Spring 5-15-2015

Strategies for increasing the applicability of biological network inference

Ezekiel John Maier
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science
Department of Computer Science and Engineering

Dissertation Examination Committee:
Michael Brent, Chair
Thomas Baranski
Jeremy Buhler
Ron Cytron
Tamara Doering
Gary Stormo

Strategies for Increasing the Applicability of Biological Network Inference

by
Ezekiel John Maier

A dissertation presented to the
Graduate School of Arts & Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2015
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Appendix Tables

Acknowledgments

I would like to thank the many wonderful people that have helped me on my journey through graduate school. My advisor, Michael Brent, has been a great and flexible advisor who motivated me when I was drifting in my early studies, and inspired my passion for utilizing network knowledge to better understand and engineer cellular state. I would also like to thank the rest of my committee members: Thomas Baranski, whose research on the conserved transcriptional responses to high sugar feeding fueled my initial passion that put me on the path toward my current research; Jeremy Buhler, who challenged me to improve academically; Tamara Doering, who is a great advocate of my computation work and is always willing to test computational predictions; Ron Cytron, who demonstrated many of the qualities of a great teacher and mentored me when I taught; and Gary Stormo, whose expertise on Transcription Factor binding was influential in my research.

I would also like to thank many others in the Washington University in St. Louis community that played important roles in my development. The current and former Brent lab members, including Aaron Tenney, Charlie Comstock, Akshat Shrivastava, Roman Sloutsky, Cole Johnson, and Holly Brown, have contributed to this work through their fellowship, discussion, and feedback. I owe great thanks to Drew Michael and Brian Haynes, who were my closest peers, collaborators, and mentors. I have greatly enjoyed my many collaborations and I’ve gained much of my biological knowledge through questions asked to those collaborators, particularly Laura Musselman and Stacey Gish. I would also like to thank the members of the Cohen lab for always bringing thoughtful scientific discussions to combined lab meetings. Finally, I would like to

thank a few of my best friends, including Brian Haynes, David Lu!!, Drew Michael, Diana O'Brien, Yi Qian, and Scott Satkin, with whom I drank much beer, played much poker and pool, and had many fun arguments.

I would like to thank my family, for always being loving and supportive, while also pushing me. My parents, Bob and Melinda, deserve much of the credit for instilling my work-ethic, and turning a 1st grade math cheater into a Computer Scientist. My siblings, Tessa and Max, have always supported me, even though I may have been a bossy older brother. My dog, Tucker, would often console me when I was feeling defeated. Finally, I would like to thank my wonderful wife, Jaci, to whom I dedicate this dissertation. Despite the challenges of living far apart for nearly two years, she was my rock, always believing in, supporting and pushing me.

Ezekiel John Maier

Washington University in St. Louis

May 2015

ABSTRACT OF THE DISSERTATION

Strategies for Increasing the Applicability of Biological Network Inference
by
Ezekiel John Maier
Doctor of Philosophy in Computer Science
Washington University in St. Louis, 2015
Professor Michael Brent, Chair

The manipulation of cellular state has many promising applications, including stem cell biology and regenerative medicine, biofuel production, and stress resistant crop development. The construction of interaction maps promises to enhance our ability to engineer cellular behavior. Within the last 15 years, many methods have been developed to infer the structure of the gene regulatory interaction map from gene abundance snapshots provided by high-throughput experimental data. However, relatively little research has focused on using gene regulatory network models for the prediction and manipulation of cellular behavior. This dissertation examines and applies strategies to utilize the predictive power of gene network models to guide experimentation and engineering efforts. First, we developed methods to improve gene network models by integrating interaction evidence sources, in order to utilize the full predictive power of the models. Next, we explored the power of networks models to guide experimental efforts through inference and analysis of a regulatory network in the pathogenic fungus *Cryptococcus neoformans*. Finally, we develop a novel, network-guided algorithm to select genetic interventions for engineering transcriptional state. We apply this method to select intervention strains for improving biofuel production in a mixed glucose-xylose environment. The contributions in this dissertation provide the first thorough examination, systematic application, and quantitative evaluation of the utilization of network models for guiding cellular engineering.

Chapter 1: Introduction

A paradigm shift has occurred in biological research in the past 15 years, from the study of only a small number of genes at a time, to the system wide study of molecular interactions in a cell. This rapidly expanding field of Systems Biology focuses on understanding cellular processes by mapping and modeling the cellular interactome. The construction of interaction maps promises to enable us to predict and manipulate cellular behavior, which will have many important applications such as human disease treatment and stress resistant crop development. Determining the transcriptional regulatory relationships between transcription factors and their target genes, which control the abundance of each gene product, is a key step towards gaining a complete view of the cell's interactome.

Gene expression profiles, or RNA profiles, yield quantitative data on the transcriptional state of a cell in a specific condition and time. Recent advances in high throughput methods for characterizing gene expression, such as DNA microarrays and RNA-seq, have made it easier, cheaper, and faster to obtain gene expression profiles. These experimental advances have spurred the development of computational methods that use gene expression profiles to infer a model of the global transcriptional regulatory network. In the last decade many algorithms have been developed to infer the structure of gene regulatory networks from gene expression profiles. A common approach taken by these algorithms to learning transcriptional regulatory network structure is to find, for each gene g_a, another gene or group of genes encoding transcription factors whose RNA levels best explains g_a's RNA level. As more sophisticated methods have been developed, the structural accuracy of inference algorithms has dramatically improved. Despite the advances in structural network recovery, transcriptional network inference has not met the need to generate predictive models that aid in experimental design and cellular engineering efforts.

Understanding, predicting, and manipulating complex cellular states requires structural and functional models of the cell. Most gene regulatory network inference algorithms provide a structural map of transcriptional interactions and an underlying mathematical model that can predict transcriptional results of genetic interventions. However, research using network models generally focuses on interpretation of the network structure, and neglects to use networks models to predict expression and physiological phenotypes. This dissertation begins to bridge the gap between inferring gene regulatory network models and utilizing the predictive power of the models to guide further research. First, we integrate the processes of inferring gene regulatory network models and *de novo* inference of transcription factor binding specificity to generate a causative, as opposed to a correlative network model, while exploring the ability of network models to aid in identifying transcription factor binding sites. We then explore the ability of network models to guide research by taking an active learning approach, which iterates between network inference and physiological phenotype prediction. Finally, we demonstrate the ability of causative regulatory network models to guide a cellular engineering effort by generating strains with regulatory interventions selected by simulated intervention expression predictions.

In chapter 2, we investigate strategies to infer more causative network models consisting of interactions that are supported by direct and functional evidence. Causative network models allow for better network-based expression and phenotype prediction than network models that rely solely on correlative analysis because they eliminate the effects of indirect regulatory interactions. To respond to the need for more causative network models for prediction applications, we develop a method to integrate the processes of network inference from gene expression profiles and *de novo* TF binding motif discovery. Through iteration, we are able to identify novel TF binding motifs and integrate them into the network to enrich the network for direct causative interactions. We evaluate this approach by applying it to the inference of the *Saccharomyces cerevisiae* and *Cryptococcus neoformans* networks and TF binding motifs. In *Cryptococcus neoformans*, we identify 18 TF binding motifs, 15 of which are novel.

In chapter 3, we map and model a regulatory network in *Cryptococcus neoformans* controlling the size of the fungal pathogen's capsule by joining the processes of computational modeling and experimentation. In this project, we explore the utility of network models as guidance for selecting costly experiments, by iterating between inferring the transcriptional regulatory network, using the network to make physiological phenotype predictions and selecting the next expression profiles to acquire. During this process we identify 18 novel regulators controlling the size of the polysaccharide capsule, which increased the number of known capsule size regulators to 27. With the finalized network map, we are able to identify many regulators controlling the enzymes responsible for the creation of capsule sugars.

Finally, in chapter 4, we fully evaluate the predictive power of causative network models by using a network model to guide transcriptome engineering. In this work, we build on our successes and knowledge gained in constructing and utilizing causative networks for prediction to develop the first validated, network-guided algorithm to select interventions for engineering transcriptional state. The NetSurgeon algorithm simulates regulator deletion and overexpression interventions on a causative network model, and selects the interventions that are predicted to force the expression state toward a goal state. We apply this algorithm to select intervention strains yielding promising results for improved biofuel production in a mixed glucose-xylose environment which have an enhanced fermentative transcriptional state in glucose-limited conditions. In addition, our generated intervention strains included strains with 120% higher xylose import rates and 31% increased ethanol production rates. The transcriptional and metabolic enhanced fermentative state of these strains demonstrates the power of using causative networks models to guide cellular engineering.

1.1 Contributions

1. **Development of methods to remove indirect regulatory network edges by inferring and integrating transcription factor binding specificity.** We present a method for *de novo* transcription factor binding motif inference by utilizing regulatory network structure and DNA-binding domain similarity. This contribution is presented in chapter 2.
2. **Inference and analysis of the Cryptococcus capsule regulatory network.** We take a network-driven iterative approach to infer the network that regulates the production and thickness of the polysaccharide capsule of *Cryptococcus neoformans*, the key virulence factor of this pathogenic fungus. This contribution is presented in chapter 3.
3. **Development of NetSurgeon, a novel algorithm for regulatory intervention selection.** We present an algorithm which aids in the rational manipulation of transcriptomes by utilizing a regulatory network to identify overexpression and deletion interventions that will force the expression state towards a desired goal state. This contribution is presented in chapter 4.
4. **Application of NetSurgeon to engineer *Saccharomyces cerevisiae* strains.** We apply NetSurgeon to *Saccharomyces cerevisiae* biofuel production, by selecting and generating intervention strains that promote a fermentative transcriptional state in the presence of xylose, an alternative carbon source. This contribution is presented in chapter 4.

Chapter 2:

Integration of Transcription Factor Binding Specificity

2.1 Abstract

Gene network models are vital to understanding how genes interact to regulate the transcriptional response to external stimuli. Within the last 15 years, much research has focused on inferring the structure of the gene regulatory networks. There have been two distinct approaches to this problem: regulatory network model inference from gene expression profiles and regulatory network model construction from identified transcription factor binding sites. In this work we have developed methods to integrate these approaches. We show that the gene regulatory network model structure can be improved by integrating known or inferred transcription factor binding specificities. We also demonstrate methods to improve DNA binding specificity inference, and we apply these methods to infer PWMs for *C. neoformans* transcription factors. Our application of these methods increased the number *C. neoformans* transcription factors with known binding specificities from 2 to 18.

2.2 Background

2.2.1 DNA Motif Discovery

One approach for constructing gene regulatory network models is to utilize the DNA binding specificities of transcription factors (TFs) to identify binding sites in the promoters of target genes. The binding specificity of a TF can be represented as a position weight matrix (PWM), which gives the probability of observing nucleotide j in position w of bound sequences. A score for an aligned substring can be calculated as the sum of log-likelihoods at each position of the alignment. By scoring substrings of DNA, PWMs can be used to identify sites with a greater potential for TF binding and transcriptional regulation.

There is a rich literature of computational methods for *de novo* DNA binding motif discovery. These methods identify similar DNA motifs within multiple input DNA sequences. DNA binding motif discovery methods are often used to infer TF binding specificity models by searching through TF bound promoters to identify enriched motifs. There are two broad approaches taken by these methods: mixture-modeling and discrimination (Lee et al. 2013). Many mixture model based methods use either expectation maximization (example: MEME, Bailey et al. 2006), or Gibbs Sampling (example: BioProspector, Liu et al. 2001) to identify TF binding sites and estimate PWMs. Discriminatory approaches, including CMF (Lee et al. 2013) and FIRE (Elemento et al. 2007), identify motifs that maximally separate a set of bound and unbound sequences. The identification of TF binding sites has been a major area of biological research and many current algorithms have been successfully demonstrated in lower model organisms (Das et al. 2007). However, computational motif discovery remains a challenging problem, beset with high false positive rates, and in need of additional prior knowledge to aid motif discovery (Simcha et al. 2012).

In recent years, the identification of transcription factor binding sites (TFBS) has improved dramatically through *in vivo* (ChIP) and *in vitro* (e.g. protein-binding microarray, PBM) methods for measuring protein DNA binding (Valouev et al. 2008 & Berger et al. 2009). There have been several large studies using these experimental techniques (Harbison et al. 2004;

Lee et al. 2002; Nègre et al. 2011; Badis et al. 2009; ENCODE Project Consortium 2012; Weirauch et al. 2014) and PWMs inferred in these studies are available in many public databases (Spivak & Stormo 2012; Shazman et al. 2013; Bryne et al. 2008; Matys et al. 2003).

The large-scale availability of TF binding specificity models has allowed researchers to identify links between protein homology and DNA binding specificity. Specifically, with the advent of large collections of inferred PWMs, several groups have conclusively shown that TFs with similar DNA binding domains (DBDs) also have similar TF binding specificities (Weirauch et al. 2014; Jolma et al. 2013). This observation was used by the Hughes group in their Cis-BP database of TF binding specificities (Weirauch et al. 2014). In the Cis-BP database they infer binding specificities for TFs without experimentally derived PWMs by assigning the PWM of the TF with the most similar DBD. Several other approaches have used machine learning techniques, including K-Nearest-Neighbors and Random Forests, to predict binding specificities for TFs from the PWMs of homologous TFs (Gupta et al. 2014; Christensen et al. 2012).

There has been a rapid rise in the number of TFs with known or inferred binding specificities. However, there is not a reliable binding specificity model for most non-model organism TFs (Weirauch et al. 2014). Even for human TFs, only 974 of the 1,734 putative TFs have a binding specificity model that has been experimentally derived or inferred (Weirauch et al. 2014). Therefore, to avoid costly experimentation, additional research is required to improve computational motif inference.

2.2.2 Gene Expression-Only Approaches for Inferring Gene Regulatory Networks

Another approach for constructing gene regulatory network models involves inferring the models from gene expression profiles. Advances in high throughput methods for characterizing gene expression have spurred the development of computational methods aimed at inferring the structure and kinetics of regulatory networks (Gardner & Faith 2005). These algorithms attempt to learn a network that best explains the measurements of all RNA transcripts in cells grown under various experimental conditions.

To evaluate the structural accuracy of an inferred network model, predicted interactions are ranked by confidence scores and compared to the true network structure, which is typically defined by ChIP-seq and ChIP-chip experiments that identify transcription factor binding sites on the promoter of target genes. A series of predicted networks is constructed by going down the list of predicted interactions, from most confident to least confident, and adding one interaction at a time to the previous network in the series. Each of these networks is evaluated using precision and recall statistics, which measure the accuracy and completeness of the inferred networks, respectively.

The most accurate and most recent expression only network inference algorithm is NetProphet (Haynes et al. 2013). This algorithm achieves superior accuracy by analyzing the expression profiles in two ways, then combining the results into a single model. First, NetProphet uses a LASSO solver to infer a regression-based regulatory network. In addition, NetProphet constructs a perturbation-based regulatory network by inferring interactions between experimentally perturbed regulators and the genes that are significantly affected by each perturbation. These two regulatory networks are weighted and combined. This approach has been demonstrated to be the best expression-only method for inferring the gene regulatory network structure of *S. cerevisiae*.

Although expression-only methods have improved dramatically, these algorithms are only able to infer a portion of the *S. cerevisiae* regulatory network with convincing accuracy.

Several challenges must be addressed in order to improve accuracy. First, expression-only inference methods rely upon co-expression of transcription factors and target genes to recover interactions. This reliance on co-expression allows these methods to recover many direct functional interactions; however, these methods can also incorrectly infer interactions between TFs and target genes that are indirectly co-expressed. A second challenge faced by these inference methods is modeling combinatorial gene regulation. Instead of direct transcriptional control by a single regulator, the transcription of many targets is controlled by the combined effects of several regulators. Combinatorial regulation makes it more difficult to infer true regulators by obscuring the relationship between regulator and target RNA concentration.

2.3 Related Work

Network inference approaches which utilize both expression-based inference and transcription factor binding site knowledge are strengthened by prioritizing interactions that the two approaches support. Support from each approach provides physical evidence of protein DNA binding and functional evidence of a regulatory effect. There are several notable methods which utilize this integrative network inference approach.

In recent work, Marbach et al. used an integrative approach to construct a gene regulatory network for *D. melanogaster* (Marbach et al. 2012). They constructed one network from interactions supported by ChIP experiments and TF binding motifs. They then constructed a second network from functional evidence by correlating expression patterns of transcription factors and target genes. Finally, they averaged the interaction scores from the functional and physical networks. Interestingly, although the physical and functional networks show little overlap in their most confident interactions, Marbach et al. found that averaging the two evidence sources generated a more accurate network.

A few methods have been proposed to integrate physical evidence directly into the regression approach for inferring gene regulatory networks. One notable example is the TILAR method, which uses knowledge of transcription factor binding sites to modify the L1 LASSO penalty (Hecker, Goertsches, Engelmann, Thiesen, & Guthke, 2009). This modification influences the order of regulator selection for each gene by biasing the selection toward regulators that are supported by prior knowledge.

Integrative network inference methods are often able infer regulatory networks with more accuracy than expression-only inference methods (De Smet et al. 2010). Moreover, the integration of binding data helps to solve some of the common problems encountered by expression-only inference methods, as binding data can be used to filter-out indirect regulatory interactions. Although integrative approaches are extremely promising, current approaches are stifled by their reliance upon prior knowledge of TF binding specificity. Due to the incompleteness of TF binding specificity knowledge, even for well studied organisms, integrative approaches must be able to infer novel TF binding specificities to reach their full potential.

2.4 Approach

In this work we investigated strategies to generate more accurate gene regulatory network models by integrating the processes of expression-only network inference and *de novo* DNA binding motif discovery. Our approach to integrating these processes started by using NetProphet (Haynes et al. 2013) to infer a gene regulatory network model from gene expression data (Fig 2.1 Top). Next, we inferred the DNA binding specificity of each TF by searching for enriched motifs in the promoters of high confidence targets of the each TF (Fig 2.1 Middle). Finally, we integrated inferred DNA binding specificities into the gene regulatory network model by combining the functional network interaction scores assigned by NetProphet with the physical network interaction scores assigned by scanning the PWMs over all promoter sequences (Fig 2.1 Bottom).

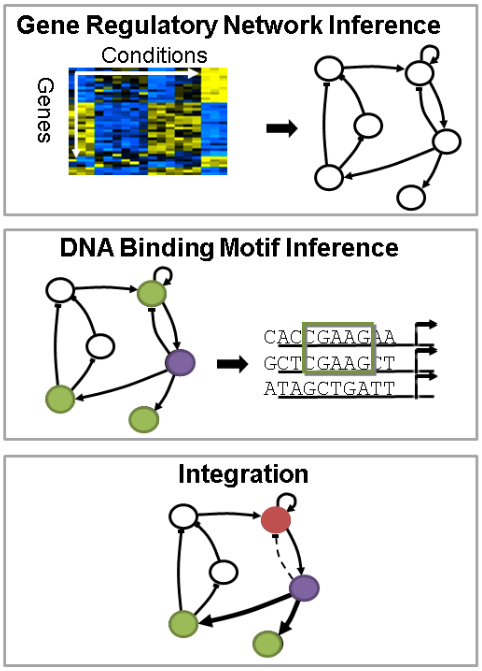

Gene Regulatory Network Inference
Conditions
Genes
DNA Binding Motif Inference
CACCGAAGAA
GCTCGAAGCT
ATAGCTGATT
Integration

Figure 2.1. Integrated regulatory network and DNA binding specificity inference. Top: Inference of a gene regulatory network model from gene expression profiles. Middle: Inference of DNA binding specificity of each TF from the promoter sequences of the predicted targets of the TF. Bottom: Modification of interaction confidence scores in the network model using inferred DNA binding specificity.

2.5 Results

2.5.1 Integration of known TF DNA binding specificities improves structural accuracy

To assess the improvements to the gene regulatory network model by integrating TF binding specificities, we constructed an integrated functional and physical network of *S. cerevisiae*. The integrated network model was constructed by combining the original published

NetProphet network model produced by Haynes et al. (Haynes et al. 2012) with a regulatory network constructed by scanning all known *S. cerevisiae* PWMs (Spivak & Stormo, 2012) over all promoter sequences. Finally, the integrated NetProphet+PWM network model was constructed by computing the geometric mean of scores in the functional and physical networks.

We evaluated the structural accuracy of the NetProphet+PWM network and compared it to the accuracy of the separate NetProphet and PWM networks. The gold standard was a network of over 29,000 chromatin immunoprecipitation (ChIP) implicated interactions (Abdulrehman, et al. 2010; Balaji et al. 2006; Harbison et al. 2004; Lee et al. 2002). We plotted a precision-recall curve for each network (Fig 2.2).

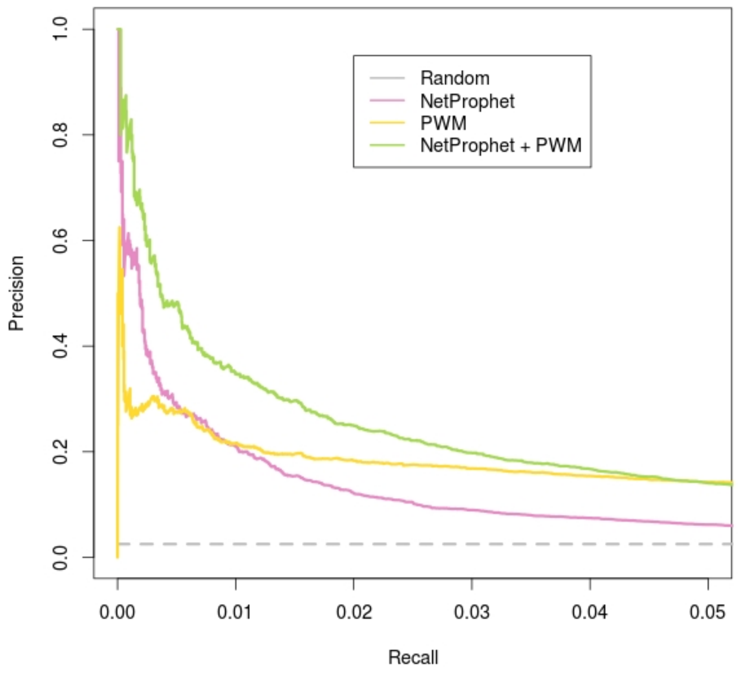


Figure 2.2. Integration of DNA binding specificity improves network structure. Precision-Recall plot showing the recovery of ChIP-supported interactions by the NetProphet inferred network (purple), binding potential network constructed using ScerTF binding models (Yellow), and a network constructed by combining NetProphet and the binding potential network (green). Random ChIP recovery is shown by the gray dotted line.

We examined the ChIP support of the most confident predicted interactions of each network by focusing on the initial 5% of the recall space. Throughout this recall space the NetProphet+PWM (Fig 2.2, Green) network integration curve dominated the NetProphet (Fig

2.2, Purple) and PWM binding specificity based (Fig 2.2, Yellow) network curves. We computed the area under the precision-recall curve (AUPRC) and found that the AUPRC for the NetProphet+PWM curve is 1.8 fold and 1.4 fold greater than the NetProphet and PWM network AUPRCs, respectively. Importantly, we have observed similar integrated network improvements when using a functional network inferred from a different large-scale gene deletion expression dataset (Kemmeren et al. 2014), and when using a physical network constructed from protein-binding microarray (PBM) obtained PWMs (data not shown), indicating that integrating TF DNA binding specificities should generally improve network models.

2.5.2 Distinguishing good from bad binding specificity models

Inspired by the observed improvements gained by combining functional and physical networks, we decided to integrate the processes of gene regulatory network inference and *de novo* DNA binding motif discovery. Performing *de novo* DNA binding motif discovery will allow us infer novel TF binding specificities and expand integrative network improvements beyond the confines of experimentally derived PWMs. Since computational motif discovery has high false positive rates, novel methods were required to select only accurate PWM models of TF binding specificity. Without the ability to distinguish accurate from inaccurate PWMs, integration of DNA binding specificities into an integrative network model would likely decrease model quality.

