THE UNIVERSITY OF EDINBURGH

SCHOOL OF MATHEMATICS

Integrative Transcriptomic and Motif Analysis Reveals Stress-Specific Regulatory Responses in *Cryptococcus neoformans* upon *FLC1* Deletion

Master of Science in Statistics with Data Science

Author: Xu Guo
Student Number: S2743905
Instructor: Tim Cannings and Serveh Sharifi Far
Date: 2025
Total Words: 4590

Exclusive Summary

This work examines the *Cryptococcus neoformans* transcriptional response to calcium-linked stress and the contribution of the *FLC1* gene. Analysis of 30 RNA-seq samples demonstrated a strong growth defect for the 37°C + CFW + *flc1*Δ condition. Principal component analysis indicated strong genotype-temperature and genotype-media interactions. Differential expression analysis pinpointed 56 genes as specifically dysregulated under the given stress. To investigate regulatory processes, we combined 4-mer promoter motif data with Random Forest and Elastic Net model analysis. Motif-based models classified the downregulated genes with moderate accuracy AUC > 0.7, but the upregulated ones were not as predictable. Future work includes orthogonal experimental designs, pathway enrichment analysis, and better motif modeling with longer k-mers. Overall, this research reveals condition-specific regulatory reactions during fungal stress adaptation and predicts candidate genes and motif features as potentially worthwhile subjects for future research.

Own Work Declaration

I declare that this thesis is entirely my own work, except where otherwise indicated. All sources have been acknowledged, and any help received has been fully credited.

Contents

1. Introduction

1.1 Biological Background

Cryptococcus neoformans is a unicellular pathogenic fungus that poses a serious threat to the health of immunocompromised individuals. It is capable of perceiving and adapting to a wide range of environmental changes both inside and outside the human host. Cell wall integrity and the ability to respond to calcium-based stress are essential survival mechanisms, particularly under elevated temperatures [1].

A key gene involved in these processes is **FLC1**, believed to encode a calcium transporter or channel required for maintaining intracellular calcium homeostasis. Previous studies have shown that deletion of *FLC1* (*flc1*Δ) causes severe growth defects when cells are exposed to high calcium concentrations at 37°C, a temperature mimicking the host environment. These defects are not observed at lower temperatures or when wild-type *FLC1* is reintroduced, indicating a specific link between *FLC1* function and calcium stress response [2].

To further characterize this phenotype, chemical perturbations such as Calcofluor White (CFW), which stresses the cell wall, and calcium chelators like BAPTA or EGTA are employed. By combining these treatments with *FLC1* deletion, researchers can investigate stress-induced changes in gene expression and regulatory mechanisms in a controlled setting. This approach aims to uncover how external stressors affect transcriptional responses and whether intrinsic DNA features—such as promoter motifs—help determine these patterns.

1.2 Experimental Design and Data

Wet Lab Experimental Design The wet-lab experiment aimed to assess how deletion of the *FLC1* gene affects growth and stress response in Cryptococcus neoformans under various environmental pressures. As *FLC1* is involved in calcium homeostasis, its deletion compromises cell wall integrity, particularly under calcium-induced stress.

To investigate this, both wild-type and *flc1*Δ strains were exposed to chemical (CFW, EGTA) and thermal (30°C vs. 37°C) stress. Spot assays revealed that *flc1*Δ mutants exhibited severe growth inhibition specifically under CFW treatment at 37°C—a condition selected for follow-up RNA-seq analysis.

RNA-seq Experimental Design Each sample is uniquely labeled and associated with a specific combination of strain, temperature, and media. However, it is important to note that the design is not fully factorial: media and temperature are not orthogonal — that is, not all media types are tested at both temperatures. Therefore, in downstream statistical modeling, media and temperature must be analyzed

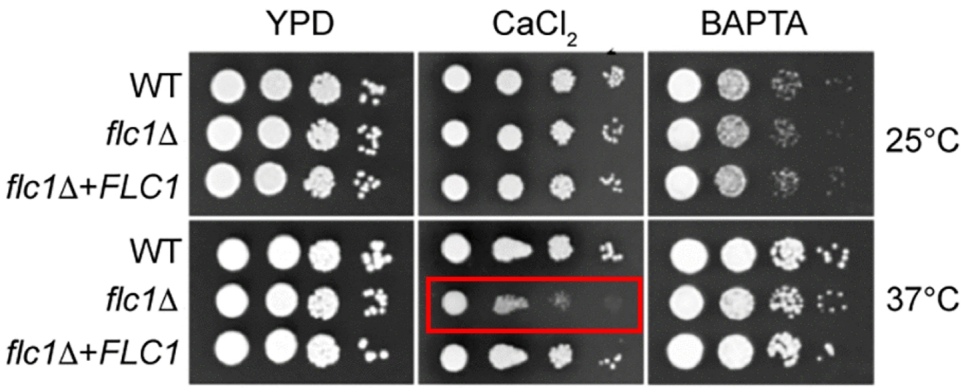

Figure 1.1: Spot assay showing growth of Cryptococcus neoformans strains under different calcium conditions and temperatures. WT, *flc1*Δ, and complemented strains were grown on YPD with $CaCl_2$ or BAPTA at 25°C and 37°C. Serial dilutions (10-fold) were spotted left to right. At 37°C with $CaCl_2$, *flc1*Δ shows impaired growth (red box), which is rescued by FLC1 complementation, highlighting its role in calcium homeostasis under heat stress.

separately.

The *FLC1* deletion, on the other hand, is orthogonal to media and to temperature, so interaction terms can be included between *FLC1* status and media or temperature. This allows one to ask how removal of a gene interacts with specific environmental stresses.

The spot set S16, S17, and S18 corresponds to the *flc1*Δ mutant treated with CFW at 37°C—the very same environment where there was suppressed growth (see Figure 1.1). This is the environment of interest and is a focus for differential expression and motif enrichment analysis.

- **Strains:** wild-type (*FLC1*) and deletion mutant (*flc1*Δ)
- **Temperatures:** 30°C (baseline), 37°C (stress condition)
- **Media types:** standard YPD, YPD + CFW, YPD + EGTA, and YPD + CFW + EGTA
- **Replicates:** 3 biological replicates for each condition

Every RNA-seq sample represents a unique combination of strain, temperature, and media. The design is not fully factorial, however—media and temperature are not orthogonally crossed—so these factors must be analyzed separately in downstream statistical models. In contrast, *FLC1* deletion is orthogonal to both media and temperature, allowing interaction terms to assess how gene deletion modulates stress responses. Notably, samples S16–S18 (*flc1*Δ under CFW at 37°C) correspond to the most severe growth defect and serve as the primary focus for differential expression and motif analysis.

Promoter Motif Data In order to examine DNA sequence-level control of stress-induced gene expression, promoter sequence information were harvested for all annotated genes within the *C. neoformans* genome. More precisely, a region of 500 nucleotides upstream of the transcription start site (TSS) of each gene was taken, as the promoter region where regulatory elements, like binding sites for transcription factors, are most likely to reside.

Representation as a motif within these promoter sequences was quantified with 4-mer counts (i.e., frequencies of all DNA words of length 4). This type of representation preserves both known and unknown regulatory elements within a model-free framework. This creates a high-dimensional matrix of the type gene-by-kmer, with each row a gene (e.g., `CNAG_00001`) and each column a given 4-mer (e.g., `ATGC`, `CGTA`, etc.).

The full dataset includes:

- **Total genes:** 8,338 genes (including all 8,152 genes in the RNA-seq expression matrix)
- **Motif dimension:** 256 possible 4-mers (i.e., 4^4 combinations)
- **File:** `H99_all_genes_promoter_500nt_4mer_counts.tsv`

The promoter region of each gene is scanned (strand-specific), and reverse-complement equivalence is applied, where appropriate, which arises from DNA being a double-stranded molecule. This allows for statistical modeling of gene expression patterns as a function of sequence features. It serves as a foundation for enrichment tests and predictive models to identify cis-regulatory elements of differentially expressed genes.

