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 dendrograms, 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**
Measures sequence variation	*Measures gene expression*
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)….

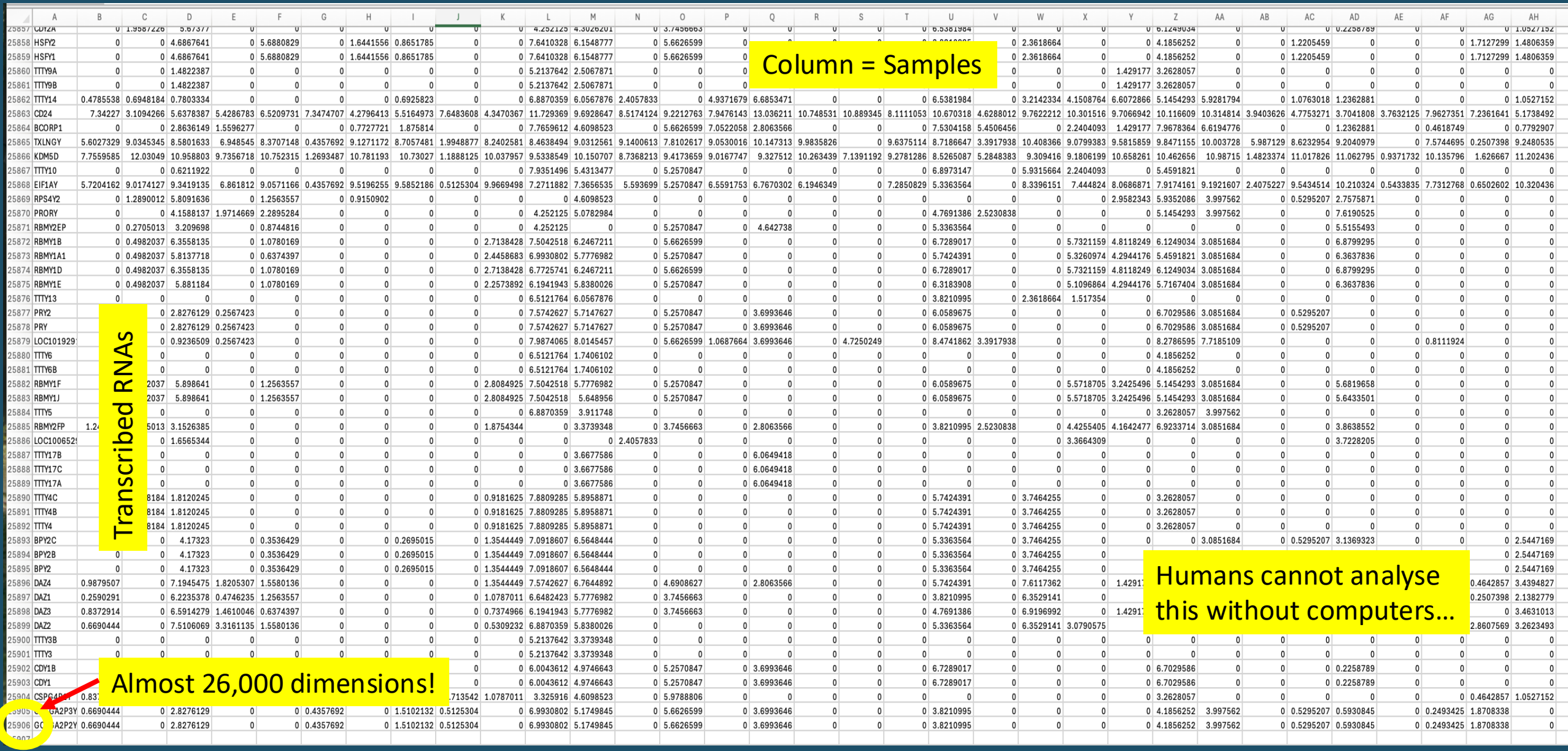
What is High-Dimensional Data?

- 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 loosing 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…



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
- Dendrograms

Dimensionality Reduction - The Issue

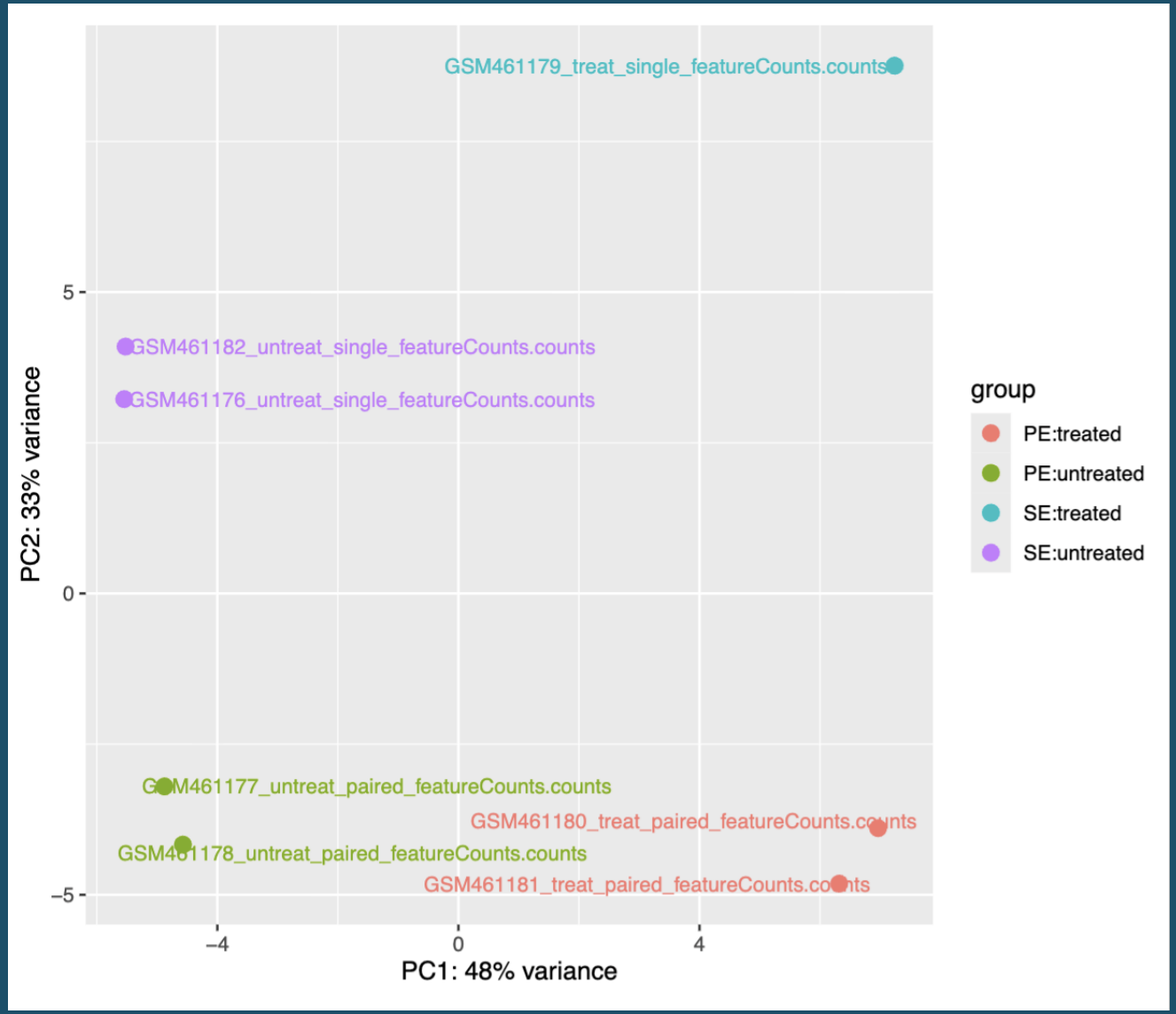
- If each sample is 20,000-dimensional:
- How do we represent it in 2D?

