# The Transcriptional Portrait of Zinc Cluster Transcription Factors in *Candida Albicans*: A Network Approach to Capture the Complicated Co-Dependencies and Regulatory Relationships

Somayeh Haji Kazem Nili

A Thesis

in

The Department

of

Biology

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Science (Bioinformatics) at
Concordia University
Montreal, Quebec, Canada
December 2018

**CONCORDIA UNIVERSITY**

**School of Graduate Studies**

This is to certify that the thesis prepared

By: Somayeh Haji Kazem Nili

Entitled: The Transcriptional Portrait of Zinc Cluster Transcription Factors in *Candida Albicans*: A Network Approach to Capture the Complicated Co-Dependencies and Regulatory Relationships

and submitted in partial fulfillment of the requirements for the degree of

**Master of Science (Bioinformatics)**

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. Malcolm Whiteway_____Chair and Examiner

Dr. AshiqKachroo_____ External Examiner

Dr. Peter Darlington_____Examiner

Dr. Michael Hallett_____Supervisor

Approved by _____

Chair of Department or Graduate Program Director

_____2019 _____

Dean of Faculty

## Abstract

### The Transcriptional Portrait of Zinc Cluster Transcription Factors in *Candida Albicans*: A Network Approach to Capture the Complicated Co-Dependencies and Regulatory Relationships

Somayeh Haji Kazem Nili

This study focuses on understanding the transcriptional regulatory relationships in the fungal organism *Candida albicans* (*C. albicans*). We are particularly focused on zinc cluster transcription factors (ZCTFs) characterized by a conserved CX2CX6CX5–12CX2CX6–8C DNA binding domain. In general, the ~82 ZCTFs are known to be involved in a range processes including invasive growth, mating strategy and drug resistance. In this study, we make use of RNA-sequencing-based transcriptional profiles for a subset of 30 of these ZCTFs that were developed previously into gain of function mutants.

Our goals were (1) to ensure that the collection of transcriptional profiles were developed into a useful resource where hypotheses could be tested quickly with the assurance that the underlying data is sound, clean and largely free of technical artifacts; (2) to catalogue the global expression patterns across the cohort of ZCTFs and provide insight into the underlying biologies present across the ZCTF family while at the same time enumerating genes, pathways and processes that are unique or nearly unique to each of the ZCTFs in order to provide insight into the specific function of each member; and (3) to produce hypotheses from our correlative analysis regarding potential causative, regulatory relationships both between the ZCTFs and with other transcription factors.

The raw and normalized data, the code used throughout our analysis and the resultant analyses are available via a github repository.

# Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Michael Hallett, who gave me the opportunity to work and learn about fascinating area of bioinformatics and obtain great experiences in research. I am sincerely thankful of his support, patience and guidance over last year. I highly appreciate his insightful suggestions and the time he dedicated for my training and completion of my graduate thesis.

It is a great pleasure to express my sincere thanks to Professor Malcolm Whiteway and his student Raha Omran for construction of the mutants and preparation of samples used in this study.

I cannot find words to express my gratitude to Dr. Vanessa Dumeaux for her valuable basic processing of the RNA-sequencing profiles and data normalization.

This thesis would not be possible without their great help.

I would also like to take this opportunity and thank Professor Malcolm Whiteway and Dr. Darlington for being my committee members, and for their patience and valuable advice.

I would like to thank the Department of Biology at Concordia University for providing me with the fellowship and opportunity to work as a teaching assistant.

I would also like to thank Thi Truc Minh Nguyen, Nicholas Nelone and Samira Mashi for their scientific helps, over last couple of months.

I would like to acknowledge other lab members; Sanny Khurdia, Alexandra Artiaga, Shawn Simpson and former lab member Daniel Del Balso for their encouragement and friendship.

Lastly, I would like to thank my family, especially my parents, for their financial and emotional support during my study. All this work would not be possible without their generous support, love and care. This thesis is dedicated to my mother, who her love and prayers have always followed me.

**Contribution of Authors**

The project was initialized by Professor Malcolm Whiteway and his lab member, Raha Omran. They transcriptionally profiled 30 out of 82 zinc cluster transcription factors. These transcription factors were selected based on their conserved zinc cluster binding domains CX2CX6CX5–12CX2CX6–8C. To profile these transcription factors, gain-of-function strains were selected from a library constructed by Schillig and Morschhäuser.

# Table of Contents

# Lists of Figures

# List of Tables

# Introduction

*Candida albicans* (*C. albicans*) is a unicellular polymorphic fungal ascomycete. Although the genus Candida is highly polyphyletic, Candida is generally round and white when cultured (Stefanie Mühlhausen 2014; Fitzpatrick *et al*. 2006). Distinguishing features of the species Albicans in contrast to the rest of the genus are its ability to ferment glucose and maltose to acid and gas, the ability to ferment sucrose to acid, and the fact that it does not ferment lactose (Meyers *et al.* 1978; Trimble 1957). *C. albicans* is a non-obligate diploid organism although it is often observed to be in tetrapoidal state (or higher) and can survive as a haploid. The genome of *C. albicans* appears to be highly neoplastic, since it is able to copy its genome via parasex. Subsequent whole or partial chromosomal loss is suspected to provide *C. albicans* with significant survival advantages (Bennet *et al.* 2014). These and other fundamental characteristics have propelled *C. albicans* as an important resource for investigating basic fungal biology and has been a particularly important foil of other fungal ascomycetes particularly *Saccharomyces cerevisiae* (*S. cerevisiae*) (Sellam and Whiteway 2016).

*C. albicans* also impacts human health. Although this pathogen is a natural component of the human flora almost ubiquitously present on skin, its opportunistic nature can cause infections ranging from superficial (e.g. thrush, balanitis, vaginal candidiasis) to life-threatening systemic forms of candidiasis especially in compromised individuals (e.g. immunosuppressed individuals including premature infants, HIV-positive patients and organ transplant recipients) (Sellam and Whiteway 2016; Nobel *et al.* 2017). The paucity of anti-fungal drugs and increasing resistance to front-line therapeutics motivate a better understanding of *C. albicans* biology in order to identify targets and strategies with clinical value.

This study uses *C. albicans* wild-type strain, SC5314, as reference. All samples been used either control or mutants have been generated from SC5314 strain. The diploid genome of the strain is approximately 29 Mb, arranged in 8 chromosomes containing approximately 6,400 protein coding genes (Assembly 22). Of relevance to this effort, there are 238 transcription factors (TFs) in *C. albicans* (Assembly 22), however we are particularly interested in a subclass of TFs with a zinc finger DNA binding domain. Zinc cluster proteins (also referred to as zinc binuclear cluster or Zn(II)2Cys6 proteins) are a subfamily of zinc fingers exclusive to fungi, characterized by a well conserved CX2CX6CX5–12CX2CX6–8C DNA binding domain motif (Maicas *et al.* 2005; MacPherson *et al.* 2006; Schillig and Morschhäuser 2013). These Zinc Cluster TFs (ZCTFs) regulate diverse cellular processes in *C. albicans* including regulation of invasive filamentous

growth, switching to the mating component opaque form, and antifungal drug resistance (MacPherson *et al.* 2006; Schillig and Morschhäuser 2013).

In some cases, the role of the ZCTF has been well investigated. For example, ASG1 is the *C. albicans* orthologue of the regulator of stress response ASG1 and ARO80 has an ortholog of the same name in *S. cerevisiae* (Ghosh 2008). In other cases, the roles of these ZCTFs are less understood and it is difficult to establish a functional ortholog in *S. cerevisiae*. For example, little is known of *C. albicans* TFs ZCF4, ZCF18, ZCF21, orf19.1604 and LYS142 beyond the fact that they are suggested in Candida Genome Database (CGD) as best hits of LYS14 in *S. cerevisiae*, a transcriptional regulator of lysine biosynthesis. Some of the *C. albicans* TFs have been well-studied directly in *C. albicans* including the adherence regulators ZCF8, ZCF31 and SUC1 (Finkel *et al.* 2012). **Table 2** lists the *S. cerevisiae* ortholog, if it is known.

Our goal here is to build upon this information from the literature and computational inferences based on sequence similarity/orthology by analyzing the transcriptional response induced by activating each of the chosen 30 ZCTFs individually. There are at least three goals. First, to identify genes, pathways and processes that are unique or near-unique to each TF. Second, to identify common genes, pathways and processes across subfamilies or subgroups of the ZCTFs. Third, to understand if and how the individual ZCTFs co-regulate each other cognizant of that fact that there has been significant transcriptional “re-wiring” between *C. albicans* and other fungi especially *S. cerevisiae* where from a great deal of our knowledge of the *C. albicans* ZCTFs has been derived (Nantel 2006; Lavoie *et al.* 2010). For example, we would like to refine our understanding of the roles of ZCF4, ZCF18, ZCF21, orf19.1604, since we only know that they are similar to LYS14 in *S. cerevisiae*. Moreover, there appear to be significant differences in the environmental stress response (ESR) between *C. albicans* and other fungi including *S. cerevisiae* (Enjalbert *et al.* 2003). Characterizing the similarity and differences in the *C. albicans* stress responses across these 30 ZCTFs might provide general insight into the *C. albicans* stress mechanisms. For example, *C. albicans* appears to have distinct transcriptional responses for thermal, osmotic and oxidative stress with few commonalities in the downstream transcriptional response. Unlike *S. cerevisiae*, the current belief is that it does not appear to have a general ESR even though 80% of the *S. cerevisiae* ESR genes (Gasch *et al.* 2017) have a *C. albicans* ortholog (Enjalbert *et al.* 2003).

Our investigation began with the quality control and normalization of the next generation Illumina-based RNA-sequencing profiles using descriptive statistics and standard data science techniques with the aim to identify outliers (control or experiment samples) and subsequent

decision making to ablate stochastic errors and biases in the experimental data. Then, using standard techniques from bioinformatics and computational biology, we identified clusters of the TFs and clusters of genes that had similar transcriptional responses. This process allowed us to identify pathways, processes and responses that were unique to TF or sub-cluster of TFs and which did not appear to be an inherent part of the bet-hedging ESR strategy. We then proposed a network approach for understanding how the different ZCTFs might regulate each other, and used bioinformatics tools for the identification of transcription factor binding sites (TFBS) to more deeply investigate these relationships. Where possible, we focused our analysis on orf19.1604, a ZCTF which was poorly characterized in *C. albicans* (Mitrovich *et al.* 2007). The reciprocal best hit for orf19.1604 in *S. cerevisiae* was the non-essential gene LYS14. Over-expression of LYS14 increases pseudohyphal growth. When LYS14 was knocked-down, *S. cerevisiae* had increased thermotolerance. The null mutant strain displayed a reduction in fitness and increased resistance to selenomethionine.

# Results

### Transcriptional profiles of gain of function *Candida albicans* transcription factors

Here, we made use of Illumina MiSeq RNA-sequencing data generated by the Whiteway lab (unpublished). The Whiteway lab in turn made use of the ZCTF gain-of-function mutants from (Schillig and Morschhäuser 2013). Briefly, all *C. albicans* transcription factors (TFs) harboring a conserved zinc cluster Zn(II)2Cys6 transcription factor ZCTF DNA binding domain with the CX2CX6CX5–12CX2CX6–8C motif were identified (n=82), and a genetically modified strain of SC5314 *C. albicans* was constructed where each target TF was fused to a GAL4 activation domain of *S. cerevisiae* and placed under the control of a ADH1 promoter. The ADH1 promoter is strongly and ubiquitously expressed in *C. albicans* and the target TF is being activated by GAL4 domain, therefore the constitutively active target TF should be over-expressed in the modified strain, as compared to wildtype SC5314 (see **Methods 1**). Of the 82 Zinc cluster TFs where gain-of-function mutants were viable, 30 poorly characterized ZCTFs were randomly chosen for transcriptional profiling using the Illumina MiSeq technology with two biological replicates. Wild-type SC5314 *C. albicans* were used as controls across four batches (see **Methods 2**). The resultant RNA-seq profiles were processed by Dr V Dumeaux using a pipeline for basic quality control and trimming of reads, followed by alignment against the *C. albicans* SC5314 haplotype

A, version A22 (**Methods 3**). **Table 1** details the TFs, batches, controls and parameters related to the RNA-seq profiles.

**The gain-of-function ZCTF profiles harbour batch effects**

We applied hierarchical clustering using 500 genes selected either for high variance or high inter-quartile range (IQR; **Methods 7**). **Figure 1** shows the expression pattern using IQR. Here colors correspond to the (log2) ratio of the sample versus control. For TFs in batches 1, 3, and 4, the control corresponds to the geometric average of the wild-type experiments performed in their respective batch. For TFs in batch 2, the control corresponds to the geometric average of the controls for batch 1 (wild-type *C. albicans* was not profiled in batch 2). The samples consistently clustered by batch number suggesting bias in the expression data. The batch effect remained present if (1) variance instead of IQR is used, (2) using more or fewer selected features (e.g. number of genes included by adjusting the IQR threshold) and (3) changing the control used for batch 2 to the controls from batch 1, 3 or 4 (data not shown).

**A pooled control strategy ablates the observed batch effect**

We asked if the batch effect was also observed when only the controls from either batch 1, 3, or 4 were used across all TFs. **Figures 2-4** show that TFs no longer clustered by batch when individual batch controls were used. Not surprisingly, this also were true when the controls from both batch 3 and 4 were combined (**Figure 5**). **Figure 6** confirms that the same TF clusters were induced regardless of whether control 3, 4 or combined controls 3 and 4 were used. However, both control 1 and the natural controls induce different clusters of the ZCTFs **(Figure 6)**.

We then investigated whether the same genes were identified as the most variable across the different controls, and found agreement when either controls 1, 3, 4 or both 3 and 4 are used (**Figure 7**). We used the term *natural controls* to refer to the default normalization, where controls from batch 1, 3, and 4 were to normalize only TFs in batch 1, 3 and 4 respectively. Batch 2 was paired with the controls from batch 1 arbitrarily. When the natural controls were used, we did not observe agreement; more than half (n=59) of the 100 most variable genes were unique to the "natural" controls. This suggests that controls 1, 3 and 4 do not significantly affect which genes are differentially expressed (d.e.), nor the magnitude of this differential expression.

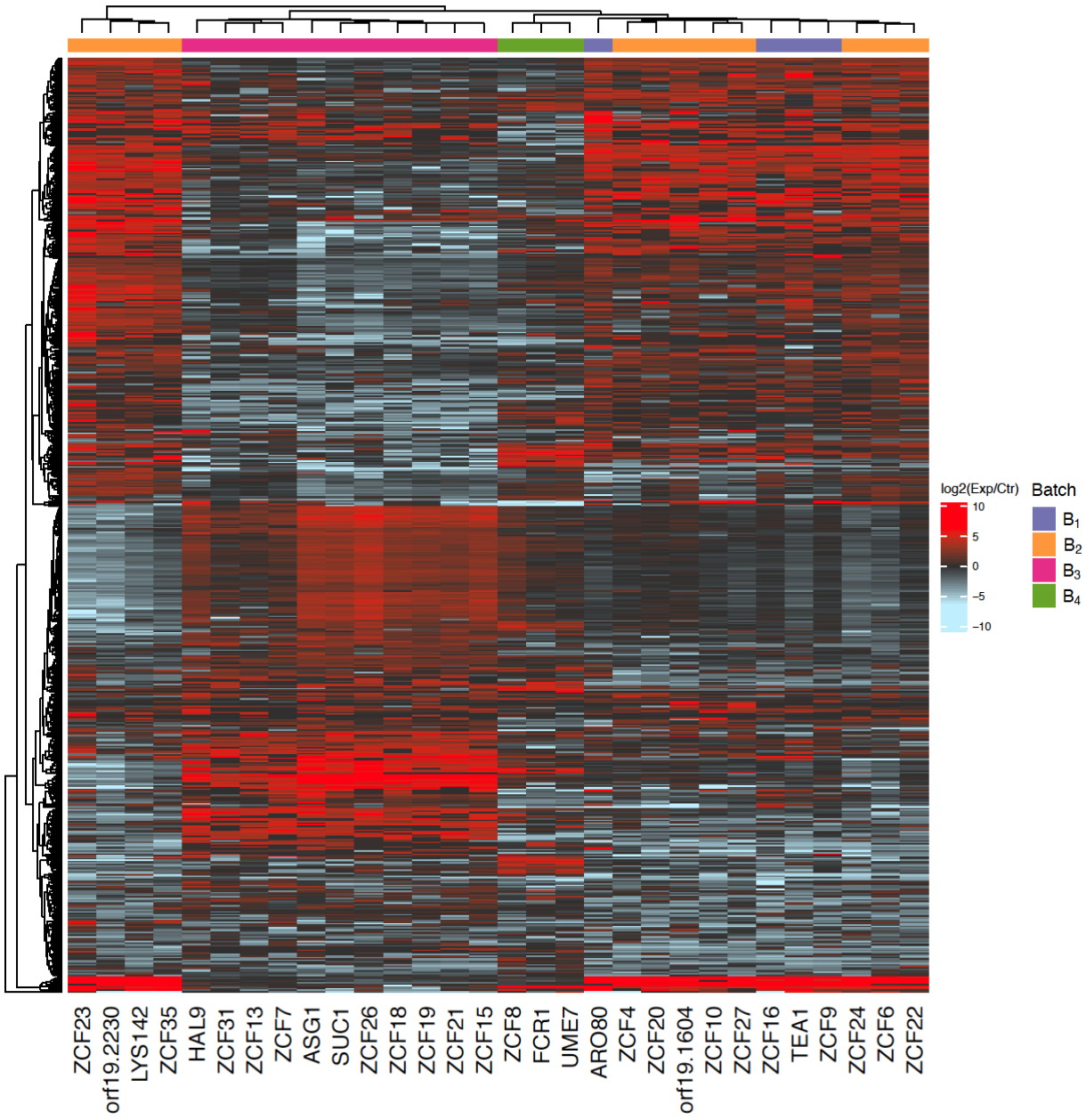



**Figure 1.TF samples cluster by batch number.**

Rows correspond to genes selected with IQR > 1.5. Columns correspond to samples. Columns are also labelled by their respective batch (see also Table 1). Depicted are the log2 ratio of the observed expression of the TF (column) to the geometric average of the two controls from the same batch. Since batch 2 did not contain controls, the controls from batch 1 were used for batch 2 samples.

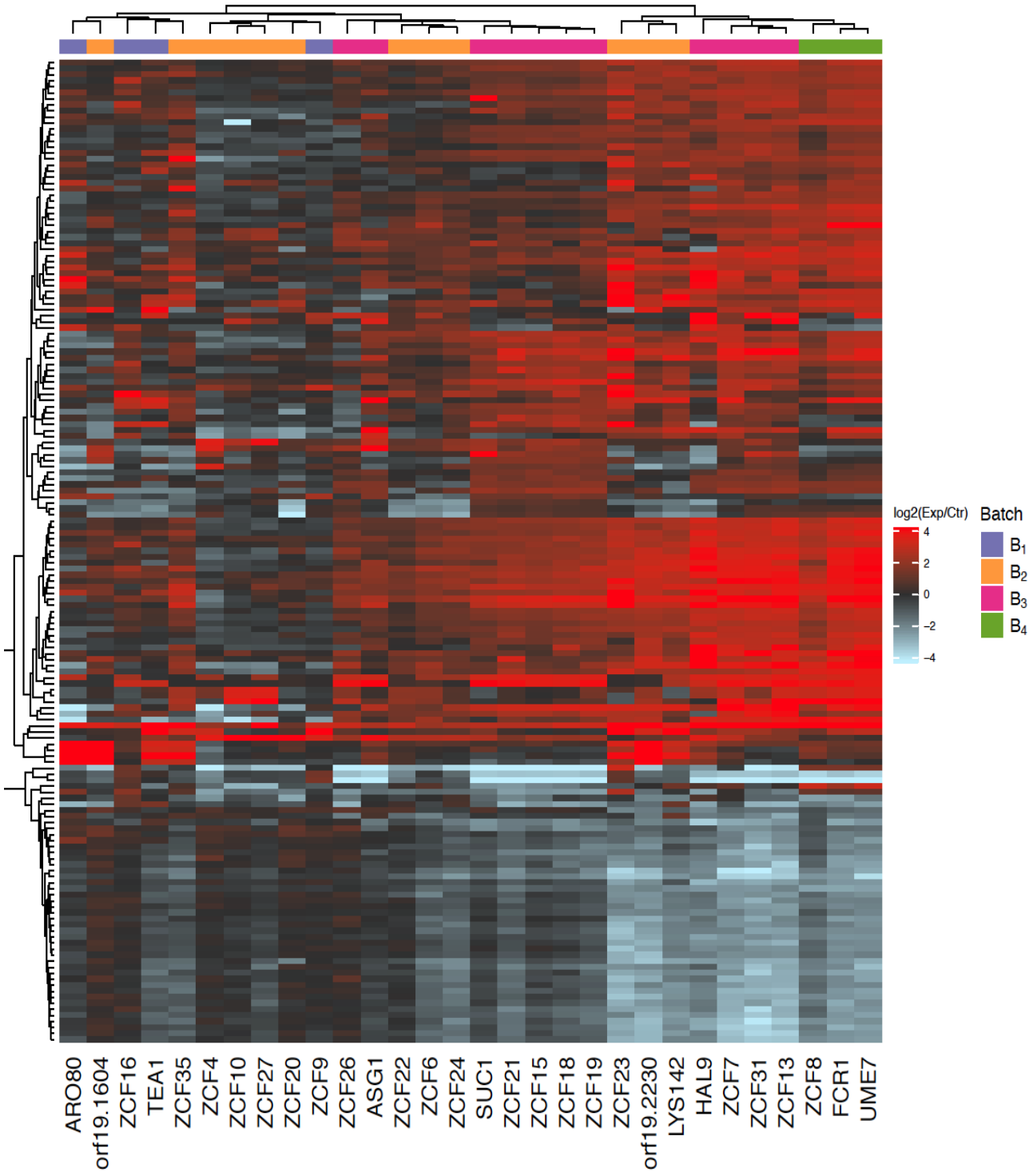



**Figure 2. Heatmap of all ZCTFs using only controls from batch 1.**

Here expression corresponds to log2 ratio of the observed measurement for a TF versus the geometric average of the two wildtype samples from control 1. Rows correspond to genes with IQR > 1.5.

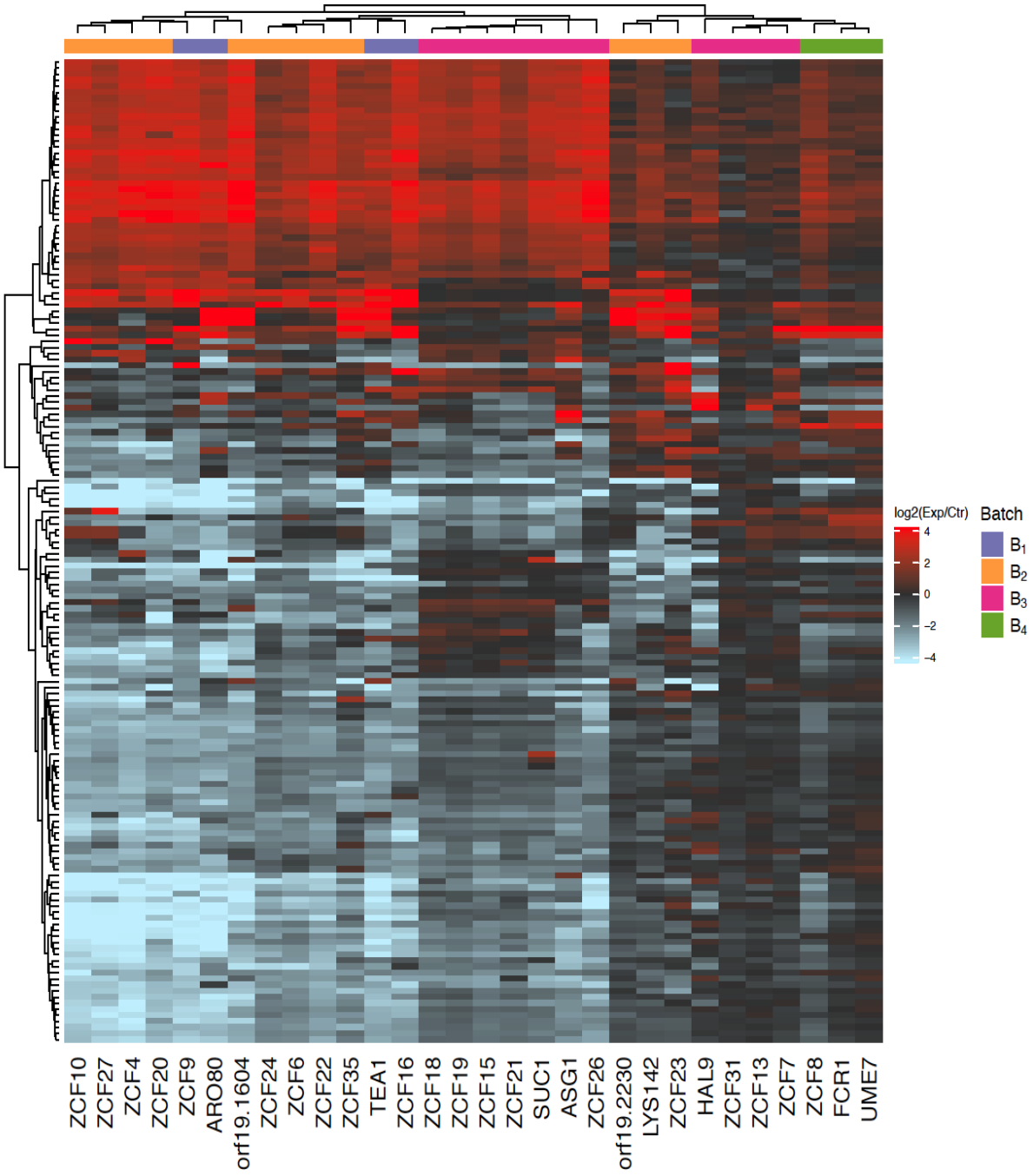



**Figure 3. Heatmap of all ZCTFs using only controls from batch 3.**

Here expression corresponds to log2 ratio of the observed measurement for a TF versus the geometric average of the two wildtype samples from control 3. Rows correspond to genes with IQR > 1.5.