1.3 Project Objectives

The objective of this work is to explore how the DNA sequence of each gene contributes to its stress-induced expression in *Cryptococcus neoformans*, particularly in the context of *FLC1* knockout and calcium-related stress responses. To achieve this, the project is organized around the following three main goals:

1. **Discover co-expression profiles across conditions.**
   This involves examining how different growth conditions—such as temperature, media composition, and *FLC1* status—affect global gene expression patterns. Dimension reduction techniques such as PCA will be applied for visualization, enabling the evaluation of sample clustering, treatment separability, and potential batch effects.

2. **Identify differentially expressed genes (DEGs) under stress conditions.**
   Genes significantly regulated by *FLC1* deletion, temperature upshift, or chemical treatments (e.g., CFW and EGTA) will be identified using appropriate statistical models, such as DESeq2 under a negative binomial framework. Special emphasis is placed on the condition where the *flc1*Δ strain exhibits a severe growth defect—growth on CFW at 37°C (Figure 1.1).

3. **Determine whether DNA sequence motifs explain stress-induced gene expression.**
   Promoter region k-mer statistics will be used to assess whether specific motifs are enriched among

DEGs. These features will also be tested for their ability to predict gene expression under stress conditions using regression or classification models. The goal is to identify a subset of sequence motifs strongly associated with expression changes, potentially corresponding to functional cis-regulatory elements.

2. Exploratory Analysis of Gene Expression Patterns

2.1 Data Pre-processing and Principal Component Analysis

The first step to address Research Question 1 was to perform quality control and then dimensionality reduction on the bulk RNA-seq data. Our research question entailed an analysis of global patterns of gene expression for all 30 samples to understand the impacts of FLC1 deletion as well as temperature and media settings on transcriptional profiles.

Through examination of NCBI GEO accession GSE292021 the dataset provided both raw count data and sample metadata. A count matrix presents gene-level read data for 8,152 genes distributed among 30 samples. The metadata comprises temperature settings (30°C or 37°C), growth media types (e.g., YPD, YPD + CFW), and strain genotypes (wild-type or *flc1*Δ).

The Variance Stabilizing Transformation (VST) technique of `DESeq2` was used to normalize counts of original reads to compensate for sequencing depth variation and compositional bias. Genes with very low expression levels (i.e., with fewer than 10 total counts among all samples) were removed before normalization.

Several quality control visualizations were generated to assess data integrity and normalization impacts: Log_2-transformed raw counts exhibited balanced distribution among samples; VST-transformed expression revealed improved homoscedasticity suitable for PCA; Total read counts per sample revealed comparable sequencing depth among samples with no obvious outliers (Figure 2.1 A-C).

The research used PCA on VST-normalized data to obtain dimension reduction as well as a visualization of global structure of the datasets. As can be seen from Figure 2.1 D, the first two principal components explained 77% and 12% of the total variance correspondingly. The samples had clear clustering according to *FLC1* status leading to sharp distinction between wild-type groups and *flc1*Δ groups. Elimination of *FLC1* triggers massive global transcriptional changes supporting the ensuing analysis of differently expressed genes.

2.2 Condition-specific and Interaction PCA Analysis

To study individual experimental factors and their interactions on overall gene expression, we performed principal component analysis (PCA) over RNA-seq data partitions stratified for temperature and media. Our study investigated interactions between environmental conditions and genetic factors (FLC1) to understand their joint actions on transcriptome composition.

Figure 2.2 shows PCA projections for six varied

A-B examine specimens cultivated under conventional YPD media conditions. Panel A displays

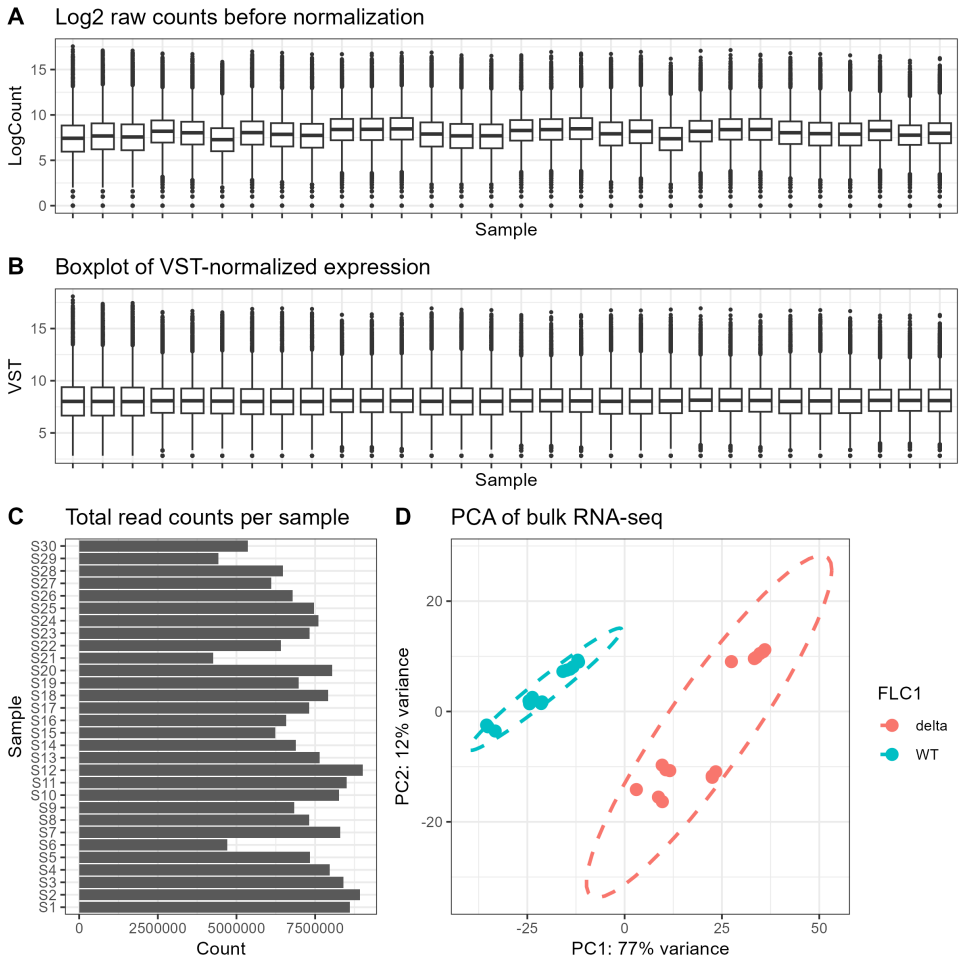

Figure 2.1: Quality control and principal component analysis (PCA) of bulk RNA-seq data.(A) Boxplot showing the log2-transformed raw counts before normalization for all samples.(B) Boxplot showing VST-normalized expression values, indicating improved comparability across samples.(C) Total read counts per sample, demonstrating sequencing depth variation among the 30 samples.(D) PCA plot based on VST-normalized expression, colored by FLC1 genotype (WT vs delta).

samples with *FLC1* status determining their coloration. Wild-type and *flc1*Δ strains exhibit distinct separations along PC1 where 97% variance is explained, indicating FLC1 deletion exerts significant transcriptomic influence under baseline conditions. Panel B displays identical samples marked by temperature variations. Despite the exclusion of all YPD samples at 30°C from sequencing processes, the 37°C group exhibits minor variations.

C-D only analyze those incubation-37°C held samples. FLC1 remains the leading source of variation for heat stress from Panel C's information. Panel D shows media composition with CFW or EGTA leading to sample-level grouping which presents clear-cut expression profiles induced by several chemicals' stresses.

E-F explicitly investigate interaction effects. Panel E shows our analysis of *FLC1*'s interaction with temperature under YPD conditions. Wild-type and mutant samples show clear separation along PC1 which validates the notion that FLC1 status dictates thermal response. Panel F shows an analysis into the *FLC1* interaction dynamics under media treatment conditions at 37°C. Each FLC1–media combination creates unique clusters which indicate that genetic background dictates cellular responses to specific

chemical stresses. The *flc1*Δ + CFW group at 37°C clearly separates from all other samples within two-dimensional PCA space, asserting that such combination shows a unique and strong transcriptome response. **The noted data corresponds to the severe growth defect phenotype observed under these conditions** (Figure 1.1).

