wellcome centre human genetics

RNA Sequencing: Processing and QC

November 2021

Ben Wright
Helen Lockstone

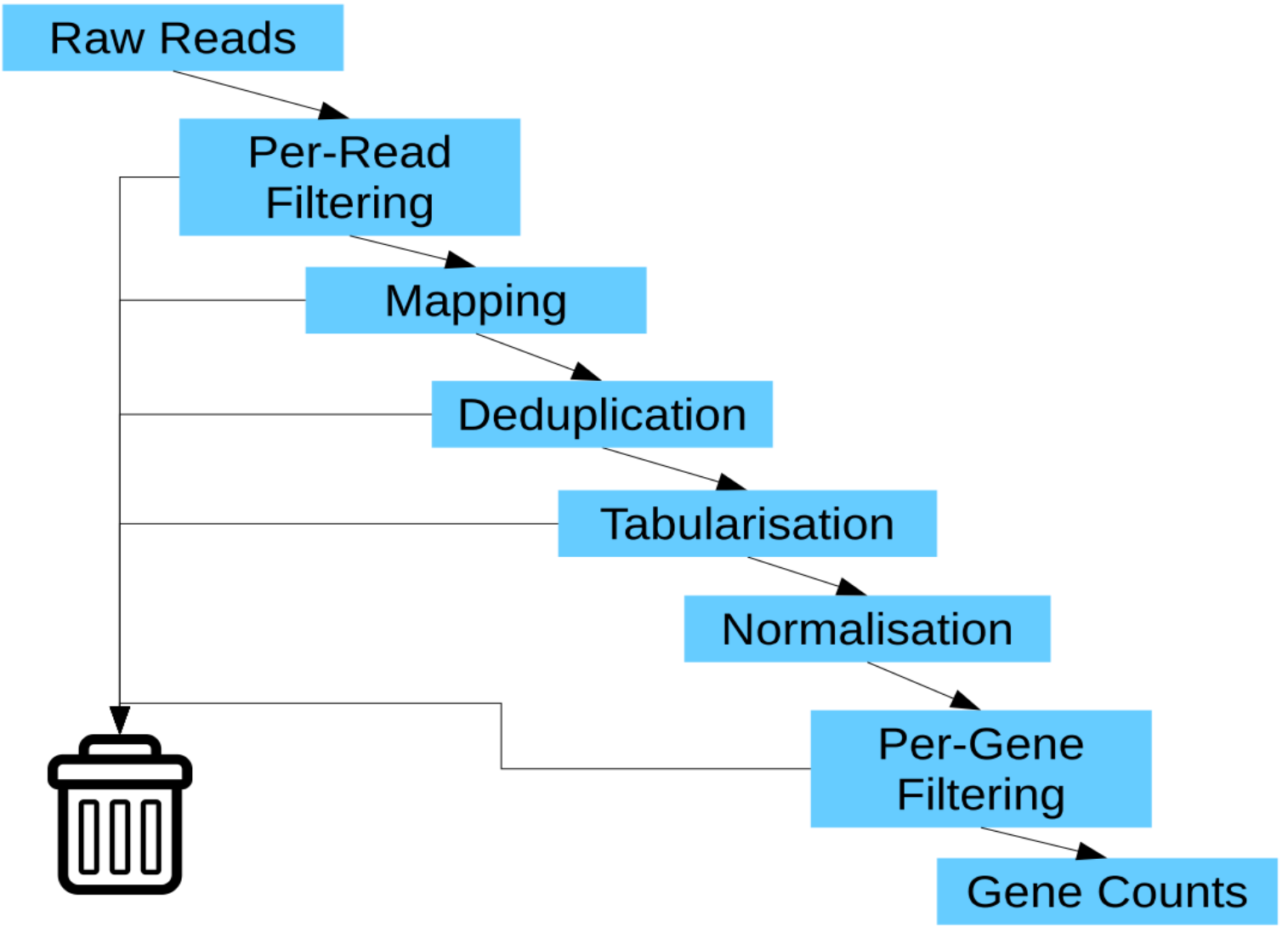
Goals

- Learn how the data from a sequencer is turned into usable gene counts for differential expression (DE) analysis
- Provide examples of what can drive differences between samples
- Recognise common Quality Control (QC) issues in the resultant data

Starting Point

- The raw data we receive from a sequencing machine has already had some QC performed on it
  - Summary metrics that look for obvious technical issues
  - Some of this may be performed internally in the machine, the rest is done as part of the regular process before the data is 'delivered' to the analyst

