Gene Counts and Data Quality

27th November 2017
WHG

Ben Wright

Oxford Genomics CENTRE

Gene Counts and Data Quality

- Goals:
  - An overview of process used to get from raw reads to count tables for differential expression analysis.
  - Description of common file formats involved.
  - Know what drives differences between samples.
  - Recognise common Quality Control (QC) issues in data.

Gene Counts and Data Quality

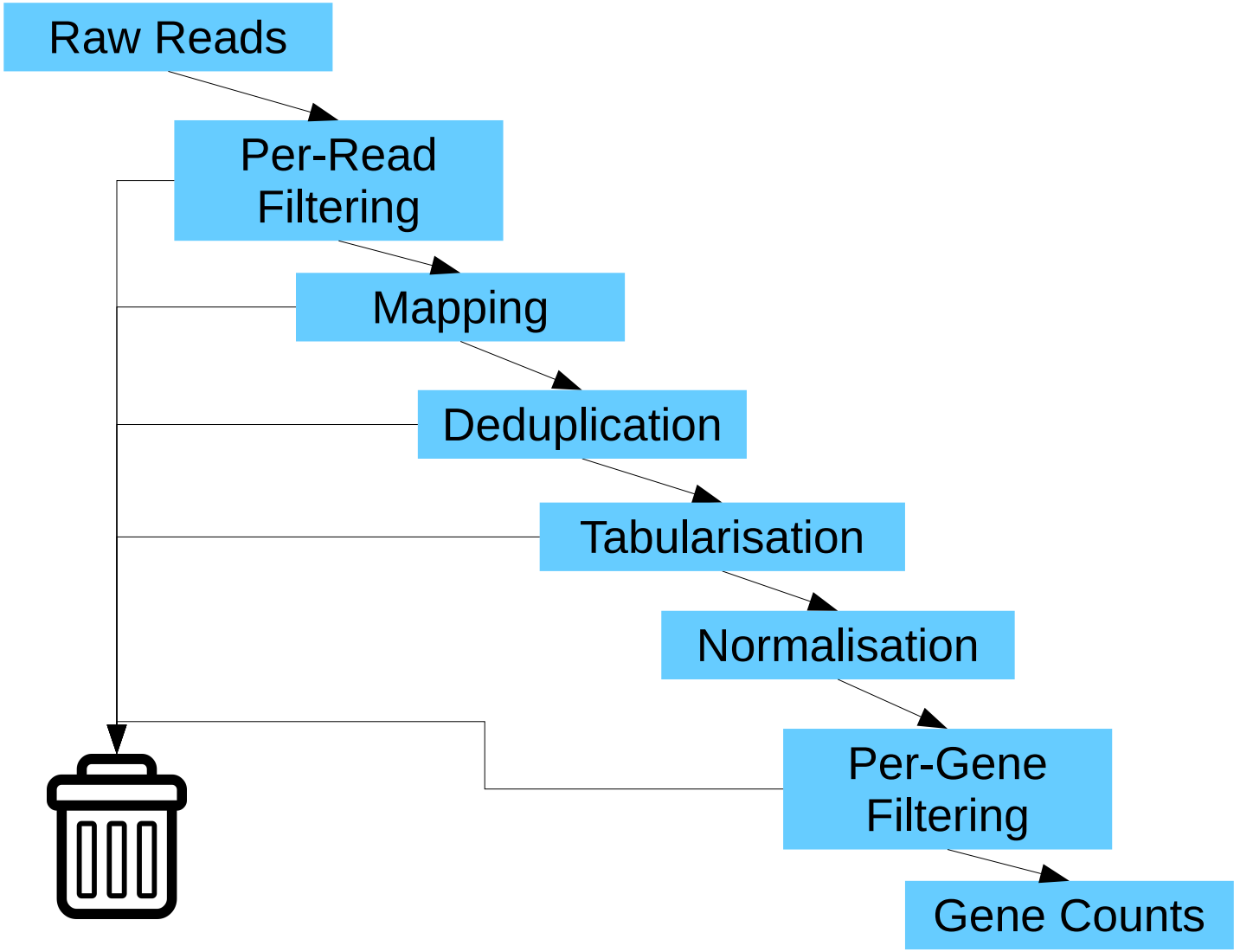
- Presentation:
  - Concentrate on differential expression projects.
  - An overview of a typical RNA Sequencing pipeline.
  - Roundup of common QC issues.
  - Examples of tools used in the default pipelines in WHG
- Practical:
  - Work with toy datasets to recognise QC problems.