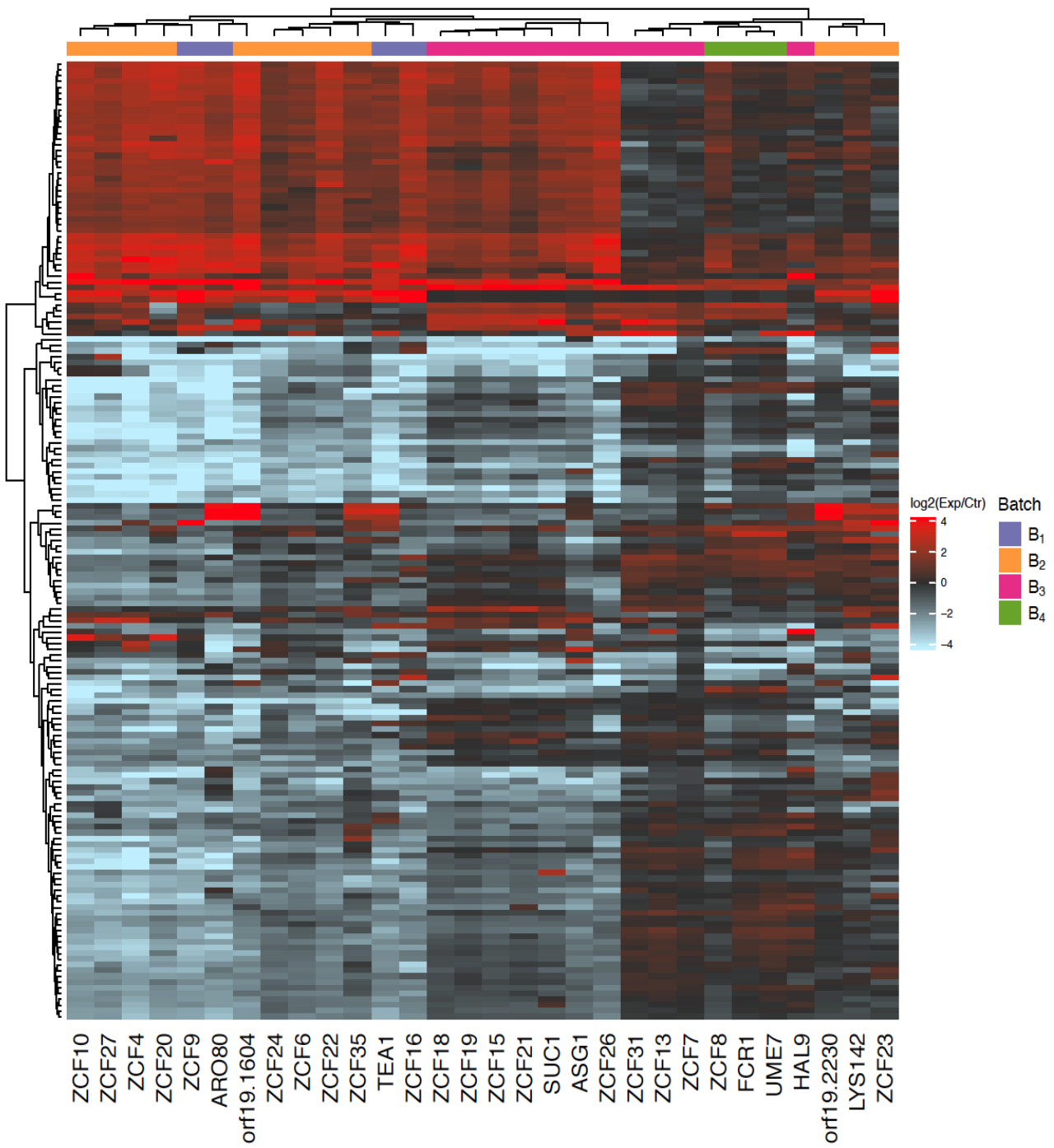



**Figure 4. Heatmap of all ZCTFs using only controls from batch 4.**

Here expression corresponds to log2 ratio of the observed measurement for a TF versus the geometric average of the two wildtype samples from control 4. Rows correspond to genes with IQR > 1.5.

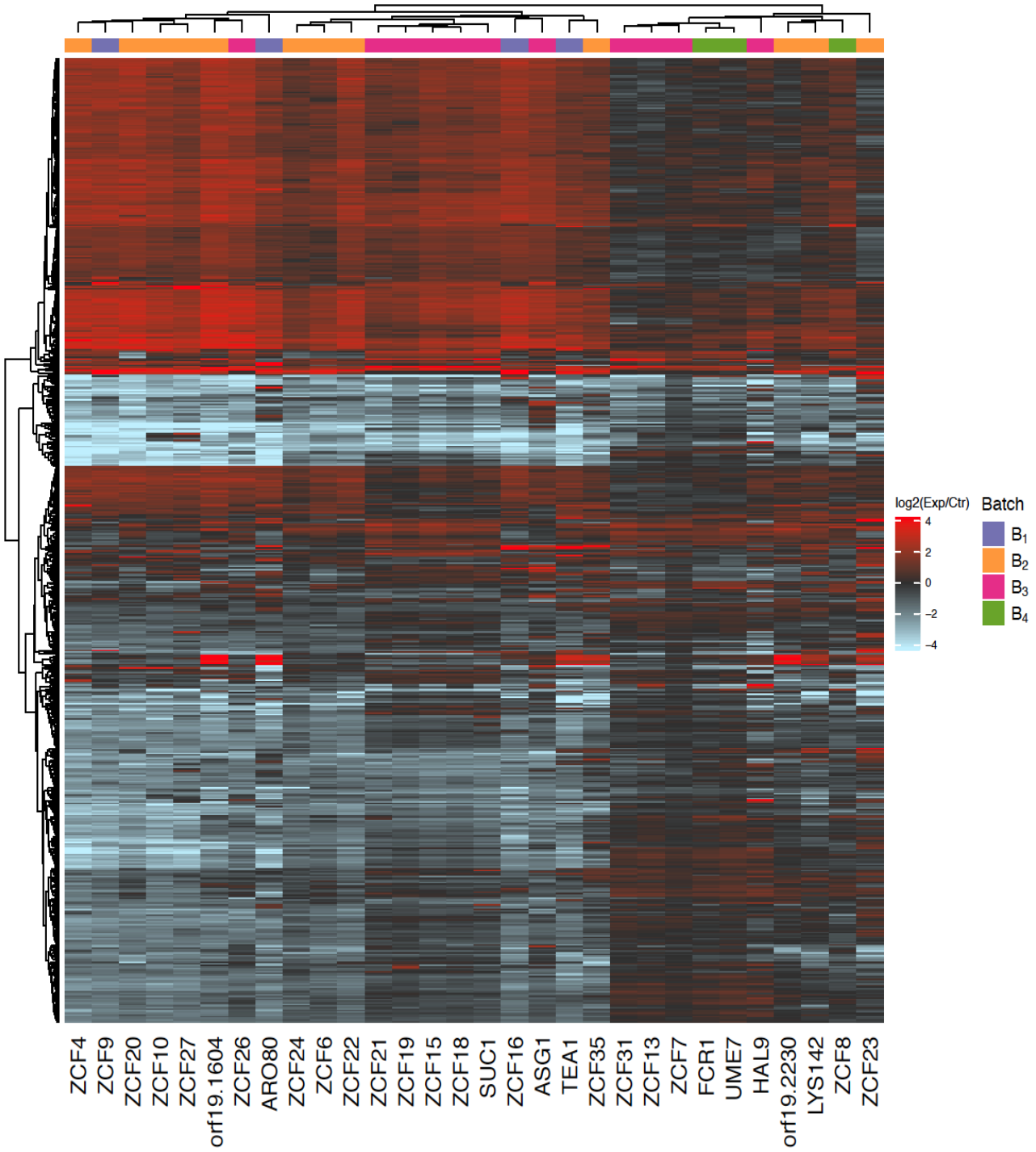



**Figure 5**. **Heatmap of all ZCTFs using only controls from both batch 3 and 4.**

Here expression corresponds to log2 ratio of the observed measurement for a TF versus the geometric average of the two wildtype samples from control 3 and 4. Rows correspond to genes with IQR > 1.5.

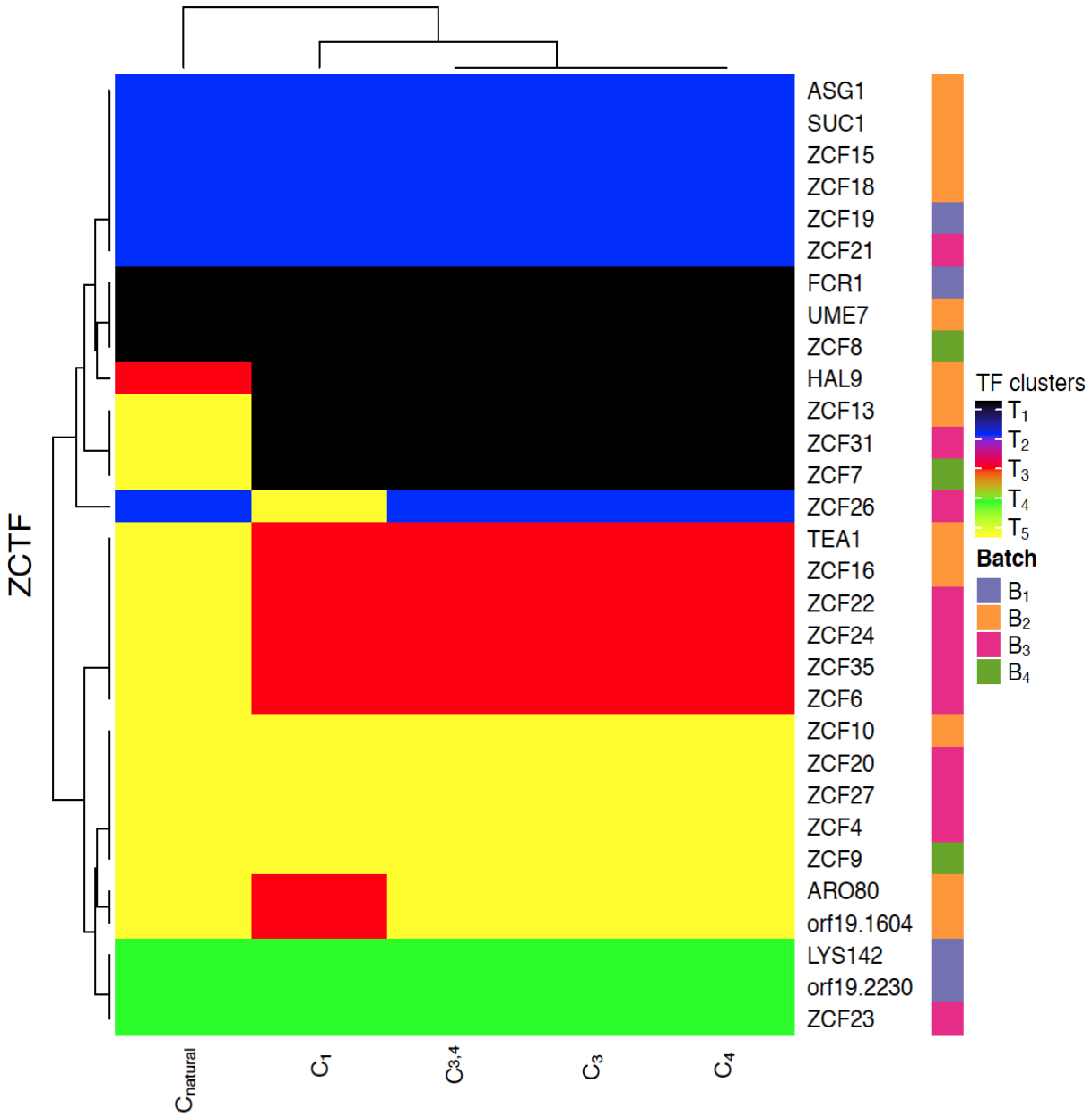



**Figure 6. Heatmap of ZCTF hierarchical clustering for differential expression analysis using different controls.**

Here five clusters for each were identified based on euclidean distance and complete linkage using 100 genes with highest IQR in each analysis.

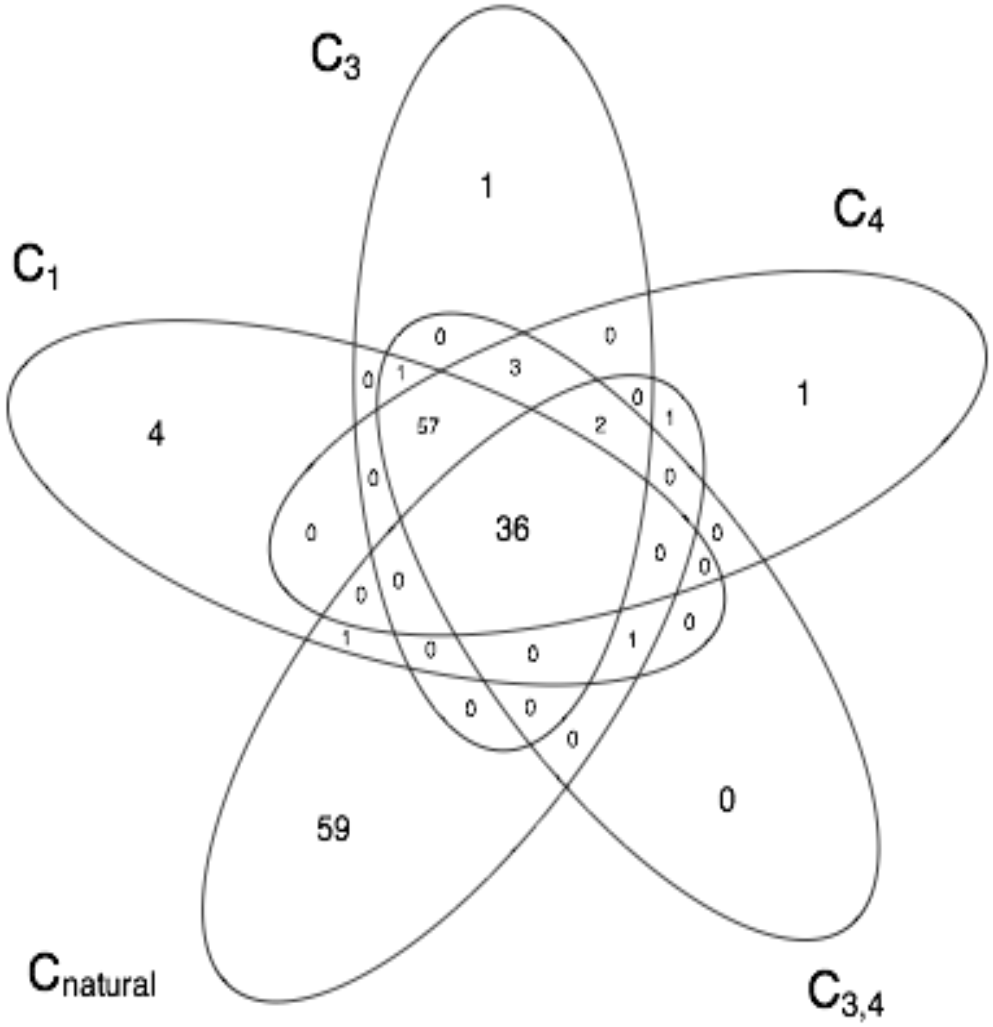



**Figure 7. Venn diagram showing agreement across different choices of control in terms of the number of genes in common in most variable.**

$C_1$, $C_3$,$C_4$, $C_{3,4}$ , and $C_{natural}$ represent the 100 highest variance genes (IQR) across all samples relative to different controls. Here $C_{natural}$ is the set of 100 highest variance genes but batch 1, batch 3 and batch 4 samples are contrasted with controls 1, 3 and 4 respectively. Batch 2 are contrasted with controls 1.

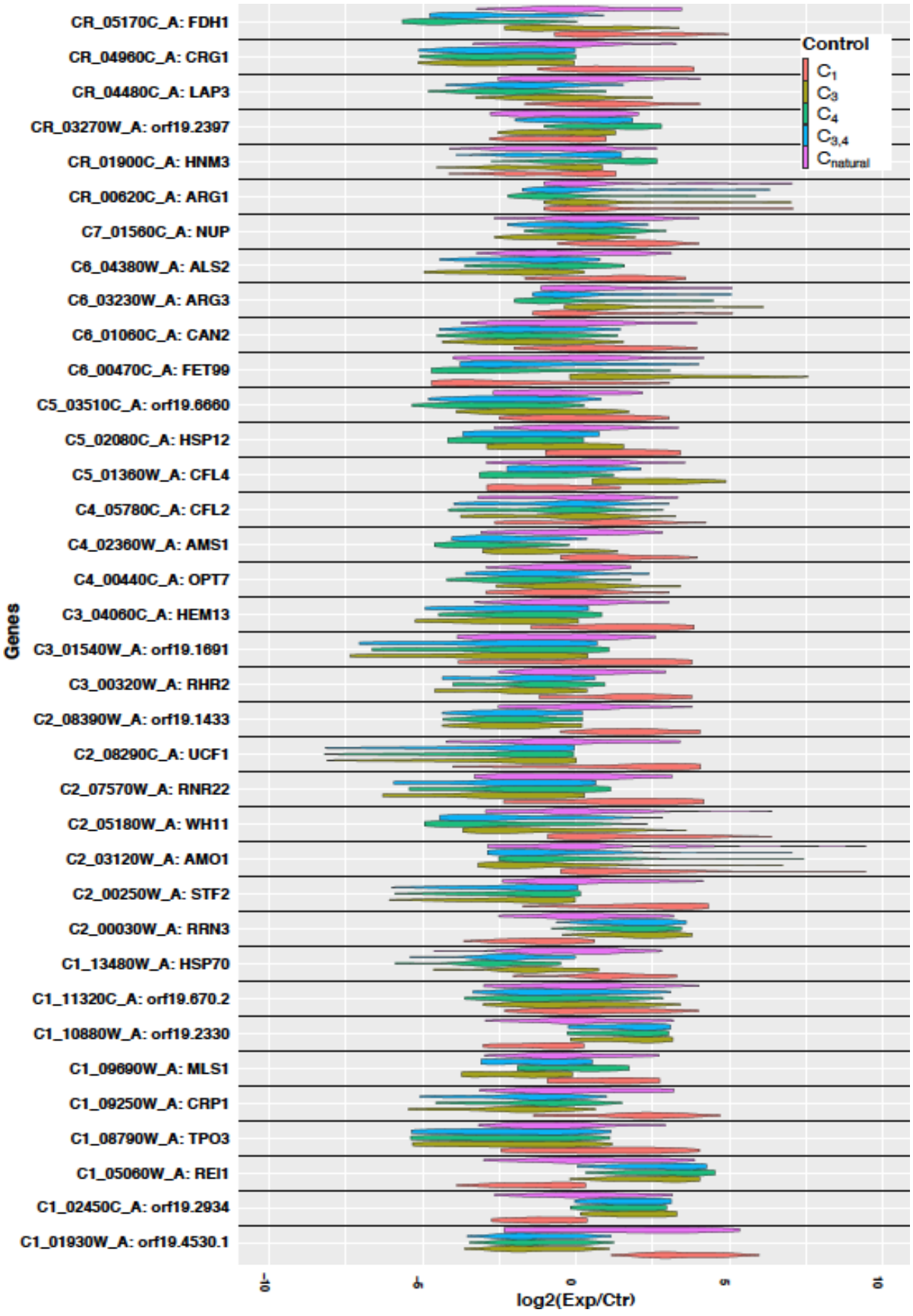

CR_05170C_A: FDH1
CR_04960C_A: CRG1
CR_04480C_A: LAP3
CR_03270W_A: orf19.2397
CR_01900C_A: HNM3
CR_00620C_A: ARG1
C7_01560C_A: NUP
C6_04380W_A: ALS2
C6_03230W_A: ARG3
C6_01060C_A: CAN2
C6_00470C_A: FET99
C5_03510C_A: orf19.6660
C5_02080C_A: HSP12
C5_01360W_A: CFL4
C4_05780C_A: CFL2
C4_02360W_A: AMS1
C4_00440C_A: OPT7
C3_04060C_A: HEM13
C3_01540W_A: orf19.1691
C3_00320W_A: RHR2
C2_08390W_A: orf19.1433
C2_08290C_A: UCF1
C2_07570W_A: RNR22
C2_05180W_A: WH11
C2_03120W_A: AMO1
C2_00250W_A: STF2
C2_00030W_A: RRN3
C1_13480W_A: HSP70
C1_11320C_A: orf19.670.2
C1_10880W_A: orf19.2330
C1_09690W_A: MLS1
C1_09250W_A: CRP1
C1_08790W_A: TPO3
C1_05060W_A: REI1
C1_02450C_A: orf19.2934
C1_01930W_A: orf19.4530.1
Genes
Control
C1
C3
C4
C3,4
Cnatural
−10
−5
0
5
10
log2(Exp/Ctr)

**Figure 8. Violin plot of 36 common genes in most variable genes across different choice of control.**

The plot is based on differential expression value ($\log_2$ fold change between ZCFT and control) across the ZCFTs . $C_1$, $C_3$,$C_4$, $C_{3,4}$ , and $C_{natural}$ represent the choice of control being used for analysis.

**Figure 8** shows via violin plots that the distribution of expression of all 36 genes common across analysis based on different controls **Figure 7**. We observe general agreement in the distribution of expression for controls 3, 4 and both 3 and 4, but observed large shifts for both the natural controls and for batch 1 controls, suggesting that control 1 was also problematic.

Together, the above results suggest that pooled controls from batches 3 and 4, excluding batch 1, are the best available. Therefore, throughout the remainder of the manuscript, for each TF we compute the log2 ratio of its observed expression versus the four control (wildtype SC5314) replicates present in batches 3 and 4, within a `DESeq2` design.

**The ZCTFs display a continuous spectrum with five levels where specific molecular processes show strong co-expression**

**Figure 9** depicts the 500 most variable genes across the TFs using our pooled control approach. Both TFs and genes showed strong sub-clusters. With respect to TF cluster, there are five large patterns (labelled $T_1$ through $T_5$) and at least five large patterns of gene expression (labelled $G_1$ through $G_5$). Broadly speaking, $T_5$ shows the highest degree of differential expression for these genes and $T_1$ shows the least, with a near continuum across $T_2$,$T_3$, and $T_4$.

Both clusters 1 and 3 ($G_1$ and $G_3$ in **Figure 9**) are statistically significantly enriched in RNA (rRNA, ncRNA) and ribosomal biogenesis and maturation, RNA and protein export, and some metabolic processes (heterocycle, organic cycle, cellular aromatic and nitrogen compound production) (**Table 3**; GO analysis adj p-value < 0.001; **Methods 10**). Except for $T_1$, this enrichment is observed in all TF clusters. The strength of over-expression follows a linear progression from $T_1$ (no expression compared to control) to $T_5$ where there is strong (>20 fold in some cases) over-expression.

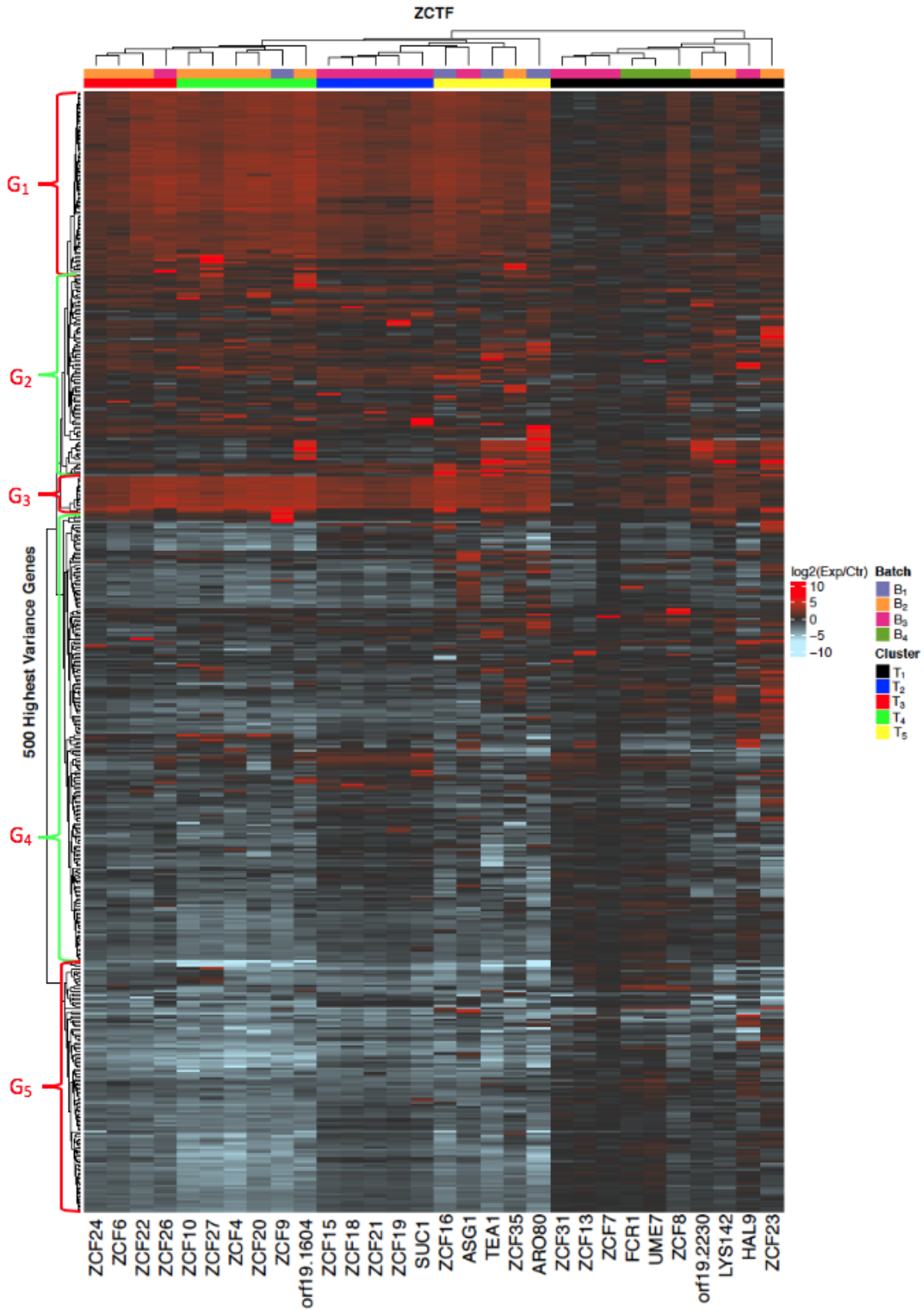

ZCTF
G1
G2
G3
G4
G5
500 Highest Variance Genes
log2(Exp/Ctr)
10
5
0
−5
−10
Batch
B1
B2
B3
B4
Cluster
T1
T2
T3
T4
T5
ZCF24
ZCF6
ZCF22
ZCF26
ZCF10
ZCF27
ZCF4
ZCF20
ZCF9
orf19.1604
ZCF15
ZCF18
ZCF21
ZCF19
SUC1
ZCF16
ASG1
TEA1
ZCF35
ARO80
ZCF31
ZCF13
ZCF7
FCR1
UME7
ZCF8
orf19.2230
LYS142
HAL9
ZCF23

**Figure 9. A global portrait of the ZCTF transcriptional response with our pooled control.**

Here the 500 most variable genes were chosen. The pooled controls from batches 3 and 4 are used across all TFs. With respect to TF cluster, there are five patterns observed, labelled $T_1$ through $T_5$. With respect to genes, there are five broad patterns labelled $G_1$ through $G_5$.

