

# Using transcriptomics to identify routes to survival under stress in bacteria

## Day 1 – Data Preprocessing

Dr Gregory Wickham

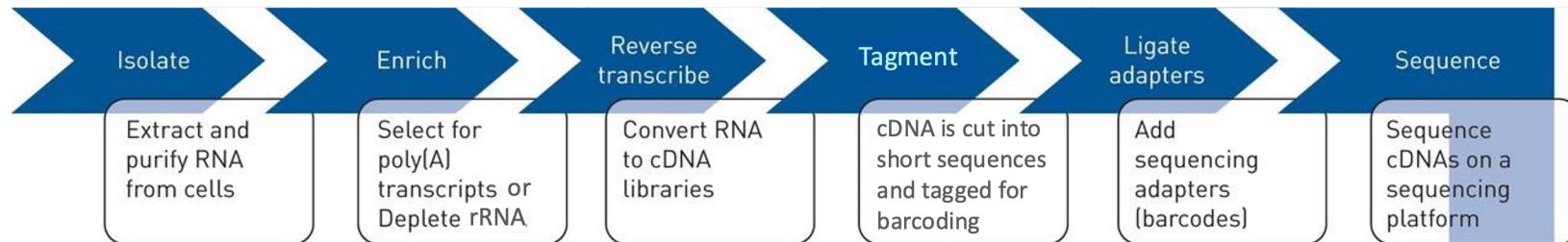
Whitchurch Group  
Quadram Institute

13<sup>th</sup> December 2023



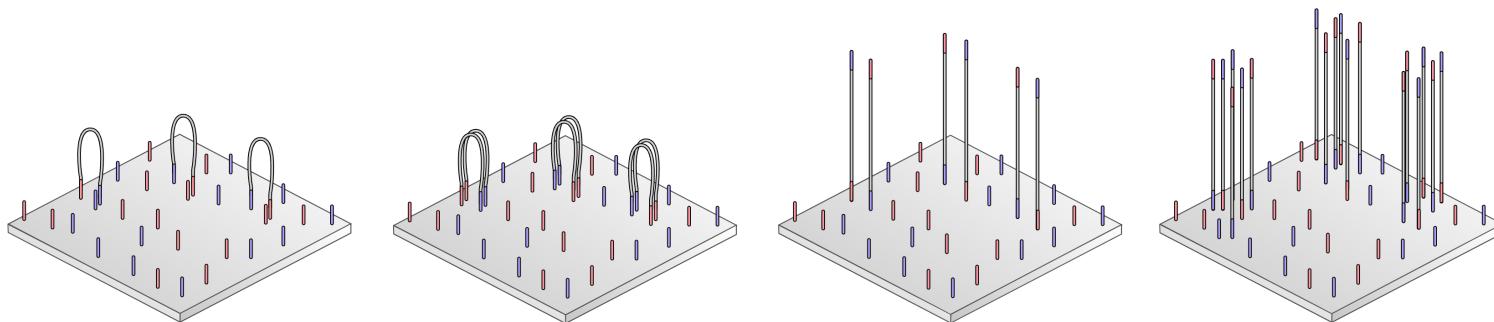
# Transcriptomics

- Transcriptomics is set of techniques which use NGS of mRNA transcripts to study global gene expression
- By quantifying which genes are being upregulated or downregulated relative to a control, the molecular basis underpinning physiological responses to specific conditions can be interrogated
- RNA-seq relies on purification of total RNA from cells, depletion of ribosomal RNA (or enrichment of mRNA) and conversion of mRNA to complementary DNA with reverse transcriptase
- When cDNA is obtained, sequencing libraries can be generated in the same manner as genomic DNA
- Sequencing by synthesis (Illumina sequencing) is the most common platform for transcriptomics
- Unlike genomic sequencing which expresses sequencing volume as genome coverage or read depth, transcriptomic sequencing just uses total read count (5 – 20 million reads is the standard read count for bacterial RNA-seq)



# Illumina Sequencing

- DNA is fragmented into ~350 bp sequences and tagged with primers which allow sequencing adapters to be ligated and fragments are amplified by PCR to generate a ‘sequencing library’
- The libraries are normalized to a standard concentration and pooled together, with the unique combination of index adapters allowing for the pool to be demultiplexed back into their respective libraries after sequencing
- The pooled libraries are transferred to a flow cell where they anneal to random oligonucleotides complementary to adapters at either end forming bridges
- Each fragment is amplified through successive cycles of PCR known as solid-phase bridge amplification forming clusters of clonal fragments which are subsequently linearized and the reverse strands are cleaved away
- The forward strands are then sequenced by detecting the emission of light from each cluster as complementary fluorescently-tagged nucleotides compete to anneal