Gene Counts and Data Quality

- Timetable:
  - 13:30 This presentation
  - 14:30 Start practical
  - 15:00 Refreshments and continued practical
  - 16:00 Finish

Oxford Genomics CENTRE

Reads to Leads



wellcome centre human genetics

UNIVERSITY OF OXFORD

Reads to Leads

- Per-read filtering
  - Use metrics from the sequencing pipeline to exclude reads of low overall quality.
  - Usually an option during the mapping step.
- Mapping
  - Map to transcriptome for the organism.
  - HiSat2

Reads to Leads

- Deduplication
  - A tool identifies duplicate reads and excludes them from further analysis.
  - Picard MarkDuplicates
- Tabularisation
  - Match mapped reads to features (i.e. genes).
  - Produce count table.
  - featureCounts

Reads to Leads

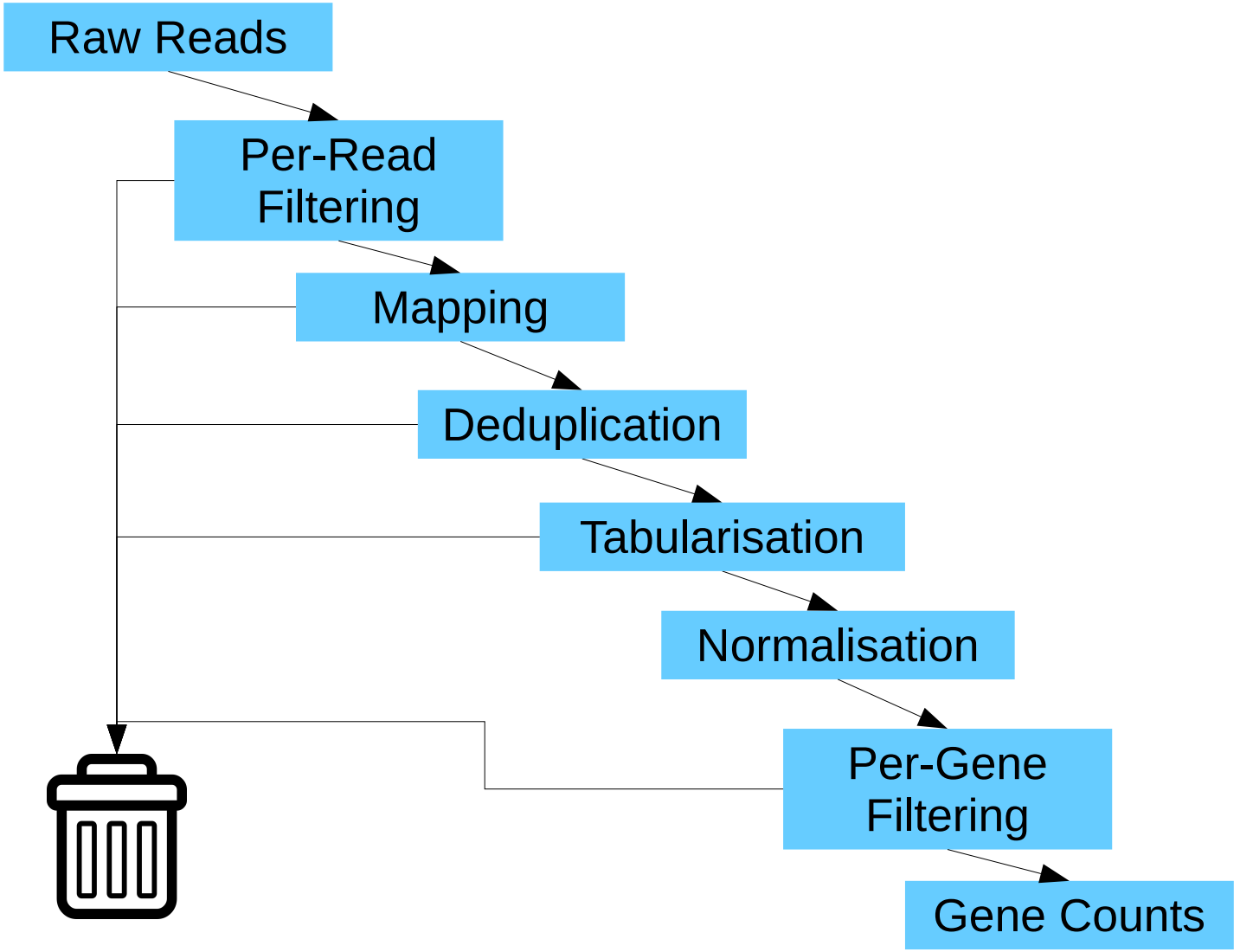
- Normalisation
  - Adjusts count table to take into account any variation in counts not due to gene expression.
  - Many methods, including:
    - Counts per million (cpm) – adjusts for library size.
    - Transcripts per million (tpm) – adjusts for library size and gene length.
    - Variance stabilisation (vsn) – adjusts so that the variance is not dependent on the mean.
  - DESeq2 (R package)

