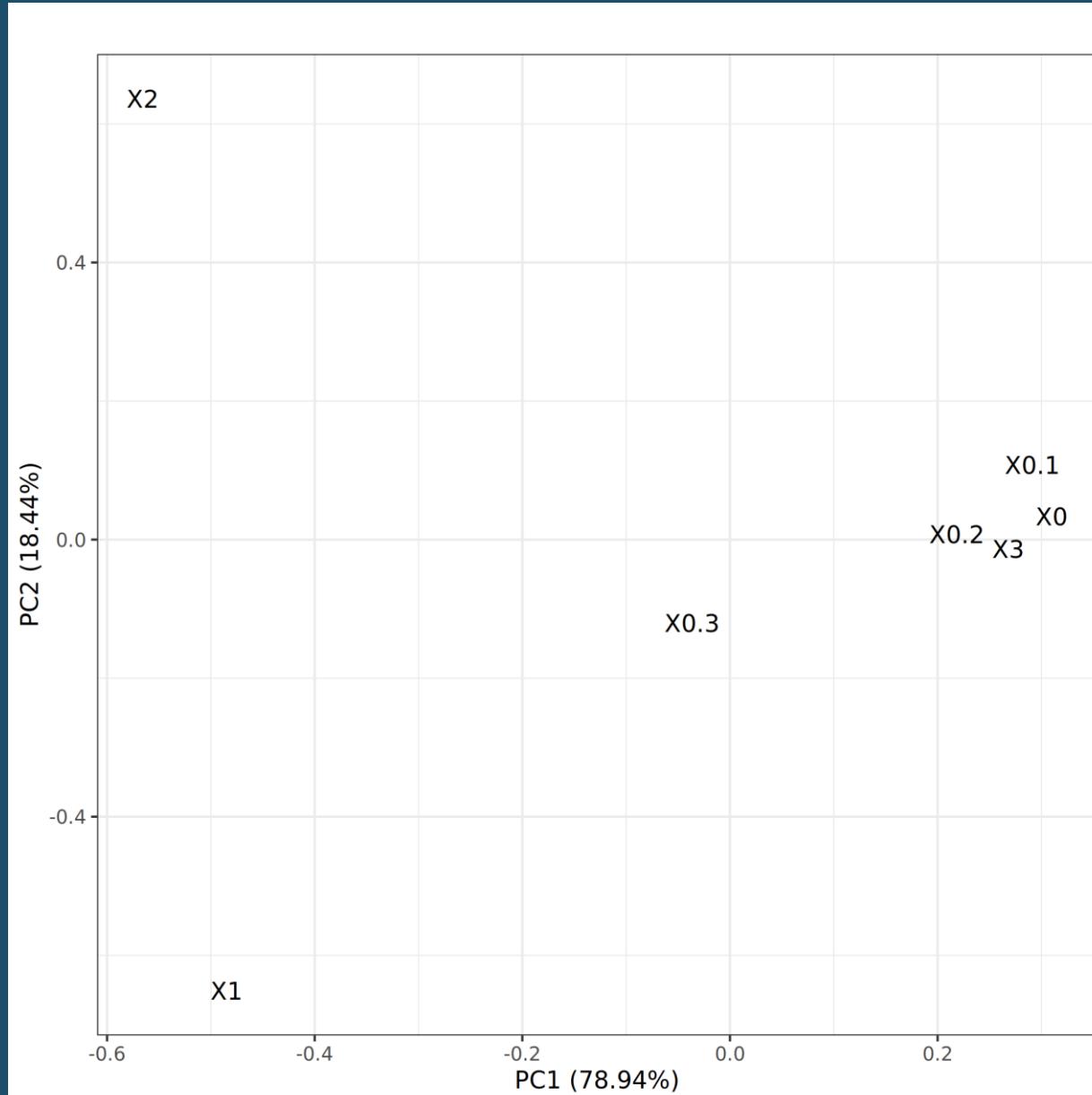
# PCA and Genomic Sequences

- Genetic differences can be compared using PCA
- The allele frequencies of all SNPs are mapped per sample