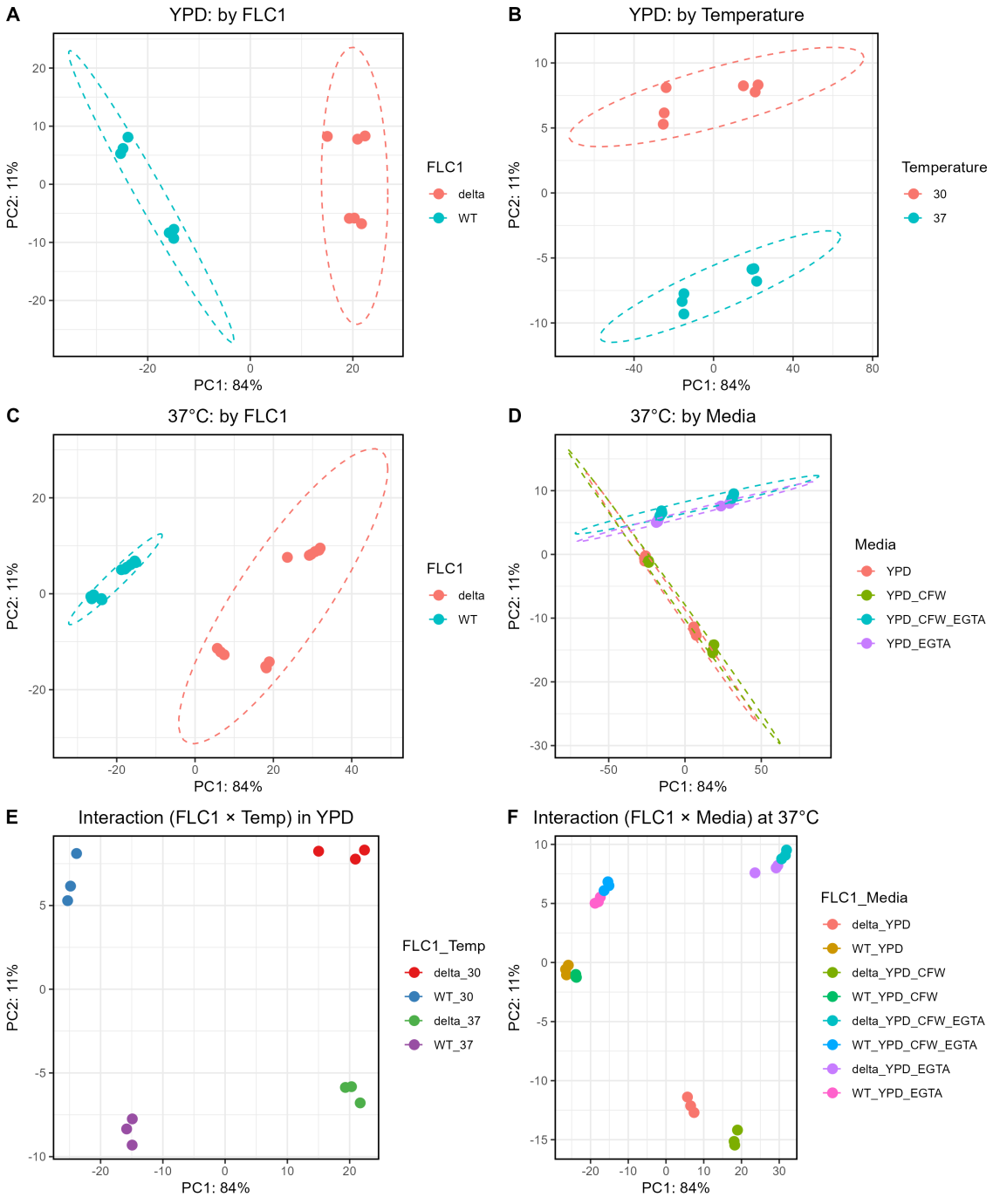

Figure 2.2: Principal component analysis (PCA) of gene expression under various stress conditions. (A) PCA of samples grown in YPD media, colored by *FLC1* genotype (WT vs. *flc1*Δ). (B) PCA of the same YPD samples, colored by temperature (30°C vs. 37°C). (C) PCA of samples incubated at 37°C, colored by *FLC1* genotype. (D) PCA of 37°C samples, colored by media treatment (YPD, YPD+CFW, YPD+EGTA, YPD+CFW+EGTA). (E) PCA of YPD-only samples, showing interaction between *FLC1* status and temperature. (F) PCA of 37°C samples, showing interaction between *FLC1* status and media. Each point represents a sample; dashed ellipses denote 95% confidence intervals. PC1 and PC2 explain 97% and 2% of the variance, respectively.

3. Differential Expression Analysis

3.1 Temperature and *FLC1* Interaction Effects on Gene Expression

To systematically identify genes influenced by temperature, *FLC1* genotype, and their interaction, we constructed a generalized linear model using the DESeq2 framework [3]. The experimental design was not fully factorial (i.e., media and temperature were not orthogonal), so we analyzed the temperature and genotype effects using a subset of 12 samples grown exclusively in YPD media at either 30°C or 37°C. This ensured orthogonality between the factors **Temperature** and **FLC1**, which allowed us to model their interaction properly using the formula:

```
~ Temperature * FLC1
```

We applied shrinkage to $\log_2$ fold changes using the `apeglm` method, and defined differentially expressed genes (DEGs) based on false discovery rate (FDR) < 0.05 and absolute $\log_2$ fold change > 0.5. Results from the Wald test for each model coefficient—temperature effect, *FLC1* effect, and their interaction—were visualized as volcano plots.

To further identify genes whose expression is significantly modulated by interaction alone, we performed a likelihood ratio test (LRT), comparing the full model including the interaction term to a reduced model without it. Genes significant in this LRT (FDR < 0.05) were recorded and compared to the Wald-based DEG sets.

As shown in Figure 3.1, panels A–C present volcano plots of DEGs associated with temperature, genotype (*FLC1* deletion), and their interaction, respectively. Temperature and *FLC1* each influence hundreds of genes, with *FLC1* exerting a particularly strong upregulatory effect. In contrast, the interaction term produces a smaller set of DEGs, but these are well separated in fold change and FDR space.

Panel D shows a Venn diagram summarizing the overlaps among DEGs identified from each effect and the LRT. Notably, there is a core set of 171 genes affected by multiple factors, including interaction-specific regulation. These genes may represent synergistic or conditional effects dependent on both genotype and temperature context.

3.2 Media and *FLC1* Interaction Effects at 37°C

In order to explore how media composition and *FLC1* deletion jointly impact transcriptional response under heat stress, we concentrated on the 18 samples cultured at 37°C under four media regimes: YPD, YPD + CFW, YPD + EGTA, and YPD + CFW + EGTA. We adopted a DESeq2 model with an interaction