Reads to Leads



Sequencing Data

- The most common format for data fresh from the sequencer is .fastq
  - Four lines of data for each read:
    - Read identifier
    - The base calls
    - '+'
    - The quality score for each of those base calls
- This file contains no information as to where in the genome these reads came from

Pre-processing

- The data may need pre-processing before the main pipeline steps
  - For example: sequences from the adapter can appear in the data, requiring the use of a tool to trim those sequences from the ends of the reads

Mapping

- A mapping tool assigns each read to its likeliest position in the transcriptome
  - Requires a reference to which each read is compared
  - Positional information is generated with respect to this reference
- These tools can have many parameters to fine-tune their algorithms, and they are in active development

Mapping

- Mappers are tolerant of single-base changes and indels (short insertions and deletions)
- Handling larger or more complicated structural variations requires a specialised tool
- Some mappers also perform local assembly to resolve more complicated small-scale variations

Mapping

- After mapping each read is in one of these categories:
  - Assigned (given a definitive position)
  - Unmapped (due to poor quality scores, a mismatch to the reference)
  - Multiply-mapped (map equally well to more than one place in the reference)

Mapping

- Output format is typically .bam
  - A block compressed .sam file
  - .sam format is complicated
  - Includes all the information from the .fastq, but also:
    - Position
    - Mapping quality score
    - Match/mismatch for each base with respect to the reference

Mapping - What Reads We Lose

- Reads with overall low quality
- Reads that otherwise can't be mapped
- (Reads that map to multiple locations)

Deduplication

- If there is a wide range of potential positions for reads, then reads should be spread fairly evenly throughout this range
- If two reads share the same position, it is likely that this is an artefact in the data and only one of those reads should be retained for further analysis
- This process is called 'deduplication'

Deduplication

- A deduplication tool takes a .bam file as input, and produces a .bam file as output
- The output is either:
  - The input with the duplicates removed
  - Identical to the input, but with duplicates annotated as such so the next tool knows to ignore them
- It also produces a summary of the overall duplication rates per sample

Deduplication

- Deduplication should only be performed when the assumption of a wide range of positions is true
  - Specifically, 3' sequencing has a much narrower range of positions and deduplication should not be performed
- Perhaps counterintuitively, higher reported duplication rates are a reason to *not* perform deduplication on the data

Deduplication - What Reads We Lose

- Reads determined to be duplicates

Tabularisation

- The remaining reads are then assigned to genomic features (usually genes)
- Requires an input file defining those features
  - Must match the reference used for mapping
- The output of this step is a plain text table
  - Simple counts of the number of reads for each feature, per sample

Tabularisation - What Reads We Lose

- Reads that fall outside one of the defined genomic features (No Feature)
- Reads that span two features and thus can't be assigned definitively to one or the other (Ambiguity)

Normalisation

- Comparison of raw read counts between samples is not informative
- Some normalisation is required
- Many algorithms to perform normalisation
  - May put the data on a log scale as a side effect
- Normalisation is usually done automatically by R's DE packages
  - It is still useful to examine normalised data before the analysis