Reads to Leads

- Per-gene filtering
  - Where the overall level of expression is very low, reliable analysis cannot be performed.
  - Identify genes with low expression across all experimental groups and filter them out.
  - If a gene of interest is filtered out in this way, no work-around other than a repeat of sequencing at greater depth.

Oxford Genomics CENTRE

Reads to Leads

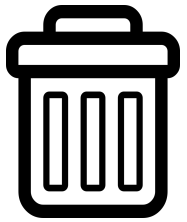


wellcome centre human genetics

UNIVERSITY OF OXFORD

What Reads are Removed?

- Poor quality reads.
- Unmapped reads.
  - Do not map at all.
- Duplicate reads.
- Reads that do not map to a unique gene.
  - Map to intron or intergenic region.
  - Do not map uniquely.
- Reads for marginally-expressed genes.



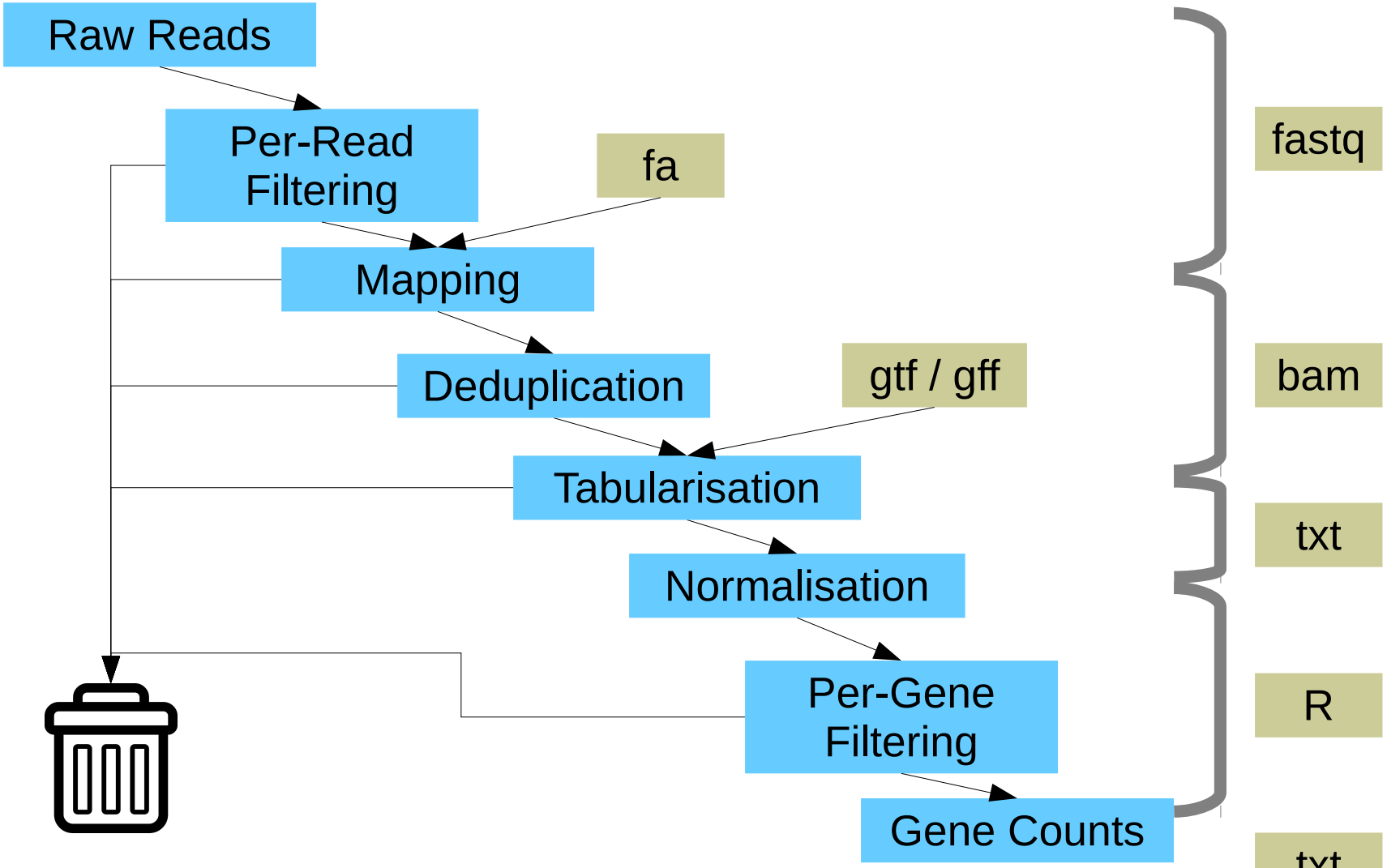
Assignment Summaries

- Tabularisation attempts to match reads to features.
  - Sometimes this cannot be done.
- In the output of featureCounts:
  - Assigned: Appears in gene counts table.
  - Ambiguity: The read maps to a unique location but spans multiple features.
  - MultiMapping: The read maps to more than one location.
  - NoFeatures: Reads that do not overlap a defined feature.
  - Unmapped: Were not mapped at all.

Assignment Summaries

- Some categories can be empty if those reads were discarded earlier in the process, or are if they only discarded if specific options are set:
  - MappingQuality
  - FragmentLength
  - Chimera
  - Secondary
  - Nonjunction
  - Duplicate

File Formats



File Formats - fa