- **Dimensionality reduction**
  - A mathematical method that reduces many variables into fewer composite variables.

Principal Component Analysis (PCA)

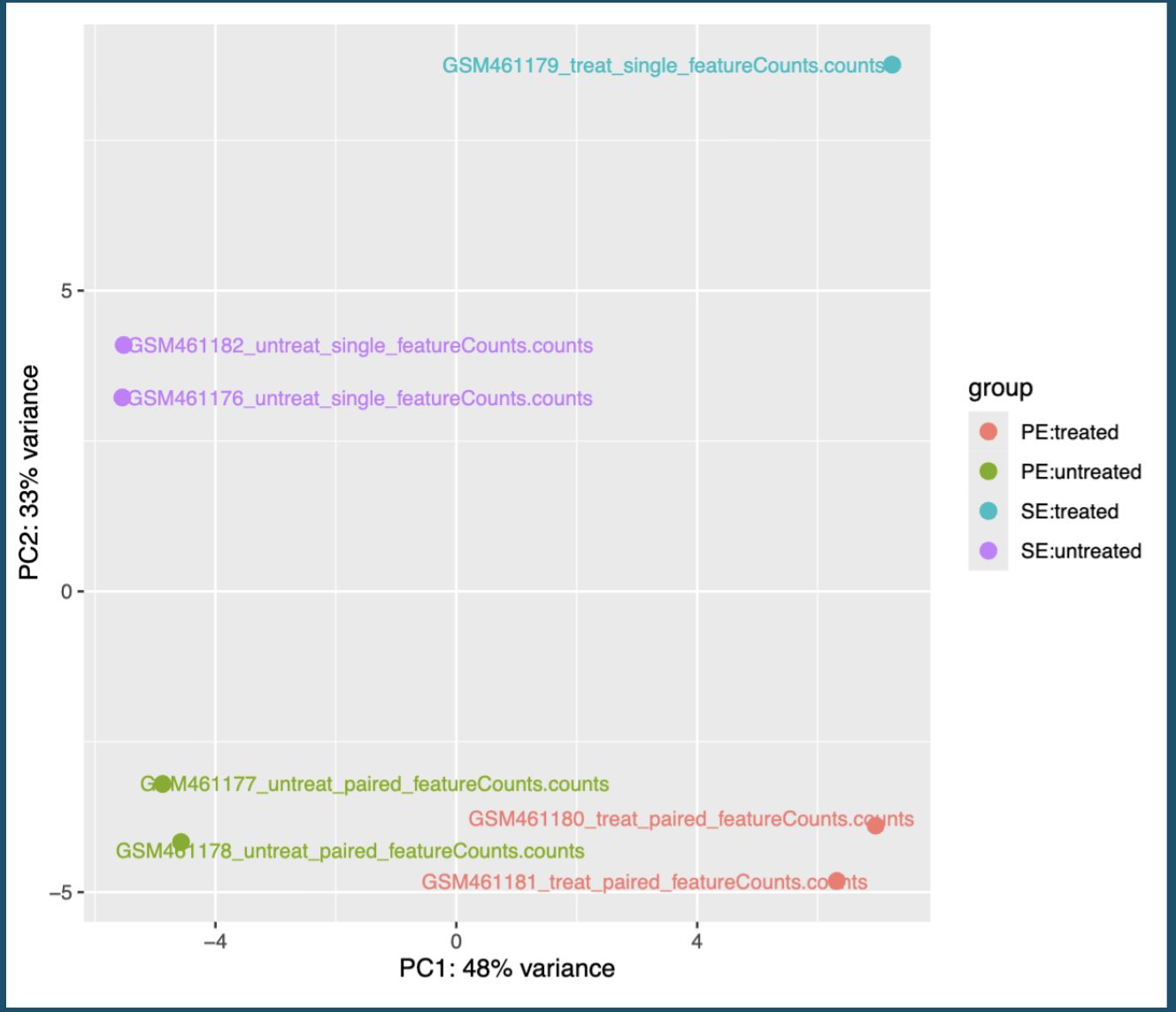
- Finds directions of maximum variance between samples
- PC1 explains the greatest variance
- PC2 explains the next greatest, orthogonal to PC1