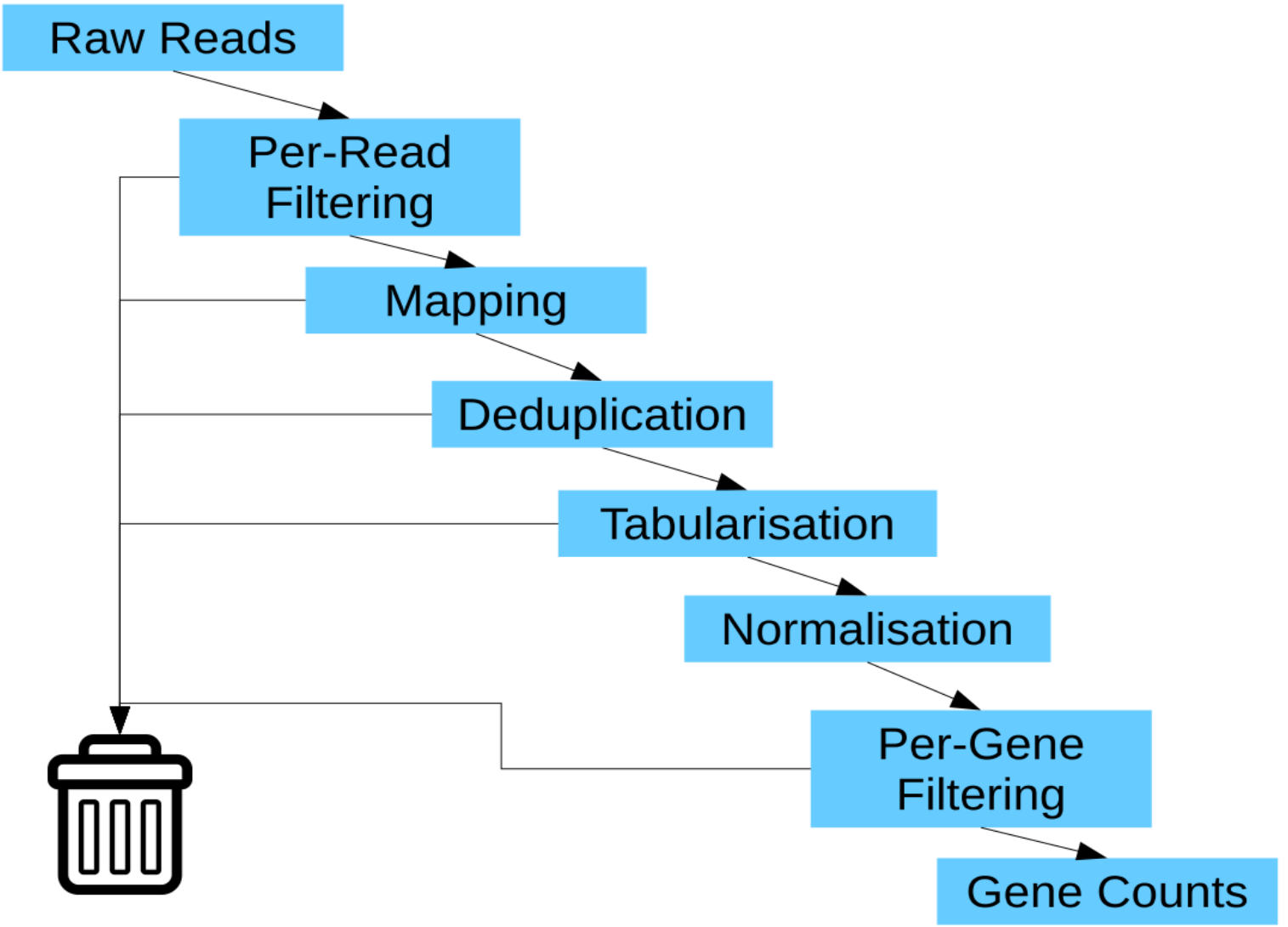
Per-Gene Filtering

- Either before or after normalisation, it is useful to discard genes with insufficient reads to merit analysis
  - Includes genes with 0 counts for all samples
  - A gene should have these low counts in *all* experimental groups to be removed this way
- Where the expression is low, DE analysis will not be reliable

Per-Gene Filtering - What Reads We Lose

- All the reads assigned to a gene with counts too low for reliable analysis

Reads to Leads



What Drives Differences?

- Quality issues
- Technical aspects
- Differential expression

Quality Issues

- Quality issues can arise at every stage of the project:
  - Sample gathering
    - Batch effects, sample labelling issues, poor experimental design, etc.
  - Lab work
    - Contamination, low input material, batch effects, etc.
  - Sequencing technicalities
    - Sequencing machine problems

Dealing with Quality Issues

- Many of these problems manifest as a systematic change in the expression profile of the affected samples
  - Where this causes a sample to 'fail', it must be excluded from analysis
  - Where this causes a sample to differ from all the others significantly, it should be also be excluded
  - Where a group of samples has a consistent difference from the rest, it may be possible to adjust for it in the analysis

Technical Aspects

- Technical aspects that can affect expression:
  - Tissue type
  - Kit type
  - Other experimental factors that should have been constant for the entire project