- fasta or fa.
- Input to mapping tools.
- Used for reference genome for an organism.
- Stores sequence information, one section per chromosome.
- Text format.

File Formats - fastq

- Typical output from sequencing machines.
- Input to mapping tools.
- Stores read data and quality information.
- Reads have unique identifiers.
- Text format.
- Archival standard.

File Formats - bam

- Output from mapping tools.
- Many tools have bam files as input and output.
- Input to tools that prepare count tables.
- Stores:
  - All data from the source fastq regarding reads and quality.
  - Mapping position for each read plus mapping quality.
- Compressed format.
- Another archival standard.

Oxford Genomics CENTRE

File Formats – gtf / gff

- Input for tabularisation.
- Contains positions of the features of interest.
  - Gene-level
  - Exon-level
- By chromosome, start, end for each feature.
- Text format.

Oxford Genomics CENTRE

File Formats – Count Tables

- Output from tools that prepare tables.
- Stores summary of read counts per gene per sample.
- No read or quality information included.
- Often includes meta-data, such as totals of unmapped reads.
- Text format.

What Drives Differences?

- Quality issues can arise at every stage of the process:
  - Sample gathering
    - Batch effects.
    - Sample labelling issues.
  - Lab work
    - Contamination.
    - Low input material.
    - Batch effects.
  - Sequencing technicalities
    - Sequencing machine problems.
    - Computational anomalies.