To identify accurate PWMs for each TF, we compared the target rankings produced by the PWM models with the target rankings produced by NetProphet. Similar to a published approach which links motifs to TFs (Verfaillie et al. 2014), we hypothesized that target gene rankings from an accurate PWM for a TF will agree with the NetProphet target rankings for the TF significantly better than target rankings produced by random PWMs. We investigated many potential metrics to measure the relationship between the NetProphet target rankings and PWM based target rankings. We computed the similarity between the rankings by correlation (Pearson, Spearman, and Kendall), and mutual information. We separated the NetProphet targets into two binary sets of bound and unbound targets, and compared the PWM-based ranking of targets to these categorical sets using AUPRC and Fisher's exact test. Also, we fit a linear regression model to explain target gene expression form PWM scored binding potential. We used the magnitude of the regression coefficient as a final measure of the PWMs accuracy.

To evaluate these metrics, we used each one to score the relationship between the NetProphet target set rankings of each TF and the target rankings produced by scanning the ScerTF PWM of each TF. For each PWM, we standardized the PWM's relationship score (converted to Z-scores) with each NetProphet target set. Then, we assessed the degree to which higher scores were assigned when the NetProphet and PWM target set rankings were generated for the same TF, rather than different TFs. We plotted the mean Z-scores of target set rankings generated for the same TF (Fig 2.3). We found that using Pearson correlation, Fisher's exact test, and the magnitude of the regression coefficient explaining gene expression from binding motifs were significantly better than other metrics at matching each NetProphet target set with the correct PWM. We decided to use the Pearson correlation coefficient for subsequent analysis due to its ease of use, and computational speed.

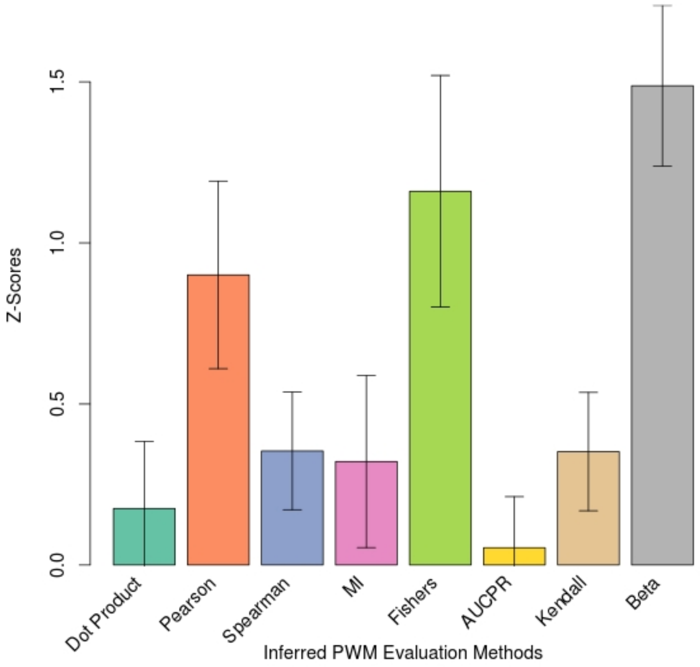


Figure 2.3. Investigation of methods to identify good binding specificity models. Barplot showing the ability of different metrics to identify the binding specificity model of each TF. The mean of the z-scores assigned to the true binding specificity models is plotted for each metric.

2.5.3 Integration of Network and DNA Binding Specificity Inference

With metrics capable of selecting accurate PWMs by comparing PWM based ranking of target genes to NetProphet ranking of target genes, we refocused on integrating *de novo* inferred PWMs into a network model. We used the NetProphet network as our functional network. We used COSMO (Bembom et al. 2007), a MEME-type motif finder, for PWM inference. In total we discovered five potentially different PWMs for each TF by inferring a single PWM for each of 5 expression pattern based clusters of NetProphet predicted targets. We assigned confidence scores to each PWM by utilizing our method to distinguish good from bad binding specificity models. Specifically, the confidence score for each PWM was the Z-score of Pearson correlation coefficient of the target set rankings produced by the PWM and matching NetProphet target set. For each TF we scanned all 5 PWMs over target promoter sequences and combined scores by weighted averaging, with each PWM's confidence score serving as its weight. We then constructed an integrated network by computing the geometric mean of scores in the functional and physical networks. To evaluate the integrated network, we plotted precision and recall of the NetProphet network (Fig 2.4.A, Purple), inferred PWM based network (Fig 2.4.A, Yellow), and the NetProphet + inferred PWM network (Fig 2.4.A Green).

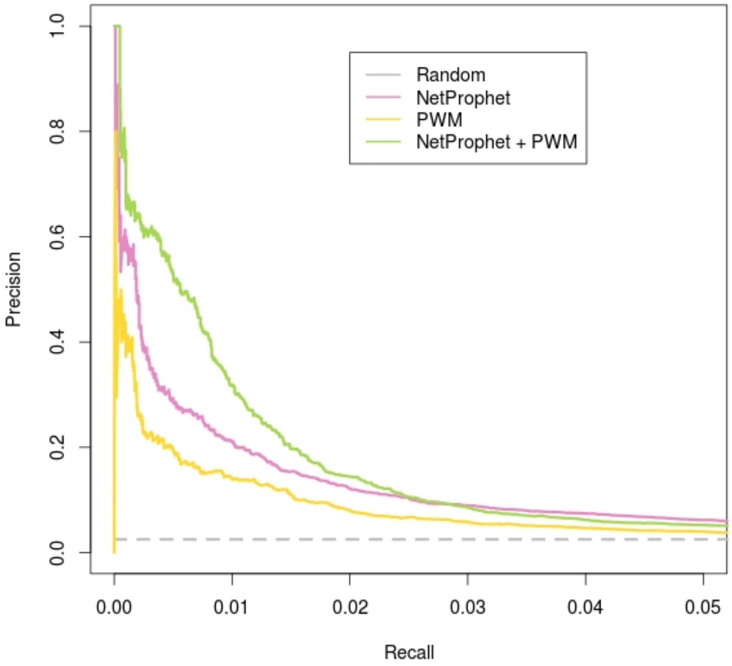


Figure 2.4.A. Precision recall curves showing the recovery of ChIP-supported interactions by the NetProphet network (purple), binding potential network constructed from inferred binding models (Yellow), and integrated network (green). Random ChIP recovery is shown by the gray dotted line.

The integrative network had an AUPRC that was 1.3 fold greater than the NetProphet alone. Interestingly, this improvement occurs even though the physical network constructed using inferred PWMs (AUPRC: 0.1) is much less accurate than the physical network constructed using known PWMs (AUPRC: 0.19).

Next, we evaluated the accuracy of the inferred PWMs. We focused on a set of 187 inferred PWMs of TFs that have a known PWM (Spivak & Stormo 2012). Our correlative method for assigning confidence scores to inferred PWMs classified 38 as strong (Bonferroni corrected P-value <= 0.01), 25 as medium (Bonferroni corrected P-value between 0.01 and 0.2), and 124 as weak (Bonferroni corrected P-value > 0.2). We evaluated these confidence classifications by comparing the classifications to the accuracy of the inferred PWMs by aligning the inferred PWMs with the true PWMs (Fig 2.4.B). We found that inferred PWMs in the strong confidence bin aligned better with their matching known PWMs then the PWMs in either the medium (Mann-Whitney *U* test p-value < 0.05) or weak bins (Mann-Whitney *U* test p-value < 10^{-4}).

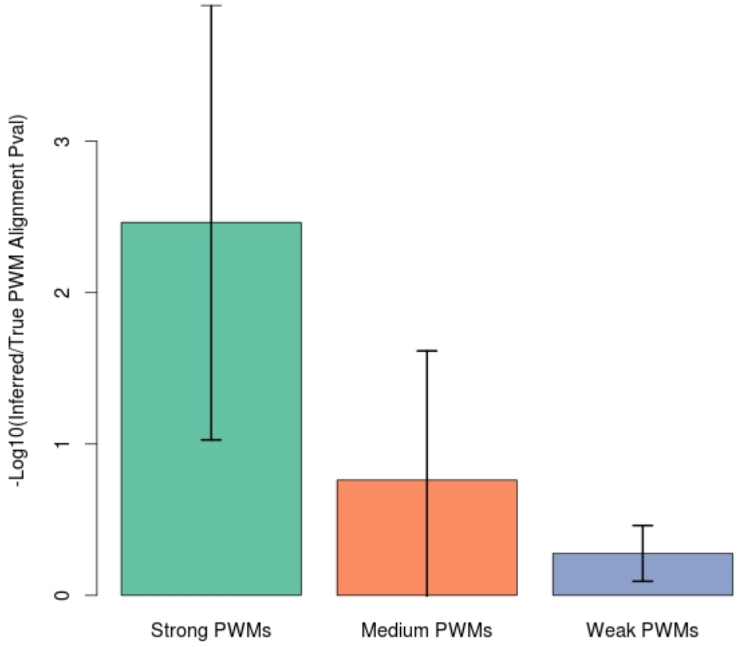


Figure 2.4.B. Accuracy of inferred PWMs is predicted by their confidence scores. Barplot showing the mean significance of inferred PWMs aligned with their matched known PWM. The inferred PWMs were binned by the confidence score of the inferred PWM. Turquoise: High confidence, Red: Medium confidence, Blue, low confidence

2.5.4 Improving DNA Binding Specificity Inference

Although we were able to improve the structural accuracy of the inferred network by integrating inferred PWMs, we were only able to infer high confidence PWMs for 38 of the 187 *S. cerevisiae* TFs. In order to approach the network model improvements demonstrated for known PWMs, we must do better. Therefore, we investigated strategies to improve PWM inference.

Many TFs are enriched to either up-regulate or down-regulate their direct targets, and these TFs are classified as activators or repressors respectively (Kemmeren et al. 2014). Activator/Repressor classification knowledge can benefit both motif discovery and network inference by modifying the confidence in each TF-target interaction based on a comparison of the expected and observed effect of the TF on the target. However, these classifications do not exist for many less well-studied organisms. Therefore, instead of using the TF classifications directly, we used the existence of these TF classifications as motivation for performing DNA motif discovery separately on the activated and repressed targets of each TF. As an initial validation of this approach, we binned NetProphet predicted targets of the *C. neoformans* TF Usv101 (Fig 2.5.A, Top) by their confidence score and regulatory sign (activated/repressed by Usv101), and identified the bins that are enriched with Usv101 motif hits (Fig 2.5.A, Bottom). We found that the most confidently repressed targets of Usv101 (top left bins) were also the bins most enriched for Usv101 motif occurrences (bottom left bins shaded red). Further, the motif recovered from the repressed targets of Usv101 was confirmed by ChIP experiments (Figure 3.6.D).

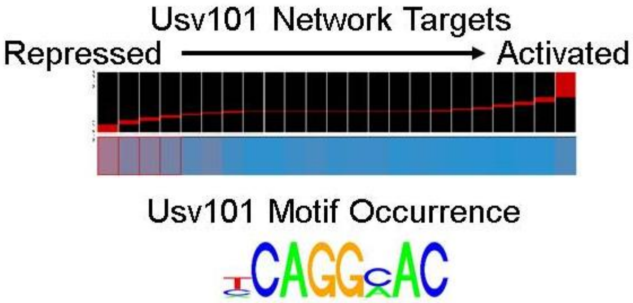


Figure 2.5.A. Interaction signs aid in DNA binding motif discovery. Top row: The targets of Usv101, a *C. neoformans* TF, binned by NetProphet confidence and predicted regulatory sign (activation or repression). Bottom row: Heatmap of occurrences of the Usv101 DNA binding motif in the promoter region of the binned genes. Red shading indicates more occurrences of the motif, blue indicates fewer.

We also investigated a second approach to improve binding specificity inference which was motivated by the observation that TFs with similar DBDs also tend to have similar PWMs (Weirauch et al. 2014; Jolma et al. 2013). To evaluate this observation in *S. cerevisiae* we computed the DBD homology and PWM similarity for all pairs of TFs. DBD homology was computed using BLASTP (Altschul et al. 2007) and PWM similarity was computed using TOMTOM (Gupta et al. 2007). For each DBD homology significance threshold, from an E-value of 1 to an E-value of 10^{-30}, we plotted the fraction of DBD homologous TFs that also have similar PWMs (Fig 2.5.B). Confirming previous publications (Weirauch et al. 2014; Jolma et al. 2013), we observed a logistic trend in which more homologous DBDs were more likely to have similar PWMs.

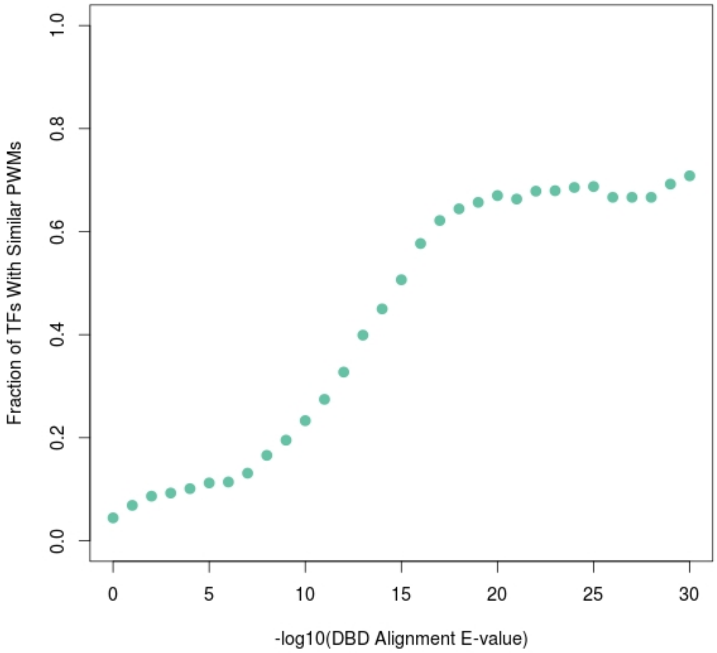


Figure 2.5.B. Relationship between DBD homology and PWM similarity for *S. cerevisiae* TFs. DBDs of all pairs of *S. cerevisiae* TFs were aligned and for each bin of DBD homology, computed as –log10 E-value), the fraction of TF pairs with similar PWMs was plotted (PWM alignment E-value < 1).

We utilized this relationship between DBD homology and PWM similarity to improve PWM inference by computing a weighted average of NetProphet target confidence scores for

TFs with similar DBDs. By taking the weighted average of NetProphet target confidence scores, the top NetProphet targets for a TF will be targets that are shared by multiple TFs with similar PWMs, enriching the top targets for genes that have the right binding motif family in their promoters. Weights were assigned to each TF's NetProphet target set fitting a logistic curve to the relationship between DBD homology and PWM similarity (Figure 2.5.B). Therefore, to generate DBD modified NetProphet target set confidence scores for a TF, the TF's target set confidence scores were given a weight of 1 and combined with other target set confidence scores for TFs that have similar DBDs (E-value < 10^{-10}), each weighted by their DBD similarity to the original TF.

Next we performed DNA motif discovery using the DBD modified NetProphet target confidence scores. The motif finding tool FIRE was used to identify sequence-specific binding motifs (Elemento et al. 2007). FIRE identifies a set of 7-mers, that can be converted to a PWM, whose presence or absence in a given regulatory region share strong mutual information with functional information about the corresponding genes (e.g. gene expression). In this case, the functional information used were the DBD modified NetProphet target confidence scores. We used FIRE to discover DNA motifs in *S. cerevisiae* and *C. neoformans*. For each of these organisms we counted the number of motifs FIRE discovers through mutual information with original NetProphet confidence scores, signed confidence scores, DBD modified confidence scores, and signed DBD modified confidence scores (Fig 2.5.C). We found that FIRE discovered an average of 2.7 fold and 1.8 fold more motifs when using the signed NetProphet scores and DBD NetProphet scores respectively compared to using the original NetProphet scores. In addition, using both sign and DBD homology information together allowed FIRE to recover the most motifs, an average of 3.7 fold more motifs than using the original NetProphet scores.

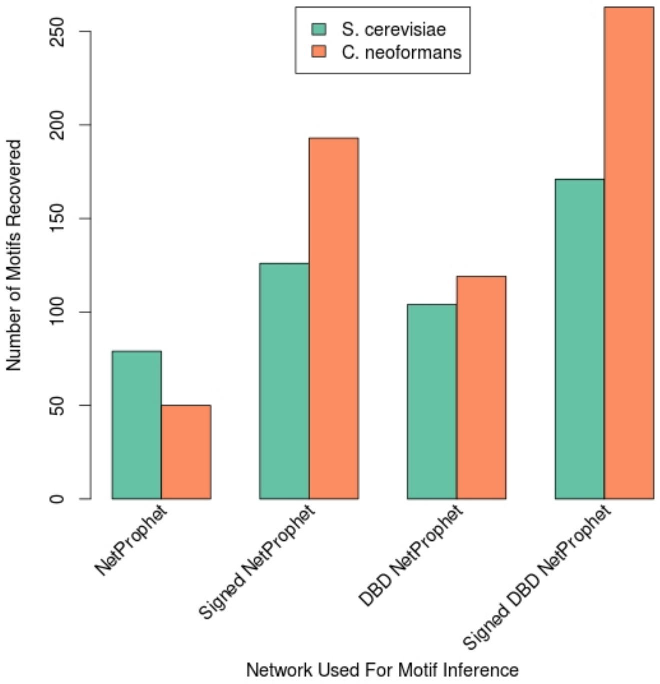
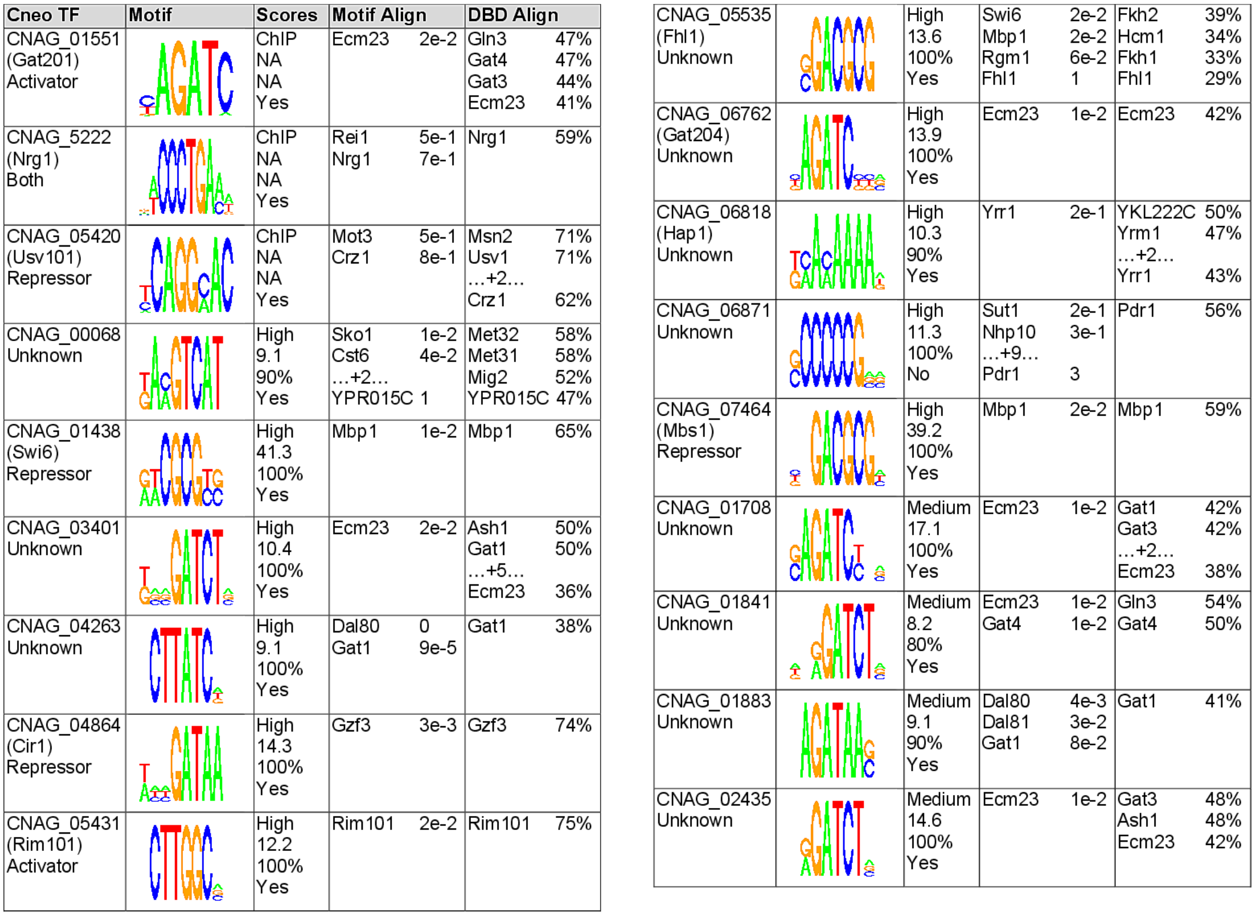


Figure 2.5.C. DBD homology and interaction sign aid DNA binding specificity inference. Barplot showing the number of DNA binding motifs inferred by FIRE when using each of four different networks. The number of motifs is displayed for *S. cerevisiae* (turquoise) and *C. neoformans* (red).

2.5.5 Inference of DNA Binding Specificities in *C. neoformans*

Although we confirmed that using regulatory signs and DBD homology allowed us to recover more PWMs, the accuracy of the inferred PWMs must still be assessed. We evaluated the correctness of each inferred *C. neoformans* motif using several metrics including FIRE reported motif confidence scores, motif occurrence conservation in *C. neoformans* JEC21, and PWM similarity for DBD homologous *S. cerevisiae* TFs. Using these evaluation metrics, we separated inferred motifs into three confidence bins (Fig 2.6). The first bin consisted of the three ChIP motifs inferred using BioProspector on the ChIP defined target sets of Gat201, Nrg1, and Usv101. The second and third bins consist of high and medium confidence motifs inferred using FIRE with the signed DBD modified NetProphet predicted target sets. The high and medium confidence motif sets both passed all initial evaluation criteria; however, the high confidence motif set had better agreement with the NetProphet predicted target scores than the medium motif confidence set. In total, the PWMs increased the number of *C. neoformans* TFs with known binding specificities from 2 to 18. In addition, for each TF with a motif, the regulatory role of the TF (activator / repressor / both) was assessed by comparing the enrichment of activated and repressed targets with likely bound targets.



Cneo TF	Motif	Scores	Motif Align	DBD Align
CNAG_01551 (Gat201) Activator		ChIP NA NA Yes	Ecm23 2e-2	Gln3 47% Gat4 47% Gat3 44% Ecm23 41%
CNAG_5222 (Nrg1) Both		ChIP NA NA Yes	Rei1 5e-1 Nrg1 7e-1	Nrg1 59%
CNAG_05420 (Usv101) Repressor		ChIP NA NA Yes	Mot3 5e-1 Crz1 8e-1	Msn2 71% Usv1 71% …+2… Crz1 62%
CNAG_00068 Unknown		High 9.1 90% Yes	Sko1 1e-2 Cst6 4e-2 …+2… YPR015C 1	Met32 58% Met31 58% Mig2 52% YPR015C 47%
CNAG_01438 (Swi6) Repressor		High 41.3 100% Yes	Mbp1 1e-2	Mbp1 65%
CNAG_03401 Unknown		High 10.4 100% Yes	Ecm23 2e-2	Ash1 50% Gat1 50% …+5… Ecm23 36%
CNAG_04263 Unknown		High 9.1 100% Yes	Dal80 0 Gat1 9e-5	Gat1 38%
CNAG_04864 (Cir1) Repressor		High 14.3 100% Yes	Gzf3 3e-3	Gzf3 74%
CNAG_05431 (Rim101) Activator		High 12.2 100% Yes	Rim101 2e-2	Rim101 75%
CNAG_05535 (Fhl1) Unknown		High 13.6 100% Yes	Swi6 2e-2 Mbp1 2e-2 Rgm1 6e-2 Fhl1 1	Fkh2 39% Hcm1 34% Fkh1 33% Fhl1 29%
CNAG_06762 (Gat204) Unknown		High 13.9 100% Yes	Ecm23 1e-2	Ecm23 42%
CNAG_06818 (Hap1) Unknown		High 10.3 90% Yes	Yrr1 2e-1	YKL222C 50% Yrm1 47% …+2… Yrr1 43%
CNAG_06871 Unknown		High 11.3 100% No	Sut1 2e-1 Nhp10 3e-1 …+9… Pdr1 3	Pdr1 56%
CNAG_07464 (Mbs1) Repressor		High 39.2 100% Yes	Mbp1 2e-2	Mbp1 59%
CNAG_01708 Unknown		Medium 17.1 100% Yes	Ecm23 1e-2	Gat1 42% Gat3 42% …+2… Ecm23 38%
CNAG_01841 Unknown		Medium 8.2 80% Yes	Ecm23 1e-2 Gat4 1e-2	Gln3 54% Gat4 50%
CNAG_01883 Unknown		Medium 9.1 90% Yes	Dal80 4e-3 Dal81 3e-2 Gat1 8e-2	Gat1 41%
CNAG_02435 Unknown		Medium 14.6 100% Yes	Ecm23 1e-2	Gat3 48% Ash1 48% Ecm23 42%

Figure 2.6. Logos representing the DNA binding specificities of 18 Cryptococcus TFs with supporting evidence. The three logos at the top of the left column were inferred from ChIP data using BioProspector and the remaining 15 were inferred from NetProphet scores using FIRE. Cneo TF: Cryptococcus TF identifiers and inferred function as activator, repressor, both, or unknown. Motif: logos. Scores: Our overall confidence in the accuracy of the motif followed by the mutual information Z-score from FIRE, the robustness score from FIRE, and whether the motif was highly conserved in the promoters of genes in Cryptococcus serotype D strain JEC21 that are orthologs of the targets of the indicated KN99 TF. Motif align: List of *S. cerevisiae* TFs whose motifs in the ScerTF database show the strongest resemblance to the inferred motif for the indicated Cryptococcus TF. *S. cerevisiae* TFs are listed in order of similarity, starting with the TF that is most similar and ending with the supporting ortholog chosen on the basis of similar DNA binding domain (DBD) and similar sequence specificity. For each *S. cerevisiae* TF, E-value output by TOMTOM for the alignment of its motif to that of the indicated *C. neoformans* TF is shown. The best possible

support for the assignment of a motif to a *C. neoformans* TF is an *S. cerevisiae* TF with a highly similar DBD that also has highly similar sequence specificity (e.g. Nrg1, Usv101, Swi6, Cir1, Rim101, and Mbs1). DBD Align: *S. cerevisiae* TFs listed in order of the similarity of their DBDs to the DBD of the indicated *C. neoformans* TF, ending with the supporting ortholog chosen on the basis of similar DBD and similar sequence specificity. For each *S. cerevisiae* TF, the percent protein identity of its DBD to that of the *C. neoformans* DBD is listed.