Differential Expression

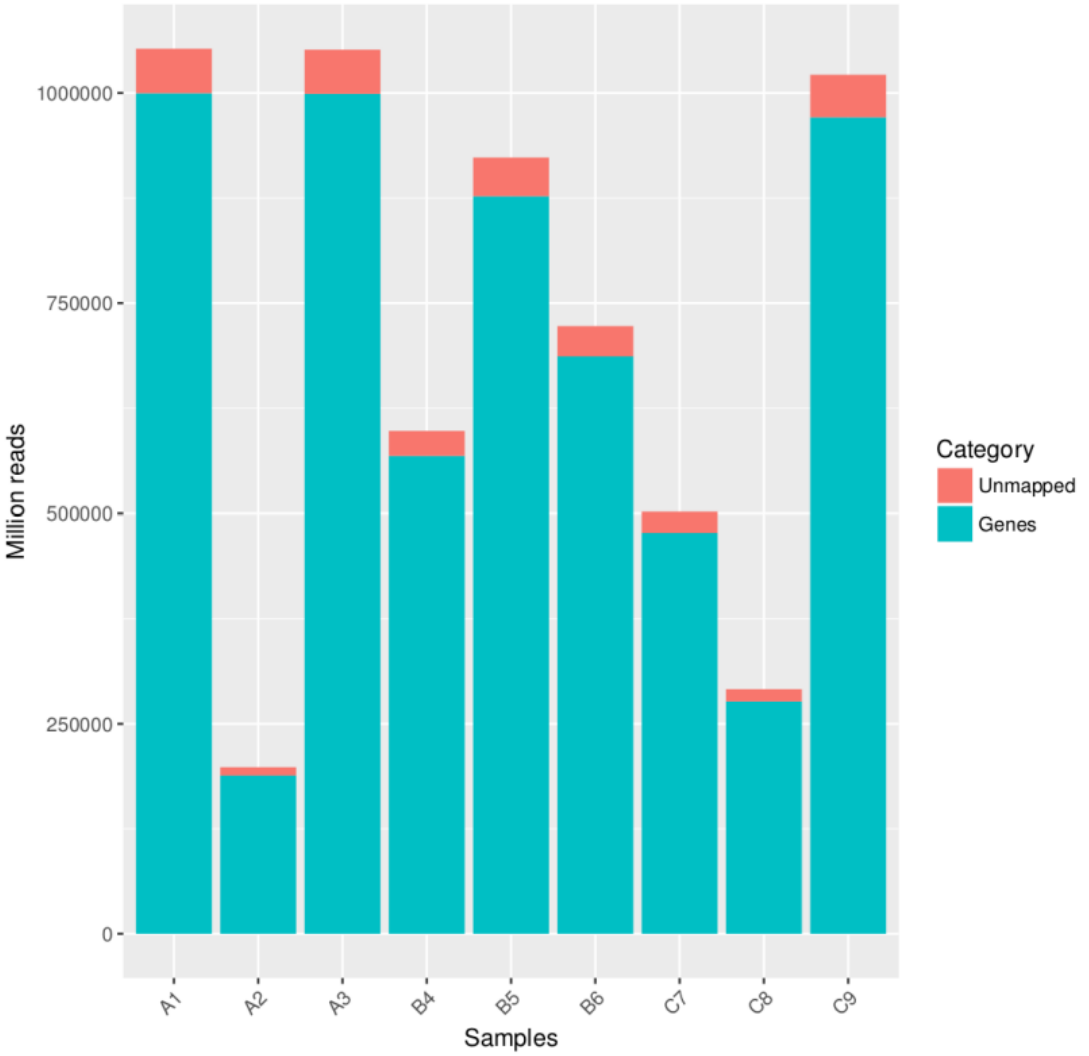
- Common different expression factors:
  - Treatment levels
  - Disease condition
  - Knockdown models
  - Time factors

Visualising QC

- Visual inspection of simple graphs can identify many quality issues
- A given problem will usually show up in more than one of these diagnostic plots
- Being able to recognise likely causes of an anomaly in these plots saves time