Dealing with Quality Issues

- Many of these quality problems manifest as a sample or set of samples having a very different profile to the rest of the data.
  - Treated as outliers.
  - Outliers generally have to be discarded before analysis.
  - If the problem is limited to a particular middle step, the sample can be sequenced again.
- Some problems have a systematic effect on the samples and can be adjusted for in the analysis.

What Drives Differences?

- Technical aspects:
  - Tissue type.
  - Kit type.
  - Other experimental variables that should have been held constant for the entire project.

What Drives Differences?

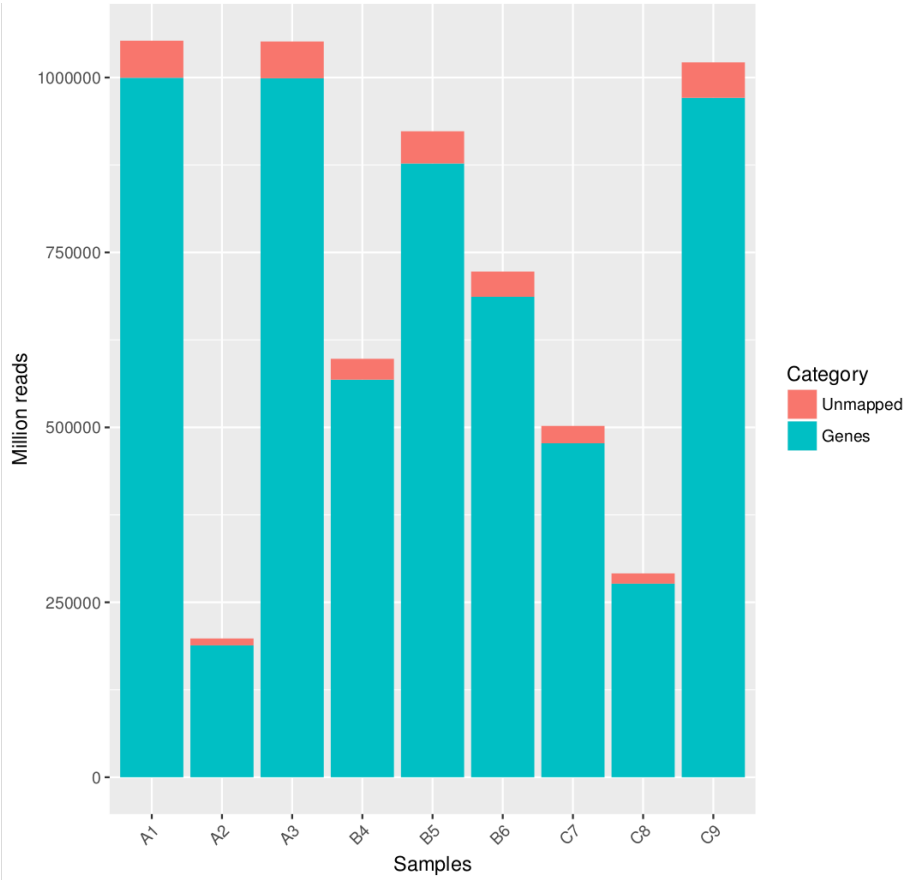
- Differential expression:
  - Treatment levels.
  - Disease condition.
  - Knockdown models.
  - Time factors.

Visualising QC

- Visual inspection can identify quality issues.
  - Outlier samples.
  - Potential batch effects.
  - Possible sample swaps or other naming problems.
- Often requires confirmation of the issue from outside the data before it can be adjusted for.
- Usually a problem can be seen in multiple visualisations.
- Being able to recognise common issues from visualisations saves time.

