The Effects of Host-like Environmental Signals and Gene Expression on Capsule Growth in Cryptococcus neoformans

Yu Min Jung
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/eng_etds

Part of the Bioinformatics Commons, Engineering Commons, and the Systems Biology Commons

Recommended Citation

This Thesis is brought to you for free and open access by the McKelvey School of Engineering at Washington University Open Scholarship. It has been accepted for inclusion in McKelvey School of Engineering Theses & Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
The Effects of Host-like Environmental Signals and Gene Expression on Capsule Growth in Cryptococcus neoformans

by

Yu Min Jung

A thesis presented to
the McKelvey School of Engineering
of Washington University
in partial fulfillment of the requirements for the degree
of Master of Science

August 2022

St. Louis, Missouri
# Table of Contents

List of Figures iii
List of Tables iv
Acknowledgements v
Abstract vi
Chapter 1: Introduction 1
Chapter 2: Dataset 4
Chapter 3: Methods 5
  - Microscopy and manual image annotation 5
  - Calculating max $\chi^2$ for each gene 5
  - Forming metagenes 5
  - XGBoost model 6
Chapter 4: Results 7
  - Growth medium has a big effect on capsule size, temperature and pH have small effects, and CO2 and cAMP have intermediate effects 7
  - Effects of signals on gene expression mirror their effects on capsule size 10
  - Effects of cyclic AMP on cell size, capsule size, and gene expression 12
  - Identifying genes whose expression levels are associated with capsule size 16
  - Testing predicted capsule-associated genes 21
Chapter 5: Discussion/Conclusion 23
References 26
Supplementary Materials 30
  - Expanded Methods 32
    - RNA-Seq lab methods and processing pipeline 32
    - Phenotyping 33
    - Selection of genes for experiments 33
List of Figures

Figure 1: Relationship between signals and capsule sizes 7
Figure 2: Relationship between signals except exogenous cAMP and gene expression 9
Figure 3: Relationship between exogenous cAMP and gene expression 13
Figure 4: Sample plots from gene selection framework 16
Figure 5: Sample results from initial set of experiments 21
Figure S1: Effects of signals and significant interactions in a logistic regression model of capsule size 30
Figure S2: Capsule index as a function of cAMP concentration 30
Figure S3: Capsule sizes of tested mutants in non-inducing conditions 31
List of Tables

Table 1: Selected genes and their preliminary experiment results from $\chi^2$ method  
Table 2: Selected genes and their preliminary experiment results from SHAP method  
Table S1: Explained variance in models with corresponding factors in the columns and labels in the rows in various combinations of environmental signals  
Table S2: Explained variance in models with corresponding factors in the columns and labels in the rows in mutant testing dataset  

iv
Acknowledgements

First and foremost, I would like to thank my thesis advisor Michael Brent for his guidance. He has always provided me with generous support and valuable advice in my research. Thank you also to everyone I have met during my time in the Brent lab.

Thank you to my family and friends who helped me manage the ups and downs of research with their unconditional support and love. This thesis would not be exist without them.

Yu Min Jung

Washington University in St. Louis

August 2022
ABSTRACT OF THE THESIS

The Effects of Host-like Environmental Signals and Gene Expression on Capsule Growth in Cryptococcus neoformans

by

Yu Min Jung

Master of Science in Computer Science

Washington University in St. Louis, 2022

Professor Michael Brent, Chair

Cryptococcus neoformans is a fungal pathogen that causes cryptococcosis, a disease that kills almost 200,000 people worldwide each year. A unique feature of this deadly yeast is its polysaccharide capsule, which is known to be associated with its virulence. Here, we systematically explore the effects of all possible combinations of 4 capsule-inducing signals on gene expression, cell size, and capsule size. These signals are medium (YPD, DMEM or RPMI), temperature (30°C or 37°C), CO₂ (room air or 5%), cAMP (0 mM or 20 mM), and pH buffer (HEPES/no HEPES). We explore the effects of exogenous cAMP at a range of concentrations and of deletions of cAMP pathway genes PKR1, PDE1, and PDE2. We present a computational framework for identifying a set of genes that are putatively associated with capsule induction. In addition, we follow up with experiments to validate and corroborate the hypotheses from the computational analysis in search of new genes that may be involved in capsule growth.
Chapter 1: Introduction

Late in the 19th century, several scientific articles described a budding yeast with a distinctive capsule, which came to be known as Cryptococcus neoformans [1, 2]. Today, we know C. neoformans as the causative agent of deadly meningitis that kills roughly 200,000 people each year worldwide [3]. We also know much about the capsule that surrounds this pathogen, including the chemical structure of the polysaccharides that compose it [4-6], its key role in disease [7-9], and the fact that it is exquisitely sensitive to environmental conditions [10-13]. Upon entry to a mammalian host, for example, the capsule dramatically increases in thickness, from a barely perceptible structure to a distinctive shell whose thickness can exceed the cell’s diameter [13, 14]. Enlarged capsules inhibit phagocytosis of the yeast by host immune cells, and shed capsule polysaccharides inhibit host defenses [8, 15]. The importance of this material in cryptococcal virulence is supported by the reduced virulence of strains in which capsule is altered or dysregulated [16].