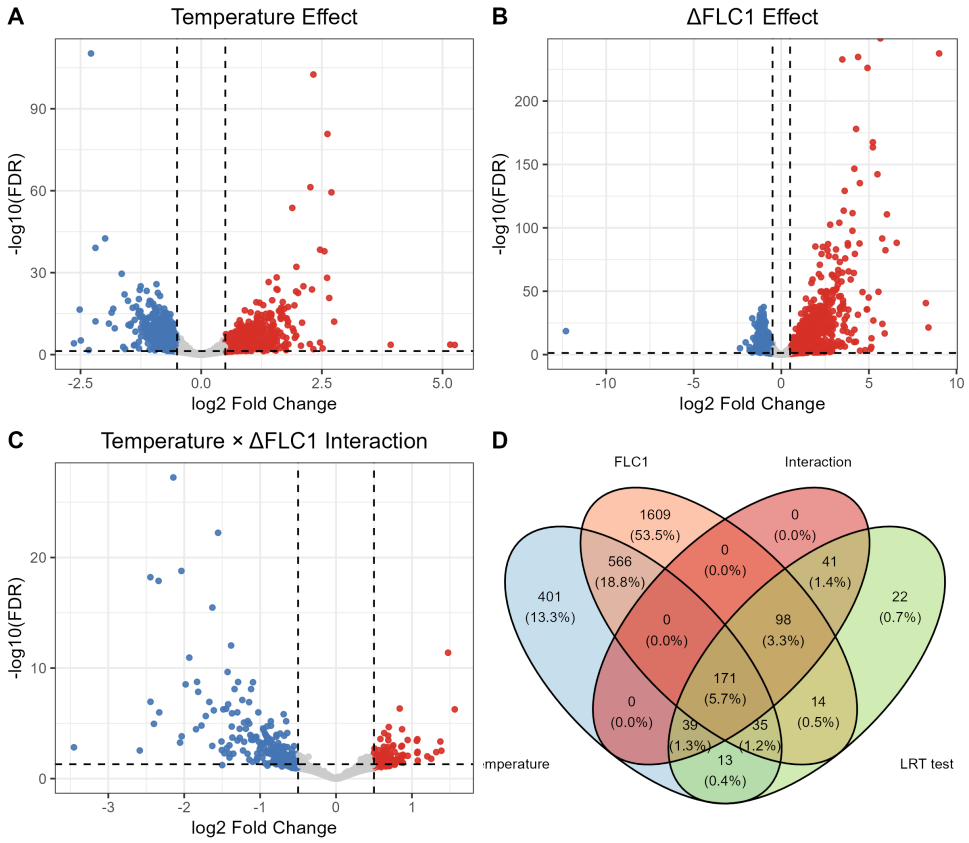

Figure 3.1: Differential expression under temperature and ΔFLC1 effects (Experiment 1). (A–C) Volcano plots showing DEGs under temperature, ΔFLC1 genotype, and their interaction. (D–E) Venn diagrams illustrate overlap of significantly upregulated (D) and downregulated (E) genes across temperature, ΔFLC1, their interaction, and LRT. Differential expression defined as $\log_2$FC > 0.5 or < -0.5 with FDR < 0.05.

term of `Media` × `FLC1` with YPD and wild-type (WT) as the reference levels. Log_2 fold changes were calculated and shrunk according to the `apeglm` method for each main and interaction term.

Figure 3.2 shows the differential expression results. A, D, and G show the reaction to CFW, EGTA, and the simultaneous treatment with CFW + EGTA compared to YPD for wild-type cells. B, E, and H show the corresponding interaction effects with *flc1*Δ, the additional expression change due to deletion of the gene under each challenge. C shows the total main effect of deletion of *FLC1* on all media.

Amongst all combinations, the interaction of *flc1*Δ with CFW (Panel B) resulted in the most pronounced differential gene expression signature, as would be expected given the strong inhibition phenotype of this combination (Figure 1.1). Most of the upregulated genes from this condition are linked to stress response and cell wall remodeling, and hence deletion of *FLC1* would enhance the effects of wall stress caused by CFW.

Figure 3.2 F and I are Venn diagrams illustrating DEGs commonly found among stressor conditions and exclusively detected with the interaction model or likelihood ratio test (LRT). We particularly found that 616 genes were commonly detected among both the three pairwise interaction comparisons and the LRT test, supporting their strong regulation with media–genotype interaction. Also, the *flc1*Δ + CFW + EGTA combination (Panel H) unveiled a characteristic but much reduced transcriptional profile, probably

because EGTA may have partially alleviated the stress.

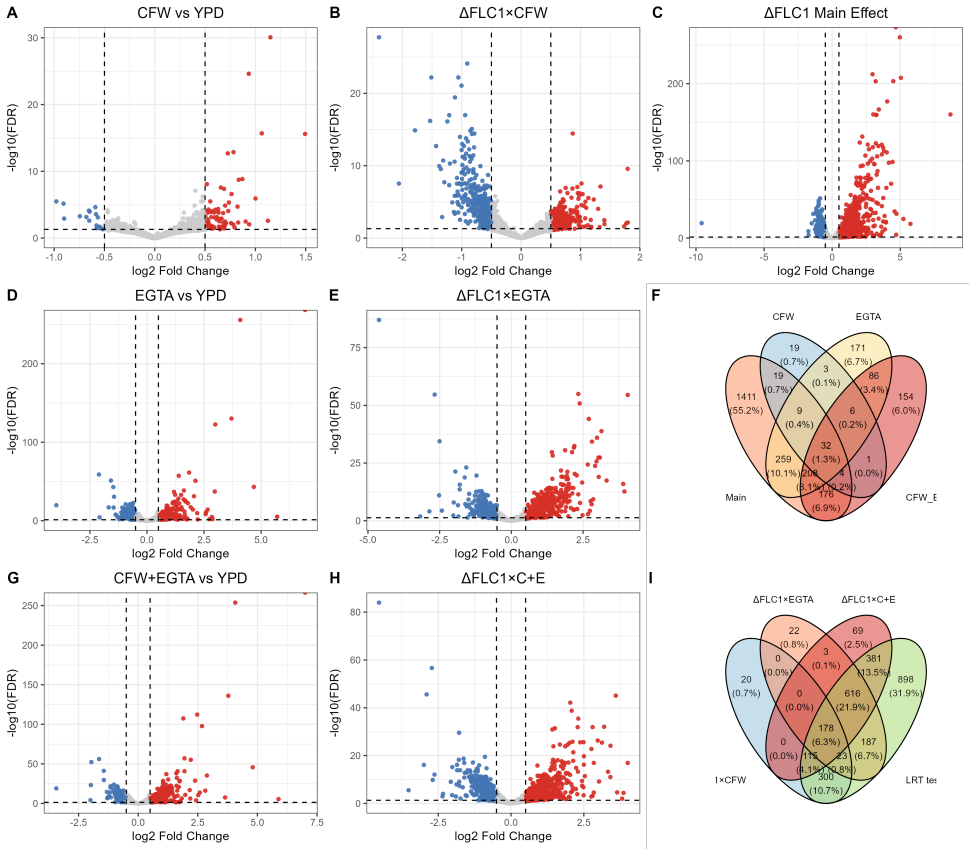

Figure 3.2: Differential expression under media stress and ΔFLC1 interaction (Experiment 2). (A–C) Volcano plots showing DEGs for media contrasts and ΔFLC1 main effect. (D–F) Volcano plots showing interaction terms between ΔFLC1 and media conditions (EGTA, CFW, Combo). (G–H) Venn diagrams display overlap of significantly upregulated (G) and downregulated (H) genes among ΔFLC1, stress conditions, their interactions, and LRT. Genes are filtered with $\log_2$FC > 0.5 or < -0.5 and FDR < 0.05.

3.3 Identification and Visualization of *FLC1* × CFW-specific Response Genes

According to pre-existing phenotypic observations (Figure 1.1), the robust growth defect and strong stress phenotype only resulted from the combined environment of *flc1*Δ, CFW treatment, and 37°C. To identify genes with expression dysregulated specifically for this combination of stresses, we adopted a set-theoretic filter-based method incorporating RNA-seq data from temperature-shift and media-based experiments.

We initially defined high- and low-expression sets of genes independently for each experiment with $\log_2$ fold change and FDR cutoffs. Subsequently, we pooled DEGs from all non-focal conditions (aside from *flc1*Δ × CFW at 37°C). This pool then was intersected with DEGs from the *flc1*Δ × CFW contrast to remove genes responding broadly to many pressures. Lastly, to obtain condition-specific signatures, we took genes uniquely differential in *flc1*Δ × CFW at 37°C by set subtraction (i.e., not present in all other genotype/treatment/interaction contrasts).