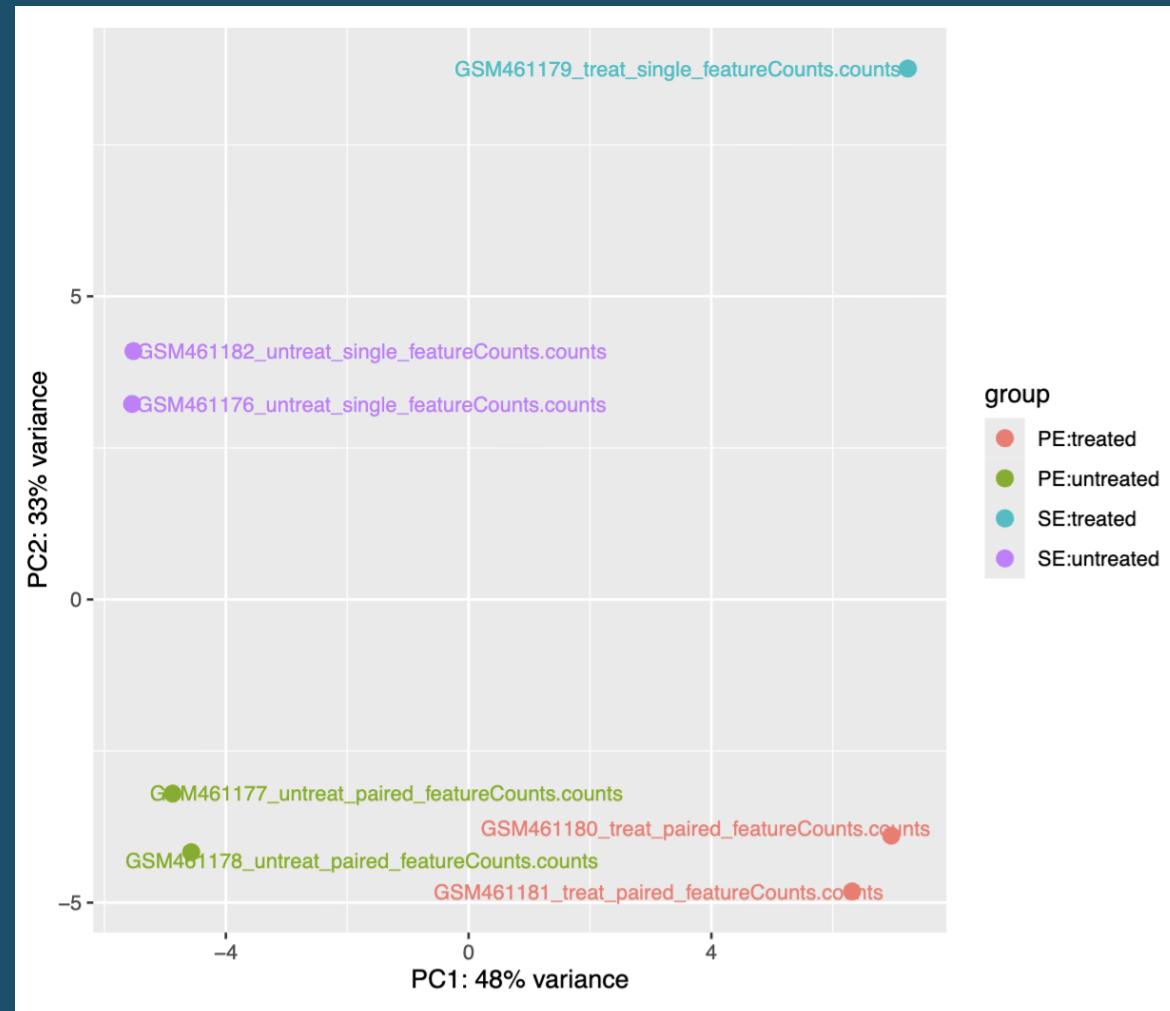
# What Happens Without Normalisation?

- PCA before normalisation...
- PC1 dominated by sequencing depth
- After normalisation, biological grouping appears
- Visualisation is only meaningful after proper normalisation.