- **Principal components are weighted combinations of all genes**



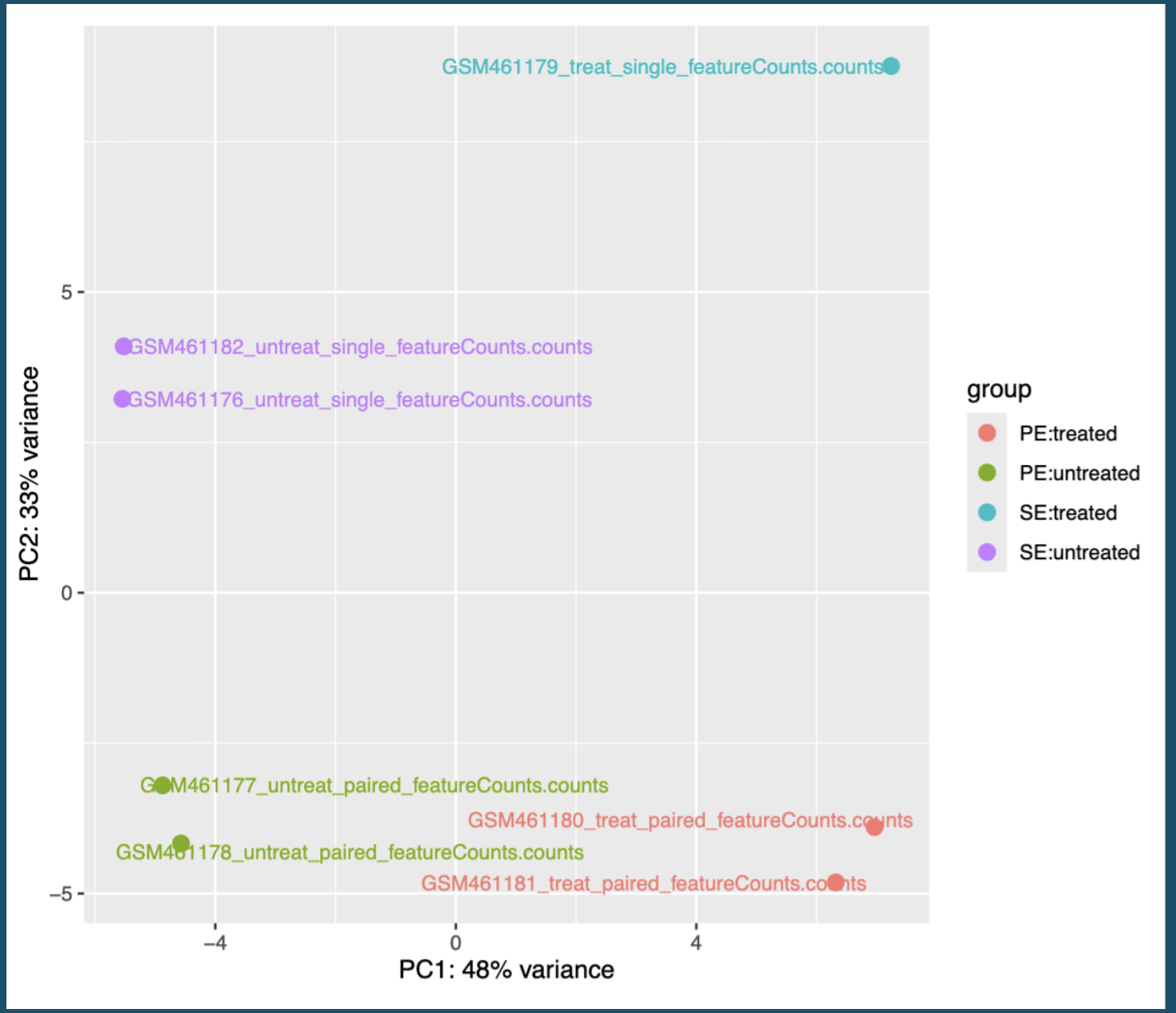
What PCA Actually Does

- 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

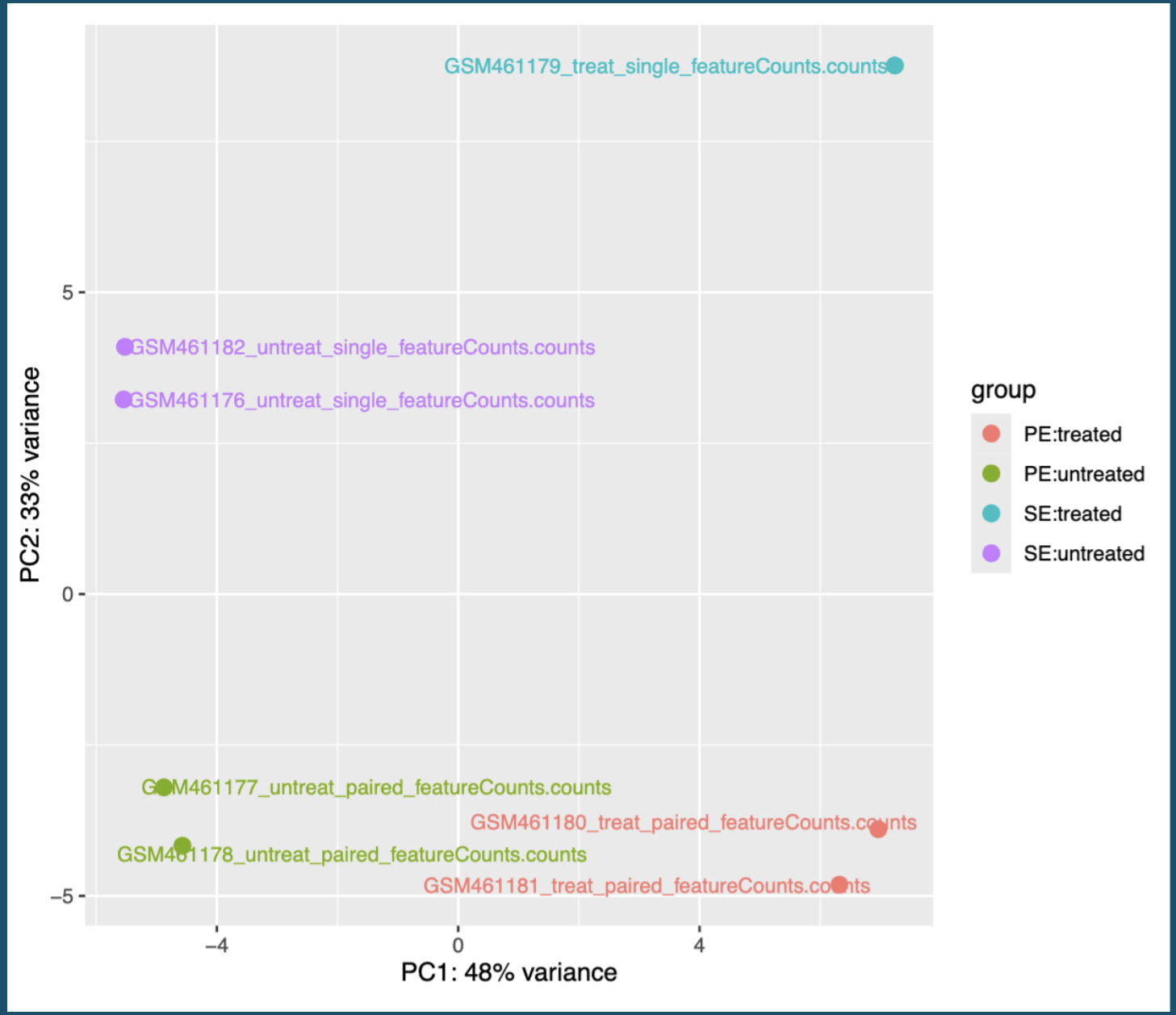
In well-controlled experiments testing distinct biological conditions, we would expect the largest source of variation to reflect the biological differences being measured. The next largest sources of variation often arise from technical factors, such as differences in sample preparation or sequencing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