In TF cluster $T_2$, $G_5$ is enriched for genes involved in oxidative reduction and sugar/carbohydrate processes; they are both under-expressed compared to control. These processes are also observed to be enriched and downregulated in TF clusters $T_3$ through $T_4$. $T_2$ through $T_5$ are also enriched for genes involved in energy reserves; they are down-regulated. Regardless of TF cluster, $G_2$ is not enriched for any GO process. However, it indeed shows some enrichment for oxidative reduction, nitrogen utilization, amino acid biosynthesis, some metabolic processes and transport of different compounds in the stronger expressing TF clusters of $T_4$ and $T_5$. In general, it appears that a few processes are nearly ubiquitous across the panel of ZCTFs with the strength of their over- or under-expression forming a linear progression from weakest ($T_1$) to strongest ($T_5$).

**There is no evidence that the gain of function ZCTFs induce a Environmental Stress Response (ESR)**

We observed that many of the 500 most variables genes (**Figure 9**) have had Sc orthologs that are present in the yeast Environmental Stress Response (ESR) (Gasch et al. 2017). The yeast ESR is a transcriptional catalogue of the genes differentially expressed when yeasts are grown in conditions that cause stress (e.g. grown in NaCl). It contains 859 genes and is divided into three broad categories called the induced ESR (iESR; genes that are differentially regulated in response to environmental xenobiotics, conditions or other challenges), the ribosomal proteins (RP) and the ribosomal biogenesis genes (RiBi; involved in rRNA production, growth and cell division). We were able to identify *C. albicans* orthologs for 642 of the 859 *S. cerevisiae* ESR genes (**Supplemental Table 1** and **Methods 8**). In total, 111 of the 500 most variable genes from **Figure 9** overlap with the 642 *C. albicans* ESR orthologs. Of the ZCTFs, only ZCF23 and ASG1 appear within these 111 genes, although 26 of 30 are amongst the 500 most variable genes. This overlap (which is roughly double what we would expect by chance) is statistically significant (hypergeometric test, p-value < $10^{-17}$, see also **Methods 4**).

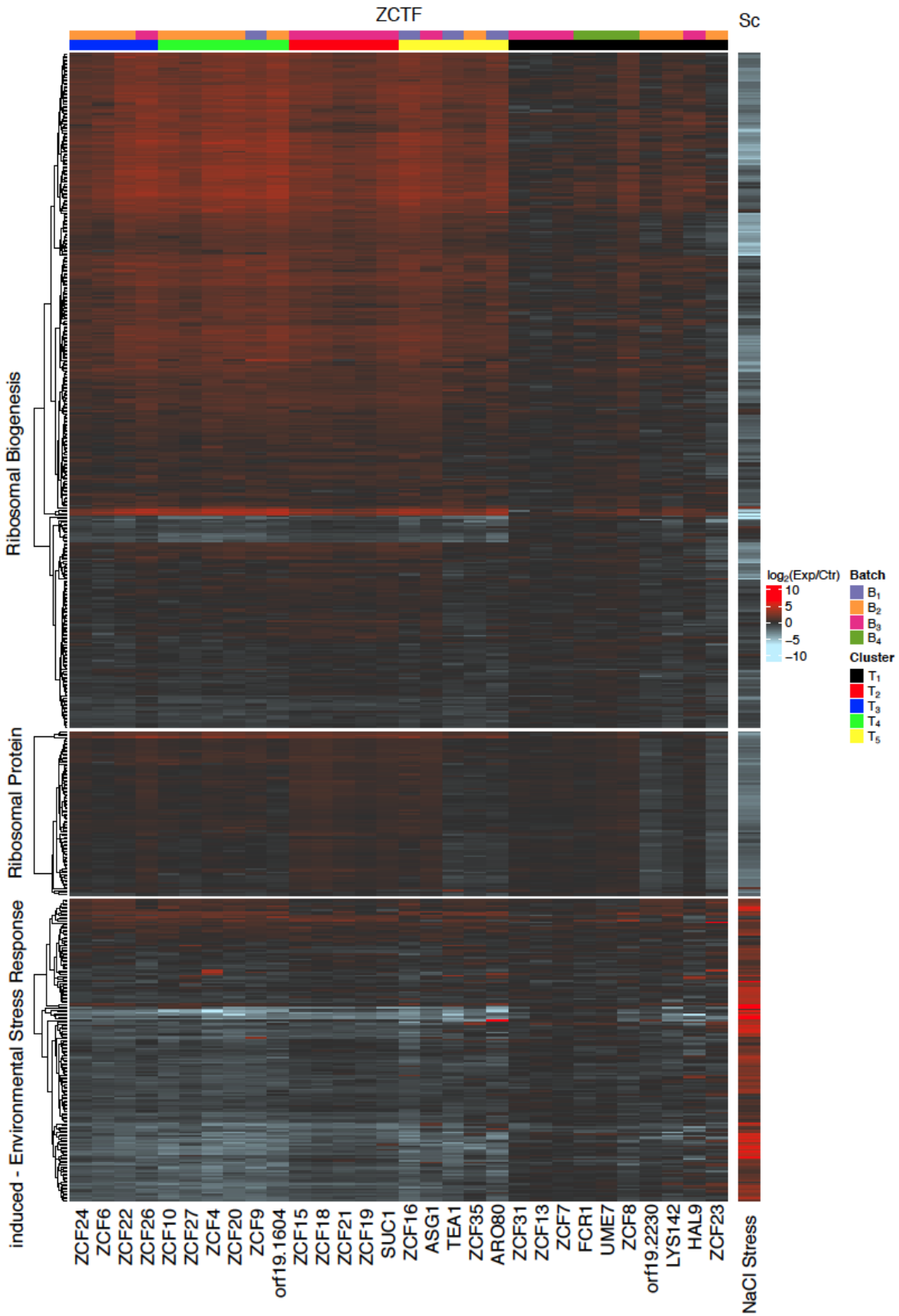

ZCTF
Sc
induced - Environmental Stress Response
Ribosomal Protein
Ribosomal Biogenesis
ZCF24
ZCF6
ZCF22
ZCF26
ZCF10
ZCF27
ZCF4
ZCF20
ZCF9
orf19.1604
ZCF15
ZCF18
ZCF21
ZCF19
SUC1
ZCF16
ASG1
TEA1
ZCF35
ARO80
ZCF31
ZCF13
ZCF7
FCR1
UME7
ZCF8
orf19.2230
LYS142
HAL9
ZCF23
NaCl Stress
log2(Exp/Ctr)
10
5
0
−5
−10
Batch
B1
B2
B3
B4
Cluster
T1
T2
T3
T4
T5

**Figure 10. The expression pattern across the ZCTFs for the *C. albicans* orthologs of the yeast ESR genes.**

The log2 expression of the ZCTFs is plotted for the *C. albicans* orthologs of the yeast ESR signature. The right most column represents the direction of expression of the gene expression in the original paper that defined the ESR (Gasch et al. 2017). Note that this image of the ESR in *C. albicans* groups the TFs in a manner that is similar to the clusters $T_1$, ... , $T_5$ introduced above. The rightmost column represents the log2 ratio of the mean expression value across NaCl stressed versus unstressed cells from (Gasch et al. 2017).

**Figure 10** depicts the expression of the *C. albicans* orthologs for the ESR signature across the ZCTF. The right most column depicts the expression (NaCl conditions versus control) of these genes in the original Gasch *et al.* manuscript.

**Evidence for aneuploidy in the gain of function ZCTF mutants**

It is well established that the *C. albicans* genome becomes neoplastic when stressed. To identify potential chromosomal aberrations in our gain-of-function mutants, we mapped gene expression data onto their chromosomal position **(Figure 11**, also **Methods 9).** The box plot depicts the distribution of expression over the entire chromosome. Significant deviations from 0 (log expression of ZCTF versus control data points) suggest that either there is concerted over-expression or under-expression of many genes along the entire chromosome that can potentially correspond to chromosomal duplication or loss. Only ZCTF members that exhibit a difference from 0 in at least one chromosome are depicted (ZCF8 is the exception and is included as a control). Approximately one third of the ZCTFs show some evidence for aneuploidy: ZCF9, orf19.1604, ZCF-4,6,10,15,19, 20, 22, 24, 26. Most of these have events on Chromosome 5 with ZCF9 exhibiting the highest degree of over-expression.

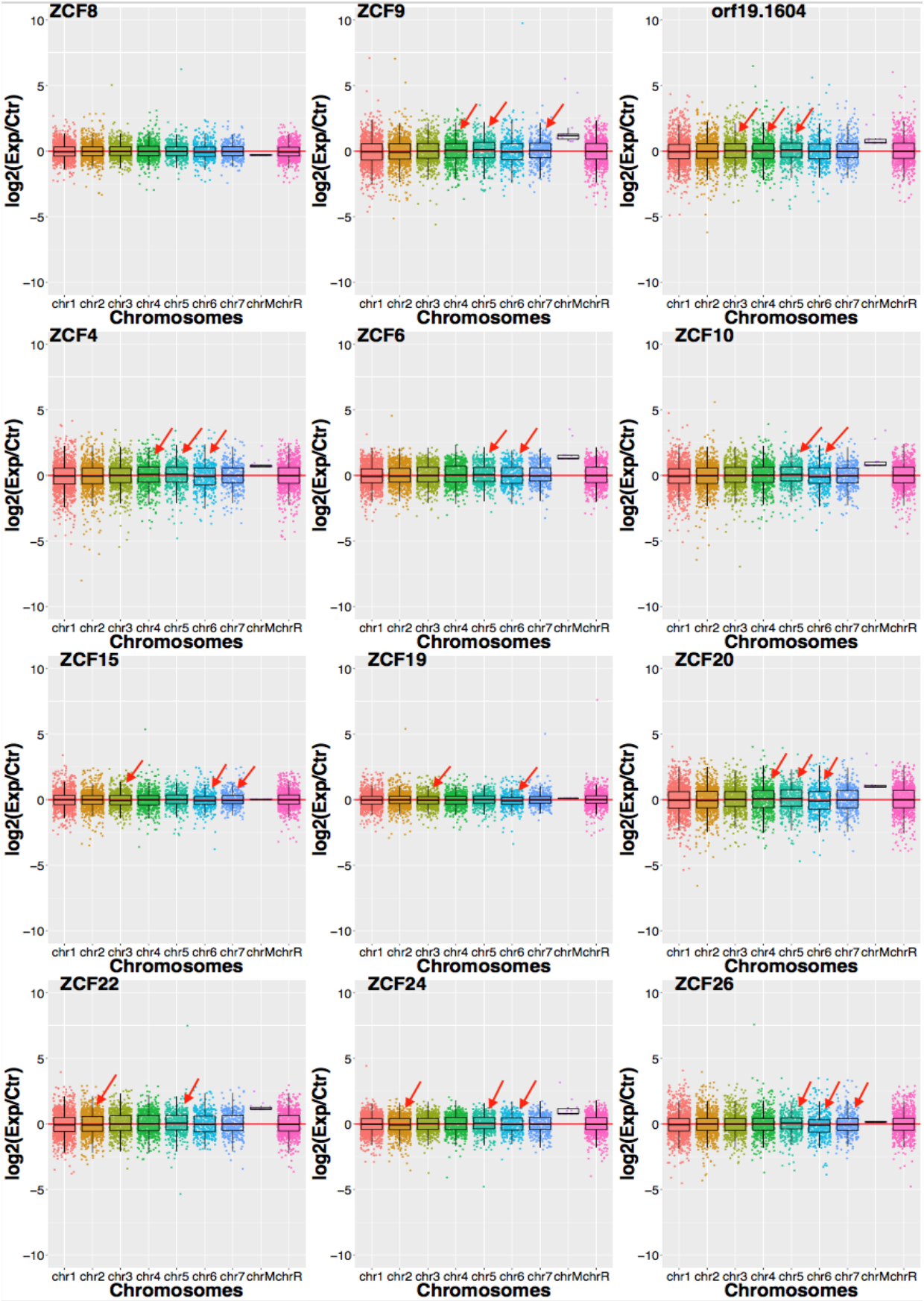

ZCF8
ZCF9
orf19.1604
ZCF4
ZCF6
ZCF10
ZCF15
ZCF19
ZCF20
ZCF22
ZCF24
ZCF26
log2(Exp/Ctr)
10
5
0
−5
−10
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chrMchrR
Chromosomes

**Figure 11. Evidence of aneuploidy across the ZCTFs.**
For each ZCTF, the expression of each gene is mapped to its chromosomal location (concatenated in the figure). Boxplots indicate the global pattern of expression for the entire chromosome. Arrows indicate chromosomes where expression appears to shifted away from the baseline. With the exception of ZCF8, only chromosomes with a least one shift from 0 are plotted.

**Differentially expressed genes for each ZCTF**

Whereas the previous analysis focused on the broad transcriptional responses that are nearly ubiquitous across all of the gain-of-function ZCTFs, we next focus on identifying genes that are differentially expressed (d.e.) uniquely or nearly uniquely for each ZCTF. In particular, for each ZCTF, we use a linear model from the DESeq2 (Love *et al.* 2014) R package (see **Methods 6**) to identify all genes whose expression is significantly different from control, with adjusted p-value < $10^{-4}$. Not surprisingly, we observe that most of genes from clusters $G_1$, $G_3$ and $G_5$ of **Figure 9** are present in the list of d.e. genes for every one of the ZCTFs. In fact, the d.e. genes significantly overlap for almost every pair of ZCTFs (Fisher's exact test, **Figure 12**). To ablate these ubiquitous signals, we developed a novel approach to stratify the d.e. genes for each ZCTF by their uniqueness. In particular, the d.e. genes for a specific ZCTF were stratified by the number of times they appeared in the list of d.e. genes for other ZCTFs. **Figure 13** displays the result of this approach, where white (0) corresponds to a gene that is d.e. only for that ZCTF and no other. Darker shades of red indicate genes that are d.e. for a greater number of ZCTFs. These correspond to genes from clusters $G_1$ - $G_5$ of **Figure 9**. **Supplemental Table 3** enumerates the d.e. genes per ZCTF for each stratification (shared with 0, 1-4 or 5-9 other ZCTFs). We sample findings for some of the ZCTFs below.

*Orf19.1604* is the only member of the ZCTF panel where the gain-of-function mutant induces the hyphae transition in *C. albicans*. The d.e. genes for orf19.1604 (unique n=37, 1-4 TFs n=138, 5-9 TFs n = 184) include ECE1 (a well-established controller of hyphal induction; log fold change 6.49), NGR1 (a known repressor of hyphal induction; log fold change -1.2), and BCR1 an established inducer of the hyphal transition (1.9 log2 fold change) (Kumamoto and Vinces 2005; Nobile and Mitchell 2005). This small sample of findings suggest the catalogue of d.e. genes for orf19.1604 is consistent with established biology for *C. albicans* from the literature.

.

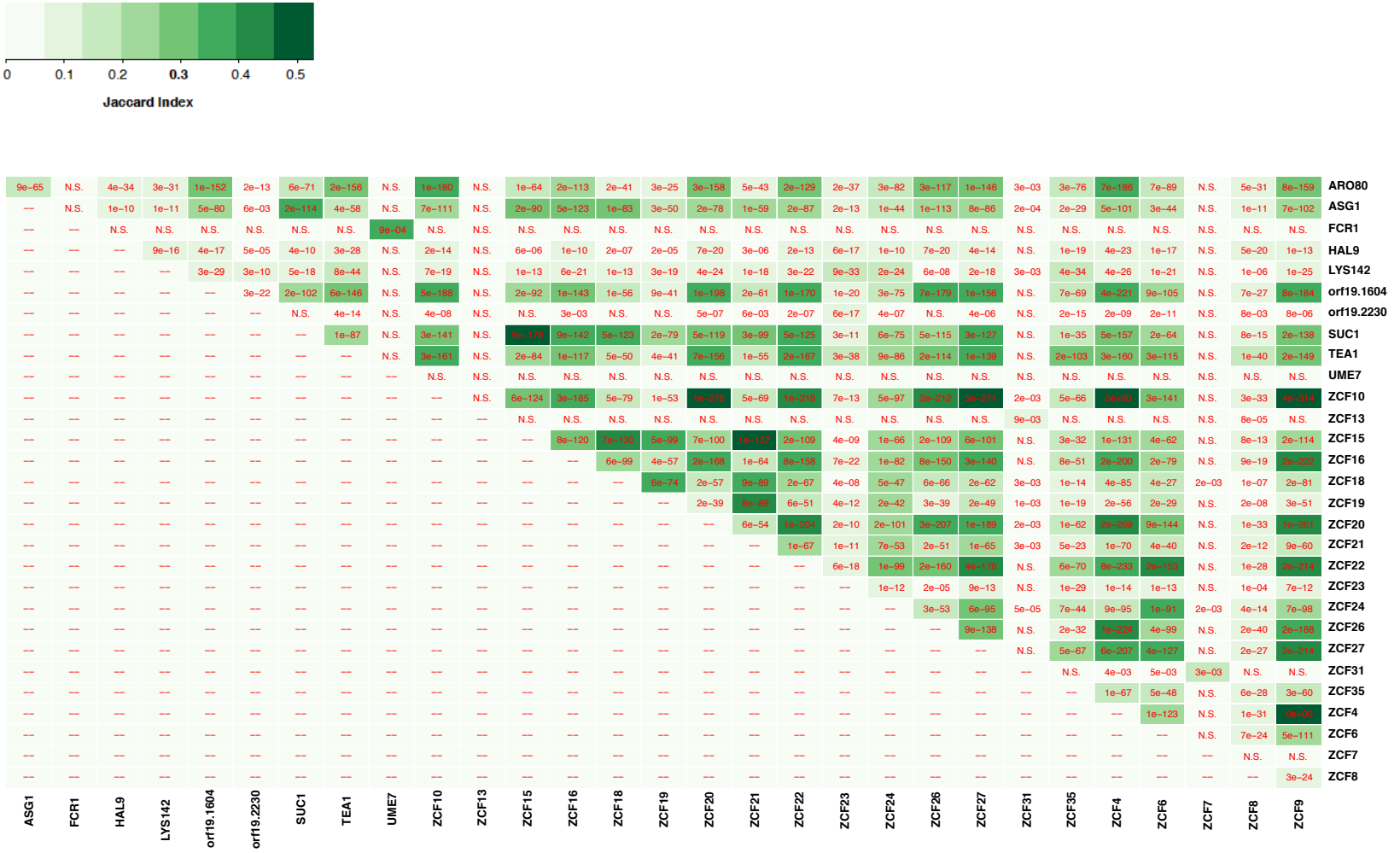



**Figure 12. Fisher's Exact test for significance of overlap between ZCTF profiles.**

Using the list of significantly d.e genes (adjusted p-value < $10^{-4}$) for each ZCTF, we compute the statistical significance of the observe overlap between all pairs of ZCTFs. Red numbers are p-values from Fisher's exact test (N.S. indicates non-significant overlap; p-value < 0.01). Colour key in the heatmap captures similarity between gene sets estimated by the Jaccard index.

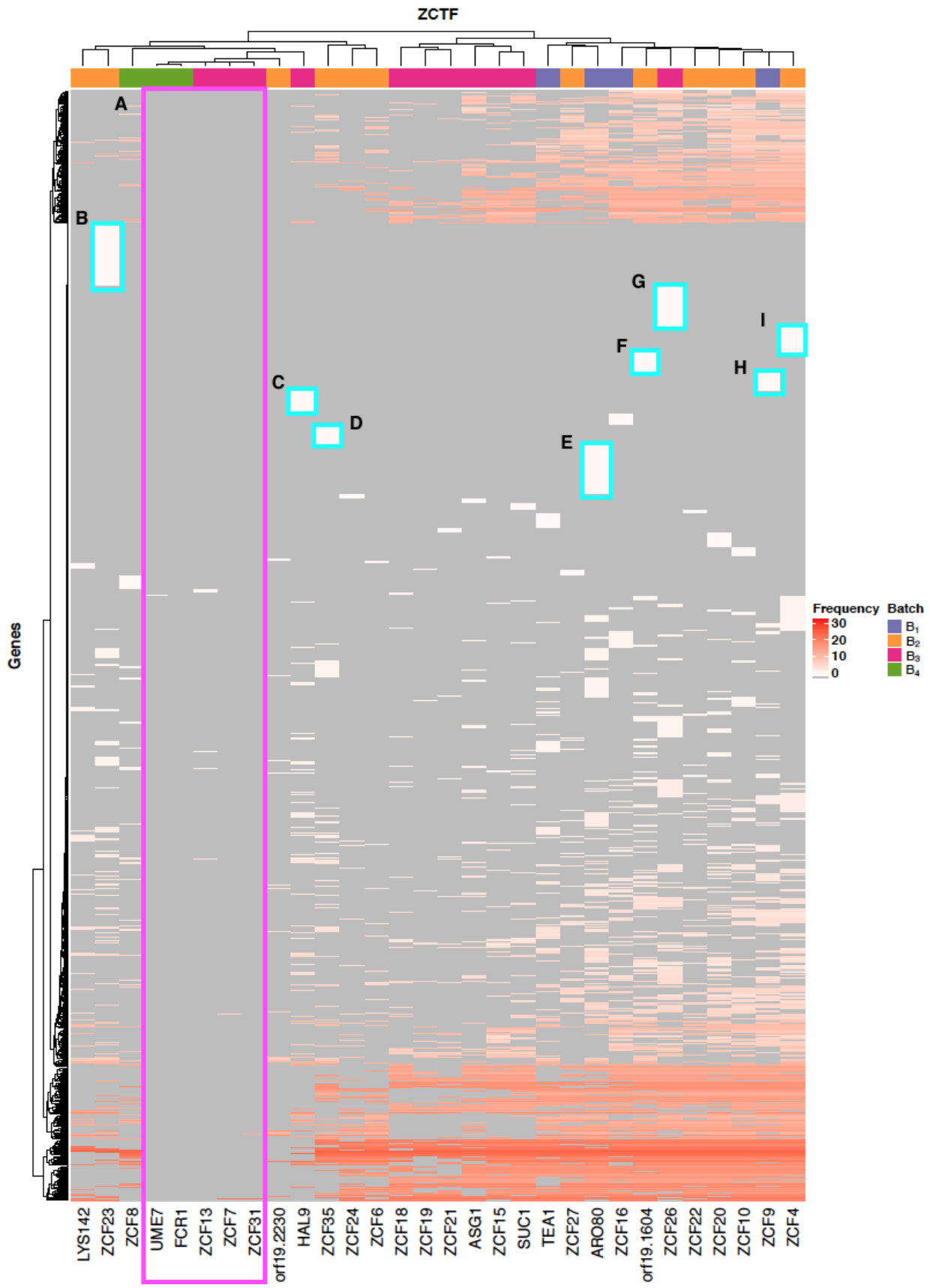

ZCTF
A
B
C
D
E
F
G
H
I
Genes
Frequency
30
20
10
0
Batch
B1
B2
B3
B4
LYS142
ZCF23
ZCF8
UME7
FCR1
ZCF13
ZCF7
ZCF31
orf19.2230
HAL9
ZCF35
ZCF24
ZCF6
ZCF18
ZCF19
ZCF21
ASG1
ZCF15
SUC1
TEA1
ZCF27
ARO80
ZCF16
orf19.1604
ZCF26
ZCF22
ZCF20
ZCF10
ZCF9
ZCF4

**Figure 13. A catalog of uniquely and near uniquely differentially expressed genes per ZCTF.** Columns correspond to ZCTFs and rows correspond to genes that are differentially expressed (versus control) in at least one ZCTF (DESeq2, adjusted p-value < $10^{-4}$). Grey denotes that the gene is not differentially expressed in the respective TF. Colour denotes the number of TFs for which the gene is differentially expressed.

*ARO80*, which has a large number of unique d.e. genes (cluster E, n=91), induces strong over expression of orf19.1473 (a 2-hydroxyacid dehydrogenase domain-containing protein) and ALD6 (an aldehyde dehydrogenase). ARO80 over-expression strongly down-regulates HIP1 (an amino acid permease), FET1 and two ZCTFs (ZCF4, ZCF1). There is significant up-regulation of several arginine biosynthesis genes including ARG1, CPA1/2, ARG5,6 and ARG8. Many of these and other arginine related genes are also d.e. in other ZCTFs including orf19.1604, orf19.2230, HAL9, LYS142, TEA1, ZCF23 and ZCF35.