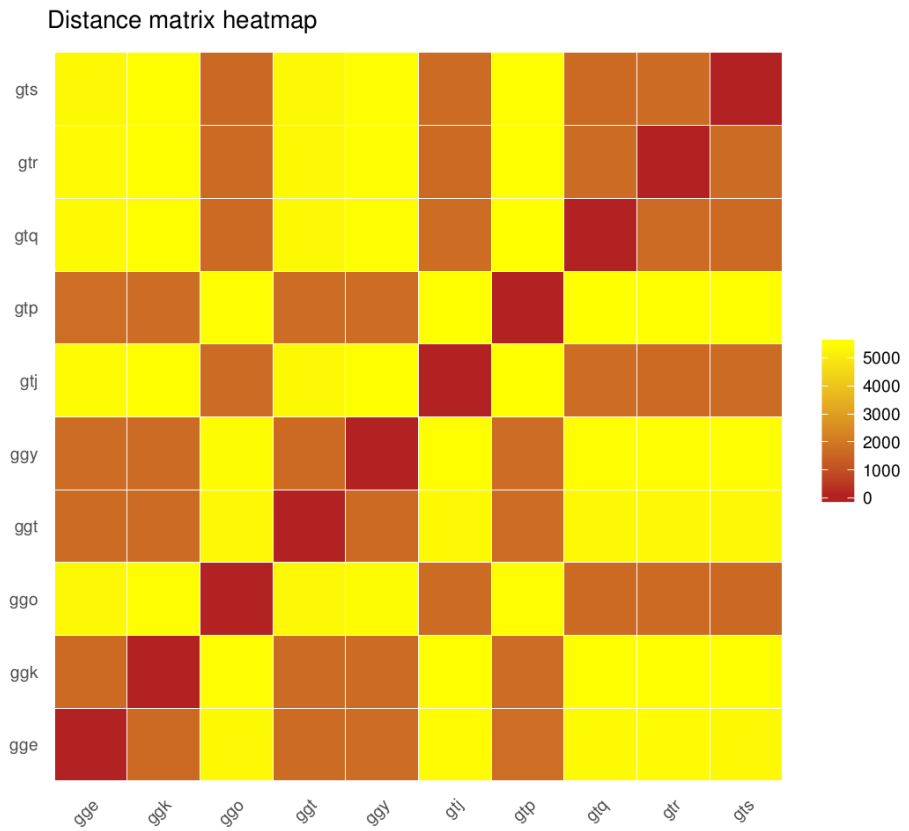
Visualising QC – Read Counts

- Shows total reads and proportion of reads assigned to different categories.
- Identifies outliers on the basis of total reads or read assignment.
- Quick way to spot failed samples.



