gene-expression

Diego Cesar Villa Almeyda

2025-06-21

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

Cryptococcus neoformans is a globally distributed invasive fungal species within the genus *Cryptococcus*, causing hundreds of thousands of deaths per year, primarily causes disease in individuals with compromised immune systems (1). Treatment remains a major challenge due to a limited arsenal of decades-old therapeutic agents and a growing problem of drug resistance to all three classes of antifungals currently in use (polyenes, azoles, pyrimidine analogues) (1). The high global incidence, mortality, dearth of drugs, toxicity, and development of resistance highlight an urgent need for new drugs and therapeutic strategies (1).

The FLC1 protein is identified as a crucial factor for C. neoformans stress responses and virulence (2). Importantly, FLC1 is considered a promising potential drug target because its homologues exist in other fungal pathogens but are poorly conserved in humans. This suggests that targeting FLC1 could lead to wider spectrum antifungal therapy with minimal toxicity to human hosts, potentially making it an "Achille's heel" for C. neoformans and other fungal infections (2).

Given the importance of this funal pathogen and the FLC1 protein, in this report we analysed the data from an experiment done by Rachel Murray at the Wallace Lab at the School of Biological Sciences, University of Edinburhg, where they wanted to understand how C. neoformans responds to cell wall stress and to factors taht modulate cell wall stress: a chemical called calcofluor white (CFW) that binds to fungal cell walls, a gene called FLC1 that is believed to be involved in calcium import to cells, and another chemical called EGTA that blocks calcium import. We observed that cells with FLC1 mutated have very weird looking cell walls, and die at 37°C, however, adding EGTA to mutated cells prevents their death.

Cryptococcus neoformans is a globally distributed opportunistic fungal pathogen responsible for hundreds of thousands of deaths annually, primarily affecting individuals with compromised immune systems (1). Treatment remains a significant challenge due to the limited availability of effective antifungal drugs, which are decades old and increasingly compromised by resistance. All three major classes of antifungal agents—polyenes, azoles, and pyrimidine analogues—face issues related to toxicity and reduced efficacy. The combination of high global disease burden, limited treatment options, and rising drug resistance underscores the urgent need for new therapeutic strategies (1).

Recent research has identified the FLC1 protein as a key factor in *C. neoformans* stress responses and virulence (2). Notably, FLC1 has homologues in other fungal pathogens but is poorly conserved in humans, making it a promising antifungal drug target. Targeting FLC1 could lead to broad-spectrum antifungal treatments with reduced risk of toxicity to human

hosts—potentially representing a therapeutic "Achilles' heel" for *C. neoformans* and related pathogens (2).

Given the clinical importance of *C. neoformans* and the potential of FLC1 as a drug target, this report presents a statistical analysis of transcriptomic data from an experiment conducted by Rachel Murray at the Wallace Lab, School of Biological Sciences, University of Edinburgh. The experiment aimed to investigate the fungal response to cell wall stress and modulators thereof, including calcofluor white (CFW), a chemical that binds to fungal cell walls; EGTA, a calcium chelator; and deletion of the FLC1 gene, which is believed to be involved in calcium import. Preliminary biological observations indicated that cells lacking FLC1 exhibit abnormal cell wall morphology and experience lethality at 37°C, a phenotype that can be suppressed by the addition of EGTA. The analyses reported here aim to characterise the gene expression patterns underlying these effects, providing statistical insights for further biological interpretation.

RNA sequencing (RNA-seq) is a high-throughput technology that enables comprehensive transcriptome profiling using deep-sequencing methods (3). Compared to earlier approaches, RNA-Seq provides more accurate and quantitative measurements of gene expression levels and transcript isoforms, allowing for genome-wide analysis at single-base resolution (3). In the experiment described above, RNA-Seq was employed to measure the abundance of RNA transcripts (expression level) across thousands of genes following a 3-hour incubation under various combinations of environmental conditions (growth media and temperature) and genetic background (presence or deletion of the FLC1 gene).

Our analysis focused on the RNA-Seq data with three primary objectives: (i) to assess the quality and consistency of replicates across experimental conditions and explore broad expression patterns among samples, (ii) to identify genes that exhibit differential expression in response to changes in environmental or genetic conditions, and (iii) to uncover and characterise distinct expression patterns within the subset of differentially expressed genes. Together, these analyses aim to generate a robust statistical evidence that can guide downstream biological interpretation, helping researchers to prioritise candidate genes for further functional investigation into the mechanisms underlying cell wall stress in *C. neoformans*.

The remainder of this report is organised as follows: Chapter 2 provides a concise yet comprehensive overview of the methods and techniques used for RNA-seq data processing and differential expression analysis. Chapter 3 presents the key findings in relation to the analytical aims. Finally, Chapter 4 summarises the main results and discusses potential limitations and opportunities for further refinement. Additional supporting information is provided in the Appendix.

2 Methods

2.1 Experimental design and data

The study involved culturing cells from both the wild-type (WT, FLC1 gene present) and FLC1Δ (FL, FLC1 gene absent) strains under various environmental conditions. Cells were grown in four different growth medium: YPD (Y), YPD with CFW (YC), YPD with EGTA (YE), and YPD with both CFW and EGTA (YCE), all at 37°C to induce stress. Additionally, both strains were grown under baseline conditions in standard YPD at 30°C. In total, 10 distinct experimental conditions were tested (5 environmental conditions × 2 strains), each with 3 biological replicates, resulting in 30 samples.

Table 2.1 presents the experimental design along with the sample labels for each condition. The 10 experimental conditions are labeled by combining the levels of the three experimental factors: strain, temperature, and growth media, in that order. For example, the condition involving the wild-type strain grown in YCE medium at 37°C is labeled WT-37-YCE, while the corresponding condition for the FLC1Δ strain is labeled FL-37-YCE.

The primary output of the experiment was a raw count matrix containing RNA abundance measurements for 6795 genes across 30 samples, resulting in a 6795 × 30 matrix. Each entry in the matrix represents the number of sequencing reads mapped to a specific gene in a given sample.

Table 2.1: Overview of the experimental design showing the combinations of temperature and growth media (environmental conditions) used to grow cells of the WT and FLC1Δ (FL) strains. Each cell lists the replicate sample labels (S1–S30) corresponding to each unique experimental condition.

	Enviromental conditions				
	30°C	**37°C**			
Strain	**Y**	**Y**	**YC**	**YE**	**YCE**
WT	S1, S2, S3	S4, S5, S6	S7, S8, S9	S10, S11, S12	S13, S14, S15
FL	S16, S17, S18	S19, S20, S21	S22, S23, S24	S25, S26, S27	S28, S29, S30

2.2 Pre-filtering

Some genes may exhibit low or zero counts across all samples, suggesting they are not expressed under the experimental conditions. These genes are unlikely to be identified as differentially expressed and are typically filtered out prior to analysis. The biological justification for pre-filtering low count or uninformative genes is that a gene typically needs to be expressed above a minimal threshold to be translated into a functional protein or to exert a meaningful biological effect (4). Moreover, low-expression genes often reflect sampling noise rather than true biological signal (5).

DE analysis involves performing a hypothesis test for each gene to evaluate whether expression levels differ between experimental conditions (see Section 2.5). This results in a multiple testing problem, which is typically addressed by adjusting p-values to control the false discovery rate (FDR). However, such corrections reduce statistical power by raising the threshold for significance. As the number of tests increases, this loss of power aggravates, further limiting the ability to detect truly DE genes, especially when they represent a small proportion of the total set (6). To mitigate this, it is recommended to filter out low-expression genes prior to DE analysis, thereby reducing the total number of hypotheses. This makes the multiple testing correction less stringent and increases the chance of correctly identifying DE genes (5,6).

Although the software used for differential expression (DE) analysis includes an internal filtering routine (see Section 2.5), additional pre-filtering was applied using an empirical method implemented in the R package `edgeR` (7), as described in (4). This method retains genes whose counts-per-million (CPM) exceed a specified threshold k in at least n samples. The CPM for each gene in a sample is calculated by dividing the raw read count by the total number of reads (i.e., the library size) in that sample and scaling by one million.

To determine the CPM threshold k, the user specifies a minimum raw count that a gene must meet in n samples, and the software computes the corresponding CPM based on the smallest library size. For this study, we required a minimum count of 10 in at least 3 samples, corresponding to the number of replicates per condition. The original, unfiltered count matrix was retained to replicate the DE analysis and compare results.

2.3 Normalisation and transformation

In RNA-seq experiments, the number of reads mapped to a gene depends not only on its expression level but also on between-sample and within-sample factors that make the row counts not directly comparable across genes or samples (8). Between-sample variation primarily arises from differences in library size, with larger library sizes producing more reads across all genes in a sample, whereas within-sample variation occurs at a gene level and can be influenced by gene length or by the proportion of guanine (G) and cytosine (C) nucleotides in the genes (GC-content, 9).

A common approach to correct for between-sample variation is total count normalisation, in which raw counts are divided by the total number of reads (library size) for each sample and scaled by a constant factor. CPM is a typical example of this method. However, total count normalisation can be problematic in DE analysis, as a small number of highly expressed genes can disproportionately inflate the library size. This can artificially lower the normalised expression of other genes within the same sample and exaggerate differences between samples, even when no real biological variation exists (8). More robust normalisation methods have been developed for DE analysis, such as the trimmed mean of M-values (TMM) method (10). In this study, we employed the normalisation procedure implemented in the R package `DESeq2` (11), which uses the median-of-ratios method to estimate sample-specific size factors for each gene count (12). For comparison purposes, we also applied CPM normalisation. No normalisation to correct for within-sample variation (e.g., gene length or GC content) was applied, following the recommendation of the project advisor.

An additional challenge in RNA-seq analysis arises during data visualization and multivariate techniques such as clustering or principal component analysis (PCA), which are sensitive to the scale of the variables being analyzed. Even after normalisation, RNA-seq data often exhibit heteroskedasticity, where genes with higher expression levels tend to show greater variance across samples than low-expressed genes. As a result, these highly variable genes can disproportionately influence the analysis, potentially obscuring meaningful biological patterns (11). A common solution is to apply a variance-stabilizing transformation (VST) to the raw or normalised counts, which aims to place lowly and highly expressed genes on a common scale (13).

One of the most common VST is the logarithm (log) transformation. However, this transformation has the issue of exaggerating variability in low-count genes, where random noise tends to overshadow true biological signal (11). In this study, we applied two transformations available in the `DESeq2` package: the regularized logarithm (rlog) transformation (11) and the VST proposed in (12). The rlog transformation is closely related to the modeling framework used for differential expression analysis in `DESeq2` and will be described in more detail in Section 2.5. The latter transformation, which we will refer to simply as VST, performs a monotonic transformation of the normalised counts so that the resulting variance is approximately independent of the mean. Both transformations were applied after normalising the counts using the `DESeq2` method. To benchmark these transformations, we also applied a log transformation with base 2 to the CPM-normalised counts having previously added a prior count of 2 to the raw counts to avoid taking logarithm of zero (undefined).

To evaluate the transformations, we plotted gene-wise means versus standard deviations for the raw counts and each transformation method. Additionally, boxplots of sample expression values were used to compare the distribution of raw and transformed counts, assessing how the transformations affected variance stabilization and overall expression profiles.

2.4 Exploratory analysis

An important first step in RNA-seq data analysis is assessing the quality and consistency of biological replicates. Ideally, replicates within the same experimental condition should exhibit similar expression profiles after normalisation and transformation. To evaluate this, we computed both the correlation matrix and the pairwise distance matrix of the samples, using the Pearson correlation and the Euclidean distance, respectively. If the replicates are consistent, the correlation and distance matrices should reveal strong similarity among samples from the same condition and clear separation from those in different conditions. Prior to this analysis, the raw counts were normalised and transformed using the methods described in Section 2.3, resulting in three versions of the data: log-transformed CPM, rlog-transformed counts, and VST-transformed counts.