2.6 Discussion

In this chapter we've demonstrated that functional network models can be improved by integrating physical TF binding data. We then showed that novel PWMs can be inferred *de novo* from NetProphet rankings of targets, and that these novel PWMs can be used to improve the network model structure. Finally, we investigated methods to improve *de novo* inference of DNA binding specificities. We showed that NetProphet interaction confidence scores modified by adding regulatory signs and utilizing DBD homology can improve the recovery of DNA binding motifs. A clear next step for this work is to re-evaluate the structural improvement to the network model by integrating inferred physical binding information (Appendix Figure 1).

In this chapter we only attempted a single round of gene network inference, *de novo* DNA binding specificity inference, and model integration. However, we believe that these inference steps could be continuously reapplied until no new PWMs are recovered. More sophisticated methods for integrating PWMs into the network model would likely be required to benefit from iterating these processes. Specifically, integrative approaches that modify the entire network, not just the targets of the TFs with inferred PWMs are required so that the target sets of TFs without a PWM are modified. A possible approach, similar to the TILAR method, would use binding information to influence the regression based inferred network model through applying adaptive weights to the L1 LASSO penalty (Zou 2006).

2.7 Methods

2.7.1 Scanning PWMs and constructing a physical binding network

Each TF binding based network was constructed by using FIMO (Grant et al. 2011) to scan PWMs over the 600 bases upstream of each gene's transcription start site. Significant PWM hits with P-value < 0.005 were used to score binding potential. For each TF, two models of regulation were considered, a strong-site model in which target genes are scored based on the maximum negative log P-value of significant PWM hits, and a weak-site model in which target genes are scored based on the sum of the negative log P-values of significant PWM hits. The strong and weak-site models were normalized on a per-TF basis so that the maximum target binding score per-TF was 1. The final physical binding network was constructed by computing the geometric mean of the strong and weak-site models of binding.

2.7.2 Combining functional and physical regulatory networks

To construct causative network models, the underlying functional and physical networks were normalized so that the maximum interaction score for each model was 1 and combined by computing the geometric mean of their scores.

2.7.3 Aligning PWMs

To assess the accuracy of inferred PWMs, we aligned the inferred PWMs against a collection of known *S. cerevisiae* PWMs (Spivak & Stormo 2012) using TOMTOM (Gupta et al. 2007). Inferred PWMs were considered accurate if they aligned well to the known correct PWM or to another PWM for a TF with a similar DBD.

2.7.4 Computing DBD homology

The DBD homology between pairs of TFs was computed by extracting the DBD of each TF, then aligning the DBDs. The DBD portion of the protein sequences of every *S. cerevisiae* (Engel, et al., 2014) and *C. neoformans* (*Cryptococcus neoformans var. grubii* H99 Sequencing Project) TF was identified using the NCBI Conserved Domain Database search tool (Marchler-Bauer et al. 2013). The DBD sequences of each TF were extracted and aligned to each other using BLASTP (Altschul et al. 1997) version 2.2.29.

2.7.5 Evaluation of *C. neoformans* inferred PWMs

Several metrics were used to evaluate the correctness of each inferred motif. First, FIRE performs extensive randomization tests to asses the relationship between the motif and functional data. A cutoff of 8 was set on the FIRE reported Z-score of the MI value between the motif and NetProphet target set, compared with random target sets. A cutoff of 7/10 was set on the FIRE robustness score, which measures the robustness of the MI significance by 10 re-calculations of MI after randomly removing one-third of the genes. Inferred motifs with scores below these cutoffs were rejected. These cutoffs help ensure that inferred motifs were likely true functional DNA recognition motifs for a TF in *C. neoformans*. Although conservation was not used as a rejection criterion, the conservation of the motifs was also assessed using FIRE by comparing the locations of motif occurrences in the promoters of *C. neoformans* H99 and *C. neoformans* JEC21. Network-level conservation of each inferred motif was assessed and motifs with conservation index >= 0.95 were defined as conserved. This index is the fraction of all possible 7-mers that are less conserved than the 7-mers permitted by the inferred motif. To also ensure that the true functional motif is linked to the appropriate TF, the inferred motif was required to align well (TOMTOM reported Q-value <= 0.5) to a known *S. cerevisiae* motif whose TF was required to share significant DBD homology (BLASTp E-value <= 1e-5) with the *C. neoformans* TF of the aligning motif.

2.7.6 Inferring regulatory sign of *C. neoformans* TFs

To assign a regulatory role for each TF with an inferred PWM, the NetProphet-predicted activated and repressed targets of the TF were compared to the targets predicted to be bound by the TF (either by ChIP, or high scoring PWM hits). Only interactions in the top 20,000 NetProphet interactions were considered to identify activated and repressed targets of a TF. The significance of binding of activated and repressed targets was computed in two ways. First, for TFs with an inferred motif, a two-sample Mann-Whitney *U* test was used to assess the significance of higher binding potential scores of activated (repressed) targets than the null background of all other possible targets. Second, for TFs with ChIP evidence, a Hypergeometric enrichment test was used to assign a P-value to the overlap between ChIP supported targets and activated (repressed) targets. The P-values were adjusted using the Bonferroni correction. If a TF had significant (p < 1e-3) enrichment for binding its activated (repressed) targets, then the TF was defined as an activator (repressor). If a TF had significant enrichment for binding both its activated and repressed targets, then the TF was defined as both an activator and a repressor. Any TF without a significant (<1e-3) adjusted p-value was not assigned a regulatory sign.

Chapter 3:

Model-driven mapping of transcriptional networks

3.1 Abstract

Key steps in understanding a biological process include identifying genes that are involved and determining how they are regulated. We developed a novel method for identifying transcription factors (TFs) involved in a specific process and used it to map regulation of the key virulence factor of a deadly fungus, its capsule. The map, built from expression profiles of 41 TF mutants, includes 20 TFs not previously known to regulate virulence attributes. It also reveals a hierarchy comprising executive, mid-level, and “foreman” TFs. When grouped by temporal expression pattern, these TFs explain much of the transcriptional dynamics of capsule induction. Phenotypic analysis of TF deletion mutants revealed complex relationships among virulence factors and virulence in mice. These resources and analyses provide the first integrated, systems level view of capsule regulation and biosynthesis. Our methods dramatically improve the efficiency with which transcriptional networks can be analyzed, making genomic approaches accessible to labs focused on specific physiological processes.

3.2 Background

In this work we present an efficient means of comprehensively mapping the network of transcription factors (TFs) that regulate a particular physiological process. Our approach cycles through deletion of TFs, expression profiling of TF mutants, model construction, and model-directed selection of TFs for the next round of deletion. This predictive genetics approach identifies TFs that affect the process of interest, providing a valuable complement to undirected mutagenesis and screening. Simultaneously, it builds a network model that explains how the TFs affect the process, yielding novel insights into the biological system under study.

Mapping the network that regulates a specific process requires knowing which TFs affect that process. One way to identify such TFs is to screen comprehensive mutant libraries, but generating such libraries is not always feasible. Furthermore, genome-scale screening assays must be fast and scalable; such assays may not exist for the process of interest or may be less sensitive than other, more laborious assays. An alternative approach is to map the targets of all TFs encoded in a genome by using methods such as chromatin- immunoprecipitation (ChIP) or large-scale TF deletion and expression analysis. However, undirected, genome-wide approaches are costly and inefficient for probing a specific biological process in detail. We report a model-guided approach that addresses all of these problems by focusing experimental effort on the TFs most likely to be involved in the process of interest. Furthermore, our approach generates a network that provides mechanistic explanations for the phenotypes of TF deletion mutants.

Our approach alternates network building by using an algorithm we call NetProphet with identifying relevant TFs by using an algorithm we call PhenoProphet. NetProphet is a validated method for mapping *direct, functional regulation* that significantly outperforms other network mapping methods (Haynes et al. 2013). It requires only gene expression profiles of strains in which TF expression has been perturbed (by gene deletion, mutation, overexpression, or RNAi) and wild-type controls, data that can be gathered in a reliable, scalable way by most molecular biology labs. It therefore offers significant advantages over alternatives such as ChIP-seq, which requires reoptimization for every TF studied (Landt et al. 2012) and follow-up experiments to determine which binding events lead to functional regulation. NetProphet works by combining differential expression (DE) analysis with co-expression analysis. In the DE analysis, genes that are strongly differentially expressed between a TF-deletion strain and a wild- type (WT) strain are considered potential targets of the TF. In the co-expression analysis, genes whose expression is strongly correlated with that of a TF (either positively or negatively) across the expression profiles are considered potential targets of the TF, enabling NetProphet to identify targets for TFs that have not been directly perturbed (see Haynes et al. 2013). PhenoProphet, described here for the first time, assigns each TF a score representing its confidence that deletion of that TF will yield some phenotypic change of interest. The score of each TF is based on the degree to which its NetProphet-predicted targets are enriched for genes associated with the phenotype of interest.

We demonstrate the power of combining NetProphet and PhenoProphet by mapping the network that regulates the major virulence factor of a pathogenic yeast, *Cryptococcus neoformans*. *C. neoformans* is a basidiomycetous yeast, with a 19 Mb genome encoding ~7000 genes (Janbon et al. 2014), that diverged from ascomycetes like *S. cerevisiae* roughly one billion years ago (Hedges et al. 2004). It is also an opportunistic pathogen that is responsible for over 600,000 deaths per year worldwide (Park et al. 2009). Multiple factors influence cryptococcal virulence (Srikanta et al. 2014), including the production of protective structures like melanin and its major virulence factor, a polysaccharide capsule. Capsule polysaccharides are both displayed on the cell surface and shed from the cell. The capsule grows large upon entry into a mammalian host, a process that can be recapitulated by a variety of host-like conditions *in vitro* (Zaragoza and Casadevall 2004).

3.3 Related Work

The initial TFs implicated in regulating the virulence factors of *C. neoformans* are almost uniformly major regulators of virulence which are directly downstream of the signaling transduction pathways related to virulence factor induction conditions. The work which identified these initial TFs relied upon protein conservation, genetic disruption, and microarray based screening. For example, the iron responsive TF Cir1, which is required for virulence in a mouse, was identified by searching the *C. neoformans* genome for an ortholog of the *S. cerevisiae* iron regulators (Jung et al. 2006). In addition, Nrg1, a TF required for capsule production and cell wall integrity, was identified through microarray analysis of cAMP activated genes (Cramer et al. 2006). Although these studies identified many of the major virulence regulators, new methods were required to expedite understanding by moving beyond the single gene study approach.

One of the largest studies of this organism was a large-scale genetic screen which identified dozens of previously uncharacterized genes that are linked with several assayed virulence factors (Liu et al. 2008). For this genetic screen, a total of 1,201 gene knockout mutants were generated and each mutant was screened for in vivo proliferation in murine lung tissue, in vitro capsule formation, in vitro melanization, and growth at body temperature. Several TFs important for regulating cryptococcal virulence factors were identified in this screen including GAT201, a key regulator of virulence, which is required for proper induction of large capsules and also regulates melanization.

Recent work on *C. neoformans* has started to incorporate Bioinformatics and Systems Biology techniques to study the capsule virulence factor. Haynes et al. identified a set of 880 capsule signature genes whose expression significantly correlated with a capsule size in various conditions (Haynes et al. 2011). One of those genes was Ada2, a regulator of stress response required for full induction of capsule. As a beginning step toward identifying the broader network controlling capsule regulation, expression profiles of *ada2*, *cir1* and *nrg1* were used to place Ada2 in a network model of capsule regulation.

Although several TFs involved in capsule regulation have been identified, however large gaps remain in our understanding of capsule regulation (Rodrigues et al. 2011; O'Meara and Alspaugh 2012; Kwon-Chung et al. 2014). Most of the downstream capsule biosynthetic machinery remains to be discovered and current knowledge of capsule regulation is incomplete and fragmented. This offers an ideal opportunity to apply model-guided network mapping. In this work, we present the first integrated, systems level view of capsule regulation and biosynthesis, which in turn produces unexpected insights into cryptococcal virulence.

3.4 Approach

We have developed a novel, predictive approach to identifying TFs related to a physiological process of interest and mapping their regulatory targets. This approach consists of a cycle (Fig. 3.1) in which TF deletion strains are subjected to phenotyping and expression profiling and network models are constructed from the expression profiles by using NetProphet. PhenoProphet is then used to predict additional TF genes which, when deleted, will influence the process of interest. These genes are then deleted and the phenotypes and expression profiles of the resulting mutants are fed back into the cycle. The genome-wide network models resulting from this process can be analyzed in multiple ways, including modeling TF binding specificity and predicting TF function based on target gene sets.

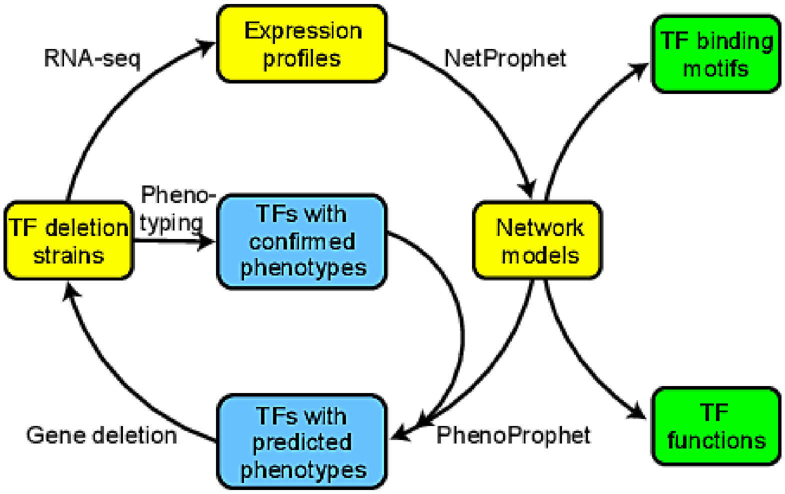


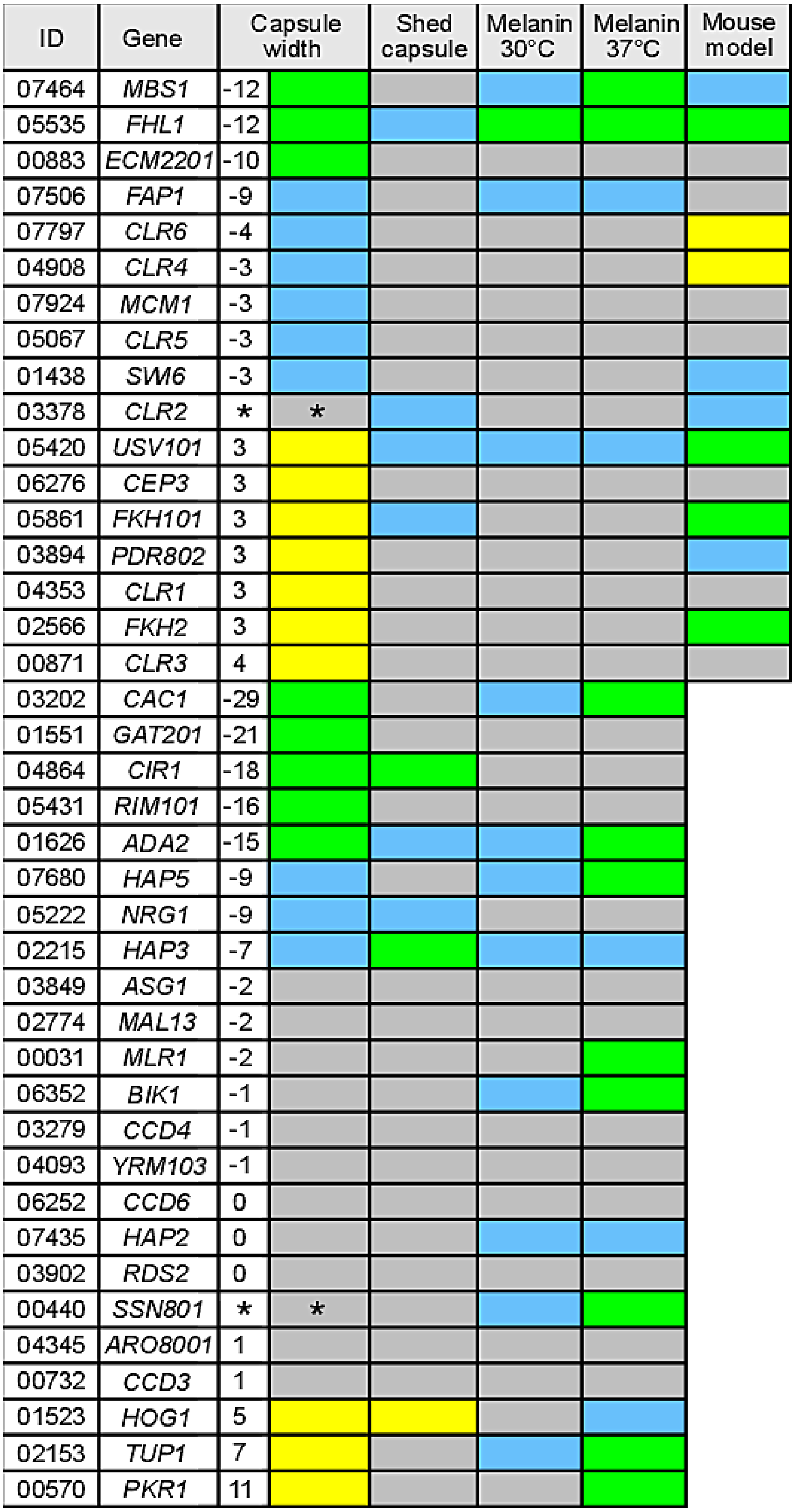
Figure 3.1. NetProphet-PhenoProphet workflow. Yellow, steps involved in network modeling; blue, steps involved in network refinement to elucidate a specific process of interest; green, products of genome-wide network analysis.

3.5 Results

3.5.1 Expression-based network mapping predicts TFs involved in capsule regulation

We selected deletion strains to make and profile in several stages. First, we deleted genes encoding 11 regulators previously reported to participate in capsule synthesis: 8 DNA-binding TFs and 3 signaling proteins (Appendix Table 1, 'Literature'). In the second stage, we deleted 17 genes encoding putative DNA binding proteins based on the correlation of their expression levels with capsule size in various conditions (Haynes et al. 2011; Appendix Table 1, 'Correlation'). We grew these 28 deletion mutants and 3 others (Appendix Table 1, 'Other') in capsule-inducing conditions and assayed them for capsule size. We developed custom software to facilitate precise measurement of capsule thicknesses, enabling statistical analysis of thickness distributions. We further subjected all of the mutant strains to expression profiling by RNA-seq in biological triplicate after a shift to capsule- inducing conditions. We used the expression profiles of these strains and WT controls as input for NetProphet to map the capsule regulation network (Fig. 3.1).

To select the final group of genes to delete, we applied PhenoProphet to the NetProphet-generated network along with 68 genes that have been reported to play a role in capsule production (Appendix Table 2.A). These genes encode a variety of enzymes, transporters, signaling factors, and proteins of unknown function. When we rank-ordered all TFs by their PhenoProphet scores, we found that 14 of the 21 highest-scoring TFs had already been deleted; 12 of these 14 had altered capsule phenotypes. We deleted an additional 10 top-ranked TFs and assayed their gene expression profiles and capsule sizes. Eight of these 10 had altered capsule (80%). For comparison, a traditional screen of mutants that included 64 regulator deletions identified only 3 required for normal capsule regulation (< 5%) (Liu et al. 2008).



ID	Gene	Capsule width		Shed capsule	Melanin 30°C	Melanin 37°C	Mouse model
07464	*MBS1*	-12	🟩	⬜	🟦	🟩	🟦
05535	*FHL1*	-12	🟩	🟦	🟩	🟩	🟩
00883	*ECM2201*	-10	🟩	⬜	⬜	⬜	⬜
07506	*FAP1*	-9	🟦	⬜	🟦	🟦	⬜
07797	*CLR6*	-4	🟦	⬜	⬜	⬜	🟨
04908	*CLR4*	-3	🟦	⬜	⬜	⬜	🟨
07924	*MCM1*	-3	🟦	⬜	⬜	⬜	⬜
05067	*CLR5*	-3	🟦	⬜	⬜	⬜	⬜
01438	*SWI6*	-3	🟦	⬜	⬜	⬜	🟦
03378	*CLR2*	*	*	🟦	⬜	⬜	🟦
05420	*USV101*	3	🟨	🟦	🟦	🟦	🟩
06276	*CEP3*	3	🟨	⬜	⬜	⬜	⬜
05861	*FKH101*	3	🟨	🟦	⬜	⬜	🟩
03894	*PDR802*	3	🟨	⬜	⬜	⬜	🟦
04353	*CLR1*	3	🟨	⬜	⬜	⬜	⬜
02566	*FKH2*	3	🟨	⬜	⬜	⬜	🟩
00871	*CLR3*	4	🟨	⬜	⬜	⬜	⬜
03202	*CAC1*	-29	🟩	⬜	🟦	🟩	
01551	*GAT201*	-21	🟩	⬜	⬜	⬜	
04864	*CIR1*	-18	🟩	🟩	⬜	⬜	
05431	*RIM101*	-16	🟩	⬜	⬜	⬜	
01626	*ADA2*	-15	🟩	🟦	🟦	🟩	
07680	*HAP5*	-9	🟦	⬜	🟦	🟩	
05222	*NRG1*	-9	🟦	🟦	⬜	⬜	
02215	*HAP3*	-7	🟦	🟩	🟦	🟦	
03849	*ASG1*	-2	⬜	⬜	⬜	⬜	
02774	*MAL13*	-2	⬜	⬜	⬜	⬜	
00031	*MLR1*	-2	⬜	⬜	⬜	🟩	
06352	*BIK1*	-1	⬜	⬜	🟦	🟩	
03279	*CCD4*	-1	⬜	⬜	⬜	⬜	
04093	*YRM103*	-1	⬜	⬜	⬜	⬜	
06252	*CCD6*	0	⬜	⬜	⬜	⬜	
07435	*HAP2*	0	⬜	⬜	🟦	🟦	
03902	*RDS2*	0	⬜	⬜	⬜	⬜	
00440	*SSN801*	*	*	⬜	🟦	🟩	
04345	*ARO8001*	1	⬜	⬜	⬜	⬜	
00732	*CCD3*	1	⬜	⬜	⬜	⬜	
01523	*HOG1*	5	🟨	🟨	⬜	🟦	
02153	*TUP1*	7	🟨	⬜	🟦	🟩	
00570	*PKR1*	11	🟨	⬜	⬜	🟩	

Figure 3.2. TF mutants have significant virulence-related phenotypes. For capsule the difference in thickness from WT (in pixels) is tabulated and also color coded: green, a decrease of ≥10 pixels compared to wild type; blue, a

decrease of 3-9 pixels; and yellow, an increase of ≥3 pixels. * denotes strains with WT mean capsule thickness but significantly increased variance. For melanin formation, green indicates colonies that were white (no melanin) to beige and blue indicates colonies that were brown but lighter than WT. For capsule shedding, green indicates that the 3 hour culture supernatant concentration of GXM was ≥8-fold lower than WT; blue, 2-4-fold lower; yellow, 2-fold higher. For short-term infectivity, fold-change in colony forming units (CFU) in 1 week was calculated (tested only for strains with newly discovered capsule phenotypes). Green, >10 times lower than WT; blue, 2 to 10 times lower; and yellow, > 2 times higher. In all columns gray indicates no significant change in phenotype.