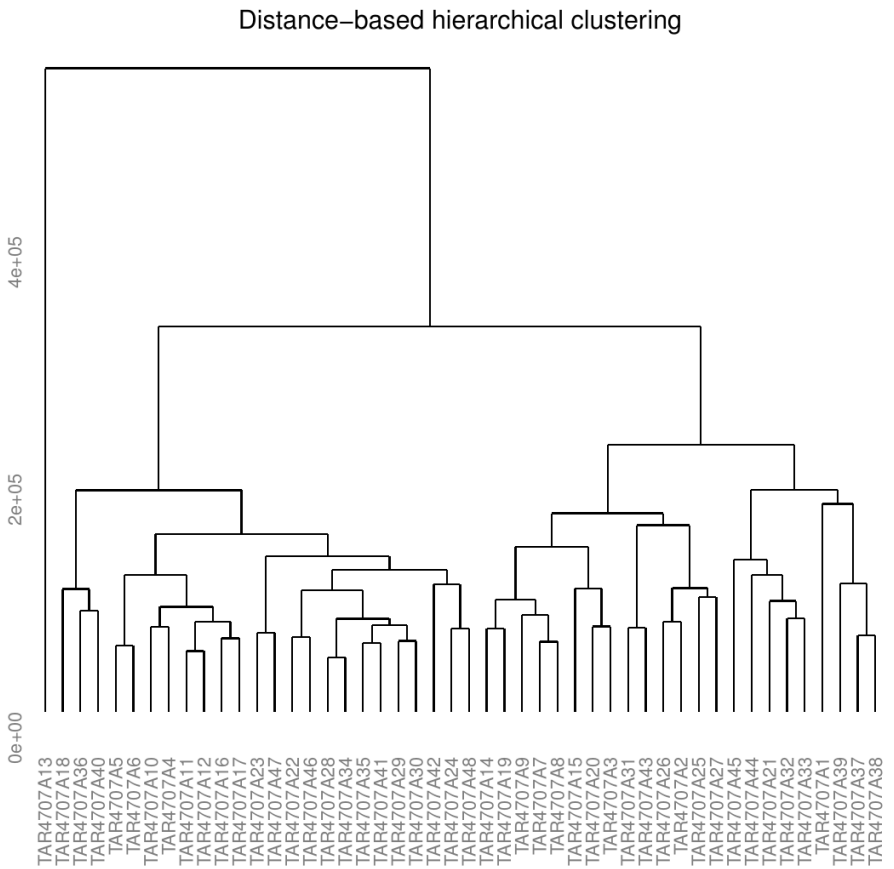
Visualising QC - Heatmaps

- Show similarity between samples.
- Many different ways of measuring that similarity.
- Can identify sample groups.
- Do not indicate underlying structure.



Visualising QC - Dendrograms

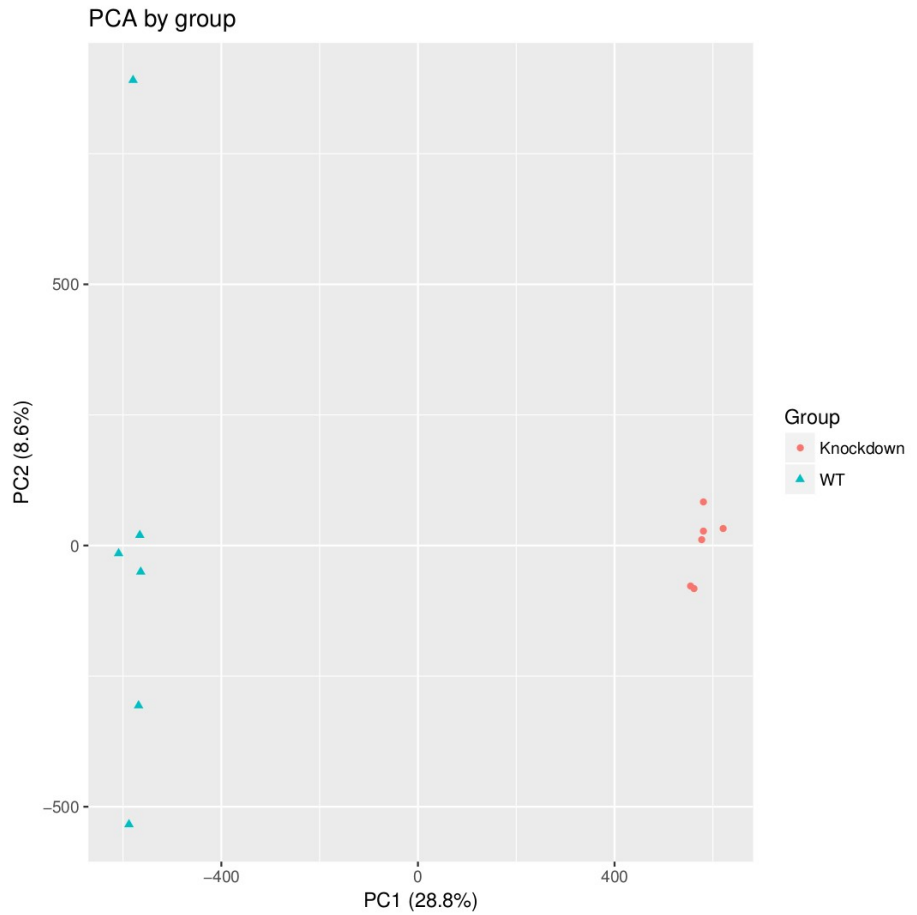
- Show similarity between samples.
- Many different ways of measuring that similarity.
- Can identify sample groups.
- Can infer hierarchical clustering from the tree.
- Often added to heatmaps.