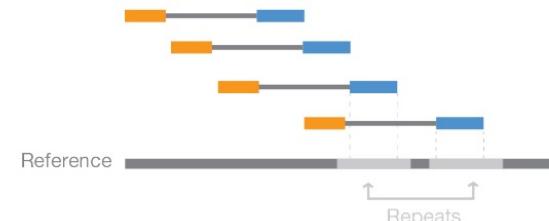
# Paired-end Sequencing

- Illumina sequencing is now generally performed in paired-end format to generate 2 reads per sample known as R1 and R2
- After the first sequencing pass, oligos on the flow cell are unblocked and the sequencing product denatured from the forward strand, bridges form again and a reverse strand is transcribed
- The forward strand is cleaved and the sequencing reaction is then performed again to generate a second read
- Reads typically have a length of typically 50 - 150 bp however fragment inserts are usually longer than the sum of the two read lengths, leaving an unsequenced part in the middle of the fragment.
- When the distance between read pairmates is known (i.e. average fragment size), paired-end sequencing can help overcome a major limitation in alignment algorithms associated with repeat sequences.

Paired-End Reads



Alignment to the Reference Sequence



# File Types

## FASTA format

- Contains sequence identifier prefixed by ‘>’
  - Contains consensus sequence
  - The format for contigs, assemblies and annotated genomes
- 
- ```
>Mus_musculus_tRNA-Ala-AGC-1-1 (chr13.trna34-AlaAGC)
GGGGGTGTAGCTCAGTGGTAGAGCGCGTGCTTAGCATGCACGAGGcCTGGGTTGATCC
CCAGCACCTCCA
```

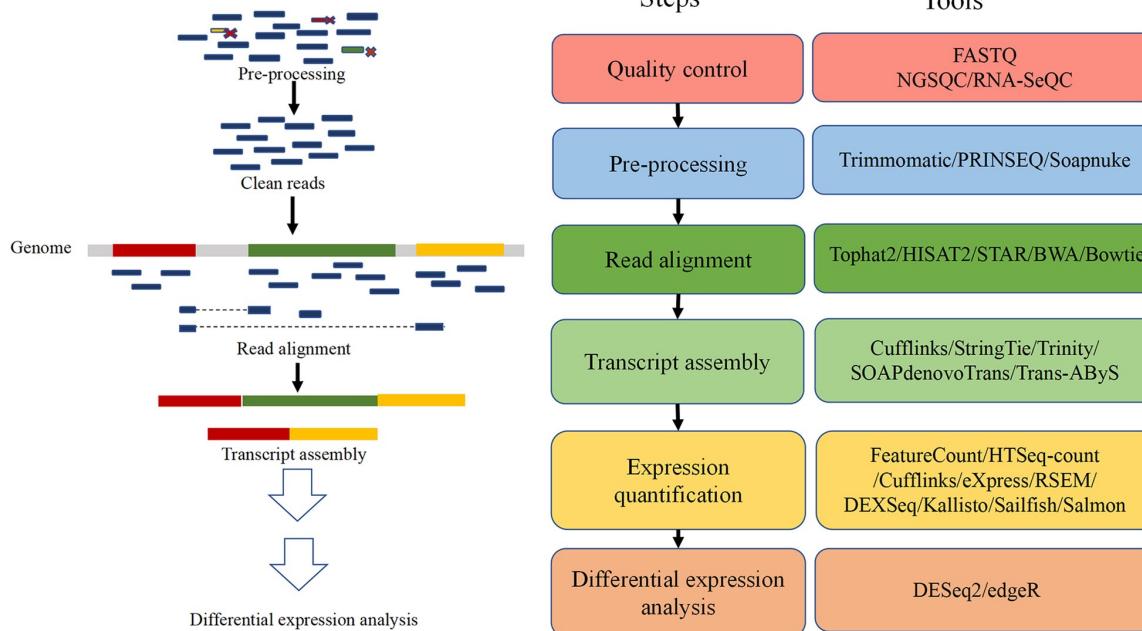
## FASTQ format

- Contains sequence identifier prefixed by ‘@’
  - Contains read sequence
  - Contains phred quality score corresponding to each base
  - Format for raw reads
  - Generally compressed in .gz files
- 
- ```
@071112_SLXA-EAS1_s_7:5:1:817:345
GGGTGATGGCCGCTGCCGATGGCGTC
AAATCCCACC
+
IIIIIIIIIIIIIIIIIIIIIIIIII
IIII9IG9IC
```

