



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

Author: Xu Guo Instructors: Tim Cannings and Serveh Sharifi Far
Station D, 13:40, School of Mathematics, University of Edinburgh, EH9 3JL, UK

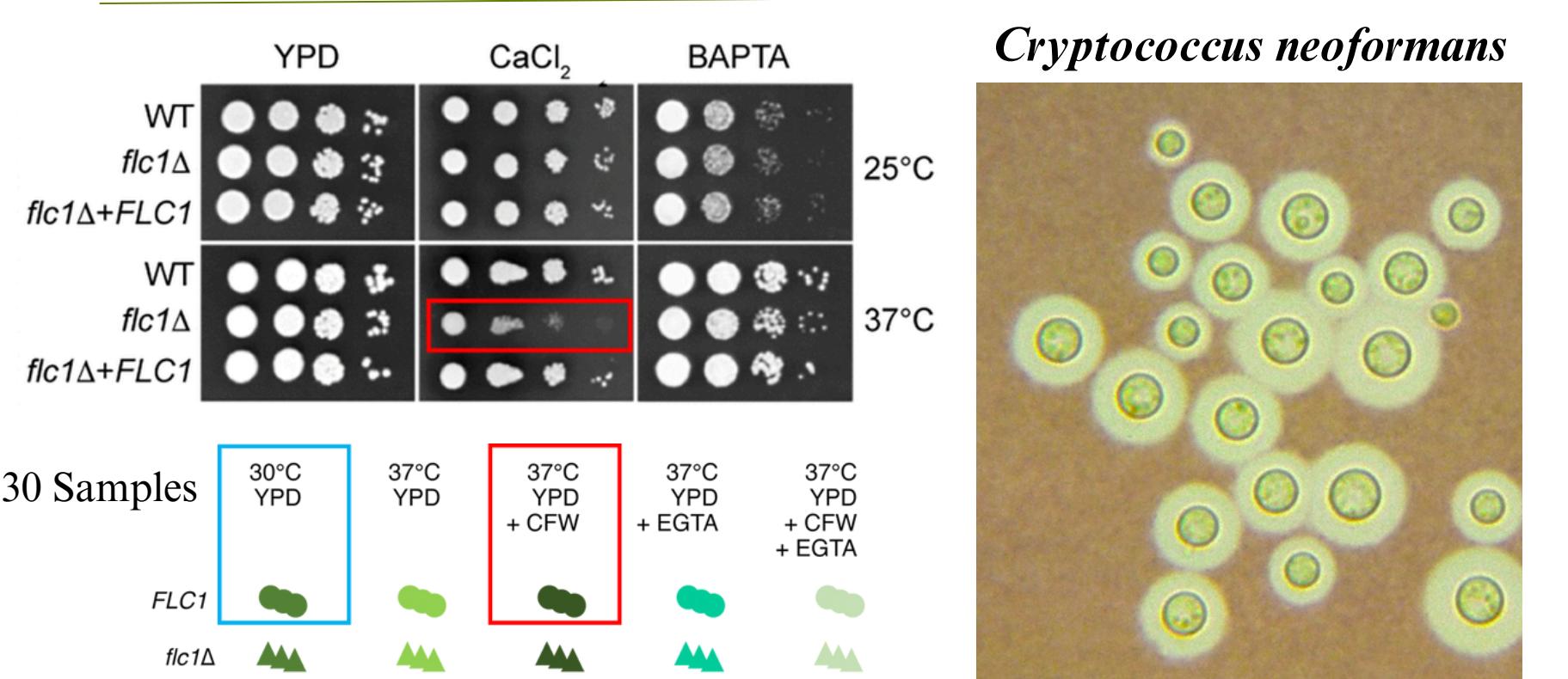
Introduction

Cryptococcus neoformans is a pathogenic yeast that adapts to environmental stress, especially in the human host.

The *FLC1* gene is crucial for calcium stability and cell wall integrity, and its deletion causes growth defects under **high calcium at 37°C**.

Researchers use chemical stressors like Calcofluor White (**CWF**) and calcium chelators to study gene expression changes and regulatory mechanisms under stress.

Which genes inhibit the growth of the fungi under conditions of 37°C+Δ*FLC1*+Stress?



Data Description

RNA-seq is a high-throughput technique used to measure genome-wide gene expression by sequencing RNA transcripts.

Read Count data show a right-skewed, overdispersed distribution with many zeros, often modeled using Negative Binomial (NB) distribution.

Motif is a short and recurring DNA sequence which is likely to be a binding site of **transcription factors**.