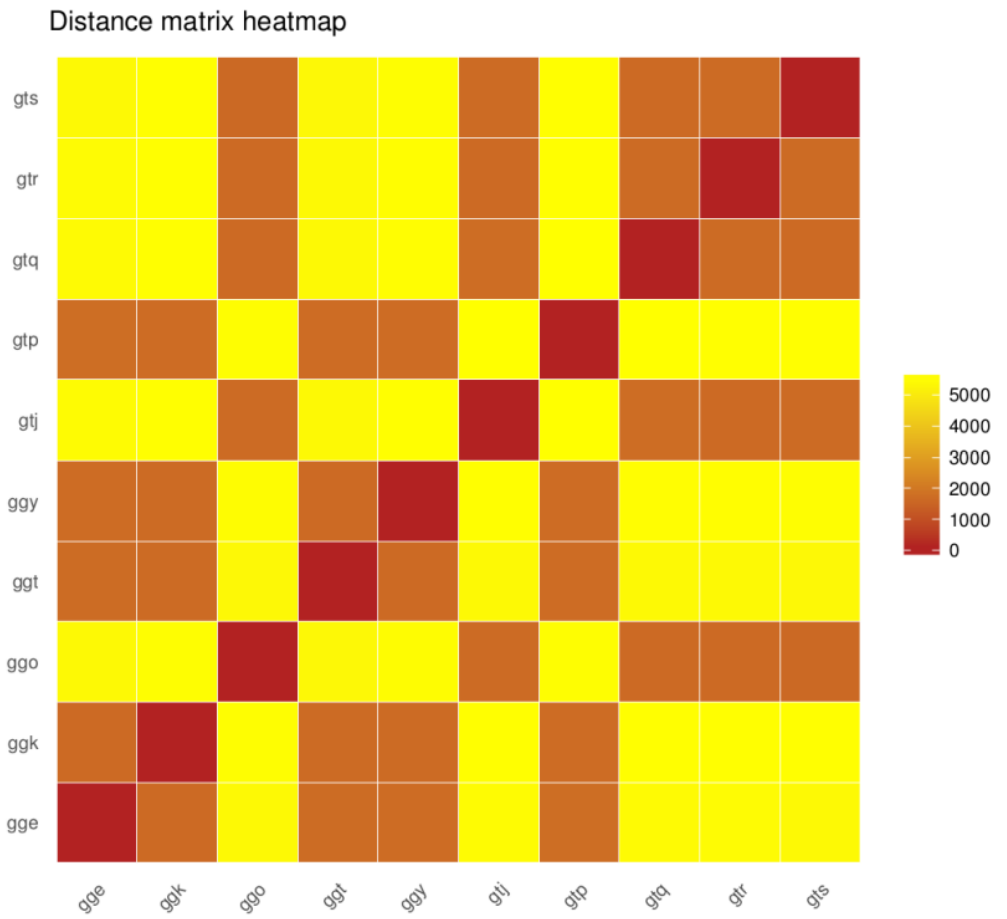
Visualising QC - Read Counts

- Shows how many reads are in each mapping and tabularisation category
- Quick way to spot failed or contaminated samples



Visualising QC - Heatmaps

- Show similarity between samples
  - Many different metrics for this similarity
- Can identify experimental groups
  - Does not show finer structure