A more visual approach to assess replicate quality is to project the samples into a lower-dimensional space and evaluate whether replicates from the same condition cluster together. We used PCA for this purpose. PCA decomposes the correlation matrix of the genes into uncorrelated components ordered by the variance they explain, which is quantified by their corresponding eigenvalues. To determine how many dimensions to retain, we applied Horn's parallel analysis: this method compares the observed eigenvalues to those obtained from randomly generated uncorrelated data sets and retains components whose eigenvalues exceed the average from the simulations (14).

We used biplots to visualise the samples alongside the five genes most strongly correlated with each component, for all pairwise combinations of the retained principal components. These correlations are quantified by the loadings. In addition, we generated loading plots displaying the top 1% of genes with the highest loading magnitudes for each retained component. To further investigate which experimental conditions or factors might be driving the separation of samples in PCA space, we produced boxplots of the sample scores (i.e., component coordinates) grouped by each experimental factor. This enabled us to identify potential groupings and assess which genes may be contributing to the observed patterns.

Prior to all PCA-related analyses, we selected the top 10% most variable genes across samples since these genes are more likely to capture biological signal, whereas low-variance genes typically contribute little or are uninformative for the purpose of visualizing the trends in the data.

2.5 Differential expression analysis

The DE analysis approach used in this study follows the methodology implemented in the `DESeq2` package. The core idea is to model the raw counts for each gene using a negative binomial (NB) distribution, where the logarithm of the normalised mean is modeled as a linear combination of coefficients corresponding to contrasts of experimental conditions. Hypothesis

testing is then performed on these coefficients to assess whether the corresponding contrasts result in statistically significant differential expression for a given gene. In the remainder of this section, we briefly outline the key features of this modeling framework, as described in (11).

Let c_{ij} denote the observed raw count for gene i in sample j. We assume that c_{ij} has a NB distribution with mean μ_{ij} and gene-specific dispersion parameter α_i; that is, $c_{ij} \sim \mathcal{NB}(\mu_{ij}, \alpha_i)$. The mean μ_{ij} is modelled as the product of a sample-specific size factor s_j and a normalised expression level q_{ij}, such that $\mu_{ij} = s_j q_{ij}$. The size factor s_j is estimated using the median-of-ratios method, as mentioned in Section 2.3. Finally, the dispersion parameter is used to model the variance of the counts via $\text{var}(c_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$.

The log-transformation of the normalised expression level, q_{ij}, is modelled in this study as

$$\begin{aligned}\log(q_{ij}) = \beta_{i0} + \beta_{i1} x_j^{(\text{FL})} + \beta_{i2} x_j^{(\text{YC})} + \beta_{i3} x_j^{(\text{YE})} + \beta_{i4} x_j^{(\text{YCE})} + \beta_{i5} x_j^{(30)} + \\ \beta_{i6} x_j^{(\text{FL})} x_j^{(\text{YC})} + \beta_{i7} x_j^{(\text{FL})} x_j^{(\text{YE})} + \beta_{i8} x_j^{(\text{FL})} x_j^{(\text{YCE})} + \beta_{i9} x_j^{(\text{FL})} x_j^{(30)},\end{aligned} \tag{2.1}$$

where the intercept term β_{i0} denotes the baseline expression level of gene i under the reference condition. The coefficients β_{i1} to β_{i5} represent the main effects of the experimental factors: $x_j^{(\text{FL})}$ is an indicator variable for the FLC1Δ strain, with the WT strain serving as the reference level; $x_j^{(\text{YC})}$, $x_j^{(\text{YE})}$, and $x_j^{(\text{YCE})}$ are indicators for the different media conditions, with Y as the reference level; and $x_j^{(30)}$ indicates the 30°C temperature condition, with 37°C as reference. Together, this implies that the reference experimental condition is WT-37-Y. The interaction terms β_{i6} to β_{i9} model how the effects of media and temperature differ under the presence or absence of the FLC1 gene. No interaction between media and temperature was included, as these factors are not fully crossed in the experimental design.

Accurate estimation of the gene-specific dispersion parameters, α_i, is crucial for robust DE analysis. However, in controlled experiments with small sample sizes (often only two or three replicates, as in this study) the maximum likelihood estimation (MLE) dispersion estimates can be highly variable, potentially compromising the reliability of the DE significance testing (11). To address this, `DESeq2` employs an an empirical Bayes shrinkage approach that dispersion estimates towards an mean expression-dependent trend. The procedure consists of the following steps:

1. Initial fitting: Estimate dispersion for each gene using the method of moments, then fit the model via MLE to obtain initial estimates $\hat{\mu}_{ij}^{(0)}$.
2. Gene-wise dispersion: Compute gene-wise dispersion estimates $\hat{\alpha}_i^{\text{GW}}$ by maximizing the Cox-Reid profile likelihood of α_i given $\hat{\mu}_{ij}^{(0)}$.
3. Trend estimation: Fit a parametric regression of $\hat{\alpha}_i^{\text{GW}}$ on the mean normalised counts $\bar{\mu}_i = \frac{1}{m} \sum_j \frac{c_{ij}}{s_j}$, where m is the number of samples. This trend, $\bar{\alpha}_i$ models the expected dispersion as a function of the mean expression.

4. Shrinkage: Treat the trend values $\bar{\alpha}_i$ as prior means in a log-normal prior, $\log(\alpha_i) \sim \mathcal{N}(\log(\bar{\alpha}_i), \sigma^2)$. The prior variance σ^2 is estimated from the residual variance of log-dispersion values around the fitted trend. For experiments with fewer than three residual degrees of freedom (samples minus model parameters), σ^2 is estimated via simulation by matching the empirical distribution of residuals to simulated densities.
5. Final estimates: The final dispersion values are obtained as the maximum a posteriori (MAP) estimates, derived by combining the Cox-Reid profile log-likelihood with the log-normal prior distribution.

While this shrinkage approach performs well on average, it can underestimate the variance for genes with genuinely high dispersion, increasing the risk of false positives. To mitigate this, `DESeq2` uses the gene-wise dispersion estimate instead of the shrunken value when the former exceeds the fitted trend by more than two residual standard deviations.

The log-fold change (LFC) quantifies the change in gene expression between two experimental conditions and corresponds to the model coefficients in Equation 2.1, typically expressed on a $\log_2$ scale. These coefficients are estimated via MLE, using the previously obtained shrunken dispersion estimates. However, MLE-derived LFC estimates can be highly variable, particularly for genes with low counts, making them noisy and potentially misleading. To enhance their stability and interpretability, `DESeq2` applies an empirical Bayes shrinkage procedure introduced by (15) and implemented in the `apeglm` package. This method proceeds through the following steps:

1. Prior specification: For each gene,a heavy-tailed Cauchy prior is assigned to the coefficients β_{ik}, for $k = 1, \ldots, p$, where p is the number of coefficients excluding the intercept. The prior is defined as $\beta_{ik} \sim \mathcal{C}auchy(0, S_k)$, where 0 is the location parameter and S_k is the scale parameter, which determines the amount of shrinkage.
2. Scale estimation: S_k is adaptively estimated from the data through an empirical Bayes approach. Specifically, that the MLEs $\hat{\beta}_{ik}$ follow a normal distribution around the true coefficient β_{ik}, with variance equal to their squared strandard errors e^2_{ik}. It further assumes that the true coefficients β_{ik} are normally distributed with mean zero an unknown variance A_k. The empirical Bayes estimate $\hat{A}_k$ is used to derive the scale: $S_k = \sqrt{\hat{A}_k}$. If the MLEs are not provided, a default scale of $S_k = 1$ is used.
3. Final estimates: The final shrunken LFC estimates are obtained by combining the NB likelihood from the observed count data with the adaptive Cauchy prior. The result is the MAP estimate of β_{ik}.
4. Uncertainty quantification: A Laplace approximation is applied to the posterior distribution of β_{ik} to approximate its variance. This provides an estimate of the posterior standard deviation, which can be used for statistical inference.

Shrinkage of LFC does not affect the number of genes identified as significantly differentially expressed, but improves the stability of estimates for downstream analyses such as visualization, gene filtering or functional analysis, where more reliable effect size estimates are needed (16).

The default approach in `DESeq2` is to perform a Wald test to assess whether a LFC is significantly different from zero, using the MLEs. In this study, the null hypothesis of no differential expression for gene i between the reference condition and the condition indicated by x_{ik} is $H_0 : \beta_{ik} = 0$, with the alternative hypothesis being $\beta_{ik} \neq 0$. While the assumption of a zero LFC may be biologically implausible for many genes due to the high connectivity of gene regulatory networks, it serves as a useful baseline for statistical testing, particularly in small-scale studies such as this one with only three replicates per condition (11). `DESeq2` adjusts p-values for multiple testing using the Benjamini–Hochberg procedure, which controls the false discovery rate (FDR) (17). Given the limited statistical power in our design due to few replicates per condition, we adopt an adjusted p-value threshold of 0.1 to determine significance.

Returning to the discussion of pre-filtering in Section 2.2, `DESeq2` applies independent filtering by default, using the mean of normalised counts across all samples as the filtering criterion. A threshold is automatically selected to maximise the number of discoveries at a target FDR, and genes with mean counts below this threshold are excluded from downstream analysis. This method is considered independent because the filter is uncorrelated with the test statistic under the null, a property shown to improve statistical power (6). The adjusted p-value threshold for significance matches the target FDR, set to 0.1. `DESeq2` also includes automatic outlier detection to identify observations that disproportionately influence LFC estimates. An observation is flagged as an outlier if its Cook's distance exceeds the 99th percentile of the $F(q, q - n)$ distribution, where q is the number of model parameters and n the number of samples. Outlier detection is skipped for conditions with two or fewer replicates. For six or fewer replicates, genes with outliers are excluded; with seven or more, outliers are replaced by imputed values (trimmed means), and the model is refitted.

The rlog transformation, introduced in Section 2.3, transforms the raw count c_{ij} of gene i at sample j as $\text{rlog}(c_{ij}) = \log_2(q_{ij}) = \beta_{i0} + \beta_{ik}$. It involves fitting a model similar to Equation 2.1 and computing shrunken LFCs relative to the baseline expression, β_{i0}, using an empirical Bayes shrinkage approach, as previously described for the full model. For this transformation, `DESeq2` uses blind dispersion estimation by default: it ignores the experimental design and treats all samples as replicates of the same condition when re-estimating dispersions. This makes the rlog-transformed data well-suited for unsupervised or exploratory analyses such as quality control, where influence from the experimental groups is undesirable. The VST transformation is also blinded by default.

Our downstream analysis focused on the four interaction terms β_{i6} to β_{i9} in Equation 2.1, rather than the main effects, to identify genes whose stress response is modulated by the presence or absence of the FLC1 gene, though main effects were also considered. To assess model fit, we examined the dispersion plot, where gene-wise dispersion estimates should scatter around the fitted mean-dependent trend, with dispersion decreasing as mean expression increases (16). After multiple testing, we inspected the distribution of p-values; a well-calibrated procedure should yield a histogram with a spike near zero and a relatively flat distribution elsewhere (18). We generated MA plots for both unshrunken and shrunken LFCs, highlighting significant and non-significant genes. This visualisation helps assess the magnitude and distribution of fold

changes across expression levels, with significant genes typically spread across the entire range (16). Volcano plots were used to report the top 10 most significant genes based on adjusted p-value. Finally, we created interaction plots of raw and normalised counts to examine the expression profiles of these top genes.

We compared the sets of genes identified as significant for the interaction terms by examining both the number of genes per interaction and the overlaps between them. To identify shared expression patterns, we focused on the top 500 genes with higher variance across the samples for all interaction terms and applied hierarchical clustering to the correlation matrix of z-scores computed from normalised counts. For this step, rlog and VST transformations were performed without blinding, meaning the design matrix was used to estimate dispersion parameters. The optimal number of clusters was determined using the Gap statistic (19) (HOW IS IT CHOSEN). Cluster expression patterns were explored through heatmaps and line plots illustrating expression characteristics across experimental conditions.