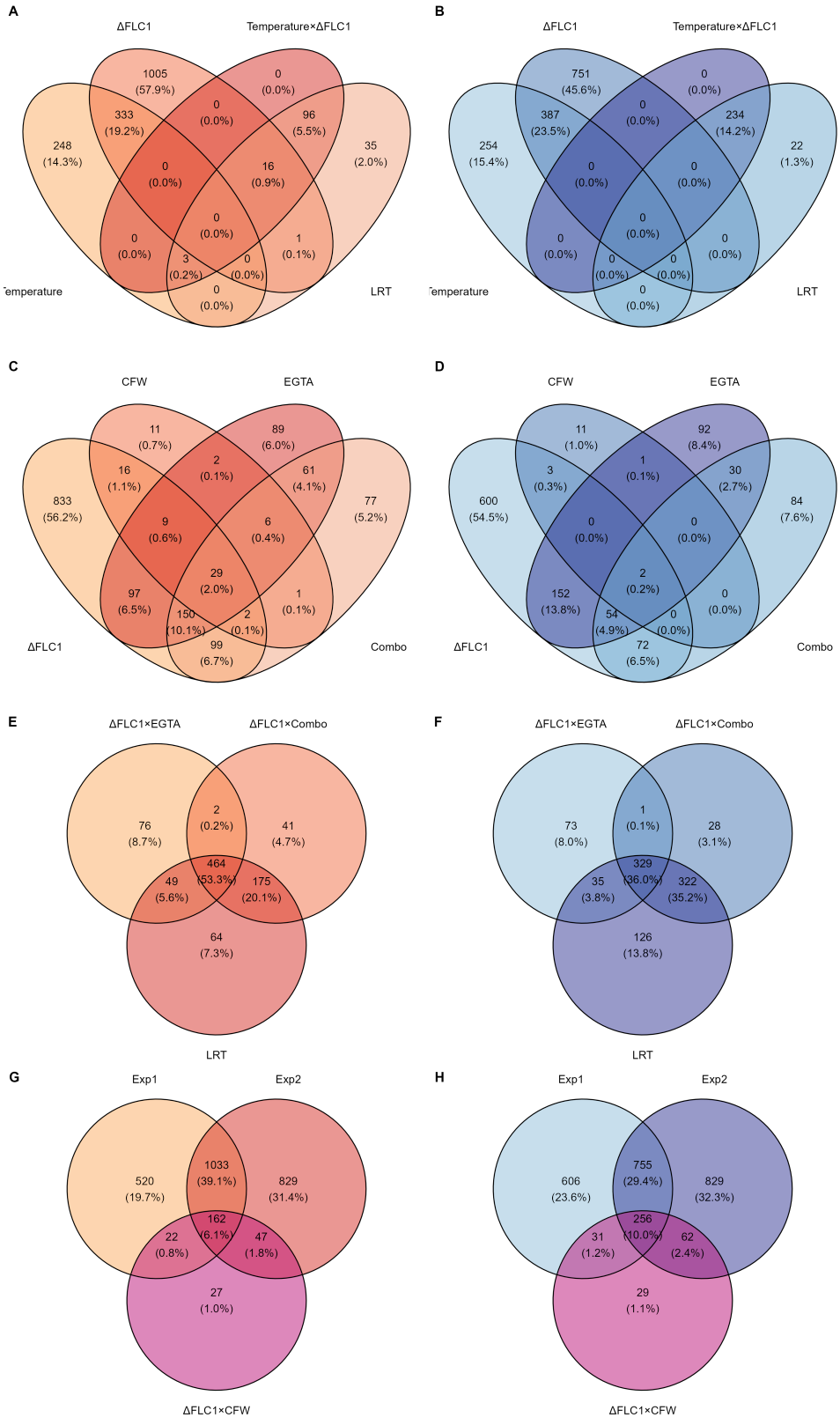

Figure 3.3: Venn diagrams of differentially expressed genes under various conditions. (A–B) Venn diagrams of significantly upregulated (A) and downregulated (B) genes in Experiment 1 (YPD), comparing the effects of temperature, ΔFLC1, their interaction, and LRT. (C–D) Upregulated (C) and downregulated (D) genes in Experiment 2 (37 °C) across ΔFLC1 and individual stress conditions (CFW, EGTA, and Combo). (E–F) Upregulated (E) and downregulated (F) genes identified from ΔFLC1×media interactions and LRT at 37 °C. (G–H) Comparison of DEGs from Experiments 1 and 2 with ΔFLC1×CFW-specific genes; genes uniquely present in this interaction are highlighted in magenta. Genes are considered differentially expressed if $\log_2$FC > 0.5 or < -0.5 with FDR < 0.05.

This stepwise refinement is represented in the following nested Venn diagrams (Figure 3.3). The left and right columns are intersections of high-expression and low-expression DEG sets, respectively. The schema reveals how the original wide DEG pools (top rows; based on LRT, main effects, and interaction terms) are subject to serial filtration. Final selection identifies 27 exclusively upregulated and 29 exclusively downregulated genes within the individual stress environment of *flc1*Δ, CFW, and 37°C.

In order to visualize these top transcripts, we created a heatmap of their expression patterns within all 30 samples (Figure 3.4). Distinct expression signatures are revealed: the 27 upregulated and the 29 downregulated genes exhibit strong induction and suppression, respectively, only within the CFW–*flc1*Δ–37°C samples (S16–S18), and expression levels within all other genotype–media–temp combinations are near baseline. These results indicate that these 56 genes form the molecular core of the fungal stress response and are only elicited when calcium homeostasis (via *FLC1*) is compromised under direct cell wall stress.

This specific set of genes offers mechanistic insights into the divergent transcriptional fates of combined genetic and environmental challenge, and identifies strong candidates for validation of their functions.

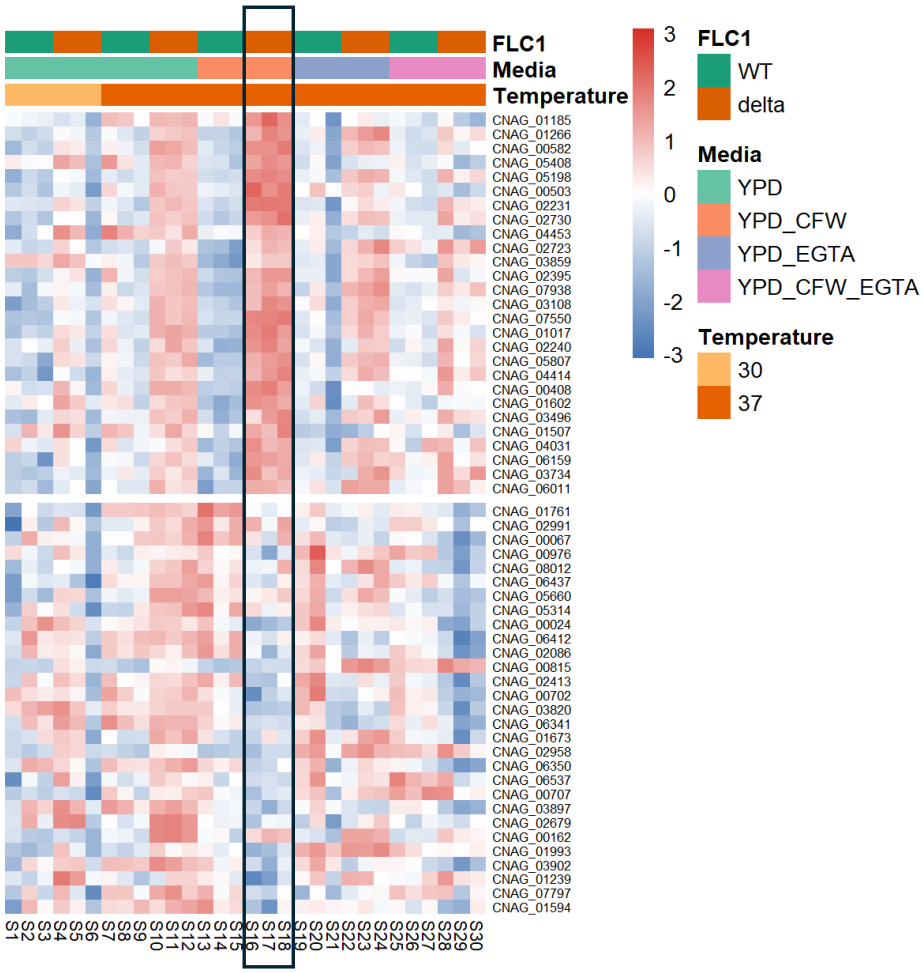

Figure 3.4: Heatmap of ΔFLC1×CFW-specific differentially expressed genes. Expression heatmap of genes uniquely regulated under the ΔFLC1×CFW interaction. Log_2-transformed expression values are shown across 30 samples (S1–S30), with top annotations indicating FLC1 genotype, media condition, and temperature. The sample group marked by the black rectangle represents the ΔFLC1 strain under CFW stress at 37 °C—the focal condition of interest. The upper panel shows 27 genes that are specifically upregulated in this group but not in others. The lower panel shows 29 genes that are specifically downregulated in this group, while showing high or undetectable expression elsewhere. All genes pass the significance threshold of $\log_2FC > 0.5$ or < -0.5 and FDR < 0.05.