## Phred scores (Q)

- Measures probability of errors in basecalling
- Phred scores range from Q2 – 40, with Q30 being the usual quality cutoff representing a 1:1000 chance of error
- The phred score is an important metric for quality control and trimming reads
- Two main schemes for phred scoring is phred+33 and phred+64 which encode different ASCII characters for each value

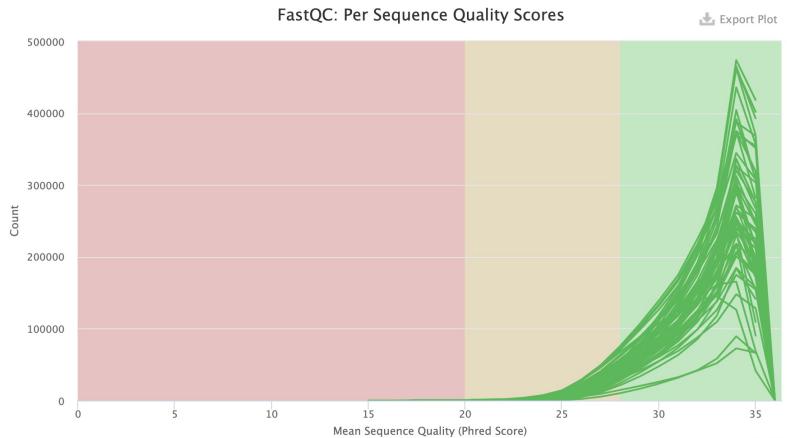
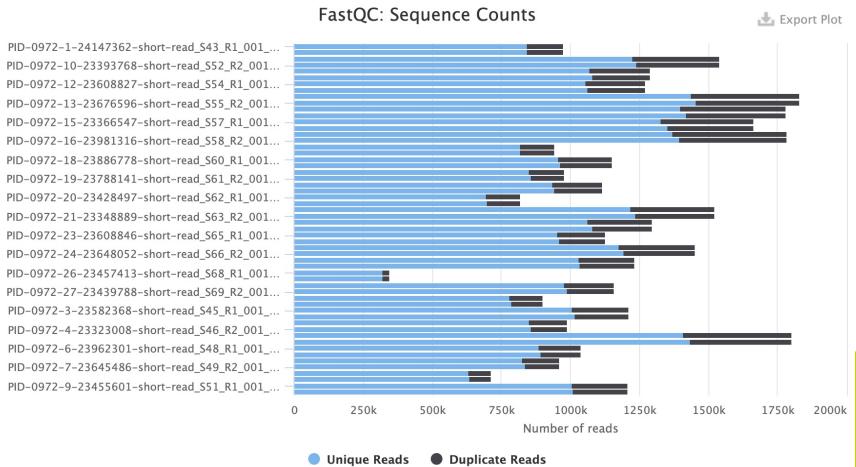
# RNA-seq Analysis Pipeline



- First-generation transcript quantifiers take alignment files as input (Cufflinks, featureCounts)
- Second-generation quantifiers are alignment free so just take reads (Kallisto, Salmon, Sailfish)
- Code-free ecosystems are now also available for performing bulk RNA-seq (iDEP, RNAlysis)

# Quality Control

- The gold standard for quality checking short read sequencing data is FastQC (LongQC and Nanoplot are good alternatives for long-read data)
- FastQC will construct summary statistics from FASTQ files for base position quality, quality per read, GC content, overrepresented sequences and sequencing adaptor contamination
- Multiple FastQC reports can be aggregated and visualised with MultiQC
- Confirming sequencing read quality should always be the first thing done when you receive sequencing data



# Read Trimming

- Low quality reads should be removed from sequencing data using filtering software such as Trimmomatic
- Trimmomatic uses several user-defined parameters to determine which sequences to trim/drop based on length and phred score

Parameter	Function
ILLUMINACLIP	Cut adapter and other illumina-specific sequences from the read
LEADING/TRAILING	Cut bases off the start/end of a read, if below a threshold quality
SLIDINGWINDOW	Scans from the 5' end and clips the read once the average quality within the window falls below a threshold
MINLEN	Drop the read if it is below a specified length
AVGQUAL	Drop the read if the average quality is below the threshold