4-mer counts matrix records the occurrence counts of each 4-mer motif. There are **256** different 4-mer motif combinations, such as AAAA, ATCG, etc.

Objective

Discover co-expression profiles

Use PCA to examine differences of global gene expression patterns across conditions

Identify differentially expressed genes

Perform differential analysis on count data via DESeq2 under a NB framework

Predict gene regulation by DNA motifs

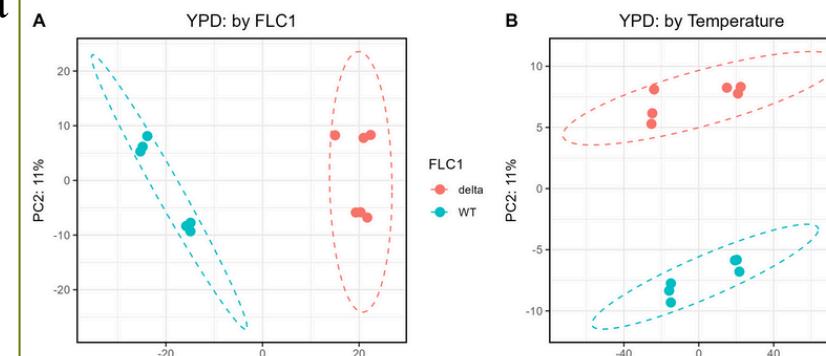
Apply Random Forest and Lasso regression to select predictors from the 4-mer motif–gene matrix.

Design 1: control media

A: **Δ*FLC1***

B: Temperature

E: **Δ*FLC1*:Temperature**



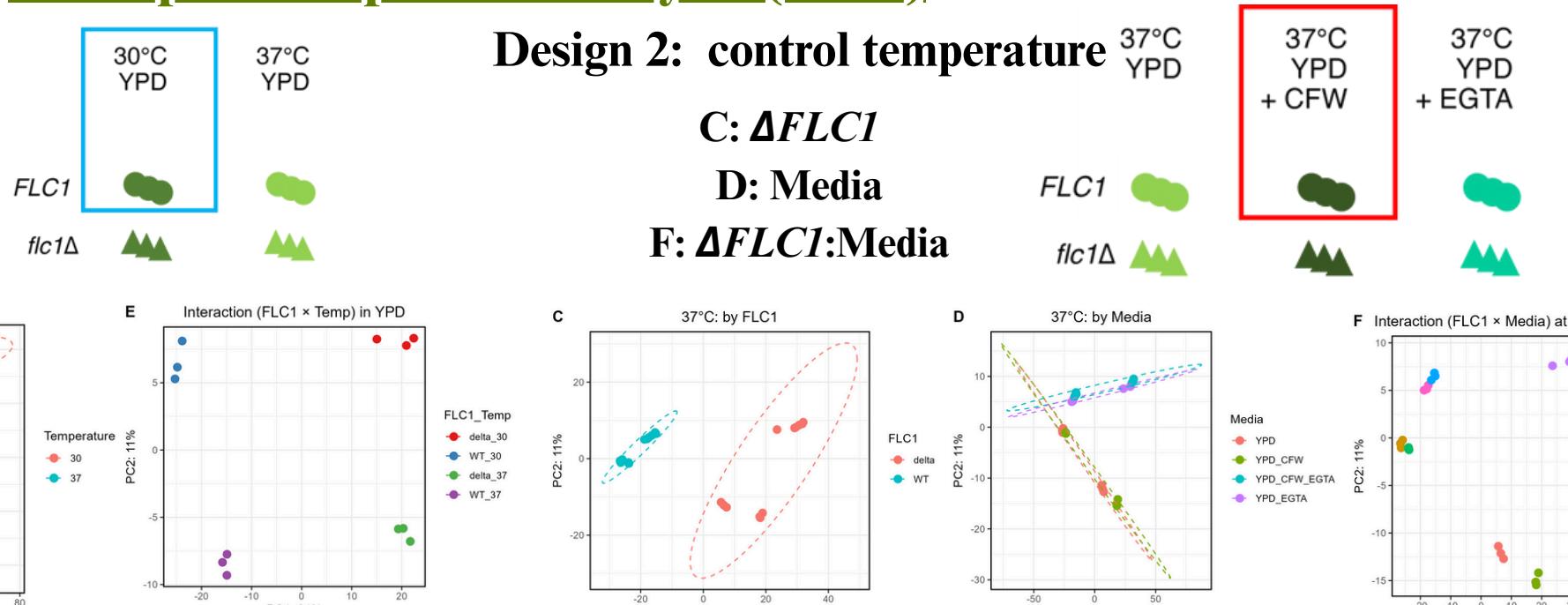
Principal Component Analysis (PCA)