4. Motif Analysis of Condition-specific Differentially Expressed Genes

4.1 Identification of Motifs Associated with *flc1*Δ + CFW-specific Expression

Trying to discover promoter sequence features indicating variation of gene expression under the *flc1*Δ + CFW + 37°C condition, we performed two-stage motif selection with Random Forest and penalized logistic regression. We did the analysis separately for genes specifically upregulated and specifically downregulated under this key condition, as identified in Chapter 3.

We started with 128 4-mer motif counts from the 500 bp upstream promoter regions of all genes. Due to reverse complement redundancy, each motif was combined with its reverse complement (e.g., AAAA and TTTT). Genes were classified as differentially expressed or not according to the *flc1*Δ × CFW comparison (Padj ¡ 0.05, $|\log_2 \mathrm{FC}| > 0.5$). A Random Forest classifier [4] was trained to rank motifs according to their importance and retained those above predefined thresholds for both Mean Decrease Accuracy and Gini index.

The selected motifs were then input into an Elastic Net logistic regression model ($\alpha = 0.5$), implemented using the `glmnet` framework [5]. Motif counts were standardized, and class imbalance was addressed by assigning weights to DEG vs. non-DEG classes. Final motif selection was performed at the $\lambda_{1\mathrm{se}}$ value identified through 10-fold cross-validation.

Separate modeling of high and low expression DEGs aimed to capture specific regulatory logic of activation and repression. This did improve interpretability and retained direction-specific signal during motif selection.

Figure 4.1 compiles analysis results for downregulated genes. Panel A presents the Random Forest importance plot, with indicated motifs chosen with joint MDA and Gini thresholding. Panel B illustrates the paths of the LASSO coefficient on $\log(\lambda)$ values, and Panel C displays cross-validation deviance and the $\lambda_{\min}$ and $\lambda_{1\mathrm{se}}$ cutpoints. Panel D ranks the chosen motifs on log odds ratio. Most of the chosen motifs (e.g., TAAA, GTGA) are AT-rich, as would be consistent with possible enrichment of Crz1-like calcium-response elements [6, 7].

A similar analysis was repeated for the set of genes specifically upregulated under the same condition (Figure 4.2). Again, important motifs were identified using both Random Forest and LASSO, but the selected motifs differ substantially from those associated with downregulation. Notably, several GC-rich 4-mers (e.g., CGCC, GCCA) were prominent among upregulated genes, suggesting different regulatory architectures between induced and suppressed genes.

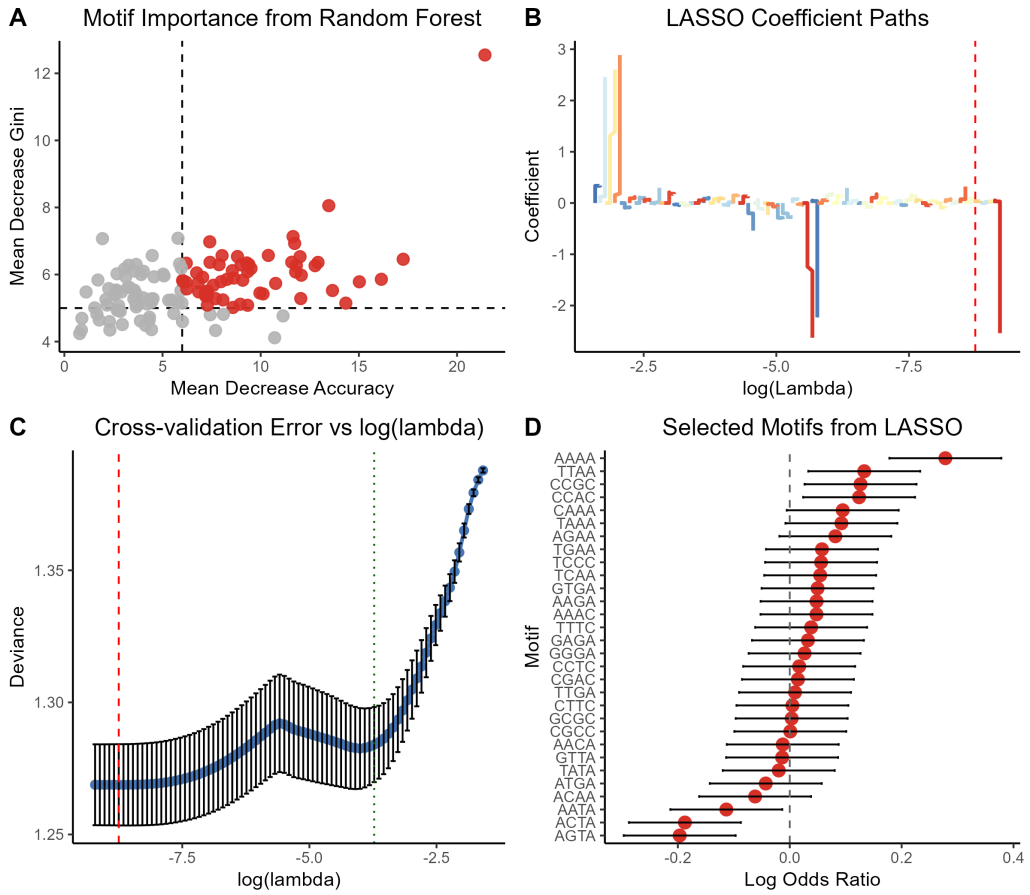

Figure 4.1: Motif selection analysis for downregulated DEGs under *flc1*Δ + CFW + 37°C condition. (A) Random Forest importance plot. (B) LASSO coefficient paths. (C) Cross-validation deviance vs. log(λ). (D) Selected motifs from LASSO ranked by log odds ratio.

Together, this analysis demonstrates basic promoter DNA sequence features predict expression response to specific disease. These features can specify possible cis-regulatory regions, which are part of stress-responsive transcription, and are hypothesis-generating towards future experimental validation.

4.2 Motif-based Classification Performance Evaluation

In order to assess the power of motif features to distinguish differentially expressed genes (DEGs) under the *flc1*Δ + CFW + 37°C condition, Receiver Operating Characteristic (ROC) curves were built on the basis of predicted probabilities from LASSO regression models trained on separate sets of high-expression and low-expression DEG sets.

For each model, we used 10-fold cross-validation to generate fitted probabilities on held-out folds, and then plotted the true positive rate versus the false positive rate across thresholds. The area under the curve (AUC) was computed to quantify classification performance, where a value of 1.0 represents perfect classification and 0.5 represents random chance.

As shown in Figure 4.3, Panel A displays the ROC curve for the model trained to distinguish high-expression DEGs. The model achieved an AUC of 0.664, indicating modest predictive power using motif