- The order in which these parameters are invoked affects which reads are trimmed, the recommended order is shown above
- In paired-end mode, Trimmomatic requires both pairmates and will output trimmed reads in both paired and unpaired formats
- For trimmed reads in paired format will, if a read is dropped, then the pairmate will be dropped irrespective of its own quality
- It is good practice to run quality control again after read trimming, however when the raw read quality is high and read filtering is low (<1% trimmed reads) it's probably overkill

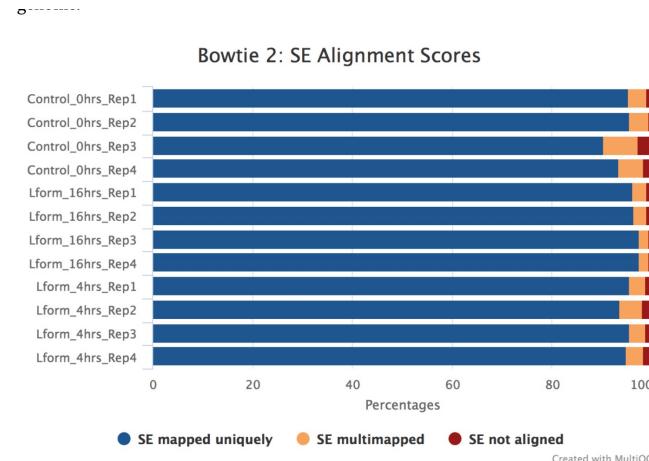
# Read alignment

- The percentage of mapped reads is a global indicator of the overall experimental quality
- When performing alignment and assembly, it is imperative to use the paired reads as these tools rely on read concordance to operate correctly (i.e. read #705,531 in R1 should be the mate of read #705,531 in R2)
- Bowtie2 and HISAT2 are two of the most popular bwa aligners for microbial RNA-seq

```
Coor 12345678901234 5678901234567890123456789012345  
ref AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCAT
```

```
+r001/1      TTAGATAAAGGATA*CTG  
+r002      aaaAGATAA*GGATA  
+r003      gcctaAGCTAA  
+r004      ATAGCT.....TCAGC  
-r003      ttagctTAGCC  
-r001/2      CAGCGGCAT
```