# Misinterpretations to Avoid

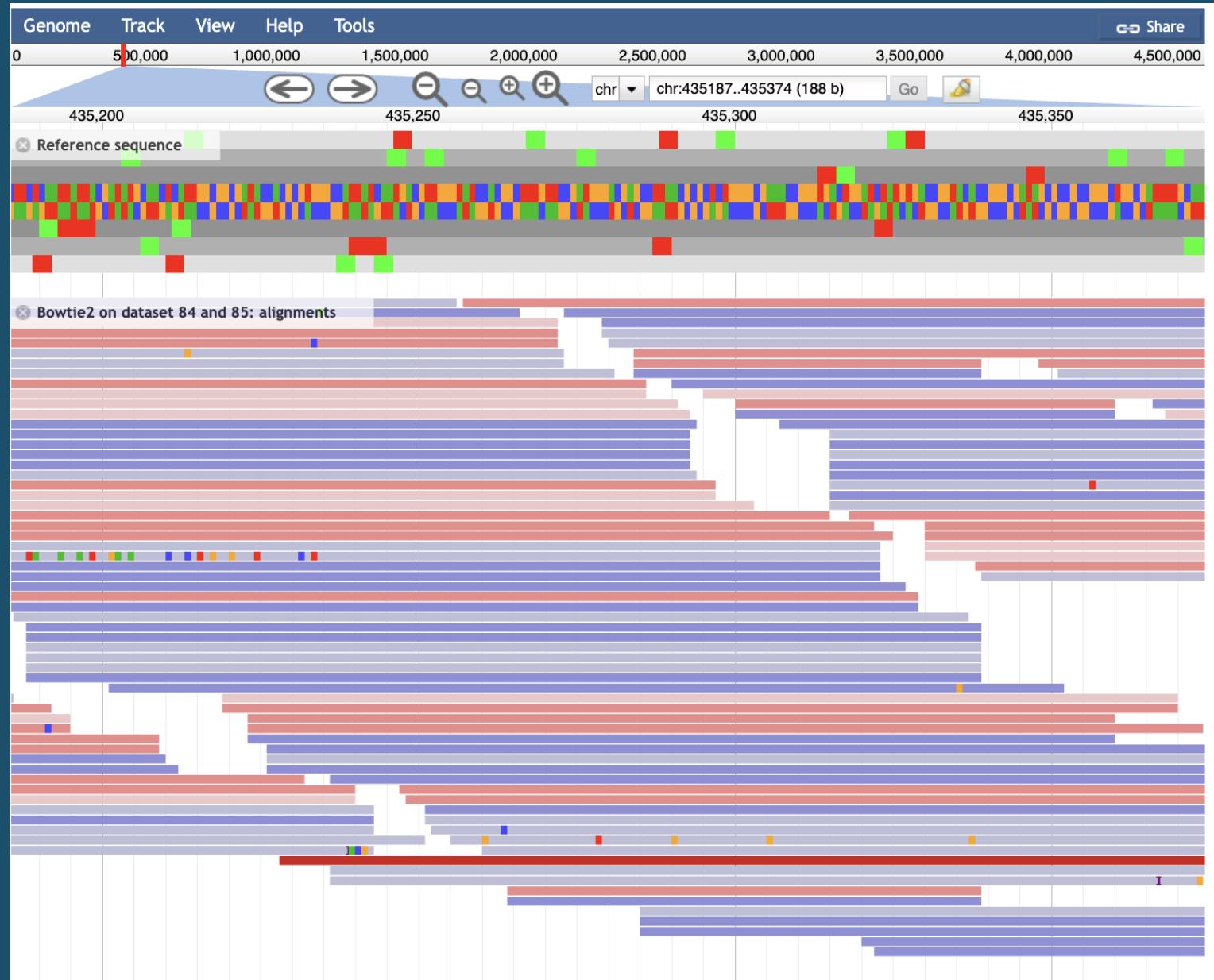
- PCA axes are not specific genes.
- PC1 is not “the most important gene”.
- Clustering does not prove causation.
- Separation does not guarantee statistical significance.



# Visualisation of Genomic Sequences

Last week we could view our reads aligned to the reference genome in the Jbrowse genome browser

While we can also do the same with transcriptomic data, we can also do something a little easier to view...