Design 2: control temperature

C: **Δ*FLC1***

D: Media

F: **Δ*FLC1*:Media**



NB Regression Model

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

Model 1:

~Temperature + $\Delta FLC1$ + Temperature: $\Delta FLC1$

Model 2:

~Media + $\Delta FLC1$ + Media: $\Delta FLC1$

K: the raw read count

μ: the expected count

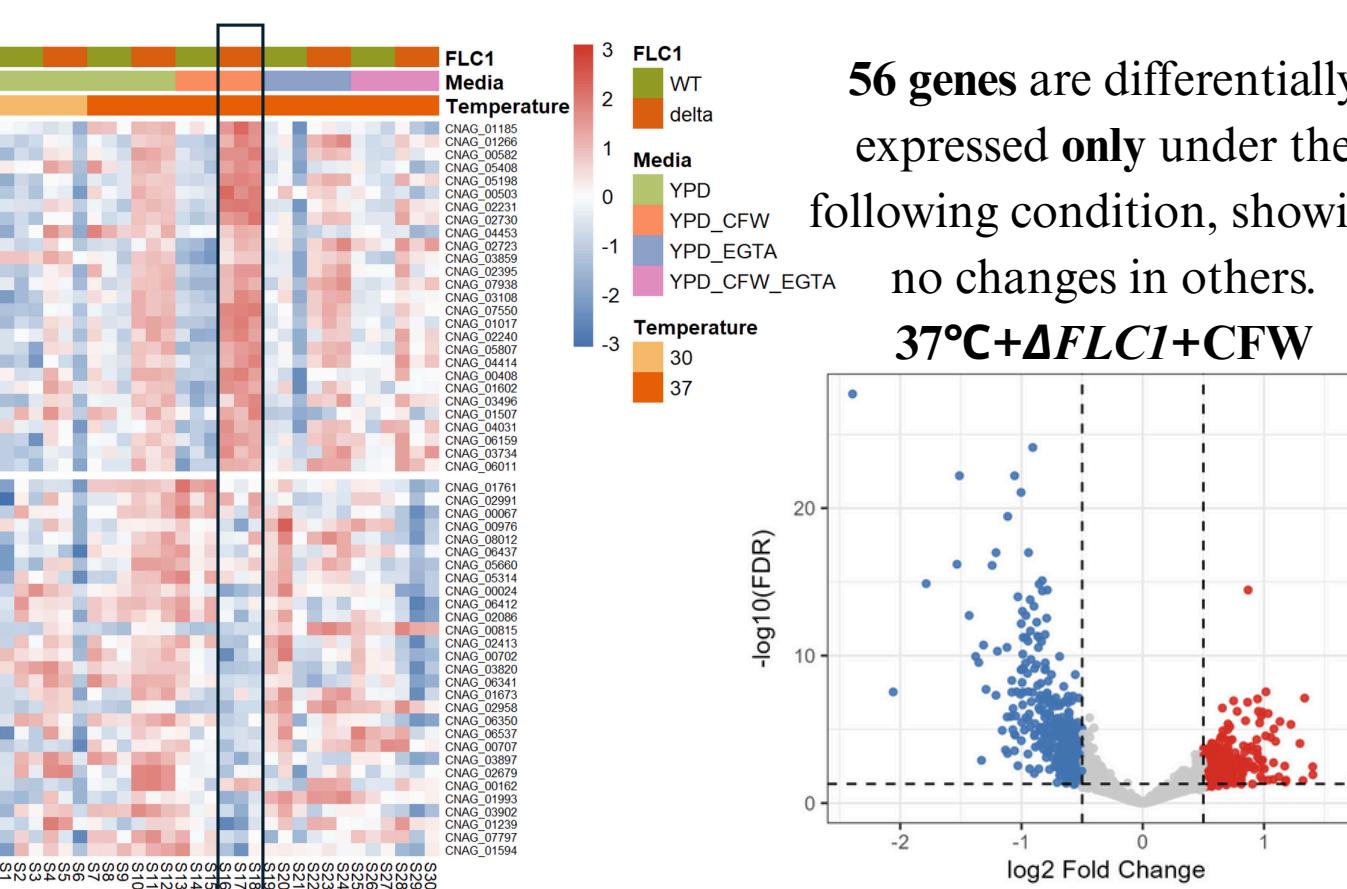
α: a gene-specific dispersion

S: the sample-specific size factor

q: the true gene expression

Differential Analysis

56 genes are differentially expressed **only** under the following condition, showing no changes in others.
37°C+Δ*FLC1*+CFW