Multiple in vitro growth conditions induce the formation of cryptococcal capsule. Conditions that reflect aspects of the mammalian host environment are of particular interest, such as those that include tissue culture media or mammalian serum, human body temperature (37°C), host-like CO$_2$ concentrations (~5% atmosphere), host-like pH, and host-like nutrient limitations. We have previously shown that transcription is required for capsule growth [17]. Here, our goal is to understand how environmental signals are transmitted via specific gene expression programs to ultimately yield enlarged capsules. To do this we have isolated key
features of capsule inducing conditions, which we call signals, and assessed their effects on cell size, capsule size, and gene expression over time during the induction process.

The signaling pathways that lead from glucose sensing to production of cyclic AMP, activation of Protein Kinase A, and some of their effects on transcription factor activity, have been extensively studied in *Saccharomyces cerevisiae* [18, 19]. When activated, Cac1 (adenylyl cyclase) converts ATP to cAMP, which causes the repressive subunits of the Protein Kinase A complex, Pkr1, to separate from the catalytic subunit, Pka1. This activates Pka1, allowing it to phosphorylate transcription factors including Nrg1 [20] and Rim101 [21]. Pde1 and Pde2 (phosphodiesterases) convert cAMP to AMP, reducing cAMP signaling. Much of this machinery is also found in *C. neoformans* [22]. In *C. neoformans*, cAMP signaling is required for capsule growth and virulence. Cells lacking Cac1 [13, 23] or Pka1 [21, 24] fail to make capsule in conditions that induce capsule in WT cells. Pkr1-deficient cells make larger capsules than WT cells in capsule-inducing conditions [13, 25], but they have not yet been shown to grow capsule in non-inducing conditions. Previous evidence suggests that the capsules of *pkr1* mutants in YPD are the same size as those of WT cells. Thus, while cAMP signaling is necessary for capsule growth, it may not be sufficient.

Here, we systematically explore the effects of all possible combinations of 4 capsule-inducing signals on gene expression, cell size, and capsule size. These signals are tissue-culture medium (DMEM or RPMI), temperature (37°C), CO₂ (5%), and pH buffer. For each combination of conditions, we grow cultures with or without exogenous cAMP, take samples for RNA-Seq at 0, 30, 90, 180, 1440 minutes. We also measure capsule thickness at 1440 min.
hrs) on images of samples that were negatively stained with India ink. We explore the effects of exogenous cAMP at a range of concentrations and of deletions of cAMP pathway genes PKR1, PDE1, and PDE2. In total, we have analyzed 745 RNA-Seq samples and 25,684 India Ink micrographs. This has given us an unprecedented ability to trace capsule-inducing signals through their effects on gene expression over 24 hours to their ultimate effects on capsule size. It has also allowed us to observe interactions between inducing signals. Notably, 5% CO2 and 20 mM cAMP increase the capsule thickness of cells growing in tissue culture medium, but have no effect on the capsules of cells growing in YPD (rich medium). Activating the cAMP pathway by deleting PKR1 and, to a lesser extent, PDE1 or PDE2, increases capsule sizes in cells growing on RPMI medium. This dataset has allowed us to see how the expression of capsule-involved genes responds to various concentrations of cAMP and to identify new genes with similar responses. Finally, it has enabled us to home in on the changes in gene expression that are most consistently associated with capsule induction across a wide range of environmental conditions and genetic perturbations. These analyses point to new genes that may be involved in processes necessary for capsule growth.
Chapter 2: Dataset

We explored the effects of various signals such as medium (YPD, DMEM or RPMI), temperature (30°C or 37°C), CO₂ (room or 5%), cAMP (0 mM or 20 mM) and pH buffer (HEPES or no HEPES) on gene expression, cell size, and capsule size. The HEPES buffer was added to adjust the pH level to 7.0. To do this, we grew cells in all possible combinations of these signals (3 * 2^4 = 48 combinations) and took samples for RNA-Seq at 0, 30, 90, 180, and 1440 minutes. In addition, we imaged the cells in the presence of India Ink and measured cell and capsule sizes at 1440 minutes. At least 4 replicates were grown in each combination of conditions.

In addition, we investigated the effects of exogenous cAMP at a range of concentrations (0 mM, 1.08 mM, 1.8 mM, 3.3 mM, 6.0 mM, 11.0 mM, and 20.0 mM) and of cAMP pathway genes PKR1, PDE1, and PDE2. At least 3 replicates were sequenced, imaged and annotated for these experiments. In total, we analyzed 745 RNA-Seq samples and 25,684 India Ink micrographs.

To test the hypothesized capsule-inducing genes from our model, we tested 47 gene deletion strains in either inducing (RPMI, 37°C, 5% CO₂), almost-inducing (RPMI, 37°C, no CO₂), or non-inducing (YPD, 37°C, 5% CO₂, 20 mM cAMP, and YPD, 37°C, 5% CO₂) conditions. Some genes were tested in more than one growth condition, resulting in total of 74 tests for the first experiment. We then performed additional experiments to confirm mutants that exhibited potential differences in capsule compared to wild-type cells in the first round of experiments.
Chapter 3: Methods

Microscopy and manual image annotation

Cells suspensions were mixed with India ink and imaged at 63X magnification. Images were then manually annotated using a custom annotation interface. Fifteen fields were annotated for each replicate of each combination, yielding an average of 92 annotated cells per replicate (SD 56).