*ZCF23* has the largest number of uniquely expressed genes (cluster B, n=112), although many of the over-expressed genes are poorly characterized, there is an enrichment for transporters and both zinc finger and zinc cluster TFs (incl. TRY5 and SUC1 which are both implicated in cell adherence, RME1 involved in fluconazole resistance, CAT8 which regulates the diauxic shift in *S. cerevisiae*, CWT1 involved in regulating the nitrosative stress response, HAL9 involved in acid response, and ZCF29 a hypocolonizer in host organs (Vandeputte *et al.* 2011)).

*HAL9* (cluster C, n=39) uniquely upregulates ADH3 (7.5 fold), several genes implicated in the white-opaque switch (PTH2, CZF1 and RPD3), genes involved in the oxidative stress response (e.g. SOD3, YCP4, RIM2, PST3 and OYE32) and in zinc sequestration from host (ZRT1, PRA). It also up-regulates ZCF31, a member of our ZCTFs. Genes that are down-regulated (incl. SOD1, CYS3, orf19.6245, SOD2, HSP60, and HSP21) appear generally to be involved in the heat shock and oxidative responses. So, genes that are involved in oxidative responses are both up and down regulated by this TF.

*ZCF4* (cluster I, n= 46) upregulates zinc clusters TFs ZCF3 and ZCF32 that are not amongst our 30 profiled ZCTFs, in addition to several other transcription factors (NSA2, PZF1, orf19.2528, orf19.1589, ADR1). We observe strong down-regulation of ZRT2 (-4 fold) which is essential for zinc uptake. Interestingly, we observe down-regulation of several genes involved in arginine biosynthesis incl. ARG5/6 and ARG8.

*ZCF9* (cluster H, n=14) causes very high over-expression of several genes but these remain poorly characterized orfs. It down-regulates several genes involved in generic stress responses and oxidative reduction.

*ZCF26* (cluster G, n=74) upregulates several genes involved in the regulation of DNA damage including orf19.6907, MMS22 and GIN1. It also upregulates genes in filamentous growth including GIR2, CLA4 and BRE1.

*ZCF8* (n=24) strongly upregulates DAL5, an allantoate transmembrane and dipeptide transporter, and other transporters (incl. orf19.1308, GDT1, HGT10, OPT8 and CCC1, QDR3, FRP3, YOR1, orf19.6976, OPT1, SUL2 and CTP1).

*ZCF10* (n=16) up-regulates several zinc finger TFs including ZCF20 (profiled here), MRR1 (not profiled here; regulates expression of the multidrug resistance gene MDR1), ROB1 (also not profiled here), and UGA3 (utilization of gamma-aminobutyrate). ZCF39 is down-regulated by ZCF10.

Approximately two-thirds of the ZCTFs have few or no unique d.e. gene. In some cases, such as SUC1, the few unique genes are strongly d.e. and belong to the same pathway. For example, both MAL2 (hydrolyzes sucrose) and MAL31 (transports maltose) are known to be regulated by SUC1 and are over-expressed in gain-of-function mutant of this ZCTF. In other cases, there is evidence of strong transcriptional changes but involve poorly characterized genes (e.g. TEA1 up-regulates by as much as 6-fold several orfs in addition to weaker up-regulation of arginine biosynthesis genes). Most severely, the ZCTFs in the cluster A of **Figure 11** have very few (from one to three) d.e. genes, suggesting the transcriptional response of the gain-of-function mutation is broadly the same as control. We investigate the apparent lack of activity next for these and other ZCTFs belong to sample cluster $T_1$.

**Low complexity transcriptional profiles from sample cluster $T_1$ revisited**

In **Figures 9** and **10**, samples from the $T_1$ cluster generally have dampened gene expression profiles and few d.e. genes. This is particularly true for ZCF7, ZCF13, ZCF31, UME7 and FCR1. We revisited the replicate expression profiles to identify potential technical biases. **Figure 14** provides PCA plots for these problematic ZCTFs (see **Methods 4**). For ZCF7, UME7 and FCR1, the two replicates segregate, suggesting that there are significant differences in the underlying gene expression patterns. We note that there was second sequencing run for ZCF7

due to a failure in the first trial (see **Table 1** and caption). No such aberrations were observed for the remaining low-complexity profiles.

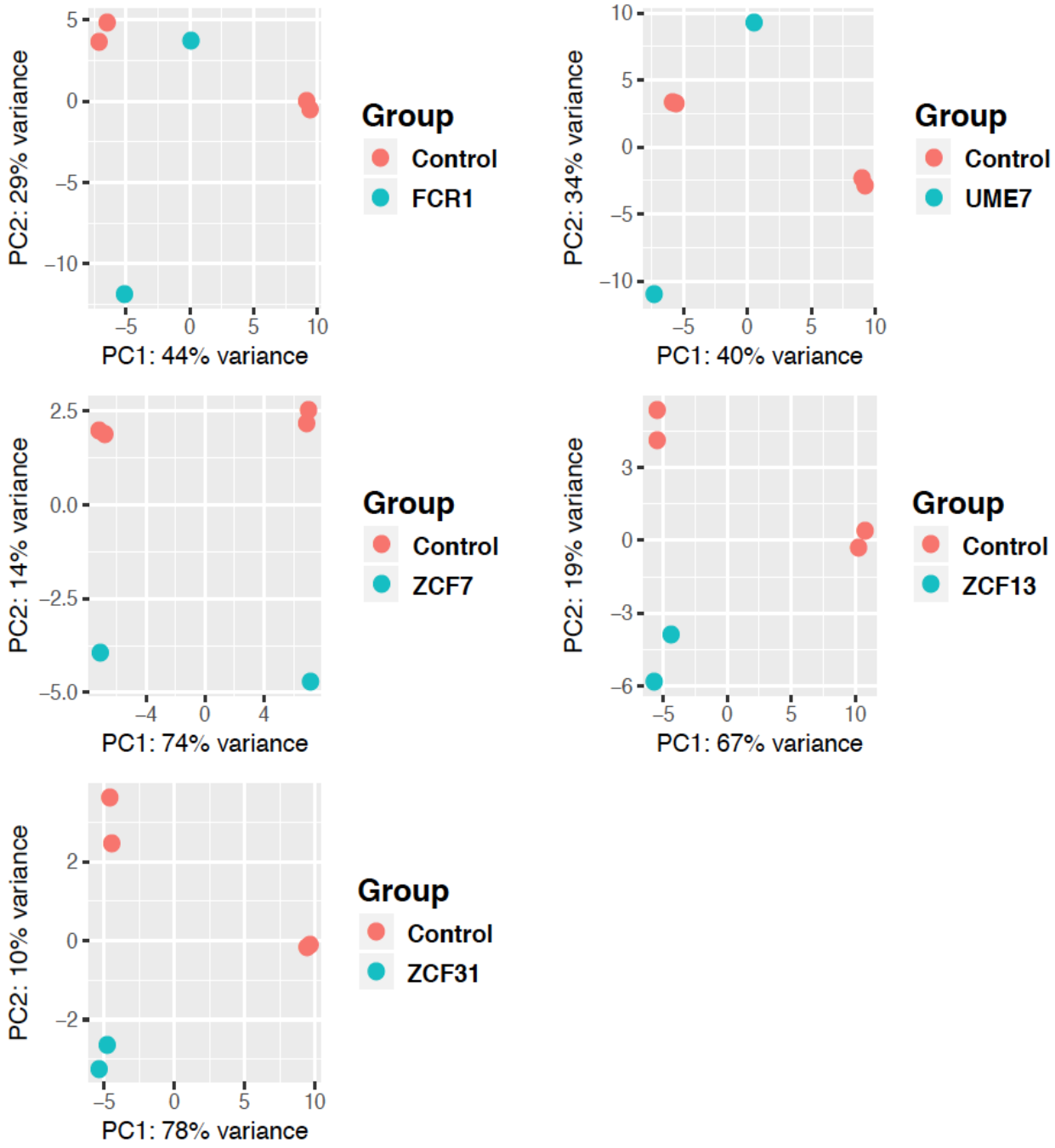



**Figure 14. PCA plots of low complexity transcriptional profiles from cluster $T_1$.**
Red dots correspond to four control replicates (from batches 3 and 4). Blue dots correspond to replicates of the target ZCTF.

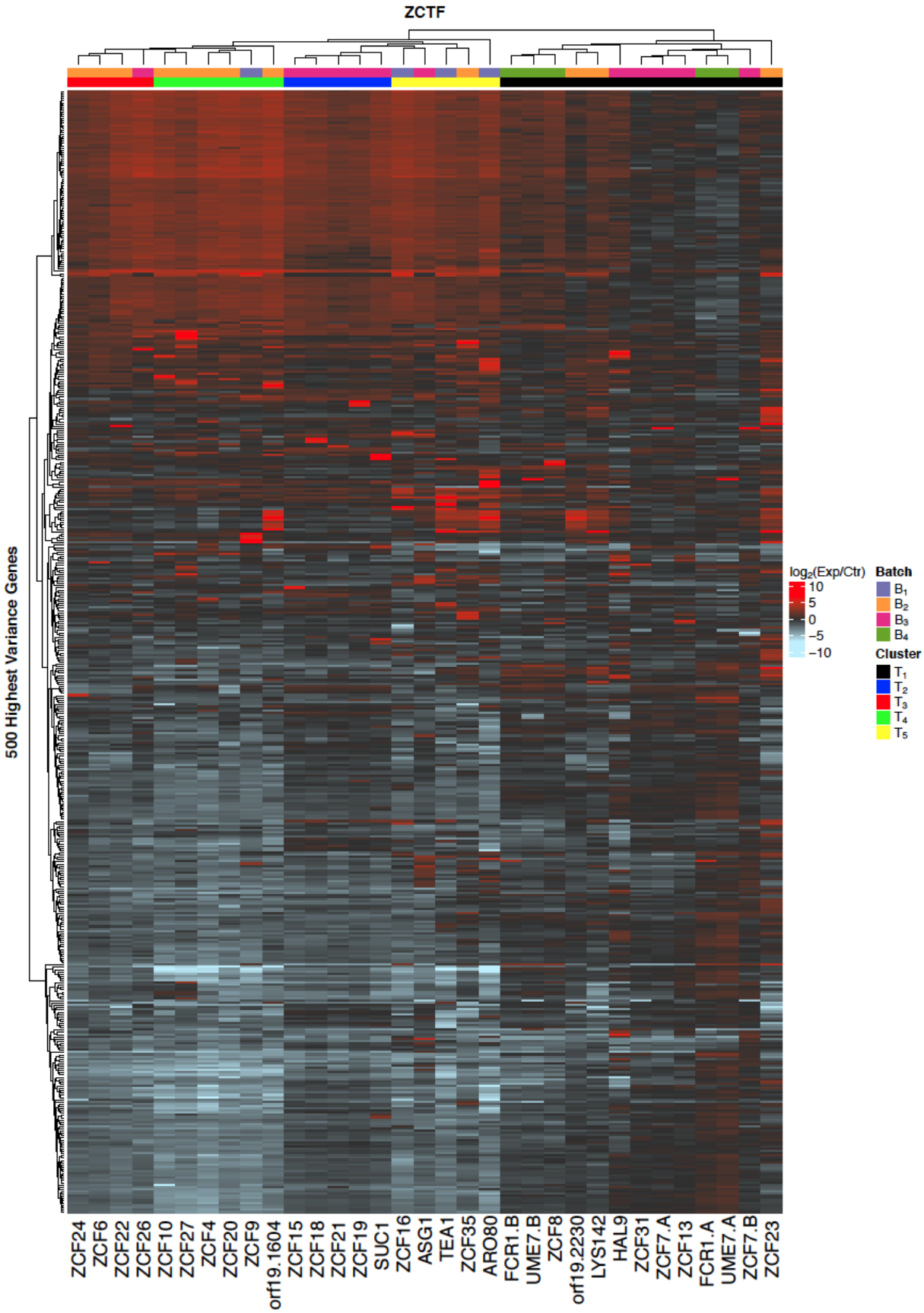

ZCTF
500 Highest Variance Genes
ZCF24
ZCF6
ZCF22
ZCF26
ZCF10
ZCF27
ZCF4
ZCF20
ZCF9
orf19.1604
ZCF15
ZCF18
ZCF21
ZCF19
SUC1
ZCF16
ASG1
TEA1
ZCF35
ARO80
FCR1.B
UME7.B
ZCF8
orf19.2230
LYS142
HAL9
ZCF31
ZCF7.A
ZCF13
FCR1.A
UME7.A
ZCF7.B
ZCF23
log2(Exp/Ctr)
10
5
0
−5
−10
Batch
B1
B2
B3
B4
Cluster
T1
T2
T3
T4
T5

**Figure 15. A global portrait of the ZCTF transcriptional response with special focus on low-complexity ZCTF profiles.**

As per Figure 9, except that the two replicates for UME7, ZCF7 and FCR1 are plotted separated (labelled A and B).

It is difficult in an absolute sense to determine which, if either, of the two replicates for each of these ZCTFs represents a better profile. However, we reasoned that if one of the two replicates showed more expression in an analogue of **Figure 9**, more d.e. genes in the re-analysis of **Figure 10**, and if processes these genes had significant overlap with the genes and processes of other ZCTFs, then we would retain only the more informative profile. Towards this end, **Figure 15** is analogous to **Figure 9** but replicates of these two genes are treated individually. By the criteria stated above, the replicates labelled B appear more informative. **Supplemental Table 4** provides the unique and near unique d.e. genes for the A and B replicates (separately) versus control. For FCR1, we observed more d.e. genes for replicate B (n=79 unique or near unique) than replicate A (n=19 unique or near unique). Since little is known regarding the function of FCR1, it is difficult to make any conclusion. However, four of the highest over-expressed FCR1.B genes are transporters (nicotinic, ammonium, oligopeptide) and some of the most under-expressed FCR1.B genes are chaperones or related to drug response and cellular stress. There is little over-expressed in FCR1.A and it is difficult to observe enrichment in the small number of under-expressed genes.

**Towards the construction of a regulatory network: direct transcriptional co-expression between ZCTFs**

Although we would expect that the transcript for each ZCTF would be over-expressed in the transcriptional profile of its respective gain-of-function mutant, we asked if the transcript levels for other members of the ZCTF were also affected. Such direct co-expression (either positive or negative) might indicate pairwise regulatory relationships that could in turn be used to define a regulatory network. **Figure 16** depicts the transcript expression (versus control) of each member of the ZCTF (rows) across each gain-of-function ZCTF mutant (columns). As expected the diagonal elements are above a log2 fold increase with the exception of ZCF27.

For ZCF27, although its transcript levels are not significantly altered, there is significant over-expression of ZCF6, 35, and 4, and under-expression of LYS142. Several explanations are

possible. If we postulate that the gain-of-function mutants failed, we would still require an explanation as to why we nevertheless observe strong differential expression for the other ZCTFs (that is, the gain-of-function mutant does seem to have perturbed the expression of several genes). Multiple testing/false discovery is a possible but unlikely explanation, since we are examining only 30 TFs and a log2 fold change is rare in the data. One possibility is that the endogenous promoter of ZCF27 is approximately as "strong" as the ADH1 promoter used in the construction of the ZCTFs, however there are other activating modifications to ZCF27 that are not related directly to transcriptional activation. Although we are unable to resolve this issue with gene expression data below, this case appears non-canonical and should be investigated in more depth.

**Towards the construction of a regulatory network: indirect transcriptional co-expression between ZCTFs**

We observe strong differences in expression in several off-diagonal entries between individual TF profiles and transcripts of the ZCTFs in **Figure 16**. For example, FCR1 has transcript levels more than 2-fold under-expressed in many of the ZCTF mutants (incl. ZCF4, 10, 6, 20, TEA1, orf19.1604). Although less pronounced, ZCF24 is also consistently under-expressed across many of the ZCTFs. Conversely, ZCF35, ZCF4 and ZCF19 are consistently over-expressed across many of the ZCTFs or show neutral change in the remaining ZCTFs.

To better identify potential regulatory relationships between members of the ZCTF family, we constructed a graph to catalogue events where ZCTF mutants cause over- and under-expression of the ZCTFs (**Figure 17**). Here an arc from X to Y indicates that the gain-of-function ZCTF mutant X over-expressed the transcript for ZCTF Y, suggesting a possible regulatory role between X and Y. The minimal change in transcript level was a log2 fold change of |2| with red and blue arcs corresponding to over- and under-expression respectively.

**Figure 17** suggest at least three interesting findings. First, FCR1 transcript is down-regulated by several ZCTFs. The only ZCTF that induces over-expression of the transcript is ASG1 (**Figure 16**). We note that FCR1 DNA binding domain is the best hit of the *S. cerevisiae* PDR1 (involved in multidrug resistance) DNA binding domain and the genes that are uniquely or nearly uniquely d.e. include many transmembrane ion pumps (**Supplemental Table 4**). Second, in contrast to FCR1, ZCF4 is strongly up-regulated by several ZCTFs, and no ZCTF causes decreased expression. Third, orf19.1604 appears to exert both positive and negative forces on other ZCTFs including FCR1 and ZCF4.

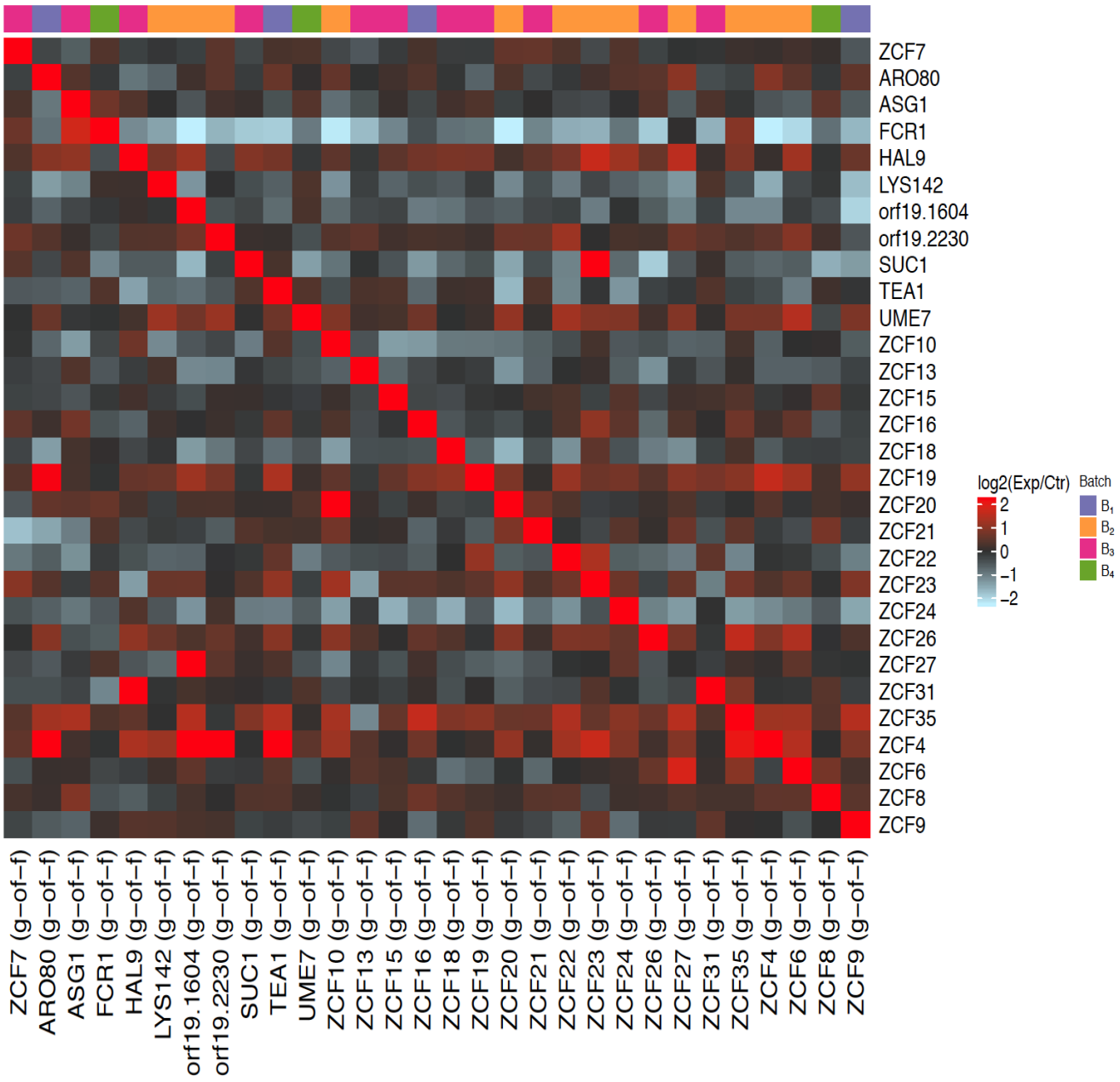



**Figure 16. Direct co-expression between members of the ZCTFs.**

Columns of the heatmap correspond to individual transcriptional profiles for each gain-of-function (g-of-f) ZCTF. Only the expression (log2 fold change versus control) of the 30 ZCTFs is shown (rows). That is, each row corresponds to the transcript of one member from the ZCTFs. For example, the ZCF4 transcript (4th row) is more than two fold over-expressed in the ARO80 gain-of-function mutant (1st column).

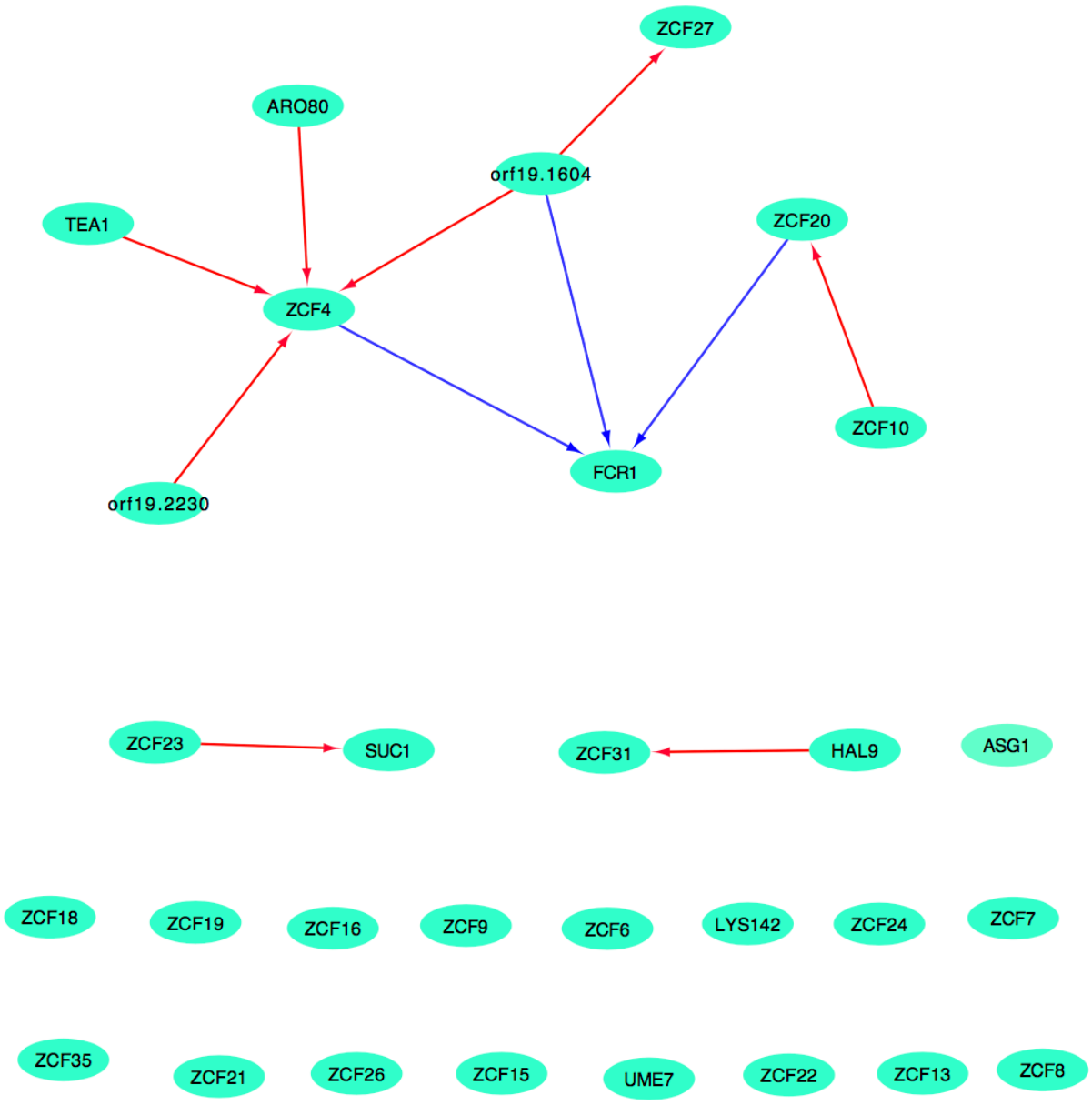



**Figure 17. Regulatory network for the ZCTFs.**

Arrows correspond to log2 fold change |2| between a ZCTF mutant and a ZCTF transcript (red is >2 and blue is < -2).

**A comparison of the ZC DNA binding domains identifies orf19.2230 as an outlier**

Using CGD and the federated domain database InterPro, we were able to identify the ZC DNA-binding domain for all ZCTFs except orf19.2230 (**Methods 8**), but a multiple sequence alignment approach using the 29 identifiable DNA-binding domains with the full length amino acid sequence of orf19.2230 identified a candidate domain where several considered cysteine residues are present with small indels. **Figure 18A** and **B** depict the phylogenetic relationships and multiple sequence alignment of the predicted binding domains for all ZCTFs. We observe that ARO80, ZCF31 and ZCF19 are outgroups in this phylogeny, and orf19.2230 is divergent with ZCF35.

The best BLAST-protein alignment of each ZCTF DNA binding domain was identified in *S. cerevisiae* in order to better identify potential functional orthologs for each ZCTF (**Table 2 column DNA b.d.**, **Methods 8**).