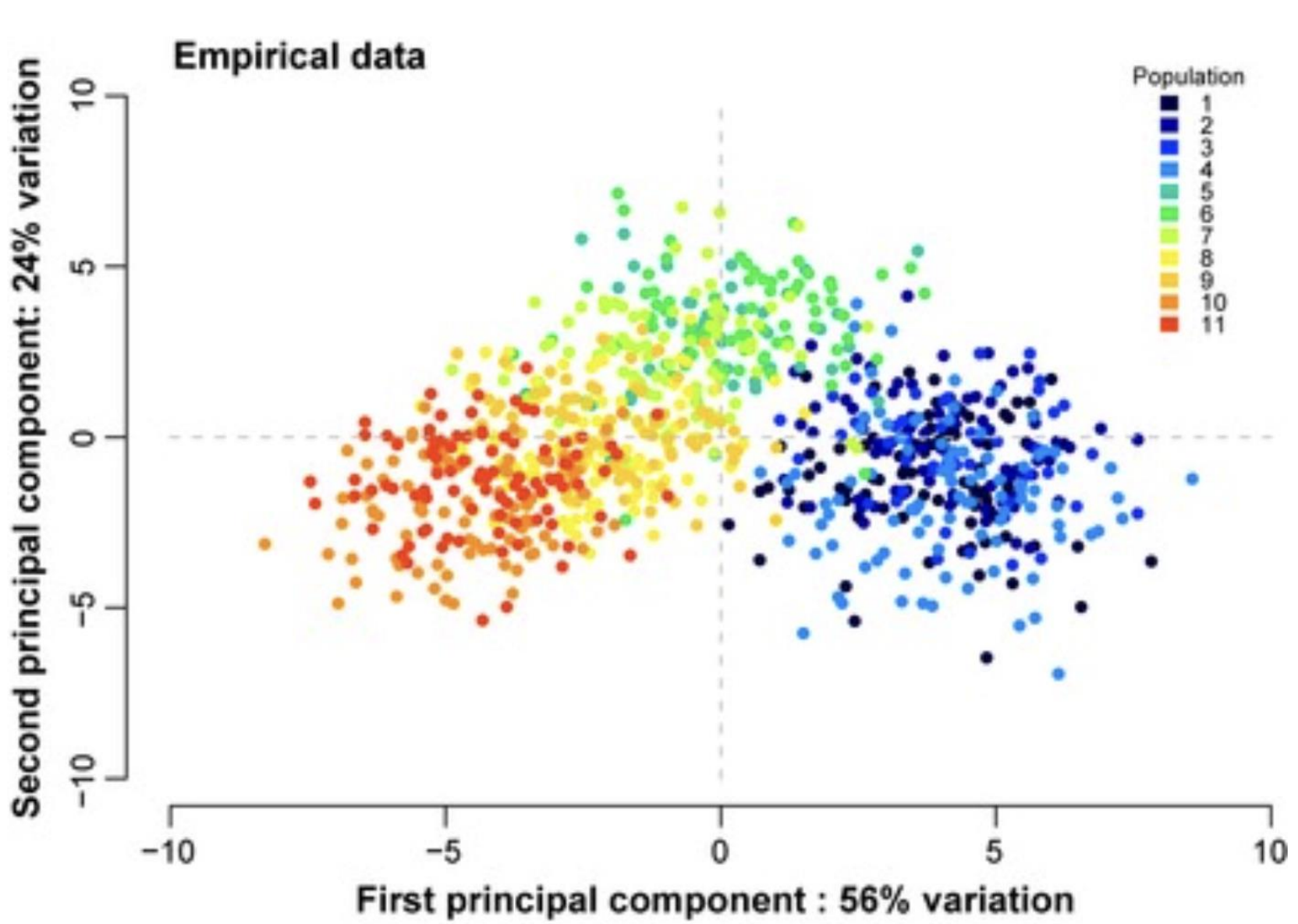
Outliers in PCA

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



PCA and Genomic Sequences

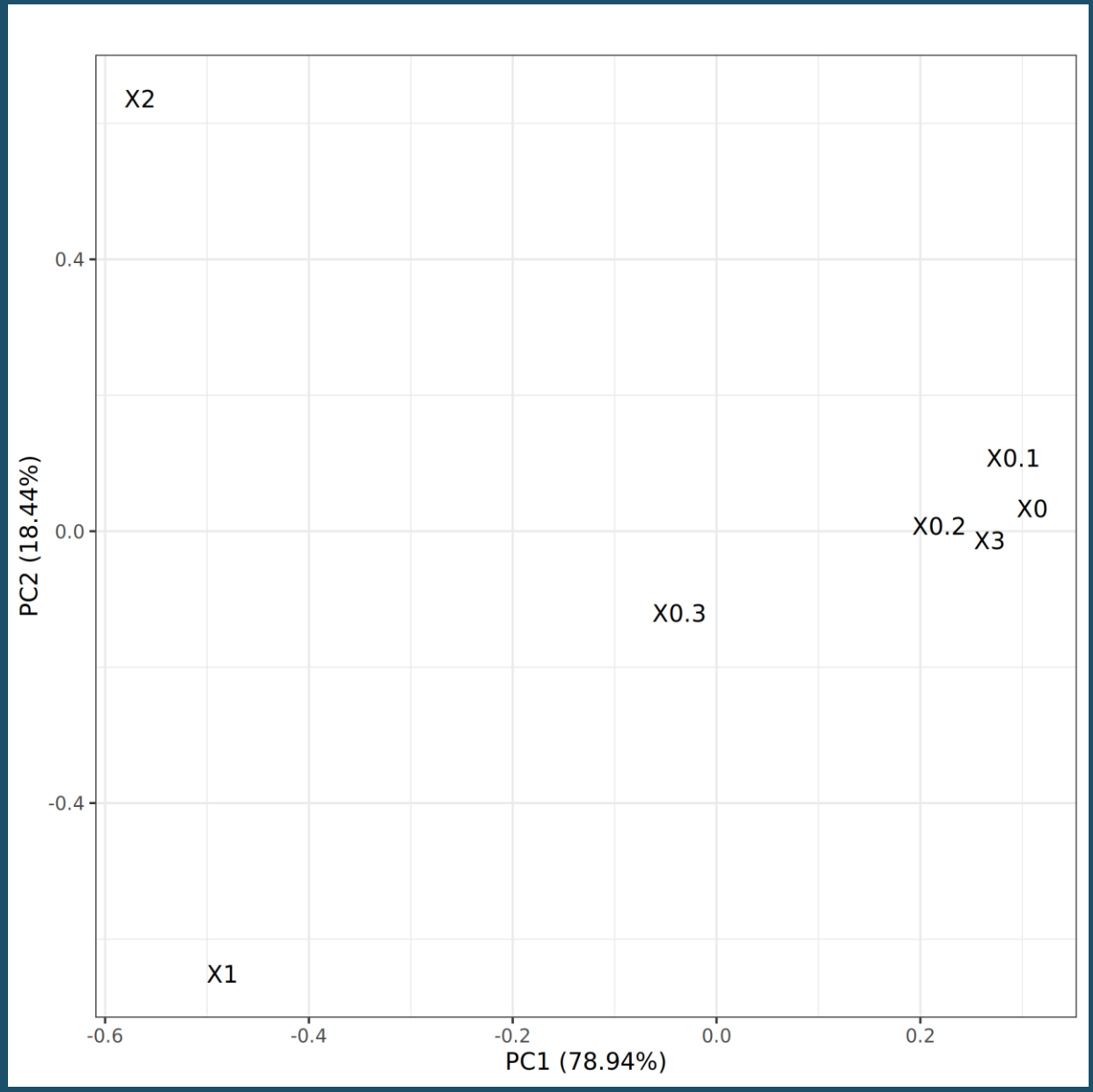
- Genetic differences can be compared using PCA
- The allele frequencies of all SNPs are mapped per sample



PCA of individuals based on the allele frequencies over all SNPs. Each dot is an individual, and colours represent the different populations sampled. Populations are numbered as in Table 2, moving from east to west along the Alaska Peninsula.