Calculating max $\chi^2$ for each gene

To calculate the $\chi^2$ statistics to associate with each gene’s expression pattern in the dataset, we compared the distribution of every gene’s expression levels for induced group to the uninduced group. We defined the threshold for induction as 1.13 µm based on the distribution of uninduced levels of capsule sizes. This threshold was calculated by taking the mean of the conditions in which we do not expect any capsule induction (any samples grown in YPD) and adding 3 standard deviations to include all YPD samples in this group. We calculated $\chi^2$ based on this induction threshold for every split of gene expression level in our data. The associated $\chi^2$ to each gene reflects the maximum $\chi^2$ statistics.

Forming metagenes

From the original gene expression data matrix ($p \times n$), where $p$ is the number of genes and $n$ is the number of samples, we sought to decrease the number of features by filtering and
combining genes. We first removed low variance genes in our data by filtering out genes where 95% of the gene expression levels were within LFC of 1 from the mean. Then, we combined highly correlated genes into metagenes (correlation threshold > 0.8). The “expression level” of a metagene was then set to the mean of the expression levels of the genes it represents. We iteratively repeated this process to combine genes into metagenes until we had combined all correlated genes into metagenes.

**XGBoost model**

Using the gene expression data matrix of metagenes $p \times n$, we trained a model for each time point that maximized model’s capsule size predictive performance. The features were normalized prior to training. The final parameters for the model were selected using 10-fold cross validation (max_depth=4, n_estimators=50). We investigated the use of regularization parameter, $\gamma$, and did not find significant improvement in the model’s predictive performance.
Chapter 4: Results

FIGURE 1: RELATIONSHIP BETWEEN SIGNALS AND CAPSULE SIZES

A) Effects of signals and significant interactions in a linear model of capsule size, B) Capsule sizes for various combinations of significant signals, C) Conditions in which each signal has the greatest influence.

Growth medium has a big effect on capsule size, temperature and pH have small effects, and CO₂ and cAMP have intermediate effects

We first examined how each individual variable affects capsule size, adjusting for the independent effects of all other variables. To do this, we built a linear regression model with the environmental variables as predictors of capsule size. The resulting regression coefficients showed that the biggest impact on capsule size was RPMI medium, which increased capsule width an average of (1.0 µm), followed by DMEM (0.62 µm), CO₂ (0.46 µm), and cAMP (0.45 µm) (Fig. 1A). All these effects were statistically significant (P< 0.001), but the much smaller
effects of adding HEPES buffer or increasing temperature from 30°C to 37°C were not significant. As a result, we combined replicates, regardless of temperature or added HEPES, and plotted capsule size for all possible combinations of medium, CO₂, and cAMP (Fig. 1B). Strikingly, the results showed that there was little to no capsule induction in any condition with YPD, regardless of CO₂ or cAMP. Capsule could be induced in either DMEM or RPMI, but capsules were bigger in RPMI in every combination of CO₂ and cAMP. DMEM alone yielded negligible induction, but the combination of DMEM with CO₂ or cAMP increased capsule sizes. CO₂ had a bigger effect than cAMP and the combination of both yielded the largest capsule. The results in RPMI were similar, except that RPMI alone yielded some induction. cAMP had a bigger effect in RPMI than in DMEM, to the extent that cAMP and CO₂ were equally effective in RPMI. As in DMEM, adding both yielded the largest capsules. A non-linear, tree-based regression algorithm and logistic regression on dichotomized capsule sizes yielded similar results (Supplemental Fig. S1).

We also investigated the combinations of conditions in which RPMI, DMEM, CO₂, and cAMP had the largest effects on capsule size. RPMI and DMEM had their biggest effects with CO₂ and cAMP at 37°C, CO₂ had the biggest effect in DMEM at 37°C, and cAMP had the biggest effect in RPMI with CO₂ at 37°C (Fig. 1C).

In summary, none of the treatments we investigated could induce capsule at all in YPD. RPMI generally led to greater induction than DMEM, holding all other variables constant. In RPMI or DMEM, CO₂ or cAMP each increased capsule size, and the combination of the two yielded the largest capsules.
FIGURE 2: RELATIONSHIP BETWEEN SIGNALS EXCEPT EXOGENOUS CAMP AND GENE EXPRESSION

A) Dendrogram of all treatments clustered by expression profile, B-E) 2D map(s) of signals by gene expression profile in each time point, F) Heatmap of DE genes, G) Lineplot of up regulated genes that are modulated downward by CO$_2$, and H) Lineplot of down regulated genes that are generally modulated upward by CO$_2$. 

A

B

C

D

E

F

G

H
Effects of signals on gene expression mirror their effects on capsule size

During the capsule inductions described above, we took samples for RNA-Seq before starting the induction and at 30, 90, 180, and 1440 min. After quality control (see Methods), we were left with RNA-Seq data on 617 samples. At least 4 replicates for each combination of environmental signals were grown on separate dates and library-prepped on separate dates to reduce the effect of day-to-day variability in growth and RNA-Seq library preparation. Each growth batch and library preparation batch included a control culture in standard laboratory conditions: YPD, 30°C, no CO$_2$, no cAMP, no added buffer. Taking advantage of this design, we regressed out the batch (date) effects. We then computed the Euclidean distances between logged, normalized counts for all pairs of samples and carried out hierarchical clustering (Fig. 2A). Remarkably, all replicates with the same medium, atmosphere, temperature, and cAMP status clustered together, indicating highly effective batch correction. Consistent with its effects on capsule size, medium had the biggest effect on gene expression state, followed by CO$_2$. Interestingly, temperature had a greater effect on gene expression than cAMP, while cAMP had the greater effect on capsule size.