**TF binding site analysis in the promoter of the ZCTFs**

In order to understand if and how individual ZCTFs coregulate, we used phylogenetic footprinting to determine regulatory elements at promoter of each ZCTF. Several conserved motifs have been identified at promoter site of ZCTFs selected from **(Table 4).** The result from phylogenetic footprinting MEME indicates that a SPT23-like binding motif is present in the promoter of most ZCTFs. This TF is involved in regulation of unsaturated fatty acid synthesis. Motif 2 is closest to the binding site of YAP5, an ATP-binding cassette (ABC) protein that is regulated by multidrug resistance TFs. It is present at promoter site of 20 of our ZCTFs.

Several ZCTFs have been shown to have common motifs at their promoters. A particularly interesting one is ZCF16. This transcription factor has different set of common motifs with four different ZCTFs; orf19.2230, orf19.1604, ZCF27 and ZCF35, suggesting possible coregulation of this TF with orf19.2230, orf19.1604, ZCF27 and ZCF35.

There are also many common motifs at promoter site of ZCF35 and orf19.2230, some different from the common set between ZCF16 and ZCF35. This event repeated for ZCF27 and ZCF35, and also orf19.2230 and ZCF35. It has been shown (**Table 4)** that there are 8 motifs common between orf19.2230 and ZCF27, and Orf19.1604 has several common motifs with ZCF35 and ZCF27 but not many with orf192230.

.

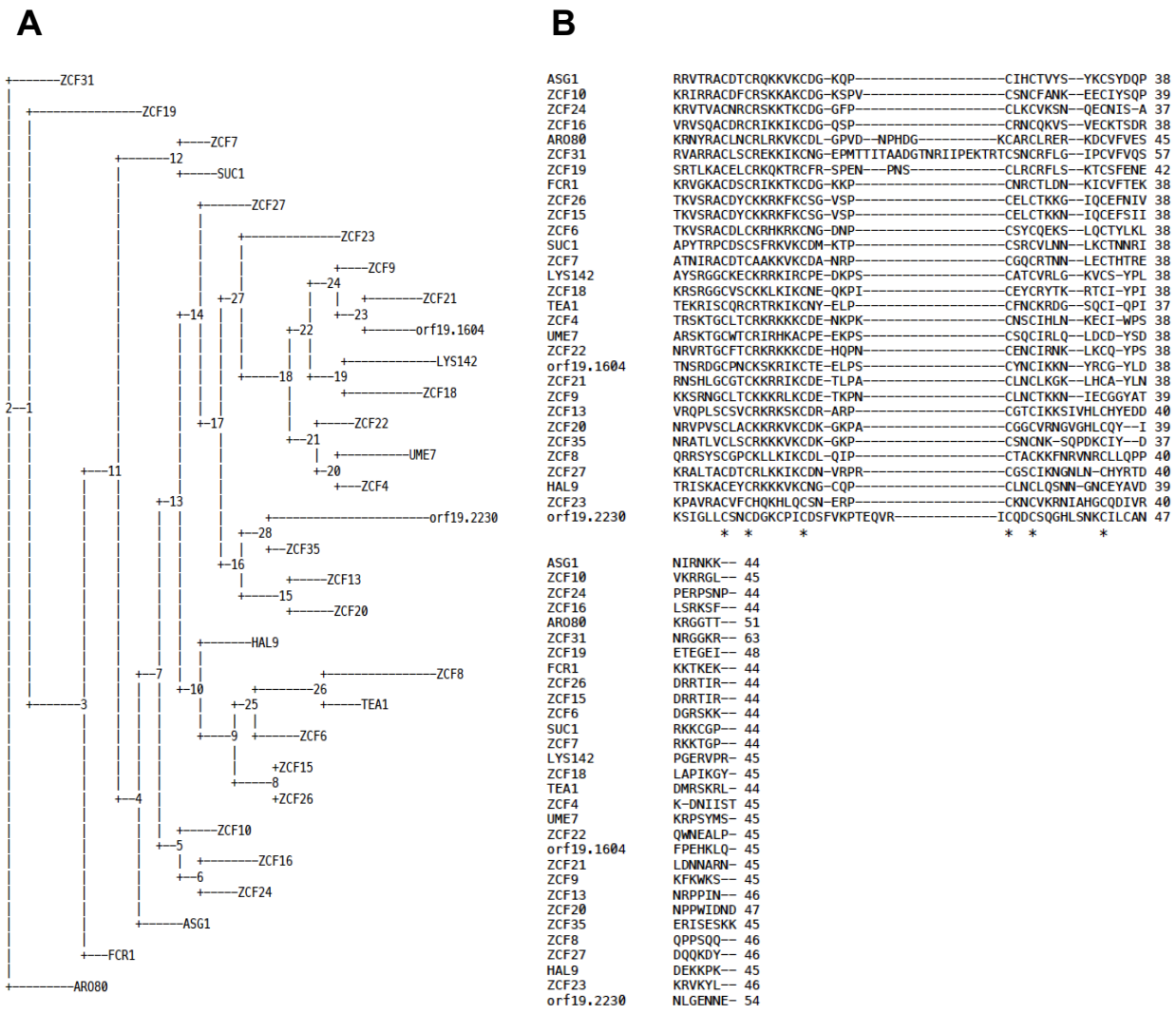



**Figure 18. Multiple alignment of protein sequences of ZCTFs zinc cluster domains.**

Panel A depicts the phylogenetic tree using only the binding domains of each ZCTF member (created with Phylip) and Panel B denotes their multiple sequence alignment (**Methods 8**). Note that the binding domain of orf19.2230 is an outlier. A star (*) denotes perfect conservation of the residue.

There are about 12 common TFBS at regulatory region of ZCF8 and ARO80, suggesting possible coregulation of these two TFs, even though they are functionally very diverged. Presence of 5 common motif at the promoter of ZCF13, ZCF19, ZCF20, ZCF27 and UME7 also suggest possible coregulation between these five TFs

Some of the 30 ZCTFs studied in this report appear to have binding sites at promoter of other ZCTFs. The binding site for ASG1, which is involved in regulation of multi-drug transport and pleiotropic drug resistance, been observed at regulatory regions of FCR1, ZCF8, ZCF16 and ZCF27 (**Table 4).** Between these ZCTFs; FCR1, ZCF16 and ZCF8 are showing to be regulated by ASG1 (**Figure 16),** Indicating potential direct interactions between ASG1 and these three ZCTFs.

In **Table 4,**binding sites for ZCF20 and ZCF7 best hit (i.e. YPR196W) have been identified at promoters of ZCF16 and ZCF35. Binding site for ZCF23 and ZCF28 also been shown to be present at the promoter of ZCF16 (**Table 4)**.

**Towards the identification of the orf19.1604 TFBS motif**

The binding site of orf19.1604 is unknown. We asked if we could use the transcriptional profile for this ZCTF to identify putative TFBSs. HOMER is a bioinformatics tools that identifies over-represented motifs that occur often in a given positive learning set of promoters compared to their frequency of occurrence in a given negative learning set of promoters (**Methods 10**). We reasoned that the genes uniquely d.e. in the orf19.1604 strain would be enriched for the orf19.1604 TFBS and therefore used its unique d.e. genes as the positive learning set (n=38). The negative learning set consisted of 1000 randomly chosen genes under the assumption that the vast majority of *C. albicans* genes are not regulated by orf19.1604 (and therefore the set should be depleted for orf19.1604 TFBSs).

The putative TFBS motifs are enumerated by significance (descending) in **Table 5**. The two most significant have bindings sites resembling PABPC1, a human polyadenylate TF, and RNA, a human RNA binding protein. However, since orf19.1604 is a zinc cluster TF, any HOMER identified motifs with similarity in sequence structure to zinc cluster (or zinc finger) motifs may be of particular interest. In this direction, the 3rd motif is close to the *S. cerevisiae* MSN4, a C2H2 zinc finger TF. It appears in approximately one quarter of all orf19.1604 unique d.e. genes and very few (<2%) of *C. albicans* genes. The 7th and 16th entries are also motifs with similarity to zinc finger TFs in *S. cerevisiae*, and have reasonable accuracies with respect to their occurrence and non-occurrence in the positive and negative learning sets respectively.

**Some of the putative orf19.1604 are present in the promoters of other ZCTFs**

In order to investigate the putative orf19.1604 motifs further, we asked if they appeared in the promoters of the ZCTF members. We once again used phylogenetic footprinting with MEME but here the promoter of each ZCTF was searched for the occurrence of each motif from **Table 5**. Since we observe that the orf19.1604 strongly upregulates ZCF4 (**Figure 17**), we started our investigation with its promoter. The phylogenetic footprint results of **Table 6** suggest that motifs 2, 5, 9, 13, and 15 appear in the ZCF4 promoter; however, none of these motifs are known to correspond to zinc finger/cluster TFs (although we have not investigated this issue thoroughly at this point). ZCF27 is also strongly upregulated by orf19.1604. **Table 7** suggests that motifs 1,7, 10-16 are present. FCR1 is strongly downregulated by orf19.1604 **(Figure 17). Table 8** suggests that motifs 1, 6-9, 11, 12 and 15 are present at promoter site of FCR1. Although none of the three clear zinc finger/cluster TFBS motifs appears in all three, there are several recurrent motifs. For example, the zinc finger motif 7 appears in both FCR1 and ZCF27.

To investigate this more globally, we repeated this analysis for the promoter of each ZCTF (**Table 9**). Motif 15 appeared within the promoters of FCR1, ZCF4 and ZCF27 in addition to other ZCTFs such SUC1, UME7, ZCF-13, 21 and 22) that had high expression in the orf19.1604 mutant (**Figure 16)**. Only motif 16 easily identified as a zinc cluster motif in *S. cerevisiae*. However, we have not exhaustively ruled out the existence of other zinc cluster motifs in the table. This motif has a binding site for orf19.2230, ZCF-9, 10, 16 and 35. Two of these 5 ZCTFs (orf19.2230 and ZCF35) are showing to be regulated by orf19.1604.

**Towards identifying a TF responsible for upregulating the arginine biosythesis pathway**

We observed that ARO80, orf19.1604, orf19.2230, HAL9, LYS142, TEA1, ZCF23 and ZCF35 induced an over-expression of many genes related to arginine biosynthesis. We attempted to decipher which TF might control this process by identifying a motif common to these promoters. Towards this end, we used HOMER with the positive set consisting of these ZCTFs and the negative set 1000 randomly chosen sequences (**Table 11**, **Method 12**). Motifs corresponding to *S. cerevisiae* TFs SFL1 and SFL2 were present in three of these ZCTFs. A motif closest to the pattern bound by PHO2 (involved in phosphate metabolism in *S. cerevisiae*) was also observed in six out of eight of these ZCTFs. Perhaps the most interesting motif is motif 7 that is similar to the zinc cluster binding site for YLR278C.

# Discussion

### Transcriptional profiles of a small set of ZCTFs

We were able to process the Illumina-based RNA-seq profiles for 30 *C. albicans* ZCTFs and ablate the batch effect in the data by using a pooled control approach. Also, we identified several ZCTFs with low complexity transcriptional profiles (FCR1, UME7, ZCF7), and were able to determine the root cause as to why these ZCTFs had so few differences in gene expression. By using only one of two replicates, we were able to identify genes that appeared to be differentially expressed. We noted however that our list of d.e. genes for these ZCTFs ought be handled with caution due to the lack of replicates. However, our preliminary ad hoc analysis of these lists did identify genes that plausibly were regulated in the gain-of-function mutant. For instance, many genes that code for transporters were up-regulated and many genes involved in stress were down-regulated for FCR1; these biologies and the direction of modulation has been observed for other ZCTFs with replicates. In summary, after quality control and normalization, the ZCTF dataset obtained from the Whiteway lab appeared statistically robust and consistent. In combination with our R code that we make available to the community, this data will serve as a useful resource where hypotheses can be tested quickly.

### The Bliss continuum

We observed a large transcriptional pattern involving hundreds of genes. It is apparent from **Figure 9** that the majority of the 500 most variable genes are either strongly positively or negatively correlated with each other across all the ZCTFs. Moreover, TFs ranged in a continuum from very weak expression of this signature (log2(TF / control) ~ 0; cluster $T_1$, black entries) to strong expression of the signature ( |log2( TF/control ) | >> 0; cluster $T_5$; red or blue entries).

We attempted to characterize the genes in this signature and observed that they had significant (combinatorial) overlap with the ESR derived in *S. cerevisiae*. Approximately ⅕ of top 500 genes belong to the ESR but many more have functions related to ribosomal biogenesis and cellular stress responses (heat shock, unfolded protein response, osmotic stress etc.) (**Figure 10**). Importantly here, $T_1$ represent strains that have healthy control levels of the ESR whereas $T_5$ represent strains that are actively suppressing the stress response. None of the ZCTFs induce a stress response. We hypothesize that the gain-of-function strains are in a conceptual “state of bliss” where all bet-hedging mechanisms normally expressed in a healthy wild-type *C. albicans* cell in anticipation of changes in their environment are turned off.

**Some evidence of aneuploidy**

We asked if there was any evidence that the *C. albicans* had genomic instability, a common response occurred when cell is placed under stress. Although we lacked DNA-sequence data and our profiling was done in bulk across thousands of members of a population, we were still able to see some coordinated differentially expression especially along chromosomes 5 where the MAT locus resides. The MAT locus controls white-opaque switch as a precursor to sexual reproduction (Hirakawa *et al.* 2015). Since we could observe general higher expression along all of chromosome 5, this suggested the possibility that the chromosome had been amplified in a significant number of members in the colony. A positive shift (above 0) was observed in approximately one-third of the ZCTFs in chromosomes 2 through 7. Perhaps interestingly, every ZCTF (orf19.1604, ZCF-4,6, 9, 10, 20, 22, 24, 26) that shows aneuploidy on chromosome 5 is a high expresser of the bliss signature (clusters $T_2$ through $T_5$).

**The ZCTF expression compendium identifies many unique or nearly unique genes**

In order to ablate the "blinding" effect of the bliss signature, we developed a method to identify genes that were either uniquely d.e. in a single ZCTF or expressed in just a few ZCTFs. Both statistical pathway analyses (via GO) or ad hoc analyses using the primary literature identified many cases, where the unique d.e. genes were in agreement with the known biologies induced by the specific TF. For example, orf19.1604 over-expresses many genes known to positively regulate hyphal transition in *C. albicans*, and many of the TFs down-regulate genes involved in various stress responses. The gain-of-function mutant also shows filamentation when grown under typically non-filament-inducing conditions. **Supplemental Table 3** (and its extension to the low complexity TFs, **Supplemental Table 4**) are excellent resources to investigate deeper the role of the individual ZCTFs, many of which are poorly characterized. We summarize some of the important findings below.

**FCR1**

The best hit for binding domain of FCR1 in *S. cerevisiae* is PDR1, a regulator of the pleiotropic drug response. Almost every ZCTF strain except ASG1, ZCF35 and FCR1 down-regulated this gene, often with more than a 2-fold decrease (**Figure 16)**. This is consistent with the idea that strains are in a state of "bliss" and down-regulate genes that might provide survival advantage when challenged (**Figure 15**). Consistent with its like role in drug resistance, many of the significant d.e. genes are transporters and channels (**Supplemental Table 3**). It is perhaps

interesting that its closest siblings in the phylogeny of binding domains (SUC1, ZCF-7, 10, 16, 24) tend to all be neutral or repressed across almost all ZCTF domains. ZCF10 also upregulates expression of MRR1, involved in regulation of multidrug resistance gene MDR1. This suggests perhaps that this clade of ZCTFs share common function likely related to drug resistance and cellular stress. In *S. cerevisiae*, PDR1 serves as both a transcriptional activator and repressor by binding to pleiotropic drug response elements (PDREs) present in the promoters of approximately 200 *S. cerevisiae* genes involved in multidrug resistance (Jungwirth and Kuchler 2005). The PDRE consensus sequence is 5'-TCCGCGGA-3'. However, HOMER analysis of the nearly unique d.e. genes from FCR1 (**Table 8**) did not identify a motif close to the PDRE.

**ZCF24**

Like FCR1, ZCF24 is down-regulated by many ZCTF strains (**Figure 16**), but strongly expresses the bliss signature. However, there are few unique or near unique d.e. genes with none of these exhibiting large differences in fold change. The majority of the d.e. genes are under-expressed.

**ZCF4**

Little is known regarding this TF beyond the fact that is located between orf19.2778 (*S. cerevisiae* ortholog is URB1) and HAP2 in *C. albicans*, and although URB1 and HAP2 are syntenically conserved with *S. cerevisiae*, ZCF4 is not. As at least 10 other Candida related species have the ortholog for ZCF4 at the same position we can conclude that S. cerevisiae lost this gene trough speciation. The unique genes d.e. for ZCF4 suggest that it in turns regulates several ZCTFs that were not profiled here. We observe that ZCF4 is in $T_4$, suggesting its over-expression sends *C. albicans* into a high state of bliss (**Figure 15**). Similar to ZCF35, it is almost always upregulated by the ZCTFs (**Figure 16**) and this upregulation is strong with four ZCTFs over-expressing it by 2-fold (**Figure 16**). It in turn down-regulates FCR1 like many other ZCTFs.

**ZCF35**

It is almost always upregulated by the ZCTFs (**Figure 16**) and strongly expresses the bliss signature ($T_5$), although without evidence of aneuploidy (**Figure 11**). It has some unique d.e. genes (**Figure 13D**) including the arginine biosynthesis pathway that is also over-expressed by orf19.1604, orf19.2230, HAL9, LYS142, TEA1 and ZCF23. To resist oxidative stress caused by host immune system *C. albicans* rapidly upregulates arginine biosynthetic genes that induces

yeast to hypha transition (Ghosh *et al.* 2009). Although synteny is generally well-conserved across fungi, there is no syntenic ortholog in *S. cerevisiae*.

**Orf19.1604**

There are few annotations regarding orf19.1604 in the literature. Syntenically, it is located between CEN2 and PMS1, and this triumvirate is generally conserved across fungi but not *S. cerevisiae*. It is the only ZCTF in our study that induces a hyphal transition. In our study here, the gene has a sizeable number of unique and nearly unique d.e. genes (n=359) with several of these genes being known regulators of different aspects of the hyphal switch. This includes over-expression of ECE1, HGC1, WOR3 and BCR1 with down-regulation of NRG1. As with ZCF35 and others, it up-regulates arginine biosynthesis. Orf19.1604 strongly upregulates both ZCF4 and ZCF27. The latter is a TF that appears to regulate filamentous growth. The ZCF27 gain-of-function mutant upregulates a number of transporters and multidrug efflux pumps.

**ARO80**

Genes that are being regulated by this TF are some permeases and aldehyde dehydrogenases. These genes are generally involved in alcohol biosynthesis via Ehrlich pathway. Indicating that perhaps ARO80 has the same function as its ortholog in *S. cerevisiae* and is involved in aromatic acid catabolism.

**UME7 and ZCF7**

We identified several ZCTFs that had low complexity transcriptional profiles including FCR1, ZCF-7, -13, -31, UME7 as they did not express the bliss signature nor had almost any d.e. genes (**Figure 9**, $T_1$). FCR1, UME7 and ZCF7 have clear differences in their replicates, motivating us to exclude only one of the two replicates in downstream analysis. For FCR1, this led to (larger) lists of d.e. genes that looked plausible from what has been reported in the literature regarding this PDR1 ortholog. It remains to determine if insight into UME7 or ZCF7 can be obtained from their modified profiles. UME7 is the *S. cerevisiae* ortholog of UME6, which is considered as a TF involved in the regulation of meiotic genes and pseudohyphal growth. ZCF7 is a positive regulator of filamentous growth in *C. albicans*. It has no obvious syntenic ortholog in other fungal species and induces very few d.e. genes.

**ZCF13, ZCF31 and orf19.2230**

For the low complexity TFs ZCF13 and ZCF31, we are not able to easily identify any differences in their respective replicates such as we did for FCR1. It is difficult to make a conclusion as to their function, when there are so few differences in these profiles (versus control). This is also true to a lesser extent for orf19.2230. The gain-of-function mutant HAL9 has up-regulated ZCF31 (**Figure 16**).

**ZCF23 and SUC1**

The gain-of-function mutation ZCF23 significantly up-regulates SUC1 transcript (**Figure 16**) and is the only member of the ZCTF family to do so. ZCF23 is included in $T_1$, the TF cluster with the weakest expression of the bliss signature (**Figure 15**). Although TFs in $T_1$ tend to have lower complexity transcriptional profiles, ZCF23 actually has the most unique and near unique genes across all ZCTFs (n=123, **Figure 13B**). Close inspection of ZCF23 in **Figure 16** shows that its pattern of expression is consistent with the bliss continuum but unique compared to other members of our ZCTFs. The expression pattern of SUC1 is consistent with other ZCTFs exhibiting moderate low expression of the bliss signature ($T_2$). ZCF23 is the syntenic ortholog of GSM1 (Glucose Starvation Modulator) in *S. cerevisiae*, a gene known to be involved in energy metabolism. SUC1 is involved in sucrose and maltose production, and has a clear syntenic ortholog with MAL13 in *S. cerevisiae*. It is possible that ZCF23, one of only two of our ZCTFs that belong to the ESR, is an important “meta-regulator” of energy production in *C. albicans*.

**ZCF10 and ZCF20**

Only the gain-of-function mutant ZCF10 up-regulates ZCF20 with a fold change exceeding 2 (**Figure 16**). It also weakly down-regulates TEA1 (-1.48 fold). Neither TF induces many unique or near unique d.e. genes (**Figure 13**). It has no clear syntenic orthology in *S. cerevisiae*, but ZCF10 appears to regulate filamentous growth in *C. albicans*. ZCF20 is syntenic with HAP in *S. cerevisiae* regulates response to levels of heme and oxygen. It localizes to both the mitochondrion as well as the nucleus. Both genes are in strong bliss clusters. ZCF20 down-regulates many chaperones, transport elements of the secretory pathway, and several components of the proteasome.

**Analysis of TFBSs identified some direct and indirect interaction between ZCTFs**

Toward the construction of the regulatory network we asked if there is any specific pattern of conserved motifs at promoter site of ZCTFs. Phylogenetic footprinting (via MEME) was used to identify conserved motifs at promoter of these ZCTFs. ZCF35 appears to have common motifs with several ZCTFs (incl. FCR1, ZCF16, ZCF27, orf19.1604 and orf19.2230). It is plausible that the presence of multiple common motifs between promoter of ZCF35 and these ZCTFs can be considered as an evidence for the coregulation and involvement of ZCF35 in their pathways (**Table 4**). With respect to the role of FCR1 in regulation of pleiotropic drug response we could identify pleiotropic drug response as one of these pathways.

It is noteworthy that the TFBS for ASG1, which is also involved in regulation of multidrug transport and pleiotropic drug resistance, is present at regulatory region of FCR1 (**Table 4**). Considering that FCR1 is upregulated by ASG1 (**Figure 16**), we can suggest the existence of a direct interaction between ASG1 and FCR1 (**Supplemental Figure 1**).

The ZCF16 ortholog in *S. cerevisiae* (SIP4) is involved in positive regulation of gluconeogenesis. Presence of many common motifs at regulatory region of ZCF16 and some ZCTFs such as ZCF27, ZCF35, orf19.1604 and orf19.2230 can be an indication for coregulation of ZCF16 with those ZCTFs, and possible involvement of glycogenesis in their pathways. Several ZCTFs such as ASG1, ZCF7, ZCF20 and ZCF23 also reportedly have binding site at ZCF16 promoter (**Table 4**). Considering the slight shift in expression of ZCF16 in gain-of-function mutant of these ZCTF (**Figure 16**), we can suggest potential direct interactions between these ZCTFs and ZCF16.

# Conclusions

After extensive quality control and normalization, the RNA-seq profiles for 30 ZCTFs are robust and identify known biologies for several members. Our analysis here focused on orf19.1604 with sporadic observations of the other 29 ZCTFs. The data set will serve as an excellent tool for further exploration of the ZCTFs. We propose that the vast majority of expression changes here are not direct regulatory relationships between the gain-of-function ZCTF and gene, but are downstream longer-term adaptive responses that may have involved many intermediate TFs. We often observed in our data that the gain-of-function ZCTFs caused changes in the expression levels of other ZCTFs and TFs not profiled here. A more complete catalog of gain-of-

function mutants across all ZCTFs or all TFs would produce a much more refined picture of the regulatory relationships between the ZCTFs.

# Methods

### 1. Construction of the gain-of-function TF strains

Schillig and Morschhäuser identified all *C. albicans* genes containing the canonical Zn2Cys6 DNA-binding motif and cloned them into an expression cassette designed to generate fusion proteins with the HA-tagged Gal4 activation domain. In total 82 putative TFs were identified, and integrated into the genome of *C. albicans* strain SC5314 but flanked by a ADH1 promoter, a construct that induces strong transcriptional expression (Schillig and Morschhäuser 2013). With the exception of ZCF29, all transformed *C. albicans* were viable. For 30 of these gain-of-function (gain-of-function) TF strains, two independent samples (biological replicates representing two independent colonies grown on the same plate derived from the same GAL4-based transformation) were preserved for downstream profiling. Wildtype SC5314 f *C. albicans* were used as control.

### 2. Transcriptional profiling

Total RNA was extracted using the QIAGEN RNeasy minikit protocol, and RNA quality and quantity were determined using an Agilent Inc. BioAnalyzer. Paired-end sequencing (150bp) of extracted RNA samples was carried out at the Quebec Genome Innovation Center located at McGill University using an Illumina MiSeq sequencing platform. The TFs were profiled in four batches with two biological replicates for each target TF. Controls profiles (n=2) were included in batches 1, 2 and 4 (**Table 1**).

### 3. Basic pipeline for processing RNA-seq files

Raw reads were pre-processed with the sequence-grooming tool cutadapt version 0.4.1 (Martin 2011) with the following quality trimming and filtering parameters (`--phred33 --length 36 -q 5 --stringency 1 -e 0.1`). Each set of paired-end reads was mapped against the *C. albicans* SC5314 haplotype A, version A22 downloaded from the Candida Genome Database (CGD) (www.candidagenome.org) using `HISAT2` version 2.0.4. `SAMtools` was then used to sort and convert SAM files. The read alignments and *C. albicans* SC5314 genome

annotation were provided as input into `StringTie` v1.3.3 (Pertea *et al.* 2015) which returned gene abundances for each sample.