Ten of our deletion strains had increased capsule thickness and 17 had reduced capsule thickness (Fig. 3.2). Only 11 of the 27 had been previously reported to influence capsule thickness. (The phenotype of cells lacking *MBS1*, which we studied because of its PhenoProphet score, was reported while our analysis was in progress (Song et al. 2012).) Fig. 3.3.A shows the altered capsule thicknesses of 17 mutants lacking TF-encoding genes that we selected based only on expression data, analyzed by either capsule size correlation or PhenoProphet. Almost half of our new mutants were hypercapsular, a phenotype that has been relatively rarely reported (D'Souza et al. 2001; Bahn et al. 2005; Lee et al. 2009). Many of the mutants with altered capsule size also showed significantly increased capsule size variability. Interestingly, two mutants, *ssn801* and *clr2*, showed a substantial increase in capsule size variability (2.9-fold, $p<10^{-78}$ and 1.6-fold, $p<10^{-18}$, respectively) with no change in mean capsule size.

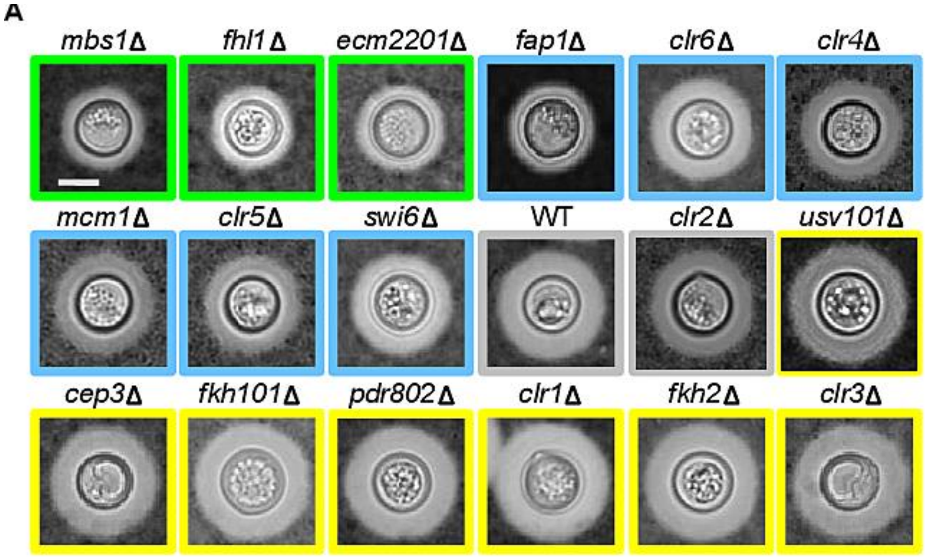


Figure 3.3.A. Capsule thicknesses of 17 novel altered capsule mutants. Representative cells of our new mutant strains, selected so all have similar cell wall diameter and each has capsule thickness very close to the average determined for that mutant. Images are all to the same scale (bar = 5 microns) and ordered by capsule size. Colors indicate capsule size groups as in Fig. 3.2.

We also assessed other virulence-related phenotypes in our uniform collection of 41 mutants. In addition to displaying capsule polysaccharide on its surface, *C. neoformans* sheds this material into the environment, with adverse effects on the host immune response (Coelho et al. 2014). We used a cryptococcal antigen latex agglutination test to assess capsule shedding in our strain set. Interestingly, both hyper- and hypocapsular strains showed alterations in capsule shedding (Fig. 3.2). Many of our new mutants also had defects in melanin production, a virulence factor (Eisenman and Casadevall 2012) (Fig. 3.2). Finally, we tested our new mutants for their ability to grow in the mouse lung; 11 showed a significant change in this characteristic (Fig. 3.3.B).

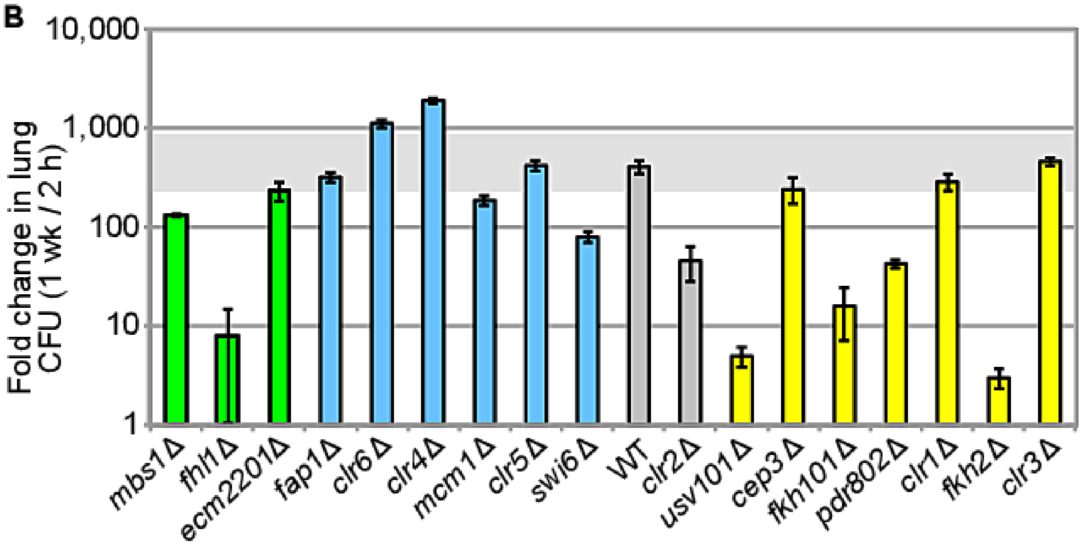


Figure 3.3.B. Mouse lung growth of 17 novel altered capsule mutants. Mean ± SEM of infectivity results; the horizontal gray bar denotes fold-increase values from 0.5 to 2-fold that of WT. All strains grew like WT on rich medium except for *fhl1*, which had a 2.5-fold higher doubling time.

About 2/3 of the new mutants with abnormal capsules also showed defects in at least one other virulence-related trait (Fig. 3.3.C, left). Several showed defects in all traits measured, including *usv101*, a novel virulence regulator (Fig. 3.2). We also identified three novel factors, Hap2, Bik1, and Mlr1, that are not involved in regulating capsule, but do yield isolated melanin defects (Fig. 3.3.C, right); Mlr1 further has no *S. cerevisiae* ortholog. Notably, all four TF mutants with reduced capsule shedding that we tested in mice had an infectivity defect. This suggests that capsule shedding is critical for infectivity, regardless of whether surface capsule is reduced or enlarged.

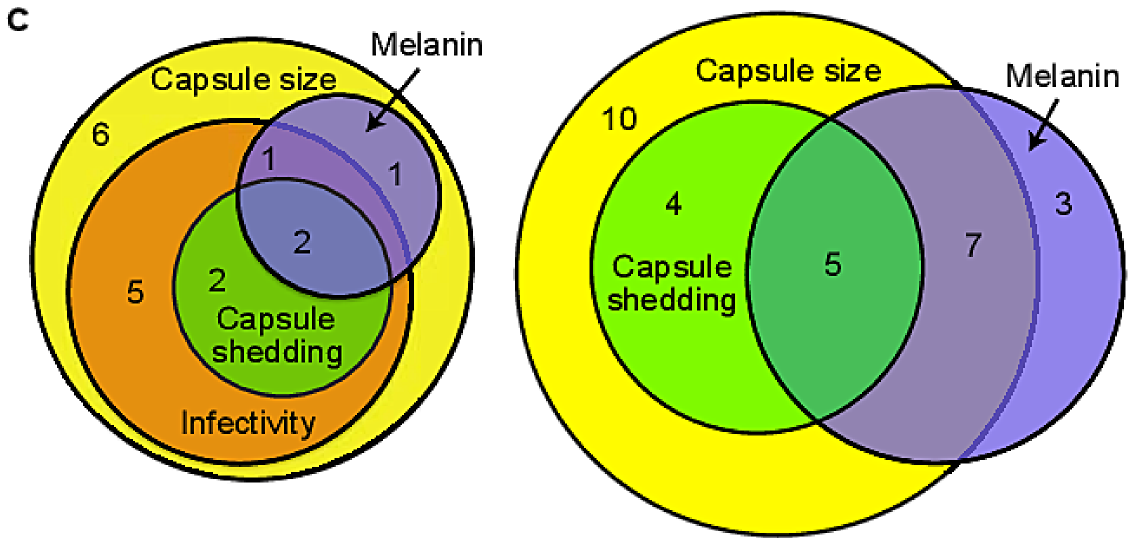


Figure 3.3.C. Aberrant phenotypes of new mutants (left) and of all the mutants in Fig. 2 (right). Melanin was scored for 37 °C phenotype.

3.5.2 PhenoProphet accurately predicts which TFs will have altered capsule thickness

A remarkably large fraction of the TFs identified by PhenoProphet were involved in capsule regulation (Appendix Table 1). We compared this result to our previous strategy of using the correlation of gene expression with capsule thicknesses to predict TF capsule involvement, using our previously published correlation scores (Haynes et al. 2011). Of the 17 TFs we selected for deletion by the correlation method, 8 had altered capsule thickness (47%); of the 10 we selected by PhenoProphet, 8 had aberrant or hypervariable capsule thickness (80%). To further compare capsule-size correlation and PhenoProphet to each other and to other methods of phenotype prediction, we applied each method to a set of TFs for which the capsule phenotype of the corresponding deletion was known. These genes had primarily been deleted because they were suspected to have a role in capsule regulation (Appendix Table 2.B). To simulate prospective phenotype prediction, we used leave-one-out cross validation, in which each mutant phenotype is predicted without using any data derived from that mutant. As a simple, baseline prediction method, we considered the hypothesis that TF genes that display significant expression changes upon capsule induction are more likely to be required for normal capsule induction than those that do not. The data did not support this hypothesis (Fig. 3.4.A, green). Next, we considered the possibility that genes whose expression is significantly correlated with capsule thickness would be more likely to encode TFs regulating capsule than genes whose expression is not correlated with capsule thickness. The data did not support this hypothesis, either (Fig. 3.4.A, red).

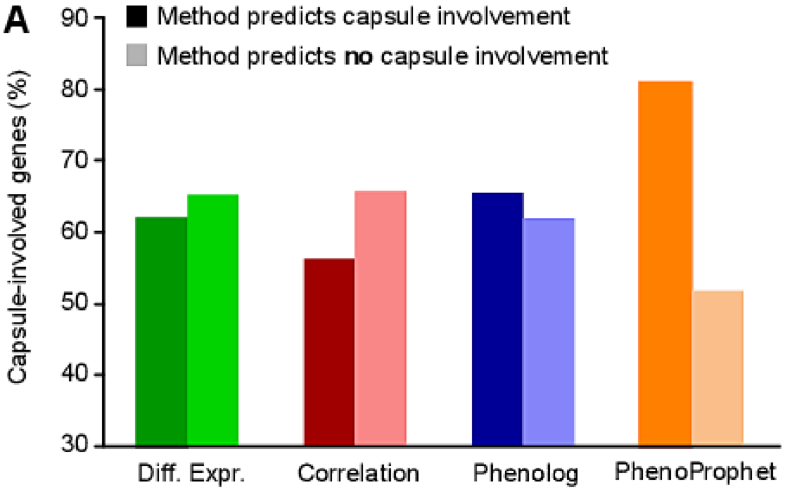


Figure 3.4.A. Comparison of methods for predicting capsule involved regulators. Four expression-based methods (x-axis) were used to predict whether cryptococcal regulators are involved with capsule formation (mutants show abnormal capsule; dark bars) or not (light bars). The set of genes in each category was then assessed for the fraction that actually does influence capsule (y-axis).

We next tried a phenotype prediction method called Phenologs, which is based on the observation that genes sharing a phenotype in one organism often share phenotypes in another organism, even when the phenotypes themselves appear to be unrelated (McGary et al. 2010). However, TFs with positive Phenolog scores were not significantly enriched for those that affect capsule thickness (Fig. 3.4.A, blue; Fisher's exact p=0.43), nor were the Phenolog scores of TFs that do affect capsule thickness greater, on average, than those of TFs that do not (Fig. 3.4.B, blue; Mann-Whitney *U* test p=0.33). Thus, Phenolog scores do not have discriminative value in this application. In contrast to all of these methods, TFs with positive PhenoProphet scores were significantly enriched for those that affect capsule thickness (Fig. 3.4.A, orange; Fisher's exact p < 0.03). Furthermore, the mean of PhenoProphet scores for TFs that affect capsule thickness was significantly greater than the mean score for TFs that do not affect capsule thickness (Fig. 3.4.B, orange; Mann-Whitney *U* test p < 0.02).

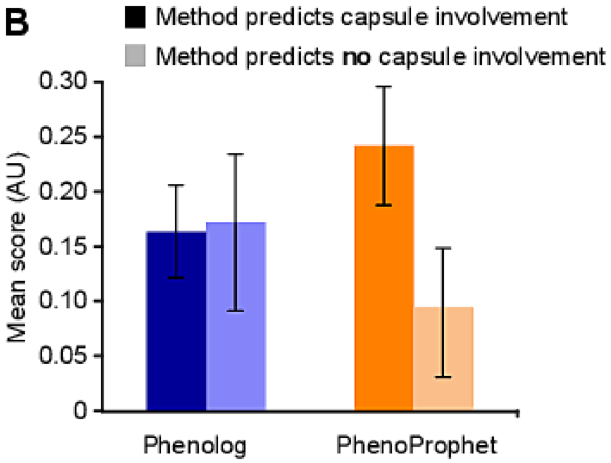


Figure 3.4.B. Comparison of scores assigned by methods for predicting capsule involved regulators. Mean and SEM of Phenolog and PhenoProphet scores for genes which are capsule involved (dark bars) or not (light bars).

The predictive power of PhenoProphet relative to the other methods is confirmed by receiver operating characteristic analyses (Fig. 3.4.C).

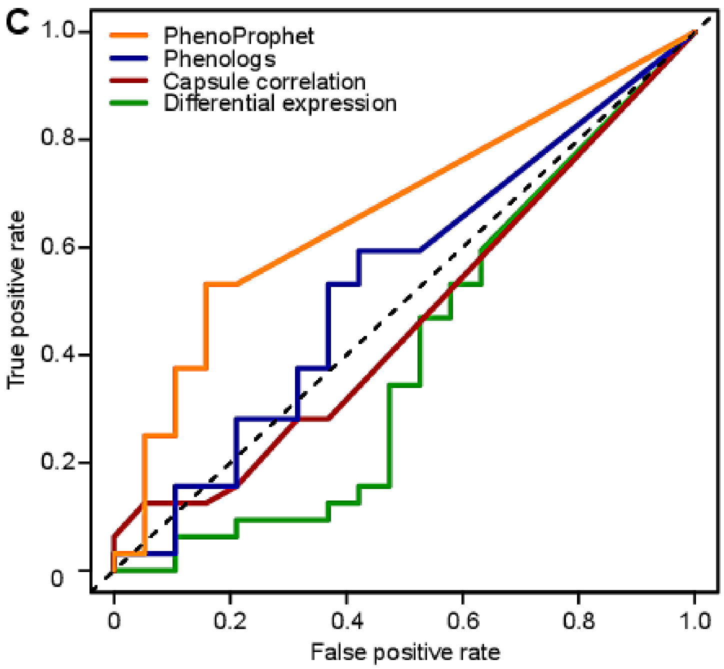


Figure 3.4.C. Receiver operating characteristic analysis comparing the indicated methods for phenotype prediction to random expectation (dotted line).

Next, we investigated the effect of the number of expression-profiled TF deletion strains on the accuracy of PhenoProphet. The results showed that the predictive accuracy of PhenoProphet exceeded chance (and the accuracy of Phenologs) even when the number of profiled TF-deletion strains given to NetProphet was reduced to 25% of the total (11 TFs deleted; Fig. 3.5.A). Providing profiles of more TF-deletion strains increased accuracy, confirming that PhenoProphet depends on the NetProphet network for its accuracy. In applications where no TF network is available and the number of TF-deletion strains that can be profiled is less than 10 another method (such as Phenolog analysis) may be most useful. We also investigated the effect of the number of known capsule-involved genes, with results similar to those described above for deletion-strain profiles (Fig. 3.5.B.)

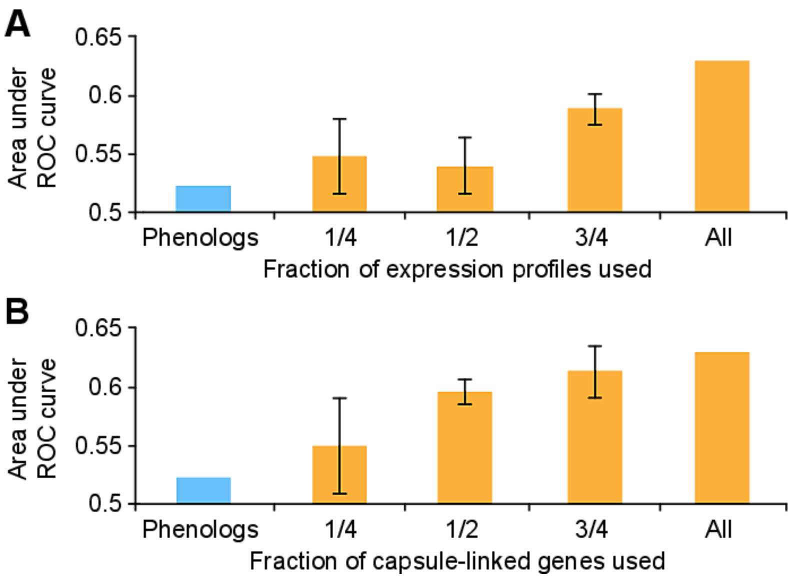


Figure 3.5.A-B. Examination of PhenoProphet accuracy in varying conditions. Panel A, effect of number of regulator-deletion expression profiles given to NetProphet on the accuracy of PhenoProphet. Panel B, effect of the number of known phenotype-linked genes on PhenoProphet accuracy.

3.5.3 NetProphet predicts functional, direct binding of TFs to their targets

We previously validated NetProphet in *S. cerevisiae* using data from ChIP-chip and protein-binding microarrays (Haynes et al. 2013). To validate NetProphet in *C. neoformans*, we focused on Gat201, the only cryptococcal TF for which ChIP data was available (Chun et al. 2011); Nrg1, a well studied capsule regulator (Cramer et al. 2006); and Usv101, a capsule regulator described here for the first time. We epitope-tagged the last two and carried out ChIP-seq. We then tested the NetProphet-predicted targets of Usv101, Gat201, and Nrg1 for significant overlap with their ChIP-positive targets. NetProphet assigns a confidence score to each potential target of a TF, so we tested top-scoring target groups of various sizes, from 40 to 200. For all three TFs, the 40 most confident NetProphet predictions were highly enriched for ChIP-positive targets, as compared to the number that would be expected if 40 genes were chosen at random (Fig. 3.6.A-C).

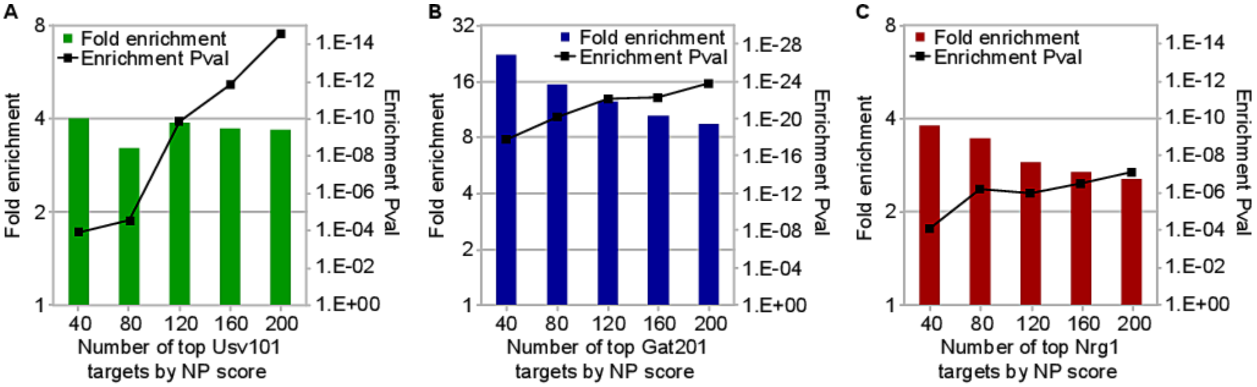


Figure 3.6.A-C. ChIP enrichment of NetProphet predicted targets. Colored bars, cumulative fold-enrichment of the top NetProphet-predicted targets of each TF for ChIP-positive targets, relative to the fraction of ChIP-positive targets among all genes; square symbols, significance of the enrichment (pval). The horizontal axis indicates the number of top-ranked NetProphet-predicted targets considered.

To further test our predictions, we compared models of TF binding specificity inferred from the NetProphet-predicted targets to specificity models derived from ChIP (Fig. 3.6.D-F, top two logos). For both Usv101 and Gat201, the Cryptococcus motifs derived from NetProphet predictions and from ChIP data were highly similar, whereas for Nrg1 there were significant differences between the two. For comparison, we extracted a motif for the closest homolog of each TF in *S. cerevisiae* from the ScerTF database (Spivak and Stormo 2012) (Fig. 3.6.D-F, bottom logo). The motif of *C. neoformans* Usv101 (confirmed independently by NetProphet and ChIP) has diverged substantially from the motif of *S. cerevisiae* Usv1 on one side. No motif is available for Gat2, the ortholog of Gat201, but the motif of Ecm23, the next best homolog, shows the expected GATA family resemblance. The motif for *S. cerevisiae* Nrg1 supports the ChIP-derived Cryptococcus motif over the NetProphet-derived motif; this is likely because the NetProphet-predicted Nrg1 targets include some indirect targets regulated by TFs downstream of Nrg1.

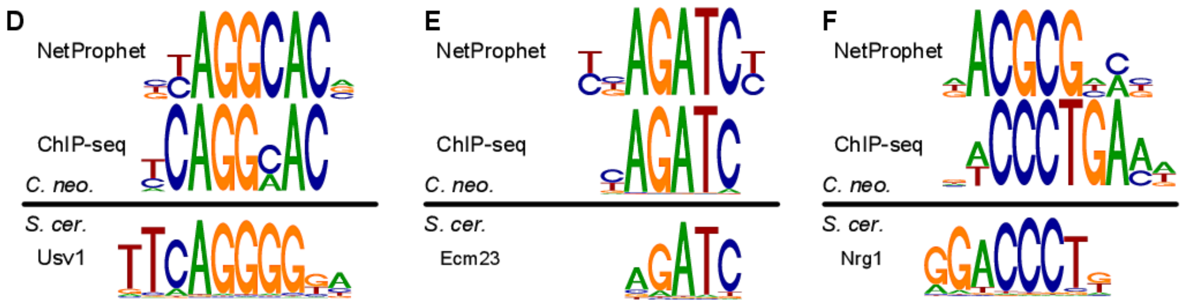


Figure 3.6.D-F. Comparison of derived binding motifs. For the indicated TFs, *C. neoformans* binding motifs derived from the promoters of NetProphet-predicted targets (top) or from the regions around ChIP-seq peaks (middle) are compared to *S. cerevisiae* binding motifs (bottom).