Next, we carried out principal components analysis (PCA) on the log normalized expression levels of all genes in all samples. We plotted samples for each time point separately (Fig. 2B-E). There is an overall migration of all points toward the lower left with time, but the points representing cells in tissue culture medium (DMEM and RPMI) move more than the YPD points. Arrows show the effects of DMEM (green) or RPMI (orange), relative to YPD, all in the absence of CO$_2$, and the effects of CO$_2$ in those tissue culture media (red lines). At 30 and 90
minutes, the effects of CO$_2$ are largely orthogonal to those of the tissue culture media, indicating very different effects on gene expression. By 24 hrs, however, the differences among all cultures are smaller and the effects of CO$_2$ more aligned with those of tissue culture media. While DMEM and RPMI move cells to slightly different areas, the further addition of CO$_2$ causes them to move to essentially the same gene expression state.

To gain further insight into the effects of tissue culture media and CO$_2$ on gene expression, we built a linear model that predicts log normalized gene expression levels from signals in DESeq2. We used the model to calculate shrunken log fold changes (see Methods) in response to each signal. Next, we carried out Gene Ontology (GO) enrichment analysis on the genes that were most responsive to DMEM or RPMI, CO$_2$, and the combination of both, at each time point. From among the enriched GO biological process terms, we selected those that were most interesting and interpretable and created a heatmap showing the average expression levels of all genes annotated with each term (not just the significantly differentially expressed genes; Fig. 2F). For each time point, we show the effects of DMEM relative to YPD, the effects of DMEM+CO$_2$ relative to DMEM alone, and the combined effects of DMEM+CO$_2$ relative to YPD alone. The GO annotations divide into two broad groups, one increased by DMEM and one decreased by DMEM. The up regulated gene sets (red) involved cell wall synthesis, adaptation to lower glucose, and stress response. For example, induced genes involved in carbohydrate biosynthesis (Fig. 2G) included those involved in gluconeogenesis, glycogen synthesis, and trehalose synthesis, all of which are typical of adaptation to lower glucose. The down regulated gene sets mainly involved growth, cell division, and carbohydrate transport. For example, many
genes involved in nuclear chromosome segregation are down regulated (Fig. 2H). Interestingly, for most gene sets in the heat map, the effect of CO\textsubscript{2}, when added to DMEM, went in the opposite direction from the effect of DMEM relative to YPD. However, the CO\textsubscript{2} response was weaker so the overall effect of DMEM+CO\textsubscript{2} goes in the same direction as DMEM alone. Exceptions to that rule are ribosome biogenesis and carbohydrate transport, which are consistently down regulated. Among the down regulated carbohydrate transport genes, we see LPI8 [26], which affects phagocytosis, and GMT2, a mannose transporter [27].

**Effects of cyclic AMP on cell size, capsule size, and gene expression**

Cyclic AMP signaling is known to be essential for capsule growth [28]. To investigate the effects of exogenous cAMP on capsule size and gene expression, we carried out experiments in conditions in which capsule growth is normally relatively small, but addition of CO\textsubscript{2} or cAMP can generate much larger capsules: RPMI, 30°C, room air, no HEPES. We added cAMP at 0, 1.08, 1.8, 3.3, 11, and 20 mM (increasing by a factor of ~1.8 at each step). Capsule width showed a consistent dose response (Fig. 3A). The logarithm of the exogenous cAMP concentration was a highly significant predictor of average capsule size (P < 10-10). Increasing cAMP by a factor of 1.8 increased the capsule size by 0.2 µm on average, but the final step from 11 mM to 20 mM had a much bigger effect. Interestingly, exogenous cAMP also increased the average radius to the cell wall (P < 3*10-5) with a 1.8 fold increase in cAMP causing an average increase of 0.11 µm (Fig. 3B). Unlike capsule width, the effect on cell size was greater at smaller concentrations – the difference between 0 mM and 3.3mM was 0.46 µm while the difference from there to 20mM was
only 0.18 µm. The effect on cell size cannot fully explain the effect on capsule, since the capsule index (capsule width as a fraction of total radius) increased with cAMP concentration (Supplemental Fig. S2).