2.6 Software

3 Results

Gene-wise total read counts ranged from 0 to 2.97 million, with gene CNAG_06125 showing the highest total count across all samples. A total of 25 out of 6795 genes had zero total counts. Library sizes (i.e., total read counts per sample) ranged from 4.26 to 9.02 million, with sample S12 from condition FL-Y-37 having the largest library. At the condition level, WT-Y-30 showed the highest total library size (25.93 million). A barplot of library sizes across samples is shown in Figure S4.1, highlighting variability across libraries and the need for normalisation prior to comparing expression levels.

After pre-filtering, we retained 6613 genes (97.32%), removing 182 genes and moderately reducing the multiple testing burden for DE analysis. We applied CPM normalisation and performed $\log_2$-CPM, rlog, and VST transformations on the filtered count matrix, as described in Section 2.3. Figure S4.2 shows boxplots of expression values by sample and condition for each transformation. CPM alone does not stabilise variance, resulting in highly skewed distributions with many outliers. In contrast, $\log_2$-CPM, rlog, and VST reduced skewness and improved centring, although outliers are still present. Among these, rlog and VST produce more uniform distributions across samples than the simple log transform, which has limitations discussed in Section 2.3. The effect of the transformations on variance stabilisation is illustrated in Figure S4.3, which plots gene-wise standard deviations against the rank of mean expression under each transformation. For CPM, the trend line (red) shows a pronounced peak at high expression levels, indicating greater variability. In contrast, the transformed data exhibit a much flatter trend, with standard deviations more uniform across the expression range, reflecting effective variance stabilisation.

PCA was performed on the transformed data of the genes at the 10% of highest variance, yielding an optimal number of three components for all methods except CPM, which retained only one. These components explained 92.86%, 93.19%, and 92.71% of the total variance for $\log_2$-CPM, rlog, and VST, respectively, while CPM explained only 75.19%. The scree plot for the rlog-transformed data, shown in Figure S4.4, illustrates the retained components and their associated variance. The score plots for all retained component pairs in Figure S4.6 show that replicates from the same condition cluster tightly, indicating high consistency and replicate quality. Since the rlog and VST transformations yield similar explained variance (slightly higher for rlog) we will present results based on the rlog-transformed data from this point forward.

In Figure 3.1, subplot A1, samples separate clearly by strain along PC1 and by media condition along PC2. Notably, samples from conditions Y and YC, as well as YE and YCE,

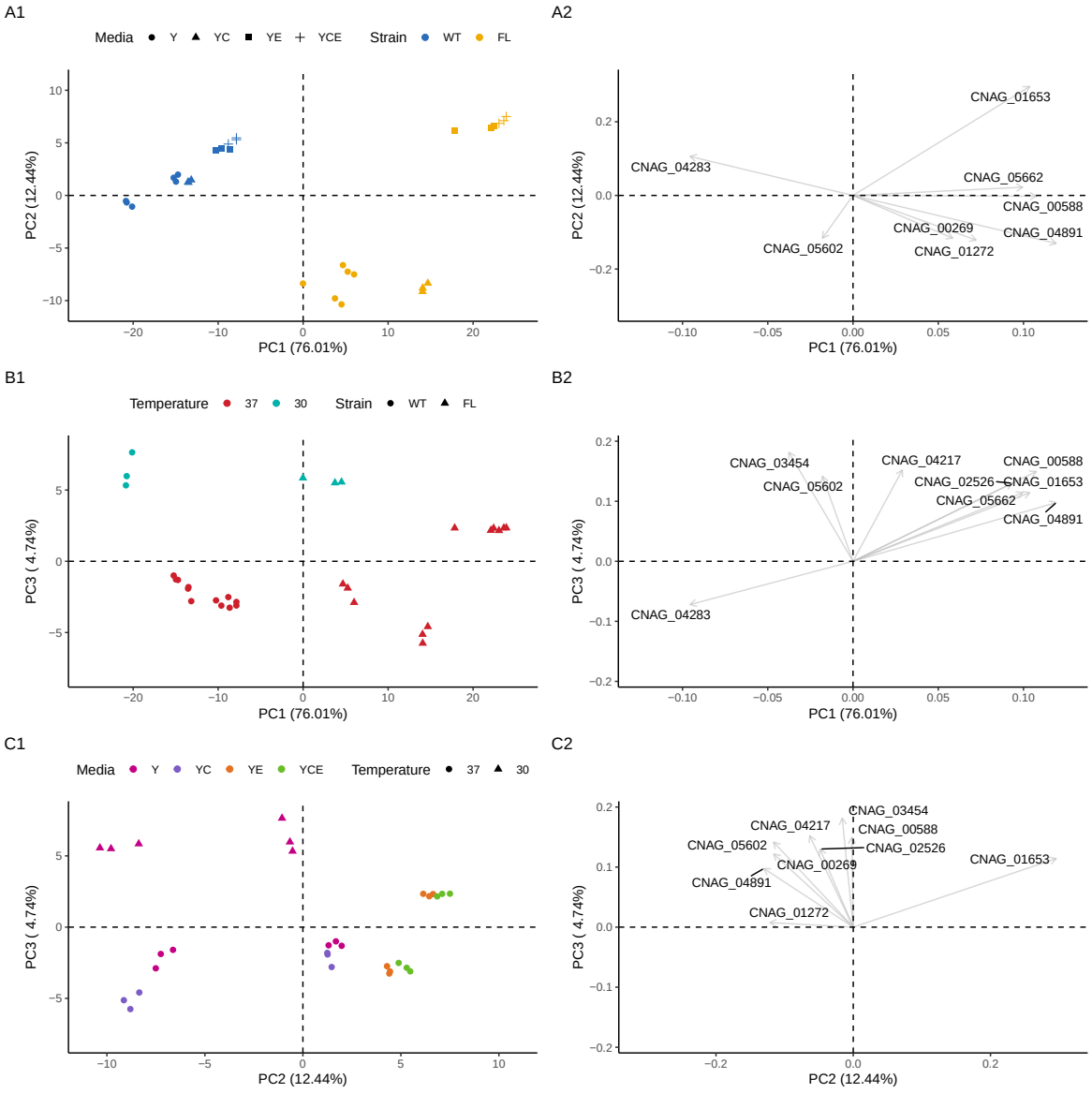

Figure 3.1: PCA plot.

cluster closely together, suggesting that EGTA drives most of the observed separation rather than CFW. Interestingly, the distinction between samples with and without EGTA is more pronounced in the FLC1Δ strain than in the WT strain. A clear separation is also evident between samples grown at the Y and YC media conditions just for the the FLC1Δ strain and not in the other conditions. These patterns suggest a potential interaction between strain and the media at which samples were grown. Figure 3.1, subplot A2, shows that CNAG_04283 and CNAG_05602 primarily drive the separation toward the WT strain, with the other genes contributing to the separation toward the FLC1Δ strain. However, the effect of CNAG_04283 is substantially stronger, which is expected since it corresponds to the FLC1 gene itself. Genes CNAG_05662 and CNAG_00588 align closely with the FLC1Δ strain samples, while CNAG_00269, CNAG_01272, and CNAG_04891 appear to contribute to the separation of FLC1Δ samples grown at the Y and YC media conditions. Finally, CNAG_01653 is strongly associated with FLC1Δ strain samples grown with EGTA, suggesting a potential gene-specific response to this experimental condition.

Moving to Figure 3.1, subplot B1, samples separate by temperature along PC3, with further separation by strain within each temperature group, suggesting a potential interaction between these factors. In Figure 3.1, subplot B2, gene CNAG_04217 appears to drive the separation of FLC1Δ samples at 30°C, while CNAG_02526, CNAG_05662, CNAG_01653, CNAG_00588, and CNAG_04891 contribute to the separation of FLC1Δ samples at 37°C. Conversely, CNAG_03454 and CNAG_05602 drive separation toward WT samples at 30°C, and CNAG_04283 (FLC1 gene) drives separation toward WT samples at 37°C. Finally, in Figure 3.1, subplot C1, we again observe temperature-driven separation along PC3 and media-driven separation along PC2. In Figure 3.1, subplot C2, Most genes with strong loadings on these components are associated with samples grown at 30°C in YPD. Notably, CNAG_01653 makes a major contribution to the separation of samples grown at 37°C in the YE and YCE media conditions.

The results of the pairwise ANOVA tests between the retained principal components and experimental factors (Figure 3.2) show that PC1 primarily captures variation associated with the strain condition. PC2 is influenced by all three experimental factors, with the strongest association observed for the media condition. PC3 reflects separation driven by both media and temperature, with the most significant effect corresponding to temperature. The top 10 genes contributing most strongly to each principal component are shown in Figure 3.3. Notably, CNAG_01653 exhibits the largest loading magnitude overall, particularly in PC2, and also shows substantial contributions to PC1 and PC3. CNAG_04283, which encodes the FLC1 gene, contributes strongly to both PC1 and PC2. Another prominent gene, CNAG_04891, is among the top contributors to both PC1 and PC2 as well. For PC3, CNAG_03454 shows the highest association, followed by CNAG_04217 and CNAG_00588, the latter also contributing notably to PC1.

Moving on to model fitting for differential expression analysis, we observed that the shrunk dispersion estimates aligned well with the mean-dependent trend, indicating a good overall fit

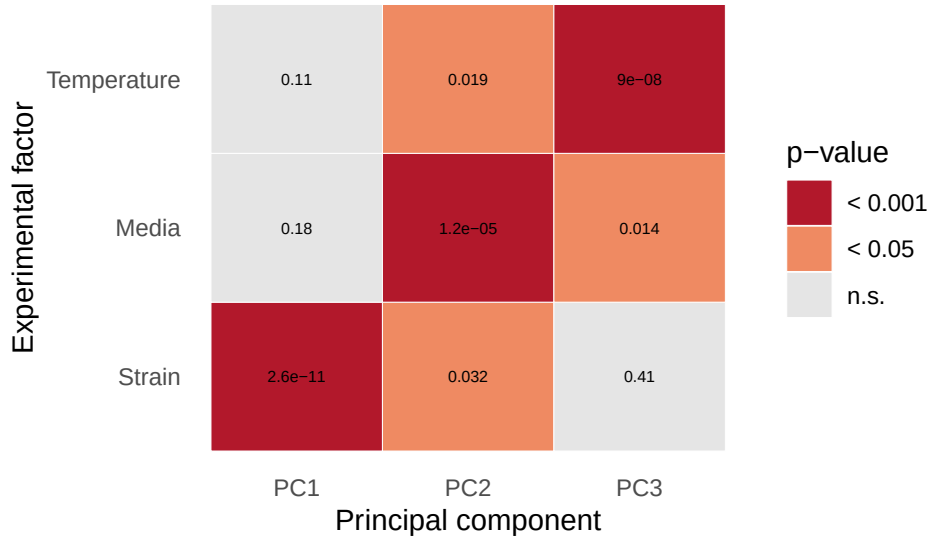

Figure 3.2: ANOVA plot.

(Figure S4.7). This pattern was more consistent in the filtered dataset, whereas in the unfiltered dataset, the trend appeared to be overly influenced by genes with low mean expression levels.

The contrasts used to test the significance of the model coefficients (excluding the intercept), which correspond to the log-fold changes (LFCs), were labelled to reflect both the main effects and interaction terms included in the DE analysis. Main effects included "Strain (FL)" for the FLC1Δ strain versus WT, "Media (YC)", "Media (YE)", and "Media (YCE)" for the addition of CFW, EGTA, and both to YPD, respectively, and "Temperature (30)" for growth at 30°C versus 37°C. Interaction terms were labelled as "Strain (FL) + Media (YC)", "Strain (FL) + Media (YE)", and "Strain (FL) + Media (YCE)" to capture the combined effect of strain and media condition, and "Strain (FL) + Temperature (30)" for the interaction between strain and temperature. As previously mentioned, our analysis focused on the interaction terms. Unless otherwise noted, results are based on the filtered dataset.