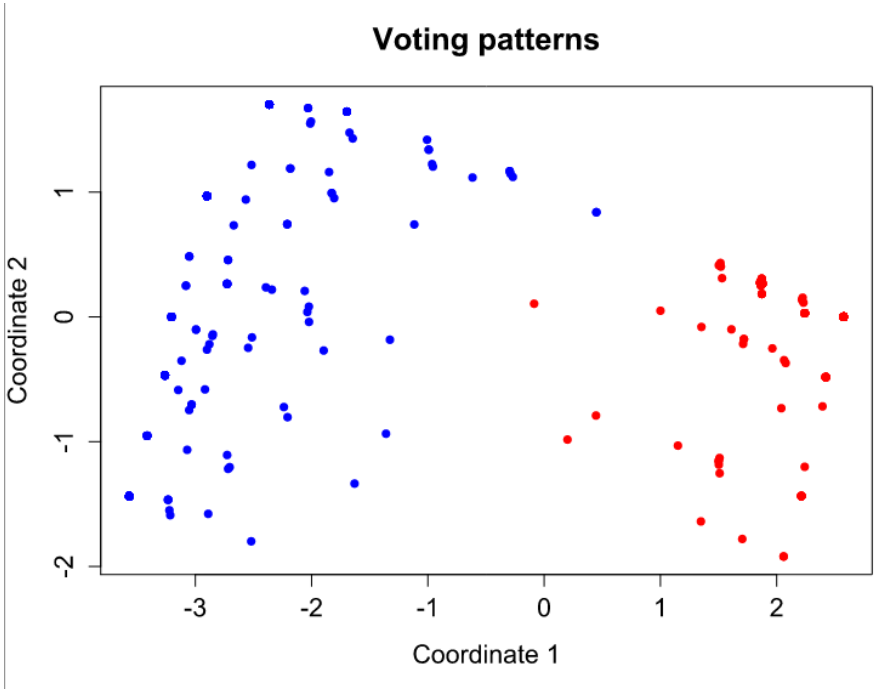


Visualising QC – PCA and MDS

- PCA
  - 'Principal Components Analysis'.
  - Multidimensional technique for revealing underlying clustering.
  - Identify multiple separate groups.
  - Provides indication of how much variability lies in each dimension.
- MDS
  - 'Multidimensional Scaling' plots.
  - Multidimensional technique for revealing underlying clustering.
  - Identify multiple separate groups.

Visualising QC – PCA and MDS





Visualising QC – PCA and MDS

- The most common plots show only the first 2 dimensions of these techniques.
- Higher dimensions can show more layers of information regarding the structure of the data.
- Ideally, each dimension will correspond 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.

Visualising QC - Limitations

- Good QC plots do not guarantee a successful project with useful analysis.
- If gene expression differs only slightly between experimental groups, the underlying pattern can be very difficult to spot visually if the visualisation is based on the full data.
- When there are multiple factors in the experiment, some may have a much larger effect on gene expression and make the differences of others harder to spot visually.

QC Issues

- QC problems of a given type affect the visualisations in a consistent way.
- Visual inspection is therefore a powerful tool for identifying what type of problem has been encountered.
- This does not replace formal statistical techniques for finding clusters or determining significant differences in expression between groups.

QC Issues – Failed Sample

- Characteristics:
  - Significantly lower total read count.
  - Proportion of read categories different from the rest of the data.
  - Isolated in PCA and MSD plots, sometimes to the point that the rest of the plot is unreadable.
- Solution:
  - Exclude that sample and re-examine QC for further issues.

QC Issues – Batch Effect

- Characteristics:
  - Experimental groups expected to be a single cluster are split into separate clusters.
  - These splits are in a consistent direction across different experimental groups.
- Solution:
  - Verify that a potential lab-based batch effect corresponds to the pattern in QC.
  - Introduce a variable for that batch effect in the analysis.

QC Issues – Sample Mix Ups

- Characteristics:
  - Clusters exist in the data, but do not correspond to the experimental groups.
  - The number of samples in each cluster make sense for the experiment.
- Solution:
  - Double check sample naming. If an error is found, correct the names and continue checking.
  - Never try to use the data to infer the correct naming.

QC Issues – Failed Project

- Characteristics:
  - No visible clustering.
  - Or variation very small.
- Solution:
  - Conduct analysis and see if there is truly nothing to be found.
  - Work with the lab to discover what went wrong.
  - Repeat the experiment.

Tools

- HiSat2
  - https://ccb.jhu.edu/software/hisat2/index.shtml
- Picard
  - http://broadinstitute.github.io/picard/
- FeatureCounts
  - http://bioinf.wehi.edu.au/featureCounts/