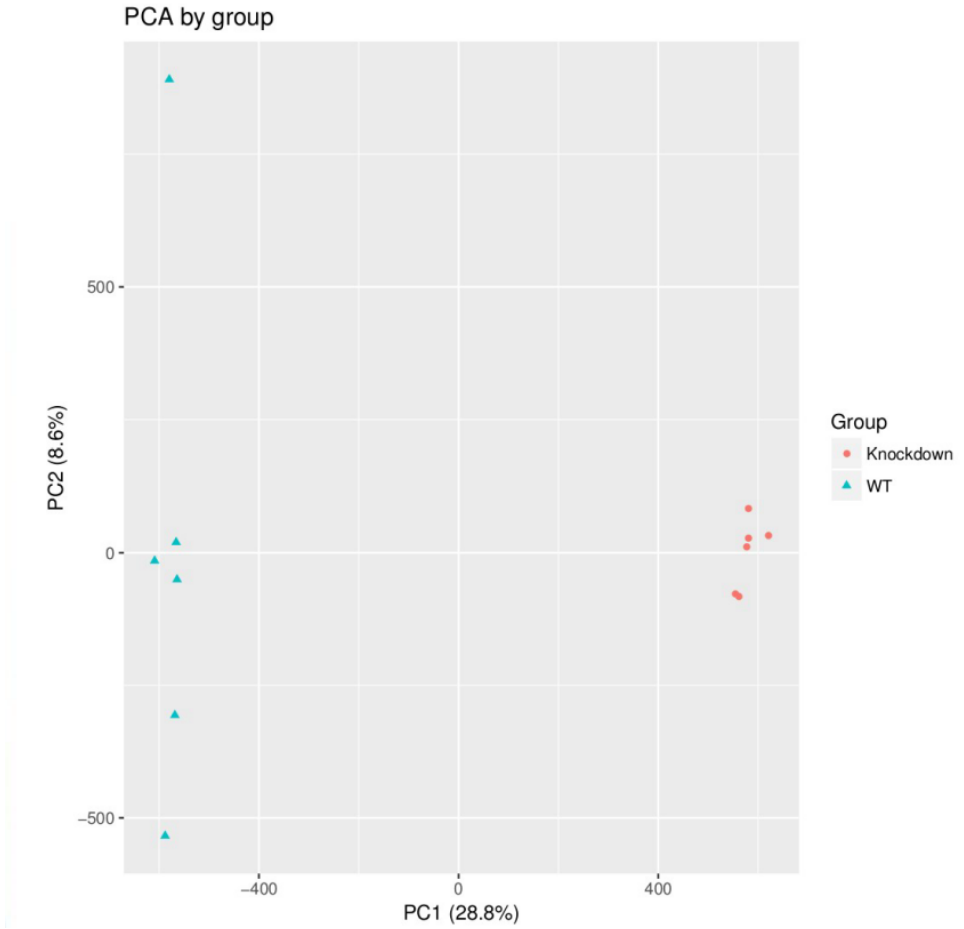
Visualising QC - PCA and MDS

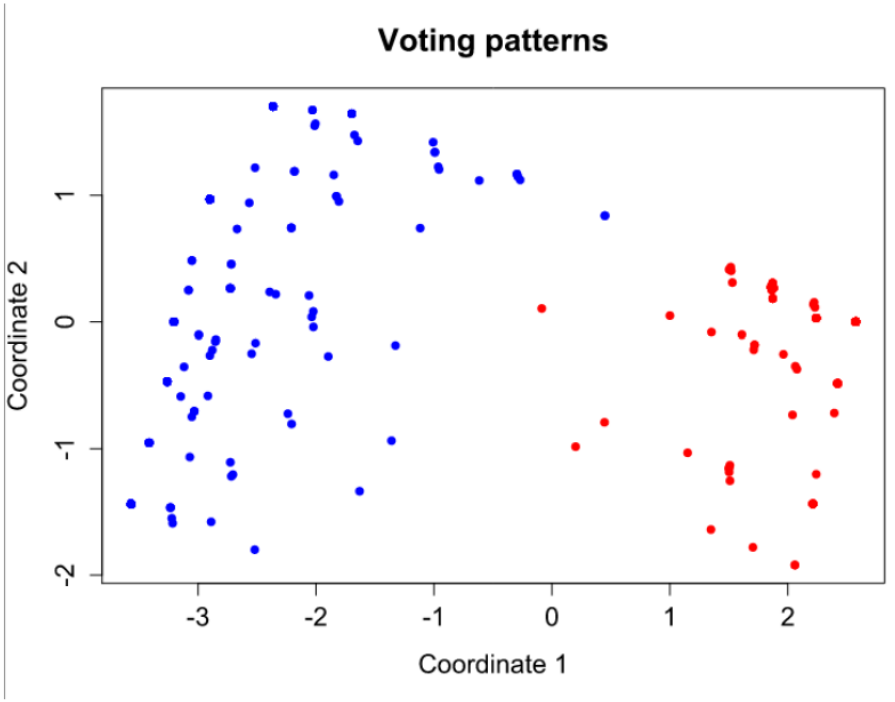
- Both principal components analysis (PCA) and multidimensional scaling (MDS) plots are 'dimensional reduction' algorithms that help expose underlying structure in multivariate data
- Can be used to identify clusters of sample visually
- The basic visualisation only shows the 2 most important dimensions of variability, but you can examine higher dimensions to find deeper structure

Visualising QC - PCA and MDS

- In a perfect world, each dimension corresponds to one source of variation:
  - Dimension 1 might be treated/untreated
  - Dimension 2 might be time since treatment
  - Dimension 3 might be a batch effect
  - Etc.
- PCA, additionally, tells you how much of the total variation is explained by each dimension

Visualising QC - PCA and MDS





Visualising QC - Limitations

- Good QC plots do not guarantee a successful project with useful results
- If gene expression differs only slightly between experimental groups, the underlying pattern may not be apparent in these visualisations
- Where there is an identifiable QC issue, it can require lateral thinking to work out what might be causing it

More Details

- There are more detailed presentations from a 2-day course run previously here:
  - https://www.well.ox.ac.uk/bioinformatics/training/RNASeq_Sept2018/