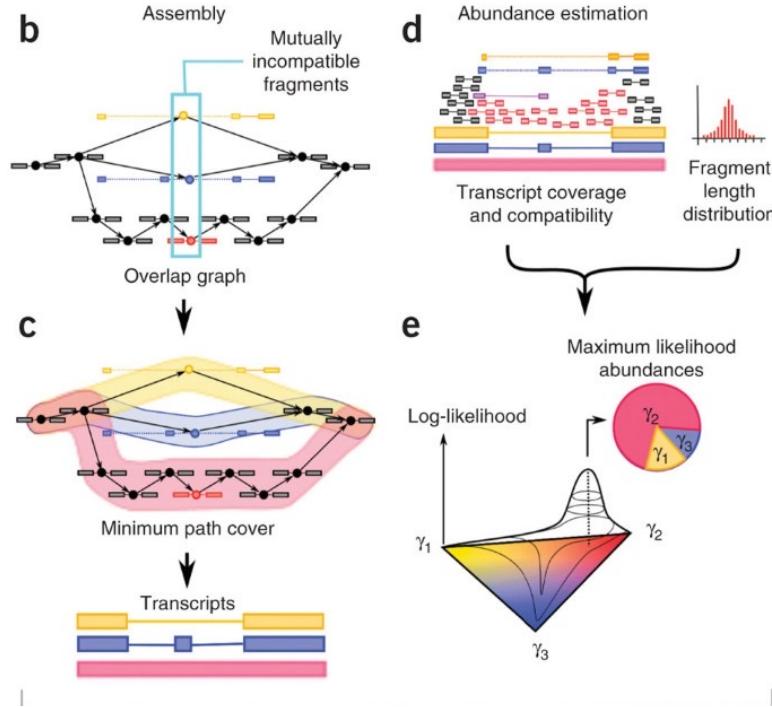
```
@HD VN:1.6 SO:coordinate  
@SQ SN:ref LN:45  
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *  
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAAGGATA *  
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;  
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *  
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;  
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



- Alignment tools typically output data in SAM/BAM format
- Sequence alignment map (SAM) files store alignment information of short reads mapped against reference sequences.
- BAM files are binary-translated SAM files which offers smaller file sizes reducing computation times

# Transcript Quantification

- Assembly of aligned reads into mRNA transcripts is largely deprecated for bacterial RNA-seq, modern counting tools perform this step automatically during transcript quantification
- Manual transcript assembly against a transcriptome reference may still be necessary in eukaryotic RNA-seq due to gene splicing
- Baseline gene expression is generally measured as RPKM (reads per kilobase per million reads) or FPKM (fragments per kilobase per million reads)
- These metrics often undergo further normalization in order to correct for transcript length, GC content and read size distributions