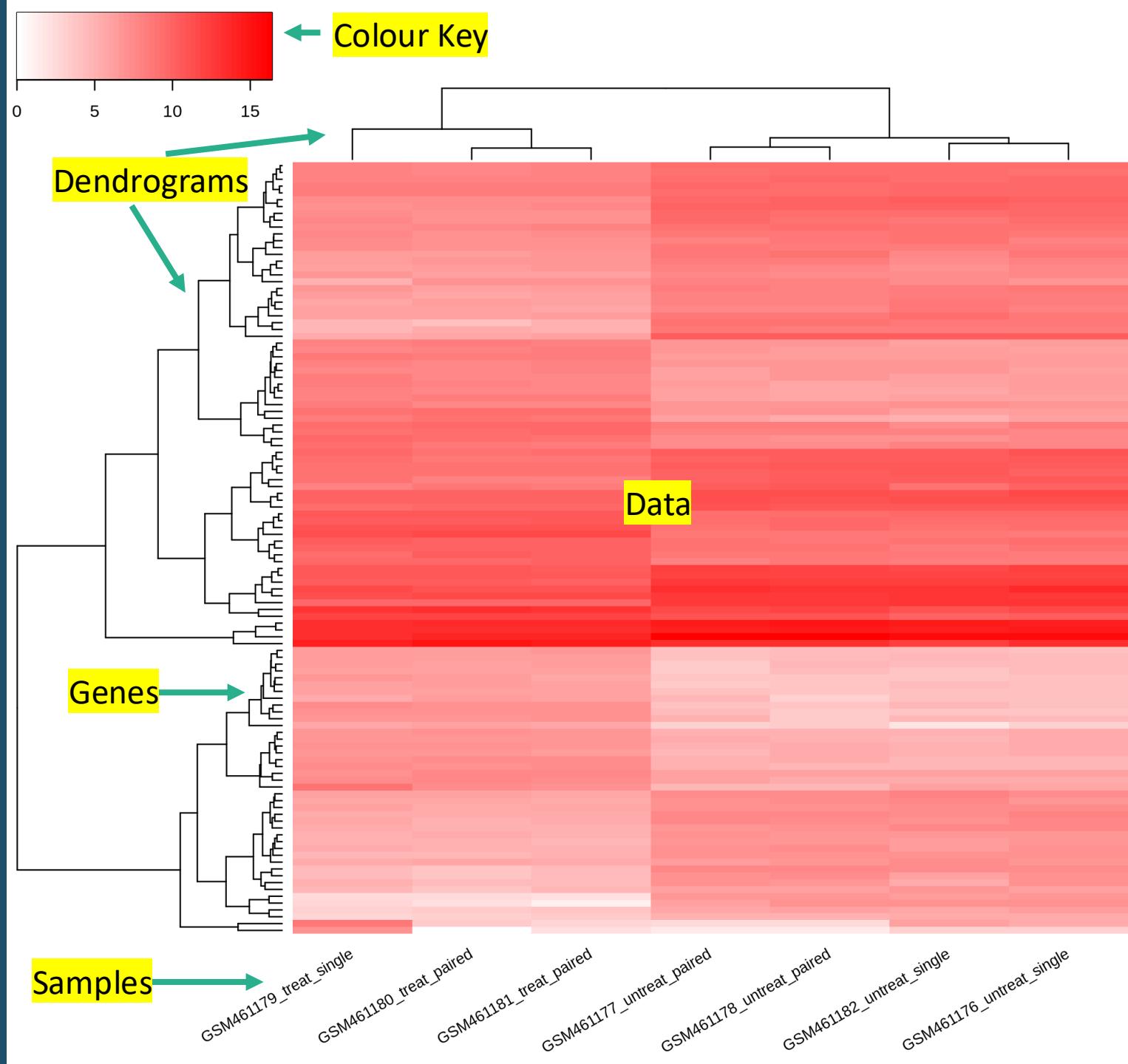
# Heatmaps

Heatmaps represents the magnitude of data points within a matrix - using colour

Colour key that represents real values

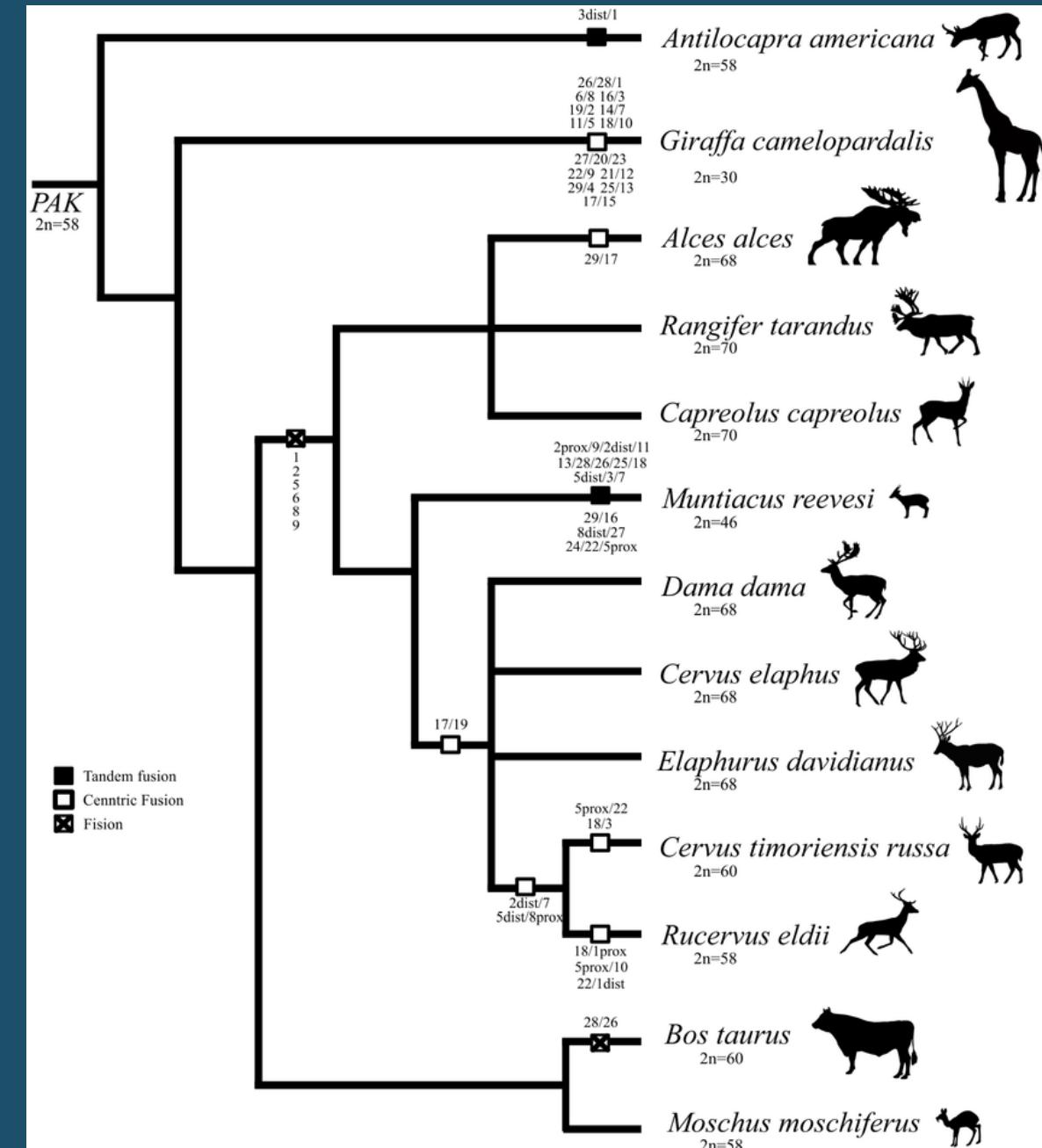
Genes and samples organised by similarity with dendograms

Data displayed for all samples



# Interpreting Dendograms

- Often used in phylogenetics to visualise relatedness between species
- Identify clusters by similarity
- Used alongside heatmaps
- Do biological replicates cluster?
- Are there unexpected groupings?
- Is there an outlier?
- Outlier Detection