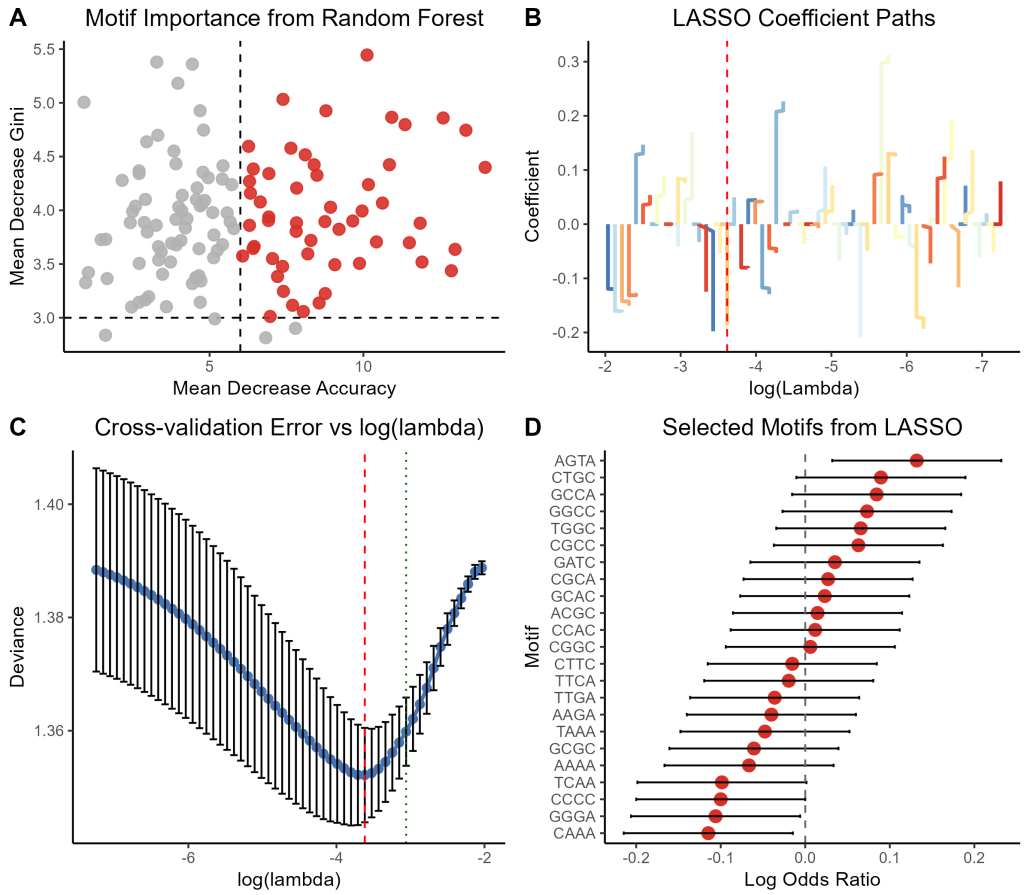

Figure 4.2: Motif selection analysis for upregulated DEGs under *flc1*Δ + CFW + 37°C condition. (A) Random Forest importance scores for all 4-mer motifs. Red points indicate motifs surpassing thresholds for both Mean Decrease Accuracy and Mean Decrease Gini. (B) LASSO coefficient trajectories across log(λ), showing how motif weights evolve with increasing regularization. (C) Cross-validation error curve with error bars, highlighting λ_{min} (red dashed line) and λ_{1se} (green dotted line). (D) Final selected motifs at λ_{1se}, ranked by their estimated log odds ratios with 95% confidence intervals.

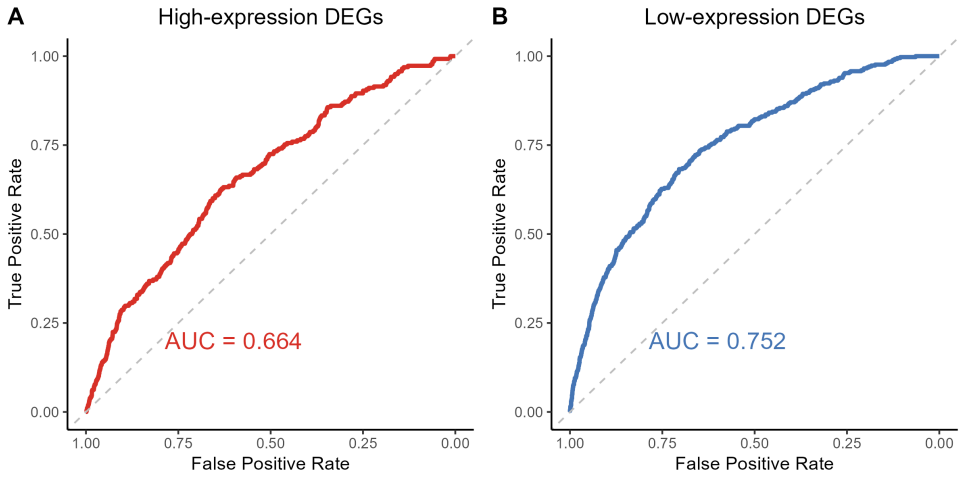

Figure 4.3: ROC curves for motif-based classification of DEGs. (A) ROC curve for high-expression DEGs under *flc1*Δ + CFW + 37°C condition (AUC = 0.664). (B) ROC curve of low-expression DEGs under the same condition (AUC = 0.752). Classifiers were trained with motif features based on cross-validated LASSO regression.

features alone. Panel B shows the ROC curve for the low-expression DEG model, which performed better with an AUC of 0.752. This result suggests that promoter sequence motifs are more predictive of

downregulation events than upregulation under this stress condition.

Taken together, these results support the utility of short promoter motifs as informative features for gene regulation prediction, although additional regulatory elements (e.g., longer motifs, chromatin context) may be required to improve predictive accuracy, especially for upregulated genes.

5. Discussion

5.1 Limitations

Limitations in experimental design. The RNA-seq design employed in this study is not fully factorial, as media and temperature were not orthogonally crossed. This limits the ability to model complete three-way interactions among *FLC1* genotype, temperature, and chemical treatment (e.g., CFW). Specifically, we cannot formally assess the interaction among all three variables (e.g., 37°C × CFW × *flc1*Δ), which is biologically the most relevant condition. In addition, the small sample size (three biological replicates per group) introduces uncertainty in parameter estimation, reducing the statistical power to detect subtle or higher-order effects.

Limitations in DEG identification strategy. Our differential expression analysis focused on identifying genes specifically dysregulated under the 37°C + CFW + *flc1*Δ condition to pinpoint causal genes linked to stress-induced growth inhibition. However, this targeted approach may overlook genes involved in combinatorial regulation. For instance, growth defects might arise from the joint effects of gene A responding to temperature and gene B to genotype. Because the filtering strategy selects only genes uniquely significant in the focal condition, it may exclude such contributors and obscure interaction-based mechanisms.

Limitations in motif-based prediction. Although the motif analysis revealed informative patterns, model performance varied between gene classes. The classifier for downregulated genes achieved acceptable accuracy (AUC > 0.7), while the model for upregulated genes performed worse (AUC < 0.7), indicating weaker motif signal. This may be due to the use of short motifs (4-mers) or the absence of broader regulatory context. Incorporating longer motifs (e.g., 5-mers), known transcription factor binding sites, or chromatin accessibility data could enhance model performance and reveal more robust regulatory features.

5.2 Conclusion

In this work, we rigorously examined the transcriptional response to calcium-linked stress in *Cryptococcus neoformans* using RNA-seq and promoter motif data, centered on the *FLC1* gene. PCA analysis demonstrated robust interaction effects of *FLC1* deletion with both temperature and media manipulations, supporting strongly genotype-specific transcriptional modulation by environmental context. We also determined that *FLC1* deletion with 37°C and CFW treatment causes a specific transcriptomic signature, corroborating the measured inhibition of growth within the wet-lab phenotype assays.

By differential expression and intersection analysis, we identified 56 signature genes (27 upregulated, 29 downregulated) specifically responsive to the *flc1*Δ + CFW + 37°C, but not other, treatment. Motif-based classification revealed various 4-mer promoter elements predictive of decreased gene expression, with moderate accuracy (AUC > 0.7). These conclusions suggest a condition-specific transcriptional program and imply promoter sequence features are an aspect of the mechanism of regulation of stress-induced gene expression, and provide the basis for future exploration of fungal adaptation and pathogenesis.

5.3 Overlook

This study lays the groundwork for a number of future directions. Experimentation with RNA-seq could be enhanced with a full orthogonal setup to more thoroughly investigate three-way interactions among temperature, media, and *FLC1* genotype, and to allow for clearer interpretation of regulatory interactions. Further, motif enrichment analysis under various environmental conditions could provide context-dependent promoter use and transcription factor activity. The here-identified differentially expressed genes could also be subject to pathway enrichment analysis to determine biological mechanisms of stress-induced inhibition of growth, and the most promising candidates could be prioritized for experimental verification. Finally, the motif-based classification framework could be improved with Bayesian models to deal better with uncertainty, or with longer k-mers (e.g., 5-mers) to increase the specificity of detection of regulatory motif.

References

[1] Robin C May, Nicholas R H Stone, Daniela L Wiesner, Tihana Bicanic, and Kirsten Nielsen. Cryptococcus: from environmental saprophyte to global pathogen. *Nature Reviews Microbiology*, 14(2):106–117, 2016.

[2] Paul R Stempinski, Matthew H Norris, Yilin Wang, et al. The cryptococcus neoformans flc1 homologue controls calcium homeostasis and confers fungal pathogenicity in the infected hosts. *mBio*, 13(6):e02253–22, 2022.

[3] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[6] Timothy L Bailey, Jody Johnson, Charles E Grant, and William Stafford Noble. The meme suite. *Nucleic Acids Research*, 43(W1):W39–W49, 2015.