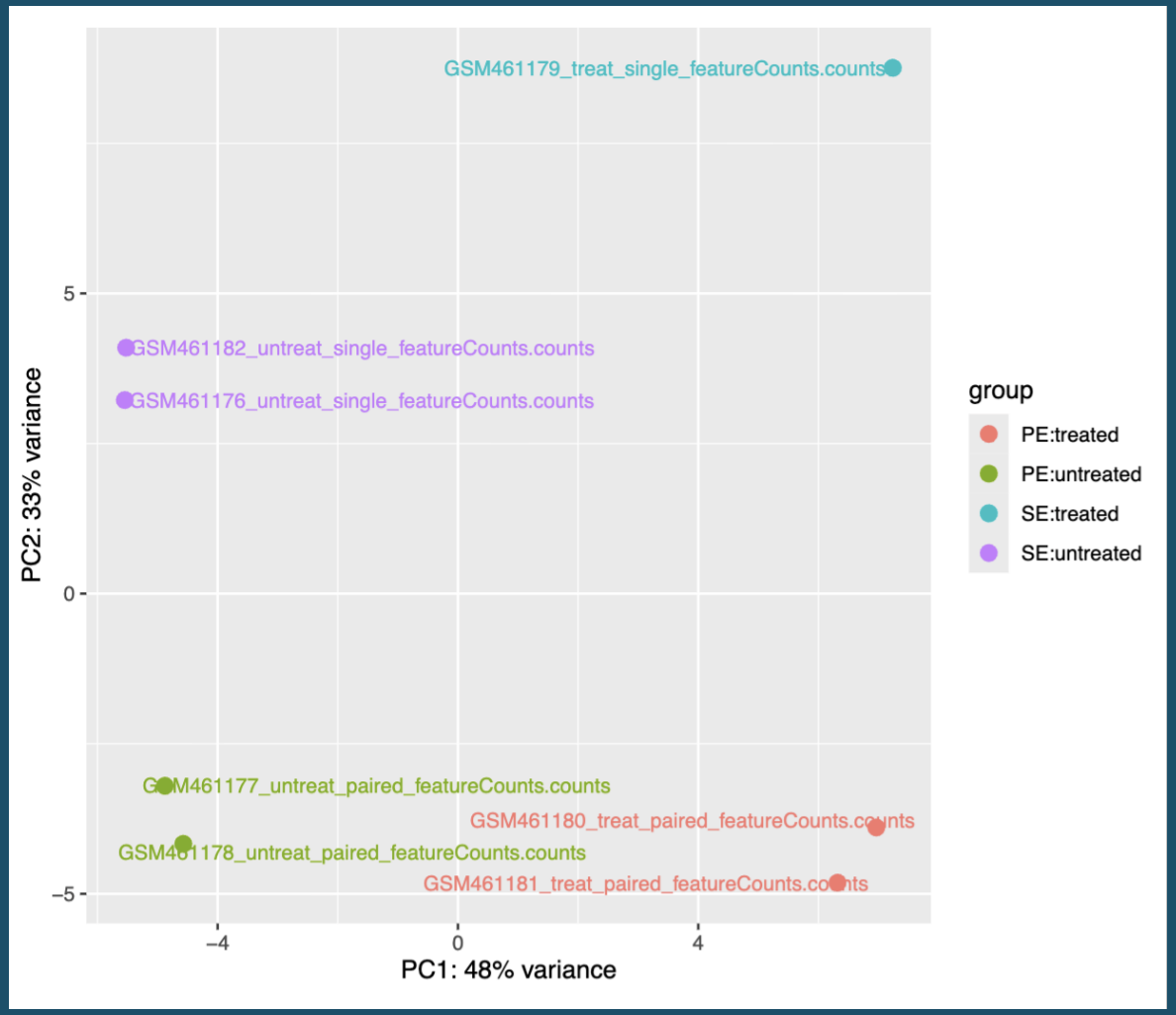
What Happens Without Normalisation?