The (unadjusted) p-value histograms for each contrast display the expected uniform distribution under the null hypothesis, indicating a well-calibrated multiple testing procedure (Figure S4.8). Similar results were observed for the unfiltered dataset (not shown). No outliers were detected in any contrast. However, 129 genes (1.95%) were automatically removed due to low mean normalized counts during the testing of the Strain (FL) + Media (YC) term, as part of `DESeq2`'s independent filtering procedure. MA plots for all contrasts showed a uniform distribution of significantly upregulated and downregulated genes across the range of mean expression levels, suggesting that the procedure is not biased towards highly expressed genes (Figure S4.9).

The Strain (FL) + Media (YCE) contrast yielded the highest number of differentially expressed

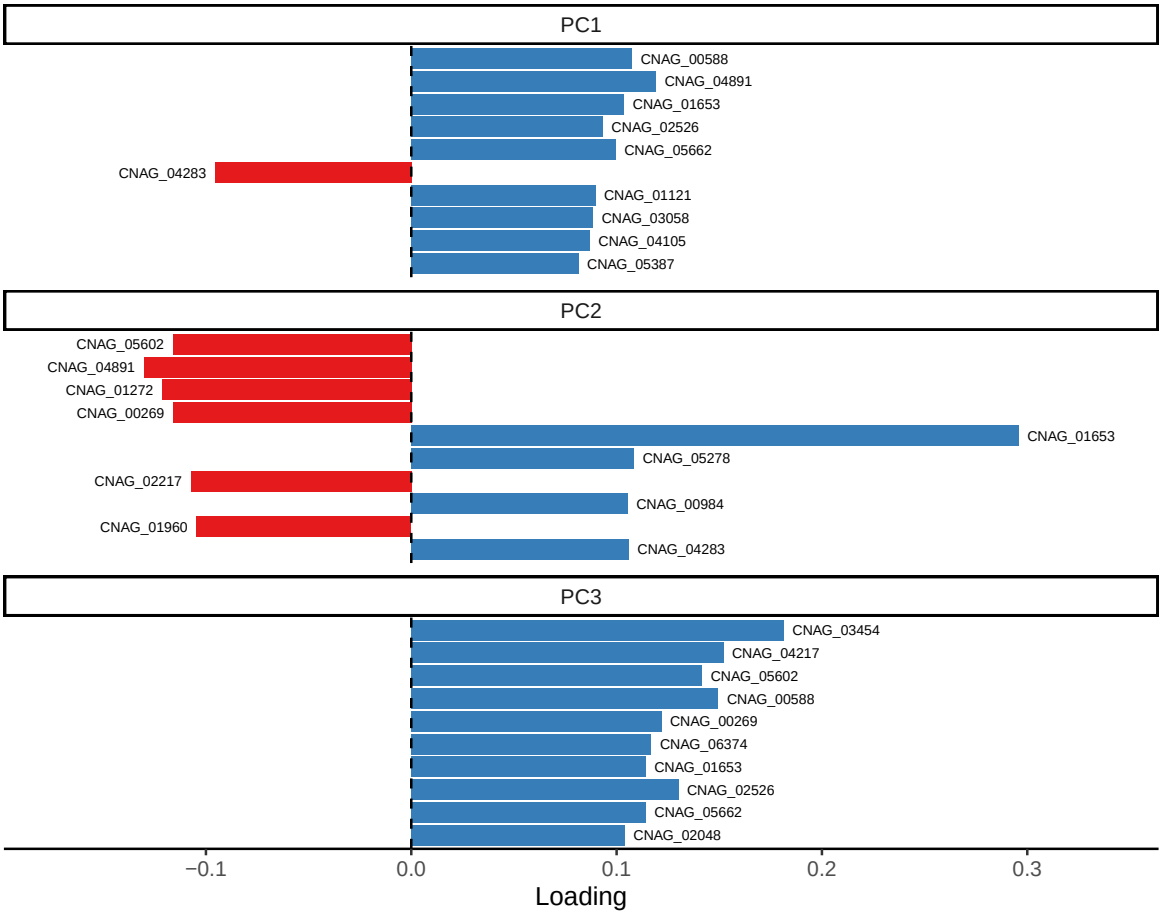

Figure 3.3: Loadings plot.

Table 3.1: Number and percentage of upregulated and downregulated genes identified at an FDR threshold of 0.1 for each contrast tested, along with the total number of DE genes.

Contrast	Upregulated	Downregulated	Total
Strain (FL) + Media (YC)	476 (37.96%)	778 (62.04%)	1254 (100%)
Strain (FL) + Media (YCE)	955 (42.96%)	1268 (57.04%)	2223 (100%)
Strain (FL) + Media (YE)	848 (48.02%)	918 (51.98%)	1766 (100%)
Strain (FL) + Temperature (30)	421 (36.39%)	736 (63.61%)	1157 (100%)

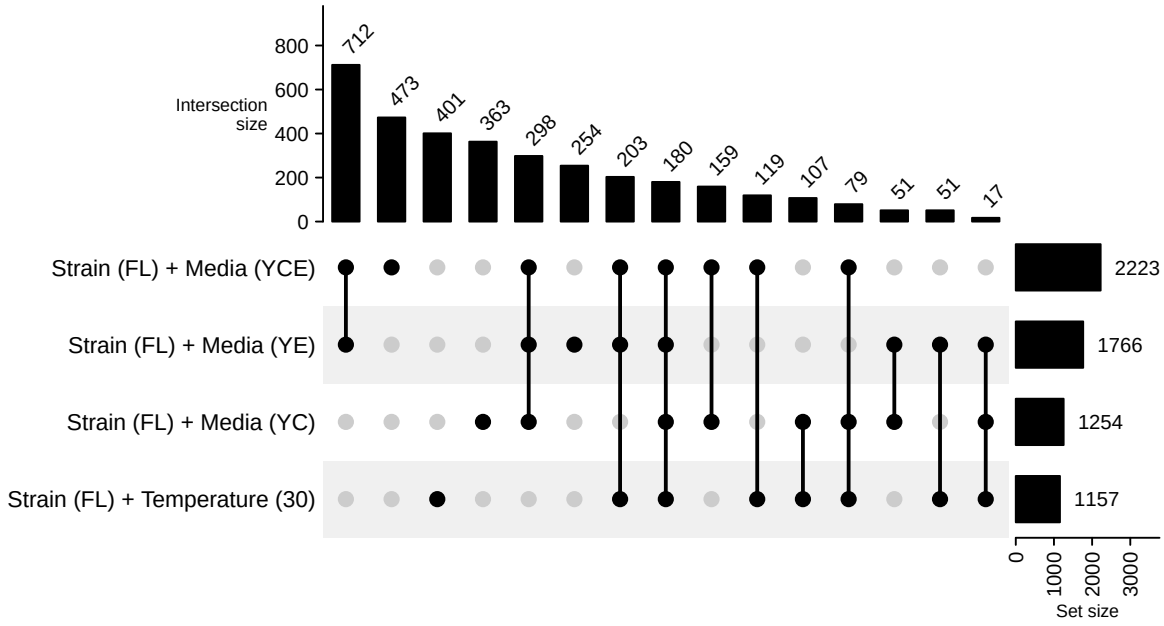

Figure 3.4: UpSet plot.

(DE) genes (2223), while Strain (FL) + Temperature (30) had the fewest (767) (Table 3.1). All contrasts showed more downregulated than upregulated genes. The highest percentage of downregulated genes was observed in Strain (FL) + Temperature (30) (63.61%), although the contrast with the greatest absolute number of downregulated genes was Strain (FL) + Media (YCE) (1268). The highest proportion of upregulated genes was found in Strain (FL) + Media (YE) (48.02%), while the highest count of upregulated genes overall was again seen in Strain (FL) + Media (YCE) (955) (Table 3.1). The Strain (FL) + Media (YCE) and Strain (FL) + Media (YE) contrasts shared the highest number of differentially expressed (DE) genes in common (712). Strain (FL) + Media (YCE) also had the greatest number of uniquely identified DE genes (473), i.e. those not shared with any other contrast. This was followed by Strain (FL) + Temperature (30) (401), Strain (FL) + Media (YC) (363), and Strain (FL) + Media (YE) (254) (Figure 3.4). Additionally, 180 DE genes were found to be shared across all contrasts (Figure 3.4).

Figure 3.5 highlights the top 10 most significant genes for each contrast. Notably, all of the top genes for the Strain (FL) + Media (YC) interaction were downregulated, while all but one of the top genes for the Strain (FL) + Temperature (30) interaction were upregulated. Gene CNAG_06576 appeared most frequently among the top genes, being significantly differentially expressed in all contrasts except Strain (FL) + Media (YC). In addition, genes CNAG_00091, CNAG_00848, CNAG_00876, CNAG_00979, and CNAG_02335 were consistently ranked among the top genes in both the Strain (FL) + Media (YE) and Strain (FL) + Media (YCE) contrasts.

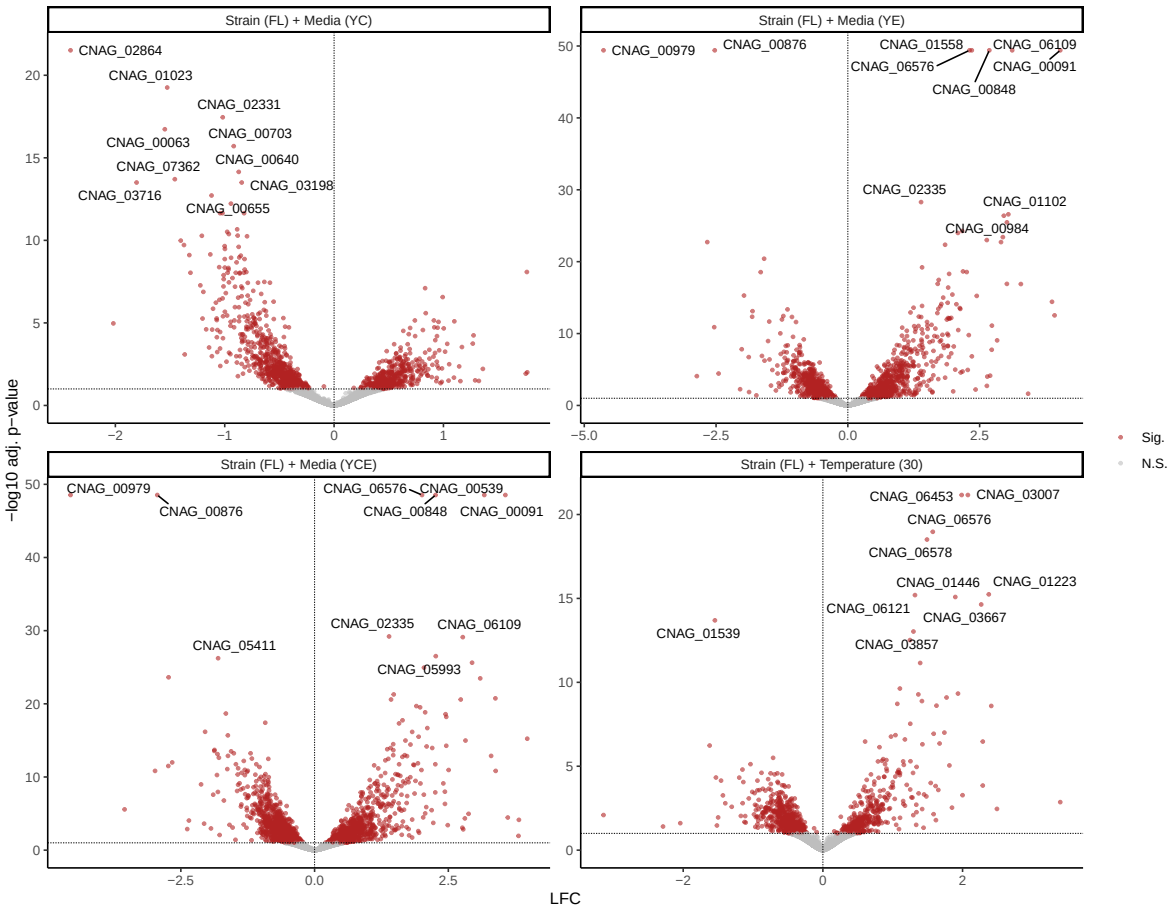

Figure 3.5: Volcano plot.