**FIGURE 3: RELATIONSHIP BETWEEN EXOGENOUS CAMP AND GENE EXPRESSION**

A) Capsule thickness as a function of cAMP concentration, B) Cell radius as a function of exogenous cAMP concentration, C) Effects of cAMP 20 mM and gene deletion mutants in the cAMP pathway (*pkr1*, *pde1*, and *pde2*) on capsule size in RPMI 30°C no CO2, D) Effects of cAMP 20 mM and KO/OE of mutants in the cAMP pathway (*pkr1*, *pde1*, and *pde2*) on cell size in RPMI 30°C no CO2, F) Heatmap of DE genes, G) Lineplot of NCR genes
To further investigate the role of the cAMP, we measured the capsules of cells lacking \( PDE1, PDE2, \) or \( PKR1, \) growing in either YPD or RPMI (30°C no CO\(_2\) or cAMP; Fig 3C). \( PDE1 \) and \( PDE2 \) encode phosphodiesterases capable of reducing the intracellular cAMP concentration, although we subsequently showed that \( PDE2 \) has a frame shift and early stop codon in the H99 background (Daniel Aghostino, personal communication). \( PDE1 \) and \( PDE2 \) deletions showed some effect on capsule sizes in RPMI in our data. \( PKR1 \) encodes the repressive moiety of the Protein Kinase A (PKA) complex. cAMP causes Pkr1 to dissociate from the complex, activating the kinase moiety. Thus, the \( pkr1 \) deletion mutant might be expected to have an enlarged capsule and indeed we see that, when grown in RPMI without CO\(_2\) or cAMP, the mutant’s capsule is similar to that of WT grown in RPMI with 20 mM cAMP (Fig. 3C). This suggests that the main effect of cAMP on capsule size is mediated by its effect on Pkr1. We also saw an increase in capsule size of \( pkr1 \) growing in YPD. Although it has been shown previously that \( pkr1 \) has enlarged capsule when growing in fully capsule-inducing conditions (low iron medium with EDDHA [29] or DMEM 37°C 5% CO\(_2\) [13]), we have found no previous evidence that its capsule is enlarged in conditions such as RPMI or YPD at 30°C without other inducing factors. No other perturbation, including 20 mM exogenous cAMP, increased capsule size in YPD (Fig. 1B). Thus, the repressive effect of YPD may be upstream of PKA activation, such that direct activation by \( PKR1 \) deletion partially overcomes that repression. Another difference between \( PKR1 \) deletion and 20 mM cAMP is that, in RPMI, the deletion does not increase cell body size as much as the cAMP (Fig. 3D). This suggests that Pkr1 in activation may not be the only pathway by which cAMP increases cell size.
We then characterized the subsets of cAMP responsive genes at each time point as in Figure 2F. Overall, the RPMI and cAMP responses were broadly similar to the DMEM and CO₂ responses, respectively (Fig. 3F). The effect of adding cAMP or CO₂ tended to moderate the responses to tissue culture media more often than it reinforced them. However, there were differences. Both tissue culture media reduced the expression of ribosome biogenesis genes, but adding CO₂ reinforced that effect at all time points, while adding cAMP moderated it at later time points. Genes involved in carboxylic acid metabolic processes (mainly nitrogen assimilation and amino acid metabolism) also responded differently, especially at early time points. We noticed that many of these genes were targets of nitrogen catabolite repression (NCR), a process that represses the expression of certain genes when preferred nitrogen sources such as glutamine (the main nitrogen source in DMEM and RPMI) are available [30]. RPMI engaged NCR, repressing NCR genes. However, adding cAMP released NCR rapidly, most notably at 30 min (Fig. 3G), suggesting that the cells responded to cAMP as though they had been moved to a less preferred nitrogen source (despite the presence of glutamine in the medium). This overcame the RPMI effect, so that the combined effect of RPMI and cAMP was to relieve NCR (consistent with less preferred carbon sources), relative to YPD.
Identifying genes whose expression levels are associated with capsule size

Next, we set out to identify genes whose expression is strongly associated with capsule development in the growth conditions tested. The first approach was to dichotomize capsule size into “induced” or “not induced”, dichotomize the expression of each gene into “high” or “low”, and analyze the relationship between expression and capsule induction for each gene separately. Since YPD completely blocks capsule induction regardless of other signals, we set a threshold for “induced” at three standard deviations above the mean for all cells in YPD. Using this
method, replicates with mean capsule width wider than 1.13 µm were deemed “induced”. We searched for the most meaningful split between high and low expression for each gene by trying each observed mean level as a possible threshold. For each possible threshold, we calculated the $\chi^2$ statistic for association between induction status and expression status. We were not concerned with P-values here, but with using the statistic itself to rank genes by their association with capsule size. The maximum $\chi^2$ over all time points was assigned to the gene (Fig. 4A, B).

The genes with the highest $\chi^2$ values were identified as possibly involved in capsule induction and slated for testing (Table 1). Each candidate gene was classified as being more highly expressed in either induced samples or uninduced samples. For genes that were more highly expressed in induced samples, we grew cells lacking the gene in capsule-inducing conditions to check for a reduction of capsule size. For genes that were more highly expressed in uninduced samples, we grew cells lacking the gene in both non-inducing conditions (YPD, 30°C, room air, no cAMP, no HEPES) and partially inducing conditions (RPMI, 5% CO$_2$, 30°C, no cAMP, no HEPES) to check for increased capsule size.

In addition to using dichotomized values and single genes, we used continuous expression and capsule size values to train a machine learning (ML) model to predict capsule size from expression levels of all genes at all time points. Specifically, we used XGBoost [31], a tree boosting algorithm, to train a random forest model that takes all genes at all time points as features and predicts capsule width. Before training, we merged features into metagenes (see Methods) by considering variability and correlation of genes across all RNA-Seq samples. This avoids the tendency of regression models to choose between correlated features arbitrarily or to
be overly sensitive to small fluctuations in correlated features. We then calculated a SHAP value for each feature in each prediction. The SHAP value explains a prediction in terms of how much each feature value influenced the model to increase or decrease its capsule width prediction for that sample, relative to the overall average capsule width (Fig. 4C, D). To calculate a single influence score for each feature, we summed the absolute SHAP values for that feature over all the samples. The genes with the largest influence scores were slated for testing (Table 2). Conditions in which to test were chosen as for the $\chi^2$ method.
**TABLE 1: SELECTED GENES AND THEIR PRELIMINARY EXPERIMENT RESULTS FROM $\chi^2$ METHOD**

Genes were selected based on the univariate $\chi^2$ analysis. The top ranking genes are shown here with the predicted direction of change in the gene deletion mutant and the observed result of the experiment in either inducing (I) or almost-inducing (AI) condition.
Genes were selected based on absolute SHAP values of each gene in the XGBoost model. The top ranking genes are shown here with the predicted direction of change in the gene deletion mutant and the observed result of the experiment in either inducing (I) or almost-inducing (AI) condition. Genes denoted with * next to the CNAG numbers indicate that they were chosen with the new XGBoost model trained without previously selected genes.