We then attempted to infer binding specificity models (PWMs) for all other Cryptococcus TFs from their NetProphet-predicted target sets (Elemento and Tavazoie 2005). We tested these PWMs for significant conservation in the genome of a related species (JEC21, serotype D). We also tested each TF to determine whether there was a highly homologous TF in *S. cerevisiae* with highly similar binding specificity, as in Figure 3.6.D, E. In total, 18 PWMs showed both types of conservation and were therefore deemed reliable models (Fig. 2.6). Previously, binding specificity was known for only 2 TFs in Cryptococcus (Chun et al. 2011; O'Meara et al. 2014), both of which strongly support our independently derived PWMs.

3.5.4 ChIP-experiments validate NetProphet predictions

We combined NetProphet and ChIP results from Usv101, Gat201, and Nrg1 to produce a high-confidence core for our model of the network regulating virulence in *C. neoformans* (Fig. 3.6.G). This reveals a highly interconnected subnetwork in which Usv101 represses *GAT201*, consistent with their opposite capsule phenotypes (large vs. small). Usv101 also represses several sugar transporters while activating *HXT1*, a hexose transporter that has a hypercapsular phenotype (Chikamori and Fukushima 2005). Usv101 represses, and Gat201 activates, *BLP1*, which is involved in a capsule-independent anti-phagocytic mechanism (Chun et al. 2011). The opposing effects of Usv101 and Gat201 on *BLP1*, which promotes fungal survival during infection, are consistent with their generally opposing roles in regulating virulence and their opposite capsule phenotypes. Usv101 represses *BLP1* both directly and via its repression of *GAT201*, forming a coherent feed-forward loop.

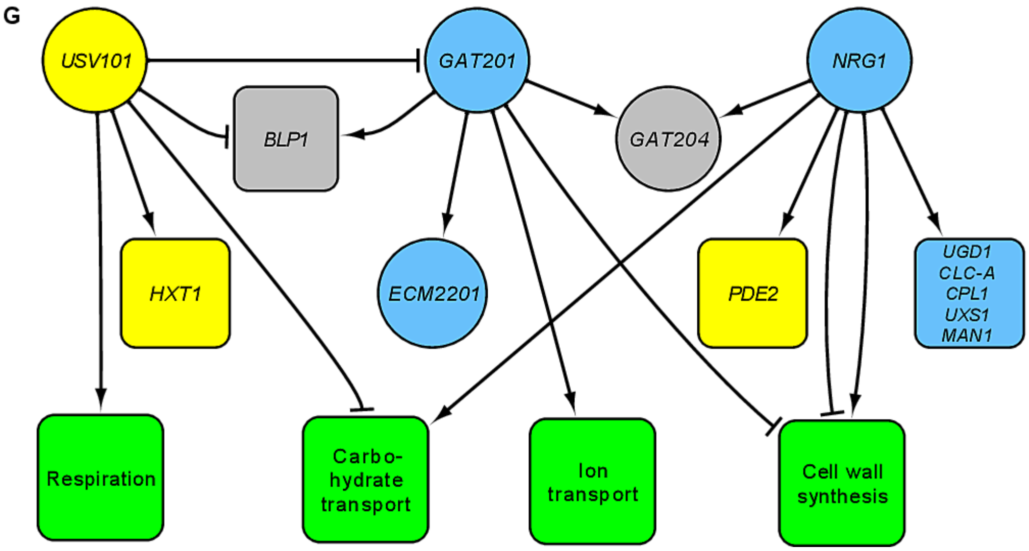


Figure 3.6.G. Network diagram showing the three validated TFs with targets supported by both NetProphet and ChIP analysis that are relevant to cryptococcal virulence. Round nodes, TFs; square nodes, target genes or gene ontology biological process terms for which the targets of the indicated TF are enriched. Blue nodes, mutants are hypocapsular; yellow nodes, mutants are hypercapsular; gray nodes, mutants are defective in capsule-independent phagocytosis. Edges with arrowheads indicate activation while those with T-heads indicate repression. ChIP evidence suggests that Usv101 binds to its own promoter but expression evidence cannot determine whether this binding results in activation, repression, or no effect.

Our data support a previous report (Chun et al. 2011) that Gat201 activates *ECM2201*, encoding a TF, and we report for the first time that the *ecm2201* mutant is hypocapsular (Fig. 3.2). Gat201 has the same effect on capsule size as Nrg1 and works with Nrg1 to stimulate expression of *GAT204*, which encodes a second TF involved in the capsule-independent anti-phagocytic mechanism (Chun et al. 2011). Both Gat201 and Nrg1 repress some genes involved in cell wall synthesis but Nrg1 also activates other genes involved in cell wall synthesis, suggesting that cell wall may be reconfigured during capsule induction. Nrg1, which is activated by cAMP signaling in Cryptococcus (Cramer et al. 2006), also activates *PDE2*, encoding a phosphodiesterase that reduces cAMP levels (Zaman et al. 2008), thus adding a slow, transcriptionally-mediated negative feedback loop in cAMP signaling to the fast, post-translational negative feedback loops that have been reported (Hicks et al. 2005; Kronstad et al. 2011).

Nrg1 further activates *UGD1*, *MAN1*, and *UXS1*, which encode glycoactive proteins (Fig. 3.6.G), as well as *CLC-A* and *CPL1*, which encode proteins involved in maintaining ion balance (Zhu and Williamson 2003 and our unpublished data). Deletion of any of these five genes results in a hypocapsular phenotype. Thus the hypocapsular phenotype of the *nrg1* mutant may be caused by its failure to activate expression of these five genes. Broadly speaking, the TFs in Fig. 3.6.G have the same phenotypes as the targets they activate. Furthermore, TFs whose absence affects capsule thickness in opposite directions (Usv101 vs. Gat201 and Nrg1) regulate their common targets in opposite directions, whereas TFs whose absence affects capsule in the same direction (Gat201 and Nrg1) also regulate their common targets in the same direction.

3.5.5 NetProphet illuminates transcriptional dynamics

To gain insight into the transcriptional dynamics of cryptococcal capsule induction, we performed RNA-seq on WT cells immediately before transfer from rich media into capsule-inducing conditions and at 1.5, 3, 8, and 24 h after transfer. Considering all genes across the time course (Fig. 3.7.A), one pattern that emerged was repression of genes involved in ribosome biogenesis, tRNA synthesis and processing, amino acid biosynthesis, and protein transport, along with induction of genes involved in specific amino acid degradation and protein degradation; this is consistent with the cells accommodating to scarcer nutrients and slower growth. Expression of some nuclear genes encoding cytochrome-C oxidase (COX) components declined while expression of mitochondrial genes encoding COX components increased. Expression of all 13 mitochondrial genes increased significantly (mean fold-change 57, median 20).

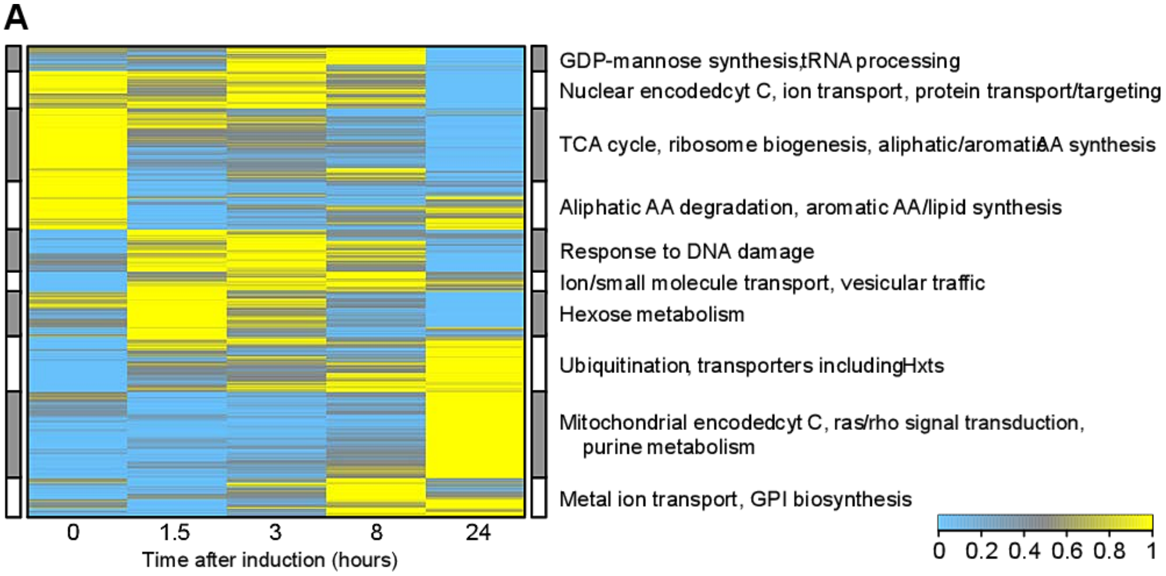


Figure 3.7.A. heat map of genome-wide expression profiling; blue, low expression; yellow, high expression. Examples of functional annotations of genes in each cluster are indicated.

To map the transcriptional dynamics onto our network, we divided TFs whose deletions alter capsule thickness into four groups based on their temporal expression patterns (Fig. 3.7.B, circles). The most “upstream” acting TFs in the network form a group with slightly increased expression at 90 minutes followed by sharply decreased expression over the next 24 hours (Fig. 3.7.B, Regulator Group 1). This group includes *activators* of ribosome biogenesis genes and *repressors* of mitochondrially-encoded respiration genes. It also contains *repressors* of a cluster of capsule-involved genes whose expression increases steadily through capsule induction (Fig. 3.7.B, Box A). This target group includes genes encoding proteins involved in nucleotide sugar synthesis and transport, polysaccharide synthesis, and maintenance of inorganic ion and osmotic balances. Group 1 includes *activators* of a set of capsule-involved genes that decreases steadily after the first few hours of capsule induction (Fig. 3.7.B, Box B). These declining genes include four that encode proteins that promote cAMP/PKA signaling, reinforcing the transcriptionally mediated negative feedback on cAMP/PKA signaling noted above.

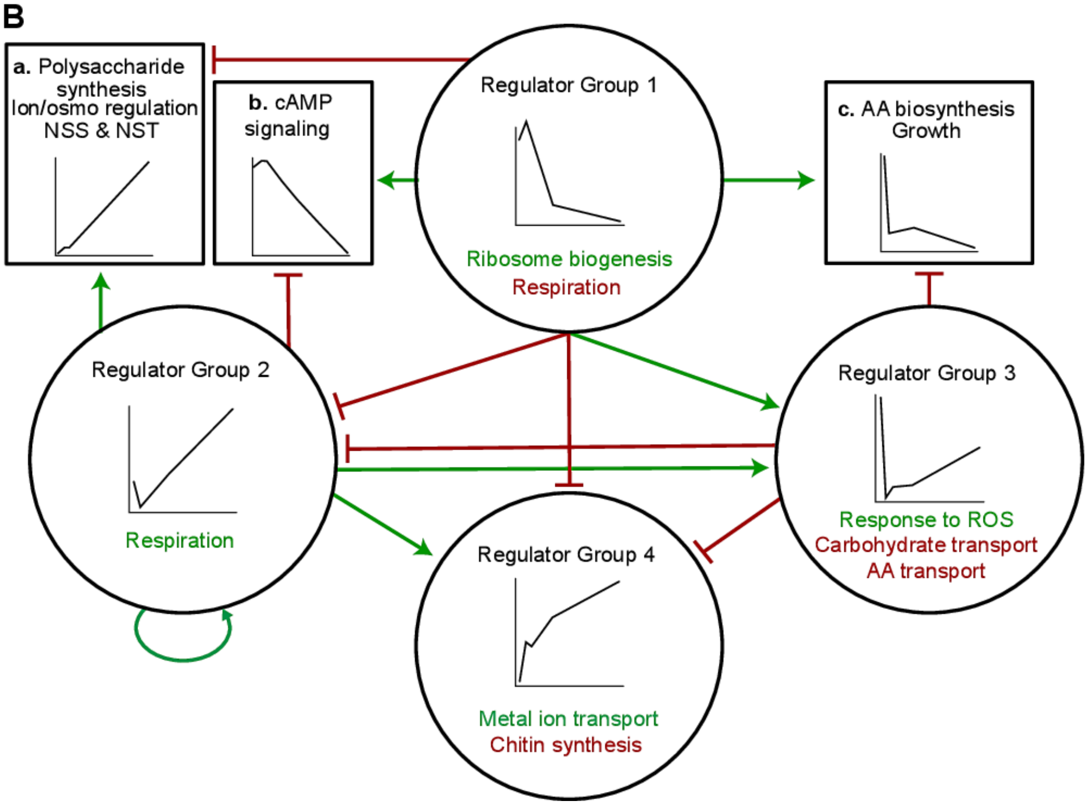


Figure 3.7.B. Regulatory relations among groups of capsule-involved genes clustered by temporal expression pattern. Circles, groups of regulators clustered by their temporal expression patterns (insets); text denotes representative functions of their activated (green) or repressed (red) targets. Boxes, *capsule-involved* target genes clustered by their temporal expression patterns (insets); text provides examples of target gene functions. Green arrows, activation of designated targets; red T-heads, repression of designated targets. Regulator Group 1 is Clr1, Hap5, Nrg1, Pkr1, and Ssn801; Regulator Group 2 is Cac1, Cep3, Cir1, Clr2, Fap1, Fhl1, Fkh2, Gat1, Mcm1, Pdr802, Sp1, Swi6, and Usv101; Regulator Group 3 is Clr6, Hog1, Mbs1, Ste12alpha, and Tup1; Regulator Group 4 is Ada2, Clr3, Clr4, Clr5, Ecm2201, Fkh101, Gat201, Hap3, and Rim101. NSS, nucleotide sugar synthesis; NST, nucleotide sugar transport; ROS, reactive oxygen species.

Group 2 is the opposite sign partner of Group 1 and cooperates with it in nearly every way: Its expression pattern is opposite that of Group 1 and it regulates mitochondrially-encoded respiration genes as well as each cluster of capsule-involved non-TFs in the opposite way from Group 1. Since Group 2 is repressed by Group 1 the indirect effects of Group 1 via Group 2 are consistent with the direct effects of Group 1.

As with Group 2, the expression of regulators in Group 3 decreases from time 0 to 90 min, then reverses course and increases from 90 min to 24 hr. The difference is that Group 2 regulators rapidly recover to well beyond their initial levels, whereas Group 3 regulators never recover their initial levels. This difference may result from the tendency of Group 2 regulators to activate each other, forming positive feedback loops. Group 3 regulators activate genes involved in the response to reactive oxygen species (ROS) while repressing certain carbohydrate and amino acid transporters. Group 3 regulates a set of capsule-involved genes that has an "L" shaped expression pattern (Fig. 3.7.B, Box C) and includes genes involved in amino acid biosynthesis and other growth- related processes.

Group 4 regulators are regulated by all of the other groups but they do not regulate other groups, putting them at the bottom of the hierarchy. Their expression increases steadily through induction and they include activators of genes involved in metal ion transport and synthesis of chitin, a component of cell wall.

Taken together, these analyses show a hierarchy of TFs (circles), with those expressed in an "inverted check-mark" (Group 1) at the top, those that first decrease and then increase in the middle, and those that increase steadily at the bottom. Capsule-involved, non-TF genes (boxes) are expressed in temporal patterns that are generally consistent with those of their regulators – the same pattern for activators and the opposite for repressors. These observations suggest that the temporal patterns of downstream genes can in many cases be explained by the patterns of their regulators shown in Figure 3.7.B. For example, Group 1 and Group 2 regulators that are connected by an edge in the underlying network show an average temporal correlation of -0.84. For Group 1 and Group 3, the average correlation is -0.16 and for Group 2 and Group 4 it is +0.60.

The fact that we do not see any delay between the changes in upstream regulators and those in downstream targets is consistent with the expectation that the translation of TF-encoding mRNAs and the initial response by target genes should occur on a faster time scale (< 0.5 hr) than the interval between samples in our time course (1.5-16.0 hr).

The hierarchical relationships among these four regulator groups were confirmed by comparing the number of factors regulating each regulator to the number of its targets (Fig. 3.7.C). This analysis shows that the regulators in Group 1 have more targets than regulators, those in Groups 2 and 3 have about the same number of targets and regulators, and those in Group 4 have more targets than regulators.

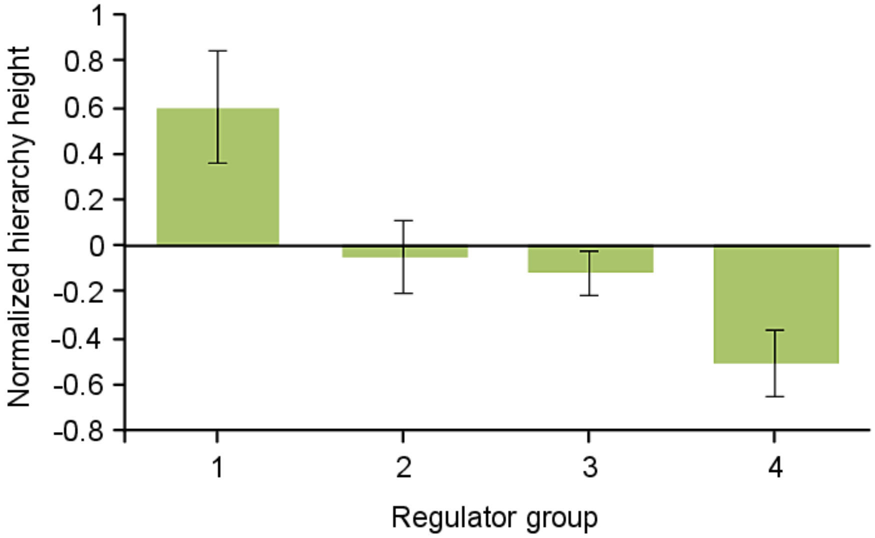


Figure 3.7.C. Normalized hierarchy heights (NHH) of regulators in Groups 1-4 of Fig. 3.7.B. NHH is the number of outgoing edges minus the number of incoming edges divided by outgoing plus incoming.

3.5.6 Network analysis reveals mechanisms of capsule biosynthesis regulation

While little is known about the glycosyltransferase reactions that generate capsule polysaccharides, the upstream pathways that form precursors for this process are well defined. Nucleotide sugar donors of mannose, galactose, glucuronic acid, and xylose are synthesized in the cytosol and transported into the Golgi for use in capsule synthesis (Fig. 3.8.A). We analyzed our network to gain insight into the regulation of these key metabolic processes. *PSA101* is the most heavily regulated of the genes we analyzed, suggesting that it is a key point at which transcriptional regulation affects biosynthesis. Of the two genes encoding GDP-mannose transporters, *GMT1*, which has a much greater effect on capsule size (Wang et al. 2014), is regulated by the major capsule regulators Cir1 (Jung et al. 2006), Rim101 (O'Meara et al. 2014), and Ada2 (Haynes et al. 2011). *GMT2*, whose deletion shows a phenotype only when *GMT1* is also inactivated (Wang et al. 2014), is repressed by Usv101 and activated by Rds2, a TF without a significant capsule thickness phenotype. The degree of regulation by capsule- involved TFs thus highlights the transporter that is more heavily involved in capsule synthesis.

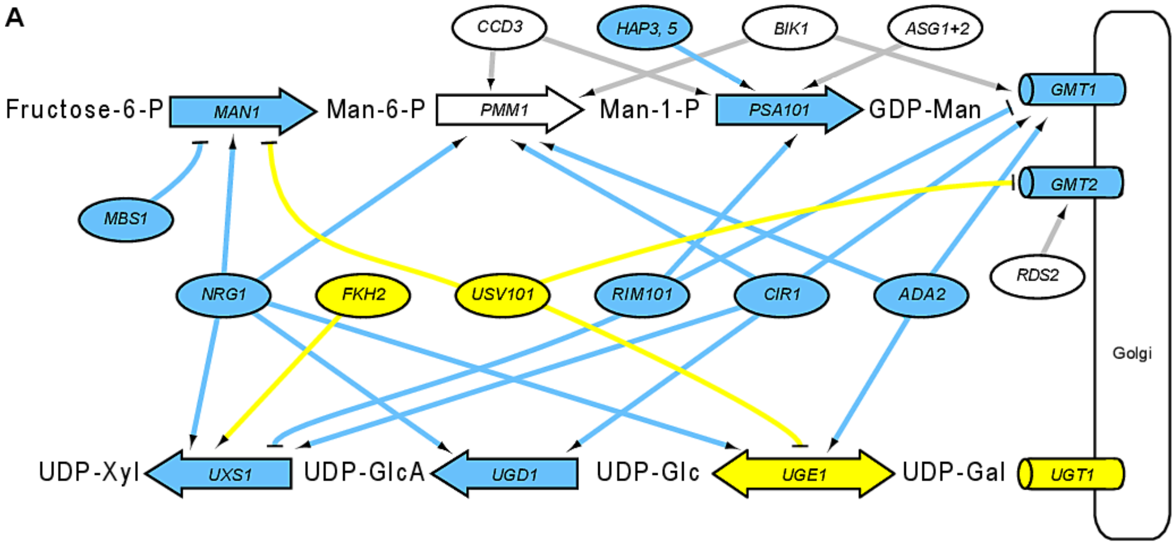


Figure 3.8.A. Regulation of upstream capsule biosynthetic pathways. Network-derived regulatory relationships between TFs and the pathways that make and localize sugar donors for capsule synthesis, considering the top 10,000 NetProphet edges. Large labeled arrows, synthetic steps; ovals, TFs; cylinders, nucleotide sugar transporters in the Golgi membrane. Shapes are labeled with the corresponding gene name and filled blue if the mutants are hypocapsular, yellow if the mutants are hypercapsular, and white if the gene has not been deleted (*PMM1*) or the mutants have normal capsule thickness (all others). Arrowheads indicate activation and T-heads repression; edge colors reflect the phenotype of the regulator. *ASG1*+2 represents 3 genes which are normocapsular when deleted: *ASG1*, *CCD6*, and *MAL13; HAP3*,*5* represents *HAP3* and *HAP5*, both hypocapsular when deleted. Man, mannose; Xyl, xylose; GlcA, glucuronic acid; Glc, glucose; Man1, phosphomannose isomerase; Pmm1, phosphomannomutase; Psa1, GDP-mannose pyrophosphorylase; Uxs1, UDP- Xyl synthase (Bar-Peled et al. 2001); Ugd1, UDP-Glc dehydrogenase (Bar-Peled et al. 2004; Moyrand and Janbon 2004); and Uge1, UDP-Glc epimerase. Gmt1 and Gmt2 are GDP-Man transporters (Cottrell et al. 2007; Wang et al. 2014) and Ugt1 is a UDP-Gal transporter (Moyrand et al. 2007); transporters for the other precursors have not been identified.

The TFs that regulate the largest number of genes involved in upstream capsule biosynthetic processes are Nrg1, whose mutant is severely hypocapsular, closely followed by

Usv101, whose hypercapsular deletion phenotype is reported here for the first time, Cir1, Rim101 and Ada2. Regulation of capsule biosynthetic enzymes and transporters is sufficient to explain the phenotypes of mutants lacking Cir1, Nrg1, Usv1, and Ada2: Usv101 (*hyper*capsular) primarily represses these pathways while the others (*hypo*capsular) primarily activate them. The hypocapsular phenotype of *rim101* is less well explained, as Rim101 appears to repress *UXS1* and *GMT1* while activating *PMM1* and *PSA1*. Likewise, the phenotypes of *mbs1* and *fkh2* are not explained by the relationships we have identified, suggesting that some of their other targets may have as-yet-unknown roles in capsule synthesis.

The TFs that regulate the metabolic pathways shown in Fig. 3.8.A also regulate one another in what is largely a feed-forward hierarchy (Fig. 3.8.B). Cir1, Nrg1, and Usv101 sit at the top of the hierarchy, each regulating multiple other TFs. The other regulator at the top of the pathway is Ccd3, which interestingly does not have a capsule size phenotype despite activating four TFs with hypocapsular phenotypes. Clr2, Mbs1, Fkh2, Hap5, and Bik1 form an intermediate layer, and Hap3, Rds2, Rim101, and Ada2 appear at the bottom of the cascade because they regulate enzymes and transporters directly but do not regulate other TFs in this context. Of the 7 TFs that regulate only one or two biosynthetic genes in Fig. 3.8.A and do not regulate other TFs in Fig. 3.8.B, only 1 is required for normal capsule thickness. In contrast, all 5 of the TFs that regulate 3 or more genes in Fig. 3.8.A are required.