**4. Basic statistics, informatics and visualization**

Analysis of the transcriptional profiles were carried out using the R programming language version 3.5.0 (R Core Team 2018) using the R Studio Interactive Development Environment `RStudio` version 1.1.423 (RStudio Team 2016). Principal component analysis (PCA) was performed using the `plotPCA` function in package `DESeq2`. Heatmaps were constructed using `Heatmap` in the `ComplexHeatmap` package (Gu *et al.* 2016). Calculations of the hypergeometric test were conducted using `phyper` function from R. Fisher's exact test was conducted by `newGOM` function from R package `GeneOverlap`. Pairwise overlap was visualised using the `drawHeatmap` function with a cutoff value of 0.01 for adjusted p-value, using the Benjamin-Hochberg method (Shen and Sinai 2013). `Cytoscape` (version 3.6.1) was used to visualize the TF networks (Shannon *et al.* 2003). It was accessed through the `CytoscapeWindow` functions of R package `Rcy3` (version 3.7). The network was plotted by `cyPlot` function of `Rcy3`, where edges were set at log2 fold change of 2 for upregulated and log2 fold change of -2 for downregulated genes by `setEdgeLineStyleRule` function of the package.

**5. Normalization and statistical models**

To ablate the effect of missing values, number 10 was added to all the values in matrix of read counts. Although there are many different normalization techniques available (eg CPM, TPM, RPKM/FPKM, EdgeR), we opted to use the so-called *median of ratios* approach from `DESeq2`, since this approach is particularly suited for cross-sample comparisons. For instance, the approach does not normalize for gene length but does normalize across the total number of reads generated per sample. The approach also normalizes for bias in the RNA composition of transcripts, a property that may affect the propensity for a transcript to be sequenced. First, the row-wise (gene-wise) geometric mean is calculated for each gene, referred to as a pseudo-reference:

```
pseudo_g <- sqrt( prod(X[g,] )
```

where X is the gene by sample count matrix. The normalized expression for gene g in sample is then X[g,s]/pseudo_g. Second, normalization between differences in counts across samples is addressed by computing the column medians. For example, the computation for sample i is:

```
normalization_factor_i <- median( X[,i] ).
```

Finally, the normalized counts for gene g in sample i are formed as follows:

```
normalized_X[g, i] <- X[g,i] / normalization_factor_i.
```

The resultant normalized was log2-transformed.

**6. Supervised analyses: differential expression**

We developed various statistical models in the `DESeq2` package (Love *et al.* 2014) in R to identify genes differentially expressed between each g-of-f TF profile versus control. The `DESeq2` package tests for differential expression by use of negative binomial generalized linear models, estimating p-values using Wald's test, and use Independent hypothesis weighting (Ignatiadis and Huber, 2017) to adjust p-values for multiple testing. Design matrices are used to define the two classes (non-overlapping subsets of samples) that are to be contrasted. Several design matrices were used especially in the quality control analyses to compare the effects of different control samples on the profiles.

**7. Unsupervised analysis: clustering**

Features (genes) were selected from (normalized, log2-transformed) count matrices using one of two approaches for feature selection. The variance-based method ranks all genes in descending order by their variance across the set of samples. Then, the first k elements of this order are selected. Alternatively, we use the interquartile range (IQR) using cutoff 1.5 (genes were selected with IQR > 1.5). The IQR for a gene is the absolute value of the difference between the median of the third and first quartile. Hierarchical clustering (of TF samples or transcripts) was used with Euclidean distance and complete linkage (`hclust`; R Core Team 2018). The resultant trees were partitioned using the `hcut` function from the `factoextra` package) (Kassambara and Mundt 2017).

**8. *S. cerevisiae* : *C. albicans* orthology**

We downloaded the *C. albicans* (strain SC531) assembly 21 and *S. cerevisiae* (S288C) homology maps from CGD. Homology mapps are from Candida Gene Order Browser (CGOB), where they were manually curated based on sequence similarity and synteny (Fitzpatrick *et al.* 2010). The mapping includes two categories (orthologs and best hits). Here the estimations of orthology were downloaded from the curated lists at CGD. They have generally been decided

based on an algorithm that considers conserved synteny. When orthology maps were not determine for a gene, we used best hit of the *C. albicans* gene in *S. cerevisiae* using BLAST-protein.

In our analysis of DNA binding sites, we first identified the DNA binding domain of each ZCTF using `InterProScan` version 5.32-71.0. Then, `BLAST-protein` was used to align this amino acid sequence (~40 aa) against each *S. cerevisiae* gene (BLOSUM 80 matrix). The longest alignment (highest coverage), with least gap, was reported.

**9. Investigations of aneuploidy using the ZCTF expression profiles**

We mapped the expression of each gene to its location along the chromosomes of *C. albicans* for each ZCTF profile in order to find evidence that chromosomes, or segments of chromosomes, have been amplified or lost. Chromosomal position for each gene was obtained from CGD and visualizations were created using the R package `ggplot2` (Wilkinson 2011).

**10. Gene enrichment analysis via the Gene Ontology**

The Gene Ontology (GO) Annotation file was obtained from CGD and used to map genes via the `GeneSetCollection` function from the `GSEABase` package (Morgan *et al.* 2018). Statistical enrichment analysis was carried out using a hypergeometric test `hyperGTest` from the `GOstats` package at cut off value of 0.05 (Falcon and Gentleman 2007). The `p.adjust` function was used to adjust the p-value below 0.001.

**11. TF motif analysis in a given promoter: phylogenetic footprinting**

In addition to the *S. cerevisiae*:*C. albicans* orthology map (**Methods 8**) we downloaded from CGD the orthology maps across many Candida species and other fungi. Using CGD for the Candida species and SGD for *S. cerevisiae*, we collected the 1Kbp region upstream of the transcription start site for our gene interest. A multiple sequence alignment was constructed across the 1kbp orthologous sequences for the genes across the species using `ClustalW` (`clustalw2` function, version 2.1 (Larkin *et.al.* 2007). We used default parameters except the *speedier rough analysis* parameter was set to off; the parameter controls the accuracy of the guide tree construction via `Phylip` version 3.695 (`dnaml` function for Nucleotide sequences and `proml` for protein tree construction). `ClustalW` outputs an estimate of the phylogenetic tree (a guide tree) generated by the `Phylip` package. We also downloaded the non-redundant collection

of known motifs for TF binding sites from the JASPAR database in the MEME format. The ClustalW alignment, guide tree and TF binding sites are then input to the motiph function in the MEME Suite version 5.0.2 (Bailey *et al.* 2009). The motiph function identifies subregions of the multiple sequence alignment that show statistically significant conservation of any of the motifs down JASPAR. A restricted maximum likelihood model is used to measure conservation (Felsenstein 1981) with a Benjamini-Hochberg adjustment for multiple testing. A q-value of 0.05 was used.

**12. TF motif analysis**

*De novo* motif discovery was performed on the d.e. genes (adjusted p-value $<10^{-4}$) unique to each ZCTF. The intuition is that this set of genes could be enriched for direct targets of the ZCTF (since no other ZCTF studied appears to perturb their expression significantly). For each such set of target d.e. genes, we created a FASTA file containing the 1Kbp upstream sequence from the transcription start site (ie the promoter). Similarly, we created a FASTA file containing the 1Kbp upstream sequence from 1000 randomly selected genes excluding the d.e. gene set above. Hypergeometric Optimization of Motif EnRichment (HOMER) was used to discover conserved motifs in target sequences (Heinz *et al.* 2010) (findMotifs.pl, default parameters). HOMER normalizes data for GC-content to avoid bias from CpG Islands. It also normalizes for imbalance in sequence content caused by codon-bias and experimental bias in A-rich stretches. Motif enrichment is estimated by zero or one occurrence per sequence (ZOOPS) scoring on target vs. background sequences based on maximum likelihood (Jiang *et al.,* 2013).

**13. Data access**

R code developed for the project is available as a github repository at https://hallettmichael@bitbucket.org/hallettlab/zctf.git.

# Tables

| Batch # | Ctro | Sample | Replicate | # Reads | # aligned reads to genome | % mapped & paired | Average length | Average quality | Average Insert Size | Insert size STD |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | SC5314 | A | 976601 | 1223110 | 94.88% | 139 | 37.4 | 221.7 | 65.5 |
| | | | B | 928824 | 1164063 | 94.63% | 138 | 37.3 | 220.9 | 66.1 |
| | | ARO80 | A | 1002687 | 1291254 | 93.18% | 138 | 37.4 | 218.6 | 62.4 |
| | | | B | 1096594 | 1407407 | 93.89% | 139 | 37.2 | 220.5 | 63.9 |
| | | TEA1 | A | 1124788 | 1370150 | 94.87% | 140 | 37.3 | 221.5 | 64.7 |
| | | | B | 989984 | 1210253 | 94.19% | 140 | 37.1 | 220.0 | 63.0 |
| | | ZCF16 | A | 1048957 | 1337705 | 94.63% | 138 | 37.3 | 221.3 | 66.6 |
| | | | B | 1143386 | 1401965 | 95.26% | 140 | 37.3 | 223.5 | 67.5 |
| | | ZCF9 | A | 808662 | 979708 | 95.22% | 140 | 37.4 | 222.0 | 66.1 |
| | | | B | 1067756 | 1371549 | 94.45% | 139 | 37.4 | 222.1 | 66.0 |
| 2 | NA | orf19.1604 | A | 1366828 | 1687066 | 95.02% | 139 | 37.6 | 219.3 | 62.9 |
| | | | B | 822489 | 1053218 | 94.25% | 137 | 37.6 | 219.0 | 63.6 |
| | | orf19.2230 | A | 731410 | 906384 | 95.20% | 139 | 37.5 | 222.2 | 65.4 |
| | | | B | 945146 | 1153786 | 94.49% | 138 | 37.6 | 221.3 | 65.8 |
| | | LYS142 | A | 781633 | 977841 | 94.62% | 139 | 37.5 | 220.4 | 65.8 |
| | | | B | 801423 | 997146 | 94.84% | 138 | 37.6 | 221.3 | 67.2 |
| | | ZCF10 | A | 1006178 | 1241005 | 95.21% | 140 | 37.6 | 221.6 | 65.5 |
| | | | B | 760891 | 951136 | 94.77% | 139 | 37.5 | 222.1 | 66.1 |
| | | ZCF20 | A | 626398 | 787690 | 94.96% | 139 | 37.6 | 220.3 | 64.0 |
| | | | B | 932542 | 1148388 | 95.23% | 139 | 37.4 | 219.8 | 64.1 |
| | | ZCF22 | A | 780488 | 943743 | 95.27% | 139 | 37.6 | 222.4 | 67.4 |
| | | | B | 703482 | 846287 | 95.06% | 139 | 37.4 | 220.2 | 66.4 |
| | | ZCF23 | A | 798982 | 1002895 | 94.09% | 139 | 37.6 | 219.6 | 61.8 |
| | | | B | 1387365 | 1684871 | 94.41% | 139 | 37.6 | 220.3 | 63.1 |
| | | ZCF24 | A | 1142695 | 1388371 | 95.30% | 139 | 37.6 | 220.2 | 65.9 |
| | | | B | 819386 | 1032431 | 95.09% | 139 | 37.5 | 220.3 | 66.7 |
| | | ZCF27 | A | 613128 | 769178 | 94.74% | 139 | 37.6 | 223.1 | 67.1 |
| | | | B | 1185356 | 1433843 | 95.10% | 140 | 37.6 | 224.2 | 67.8 |
| | | ZCF35 | A | 800463 | 984414 | 94.82% | 139 | 37.6 | 219.5 | 64.7 |
| | | | B | 1336746 | 1648681 | 94.83% | 139 | 37.6 | 223.4 | 68.1 |
| | | ZCF4 | A | 968193 | 1219530 | 94.70% | 139 | 37.6 | 220.6 | 64.4 |
| | | | B | 918653 | 1148822 | 94.81% | 139 | 37.6 | 221.9 | 65.2 |
| | | ZCF6 | A | 575900 | 709362 | 94.81% | 139 | 37.6 | 221.0 | 64.9 |
| | | | B | 708292 | 871813 | 94.65% | 138 | 37.5 | 220.0 | 64.5 |

| Batch # | Ctro | Sample | Replicate | # Reads | # aligned reads to genome | % mapped & paired | Average length | Average quality | Average Insert Size | Insert size STD |
|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 3 | SC5314 | A | 2139161 | 1432961 | 96.97% | 133 | 37.6 | 212.6 | 59.5 |
| | | | B | 2133239 | 1420702 | 96.91% | 135 | 37.7 | 214.3 | 59.9 |
| | | ASG1 | A | 2059551 | 1362084 | 96.93% | 135 | 37.7 | 216.6 | 62.8 |
| | | | B | 2078470 | 1366282 | 96.93% | 135 | 37.7 | 216.4 | 62.4 |
| | | HAL9 | A | 1868168 | 1183563 | 96.70% | 138 | 37.6 | 225.5 | 69.1 |
| | | | B | 1856763 | 1164114 | 97.21% | 139 | 37.7 | 227.1 | 70.0 |
| | | SUC1 | A | 2016800 | 1326439 | 96.97% | 135 | 37.7 | 218.5 | 64.6 |
| | | | B | 2010195 | 1343427 | 96.52% | 134 | 37.7 | 218.0 | 64.9 |
| | | ZCF13 | A | 1937024 | 1295015 | 96.83% | 133 | 37.7 | 215.3 | 61.6 |
| | | | B | 1789998 | 1193186 | 97.00% | 133 | 37.7 | 212.8 | 59.7 |
| | | ZCF15 | A | 1742208 | 1160075 | 97.01% | 133 | 37.6 | 213.8 | 61.2 |
| | | | B | 2088533 | 1386319 | 96.98% | 134 | 37.6 | 215.2 | 62.4 |
| | | ZCF18 | A | 2009768 | 1306154 | 97.01% | 135 | 37.6 | 220.7 | 67.2 |
| | | | B | 2069304 | 1331604 | 97.24% | 136 | 37.8 | 224.6 | 69.9 |
| | | ZCF19 | A | 1857986 | 1242159 | 96.94% | 133 | 37.7 | 213.0 | 60.1 |
| | | | B | 1638066 | 1104791 | 96.93% | 133 | 37.6 | 214.6 | 61.7 |
| | | ZCF21 | A | 1771358 | 1161696 | 96.84% | 135 | 37.5 | 217.2 | 63.8 |
| | | | B | 1909846 | 1246833 | 97.10% | 135 | 37.7 | 219.4 | 65.4 |
| | | ZCF26 | A | 1627453 | 1083430 | 97.04% | 133 | 37.7 | 213.9 | 61.2 |
| | | | B | 1745479 | 1156668 | 97.02% | 133 | 37.5 | 213.4 | 60.8 |
| | | ZCF7* | A | 2092046 | 1345585 | 96.02% | 138 | 37.7 | 227.7 | 71.2 |
| | | | B | 1892163 | 1244314 | 96.81% | 135 | 37.7 | 216.5 | 63.2 |
| | | ZCF31 | A | 2115321 | 1396307 | 96.92% | 135 | 37.7 | 215.0 | 60.8 |
| | | | B | 1880470 | 1257531 | 96.70% | 134 | 37.5 | 213.7 | 60.7 |
| 4 | 4 | SC5314 | A | 2335105 | 1542349 | 91.63% | 140 | 36.2 | 219.9 | 61.5 |
| | | | B | 1965842 | 1301818 | 93.05% | 138 | 36.3 | 216.6 | 60.2 |
| | | FCR1 | A | 2080586 | 340011 | 93.23% | 141 | 36.6 | 219.7 | 60.5 |
| | | | B | 2182140 | 1417342 | 92.58% | 140 | 36.0 | 217.9 | 60.3 |
| | | UME7 | A | 3699435 | 2415697 | 93.05% | 141 | 36.4 | 220.8 | 60.9 |
| | | | B | 2231555 | 1458833 | 93.58% | 140 | 36.4 | 219.1 | 60.7 |
| | | ZCF8 | A | 3375776 | 183977 | 93.45% | 141 | 36.4 | 220.4 | 61.3 |
| | | | B | 2162897 | 402908 | 93.55% | 140 | 36.4 | 218.4 | 60.4 |

**Table 1. Profiling of TFs in four batches.**

Each TF was subject to two biological replicates. Each batch had two controls representing wild type SC5314 *C. albicans* with the exception of batch 2. *One replicate for ZCF7 was sequenced a second time due to failure in the first trial. The two sequenced samples originate from distinct colonies. Except for ZCF7, each sequenced sample comes from two colonies harvested from a single agar plate.

| ZCTF (*C.a.*) | Type | *S.c.* | DNA b.d. | Description |
|---|---|---|---|---|
| orf19.1604 | Best hit | LYS14 | LYS14 | Predicted Gal4-like DNA-binding transcription factor |
| ZCF21 | Best hit | LYS14 | LYS14 | LYS14 is transcriptional activator of lysine pathway genes that regulates lysine biosynthesis in *S.c.* |
| ZCF18 | Best hit | LYS14 | LYS14 | LYS14 is transcriptional activator of lysine pathway genes that regulates lysine biosynthesis in *S.c.* |
| ZCF4 | Best hit | LYS14 | ECM22 | LYS14 is transcriptional activator of lysine pathway genes that regulates lysine biosynthesis in *S.c.* |
| LYS142 | Best hit | LYS14 | LYS14 | LYS14 is transcriptional activator of lysine pathway genes that regulates lysine biosynthesis in *S.c.* |
| ZCF8 | -- | -- | CEP3 | Required for yeast cell adherence to silicone substrate (Finkel et al. 2012) |
| ZCF31 | -- | -- | YJL206C | Required for yeast cell adherence to silicone substrate (Finkel et al. 2012) |
| SUC1 | Ortholog | MAL13 | MAL13 | A positive regulator of cell surface targets of adherence regulators and hyphal growth or virulence, and negative regulator of ZAP1 targets, involved in cell adherence (Finkel et al. 2012) |
| ZCF23 | Ortholog | GSM1 | ERT1 | Induced in flow model biofilm. The ortholog is involved in regulation of energy metabolism in *S.c.* |
| ZCF10 | Best hit | CAT8 | CAT8 | CAT8 activator of genes involved in non-fermentative growth and diauxic shift in *S. cerevisiae* |
| FCR1 | Best hit | CAT8 | PDR1 | Transcription factor involved in resistance to fluconazole/ketoconazole/brefeldin A. Transposon mutagenesis enhances filamentation (Uhl et al. 2003) |
| ZCF20 | Ortholog | HAP1 | HAP1 | Regulated by Sef1 and Sfu1and repressed by Hap4. The ortholog (HAP1) in S.c. is involved in complex regulation of gene expression in response to heme and oxygen level |
| ZCF13 | Best hit | HAP1 | HAP1 | Involved in filament growth and invasion. Inactivated mutants are hypersusceptible to heat stress (Vandeputte *et al.* 2011) |
| ZCF6 | Best hit | ASG1 | ASG1 | Involved in virulence |
| ZCF24 | Best hit | ASG1 | CAT8 | Induced by caspofungin and repressed by Hap43 |
| ZCF9 | Best hit | ARG81 | ARG81 | Hypersensitive to toxic ergosterol analog ECC69 and/or ECC1384 |
| ZCF19 | -- | -- | TBS1 | Uncharacterized |

| ZCTF (*C.a.*) | Type | *S.c.* | DNA b.d. | Description |
|---|---|---|---|---|
| HAL9 | Best hit | HAL9 | OAF3 | Induced by Mnl1 to regulate weak acid stress responses in *C.a.* (Ramsdale *et al.* 2008) |
| orf19.2230 | Ortholog | RDS3 | RDS3 | Essential pre-mRNA-splicing factor; decreased transcription is observed upon benomyl treatment |
| ZCF15 | Best hit | PDR1 | STB4 | Transcription factor of unknown function. PDR1 is the master regulator of pleiotropic drug response elements (PDREs) in *S.c.* (Jungwirth & Kuchler 2006) |
| ZCF27 | Best hit | OAF1 | YKL222C | OAF1 is involved in beta-oxidation of fatty acids, peroxisomal proliferation and chromatin silencing at telomeres in *S. cerevisiae*. Acting as negative regulator of general stress response and positive regulator of fatty acid metabolism response (Smith *et al.* 2007) |
| ZCF35 | Best hit | OAF3 | STB4 | OAF3 is a transcription repressor involved in regulation of multiple cellular responses in *S.c.* |
| ASG1 | Ortholog | ASG1 | ASG1 | Gal4p family zinc-finger transcription factor, involved in regulation of multidrug transport and pleiotropic drug resistance (Coste et al. 2008) |
| ARO80 | Ortholog | ARO80 | ARO80 | Transcription activator of aromatic amino acid catabolism that regulates aromatic alcohol biosynthesis via the Ehrlich pathway. Involved in adaptation to different PH condition, and aerobic, anaerobic, and hypoxic microenvironments of host (Ghosh 2008) |
| ZCF26 | Best hit | GAL4 | STB4 | Induced by $\alpha$ pheromone in early stages of mating |
| ZCF22 | Best hit | UPC2 | UPC2 | UPC2 induces steroid alcohol (sterol) biosynthetic genes and acts as sterol sensor upon sterol depletion in *S.c.* |
| TEA1 | Ortholog | TEA1 | TEA1 | Ty enhancer activator. The ortholog is required for full levels of Ty enhancer-mediated transcription activates Ty1 retrotransposon in S.c. (Gray and Fassler 1996; MacPherson *et al.* 2006) |
| ZCF7 | Best hit | YPR196W | YFL052W | Mutants are unable to utilize mannitol as a carbon source |
| UME7 | Ortholog | UME6 | UME6 | UME6 is the regulator of early meiotic genes involved in chromatin remodeling and transcription repression via DNA looping in *S. cerevisiae* |
| ZCF16 | Ortholog | SIP4 | SIP4 | The ortholog (SIP4) is involved in positive regulation of gluconeogenesis in *S.c.* |

**Table 2. Description on individual ZCTFs understudy.**

Manually curated from CGD and primary literature. S.C. indicates *S. cerevisiae* and C.a. indicates *C. albicans*.

| | Enriched pathways and processes | $T_1$ | $T_2$ | $T_3$ | $T_4$ | $T_5$ |
|---|---|---|---|---|---|---|
| $G_1$ | 1. rRNA & ncRNA biogenesis, maturation, processing and metabolic processing<br>2. RNA and ribosome biogenesis, processing and localization<br>3. Nucleic acid metabolic process<br>4. Gene expression<br>5. Cellular component biogenesis and metabolic process<br>6. Heterocycle, organic cyclic, cellular aromatic and nitrogen compound metabolic processes<br>7. RNA, nucleic acid and protein transport<br>8. Nuclear export<br>9. Drug response | - | ↑ | ↑ | ↑↑ | ↑↑ |
| $G_2$ | ---- | - | - | - | - | - |
| $G_3$ | 1. rRNA and ncRNA biogenesis, maturation, processing and metabolic process<br>2. RNA processing<br>3. Gene expression<br>4. Cellular nitrogen compound metabolic process<br>5. Cellular component biogenesis | - | ↑↑ | ↑ | ↑ | ↑↑ |
| $G_4$ | 1. Oxidation-reduction processes<br>2. Nitrogen utilization<br>3. a-aminoacid biosynthesis and metabolic process<br>4. Oxoacid, carboxylic acid, organic acid, carbohydrate, biotin & drug metabolic processes<br>5. Transmembrane transport; Cation transport<br>6. Copper, ammonium and metal ion transport | - | - | - | ↓ | ↓↓ |
| $G_5$ | 1. Oxidation-reduction processes<br>2. Sugar and carbohydrate metabolic processes<br>3. Energy reserve | - | ↓ | ↓ | ↓↓ | ↓↓ |

**Table 3. Biological processes and pathways enrichment analysis for each gene and ZCTF cluster across the 500 most variable genes.**

Gene ontology enrichment analysis for each gene and ZCTF clusters from **Figure 9**, Here ↑ and ↑↑ represent the degree of over-expression. Similarly, ↓ and ↓↓ represent under-expression.