# Using transcriptomics to identify routes to survival under stress in bacteria

## Day 2 – Differential Expression Analysis

Dr Gregory Wickham

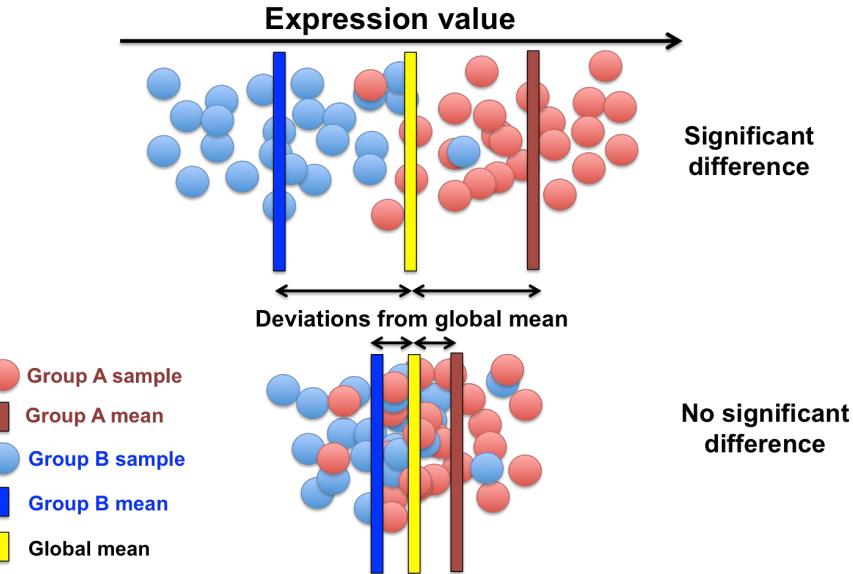
Whitchurch Group  
Quadram Institute

14<sup>th</sup> December 2023



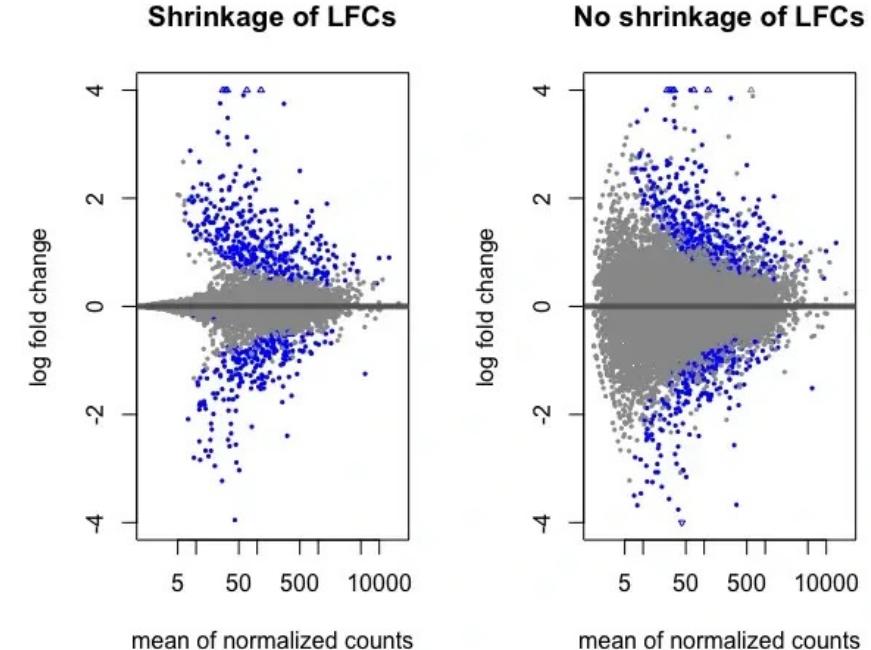
# Differential Expression Analysis

- Differential expression analysis takes count data and fits models to discover changes in expression between experimental groups.
  - Differential expression analysis is specifically designed for comparing expression **between** samples
  - For comparisons of genes **within** a sample, RPKM/FPKM are more reliable
- edgeR and DEseq2 are the two most common DE tools. DEseq2 is more conservative and its estimations are considered to have lower rate of type I errors although can miss smaller changes in expression.
- Differential expression tools allow for the modelling of complex experimental designs, such as multiple groups with multiple replicates.
- Differential expression is measured in log<sub>2</sub> fold change between the test and control groups, and an adjusted p-value is calculated based on the Wald test