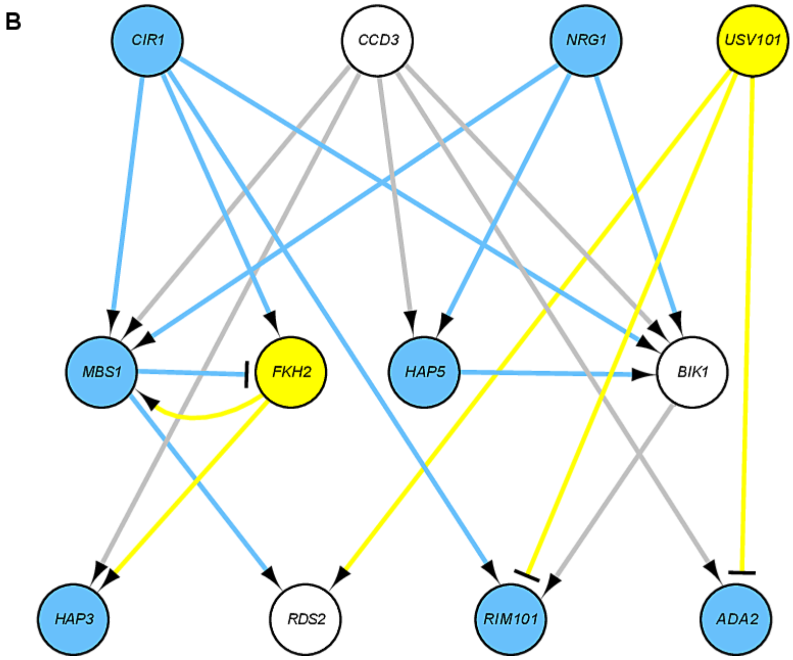


Figure 3.8.B. Network-derived regulatory relationships between the TFs shown in Figure 3.8.A, using the same colors and symbols except that TFs are circles. For clarity, only the top 10,000 NetProphet edges were used.

3.6 Discussion

NetProphet and PhenoProphet enable individual labs lacking the resources of a genome center to systematically and efficiently study the transcriptional regulation of a specific physiological process. Currently, the TFs that regulate a process of interest are typically discovered by large-scale mutant screens and TF-target relations are mapped in big-science projects that do not focus on TFs with specific biological functions (Harbison et al. 2004; Hu et al. 2007; Kemmeren et al. 2014). Our approach brings TF discovery and mapping together through focused, iterative network construction and analysis. We demonstrated this approach by mapping the network that regulates the major virulence factor of *Cryptococcus neoformans*, a deadly human pathogen. Key to the success of this effort was PhenoProphet’s accuracy in identifying TFs that are required for normal capsule growth. This enrichment for TFs involved in capsule regulation enabled us to perform quantitative capsule-size assays that are more sensitive, but also more labor-intensive, than those used in traditional screens. Our approach enables TF networks to be mapped using only gene perturbation and expression profiling, both of which are straightforward in most experimental systems. Indeed, the number of TF-perturbation expression profiles for mammalian systems is growing rapidly, facilitating the application of our approach to mammals. No single approach has perfect sensitivity and specificity, so large-scale mutant screens and ChIP-seq remain important complementary methods. Nonetheless, we have filled a significant methodological gap between single-gene approaches and undirected genomic approaches.

Using the NetProphet-PhenoProphet approach, we produced a comprehensive map of the TF network that regulates cryptococcal capsule size, increased the number of TFs known to regulate capsule from 11 to 27, and increased the number of *C. neoformans* TFs with known sequence specificity from 2 to 18. In the course of this work, we generated a rich resource for systems biology of fungal virulence. We increased the number of publically available RNA-seq profiles from *C. neoformans* TF-deletion studies 20-fold, more than doubled the total number of Cryptococcus expression profiles (including microarrays), presented the first time course of expression during capsule induction, and generated virulence-related phenotypes for 41 regulator deletion mutants under identical conditions, including all known TF mutants that affect capsule size. Taken together, our data sets form the most comprehensive resource for regulatory systems biology available for any fungal pathogen. We expect that this data set, like large-scale data sets for *S. cerevisiae* (Harbison et al. 2004; Hu et al. 2007), will catalyze the development of powerful new network analysis and phenotype prediction algorithms.

Our kinetic evaluation of gene expression during capsule induction allowed us to cluster major regulators based on their temporal expression patterns. We found that the TFs comprising

Group 1 decrease in expression during capsule induction, releasing repression of Group 2 TFs, which correspondingly increase in expression. Groups 1 and 2 have strikingly similar net effects through opposite expression patterns and opposite effects on target expression, forming coherent feed forward loops. Our dynamic analysis also revealed how regulators interact to influence general cellular processes as well as capsule synthetic pathways. For example, the Group 1 and 2 regulators cooperate to induce mitochondrially-encoded respiration genes, resulting in massive upregulation of these genes (mean, 57 fold; median, 20 fold). This is interesting because host conditions are hypoxic (Erecinska and Silver 2001) and the virulence of *Cryptococcus gattii*, which can cause fatal infections in immunocompetent individuals, is closely associated with upregulation of mitochondrial gene expression (Ma et al. 2009). We also integrated our broad analysis of transcriptional dynamics with our focused analysis of nucleotide sugar regulation. Both analyses highlight hierarchies of transcription factors that are largely consistent with one another. They agree that Nrg1 is at the top, regulating many capsule-involved TFs but not itself transcriptionally regulated. They further agree that Hap3, Rim101, and Ada2 are at the bottom, regulated by many TFs and regulating relatively few, and that Usv101, Mbs1, and Fkh2 play both roles, integrating signals from master regulators and distributing them to lower level regulators.

Much of the downstream machinery required for capsule polysaccharide synthesis has not yet been identified. We expect that the NetProphet network, in addition to efficiently identifying novel capsule regulators, will address this gap. For example, we noticed that CNAG_03320 (the more diverged of two cryptococcal homologs of the *S. cerevisiae* GDP-mannose pyrophosphorylase Psa1) was regulated by large numbers of capsule-involved TFs, suggesting that it might have a role in synthesizing capsule precursors. When we tested this idea by deleting the gene, now named *PSA101*, the mutants were indeed severely hypocapsular. Another way in which the network can highlight relevant biosynthetic machinery is illustrated by Ecm2201, a TF that is required for normal capsule growth but does not regulate any other genes known to be required for normal capsule. We anticipate that the targets of Ecm2201 and other TFs with unexplained phenotypes include missing elements of capsule biosynthetic pathways. Filling in these gaps in knowledge about synthesis of a major virulence factor that has no parallel in human cells may help identify targets for future antifungal therapy.

In addition to capsule size, we assayed our matched set of mutants for other virulence-related phenotypes, including capsule shedding, melanization, and infectivity in a short-term mouse model. This revealed some surprising relationships among phenotypes. Many of our novel TF mutants that affected capsule thickness also affected infectivity, with both *hyper*capsular and *hypo*capsular mutants showing reduced infectivity. Hypercapsular mutants were particularly impaired in this regard. This is consistent with a recent report that virulence in *C. neoformans* is positively correlated with rate of uptake by macrophages, which is negatively correlated with capsule size (Sabiiti et al. 2014). Reduced capsule shedding was a strong and significant predictor of reduced infectivity ($p. < 0.03$). Thickness and shedding were not clearly related, suggesting that these processes are independently regulated and that enlarged capsules might result from increased production in some cases and reduced shedding in others. We also observed that deletion of TFs frequently increases capsule size variability, showing that variability is controlled by TFs, probably through negative feedback loops.

In this work, we report a significant advance in the efficiency with which TFs that regulate a specific biological process can be identified and their regulatory networks mapped. We further used that technical advance to gain major insights into fungal virulence regulation. In the process we produced a valuable resource for regulatory systems biology of fungal pathogens, comprising high quality gene expression and phenotype data produced by a single laboratory using a consistent strain background. We expect that our methodological advances will have a broad impact in systems biology and that our discoveries and data resources will transform our understanding of fungal virulence.

3.7 Methods

3.7.1 Materials, strains, and cell growth

Cell culture media (i.e. Dulbecco's Modified Eagle's Medium, D6429), chemicals (i.e. L-DOPA for melanization, D9628), and PCR primers were from Sigma- Aldrich, PCR purification (28106) and gel extraction (28706) kits from Qiagen, and reagents used for RNA-seq, such as the SuperScript III Kit (18080) and the mRNA Catcher Plus Kit (K1570), from Life Technologies. Strains were made in *C. neoformans* KN99α (Nielsen et al. 2005) with standard growth at 30 °C in yeast peptone dextrose (YPD) medium. For capsule induction, an overnight culture in YPD was washed, resuspended in DMEM, and grown at 37°C in 5% CO_2.

3.7.2 Gene manipulation and naming

A split-marker strategy (Fu et al. 2006) was used to replace specific genomic targets with drug resistance cassettes as in (Haynes et al. 2011) and to incorporate HA tags; New gene names were *CLR*, capsule-linked regulator; *MLR*, melanin- linked regulator; and *CCD*, capsule-correlated DNA-binding protein.

3.7.3 Phenotyping

Growth *in vitro* was assessed by cell counts, melanization by colony color on L- DOPA agar, and shed capsule polysaccharide by the Cryptococcal Antigen Latex Agglutination System

(CALAS®, Meridian Bioscience, Cincinnati, Ohio). To assess capsule thickness, duplicate cultures of cells grown for 24 h in inducing conditions were washed and mixed 3:1 (v/v) with India ink for imaging. The cell wall and capsule edge of each cell were manually annotated (≥10 images per culture) using custom software, and the capsule thickness (outer capsule edge diameter minus cell wall diameter) of mutants relative to WT cells grown in parallel was calculated. Only significant differences ($p<10^{-7}$) of more than 2.5 pixels were reported as altered capsule thickness.

3.7.4 RNA Isolation, RNA-seq, and ChIP-seq

RNA was isolated by standard methods from ≥3 biological replicates for each strain grown for 90 min in capsule-inducing conditions. Libraries for RNA-seq were prepared as in (Haynes et al. 2011), barcoded, and pooled in equimolar ratios for multiplex sequencing. ChIP studies were performed as in (Haynes et al. 2011), using WT and HA-tagged strains and comparing samples subjected to immunoprecipitation (IP) to input material and mock precipitated samples.

3.7.5 Animal Studies

All animal studies were reviewed and approved by the Animal Studies Committee of Washington University School of Medicine and conducted according to NIH guidelines. Groups of six 6 week-old female C57Bl/6 mice were inoculated intranasally with 1.25×10^4 cryptococci, and lung CFU were determined at 2 h and 7 d post-infection.

3.7.6 Comparison of phenotype prediction methods

The accuracy of each method was assessed by comparing its predictions to the phenotypes of 50 single-regulator deletion strains that have been analyzed for capsule thickness either by us or in published works (Appendix Table 2.B). Most of these genes were deleted because they were thought likely to have capsule phenotypes – they are not a random sample of all cryptococcal TFs and 32 of them had altered capsule thickness. Differential expression was assessed using standard methods. Capsule size correlation was assessed as in (Haynes et al. 2011). Phenologs were assessed as in (McGary et al. 2010). The PhenoProphet score of a TF for capsule thickness is –log p-value from the hypergeometric test for enrichment of its NetProphet-predicted targets for genes that are known to have capsule thickness phenotypes. Specifically, the PhenoProphet score is the maximum –log p-value over all networks consisting of the top n NetProphet predictions, with n ranging from 500 to 40,000 in increments of 500. To compute the NetProphet and PhenoProphet scores of a TF we did not use any information about the phenotype or expression profile of the corresponding deletion mutant.

3.7.7 Network validation

For Gat201 (CNAG_01551), we used published ChIP data (Chun, Brown, & Madhani, 2011). For Nrg1 (CNAG_05222) and Usv101 (CNAG_05420) we carried out ChIP-seq as above. For each TF, a ChIP-based sequence-specific binding motif was inferred using BioProspector (Liu et al. 2001). Promoter regions were defined as the 1,000 bp upstream of the start codon. A NetProphet- based binding motif was also inferred by inputting NetProphet's target confidence scores for each TF to FIRE (Elemento, Slonim, & Tavazoie, 2007). The motifs of orthologous TFs from *S. cerevisiae* were obtained from ScerTF (Spivak and Stormo 2012). If the motif for the best *S. cerevisiae* match was unknown the next best match was used. A network of interactions that were supported by both ChIP and NetProphet was constructed using the top 10,000 NetProphet predictions.

3.7.8 Transcriptional dynamics of capsule induction

Triplicate cultures of WT cells were sampled for RNA-seq at 0, 1.5, 3, 8, and 24 h after a shift to capsule-inducing conditions. For each gene, a temporal expression signature was constructed from its median expression level at each time point. For each pair of genes the correlation between their temporal signatures was converted to a dissimilarity. Gene clusters were formed by applying hierarchical agglomerative clustering and cutting the resulting dendrogram at the 10-branch level. For each cluster, GO and KEGG functional enrichment analysis were performed, over-represented terms were examined in detail, and relevant terms were selected. The heatmap was created by scaling the expression of each gene to span the range from 0 to 1.

Temporal expression signatures for capsule-involved regulators and capsule-involved non-regulators were clustered separately into four groups each (Fig. 3.7.B, circles and boxes, respectively). One of the groups of non-regulators is not shown, as we had no comment on it. A combined signature was generated for each cluster by taking the median expression level of all genes in the cluster at each time point. If the number of NetProphet-predicted activating (repressing) edges from one regulator group to another was enriched 1.5-fold relative to an even distribution of the activating (repressing) edges among regulator groups then the corresponding activating (repressing) edge was shown in Figure 3.7.B. The analogous calculations were made for edges from regulator groups to non- regulator groups.

3.7.9 Data access

All generated RNA-seq and ChIP-seq data have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE60398. A software package that implements PhenoProphet is available at http://mblab.wustl.edu/.

Chapter 4:

Transcriptome Engineering Promotes a Fermentative Transcriptional State

4.1 Abstract

The rational manipulation of transcriptomes offers the possibility to engineer the cell as a collective unit toward specified goals, revolutionizing medicine and bioengineering. Progress in transcriptome engineering has primarily consisted of experimental approaches that are iterative, slow, and expensive. We have developed a novel algorithm, NetSurgeon, which utilizes genome-wide gene regulatory networks to identify interventions that will force a cell toward a desired expression state. Following extensive *in silico* validation, we applied NetSurgeon to *S. cerevisiae* biofuel production, generating interventions designed to promote a fermentative state during xylose catabolism. Our selected interventions successfully promoted a fermentative transcriptional state in the absence of glucose and generated strains with 120% higher xylose import rates, improved xylose integration into central carbon metabolism by 303%, and increased ethanol production rates by 31%. We conclude by presenting an integrated model of transcriptional regulation and metabolic flux that will enable metabolic engineering efforts to prioritize functional regulators of central carbon metabolism.

4.2 Background

The central promise of regulatory systems biology is that a map of the cell's global connectivity will enable us to understand, predict, and rationally manipulate cellular behavior. The manipulation of cellular state has many promising applications, including stem cell biology and regenerative medicine, biofuel production, and gene therapy. Fundamental progress toward the goal of cellular state control has been advanced via systems biology - the study of cellular behavior as a complete unit, and synthetic biology - a rapidly advancing discipline which aims to design regulatory and effector molecules with defined behaviors. In systems biology, immense resources have been invested in genome sequencing, systematic deletion collections, and massively parallelized data acquisition, leading to network maps and improved understanding of the cell as a complete system (Gerstein, et al., 2012). However, relatively little research has focused on using these network models for the prediction and manipulation of cellular behavior (Chuang, Hofree, & Ideker, 2010). Synthetic biology has focused on creating molecular components that can be placed into a system to modify the transcriptional state of a small number of genes. However, genome-scale regulatory engineering is still rare, with most systems restricted to a small number of regulators and a limited set of controlled targets (Cameron, Bashor, & Collins, 2014). Bridging the gap between these two disciplines, we demonstrate that the integration of functional transcriptional network mapping, gene expression profiling, and computational modeling can be used to rationally engineer cellular state.

4.3 Related Work

Transcriptome engineering focuses on the manipulation of extant cellular networks and regulatory systems to enforce a state associated with a desired cellular phenotype. The use of native cellular regulatory mechanisms and network models enables the investigator to access evolutionarily optimized states and avoid the extensive iteration often associated with the integration of a synthetic regulatory circuit into a host system (Cardinale & Arkin, 2012) (Litcofsky, Afeyan, Krom, Khalil, & Collins, 2012). The majority of transcriptome engineering thus far has taken place within the context of developmental stem cell engineering, with the generation of induced pluripotency being the best example (Takahashi & Yamanaka, 2006). Since the development of induced pluripotent stem cells, many transcriptional interventions have been identified that move cells at various developmental stages along a specified lineage (Morris & Daley, 2013). However, current strategies for direct lineage conversion are often unable to fully convert cells to the state of the goal cell fate (Feng et al. 2008; Marro et al. 2011; Morris et al. 2014).

The CellNet algorithm was developed in response to current deficiencies in cellular engineering. CellNet is a network-guided algorithm for determining how completely an engineered cell recapitulates a target cell state and identifying transcriptional interventions to guide further engineering (Cahan et al. 2014). CellNet identifies sub-networks within mouse and human cell-type-specific regulatory networks whose expression state is predictive of the cell type. These predictive sub-networks are used as features for classifying novel gene expression profiles according to the cell type they most resemble. In addition, CellNet selects TF interventions for transcriptome engineering by computing a Network Influence Score for each TF which is the sum of two components: the dysregulation of the regulator weighted by its expression level and the dysregulation of its targets weighted by their expression levels. This approach for target selection intervention was used to guide B cell to macrophage conversion by knocking down B cell regulators (Morris et al. 2014). The generalizability of this method remains unclear due to the limited number of interventions and evaluations performed. These are exciting demonstrations of the power of transcriptome engineering, but studies in these complex developmental systems are limited by incomplete transcriptional network maps, complex cell culture requirements, and a lack of quantitative phenotypes directly linked to molecular effectors. These issues have thus far prevented a quantitative assessment of transcriptome engineering efforts.

4.4 Approach

In order to quantitatively assess the current state of transcriptome engineering and establish benchmarks, we utilized *S. cerevisiae* as a model system. 196 of the 209 transcription factors (TF) with an annotated DNA-binding domain in the *S. cerevisiae* genome possess a known DNA binding specificity (Spivak & Stormo, 2012) (Weirauch, et al., 2014) and the genome-wide effect of TF removal on expression has been quantitated through microarray profiling (Hu, Killion, & Iyer, 2007) (Kemmeren, et al., 2014). These data provide us with the ability to generate an accurate network model and to validate our algorithmic approaches. The simplicity of *S. cerevisiae* culture enables quantitative modeling and assessment, with the input/output metabolic function measurable by HPLC and the transcriptional state of the cell quantitated by RNA sequencing.

We identified the industrially relevant fermentation of the pentose carbohydrate xylose as a prototype application that met all our criteria for the quantitative assessment of transcriptome engineering. Xylose is a component of hemicellulose, a polymer that represents approximately 23% of lignocellulosic biomass and is not efficiently fermented by *S. cerevisiae* into ethanol (Chandel & Singh, 2011). Biochemical research has identified all enzymes required for the integration of xylose into the cell’s central carbon metabolism. However, recombinant yeast strains expressing these enzymes, and grown in mixed glucose/xylose cultures, rapidly ferment all available glucose and then undergo a diauxic shift into a respiratory metabolic state. Salusjarvi et al demonstrated through transcriptional and proteomic analysis that cells grown on xylose exist in a hybrid fermentative/respiratory state (Salusjärvi, et al., 2008). The abundance of systems-level data, known metabolic pathways, clear regulatory constraint and quantitative phenotypes enabled us to utilize the transcriptome engineering of xylose metabolism to evaluate the current state of transcriptome engineering.

In this work we present a novel algorithm, NetSurgeon, designed to enable transcriptome engineering. We ran this algorithm over genome-wide gene regulatory networks (GRN) generated using NetProphet (Haynes et al., 2013), and assessed its performance at selecting TFs whose deletion or overexpression will move the transcriptional state of the cell toward a desired goal. Following algorithmic validation, we applied the algorithm to engineer a fermentative xylose transcriptional state and assessed global cellular response to our transcriptional interventions by using analytical chemistry. Our results demonstrate that transcriptome engineering can be efficiently guided using network models and reveal the degree of transcriptional control over a quantitative multi-factorial phenotype.

Our transcriptome engineering method, NetSurgeon, simulates interventions on a transcriptional network model to prioritize those that are likely to move the transcriptional state towards a goal state. Our transcriptome engineering efforts consisted of three steps. First, a map of the network of direct, functional regulation is built (Fig. 4.1.A). Second, starting and goal transcriptional states are defined and our algorithm for prioritizing TF interventions, NetSurgeon, searches through the all possible interventions (deletion or overexpression) to identify interventions that are likely to move the transcriptional state towards the goal state (Fig. 4.1.B).

Finally, strains containing the predicted best interventions are created and RNA-seq and HPLC are used to quantitatively assay their transcriptional and metabolic state (Fig. 4.1.C).

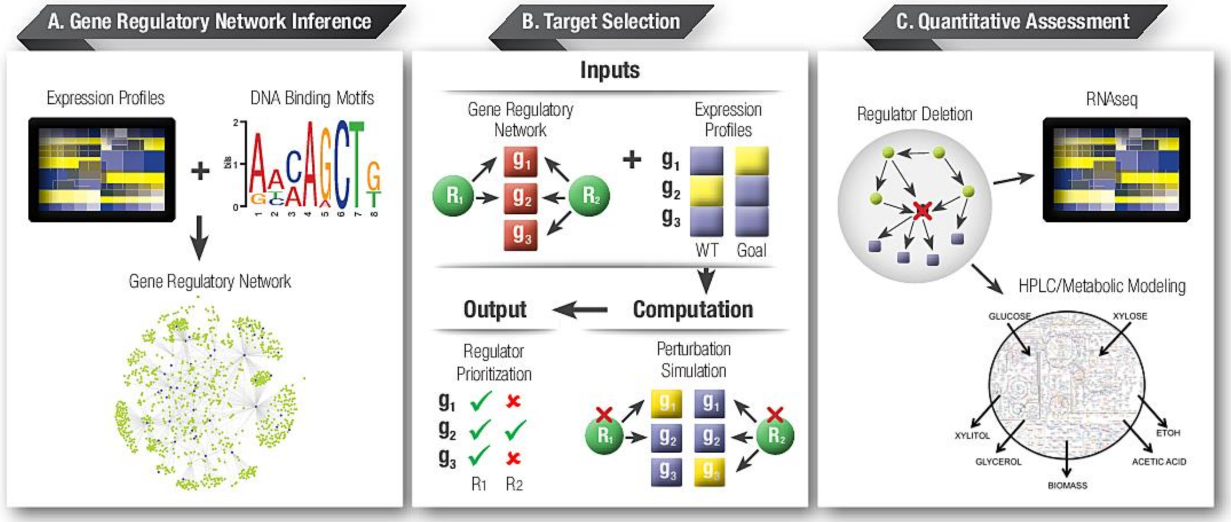


Figure 4.1. Overview of the computational and experimental approaches for rational control of transcriptional state. Panel A: Approach for generation of a gene regulatory network model from DNA binding specificity information and gene expression profiling. Panel B: Approach for target selection through intervention simulation and regulator prioritization. Panel C: Approach for quantitative assessment of intervention effect via RNA sequencing and HPLC metabolite profiling and modeling.