<table>
<thead>
<tr>
<th>CNAG</th>
<th>Name</th>
<th>Ortholog</th>
<th>Description</th>
<th>SHAP in I</th>
<th>SHAP in AI</th>
<th>Prediction</th>
<th>Algorithmic Call I</th>
<th>Algorithmic Call AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>03078</td>
<td>NPP1</td>
<td>NPP1</td>
<td>type I phosphodiesterase/nucleotide pyrophosphatase</td>
<td>-0.299</td>
<td>HYPO</td>
<td>No mutant</td>
<td>No mutant</td>
<td></td>
</tr>
<tr>
<td>01119</td>
<td>PTR2</td>
<td>POT1</td>
<td>POT family proton-dependent oligopeptide transporter</td>
<td>0.2</td>
<td>HYPO</td>
<td>No mutant</td>
<td>No mutant</td>
<td></td>
</tr>
<tr>
<td>03045</td>
<td>MVP1</td>
<td>hypothetical</td>
<td></td>
<td>0.179</td>
<td>HYPER</td>
<td>No effect</td>
<td>No effect</td>
<td></td>
</tr>
<tr>
<td>02005</td>
<td>YBR190W</td>
<td>hypothetical</td>
<td></td>
<td>-0.179</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04677</td>
<td>TCO89</td>
<td>hypothetical</td>
<td></td>
<td>-0.179</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02545*</td>
<td>IPP1</td>
<td>IPP1</td>
<td>Inorganic pyrophosphatase</td>
<td>-0.179</td>
<td>HYPO</td>
<td>No mutant</td>
<td>No mutant</td>
<td></td>
</tr>
<tr>
<td>05050*</td>
<td>GCS2</td>
<td>GCS1</td>
<td>ADP-ribosylation factor GTPase-activating protein 1</td>
<td>-0.179</td>
<td>HYPO</td>
<td>No mutant</td>
<td>No mutant</td>
<td></td>
</tr>
<tr>
<td>01088</td>
<td>BNI1</td>
<td>hypothetical</td>
<td></td>
<td>-0.141</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>01993</td>
<td>HAA1</td>
<td>hypothetical</td>
<td></td>
<td>-0.141</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>06241</td>
<td>CFO1</td>
<td>FET5</td>
<td>acid-laccase</td>
<td>-0.141</td>
<td>HYPO</td>
<td>Hyper</td>
<td>Indeterminate</td>
<td></td>
</tr>
<tr>
<td>06242</td>
<td>CFT1</td>
<td>FTR1</td>
<td>Cryptococcus ferric transporter 1</td>
<td>-0.141</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>07870*</td>
<td>MGM1</td>
<td>hypothetical</td>
<td></td>
<td>0.121</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>03493</td>
<td>PMA1</td>
<td>1-alkyl-2-acetylglycerophosphocholine esterase</td>
<td>-0.112</td>
<td>HYPO</td>
<td>No effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06509*</td>
<td>UGE2</td>
<td>GAL10</td>
<td>UDP-glucose 4-epimerase</td>
<td>0.096</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02539</td>
<td>DIP5</td>
<td>AAT family amino acid transporter</td>
<td>0.090</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>02540</td>
<td>TNA1</td>
<td>hypothetical</td>
<td></td>
<td>0.086</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00368</td>
<td>VPS53</td>
<td>vacuolar-sorting protein 53 long isoform</td>
<td>-0.081</td>
<td>HYPO</td>
<td>Indeterminate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06509*</td>
<td>SMC6</td>
<td>hypothetical</td>
<td></td>
<td>-0.079</td>
<td>HYPO</td>
<td>Indeterminate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>06567*</td>
<td>QIR7</td>
<td>hypothetical</td>
<td></td>
<td>-0.076</td>
<td>HYPO</td>
<td>Indeterminate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04794*</td>
<td>TPO3</td>
<td>spermine transporter</td>
<td>0.071</td>
<td>HYPER</td>
<td>Indeterminate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>04887*</td>
<td>TGL4</td>
<td>hypothetical</td>
<td></td>
<td>0.067</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04103*</td>
<td>ICE2</td>
<td>hypothetical</td>
<td></td>
<td>0.052</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>02478*</td>
<td>GCY1</td>
<td>hypothetical</td>
<td></td>
<td>-0.051</td>
<td>HYPO</td>
<td>Indeterminate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>00871*</td>
<td>CLR3</td>
<td>HAC1</td>
<td>hypothetical (with bZIP DNA binding domain)</td>
<td>-0.050</td>
<td>HYPO</td>
<td>No effect</td>
<td>No effect</td>
<td></td>
</tr>
<tr>
<td>01230</td>
<td>MP98</td>
<td>CDA2</td>
<td>chitin deacetylase 2</td>
<td>0.048</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
<tr>
<td>03007*</td>
<td>ANT1</td>
<td>hypothetical</td>
<td></td>
<td>0.048</td>
<td>HYPER</td>
<td>No effect</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2: SELECTED GENES AND THEIR PRELIMINARY EXPERIMENT RESULTS FROM SHAP METHOD**
FIGURE 5: SAMPLE RESULTS FROM INITIAL SET OF EXPERIMENTS