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences (5' TO 3') | ZCF27 | UME7 | ZCF19 | ZCF20 | ASG1 | HAL9 | ZCF21 | ZCF22 | orf19.1604 | ZCF4 | ZCF7 | LYS142 | SUC1 | TEA1 | ZCF6 | ZCF9 | ZCF8 | ARO80 | ZCF10 | ZCF13 | orf19.2230 | ZCF16 | ZCF15 | ZCF26 | FCR1 | ZCF24 | ZCF18 | ZCF23 | ZCF31 | ZCF35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0388.1 | SPT23 | SPT23* | RAAATCAA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0417.1 | YAP5 | -- | ARRCAT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0370.1 | RME1 | RME1* | TYNAAAGGNA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0426.1 | YHP1 | -- | TAATTG | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0274.1 | ARR1 | -- | ANYTGAAT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0317.1 | HCM1 | HCM1* | ATAAACAA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0287.1 | CUP2 | -- | CAGCARAAAWG | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0398.1 | SUM1 | -- | NWWATTTTT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0356.1 | PHO2 | GRF10* | WTAWTW | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0387.1 | SPT2 | orf19.6726* | WNTTAAVYAR | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0297.1 | FKH2 | -- | GTAAACA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0393.1 | STE12 | CPH1* | TGAAACR | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0277.1 | AZF1 | orf19.173* | AAAAAGAAA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0316.1 | HAP5 | HAP5* | NSSNNKCTNATTGGY | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0327.1 | HMRA1 | MTLA1* | RCACAAT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0371.1 | ROX1 | RFG1* | YCNATTGTTCTC | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0307.1 | GLN3 | GLN3* | GATAA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0379.1 | MOT2 | NOT4* | ATATA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0407.1 | THI2 | -- | GGMAACYSWAAGARC | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0313.1 | HAP2 | HAP2* | TTGGY | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0319.1 | HSF1 | CTA8* | ATGGAACN | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0390.1 | STB3 | STB3* | GNYNAAAWTTTTTCACTNHNN | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0433.1 | YOX1 | YOX1* | TTAATTAA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences (5' TO 3') | orf19.1604 | ZCF35 | ZCF27 | ZCF16 | FCR1 | ZCF7 | ZCF18 | ZCF4 | ARO80 | ASG1 | HAL9 | LYS142 | orf19.2230 | SUC1 | UME7 | ZCF6 | TEA1 | ZCF15 | ZCF8 | ZCF9 | ZCF13 | ZCF19 | ZCF20 | ZCF21 | ZCF22 | ZCF10 | ZCF26 | ZCF23 | ZCF24 | ZCF31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0391.1 | STB4 | CTA7* | YTCGGAA | ■ | ■ | | ■ | ■ | | | | | | ■ | | | ■ | | | | ■ | | | | | | | | | | | | |
| MA0419.1 | YAP7 | -- | ATTAGTAAYCA | ■ | ■ | ■ | | | | | ■ | ■ | | | | | | | | | | ■ | | | | | | | | ■ | | | |
| MA0367.1 | RGT1 | RGT1* | NNWTNWTCCGN | ■ | ■ | ■ | ■ | | | | | ■ | | | | | | | | | | | | | | | | ■ | | | | | |
| MA0408.1 | TOS8 | CUP9* | NTGTCAAA | ■ | | ■ | ■ | | | | | | ■ | | | ■ | | | | | | | | | | | | | ■ | | | | |
| MA0298.1 | FZF1 | -- | CTATCA | ■ | | | | | | | | | | ■ | | | | ■ | | | | | | ■ | | | | | ■ | | | | ■ |
| MA0382.1 | SKO1 | SKO1* | ACGTAWTG | | | ■ | ■ | ■ | | | | | | | | ■ | | ■ | | | | | | | | | | ■ | | | | | |
| MA0326.1 | MAC1 | MAC1* | TTTGCTCR | | | | | ■ | ■ | ■ | | ■ | | | | | | | | ■ | | | | ■ | | | | | | | | | |
| MA0331.1 | MCM1 | MCM1* | CCNNWTTRGGAA | | | | | ■ | ■ | ■ | | | | | | | | | | | | | | | | | | ■ | | ■ | ■ | | |
| MA0378.1 | SFP1 | SFP1* | NNNNDRAAAAWTTTTYYNNNN | | | | ■ | ■ | | | | ■ | | | | | ■ | | ■ | ■ | | | | | | | | ■ | | | | | |
| MA0336.1 | MGA1 | SFL2* | NNNNHNATAGAACAYNHHNNN | | ■ | ■ | | | | | | ■ | | | | | | | ■ | | | ■ | ■ | ■ | | | ■ | ■ | | | | | |
| MA0288.1 | CUP9 | -- | TGACACAWW | | ■ | ■ | | | | | | | ■ | ■ | | ■ | | | | | | | | | | | | | ■ | | | | |
| MA0314.1 | HAP3 | HAP31* | TCTSATTGGYYVRRA | | | ■ | | | | ■ | | ■ | | | | ■ | | | | | | | | | ■ | | | | | ■ | | | ■ |
| MA0328.2 | MATALPHA2 | -- | CRTGTAAW | | ■ | ■ | ■ | | | | | | | | | ■ | | ■ | | | | | | | ■ | | ■ | ■ | | | | | |
| MA0377.1 | SFL1 | SFL1* | NNNNNNATMGAAGAAANNWNW | | | ■ | | | | | | | | | | | | | | ■ | ■ | | ■ | | | ■ | ■ | | | | | | |
| MA0406.1 | TEC1 | TEC1* | RCATTCCN | | | | | | | | | ■ | | | ■ | ■ | | ■ | | ■ | | | | ■ | ■ | | | | | | | | |
| MA0279.1 | CAD1 | -- | ATTAGTAAYC | | | | | | | | | ■ | | | | | | ■ | | | ■ | | | | | | ■ | ■ | ■ | | ■ | | |
| MA0296.1 | FKH1 | FKH2* | NNWWWTGTAAACAAANNNNN | | | | | | | | | | ■ | | | ■ | | | | | ■ | | | | | | | ■ | ■ | | | | |
| MA0383.1 | SMP1 | -- | NNACCTWTAATTAWANBWNNN | | | | | | | | | | ■ | | | | | | | | ■ | | ■ | | ■ | | ■ | | | ■ | | | |
| MA0309.1 | GZF3 | GZF3* | YGATAASN | | | | ■ | | | | | | | ■ | | | | | | ■ | ■ | ■ | | | | | | ■ | | | | | ■ |
| MA0346.1 | NHP6B | -- | NNTNNNWATATATWWWRNDV | | ■ | | | | | | | | | | | ■ | ■ | | | ■ | ■ | ■ | | | | ■ | | | | | | | |
| MA0289.1 | DAL80 | -- | CGATAAG | | | | ■ | | | | | | | | | | | | | ■ | ■ | | ■ | | | | ■ | | | | | | |
| MA0335.1 | MET4 | MET4* | AACTGTGG | | | | ■ | | | | | | | | | | | ■ | | | | ■ | | | | | | | ■ | ■ | ■ | ■ | |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences (5' TO 3') | orf19.1604 | FCR1 | ZCF27 | ZCF16 | SUC1 | ZCF7 | ARO80 | HAL9 | ASG1 | ZCF8 | orf19.2230 | TEA1 | UME7 | ZCF9 | ZCF6 | ZCF10 | ZCF13 | ZCF15 | ZCF4 | ZCF18 | ZCF21 | ZCF19 | ZCF22 | ZCF23 | ZCF26 | LYS142 | ZCF24 | ZCF31 | ZCF20 | ZCF35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0294.1 | EDS1 | -- | CGGAANAAT | ■ | | | ■ | | ■ | | | | | | | | ■ | | ■ | | | | | | | | | | | | | | |
| MA0300.1 | GAT1 | GAT1* | NYGATAAG | ■ | | ■ | | | | | | | | | ■ | | | | | | | ■ | | | | | | | | | ■ | | ■ | |
| MA0321.1 | INO2 | -- | GCATGTGAA | ■ | ■ | | | | | | | | ■ | | | | | | | | | | | | | | ■ | | | ■ | | | |
| MA0275.1 | ASG1 | ASG1* | CCGGAW | | ■ | | ■ | | | | | | | | | | ■ | | | | | | ■ | | | | | | | | | | |
| MA0365.1 | RFX1 | RFX1* | NGTTGCYA | | ■ | | | | ■ | | | | | | | ■ | | | | | | | | | ■ | | | | | | | ■ | |
| MA0411.1 | UPC2 | UPC2* | NAWACGA | | | ■ | | | | | ■ | | | | | | | | | | | | | ■ | | | ■ | | | | | | |
| MA0340.1 | MOT3 | CAS5* | HAGGYA | | | | ■ | | | | | ■ | | | | | | | | | | | | | | ■ | | | | ■ | | ■ | |
| MA0386.1 | SPT15 | TBP1* | VNHNAGNWATATATATNSNNN | | | | ■ | | | | | | ■ | | ■ | | | ■ | | | | | ■ | | | | | | | | | | |
| MA0929.1 | NCU00019 | -- | NNNGTAAAYANN | | | | ■ | | | | | | | ■ | | | | | | ■ | | | ■ | | | | | | | | ■ | | |
| MA0272.1 | ARG81 | ARG81* | NTGACTCH | | | | ■ | | | | | | | ■ | | | | | | | ■ | | | | | | | | | | | | ■ |
| MA0293.1 | ECM23 | -- | NNAGATCTNNN | | | | ■ | | | | | | | | ■ | ■ | | | | | | | | | | | ■ | | | | | | |
| MA0302.1 | GAT4 | -- | NHAGATCTNNN | | | | ■ | | | | | | | | ■ | ■ | | | | | | | | | | | | | | | | | ■ |
| MA0385.1 | SOK2 | EFG1* | NNMTGCAKGNN | | | | ■ | | ■ | | | | | ■ | | | | | | | | | | | | | | ■ | | | | | |
| MA0305.1 | GCR2 | -- | GCTTCCH | | | | | | ■ | | ■ | | | | | | ■ | | | | | | ■ | | | ■ | | | | | | | | |
| MA0349.1 | OPI1 | OPI1* | NGAACCV | | | | | | ■ | | ■ | | | | | ■ | | | ■ | | | | | | | | | | | | | | |
| MA0409.1 | TYE7 | TYE7* | CACGTGA | | | | | | ■ | | | | ■ | | | | | | | | | | | | | | | ■ | ■ | | | | |
| MA0369.1 | RLM1 | RLM1* | VNTTCTAWWWATAGMYYN | | | | | | | ■ | | | | | | | ■ | | | ■ | | | | | | ■ | | | | | | | |
| MA0332.1 | MET28 | MET28* | CTGTGG | | | | | | | ■ | | | | | | | | ■ | | | | | ■ | | | | | | | | | ■ | ■ |
| MA0334.1 | MET32 | orf19.1757* | MGCCACA | | | | | | | | | | ■ | | | ■ | | | ■ | | | | | | | | | | ■ | | | | |
| MA0345.1 | NHP6A | NHP6A* | NNNNHYWNTATATAANNNNNH | | | | | | | | | | | ■ | ■ | | ■ | | | | | | | | | | | | | | ■ | | |
| MA0439.1 | YRR1 | -- | NTTATHTCCGY | | | | | | | | | | | | | | | | | | | | ■ | ■ | | ■ | | | | | | | ■ |
| MA0438.1 | YRM1 | -- | ACGGAAAT | | | | | | | | | | | | | | | | | | | | | | | ■ | | ■ | | ■ | | | ■ |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences (5' TO 3') | orf19.1604 | ZCF16 | FCR1 | ZCF4 | ZCF27 | ZCF7 | ARO80 | ZCF10 | ASG1 | HAL9 | LYS142 | ZCF23 | orf19.2230 | SUC1 | TEA1 | UME7 | ZCF6 | ZCF8 | ZCF9 | ZCF13 | ZCF15 | ZCF18 | ZCF19 | ZCF20 | ZCF21 | ZCF22 | ZCF24 | ZCF26 | ZCF31 | ZCF35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0422.1 | URC2 | -- | NCGGANWTAN | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | | | ■ |
| MA0432.1 | YNR063W | -- | TCGGAGAW | ■ | ■ | | | | | | | | ■ | | | | | | | | | | | | | | | | | | | | |
| MA0350.1 | TOD6 | -- | NNBNNNASCTC ATCGCNNNNN | | | ■ | | | ■ | | | | | | | ■ | | | | | | | | | | | | | | | | | |
| MA0323.1 | IXR1 | -- | AARCCGGRAGC GGYG | | ■ | ■ | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | |
| MA0354.1 | PDR8 | -- | RCGGAGAT | | | ■ | | | | | | | ■ | | | | | | | | | | | | | | ■ | | | | | | |
| MA0368.1 | RIM101 | RIM101* | YGCCAAG | | | ■ | | | | | | | | | | ■ | | | | | | | | | ■ | | | | | | | | |
| MA0362.1 | RDS2 | CWT1* | NTCGGGG | | ■ | ■ | | | | | | | | | | | ■ | | | | | | | | | | | | | | | | |
| MA0440.1 | ZAP1 | CSR1* | ACCYTMAAGGT NATG | | | | | ■ | ■ | | | | | | | | | | | | | | | | ■ | | | | | | | | |
| MA0311.1 | HAL9 | TAC1* | CGGAR | | ■ | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | ■ |
| MA0304.1 | GCR1 | -- | TGGAAGCC | | | | | | | ■ | ■ | | | | | | | | | | | | | | | | ■ | | | | | | |
| MA0355.1 | PHD1 | -- | NSMTGCANNN | | | | | | | ■ | | | | | | | ■ | | | | | | | | | | | ■ | | | | | |
| MA0414.1 | XBP1 | orf19.5210* | YTCGARN | | ■ | | | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | |
| MA0318.1 | HMRA2 | -- | CRTGTAAW | | ■ | | | | | | | | | ■ | | | | | ■ | | | | | | | | | | | | | | |
| MA0301.1 | GAT3 | -- | AGATCTANN | | | | | | | | | | | | | ■ | | ■ | ■ | | | | | | | | | | | | | | |
| MA0322.1 | INO4 | INO4* | GCATGTGAA | | | | | | | | | | | | | ■ | | | | ■ | | | | | | | | ■ | | | | | |
| MA0325.1 | LYS14 | LYS143* | NCGGAATT | | ■ | | | | | | | | | | | ■ | | | | | | | | | | | | | | | | ■ |
| MA0357.1 | PHO4 | PHO4* | SCACGTGS | | | | | | | | | | | | ■ | ■ | ■ | | | | | | | | | | | | | | | | |
| MA0436.1 | YPR022C | orf19.7397* | CCCCACN | | | | | | | | | | | | ■ | ■ | | | | | | | | | | | | | | | ■ | | |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences (5' TO 3') | orf19.1604 | FCR1 | ZCF35 | ZCF16 | ZCF4 | ZCF27 | ZCF7 | ARO80 | ASG1 | ZCF10 | HAL9 | LYS142 | orf19.2230 | SUC1 | ZCF9 | TEA1 | UME7 | ZCF6 | ZCF8 | ZCF18 | ZCF13 | ZCF15 | ZCF19 | ZCF20 | ZCF21 | ZCF22 | ZCF23 | ZCF24 | ZCF26 | ZCF31 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0268.1 | ADR1 | ADR1* | NCCCCAM | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0339.1 | MIG3 | -- | CCCCRCN | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0351.1 | DOT6 | DOT6* | NBNNNNWSCTC ATCGCVNNNN | | ■ | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | |
| MA0363.1 | REB1 | -- | RTTACCCYG | | | | | | | ■ | | | | | | | | | | ■ | | | | | | | | | | | | | |
| MA0401.1 | SWI4 | SWI4* | ACGCGAAA | | | | | | | ■ | | | | | | | | ■ | | | | | | | | | | | | | | | |
| MA0270.1 | AFT2 | AFT2* | NACACCCN | | | | | | ■ | | | | | | | ■ | | | | | | | | | | | | | | | | | |
| MA0271.1 | ARG80 | -- | WGACKC | | | | | | ■ | | | | | | | | ■ | | | | | | | | | | | | | | | | |
| MA0364.1 | REI1 | REI1* | CCCCTGA | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | ■ | | |
| MA0434.1 | YPR013C | -- | YGTARATCN | | | | | | | | | | ■ | ■ | | | | | | | | | | | | | | | | | | | |
| MA0281.1 | CBF1 | CBF1* | GCACGTGA | | | | | | | | | | | | | ■ | | | | | | | | | | | | | | ■ | | | |
| MA0310.1 | HAC1 | HAC1* | GACACGTN | | | | | | | | | | | | | ■ | | | | | | | | | | | | ■ | | | | | |
| MA0282.1 | CEP3 | -- | YTCGGAAN | | | | ■ | | | | | | | | | | ■ | | | | | | | | | | | | | | | | |
| MA0396.1 | STP3 | STP4* | NNTAGCGCN | | | | | | | | | | | | | | ■ | ■ | | | | | | | | | | | | | | | |
| MA0397.1 | STP4 | -- | GNTAGCGCA | | | | | | | | | | | | | | ■ | ■ | | | | | | | | | | | | | | | |
| MA0284.1 | CIN5 | CAP4 | TTAYGTAAKC | | | | | | | | | | | | ■ | | | | | ■ | | | | | | | | | | | | | |
| MA0373.1 | RPN4 | RPN4* | GGTGGCG | | | | | | | | | | | | | | | | | ■ | | | | | | | | ■ | | | | | |
| MA0290.1 | DAL81 | DAL81* | AAAAGCCGCGG GCGGGATT | | | | | | | | | | | | | | | | | | | | ■ | ■ | | | | | | | | | |
| MA0392.1 | STB5 | STB5* | CGGNNNTA | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | ■ |
| MA0312.1 | HAP1 | ZCF20* | CGGAGWTA | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0348.1 | OAF1 | CTA4* | YCGGRGATA | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0420.1 | ERT1 | ZCF11* | AYCGGAAC | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0430.1 | YLR278C | -- | NCGGAGTT | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0437.1 | YPR196W | ZCF7 | ATTTNYCCG | | | ■ | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences | orf19.1604 | FCR1 | ZCF4 | ZCF27 | ZCF7 | ARO80 | ASG1 | HAL9 | LYS142 | orf19.2230 | SUC1 | TEA1 | UME7 | ZCF6 | ZCF8 | ZCF9 | ZCF10 | ZCF13 | ZCF15 | ZCF16 | ZCF18 | ZCF19 | ZCF20 | ZCF21 | ZCF22 | ZCF23 | ZCF24 | ZCF26 | ZCF31 | ZCF35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0292.1 | ECM22 | ZCF28 | HTCCGGA | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0308.1 | GSM1 | ZCF23* | NNNNNNWANCT CCGGANNNNN | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0424.1 | YER184C | -- | HTCCGGAN | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0428.1 | YKL222C | -- | NACGGARAT | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0429.1 | YLL054C | orf19.6888 | CGGCCGA | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0358.1 | PUT3 | PUT3* | CCCGGGAN | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0380.1 | SIP4 | ZCF16* | YTCCGGA | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | |
| MA0276.1 | ASH1 | ASH1* | CCGNATCRGG | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | |
| MA0341.1 | MSN2 | MSN4* | RGGGG | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | |
| MA0342.1 | MSN4 | -- | AGGGG | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | |
| MA0343.1 | NDT80 | NDT80* | HNNNNKGMCAC AAAANCSVNN | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | |
| MA0366.1 | RGM1 | -- | AGGGG | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | | |
| MA0375.1 | RSC30 | -- | NSCGCGCG | | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | | |
| MA0431.1 | TDA9 | orf19.5026* | NCCCCDCWN | | | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | | |
| MA0372.1 | RPH1 | orf19.2743* | ACCCCTAA | | | | | | | | | | | | | | | | | | | | | | | | | | ■ | | | | |
| MA0410.1 | UGA3 | UGA3* | NGGCGGGA | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ■ | |
| MA0286.1 | CST6 | RCA1* | RTGACGTNN | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | ■ |

| Motif ID | *S.c.* Gene | *C.a.* Gene | Consensus sequences | orf19.1604 | FCR1 | ZCF4 | ZCF27 | ZCF7 | ARO80 | ASG1 | HAL9 | LYS142 | orf19.2230 | SUC1 | TEA1 | UME7 | ZCF6 | ZCF8 | ZCF9 | ZCF10 | ZCF13 | ZCF15 | ZCF16 | ZCF18 | ZCF19 | ZCF20 | ZCF21 | ZCF22 | ZCF23 | ZCF24 | ZCF26 | ZCF31 | ZCF35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MA0403.1 | TBF1 | TBF1* | ARCCCTAN | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0295.1 | FHL1 | FHL1* | GACGCANA | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0413.1 | USV1 | BCR1* | DNNTTMCCCCT GAANNNNNN | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0415.1 | YAP1 | CAP1* | NNNNNMTTACG TAAYNNNNN | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0266.1 | ABF2 | -- | NTCTAGA | | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0389.1 | SRD1 | -- | AGATCTMN | | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | | | | |
| MA0402.1 | SWI5 | orf19.2612 | TGCTGGTN | | | | | | | | | | ■ | | | | | | | | | | | | | | | | | | | | | |
| MA0278.1 | BAS1 | BAS1* | NCWNRGCCVG AGTCARDWNNN | | | | | | | | | | | | | ■ | | | | | | | | | | | | | | | | | | |
| MA0435.1 | YPR015C | -- | TNNNNACGTAA ATCMTNNHH | | | | | | | | | | | | | ■ | | | | | | | | | | | | | | | | | | |
| MA0360.1 | RDR1 | -- | TGCGGAAN | | | | | | | | | | | | | | | | ■ | | | | | | | | | | | | | | | |
| MA0285.1 | CRZ1 | CRZ1* | NNNMGCCNC | | | | | | | | | | | | | | | | | | ■ | | | | | | | | | | | | | |

**Table 4. Promoter analysis of the ZCTFs.**

Phylogenetic footprinting was used to identify conserved motifs present in the promoter of each of the ZCTFs. Here several fungi including *S. cerevisiae* were used in the analysis. The TFBS motifs originate from JASPAR. Only motifs (rows) appearing in at least one ZCTF are included here. "*S. cerevisiae* Gene" refers to the *S. cerevisiae* TF that binds the motif. *C. albicans* gene is the known ortholog of the *S. cerevisiae* TF, or best hit. Here dashes (--) indicate cases where we could not identify a full-length alignment of the *S. cerevisiae* TF with a *C. albicans* gene. N corresponds to any nucleotide; Y indicates pyrimidine (C or T) and R for purine (A or G). W is either A or T and S is either C or G. B corresponds to (G, C or T) ; D indicates (G, A or T); H indicates (A, C or T) and V indicates (G, A or C). S.C. indicates S. cerevisiae and C.a. indicates C. albicans.

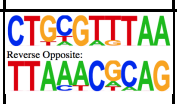
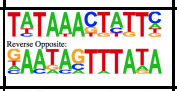
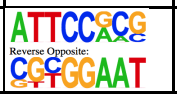
| Motif Rank | Motif logo | P-value | % d.e. unique | % Back-ground | Possible TF that binds the motif | *S. cerevisiae* Matches |
|---|---|---|---|---|---|---|
| 1 |  | $1e^{-9}$ | 40.54 | 4.54 | PABPC1<br>Homo sapiens | -- |
| 2 |  | $1e^{-8}$ | 18.92 | 0.29 | HuR<br>Homo sapiens | -- |
| 3 |  | $1e^{-8}$ | 27.03 | 1.76 | MSN4<br>*S. cerevisiae* | -- |
| 4 |  | $1e^{-7}$ | 18.92 | 0.46 | Ct<br>Drosophila melanogaster | SPT2<br>STE12<br>HMRA2 |
| 5 |  | $1e^{-7}$ | 21.62 | 1.03 | Foxq1<br>Rat | TATA-box |
| 6 |  | $1e^{-7}$ | 27.03 | 2.29 | HHO6<br>Arabidopsis thaliana | SPT23 |
| 7 |  | $1e^{-7}$ | 13.51 | 0 | GLN3<br>*S. cerevisiae* | DAL82<br>GZF3 |
| 8 |  | $1e^{-7}$ | 27.03 | 2.46 | TBX20<br>Homo sapiens | |
| 9 |  | $1e^{-7}$ | 21.62 | 1.14 | SOX10<br>Homo sapiens | -- |
| 10 |  | $1e^{-6}$ | 18.92 | 0.75 | PFF0320c<br>Plasmodium falciparum | YAP1,3,5-7<br>ARR1 |
| 11 |  | $1e^{-6}$ | 13.51 | 0.15 | Sox13<br>Mus musculus | -- |
| 12 |  | $1e^{-6}$ | 18.92 | 1.04 | Brn2<br>Mus musculus | -- |
| 13 |  | $1e^{-4}$ | 51.35 | 18.50 | Initiator Promoters<br>Drosophila melanogaster | -- |
| 14 |  | $1e^{-3}$ | 8.11 | 0.38 | SVP<br>Arabidopsis thaliana | -- |
| 15 |  | $1e^{-2}$ | 16.22 | 3.23 | Sox1<br>Mus musculus | CBF1<br>ROX1 |
| 16 |  | $1e^{-2}$ | 24.32 | 8.26 | LYS14<br>*S. cerevisiae* | LYS14 |

**Table 5. Identifying potential TFBSs for orf19.1604.**

%unique d.e. corresponds to the percentage the motif appears in a the positive learning set of unique d.e. genes for orf19.1604, and %background corresponds to the percentage that the same motif appears in the negative learning set consisting of 1000 randomly chosen genes. All p-values here are above $1e^{-10}$. The last column gives the TF name and species for the motif.

| Rank | Pattern (5' TO 3') | Best Guess | *S. cerevisiae* Matches | Matched Sequence | Q-value |
|---|---|---|---|---|---|
| 2 | TCCWTTATTTTA | HuR<br>Homo sapiens | -- | 3'-TAAAATATTTGA-5' | 0.00289 |
| 5 | TATAAACTATTC | Foxq1<br>Rattus rattus | TATA-box | 3'-AAAATATTTGAA-5' | 0.0035 |
| 5 | TATAAACTATTC | Foxq1<br>Rattus rattus | TATA-box | 3'-AATTTATTTATA-5' | 0.00399 |
| 5 | TATAAACTATTC | Foxq1<br>Rattus rattus | TATA-box | 3'-GGATCAAGTAAA-5' | 0.00399 |
| 9 | CTTTGTTCGT | SOX10<br>Homo sapiens | -- | 5'- ATTTATTTAT-3' | 0.00809 |
| 9 | CTTTGTTCGT | SOX10<br>Homo sapiens | -- | 5'-ATTTTTATTTAATTAAGT-3' | 0.00809 |
| 15 | CAATACAATAVA | Sox1<br>Mus musculus | CBF1<br>ROX1 | 5'-TAAAATATTTGA-3' | 0.0165 |
| 13 | TTGRACTGAA | Initiator Promoters<br>Drosophila melanogaster | -- | 5'-TTGAAACGAA-3' | 0.0229 |

**Table 6. The subset of putative TFBSs for orf19.1604 that are present in the promoter of ZCF4.**

Phylogenetic footprinting was used to estimate whether each putative TFBS motifs for orf19.1604 **(Table 5)** was or was not present in the promoter of ZCF4. N corresponds to any nucleotide; Y indicates pyrimidine (C or T) and R for purine (A or G). W is either A or T, and S is either C or G.