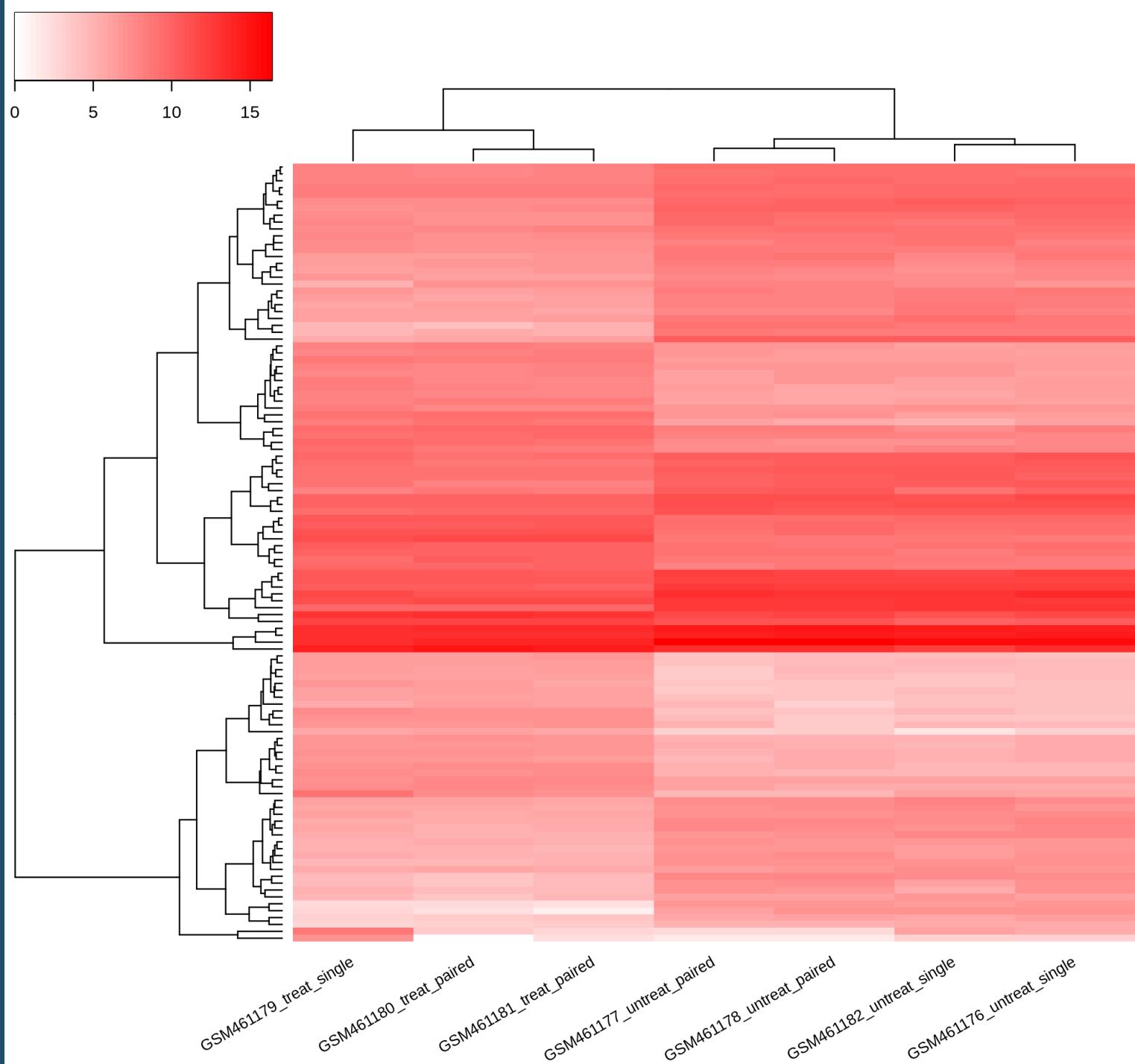


# Interpreting Heatmaps

This heatmap is produced to show the number of gene counts - therefore scale is 0 - max gene count

0 = white = no counts for that gene in that sample

Max count in data = most red.