A-C) Box-and-whisker plots of capsule sizes for WT and mutants, D-E) Pie chart summary of experiment results

Testing predicted capsule-associated genes

We selected genes to test based on $\chi^2$ values or mean absolute SHAP values across all samples. We tested the genes in Table 1 by growing gene deletion mutants, taking India ink micrographs, and annotating them manually (see Methods). The tests included non-inducing conditions (YPD, 37°C, 5% CO$_2$, 20 mM cAMP, no HEPES and YPD, 37°C, 5% CO$_2$, no cAMP, no HEPES), almost-inducing conditions (RPMI, 37°C, room air, no cAMP, no HEPES), or inducing conditions (RPMI, 37°C, 5% CO$_2$, no cAMP, no HEPES). In YPD, we tested mutants in two different levels of exogenous cAMP based on the SHAP values from XGBoost.

We show here three of the tested genes that resulted in notable phenotypes (Fig 5A-C). We note hypercapsular phenotypes for 00586Δ and 02455Δ in almost-inducing conditions and
potentially hypocapsular phenotype in \textit{00328A} in inducing condition. Overall, we tested 27 deletion mutants in almost-inducing conditions and 25 in inducing conditions and identified 2 that made significantly larger capsules in almost-inducing conditions and 3 in inducing conditions (Fig. 5D-E). There were also a few that were labeled indeterminate that will get resolved with more replicates in future experiments. We did not identify any mutants that induced capsule when grown in YPD (Supplemental Fig. S3).
Chapter 5: Discussion/Conclusion

In this thesis, we proposed a computational framework for identifying a set of genes more likely to be associated with capsule induction in *Cryptococcus neoformans*. After observing that there were gene expression patterns that described supposed restrictive and permissive range of gene expression levels associated with resulting capsule sizes, we used $\chi^2$ statistics to quantify the strength of the observed gene expression patterns. Moreover, by training a machine learning model that takes all genes at each time point as features and predicts capsule width, we proposed another method of selecting genes more likely to be involved in capsule induction. The machine learning model was regularized through dimensionality reduction by merging features into metagenes. This allowed the model to avoid arbitrary selection among correlated features and made the model more robust to fluctuations in correlated features.

We used two complementary methods to make predictions about the role of genes in capsule induction. My initial studies support some of these predictions. Through additional experiments, we anticipate that more of these predictions will hold. Although the described application of the framework was for a specific dataset regarding *Cryptococcus neoformans*, the computational framework is more widely applicable to other datasets to analyze any similar phenomenon to uncover causal features of a phenotype in the data.

In addition to the computational framework, we reviewed some of the findings from a comprehensive dataset that Brent Lab generated in order to probe the effect of various environmental signals of interest as they relate to capsule induction. Notably, cells grown in YPD medium did not grow any capsule in our data, suggesting YPD as an effective inhibitor of
capsule induction. In each of the tissue culture medium (DMEM and RPMI), two independent signals, CO₂ and cAMP showed additive effect of capsule induction. We also showed the modulation of capsule sizes with various levels of cAMP concentration.

Gene expression and phenotypic data described in this paper will be a valuable resource for the further study of fungal pathogens. The processed data will be available to the public with this thesis, and Brent Lab hopes to release all raw RNA-seq reads file and image files with the annotations in a public database in the near future. The unprecedented scale of annotation of capsule sizes across a diverse set of conditions from a single laboratory will undoubtedly be of incredible value to the community. It is worth noting that we looked at replicability of capsule induction and how much noise to signal was apparent in our experimental data (Supplemental Table S1). Replicability was high ($r^2 = ~0.90$) within replicate sets in various capsule size measurements in our dataset of various combinations of environmental signals.

We found that although replicability was high in our experiments where we tested capsule induction in various combinations of environmental signals, same could not be said for our mutant experiments (Supplemental Table S2). One possible explanation for such a difference is that the mutant strains could be bad. This was not addressed in this thesis due to time constraints, but strains can be checked through our sequencing data.

With additional gene deletion strains testing on the horizon, we believe that both the univariate $\chi^2$ analysis and multivariate SHAP in XGBoost model analysis will benefit from more diverse set of input data. Gene deletion strains and their expression data allows the model to probe not yet included gene expression space. Despite the incredible scale of the dataset
presented in this thesis, modeling proposes a challenge as the number of features is far greater than the number of samples in our dataset. Moreover, the sample size is significantly smaller if we consider that the samples are not independent from one another as many are biological replicates. This significantly limits our statistical power/ability to make conclusions about individual genes. Perhaps one of the byproducts of this is that we do not see any evidence that the more complex multivariate model is indeed better than the simple univariate model.