We built an integrated gene regulatory network map by building and combining separate functional and physical maps. The functional map was itself constructed by combining maps inferred by NetProphet from three large expression datasets (Chua, et al., 2006) (Gasch, et al., 2000) (Hu, Killion, & Iyer, 2007). NetProphet is a state-of-the-art GRN mapping algorithm that combines a differential expression (DE) analysis and a co-expression analysis. The physical map was built using a combination of TF binding information from both chromatin immunoprecipitation (ChIP) implicated TF target interactions (Abdulrehman, et al., 2010) (Balaji, Babu, Iyer, Luscombe, & Aravind, 2006) (Harbison, et al., 2004) (Lee, et al., 2002) and TF binding potential estimated by scanning a collection of position weight matrix (PWM) models over all yeast promoters (Spivak & Stormo, 2012). An integrated functional and physical GRN map was built by assigning a score to each TF-target gene pair that was equal to the geometric mean of the scores assigned to it in the functional and physical networks. The geometric mean ensures that high scoring interactions are supported by both binding and expression evidence.

To select interventions that will shift transcriptional state toward the goal state, we applied NetSurgeon. This algorithm assigns a score to each possible intervention representing its confidence that the intervention will yield a shift toward the goal state. The score assigned to each intervention is based on the number of targets of the regulator that are predicted to move toward the goal state and degree to which the initial and goal states differ for the regulator and targets. Deletion of a TF is predicted to increase expression targets it represses and decrease expression of targets it activates. Conversely, overexpression of a TF is predicted to decrease expression targets that the TF represses it and increase expression targets it activates. High-scoring interventions are those that are predicted to change many genes in the right direction, with greater weight given to targets that are the most significantly differentially expressed genes between the initial state and goal state.

4.5 Results

4.5.1 Network models can efficiently guide transcriptome engineering efforts

To assess the ability of NetSurgeon to select interventions that will move the initial transcriptional state toward the goal state, we used NetSurgeon to select regulator interventions for single regulator intervention goal states from publically available gene expression datasets. We choose to use independent single regulator intervention expression profiles for validation goal states, rather than randomly generated expression states, because randomly generated expression states may not be biologically achievable. We constructed the GRN used for validation in a similar fashion as previously described, except the functional network was inferred from only one of the three gene expression datasets previously used, the dataset consisting of 269 regulator deletions strains grown in YPD (Hu, Killion, & Iyer, 2007).

We initially examined NetSurgeon's performance on goal states that we knew could be achieved by a single TF deletion mutant growing in synthetic complete medium (SC). This medium was different from the rich medium (YPD) in which the expression profiles used to build the network were obtained, but the two media featured the same sugar: 2% glucose. The goal states were an independent set of expression profiles from regulator deletion mutants (Kemmeren, et al., 2014). For each of the 245 goal states, NetSurgeon used the NetProphet+PWM network to assign scores to all 320 possible regulator deletions. We plotted the number of goal states for which NetSurgeon ranked the best intervention (the one that actually produced the goal state profile) at or above each rank (Fig 4.2.A, green). We compared this to a random assignment of rankings for each of the deletion goal states, by running NetSurgeon on 100 random networks of the same topology (Fig. 4.2.A, gray). We found that NetSurgeon is able to assign higher scores to the correct interventions compared with ranks assigned by running NetSurgeon over randomly generated networks (Mann-Whitney U test P < 10-46). Further, we observed that NetSurgeon performed at random chance levels using the permuted networks, indicating network structural accuracy is critical for NetSurgeon performance. We also assessed the ability of NetSurgeon to identify the best intervention within the top 5 scoring interventions, a reasonable number of interventions to test experimentally. NetSurgeon ranks the best intervention in the top 5 for 91 goal states, which is 29-times better than random networks scores (P<10-165).

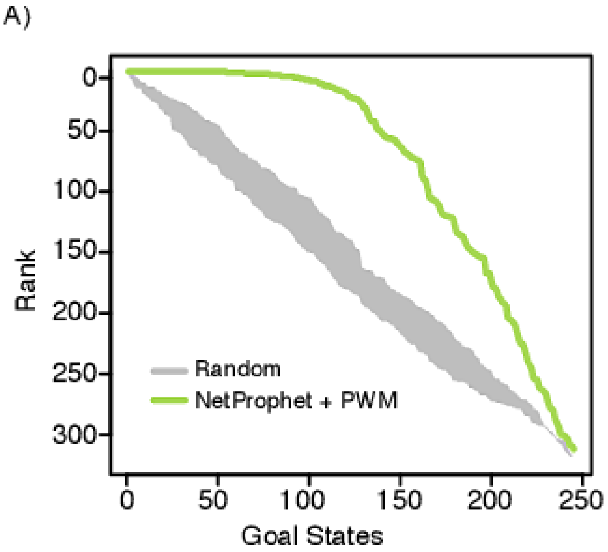


Figure 4.2.A. *In silico* assessment of NetSurgeon using 245 deletion mutant expression profiles grown in synthetic complete medium. Plotted curves show the number of goal states for which NetSurgeon ranked the best intervention at or above each rank (green), compared with random ranks (gray).

We also evaluated the ability of NetSurgeon to indentify interventions in cells cultured in conditions even further from those used to construct the GRN. The goal states consisted of 63 expression profiles obtained from regulator overexpression strains grown in selective synthetic medium supplemented with 2% galactose (Chua, et al., 2006). We assessed the scores assigned to the best regulator for each overexpression goal state and compared the outcome to scores generated using random networks (Fig. 4.2.B). We found that NetSurgeon is able to assign higher scores to the correct interventions compared with random network generated scores (Mann-Whitney U test P < 10-6). NetSurgeon is also able to assign the best intervention a top 5 rank for 8 of the 63 goal states (13%), a 10 fold improvement over the mean of the random network

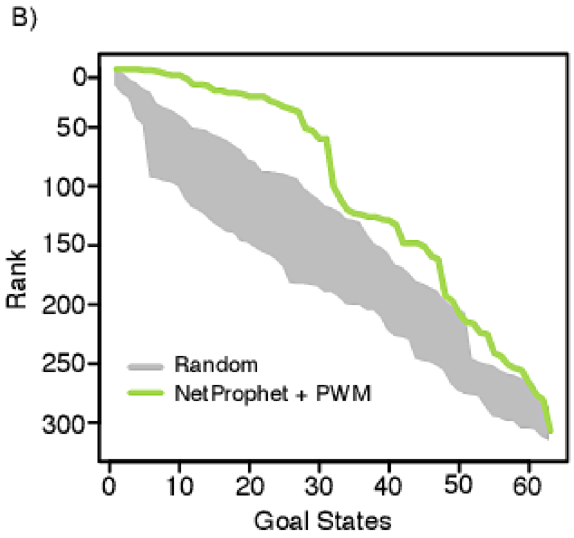


Figure 4.2.B. *In silico* assessment of NetSurgeon using 63 overexpression strains grown in grown in selective synthetic medium supplemented with 2% galactose. Plotted curves show the number of goal states for which NetSurgeon ranked the best intervention at or above each rank (green), compared with random ranks (gray).

In order to evaluate the effect of network accuracy on NetSurgeon performance, we applied NetSurgeon to GRNs inferred from the same expression data sets by CLR (Faith, et al., 2007), regression (Bonneau, et al., 2006), NetProphet (Haynes, et al., 2013), and NetProphet integrated with PWM scores. We first evaluated the structural accuracy of the five GRNs by determining the level of ChIP support for high confidence interactions in each GRN. We then evaluated the performance of NetSurgeon when using each of these five GRNs on our two test data sets: the TF-deletion in SC glucose and TF-overexpression in SC galactose. We plotted the structural accuracy of each of the five GRNs against the NetSurgeon's accuracy when using that GRN (Fig. 4.2.C). We observed a clear pattern of improved NetSurgeon performance with more structurally accurate GRNs. A level-log regression model was fit to test this observation (Multiple R2 = 0.853, P=0.00014) and forecasted a maximum NetSurgeon intervention recovery of 0.85 AUC with a perfect network model, which is a 33% improvement over current NetSurgeon performance.

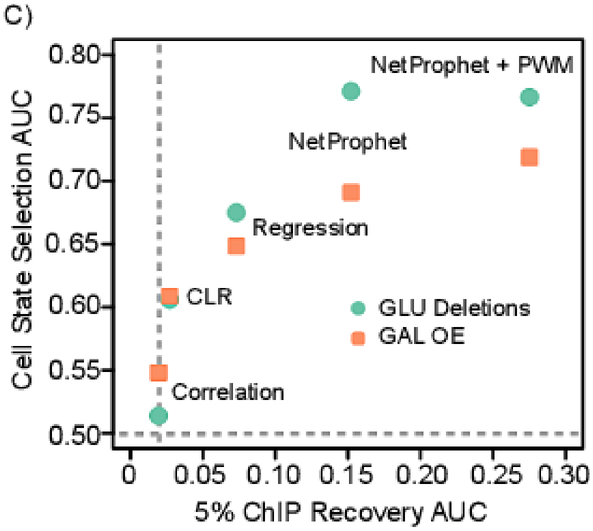


Figure 4.2.C. *In silico* assessment of the effect of network structural accuracy on NetSurgeon target intervention selection accuracy. Network structural accuracy of five GRNs, summarized by area under the precision recall curve at 5% ChIP recovery (x-axis), is compared with NetSurgeon intervention target selection accuracy, summarized by area under the curve of the number of goal states for which NetSurgeon ranked the best intervention at or above each rank (y-axis). Gray dotted lines indicate chance 5% ChIP recovery AUC and cell state selection AUC.

To assess the practicality of NetSurgeon-guided engineering, we ran NetSurgeon on the NetProphet+PWM network and computed the median number of interventions needed to identify the first, the best, and all deletion genotypes that reduce the distance between the wild-type cells and the goal by at least 10% (Fig. 4.2.D). A median of 12, 22 and 51 mutant strains were required to recover the first, the best, and all interventions (10-, 7-, and 4-fold better than random, respectively).

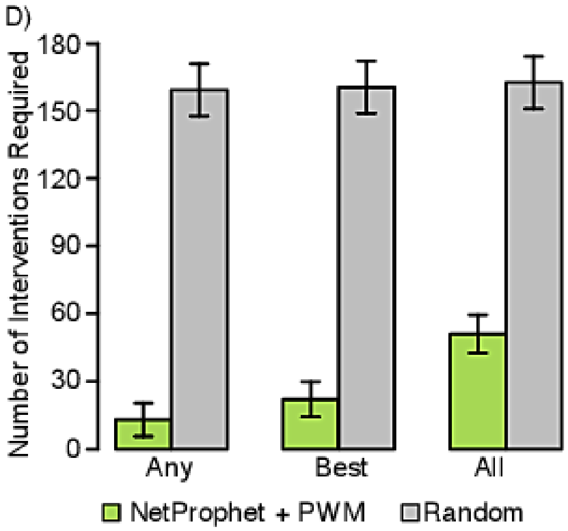


Figure 4.2.D. *In silico* assessment of the median number of NetSurgeon interventions required to generate any strain, the best strain, or all strains, that will converge expression state at least 10% towards the goal state (green), compared with random ranking (gray).

4.5.2 Application of transcriptome engineering to biofuel production

Following the successful in silico validation of our approach for transcriptome engineering, we applied the algorithm to the industrially relevant problem of ethanol production in a mixed glucose-xylose co-culture. Principle components analysis of RNA-seq data from *S. cerevisiae* cells grown with xylose as the sole carbon source indicated that the system was in a hybrid transcriptional state with some characteristics of cells grown in 2% glucose, a fermentative state, and some characteristics of cells grown in 1.3% ethanol, a respiratory state (Fig. 4.3.A). As *S. cerevisiae* cells do not natively consume xylose, we hypothesized that the system was unable to recognize the pentose carbohydrate as a fermentable carbon source and therefore entered into a transcriptional state that was non-optimal for fermentative metabolism.

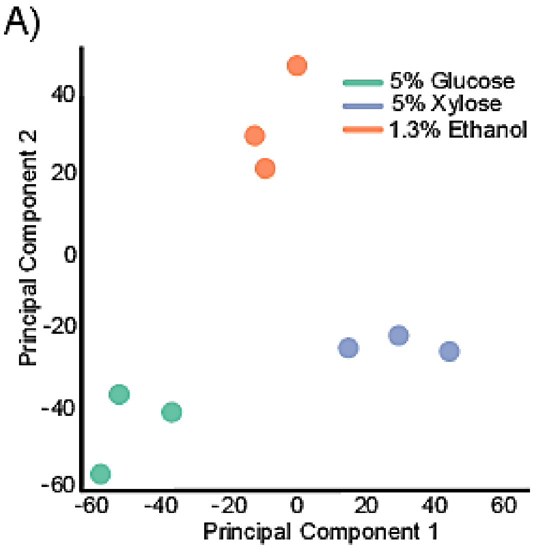


Figure 4.3.A. Principal component analysis of RNA expression profiles reveals a state transition between cells grown on 5% glucose (green), 5% xylose (blue), and 1.3% ethanol (red).

We therefore sought to identify interventions that would shift the system from the xylose-only transcriptional state (origin state) to the high-glucose state (goal state). In order to apply NetSurgeon to this problem, we generated the integrative, genome-wide network map described above. Using this map, NetSurgeon produced a rank-ordered list of regulators whose deletion was predicted to force the system toward the 2% glucose transcriptional state. From this rank ordered list, we selected the top eight predicted interventions for biological validation via PCR-mediated genetic deletion of the selected regulators in the H2217-7 yeast strain (Table 4.1). In order to assess the combinatoric effect of the predicted deletions, we generated an additional three strains carrying deletions in two of the NetSurgeon-selected regulators (*cat8/hap4*, *cat8/adr1*, *cat8/aft2*). For a limited comparison of our algorithmically selected deletions with expert intuition, we deleted the master regulator *SNF1*, the yeast ortholog of AMP kinase and a critical regulator responsible for glucose repression and other features of fermentative metabolism.

WILD-TYPE	SINGLE KO	DOUBLE KO
H2217-7	*snf1*	*cat8/adr1*
	adr1	*cat8/hap4*
	cat8	*cat8/aft2*
	usv1	
	gis1	
	msn2	
	hap4	
	msn4	
	aft2	

Table 4.1. Wild-type and deletion mutant strains profiled

We found that the NetSurgeon-selected targets were supported by existing literature. Cat8 and Hap4 are respiratory factors active in the general cellular response to xylose and deletion of *HAP4* was recently shown to improve cellobiose consumption rates (Salusjärvi, et al., 2008) (Lin, et al., 2014). *MSN2* and *MSN4*, encoding stress associated factors, were observed to be highly upregulated in xylose and their transcriptional targets misregulated (Matsushika, Goshima, Hoshino, & others, 2014). Usv1, Gis1 and Aft2 were all found to have clear roles in the yeast transcriptional response to non-fermentable carbon sources and general stress response (Hlynialuk, Schierholtz, Vernooy, & der, 2008) (Pedruzzi, Bürckert, Egger, & Virgilio, 2000) (Blaiseau, Lesuisse, & Camadro, 2001).

Aerobic batch fermentations were used to assess the outcome of our transcriptome interventions at the transcriptional and metabolic levels. Cells were inoculated into synthetic complete medium supplemented with 2% glucose and 5% xylose at an OD600 of 1.0+/- 0.2 and grown for 48 hours. Samples were taken for RNA-sequencing at 4 hours and 24 hours, representative of the glucose-xylose and xylose-only metabolic states. Aliquots were acquired for

HPLC metabolite analysis across the 48 hour fermentation (Fig. 4.3.B). Using this data, we examined the NetSurgeon's ability to control transcriptional state and quantitatively assessed the effect of transcriptome transcriptional state change on a complex phenotype.

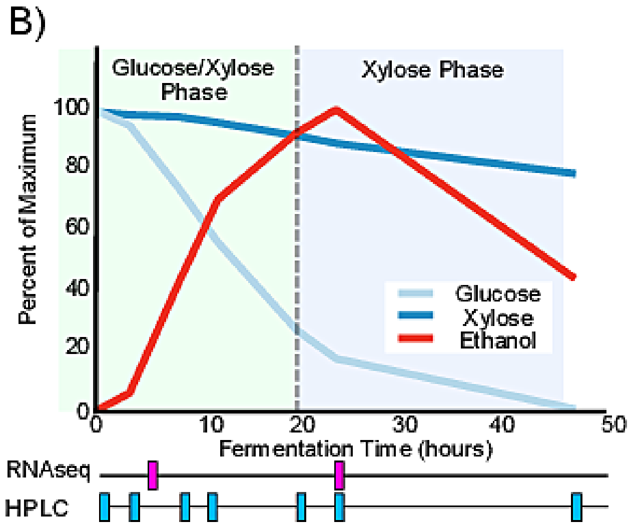


Figure 4.3.B. Top: Glucose (light blue), xylose (dark blue), and ethanol (red) metabolite concentration profiles from the fermentation of the wild-type H2217-7. Bottom: Overview of RNA-seq (magenta) and HPLC (turquoise) sampling strategy for aerobic batch fermentations used in this study.

4.5.3 Transcriptome engineering successfully promotes a fermentative state:

Differential expression analysis revealed that 2,887 genes are differentially expressed by at least two fold in wild-type cells as a result of glucose depletion (42% of all genes. Fig. 4.4.A). Six of the eight NetSurgeon-selected interventions lowered the number of differentially expressed genes. The cat8 mutant was the best, preventing the change in expression of 1,182 of 2,887 DE genes while creating only 526 new DE genes, for a net reduction of 656 DE genes. Notably, the deletion of *CAT8* reduced differential expression better than the deletion of *SNF1*, a master regulator of the *S. cerevisiae* glucose repression system.

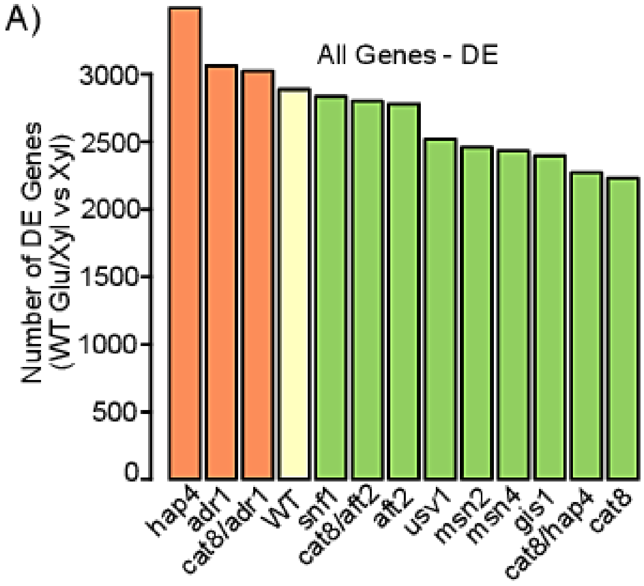


Figure 4.4.A. Number of 2-fold or greater differentially expressed genes between the wild-type strain in the fermentative state and each strain in the respiratory state. Green and red bars indicate strains with less and more differentially expressed genes than wild-type respectively.

Next, we calculated the Euclidean distance between the wild-type expression state in the glucose-xylose phase and the deletion strain's expression state in the xylose-only phase (Fig. 4.4.B). Six of the eight NetSurgeon interventions lowered the Euclidean distance between the two phases. The single deletion mutant *cat8* reduced the genome-wide expression distance between the glucose-xylose phase and the xylose-only phase by 28.4%. The mean reduction in Euclidean distance of the six successful NetSurgeon selected interventions was 20.8%. As in the DE analysis, the deletion of *HAP4* and *ADR1* increased the total distance between the two state vectors.

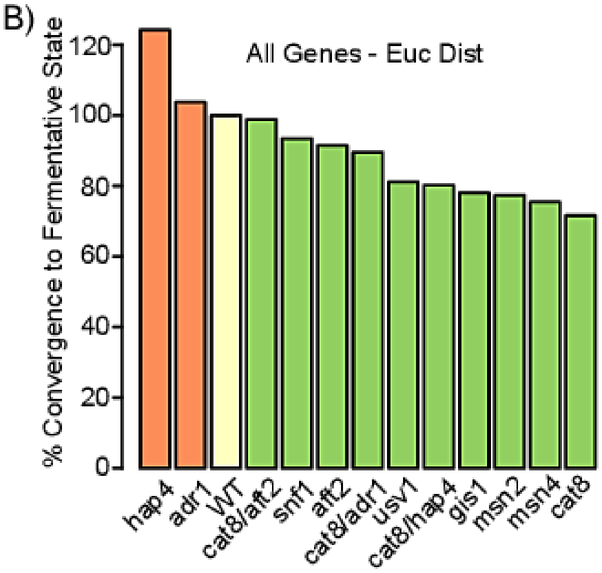


Figure 4.4.B. Euclidean distance between the full expression profile of wild-type strain in the fermentative state, and the full expression profiles of all strains in the respiratory state. Green and red indicate reduced and increased Euclidean distance compared with wild-type respectively.

The NetSurgeon-selected interventions were specifically targeted at optimizing the expression state of 445 genes involved in carbon metabolism. Among these 445 genes, the *cat8* deletion mutant reduced the Euclidean distance by 36% (Fig. 4C). On average, the six successful NetSurgeon-selected interventions reduced the Euclidean distance between the two state vectors by 24%.

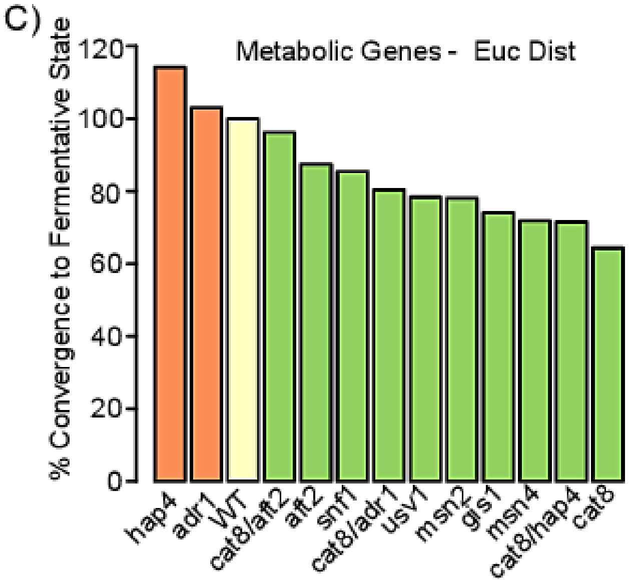


Figure 4.4.C. Euclidean distance between the expression 445 metabolically active computationally optimized genes in the fermentative state of the wild-type strain, and the matching optimized gene expression profiles of all strains in the respiratory state. Green and red indicate reduced and increased Euclidean distance compared with wild-type respectively.

Each of the eight NetSurgeon selected transcription factors had known roles in the regulation of the cellular stress response or respiratory processes. We evaluated the ability of each transcription factor to promote a fermentative state across specific metabolic pathways (Fig 4.4.D). With the exception of *adr1*, each deletion mutant affected the expression of genes across many of the metabolic pathways in central carbon metabolism. Seven of the eight NetSurgeon-selected interventions lowered the Euclidean distance in at least one of the central carbon metabolism pathways evaluated. All six of the interventions that reduced differential expression and global Euclidean distance moved the expression of glycolytic genes toward a fermentative state.

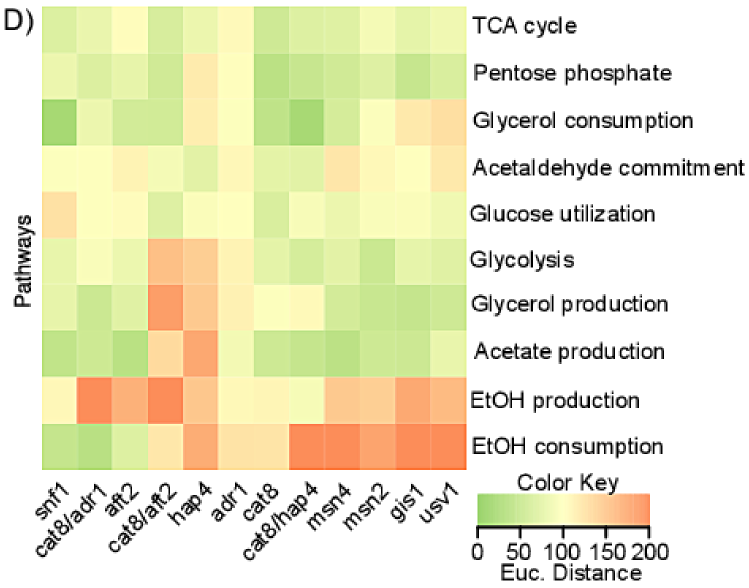


Figure 4.4.D. Euclidean distance between the expression profiles of central carbon metabolic pathways in the fermentative state of the wild-type strain, and the matching central carbon metabolic pathway expression profiles of all strains in the respiratory state. Green and red indicate reduced and increased Euclidean distance compared with wild-type respectively.