- 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…

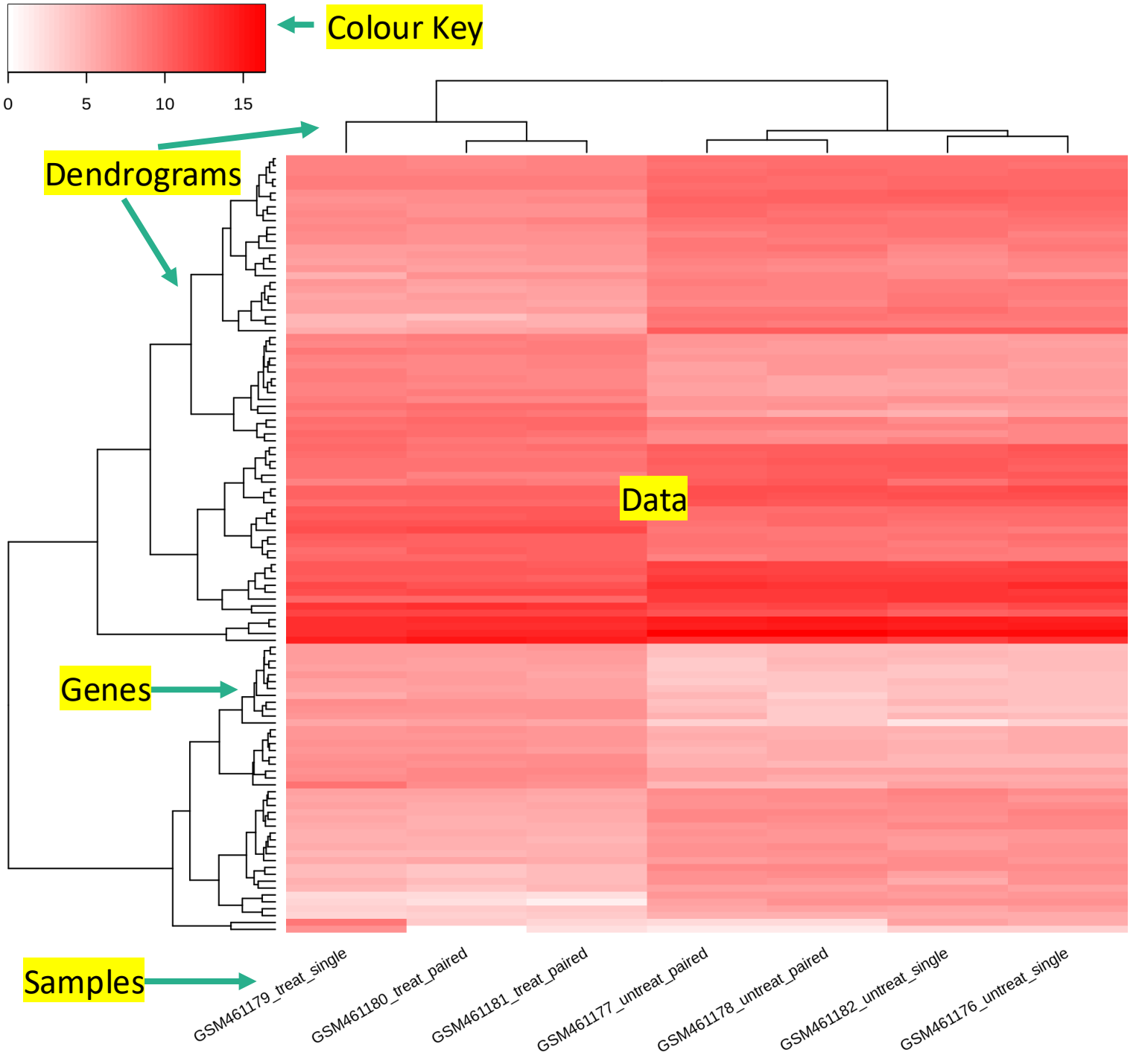
Heatmaps

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

Colour key that represents real values

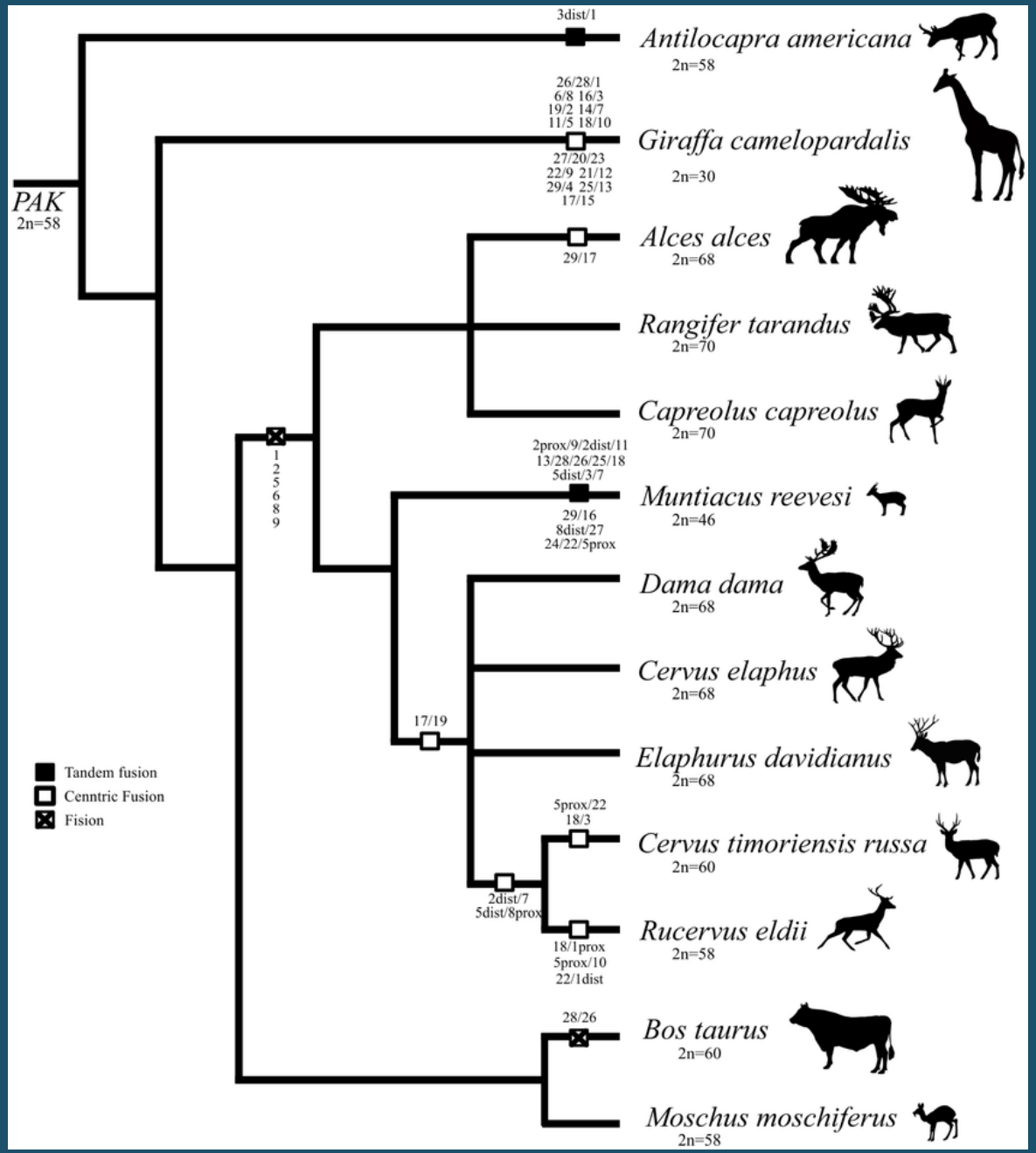
Genes and samples organised by similarity with dendrograms