- A differentially expressed gene (DEG) is called when the log<sub>2</sub>FC is > 2 and the adj p-value (also referred to as false discovery rate (FDR)) is < 0.05



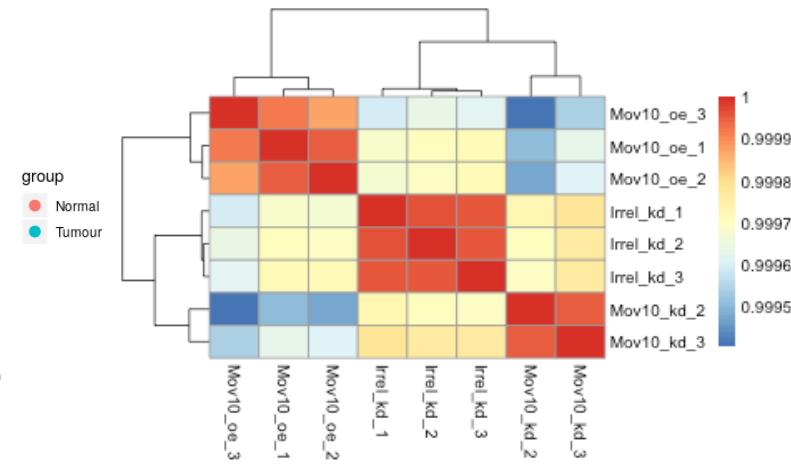
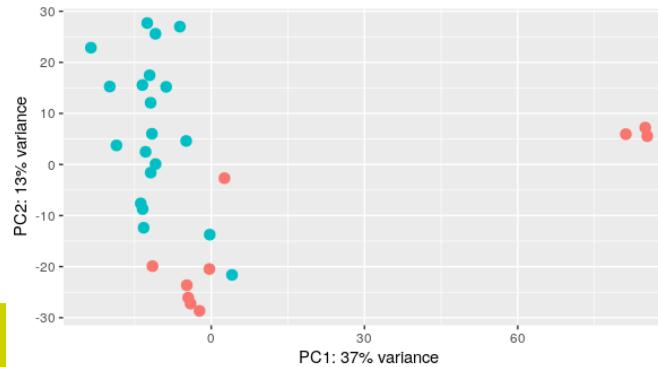
# Modelling Differential Expression

- DESeq2 generates a negative binomial model from the median of ratios of normalized counts against a ‘size factor’ – a statistic calculated to correct for library size, transcript length and composition bias.
- Shrinkage estimator is used stabilize parameter estimates by “shrinking” each point closer to the mean (zero or very close to), with larger variances shrinking more.
- Typically, the number of replicates for an RNA-seq experiment is small (2 – 4) and the number of dimensions within the data (i.e. genes) is large, therefore it is difficult to reliably estimate variance.
- As genes of similar expression can be assumed to possess similar variance, clustering is performed to increase the effective sample size which increases statistical power of significance calculations.



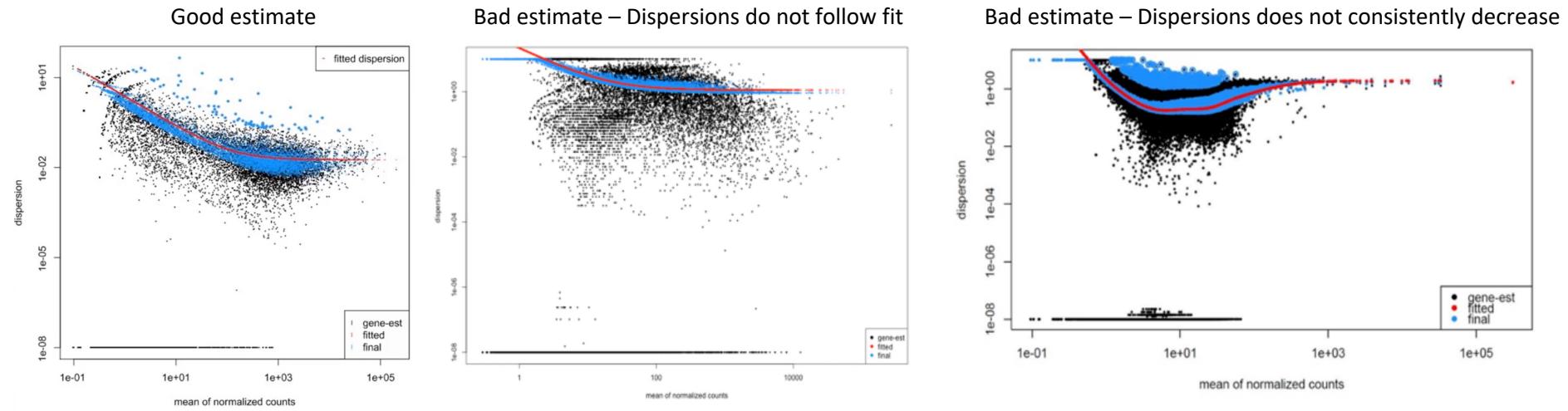
# Validating Differential Expression Analysis

- Principal component analysis (PCA) is a dimensionality reduction technique used to determine inter-group variation
  - PCA calculates the maximum variance across each dimension
  - A principal component is generated from each dimension