Overall, the findings regarding the effects of various environmental signals in our novel dataset and the computational framework described in this paper to select and test set of genes critically associated with capsule induction bring us closer to understanding the transcriptional changes that are required for capsule induction.
References


28


Supplementary Materials

FIGURE S1: EFFECTS OF SIGNALS AND SIGNIFICANT INTERACTIONS IN A LOGISTIC REGRESSION MODEL OF CAPSULE SIZE

FIGURE S2: CAPSULE INDEX AS A FUNCTION OF CAMP CONCENTRATION
FIGURE S3: CAPSULE SIZES OF TESTED MUTANTS IN NON-INDUCING CONDITIONS

TABLE S1: EXPLAINED VARIANCE IN MODELS WITH CORRESPONDING FACTORS IN THE COLUMNS AND LABELS IN THE ROWS IN VARIOUS COMBINATIONS OF ENVIRONMENTAL SIGNALS

<table>
<thead>
<tr>
<th></th>
<th>Combination of growth conditions</th>
<th>Day effects</th>
<th>Growth conditions + Day effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean capsule width</td>
<td>0.80</td>
<td>0.29</td>
<td>0.88</td>
</tr>
<tr>
<td>Log2(mean capsule width)</td>
<td>0.77</td>
<td>0.24</td>
<td>0.86</td>
</tr>
<tr>
<td>80th percentile of capsule width</td>
<td>0.80</td>
<td>0.31</td>
<td>0.87</td>
</tr>
<tr>
<td>Capsule index</td>
<td>0.82</td>
<td>0.21</td>
<td>0.90</td>
</tr>
<tr>
<td>Cell size</td>
<td>0.38</td>
<td>0.64</td>
<td>0.81</td>
</tr>
</tbody>
</table>

TABLE S2: EXPLAINED VARIANCE IN MODELS WITH CORRESPONDING FACTORS IN THE COLUMNS AND LABELS IN THE ROWS IN MUTANT TESTING DATASET

<table>
<thead>
<tr>
<th></th>
<th>Growth conditions and mutant</th>
<th>Day effects</th>
<th>Growth conditions and mutant + Day effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean capsule width</td>
<td>0.21</td>
<td>0.20</td>
<td>0.34</td>
</tr>
<tr>
<td>Log2(mean capsule width)</td>
<td>0.21</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>80th percentile of capsule width</td>
<td>0.26</td>
<td>0.18</td>
<td>0.39</td>
</tr>
<tr>
<td>Capsule index</td>
<td>0.21</td>
<td>0.19</td>
<td>0.35</td>
</tr>
<tr>
<td>Cell size</td>
<td>0.32</td>
<td>0.11</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Expanded Methods

RNA-Seq lab methods and processing pipeline

The reads were aligned with novoalign (version 4.03.02) and quantified with HTSeq (0.9.1) using the FungiDB KN99 genome. The genome annotations were augmented with noncoding RNA regions from the current H99 genome. The quantification for downstream analysis was calculated over gene feature exons, but CDS quantification was also performed for QC purposes. Samtools (version 1.12) was used to convert the SAM output from novoalign to BAM. The Bams were indexed using samtools index. Novoalign and HTSeq logs were collated by MultiQC (version 1.2). Custom scripts were used to calculate coverage over perturbed loci and markers. QC was conducted in two rounds. In the first, we labeled libraries as ‘passing’ or ‘failing’ based on protein coding total reads > 1e6 and not aligned total percent > 7%.

In the second round, we evaluated replicate agreement using the metric Regularized Log Expression [32]. Briefly, we used DESeq (version 1.34.0) to estimate the effect of the libraryDate (the known batch effect). We removed the batch effect using the DESeq coefficients for the libraryDates such that the data was standardized to a single date. The result were quantification, on a log2 normalized scale, with the batch effect removed. Next, the gene-wise median was subtracted from the expression of the given gene in each sample, yielding column vectors which represented how much each gene deviated from the median expression of that gene. If the interquartile range of the deviations from the median of a given sample was greater than 1, then the sample was considered an outlier and failed for replicate agreement.
Phenotyping

Prior to induction the strains were removed from -80°C freezer, streaked on YPD and incubated 2-3 nights at 30°C. The night before induction (1/24/22), single colonies from each plate were inoculated in 50 ml YPD and grown overnight at 30°C with shaking at 250 rpm. The next day (1/25/22), strains were spun down, washed in 25 mls room-temperature respective medium (YPD/DMEM/RPMI), then resuspended in 10 mls preconditioned respective medium. Flasks were incubated in the combination of various environmental signals for 24 hours. Cells were transferred to 1.5mL tubes and fixed and imaged using India ink on a fluorescence microscope at 63X. At least 10 images were collected for each biological replicate. Annotations were made using custom lab software. The capsule thickness was computed as the difference between the diameter of the outer capsule edge and the diameter of the cell wall.

Selection of genes for experiments

For selection of genes based on $\chi^2$, all genes with the highest max $\chi^2$ statistics above 200 were selected. The conditions to test these were selected according to the sign of the correlation of gene expression levels to capsule sizes. For positively correlated genes, the gene deletion mutant was tested in inducing conditions. For negatively correlated genes, the gene deletion mutant was tested in almost inducing conditions and non-inducing condition without added cAMP.
For SHAP value analysis, we trained the XGBoost model on the data in each timepoint. Then, we calculated the SHAP value for each gene in each sample. For SHAP values attributed to a metagene, we gave each gene in the group the same SHAP value. With this, each gene in our data ended up with a SHAP value in each timepoint. We selected the maximum SHAP value across timepoints to associate with each gene. For each condition we sought to test gene deletion mutants in (non-inducing with added cAMP, almost-inducing and inducing), genes with the absolute mean associated SHAP value greater than 0.05 in those conditions were selected for testing. In both analyses, some gene deletion mutants that were not in our collection were omitted from our experiments. To compensate for the number of omitted genes in our selection, we repeated the selection based on SHAP values by training an XGBoost model without the genes already included in the experiments. These genes are denoted with a * next to their CNAG numbers in the Table 2.