[7] Jie Cheng, Kristina C Maier, Ziga Avsec, Yael Berenstein, Yarden Lubling, Arnau Sebé-Pedrós, and Shalev Itzkovitz. Cis-regulatory elements explain most of the mrna stability variation across genes in yeast. *RNA*, 23(11):1648–1659, 2017.

[8] Alina Zhu, Joseph G Ibrahim, and Michael I Love. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, 35(12):2084–2092, 2019.

Appendix

A Method

A.1 Differential Gene Expression Analysis Using DESeq2

To determine the genes that are differently expressed among environmental and genetic perturbations, we carried out two individual differential expression analyses with the DESeq2 package [3]. DESeq2 fits the count data with a negative binomial generalized linear model (NB-GLM) while compensating for library size and biological variation.

Statistical Framework. Let K_{ij} represent the raw read count for gene i in sample j. DESeq2 assumes:

$$K_{ij} \sim \mathrm{NB}(\mu_{ij}, \alpha_i), \quad \mu_{ij} = s_j \cdot q_{ij},$$

where μ_{ij} is the expected count, s_j is the sample-specific size factor, and α_i is a gene-specific dispersion. The logarithm of the true expression q_{ij} is modeled linearly:

$$\log(q_{ij}) = X_j \beta_i,$$

where X_j is the covariate vector for sample j, and β_i is the gene-specific coefficient vector.

Model 1: Temperature and Genotype Interaction (Experiment 1). To investigate the effect of temperature (30°C vs 37°C), genotype (WT vs Δ*FLC1*), and their interaction, we modeled the design:

$$\texttt{~Temperature + FLC1 + Temperature:FLC1}.$$

The reference levels were set as 30°C for temperature and WT for strain. Shrinkage of $\log_2$ fold changes was applied using the `apeglm` method.

Model 2: Media Condition and Genotype Interaction (Experiment 2). In a separate experiment, cells were exposed to chemical stressors: calcofluor white (CFW), EGTA, and their combination. To assess main and interaction effects, we modeled:

$$\texttt{~Media + FLC1 + Media:FLC1},$$

with "YPD" and "WT" as reference levels for `Media` and `FLC1`, respectively.

Thresholding and Filtering. In both models, genes with low counts (total across all samples < 10) were removed. Significant DEGs were identified based on:

$$|\log_2 \text{Fold Change}| > 0.5 \quad \text{and} \quad \text{FDR-adjusted } p < 0.05,$$

using the Benjamini–Hochberg method for multiple testing correction.

Effect Size Shrinkage To enhance the stability and interpretability of $\log_2$ fold change (LFC) estimation, particularly for low-count or highly dispersed genes, we utilized the `lfcShrink` function [8] with the adaptive prior-based method `apeglm`. The use of this technique dampens the noise within fold change estimation but maintains ranking and direction, lending stability to downstream visualization and selection of differentially expressed genes.

Interaction Significance Testing. To evaluate the importance of interaction terms, we added to coefficient-based Wald tests a likelihood ratio test (LRT). For this, we compared, in particular, nested models with and without interaction terms:

$$H_0 : \text{Design} = \texttt{main effects only} \quad \text{vs.} \quad H_1 : \text{Design} = \texttt{main effects + interaction}.$$

The LRT determines whether adding the interaction makes a substantive difference to model fit, and provides an orthogonal test to significance testing for coefficients. This two-pronged approach guarantees statistical robustness and biological interpretability.

A.2 Motif Selection via Random Forest

We used the Random Forest method [4] for feature selection for identifying DNA motifs associated with differential gene expression. Random Forest, an ensemble method based on bootstrap aggregation and decision trees, is well-suited for high-dimensional inputs like motif occurrence matrices. For a training set $\mathcal{D} = (x_i, y_i)i = 1^n$, where $x_i \in \mathbb{R}^p$ is the motif count vector for gene i and $y_i \in 0, 1$ is an indicator of whether gene i is differentially expressed, B trees $T_b b = 1^B$ are produced from bootstrap samples. At each internal node, all features from any random subset of size m_{try} are considered for splitting and are used for robust selection of informative sequence patterns.

The final prediction is obtained by majority vote across all trees:

$$\hat{f}_{\text{RF}}(x) = \text{majority vote } \{T_b(x)\}_{b=1}^{B}.$$

Feature importance is evaluated using two metrics:

- **Mean Decrease in Accuracy (MDA):** This metric quantifies the reduction in classification accuracy

after permuting a given feature. Let $\text{Acc}_b^{\text{orig}}$ and $\text{Acc}_b^{\text{perm}(j)}$ denote the accuracy of tree b before and after permuting feature j, respectively. Then the importance of feature j is:

$$\text{MDA}_j = \frac{1}{B}\sum_{b=1}^{B}\left(\text{Acc}_b^{\text{orig}} - \text{Acc}_b^{\text{perm}(j)}\right).$$

- **Mean Decrease in Gini Index (MDG):** For each feature, we compute the total reduction in Gini impurity across all nodes where the feature is used for splitting:

$$\text{MDG}_j = \sum_{t \in T_j} p(t) \cdot \Delta\text{Gini}(t),$$

where T_j is the set of nodes where feature j is used, $p(t)$ is the proportion of samples at node t, and $\Delta\text{Gini}(t)$ is the impurity decrease due to the split.

To select motif features, we ranked all motifs based on both MDA and MDG scores and applied empirical thresholds to retain only those motifs that were consistently informative across both metrics.

A.3 Regularized Logistic Regression

In order to find motif patterns with predictive power for differential gene expression, we carried out penalized logistic regression with the package `glmnet` [5]. The response variable y is a binary variable representing whether a gene is a DEG or not, and the predictor X is motif occurrence counts from upstream promoter sequences. We employed a subset of pre-selected motifs using Random Forest analysis, hereafter referred to as X_{topN}.

Denote $X \in \mathbb{R}^{n\times p}$ as the scaled motif count matrix for n genes and p chosen motifs, and $y \in \{0,1\}^n$ as the DEG labels. We adopt a regularized logistic regression model:

$$\min_{\beta_0,\boldsymbol{\beta}}\left\{-\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{p}_i + (1-y_i)\log(1-\hat{p}_i)\right] + \lambda\left(\alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2\right)\right\},$$

where $\hat{p}_i = \text{logit}^{-1}(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})$, λ is the regularization parameter, and α controls the mixing between LASSO (L_1) and Ridge (L_2) penalties.

We set $\alpha = 0.5$ to apply elastic net regularization, balancing LASSO's sparsity with Ridge's stability—especially important in motif analysis, where k-mer features are often correlated due to sequence redundancy or reverse complementarity. Using 10-fold cross-validation, we selected the regularization parameter λ that minimized binomial deviance. Motifs with non-zero coefficients at λ_{min} were retained as potential predictors of DEG status, interpreted as enriched or depleted motifs and considered biologically meaningful signatures.

B Matericals

B.1 Data Sources

Gene expression data: RNA-seq was performed on 30 samples of *Cryptococcus neoformans*, covering combinations of temperature (30°C or 37°C), *FLC1* genotype (wild-type or *flc1*Δ), and media treatments (YPD, YPD+CFW, YPD+EGTA, YPD+CFW+EGTA), with three biological replicates per condition. Raw count matrices are provided in `GSE292021_counts_quantseq_CryptoRNAseqFLC1.txt`, and sample metadata in `GSE292021_CryptoRNAseqFLC1SampleKey_tabseparated.txt`.

Promoter motif data: Promoter sequences (500 bp upstream) were processed to compute 4-mer motif counts per gene. Reverse-complement motifs were merged to reduce redundancy. The resulting motif matrix (`H99_all_genes_promoter_500nt_4mer_counts.tsv`) includes 128 unique 4-mers for 8338 annotated genes.

B.2 R Packages

[1] DESeq2	1.38.3	[2] apeglm	1.20.0
[3] ggplot2	3.4.0	[4] ggpubr	0.6.0
[5] ggvenn	0.1.9	[6] randomForest	4.7-1.1
[7] glmnet	4.1-6	[8] dplyr	1.0.10
[9] tidyr	1.2.1	[10] reshape2	1.4.4
[11] RColorBrewer	1.1-3	[12] pheatmap	1.0.12
[13] pROC	1.18.0	[14] gtools	3.9.4
[15] tibble	3.1.8		