| Rank | Pattern (5' TO 3') | Best Guess | *S. cerevisiae* Matches | Matched Sequence | Q-value |
|---|---|---|---|---|---|
| 13 | TTGRACTGAA | Initiator Promoters Drosophila melanogaster | -- | 3'-TTAAACCCTG-5' | 0.000125 |
| 11 | AAGAACCATTTC | Sox13 Mus musculus | -- | 5'-TAGTATCATTTA-3' | 0.00225 |
| 7 | CTATCWTATCCC | GLN3 *S. cerevisiae* | DAL82 GZF3 | 5'-TAACAACATTCG-3' | 0.00449 |
| 12 | ACACATATTCAT | Brn2 Mus musculus | -- | 5'-ACATTCGGTTAA-3' | 0.00503 |
| 15 | CAATACAATAVA | Sox1 Mus musculus | CBF1 ROX1 | 3'-TTTAATGCATAG-5' | 0.0056 |
| 15 | CAATACAATAVA | Sox1 Mus musculus | CBF1 ROX1 | 3'-TTGCCTGCTTGC-5' | 0.0056 |
| 1 | CTATTTTTCTTC | PABPC1 Homo sapiens | -- | 5'-TTTTCTTTCTTC-3' | 0.00643 |
| 13 | TTGRACTGAA | Initiator Promoters Drosophila melanogaster | -- | 3'-CTCTATTTTA-5' | 0.0159 |
| 10 | GTTTAGTGAGAA | PFF0320c Plasmodium falciparum | YAP1,3,5-7 ARR1 | 3'-TTTTCTAAATAT-5' | 0.0167 |
| 14 | TTCCTTTCTTGG | SVP Arabidopsis thaliana | -- | 5'-CTTTCTTCTTGG-3' | 0.0246 |
| 16 | ATTCCRCG | LYS14 *S. cerevisiae* | LYS14 | 5'-ATTCCAAT-3' | 0.0348 |

**Table 7. The subset of putative TFBSs for the ZCTF orf19.1604 that are present in the promoter of ZCF27.**

Motifs that identified as potential binding sites of orf19.1604 **(Table 5)** were tested against multiple alignment of ZCF27 promoter from different strain, to identify if there is a conserved motif at promoter site of ZCF27 that match any of those motifs. N corresponds to any nucleotide; Y indicates pyrimidine (C or T) and R for purine (A or G). W is either A or T and S is either C or G.

| Rank | Pattern (5' TO 3') | Best Guess | *S. cerevisiae* Matches | Matched Sequence | Q-value |
|---|---|---|---|---|---|
| 12 | ACACATATTCAT | Brn2 Mus musculus | -- | 5'-TTTATTTTTTAT-3' | 0.000246 |
| 12 | ACACATATTCAT | Brn2 Mus musculus | -- | 5'-CCAAACATTCTT-3' | 0.00175 |
| 12 | ACACATATTCAT | Brn2 Mus musculus | -- | 5'-AAACATTCTTCT-3' | 0.00175 |
| 1 | CTATTTTTCTTC | PABPC1 Homo sapiens | -- | 5'-ATCTTTTTATTT-3' | 0.00246 |
| 1 | CTATTTTTCTTC | PABPC1 Homo sapiens | -- | 5'-TTTTATTTTTTA-3' | 0.00246 |
| 1 | CTATTTTTCTTC | PABPC1 Homo sapiens | -- | 5'-CTTGATTTCTTT-3' | 0.00246 |
| 6 | TAMAGAATCAAA | AT1G49560 Arabidopsis thaliana | SPT23 | 3'-TTTTATTTTTTA-5' | 0.00347 |
| 7 | CTATCWTATCCC | GLN3 *S. cerevisiae* | DAL82 GZF3 | 5'-ATTTTTTATTTA-3' | 0.0074 |
| 11 | AAGAACCATTTC | Sox13 Mus musculus | -- | 3'-TTATTTTTTATT-5' | 0.00808 |
| 9 | CTTTGTTCGT | SOX10 Homo sapiens | -- | 5'-ATTATTCAAT-3' | 0.00809 |
| 9 | CTTTGTTCGT | SOX10 Homo sapiens | -- | 5'-TTTTATTTTT-3' | 0.00809 |
| 9 | CTTTGTTCGT | SOX10 Homo sapiens | -- | 5'-ATCTTTTTAT-3' | 0.00809 |
| 11 | AAGAACCATTTC | Sox13 Mus musculus | -- | 3'-AATCTTTTTATT-5' | 0.00815 |
| 8 | TTCACACCCA | TBX20 Homo sapiens | -- | 5'-TTTATTTACA-3' | 0.0106 |
| 15 | CAATACAATAVA | Sox1 Mus musculus | CBF1 ROX1 | 3'-TTTGCTTGATTT-5' | 0.0123 |

**Table 8. The subset of putative TFBSs for the ZCTF orf19.1604 that are present in the promoter of FCR1.**

Motifs that identified as potential binding sites of orf19.1604 **(Table 5)** were tested against multiple alignment of FCR1 promoter from different strain, to identify if there is a conserved motif at promoter site of FCR1 that match any of those motifs. N corresponds to any nucleotide; Y indicates pyrimidine (C or T) and R for purine (A or G). W is either A or T and S is either C or G.

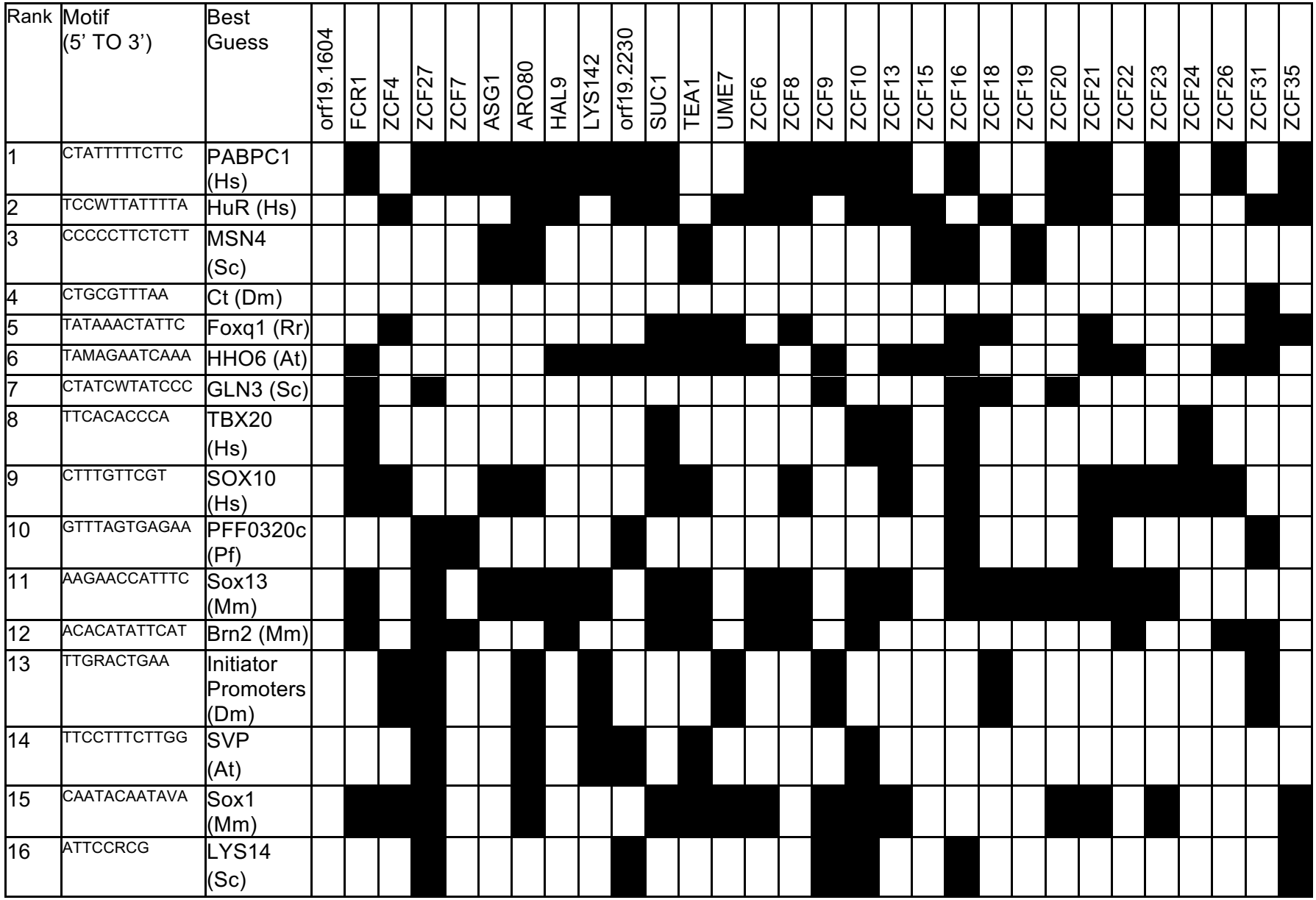

Rank
Motif (5' TO 3')
Best Guess
orf19.1604
FCR1
ZCF4
ZCF27
ZCF7
ASG1
ARO80
HAL9
LYS142
orf19.2230
SUC1
TEA1
UME7
ZCF6
ZCF8
ZCF9
ZCF10
ZCF13
ZCF15
ZCF16
ZCF18
ZCF19
ZCF20
ZCF21
ZCF22
ZCF23
ZCF24
ZCF26
ZCF31
ZCF35
1 CTATTTTTCTTC PABPC1 (Hs)
2 TCCWTTATTTTA HuR (Hs)
3 CCCCCTTCTCTT MSN4 (Sc)
4 CTGCGTTTAA Ct (Dm)
5 TATAAACTATTC Foxq1 (Rr)
6 TAMAGAATCAAA HHO6 (At)
7 CTATCWTATCCC GLN3 (Sc)
8 TTCACACCCA TBX20 (Hs)
9 CTTTGTTCGT SOX10 (Hs)
10 GTTTAGTGAGAA PFF0320c (Pf)
11 AAGAACCATTTC Sox13 (Mm)
12 ACACATATTCAT Brn2 (Mm)
13 TTGRACTGAA Initiator Promoters (Dm)
14 TTCCTTTCTTGG SVP (At)
15 CAATACAATAVA Sox1 (Mm)
16 ATTCCRCG LYS14 (Sc)

**Table 9. ZCTFs that are potentially regulated by orf19.1604.**

We used phylogenetic footprinting to estimate whether each potential TFBS for orf19.1604 **(Table 5)** was present or absent in the promoter of each ZCTF. Sc = *S. cerevisiae;* At = Arabidopsis thaliana; Hs =Homo sapiens; Mm = Mus musculus; Dm = Drosophila melanogaster and Pf = Plasmodium falciparum.

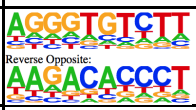
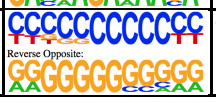
| Motif Rank | Motif logo | P-value | % d.e. unique | % Back-ground | Possible TF that binds the motif | *S. cerevisiae* Match |
|---|---|---|---|---|---|---|
| 1 |  | $1e^{-9}$ | 50% | 1.79% | TCF7L1<br>Homo sapiens | -- |
| 2 |  | $1e^{-7}$ | 31.25% | 0.45% | MET31<br>*S. cerevisiae* | MET32<br>YML081W |
| 3 |  | $1e^{-7}$ | 31.25% | 0.45% | Bc111a(Zf)<br>Homo sapiens | RGT1<br>RLR1<br>EDS1 |
| 4 |  | $1e^{-6}$ | 31.25% | 0.65% | WT1(Zf)<br>Homo sapiens | HAP3 |
| 5 |  | $1e^{-6}$ | 50% | 4.3% | Aef1<br>Drosophila Melanogaster | HAP2 |
| 6 |  | $1e^{-6}$ | 25% | 0.16% | SPL11(SBP)<br>Arabidopsis thaliana | ROX1 |
| 7 |  | $1e^{-5}$ | 43.75% | 4.59% | Six1<br>Homo sapiens | -- |
| 8 |  | $1e^{-4}$ | 43.75% | 5.69% | ZmHOX2a<br>Zea mays | ECM23<br>RTG3 |
| 9 |  | $1e^{-4}$ | 62.5% | 15.84% | KLF4<br>Homo sapiens | AFT2<br>ARG80 |
| 10 |  | $1e^{-3}$ | 43.75% | 9.93% | CG11360<br>Drosophila thaliana | -- |
| 11 |  | $1e^{-2}$ | 31.25% | 5.03% | Lm_0212(RRM)<br>Leishmania major | CST6<br>GAT3 |
| 12 |  | $1e^{-2}$ | 25% | 4.53% | GATA6<br>Homo sapiens | GAT1 |
| 13 |  | $1e^{-1}$ | 12.5% | 2.51% | Sequence Bias<br>polyC-repeat | -- |

**Table 10. Potential transcription factor binding site motifs for FCR1.**

%unique d.e. corresponds to the percentage the motif appears in a the positive learning set of unique d.e. genes for FCR1, and %background corresponds to the percentage that the same motif appears in the negative learning set consisting of 1000 randomly chosen genes. All p-values here are above $1e^{-10}$. The name of the TF and species for the motif are given in the last column.

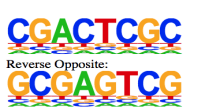
| Motif Rank | Motif logo | P-value | % Target | % Back-ground | Possible TF that binds the motif | *S. cerevisiae* Match |
|---|---|---|---|---|---|---|
| 1 |  | $1e^{-7}$ | 37.50 | 0 | YBX1<br>Homo sapiens | HAP3<br>ACE2<br>SWI5 |
| 2 |  | $1e^{-7}$ | 37.50 | 0 | SFL1<br>*S. cerevisiae* | SFL1 |
| 3 |  | $1e^{-7}$ | 37.50 | 0 | HRB27C<br>Drosophila melanogaster | MGA1 |
| 4 |  | $1e^{-6}$ | 62.50 | 2.25 | CG34031<br>Drosophila melanogaster | PHO2 |
| 5 |  | $1e^{-6}$ | 50 | 0.93 | SeqBias: GA-repeat | GAGA-repeat promoter |
| 6 |  | $1e^{-5}$ | 50 | 1.29 | ZmHOX2a<br>Zea mays | GCR1 |
| 7 |  | $1e^{-3}$ | 37.50 | 1.46 | YLR278C<br>*S. cerevisiae* | HAP1<br>CHA4<br>SUT2<br>PDR8 |
| 8 |  | $1e^{-3}$ | 50 | 5.03 | Ng_0261<br>Naegleria gruberi | SUM1 |
| 9 |  | $1e^{-3}$ | 25 | 0.38 | Pan<br>Drosophila melanogaster | MAC1 |
| 10 |  | $1e^{-2}$ | 25 | 0.77 | G3BP2<br>Homo sapiens | -- |
| 11 |  | $1e^{-2}$ | 12.50 | 0 | HIC1<br>Homo sapiens | VTS1<br>ACE2 |
| 12 |  | $1e^{-2}$ | 12.50 | 0 | Lm_0254<br>Leishmania major | XBP1<br>OPI1 |
| 13 |  | $1e^{-2}$ | 12.50 | 0 | Deaf1<br>Drosophila melanogaster | BAS1<br>SWI6<br>ARG80 |

**Table 11.TFBS for ZCTFs that are involved in arginine biosynthesis.**

%targets corresponds to the percentage the motif appears in a the positive learning set (promoters of orf19.1604, orf19.2230, ZCF23, ZCF35, ARO80, HAL9, LYS142 and TEA1) , and %background corresponds to the percentage that the same motif appears in the negative learning set consisting of 1000 randomly chosen genes. All p-values here are above $1e^{-10}$. The last column gives the TF name for the motif in *S. cerevisiae*.

# Supplemental Tables

**Supplemental Table 1. *C. albicans* orthologs of the *S. cerevisiae* ESR genes.**

The 642 orthologs of ESR genes in *C. albicans* : *S. cerevisiae* Divided to three groups, RiBi; Ribosomal Biogenesis, RP; Ribosomal Protein, iESR; induced environmental protein.

**Supplemental Table 2. The collection of transcription factor binding site motifs used in the cross-validation of the ZCTF regulatory network.**

The 177 motifs were obtained from the JASPAR database (release 2018).

**Supplemental Table 3. Description of differentially expressed genes.**

Genes from each ZCTFs with (adj p-value < $10^{-4}$) been stratified in different categories (0), (1-4) and (5-9) based on their frequency between TFs.

**Supplemental Table 4. Description of differentially expressed genes in two replicates of FCR1, UME7 and ZCF7.**

Genes from each ZCTFs with (adj p-value < $10^{-4}$) been stratified in different categories (0), (1-4) and (5-9) based on their frequency between TFs.

# Supplemental Figures

**Supplemental Figure 1. The guide alignment and phylogeny used for the transcription factor binding site analysis in *C. albicans* for FCR1.**

The alignment was derived using CLUSTALW and the tree was derived using Phylip. See **Methods 11**.

# References

Fitzpatrick DA, Logue ME, Stajich JE, Butler G. "A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis". BMC Evolutionary Biology. 2006; 6: 99. doi:10.1186/1471-2148-6-99

Stefanie Mühlhausen MK. Molecular Phylogeny of Sequenced Saccharomycetes Reveals Polyphyly of the Alternative Yeast Codon Usage. Genome Biol Evol. Oxford University Press; 2014;6: 3222.

Meyers FH, Jawetz E, Goldfien A. Review of Medical Pharmacology (6th ed.). Lange Medical Publications. ISBN 1978; 978-0-87041-151-9.

Trimble JR. The use of a precipitin test to differentiate Candida albicans from Candida stellatoidea. J Invest Dermatol. 1957;28: 349–358.

Bennett RJ, Forche A, Berman J. Rapid Mechanisms for Generating Genome Diversity: Whole Ploidy Shifts, Aneuploidy, and Loss of Heterozygosity. Cold Spring Harb Perspect Med. Cold Spring Harbor Laboratory Press; 2014; a019604.

Sellam A, Whiteway M. Recent advances on *Candida albicans* biology and virulence. F1000Res. 2016;5: 2582.

Noble SM, Gianetti BA, Witchley JN. Candida albicans cell-type switching and functional plasticity in the mammalian host. Nat Rev Microbiol. 2017;15: 96–108.

Odds FC, Brown AJP, Gow NAR. Candida albicans genome sequence: a platform for genomics in the absence of genetics. Genome Biol. 2004;5: 230.

Maicas S, Moreno I, Nieto A, Gómez M, Sentandreu R, Valentín E. In silico analysis for transcription factors with Zn(II)(2)C(6) binuclear cluster DNA-binding domains in Candida albicans. Comp Funct Genomics. 2005;6: 345–356.

Klimova N, Yeung R, Kachurina N, Turcotte B. Phenotypic analysis of a family of transcriptional regulators, the zinc cluster proteins, in the human fungal pathogen Candida glabrata. G3 . 2014;4: 931–940.

Schillig R, Morschhäuser J. Analysis of a fungus-specific transcription factor family, the Candida

albicans zinc cluster proteins, by artificial activation. Mol Microbiol. 2013;89: 1003–1017.

Tebung WA, Choudhury BI, Tebbji F, Morschhäuser J, Whiteway M. Rewiring of the Ppr1 Zinc Cluster Transcription Factor from Purine Catabolism to Pyrimidine Biogenesis in the Saccharomycetaceae. Curr Biol. 2016;26: 1677–1687.

Ghosh S, Kebaara BW, Atkin AL, Nickerson KW. Regulation of aromatic alcohol production in Candida albicans. Appl Environ Microbiol. 2008;74: 7211–7218.

Finkel JS, Xu W, Huang D, Hill EM, Desai JV, Woolford CA, et al. Portrait of Candida albicans adherence regulators. PLoS Pathog. 2012;8: e1002525.

Nantel A. The long hard road to a completed Candida albicans genome. Fungal Genet Biol. 2006;43: 311–315.

Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol. 2010;8: e1000329.

Enjalbert B, Nantel A, Whiteway M. Stress-induced gene expression in Candida albicans: absence of a general stress response. Mol Biol Cell. 2003;14: 1460–1467.

Gasch AP, Yu FB, Hose J, Escalante LE, Place M, Bacher R, et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLoS Biol. 2017;15: e2004050.

Mitrovich QM, Tuch BB, Guthrie C, Johnson AD. Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans. Genome Res. 2007;17: 492–502.

Saccharomyces Genome Database (SGD) [2 November 2018] LYS14.

Arbour M, Epp E, Hogues H, Sellam A, Lacroix C, Rauceo J, et al. Widespread occurrence of chromosomal aneuploidy following the routine production of Candida albicans mutants. FEMS Yeast Res. Wiley-Blackwell; 2009;9: 1070.

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15: 550.

Kumamoto CA, Vinces MD. Contributions of hyphae and hypha-co-regulated genes to Candida albicans virulence. Cell Microbiol. 2005;7: 1546–1554.

Nobile CJ, Mitchell AP. Regulation of Cell-Surface Genes and Biofilm Formation by the C. albicans Transcription Factor Bcr1p. Curr Biol. 2005;15: 1150–1155.

Vandeputte P, Ischer F, Sanglard D, Coste AT. In Vivo Systematic Analysis of Candida albicans Zn2-Cys6 Transcription Factors Mutants for Mice Organ Colonization. PLoS One. Public Library of Science; 2011;6: e26962.

Jungwirth H, Kuchler K. Yeast ABC transporters - A tale of sex, stress, drugs and aging. FEBS Lett. 2005;580: 1131–1138.

Hirakawa MP, Martinez DA, Sakthikumar S, Anderson MZ, Berlin A, Gujja S, et al. Genetic and phenotypic intra-species variation in Candida albicans. Genome Res. Cold Spring Harbor Laboratory Press; 2015;25: 413.

Ghosh S, Dhammika H M L, Roberts DD, Cooper JT, Atkin AL, Petro TM, et al. Arginine-Induced Germ Tube Formation in Candida albicans Is Essential for Escape from Murine Macrophage Line RAW 264.7. Infect Immun. American Society for Microbiology Journals; 2009;77: 1596–1605.

Tao L, Zhang Y, Fan S, Nobile CJ, Guan G, Huang G. Integration of the tricarboxylic acid (TCA) cycle with cAMP signaling and Sfl2 pathways in the regulation of CO2 sensing and hyphal development in Candida albicans. PLoS Genet. Public Library of Science; 2017;13. doi:10.1371/journal.pgen.1006949

Candida Genome Database [28 October 2018] C_albicans_SC5314_A22_current_chromosomal_feature.tab Assembly version A22.

Candida Genome Database [cited 29 Oct 2018] C_albicans_SC5314_A22_current_chromosomes.fasta.gz Assembly version A22.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33: 290–295.

R Core Team. R: A language and environment for statistical computing. 2018; R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ .

RStudio Team. RStudio: Integrated Development for R. 2016; RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32: 2847–2849.

Shen L, Sinai M. GeneOverlap: Test and visualize gene overlaps. 2013; R package version 1.16.0, http://shenlab-sinai.github.io/shenlab-sinai/

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003; 13 (11):2498-504. https://doi.org/10.1101/gr.1239303 PMID: 14597658; PubMed Central PMCID: PMCPMC403769.

Ono K, Muetze T, Kolishovski G, Shannon P, Demchak B. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. F1000Res. 2015;4: 478.

Ignatiadis N, Huber W . Covariate-powered weighted multiple testing with false discovery rate control. arXiv. 2017 01. doi: arXiv:1701.05179.

Kassambara A, Fabian M. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. 2017 R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

Candida Genome Database [10 November 2018] C_albicans_SC5314_S_cerevisiae_orthologs.txt Assembly version A21.

Candida Gene Order Browser [29 October 2018].

Fitzpatrick DA, O'Gaora P, Byrne KP, Butler G. Analysis of gene evolution and metabolic pathways using the Candida Gene Order Browser. BMC Genomics. BioMed Central; 2010;11: 290.

Candida Genome Database [10 November 2018] C_albicans_SC5314_S_cerevisiae_best_hits.txt Assembly version A21.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–1240.

BLAST: Basic Local Alignment Search Tool [16 November 2018].

Wilkinson L. ggplot2: Elegant Graphics for Data Analysis by WICKHAM, H. Biometrics. 2011;67: 678–679.

Candida Genome Database [31 October 2018]. gene_association.cgd.gz.

Morgan M, Falcon S, Gentleman R GSEABase: Gene set enrichment data structures and methods. 2018; R package version 1.44.0.

Falcon S, Gentleman R, Using GOstats to test gene lists for GO term association. Bioinformatics, 2007; 23(2):257-8.

Candida Genome Database [4 November 2018] All_Species_Orthologs_from_CGOB.txt Assembly version A21.

Dineen D. Clustal W and Clustal X Multiple Sequence Alignment [4 November 2018].

Larkin MA E al. Clustal W and Clustal X version 2.0. - PubMed - NCBI. [4 November 2018].

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202–8.

Felsenstein J. Evolutionary Trees From Gene Frequencies and Quantitative Characters: Finding Maximum Likelihood Estimates. Evolution. 1981;35: 1229.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010;38: 576–589.

Jiang R, Zhang X, Zhang MQ. Basics of Bioinformatics: Lecture Notes of the Graduate Summer School on Bioinformatics of China. Springer Science & Business Media; 2013.

Uhl MA, Biery M, Craig N, Johnson AD. Haploinsufficiency-based large-scale forward genetic analysis of filamentous growth in the diploid human fungal pathogen C. albicans. EMBO J. 2003;22: 2668–2678.

Ramsdale M, Selway L, Stead D, Walker J, Yin Z, Nicholls SM, et al. MNL1 Regulates Weak Acid–induced Stress Responses of the Fungal Pathogen Candida albicans. Mol Biol Cell. 2008;19: 4393–4403.

Smith JJ, Ramsey SA, Marelli M, Marzolf B, Hwang D, Saleem RA, et al. Transcriptional responses to fatty acid are coordinated by combinatorial control. Mol Syst Biol. 2007;3: 115.

Coste AT, Ramsdale M, Ischer F, Sanglard D. Divergent functions of three Candida albicans zinc-cluster transcription factors (CTA4, ASG1 and CTF1) complementing pleiotropic drug resistance in Saccharomyces cerevisiae. Microbiology. Microbiology Society; 2008;154: 1491–1501.

Gray WM, Fassler JS. Isolation and analysis of the yeast TEA1 gene, which encodes a zinc cluster Ty enhancer-binding protein. Mol Cell Biol. 1996;16: 347–358.

MacPherson S, Larochelle M, Turcotte B. A fungal family of transcriptional regulators: the zinc cluster proteins. Microbiol Mol Biol Rev. 2006;70: 583–604.