  - The top 2 or 3 principal components (which reflect the greatest amount of variance able to be visualized on a 2 or 3 dimensional axis) are used to address how well groups cluster
  - The greater variance accounted for in a PCA plot, the more robust the conclusions on inter-group variation are
  - Homogenous variance across principal components is often due to non-linear relationships, in which case non-linear dimensionality reduction methods should be used such as t-SNE or UMAP.
  - Hierarchical clustering can also be used to determine this

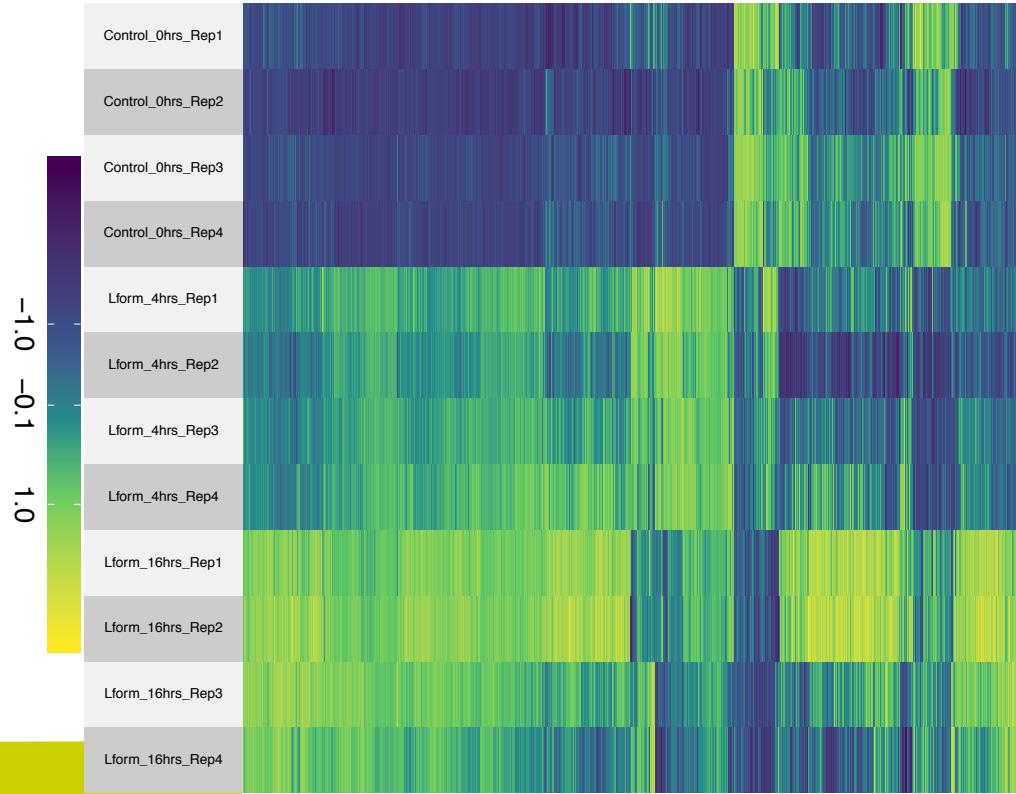


# Validating Differential Expression Analysis

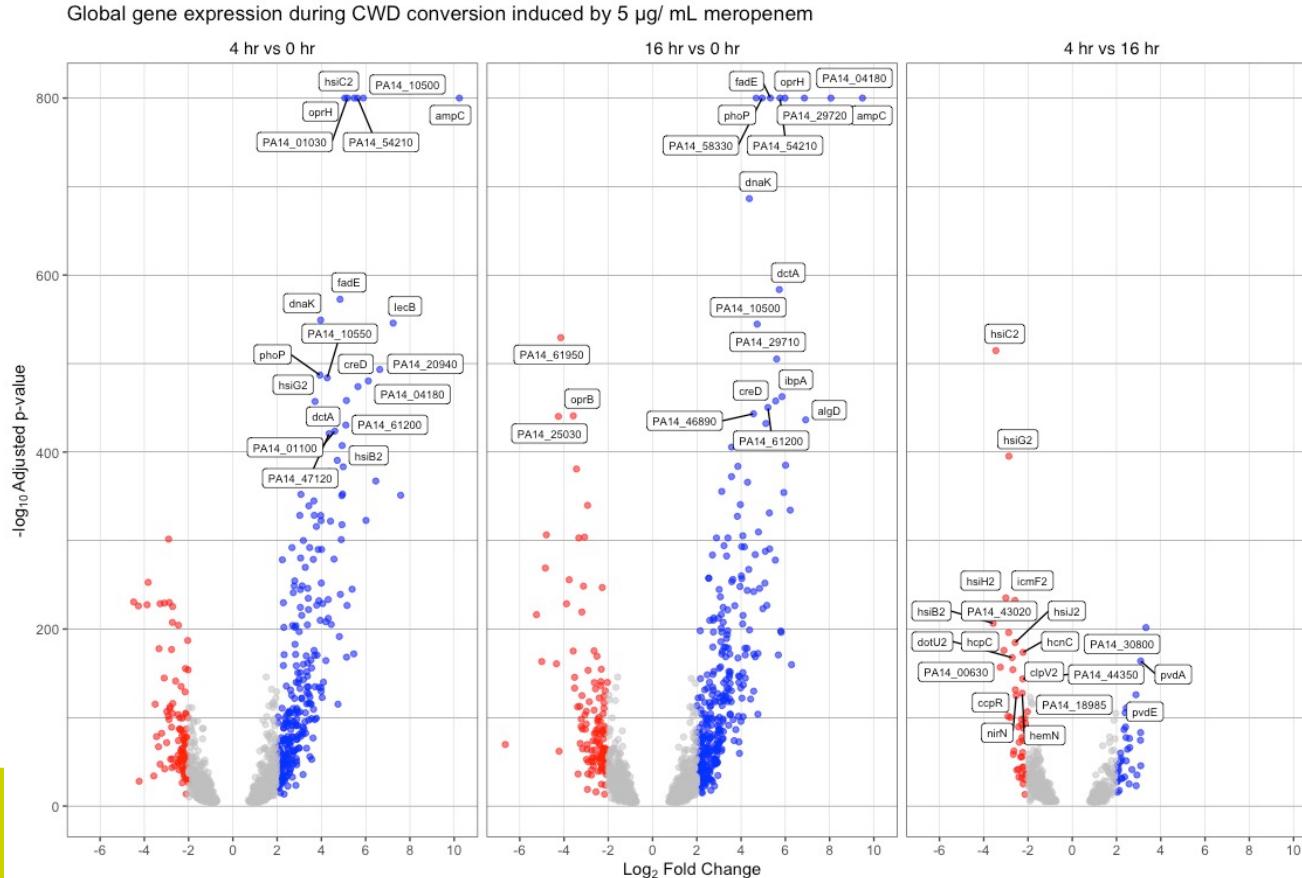
- Dispersion plots are used to estimate within-group variation, which shows the data pre- and post-shrinkage
  - Good models should cluster dispersion estimations around the fitted maximum likelihood (ML)
  - Dispersion should decrease with increasing mean count



# The heatmap is the highest level for visualization of global gene expression changes



# The volcano plot allows for the most significantly differentially-expressed genes to be visualised



# Shared functions can also be classified in differentially expressed genes through enrichment analysis (KEGG, GO, COG)