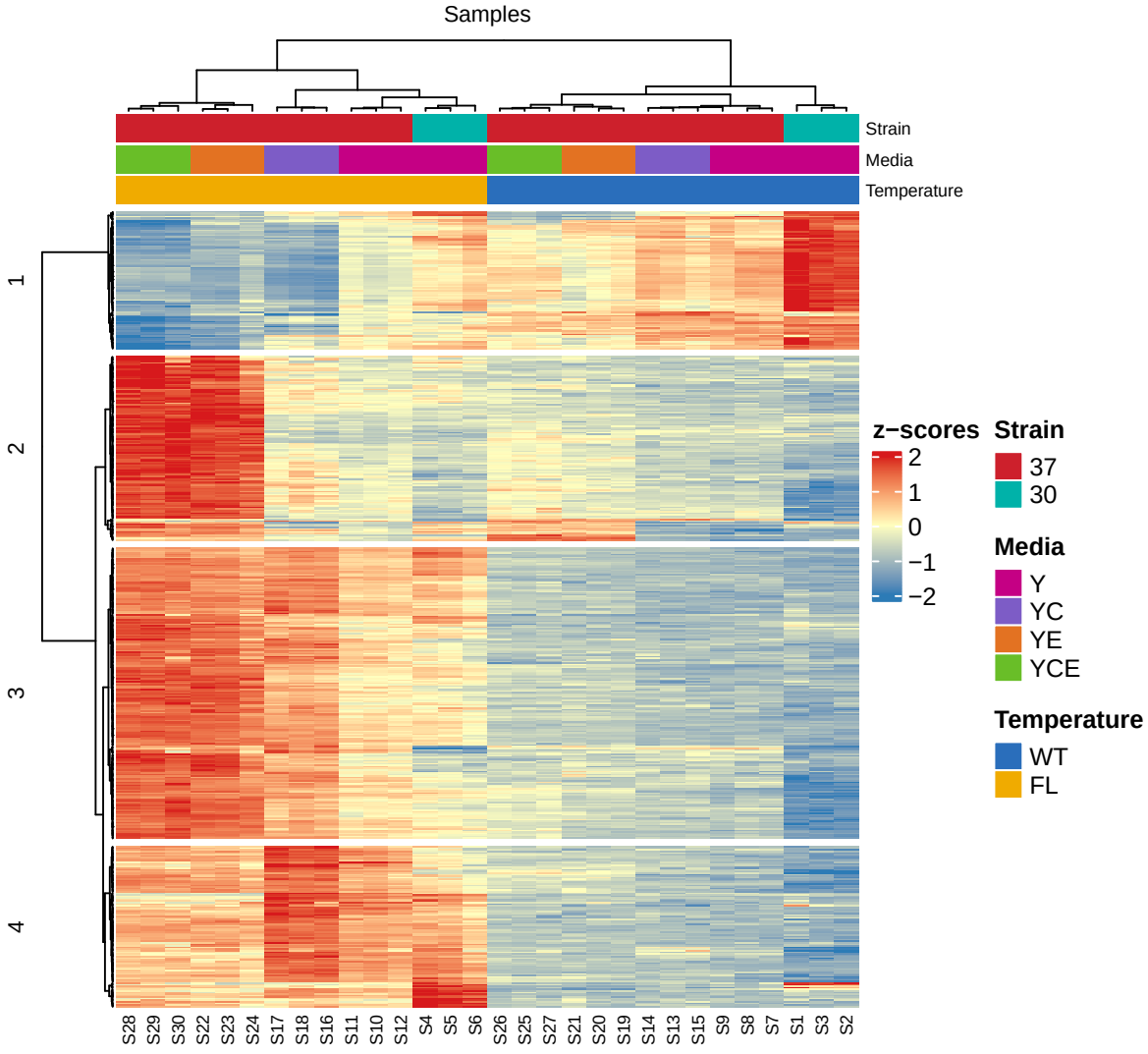

Figure 3.6: Heatmap.

The optimal number of clusters for the top 500 differentially expressed (DE) genes with the highest variance, based on the Gap statistic, was four when using z-scores of the rlog-transformed data (Figure S4.10). The same number of clusters was obtained using VST-transformed data. In contrast, the $\log_2$-CPM transformation produced an optimal number of one, which is uninformative and suggests that $\log_2$-CPM may be less suitable for capturing meaningful expression patterns in this context. A heatmap of the z-scored expression values across the four clusters (labelled 1–4) is shown in Figure 3.6. Since the heatmap is based on z-scores, positive values indicate expression levels above the gene's mean across all conditions,

while negative values indicate below-average expression. Cluster 1 stands out from the others, with genes showing above-average expression in the WT strain and below-average expression in the FL strain, especially under the WT-Y-30 condition, which is the basal condition. In contrast, clusters 2–4 contain genes with the opposite pattern (lower z-scores in WT and higher in FL) suggesting a strain-specific shift in relative expression. Within these, genes in cluster 2 tend to show above-average expression in the FL-YE-37 and FL-YCE-37 conditions and near or below-average expression in other FL combinations. Cluster 3 follows a similar pattern but with higher relative expression across more FL conditions than cluster 2. Cluster 4 displays an inverse trend to cluster 2, with higher average expression levels at the FL-Y-30, FL-Y-37, and FL-YC-37 conditions.

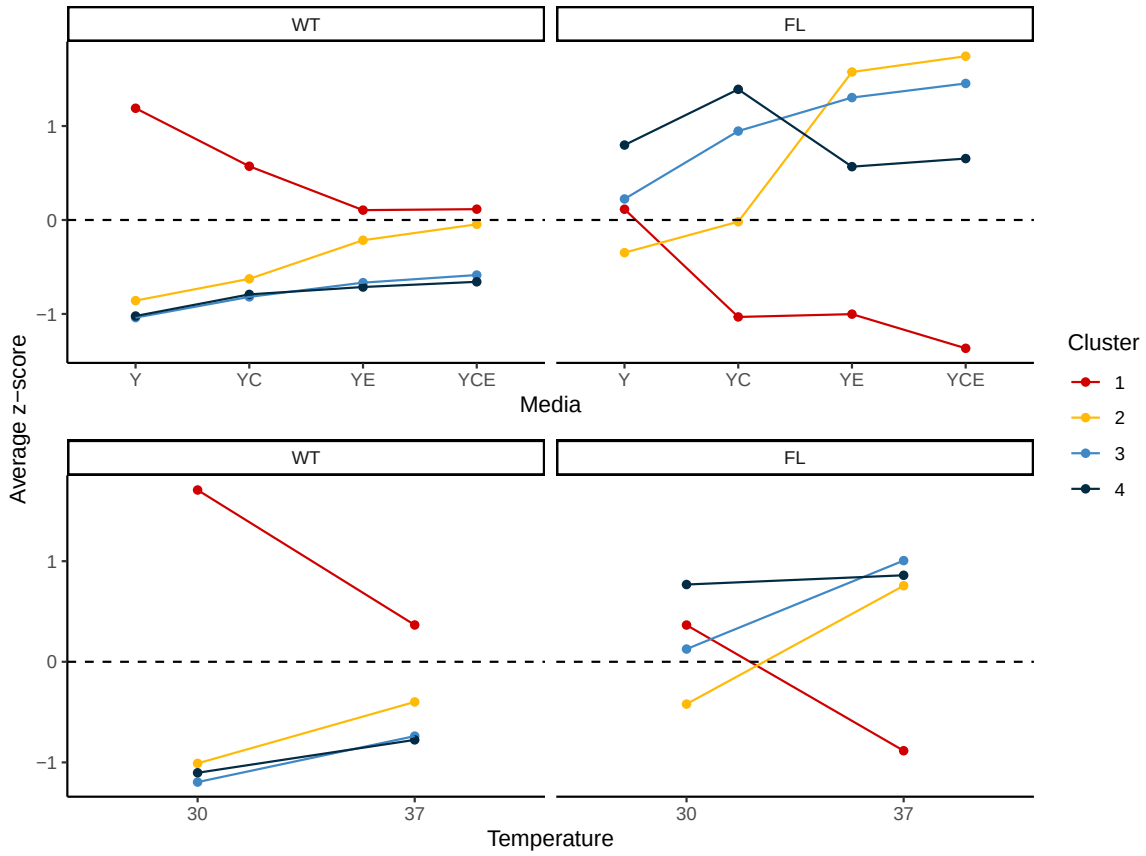

Figure 3.7: Cluster profile plot.

In Figure 3.7 we can observe the different average expression patterns across the levels of media and temperature by the WT and FL strain conditions. Cluster 1 have above-average expression levels across all the levels of media, while the other clusters have below-average expression levels. Cluster 1 show lower expression levels at the media conditions with EGTA than in the media condition without EGTA. Also the it is less expressed in YC than in Y, and

have similar average expression at the YE and YCE conditions where the expression level is close to average. The other clusters show an opposite pattern, with higher expression levels at the conditions with EGTA. For the FLC1Δ strain condition (FL), the patterns of the clusters across the media conditions change drastically. At the FL condition, now cluster 1 have less than average expression levels and the other higher than average. The average expression levels in cluster 1 reduce when adding CFW to the media conditions (from Y to YC, and from YE to YCE), while the expression levels are similar for the YC and YE conditions. Genes at cluster 2 increase their expression level when adding CFW to the media condition (from Y to YC, and YE to YCE), and have higher expression levels when EGTA is added (YE and YCE relative to Y and YC). Cluster 3 shows an increase when CFW is added (from Y to YC, and YE from YCE), whith highest expression levels where EGTA is added. Cluster 4 shows an increase when CFW is add (from Y to YC, and YE from YCE), although the increase is superior when adding CFW to the Y media condition, and have lowest expression levels when EGTA is present. Moving on to temperature, at WT, cluster 1 has above-average expression levels, while the other clusters have below-average expression levels. Cluster 1 has lower expression levels on average at 37°C than in 30°C, while the opposite pattern is shown at the other clusters. At the FL condition, the overall expression pattern switch, with cluster 1 having less expressed genes than the other clusters. Cluster 1 reduce its average expression level when passing from 30°C to 37°C, while the other clusters have the opposite pattern. Notably, genes at cluster 4 have similar expression level at 30°C and 37°C.

Figure 3.7 illustrates the average z-scored expression patterns of the four gene clusters across media and temperature conditions, separately for the WT and FL (FLC1Δ) strains. In the WT strain (top-left panel), cluster 1 genes show consistently above-average expression across all media types, with highest values at the Y media condition and a gradual decrease through YC and YE to YCE, although the difference is greater from Y to YC than froM YE to YCE. In contrast, Clusters 2, 3, and 4 show consistently below-average expression in WT, with expression increasing slightly with the addition of CFW and EGTA. Under the FL strain (top-right panel), the expression patterns invert: cluster 1 now shows below-average expression in all media, with a marked decrease when CFW and EGTA are present (from Y to YC, and from YE to YCE). Meanwhile, clusters 2 and 3 show progressively increasing expression across media, particularly in YE and YCE. Cluster 2 shows a marked increase when EGTA is added, while cluster 3 has a similar trend but with a less dramatic increase when EGTA is added. Cluster 4 shows increased expression when CFW is added (from Y to YC and from YE to YCE), with a stronger effect observed in the Y medium. However, its expression is lowest in conditions where EGTA is present.