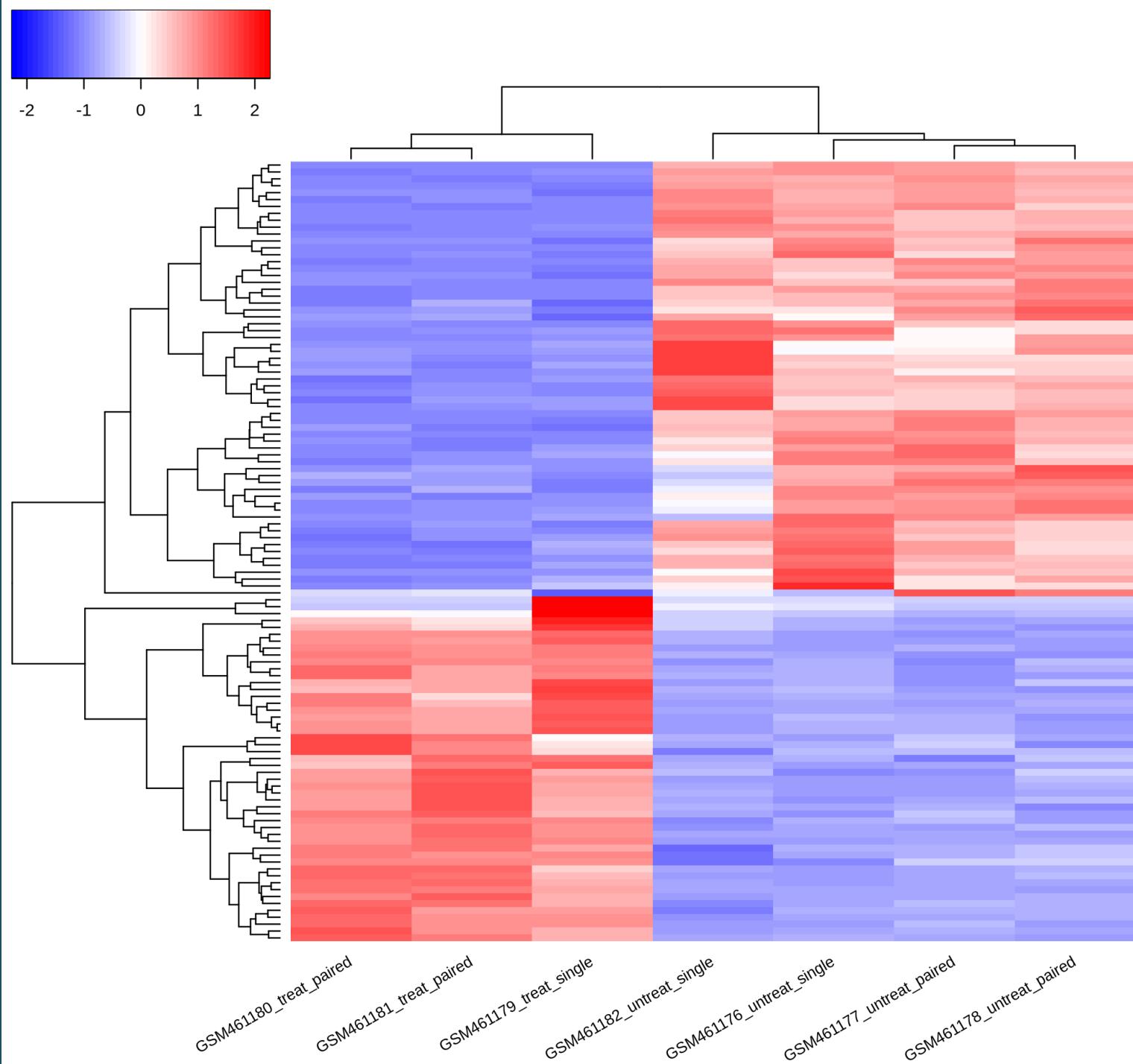
# Interpreting Heatmaps

This heatmap is produced to show differentially expressed genes - therefore scale is multidirectional - negative, zero, and positive.

0 = white = expression value is 0 = no difference to normal

Positive = red = expression is greater to other condition

Negative = blue = expression is less than other conditions



# Heatmap Limitations

- Does not reduce dimensionality
- Better to subset data - lots of genes/samples difficult to interpret
- Relies on colour gradients
- Visual intensity does not mean statistical significance

# To Conclude

- Omics data is high-dimensional.
- Visualisation helps detect structure.
- PCA summarises variation.
- QC & normalisation must come first.
- Visualisation is exploratory, not confirmatory.

Any Questions?