Data displayed for all samples



Interpreting Dendrograms

- 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

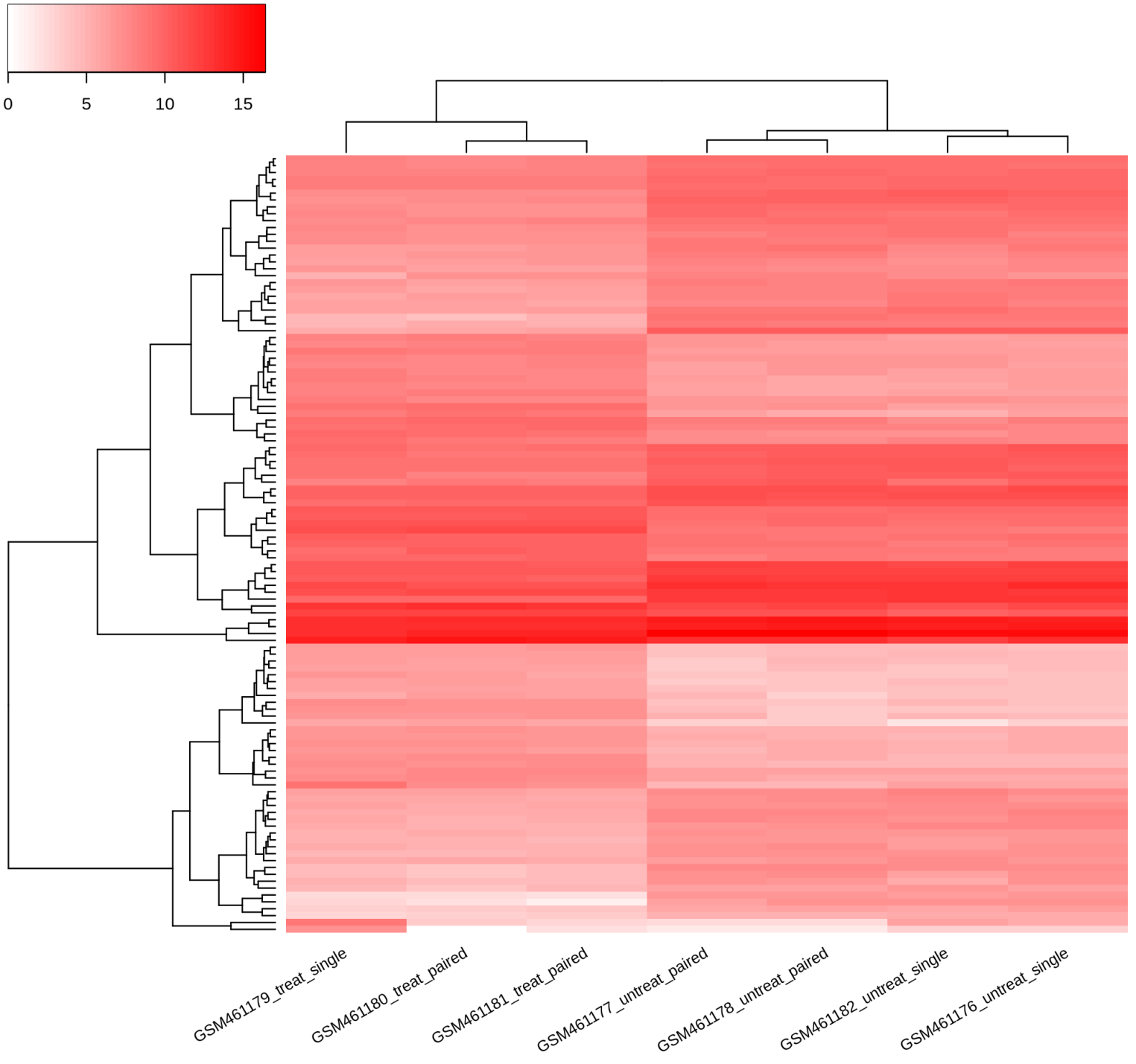


Interpreting Heatmaps

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.



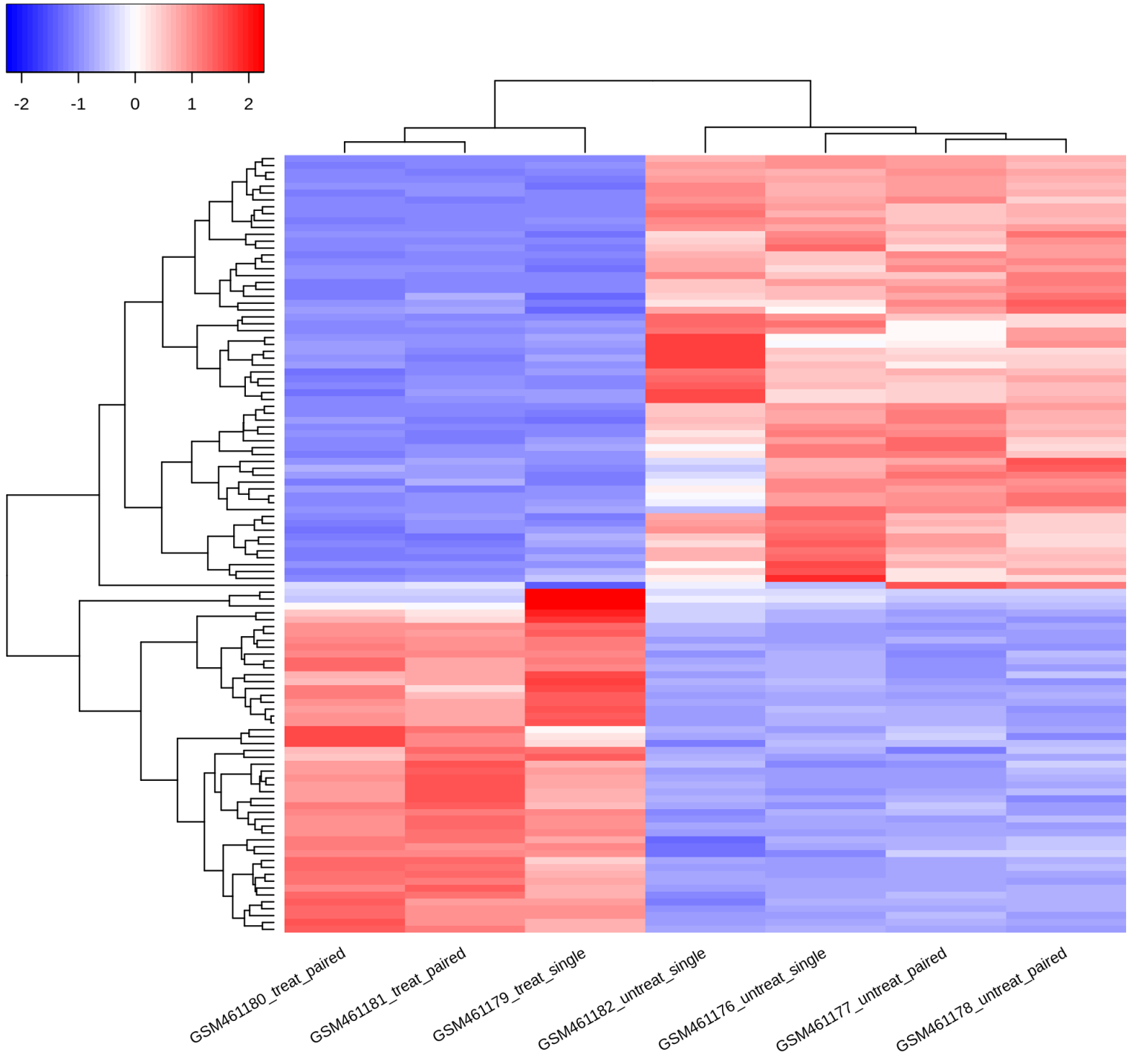
Interpreting Heatmaps

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?