Now, we move on to temperature. In the WT strain (bottom-left panel), Cluster 1 genes are expressed above average at both temperatures but show reduced expression at 37°C. Clusters 2–4 remain below average, though expression increases with temperature. In the FL strain (bottom-right panel), expression patterns again invert. Cluster 1 genes show near or below-average expression, decreasing further at 37°C. Meanwhile, clusters 2 and 3 show near or above-average expression, with expression increasing with temperature. Cluster 4 shows above-average expression, however, it maintains nearly constant expression across temperatures.

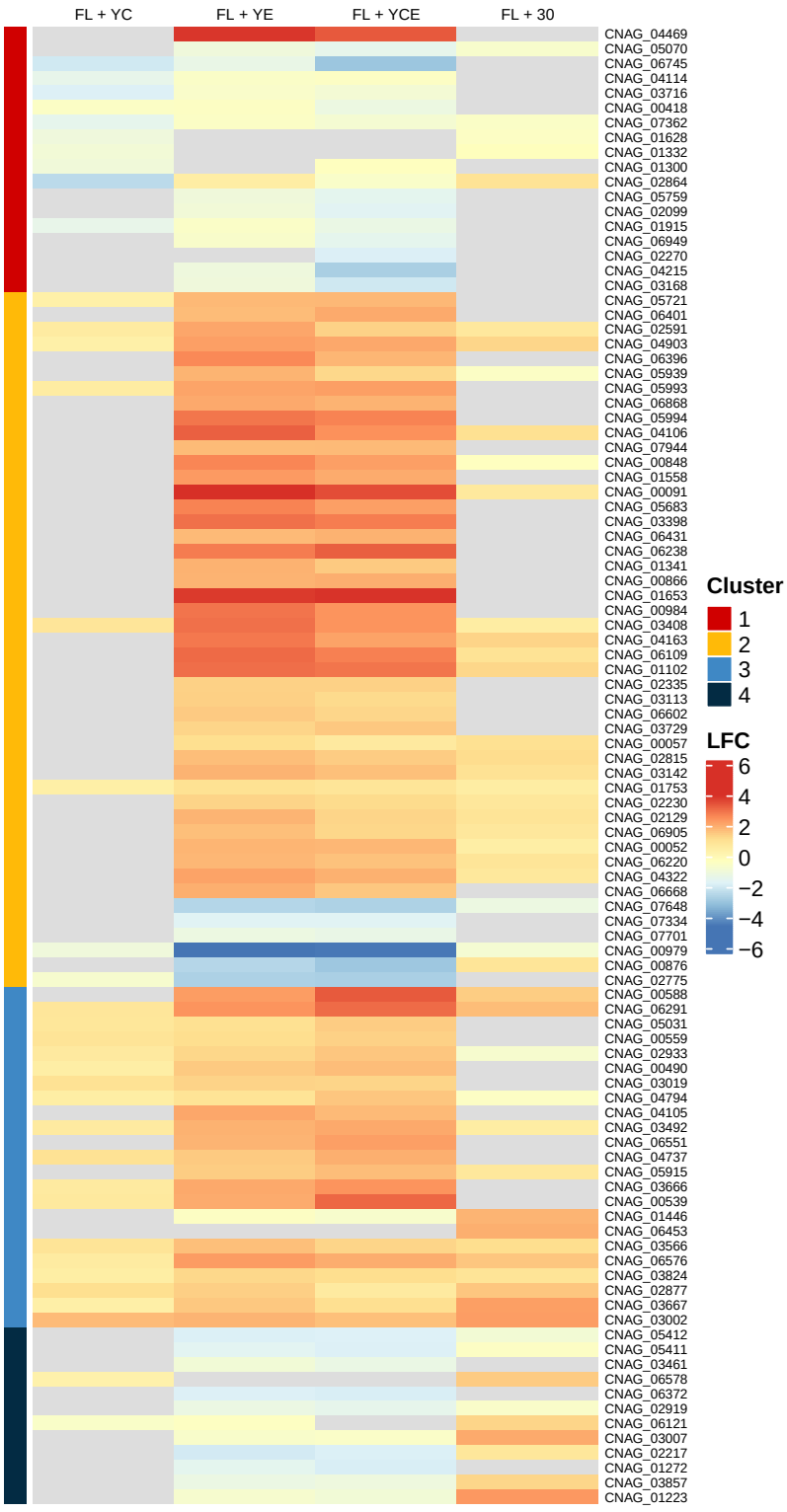

Figure 3.8: Heatmap LFC.

Finally, we ranked the genes we used for clustering according to their p-values (lowest first) and LFC magnitude (highest first), and computed an average ranking to select the top 100 genes. These genes are shown in Figure 3.8, which is a heatmap of the LFC across the contrasts for all the genes, alongside with the cluster they belong to. This plot was intented as a tool to select potentially important genes using most of the results presented so far. For example, we can observe that gene CNAG_04469 at the top is significant for both the interaction of strain and media at YE and YCE levels (and no significant for the other interactions), and it is upregulated for both interactions, so this gene gets upregulated at the FL strain condition and YE and YCE conditions relative to the WT-Y-37 condition. We can also see that this gene belongs to cluster 2, and we can expect to have an expression pattern similar to the described using Figure 3.7 for that cluster. In addition, we see that most of these genes have significant LFCs for the YE and YCE levels, and that genes that significant to YE are in most cases also significant to YCE.

To prioritise the most informative genes, we ranked the 500 genes used for clustering by two criteria: smallest adjusted p-value and largest absolute log2 fold change (LFC), and then computed an average rank across both metrics. The top 100 genes according to this combined ranking are displayed in Figure 3.8, which presents a heatmap of their LFCs across all contrasts, along with their corresponding cluster membership. This plot serves as a comprehensive summary tool, allowing us to identify candidate genes of interest by integrating statistical significance, expression changes, and clustering results. For instance, gene CNAG_04469, which appears at the top of the heatmap, is significantly differentially expressed in the interaction between strain and media at the YE and YCE levels, but not at other levels. In both contrasts, the gene is upregulated in the FL strain at the YE and YCE levels relative to the WT-Y-37 reference. This gene is a member of cluster 2, so we can expect an expression profile similar to the described for the genes in cluster 2 in Figure 3.7. Moreover, we observe that many of the top-ranked genes show significant LFCs for the YE and YCE conditions, and that significance in one is often accompanied by significance in the other.

4 Conclusions

References

1. May RC, Stone NR, Wiesner DL, Bicanic T, Nielsen K. Cryptococcus: From environmental saprophyte to global pathogen. Nature Reviews Microbiology. 2016;14(2):106–17.

2. Stempinski PR, Goughenour KD, Plooy LM du, Alspaugh JA, Olszewski MA, Kozubowski L. The cryptococcus neoformans Flc1 homologue controls calcium homeostasis and confers fungal pathogenicity in the infected hosts. Mbio. 2022;13(5):e02253–22.

3. Wang Z, Gerstein M, Snyder M. RNA-seq: A revolutionary tool for transcriptomics. Nature reviews genetics. 2009;10(1):57–63.

4. Chen Y, Lun TLL, Smyth GK. From reads to genes to pathways: Differential expression analysis of RNA-seq experiments using rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016;5(1438).

5. Sha Y, Phan JH, Wang MD. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In: 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE; 2015. p. 6461–4.

6. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. Proceedings of the National Academy of Sciences. 2010;107(21):9546–51.

7. Chen Y, Chen L, Lun ATL, Baldoni P, Smyth GK. edgeR v4: Powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. Nucleic Acids Research. 2025;53(2):gkaf018.

8. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31(22):3625–30.

9. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics [Internet]. 2012 Sep;14(6):671–83. Available from: https://doi.org/10.1093/bib/bbs046

10. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology. 2010;11:1–9.

11. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 2014;15:1–21.

12. Anders S, Huber W. Differential expression analysis for sequence count data. Nature Precedings. 2010;1–1.

13. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome biology. 2019;20(1):296.

14. Dinno A. Exploring the sensitivity of horn's parallel analysis to the distributional form of random data. Multivariate behavioral research. 2009;44(3):362–88.

15. Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: Removing the noise and preserving large differences. Bioinformatics. 2019;35(12):2084–92.

16. Mistry M, Piper M, Liu J, Khetani R. Hbctraining/DGE_workshop_salmon_online: Differential gene expression workshop lessons from HCBC (first release) [Internet]. Zenodo; 2021. Available from: https://doi.org/10.5281/zenodo.4783481

17. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological). 1995;57(1):289–300.

18. The Pennsylvania State University. STAT 555: Statistical analysis of genomics data. https://online.stat.psu.edu/statprogram/stat555; 2025.

19. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the royal statistical society: series b (statistical methodology). 2001;63(2):411–23.

Appendix

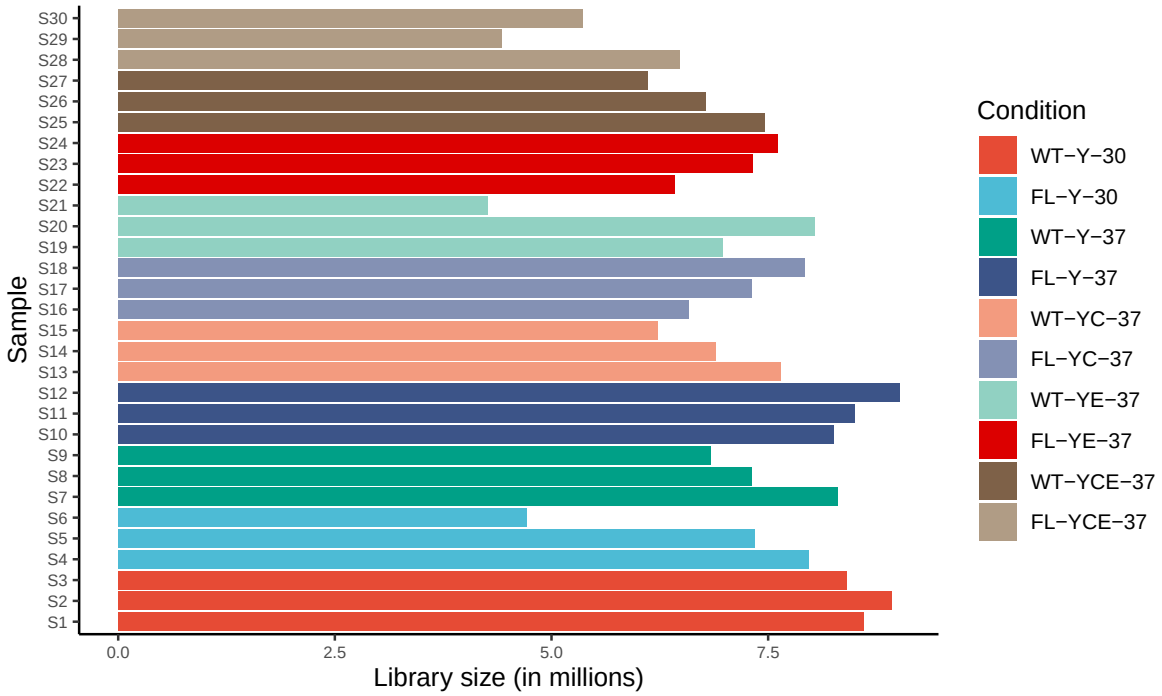

Figure S4.1: Library sizes.