High

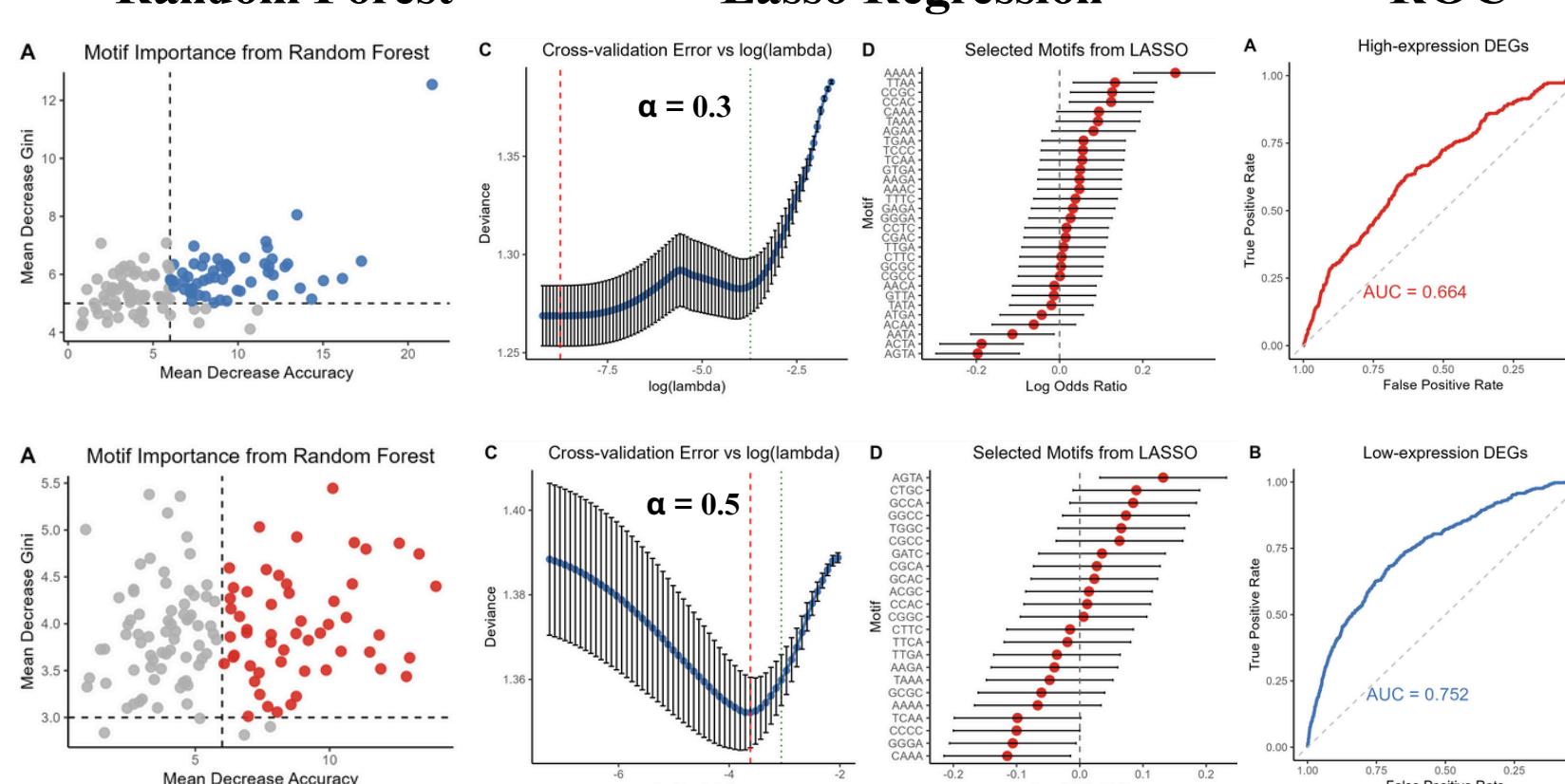
Low

Motif Prediction Model

Random Forest

Lasso Regression

ROC



Limitation

1. Incomplete control experiment (30°C+media).
2. Lack of accuracy for high-expression predictors.
3. Lack of external model validation.
4. Overlooks genes involved in complex regulation.

Conclusion

PCA demonstrated that temperature, *FLC1* deletion, medium composition, and their interactions significantly shaped global gene expression profiles. Notably, **56 genes** were differentially expressed exclusively under the condition of **37 °C + Δ*FLC1* + CFW**. Furthermore, **motif-based prediction** revealed that specific 4-mer DNA motifs were associated with stress-induced regulation of gene expression.