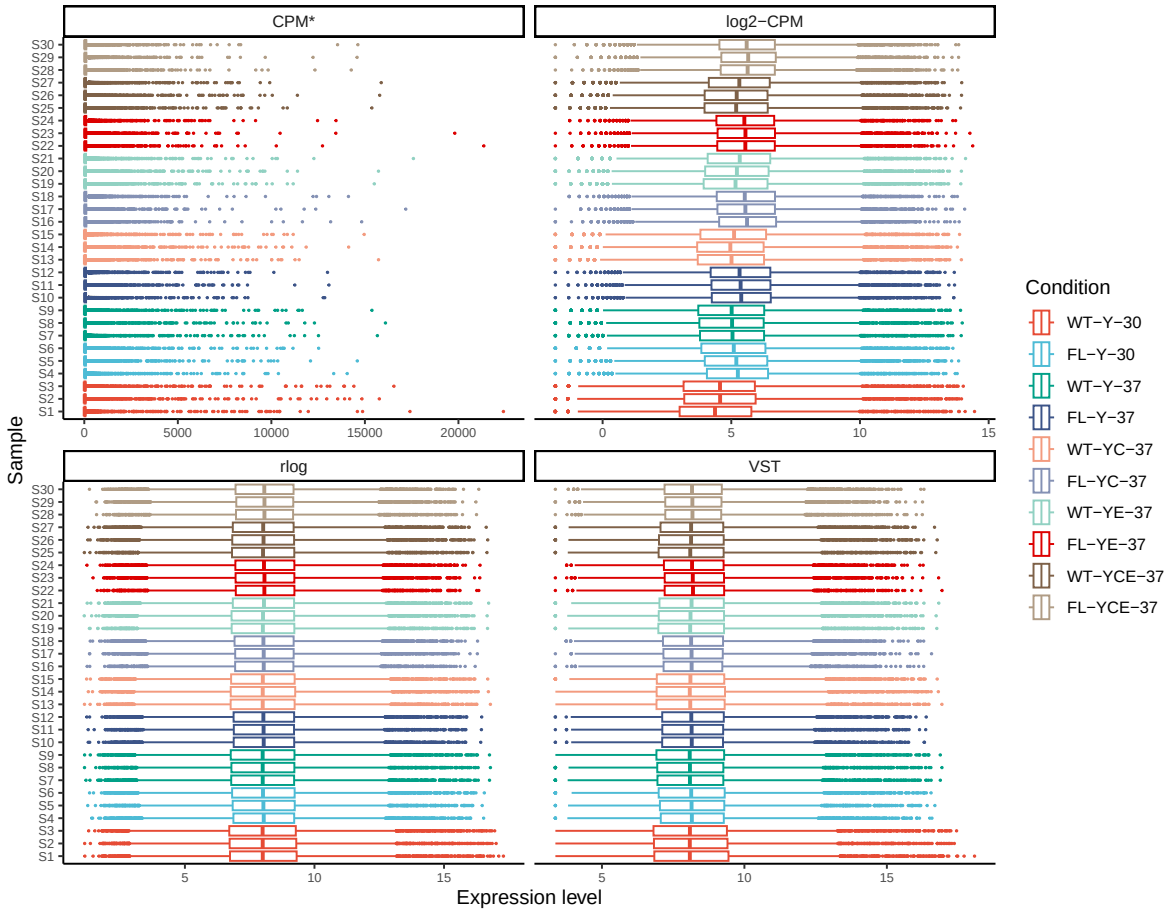

Figure S4.2: Boxplots.

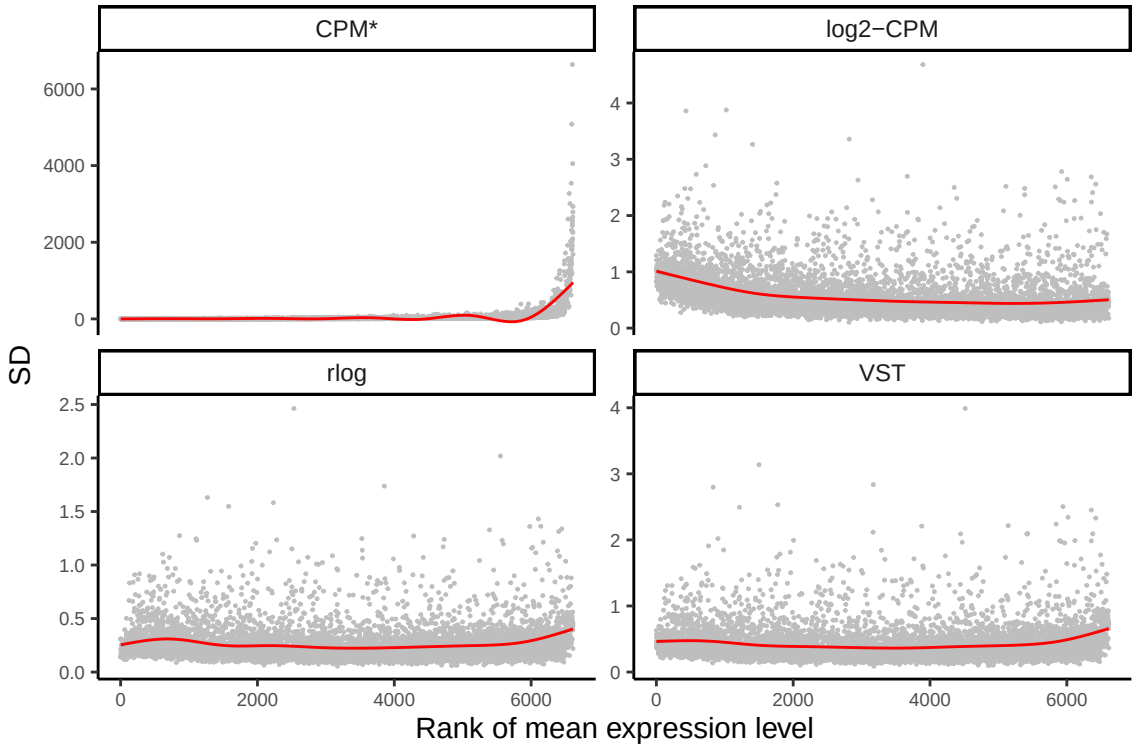

Figure S4.3: Mean-SD plot.

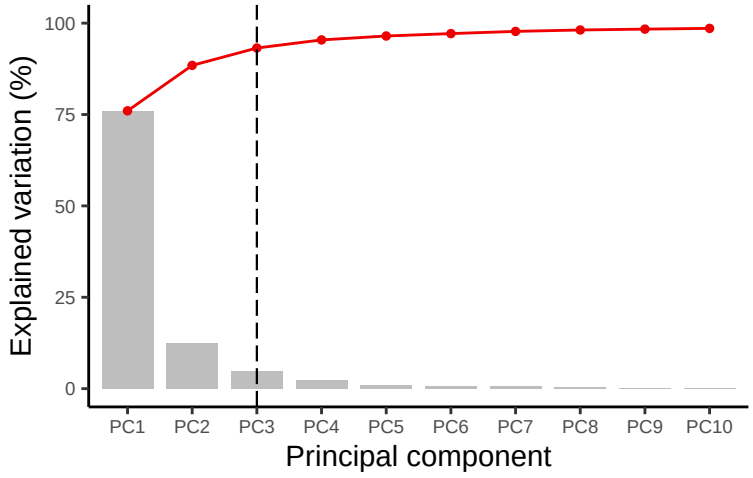

Figure S4.4: Scree plot.

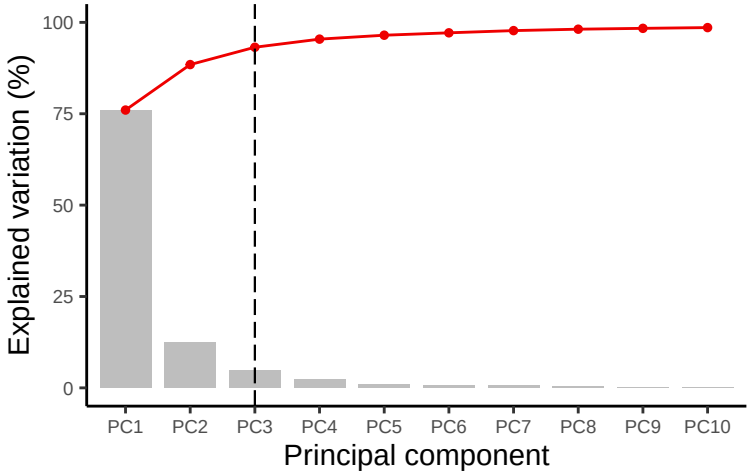

Figure S4.5: PCA plot.

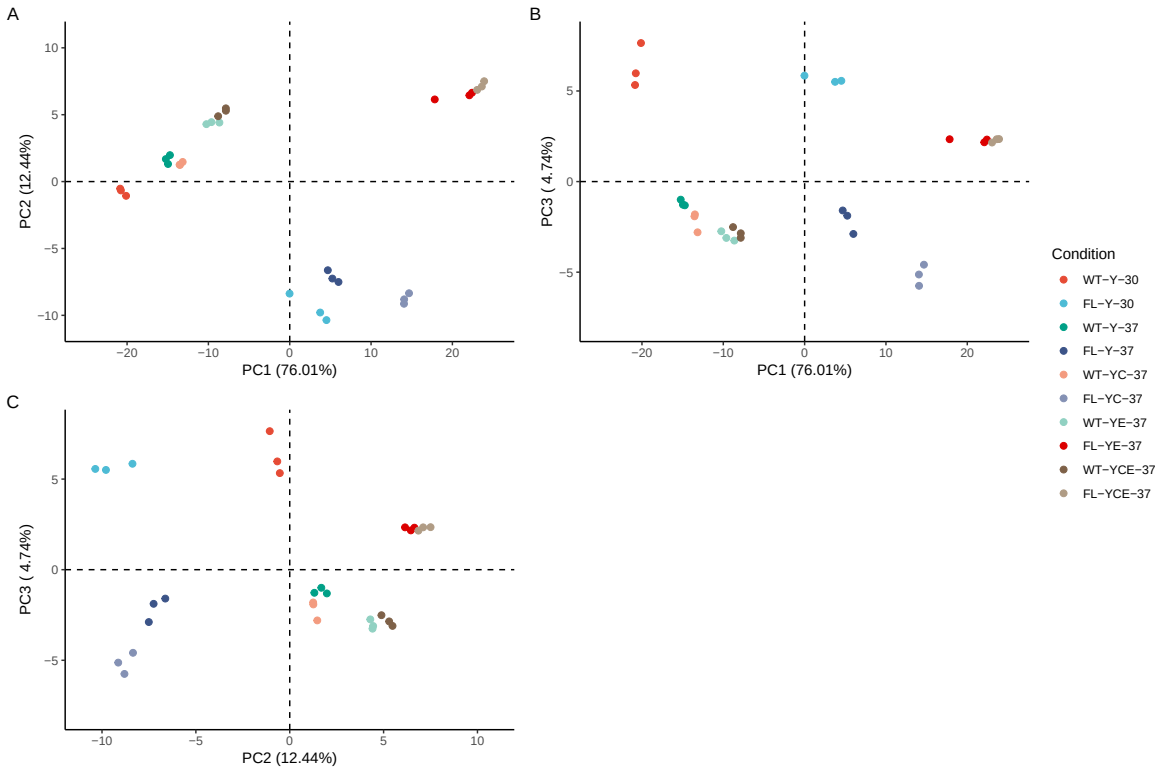

Figure S4.6: PCA plot. displays the samples projected onto principal component axes 1 vs 2 and 2 vs 3, coloured by experimental condition, using the rlog-transformed data

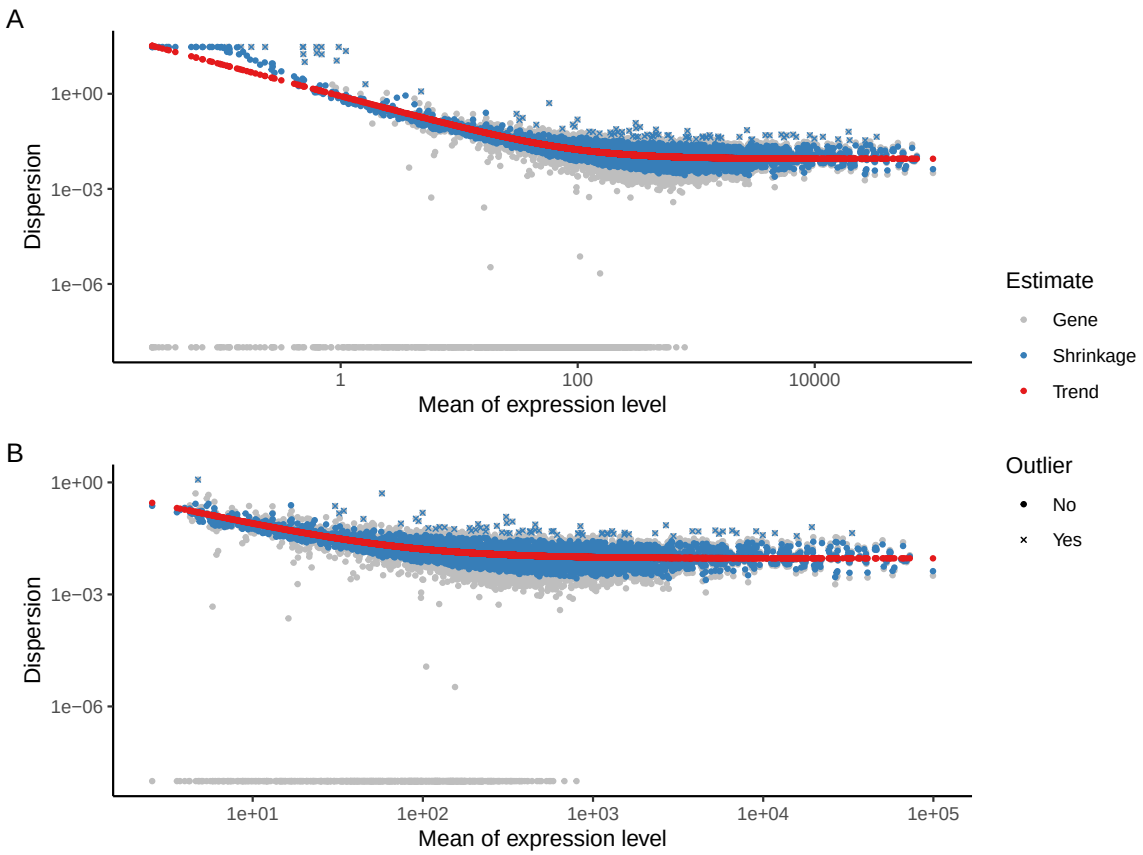

Figure S4.7: Dispersion plot.

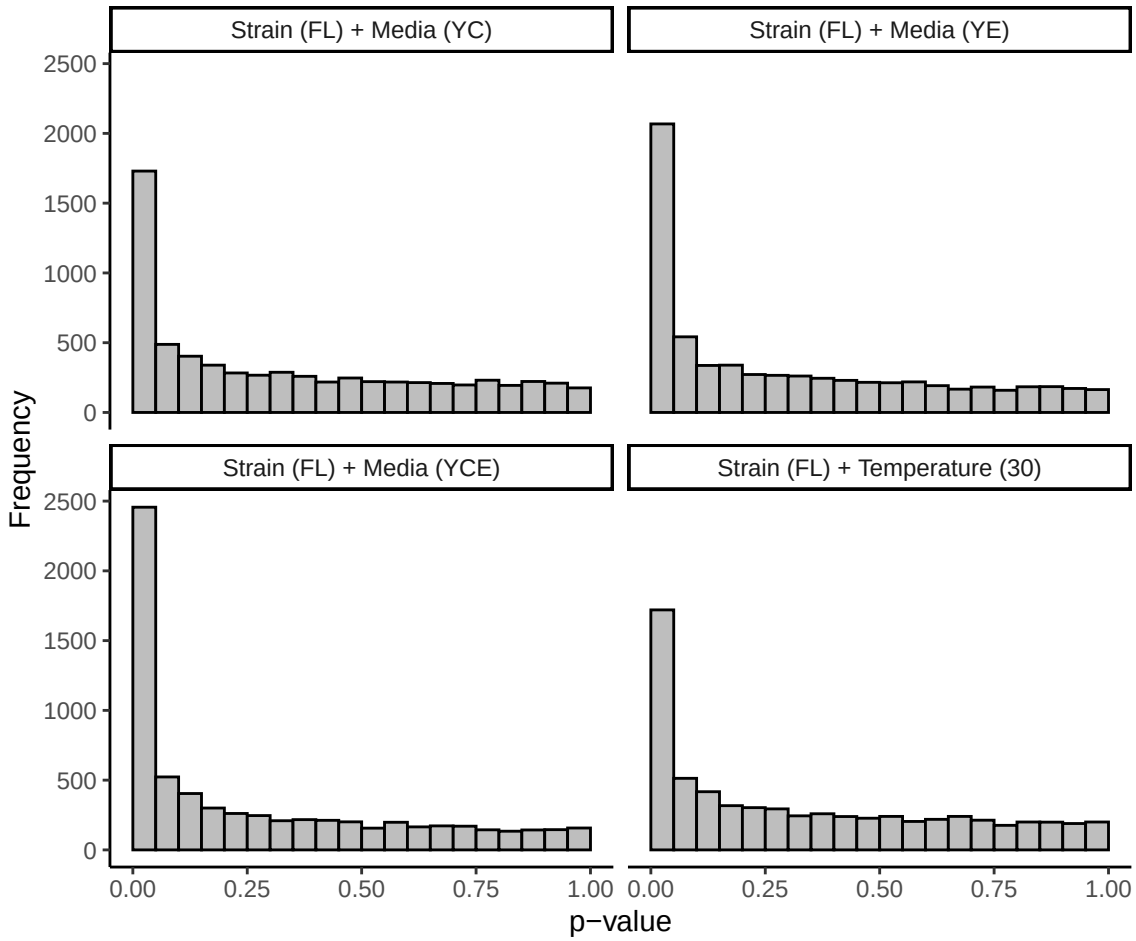

Figure S4.8: p-value histograms.

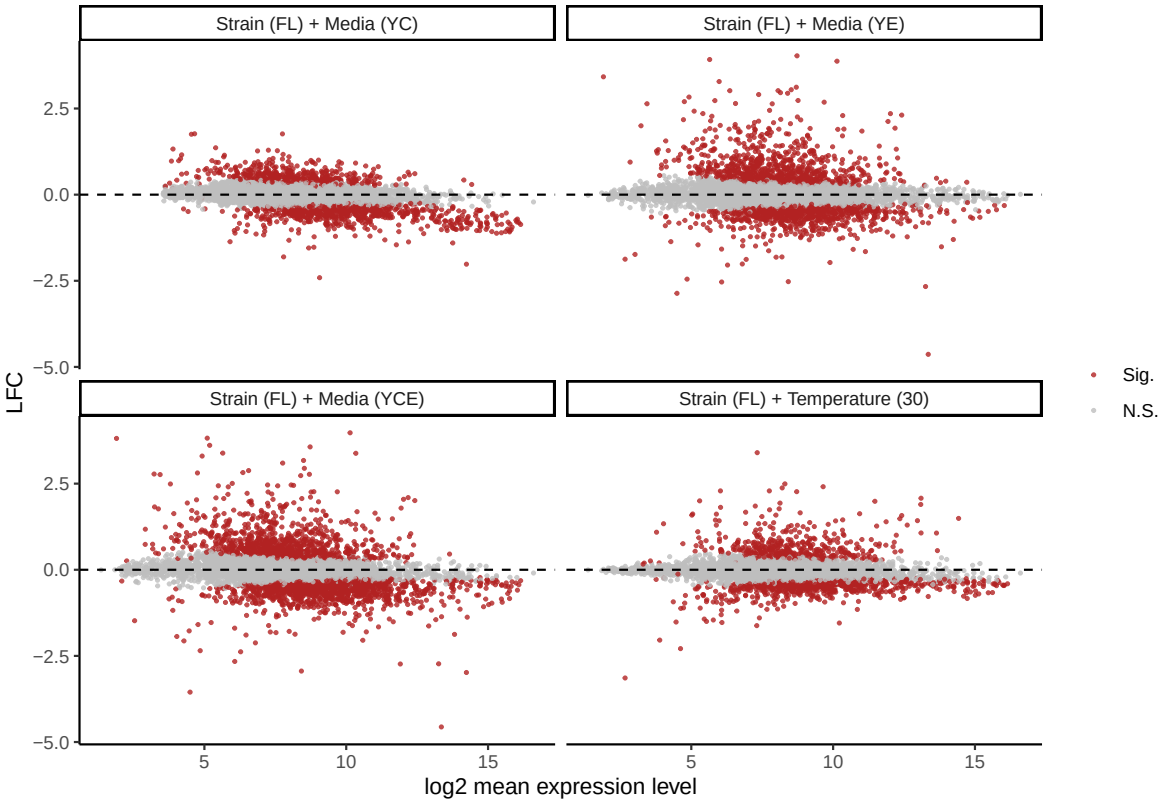

Figure S4.9: MA plot.

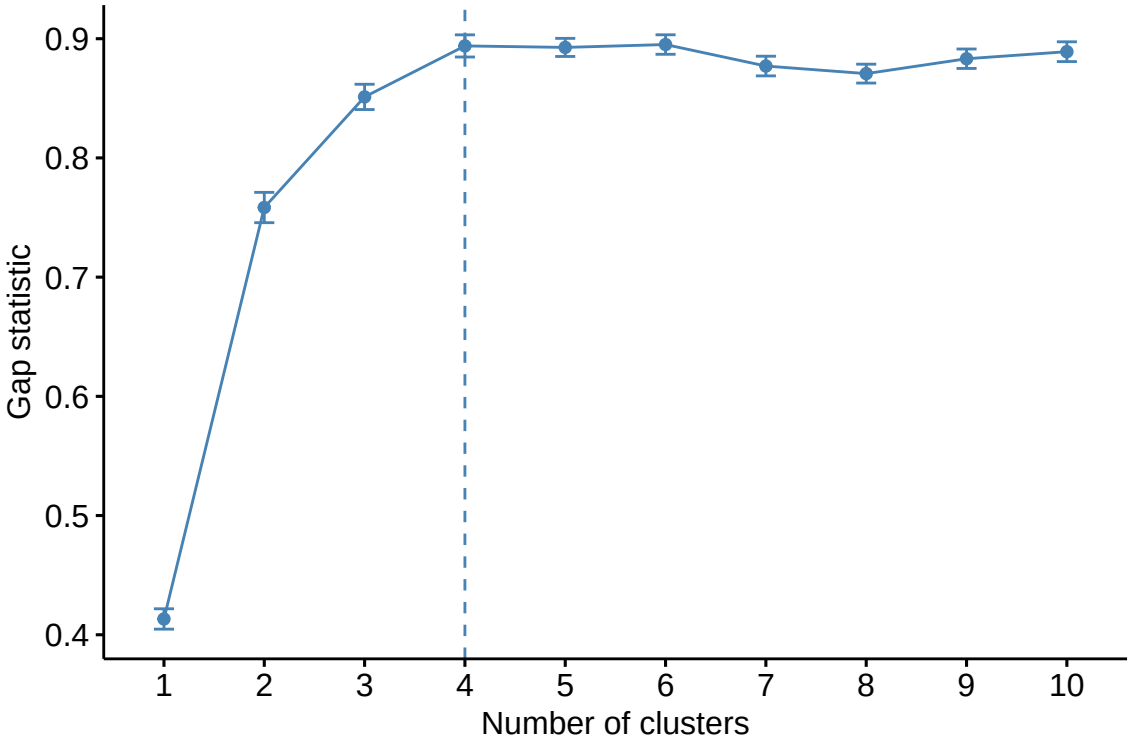

Figure S4.10: Gap plot.