Three of these interventions shifted the expression of TCA cycle genes toward a fermentative state. Deletion of *CAT8* promoted a fermentative state in many metabolic pathways essential for xylose fermentation, including genes involved in glucose utilization, the pentose phosphate pathway, glycolysis, the TCA cycle, and acetate/glycerol production. All of the TCA cycle genes were moved toward the expression level associated with fermentative metabolism (Fig 4.4.E). Deletion of *CAT8* also reduced the Euclidean distance of all TCA genes from the fermentative state by 60%. These observations highlight the power of transcriptome level interventions to modulate the expression of many more genes than is feasible by traditional, one-gene-at-a-time genetic engineering.

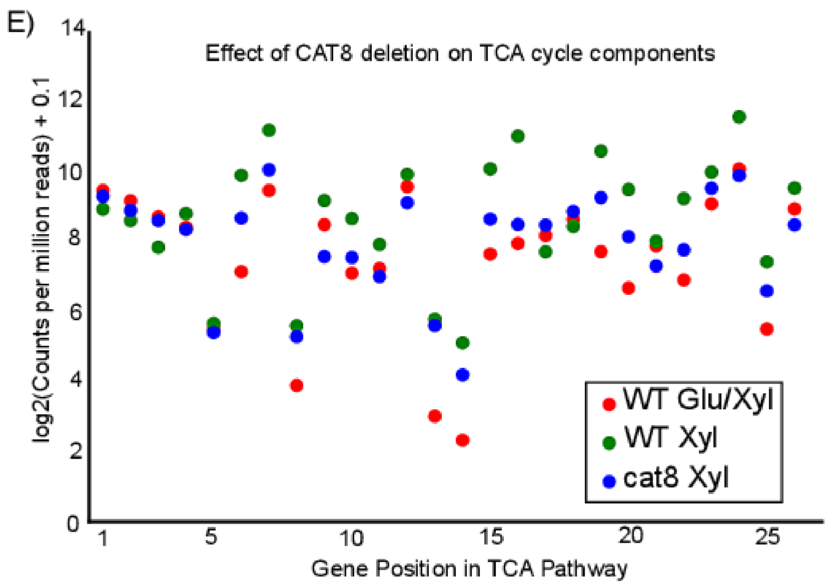


Figure 4.4.E. Comparison of the expression TCA cycle genes between the fermentative state of the wild-type strain (red), the respiratory state of the wild-type strain (green), and the respiratory state of the cat8 deletion mutant strain (blue).

4.5.4 Identification of transcriptional states associated with improved fermentation

In order to assess the change in cellular metabolic behavior following our transcriptional interventions, we profiled metabolic intake and output via HPLC. HPLC analysis identified three metabolic states associated with high glucose, low glucose, and respiratory metabolic phases. We focused our downstream analyses on the high glucose and respiratory phases of the fermentation during which we had carried out RNA-seq.

To examine the ability of the selected transcriptional interventions to control the metabolic state of the cell, we calculated the percentage of input carbon that end up in each of the major carbon fates in each phase (Fig 4.5.A, 4.5.B). Carbon import rates significantly declined in the absence of high glucose, with a mean reduction of import across all assayed genotypes by 86%. In addition to changes in import rate, the cells significantly upregulated their commitment of carbon to respiratory processes in the xylose-only phase. Carbon commitment to respiration changed from a mean of 24% in the glucose-xylose phase to 89% in the xylose-only phase. This indicated that the metabolism of all strains had shifted into a respiratory mode during the xylose-only phase (Fig. 4.5.B).

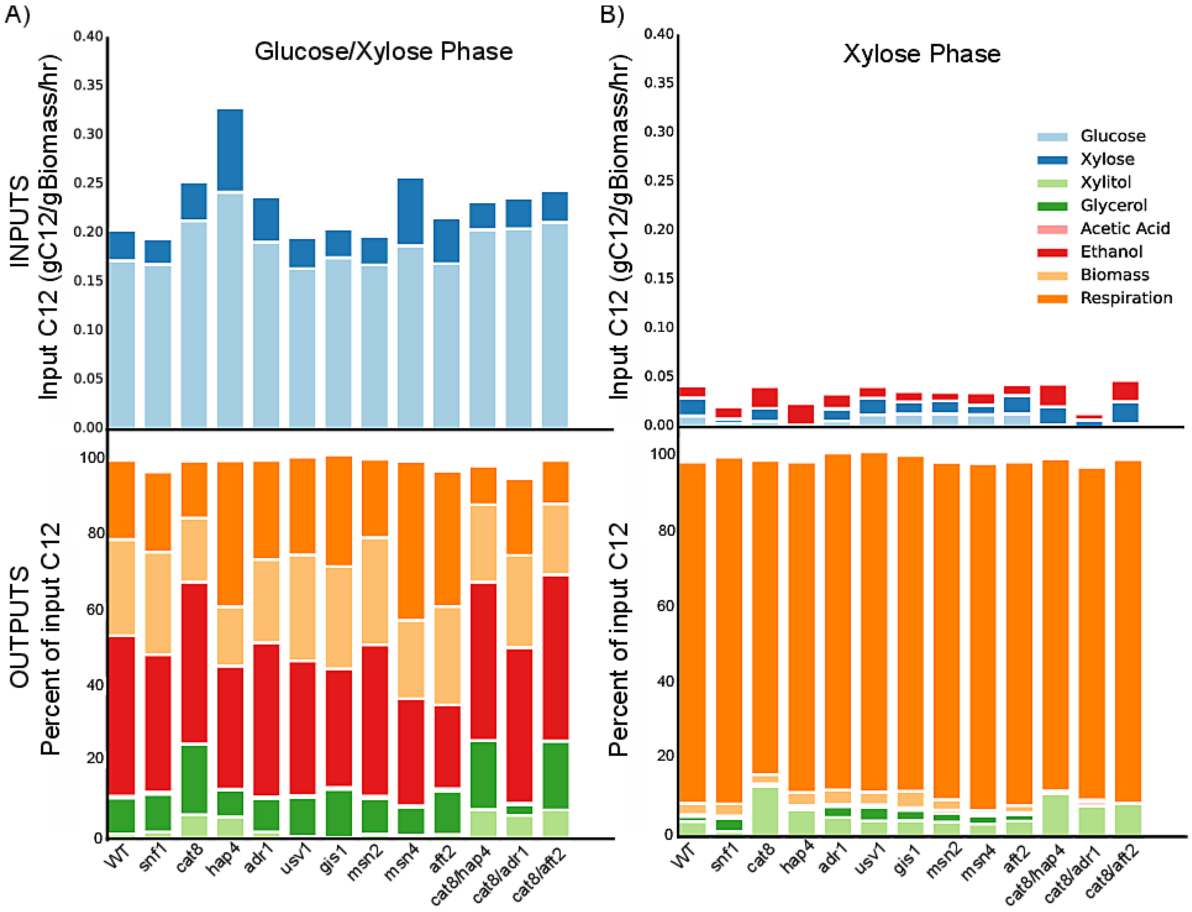


Figure 4.5.A-B. Transcriptome interventions alter carbon intake rates, but do not prevent a transition to a respiratory metabolism. Panel A: Cellular metabolic input and output across profiled genotypes during the glucose-xylose phase of aerobic fermentation. Panel B: Cellular metabolic input and output within the xylose phase.

Although the tested interventions did not prevent the transition to a respiratory metabolic state, they did affect the cell’s commitment of carbon to output metabolites significantly. We observed 41 statistically significant changes in carbon commitment across the 13 profiled genotypes ($p<0.05$, t-test, Benjamini-Hochberg corrected). 28 of these changes were within the glucose-xylose phase. Carbon commitment to all of the profiled metabolites and phenotypes was altered in at least one of our transcriptional interventions, indicating that changes in transcriptional state have the power to impact all dimensions of cellular metabolism. Carbon commitment to xylitol was significantly increased in transcriptional interventions associated with respiratory processes, a potential side effect of the respiratory factors modulating the ratio of the Xyl1, Xyl2, and Xks1 enzymes required for xylose integration into central carbon metabolism. Interestingly, all significant changes in carbon commitment to ethanol and biomass were reductions. The deletion of *SNF1*, *HAP4*, *USV1*, *GIS1*, *MSN4* and *AFT1* significantly reduced carbon commitment to ethanol, with a mean reduction in carbon flux by 26%. Deletions involving *CAT8* or *HAP4* significantly reduced carbon commitment to biomass by 33% and 38%, respectively, in the glucose-xylose phase of the fermentation.

We also observed 57 statistically significant changes in the specific rates of metabolite production or consumption across the glucose-xylose and xylose-only phases of the fermentation ($p < 0.05$, t-test, Benjamini-Hochberg corrected). Within the glucose-xylose phase, we identified industrially relevant changes in glucose and xylose consumption rates, acetic acid output and ethanol production. All of the profiled interventions on respiratory regulators (*cat8*, *hap4*, *adr1*) improved the specific rate of glucose consumption between 11% and 40% (Fig. 4.5.C). We found that *hap4* and *msn4* mutants improved the specific rate of xylose consumption by 170% and 120% respectively (Fig. 4.5.D). Acetic acid, a fermentation byproduct demonstrated to inhibit glycolysis, was also produced at 53%-83% lower specific rates in the *hap4* and *cat8* mutants (Fig. 4.5.E) (Pampulha & Loureiro-Dias, 1990). Importantly, the specific rate of ethanol production was significantly increased by 22% and 31% in the cat8 and hap4 mutants (Fig 4.5.F). Within the set of stress associated factors, we found that the deletion of *USV1*, *MSN2*, *MSN4* and *AFT2* significantly reduced the specific rate of ethanol production, with a mean rate reduction of 22% (Fig 4.5.F). Taken together, these data demonstrate the ability of transcriptome engineering to generate significant changes in cellular behavior, even in the absence of complete phenotypic conversion.

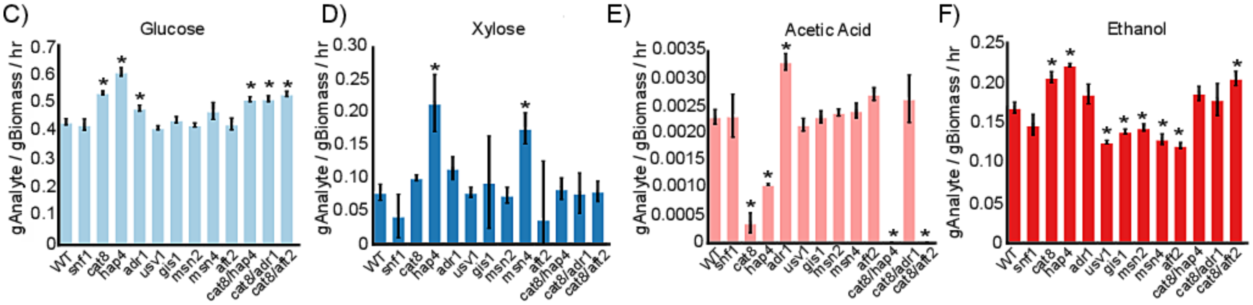


Figure 4.5.C-F. Transcriptome interventions alter specific rates of metabolite production or consumption in the glucose-xylose phase. Panel C: Specific rate of glucose consumption. Panel D: Specific rate of xylose consumption. Panel E: Specific rate of acetic acid production. Panel F: Specific rate of ethanol production.

4.5.5 An integrated model of transcriptional regulation and metabolic flux

The lack of data linking transcriptional state with metabolic phenotypes has prevented the use of transcriptional interventions for effective engineering of metabolism. In order to address this issue, we utilized our dataset to construct an integrated model of *S. cerevisiae* central carbon metabolic flux and expression. We identified regulators linked to flux by correlating their expression with pathway carbon flux. From this set of regulator-flux correlations, we identified regulators putatively controlling metabolic flux outcomes via network-predicted direct regulatory relationships (Fig. 4.6).

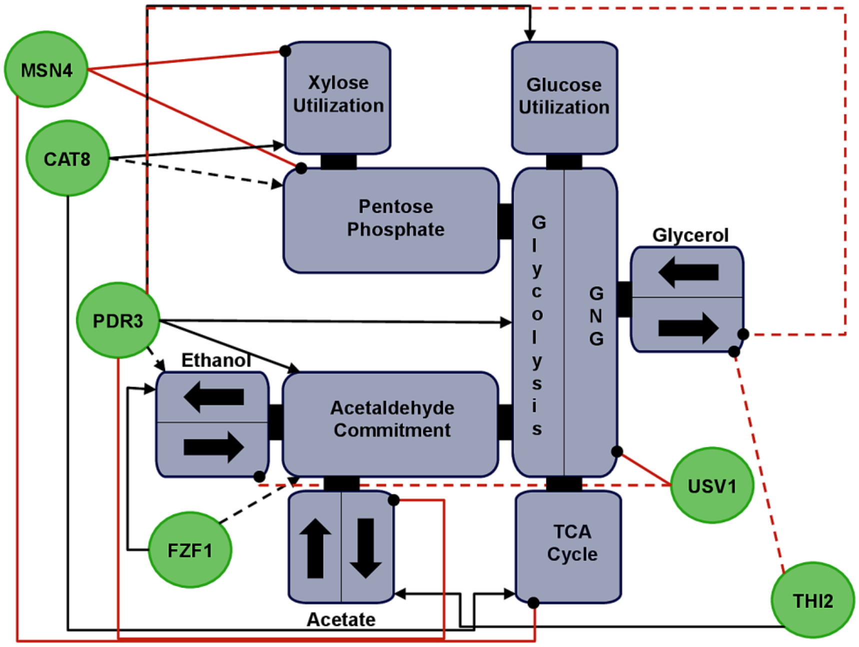


Figure 4.6. An integrated map relating transcription factors to central carbon metabolism flux. Blue rounded rectangles: pathways in central carbon metabolism. Green ovals: transcription factors. Links between transcription factors and pathways denotes transcription factor expression correlation with increased (black arrow headed link) or decreased (red circle headed link) flux through the pathway. Solid link lines: transcription factor directly regulates the expression of genes in the pathway.

This analysis revealed that three transcriptional regulators were deeply interconnected with biochemical pathways important for xylose metabolism and fermentation. *CAT8* expression was correlated with genes associated with xylose utilization, the pentose phosphate pathway, acetate production and the TCA cycle. Msn4 was predicted to directly regulate genes involved in xylose utilization, the pentose phosphate pathway, and the TCA cycle, and flux through these pathways was anti-correlated with *MSN4* expression. Pdr3 was revealed to be a regulator of glycolytic genes, and flux through these pathways was positively correlated with *PDR3* expression. This integrated model of transcriptional regulation and metabolic flux is an important step toward the rational engineering of *S. cerevisiae* metabolism.

4.6 Discussion

We have demonstrated that transcriptional network maps can be used to rationally manipulate cellular state by identifying the crucial regulators mediating a state transition and prioritizing them for genetic intervention. The formalization of this process of rational state manipulation is expected to enable future developments in personalized medicine, improve approaches to stem cell engineering, and reduce the costs associated with these efforts. Our work establishes quantitative benchmarks in this new field, enabling the rapid progress generally associated with clear benchmarks (Stolovitzky, Monroe, & Califano, 2007).

The availability of deletion and overexpression collections in *S. cerevisiae* has enabled us to assess the state of the art in network-guided transcriptome engineering. We found that NetSurgeon can identify the best intervention within a median of 22 guesses, a 7-fold improvement over random guessing. We observed that network amps built from data on one environmental condition can be successfully used to predict interventions in different conditions. This is important for applications that deviate from standard environmental conditions. Finally, we have demonstrated the utility of TF-network maps enriched with direct regulatory relationships; maps generated by NetProphet together with PWM models led to selections that were substantially better than those made by using maps expression correlation or CLR.

We applied NetSurgeon to optimizing yeast for ethanol production from glucose-xylose co-culture. NetSurgeon selected critical regulators highlighted in the literature and six of the eight promoted a fermentative transcriptional state. Although the single deletions were insufficient to entirely prevent a state transition involving 43% of the yeast genome, it succeeded in significantly changing the rate and ratio of cellular carbon commitment. We found that regulators associated with respiratory processes had significant metabolic effects in the fermentative phase of the culture. We also found that deletion of transcription factors controlling stress response lowers the rate of production and the total ethanol yield. In addition, our dataset of 8,055 metabolic measurements with 73 matched RNA sequencing profiles across 14 genotypes will enable future engineering efforts to identify and rationally manipulate the critical regulators of metabolic flux in order to maximize biofuel production.

One of the advantages of transcriptome engineering is the possibility of accessing evolutionarily optimized states associated with specific phenotypes. The expression levels of genes within linear metabolic pathways such as glycolysis and the TCA cycle are highly regulated in order to maintain a correct ratio of enzyme products necessary for avoiding intermediate metabolite accumulation and allosteric inhibition of upstream processes. The engineering of optimal expression levels across entire pathways is a challenging problem that is often addressed through iterative selection strategies (Wang, et al., 2009). We observed that manipulation of regulator expression levels is a promising strategy to access pre-defined expression states across entire pathways. The effect of *CAT8* deletion on TCA gene expression is one example of an interventions reconfiguring the expression of an entire pathway toward a fermentative state. The TCA cycle within *S. cerevisiae* consists of twenty-six genes, making optimization of this pathway's expression level a difficult task through one-gene-at-a-time engineering. Cat8 was predicted by NetProphet to regulate four genes within the TCA cycle and the glyoxylate pathway, and removal of this factor was predicted to move the TCA cycle toward a fermentative expression configuration. We found that *CAT8* deletion moved all twenty-six genes of the TCA cycle toward a fermentative state, providing evidence that naturally evolved transcriptional states can be leveraged for transcriptome engineering.

Our analysis of the double deletion strains highlighted the complexity of epistatic effects within gene regulatory networks. Although the generation of strains with multiple regulatory perturbations offers the possibility of large scale reconfiguration of cellular state, we observed that the three double deletion strains failed to reduce differential expression and Euclidean distance as much as their component single deletions. This non-additivity between genotypes indicates that a more sophisticated approach to modeling the effect of multiple regulator perturbations will be required to expand target selection approaches multiple perturbations.

4.7 Methods

4.7.1 Network guided target selection

To rank possible regulator interventions for convergence towards a goal expression state, NetSurgeon uses a GRN to simulate interventions for all regulators, and for each simulated regulator intervention a score is assigned representing the confidence that the regulator intervention will converge the expression state towards the goal state. The score for a simulated regulator intervention is based on the enrichment of the regulator's simulated intervention effects to fix the total dysregulation of all genes between the initial and goal expression states, where the total dysregulation of all genes is quantified by the sum of the negative log pvals of significance of differential expression. Specifically the NetSurgeon network intervention score for a regulator is:

NetSurgeon network intervention score (R_i) = max($-\log_{10}$ (hypergeometric distribution(X_{ij} * (W/D), W, U-W, (X_{ij}+Y_{ij}) * (W/D) + C_{ij} - Z_{ij}))) for network cutoff j = 500, …, 40,000

where U is the total number of genes in the network, W is the number of dysregulated genes, D is total amount of dysregulation, X_{ij} is the total amount of dysregulation that the intervention of regulator R_i will remove when considering only the top j interactions in the network, Y_{ij} is the total amount of dysregulation that the intervention of regulator R_i will make worse when considering only the top j interactions in the network, C_{ij} is the total number of genes regulated by regulator R_i when considering only the top j interactions in the network, and Z_{ij} is the total number of dysregulated genes regulated by regulator R_i when considering only the top j interactions in the network.

4.7.2 Strain engineering

The xylose metabolizing strain VTT-C-99318 (CEN.PK2-1D ura3::XYL1 XYL2 his3::XKS1 kanMX) was acquired from Salusjarvi et al. and used as the base strain for all experiments in this study (Salusjärvi, et al., 2008). The At5g17010 xylose transporter from A. thaliana was transformed into the VTT-C-99318 strain and maintained through the use of dropout media (Hector, Qureshi, Hughes, & Cotta, 2008). The genetic deletion of algorithmically selected transcription factors was accomplished through PCR amplification and targeting of drug cassettes to the selected ORF via the addition of 45 base pairs of homologous sequence to the 5'/3' amplifying oligos (Baudin, Ozier-Kalogeropoulos, Denouel, Lacroute, & Cullin, 1993). Prior to use in experimentation, all strains were freshly plated onto selectable media from frozen stocks.

4.7.3 *S. cerevisiae* fermentations

All *S. cerevisiae* strains were grown aerobically in 60 mL of synthetic complete at 30℃ in 250 mL baffled erlenmeyer culture flask shaken at 225 RPM. Cultures for identification of differential expression associated with carbon sources were grown in triplicate in either 50 g/L glucose, 50 g/L xylose or 1.3 g/L ethanol for 8 hours prior to collection of biomass for RNA sequencing. Culture of cells for evaluation of the impact of transcriptional interventions were performed in triplicate and initiated by inoculating 1.0+/- 0.2 OD600 units of biomass into 60 mL of synthetic complete media supplemented with 20 g/L glucose and 50 g/L xylose. Samples taken for RNA-seq analysis were aliquoted from the primary culture, spun down at 3000xg and frozen in liquid nitrogen prior to downstream analysis. The supernatant of samples for HPLC was collected by centrifugation of culture samples at 12,000xg for 3 minutes prior to snap freezing for storage in a dry ice/ethanol bath. All samples were stored at -80℃. At least two independent experiments of three biological replicates was performed for each genotype evaluated by HPLC. Cellular density was quantitated through analysis of culture turbidity at 600 nm.

4.7.4 Metabolite analysis

The concentration of input and output cellular metabolites was analyzed using HPLC. Supernatant solutions were stored at -80℃ and filtered through the use of 0.22 um syringe prior to HPLC analysis. Metabolites were eluted from an Aminex HPX-87H column maintained at 65℃ and peaks detected by refractive index. Identified peaks were quantified through integration and interpolated against serial dilutions of standards for glucose, xylose, xylitol, glycerol, acetic acid and ethanol. Analysis of HPLC data was performed on a per biological replicate basis, with metabolic input/output relationships quantified across each fermentation and pooled into a single distribution based on genotype. Turbidity measurements were converted into units of g/biomass based on the turbidity to biomass conversion factor published (Hector, Qureshi, Hughes, & Cotta, 2008). Calculations of analyte rate and specific rate of change were performed across steady states identified in the ethanol dimension. In order to evaluate internal carbon flux, a system of linear equations was developed to describe central carbon metabolism in *S. cerevisiae*.

The system of equations was fit to experimentally measured parameters of carbon import and export for each genotype across the glucose/xylose and xylose-only phases of each fermentation.

4.7.5 RNA sequencing and analysis

Total mRNA was isolated using the yeast RiboPure kit (Life Technologies, Carlsbad CA). Libraries for RNA-Seq were prepared as in (Haynes, et al., 2011). Briefly, poly(A) RNA was selected from the total RNA isolated as above using the mRNA Catcher Plus Kit (Life Technologies) with an epMotion 5075 automated pipettor (Eppendorf). The poly(A) RNA was subsequently sheared by incubating in TURBO DNA-free buffer at 75°C for 10 minutes and purified with the QIAquick PCR Purification Kit (Qiagen). First strand cDNA synthesis was performed using random hexameric primers and SuperScript III Reverse Transcriptase, followed by treatment with E. coli DNA ligase, DNA polymerase I, and RNase H for second-strand synthesis, all using standard methods. The cDNA libraries were end-repaired with a Quick Blunting kit and A-tailed using Klenow exo- with dATP (New England Biolabs). Illumina adapters were ligated to the cDNA and fragments ranging from 150-250 bp in size were selected using gel electrophoresis. The libraries were enriched and indexed in a 10-cycle PCR using Phusion Hot Start II High-Fidelity DNA (Fermentas), purified, and pooled in equimolar ratios for multiplex sequencing on an Illumina HiSeq 2500.

4.7.6 RNA/metabolic data integration and analysis

We utilized two different complementary methods for integrating RNA expression profiles and metabolic data in order to gain a better understanding of the molecular mechanisms controlling metabolic phenotypes. First, we used the limma software package (Ritchie, et al., 2015) to identify differentially expressed genes within the fermentative and respiratory states between the wild-type strain and each deletion strain. We then used this differential expression analysis to putatively link genes mechanistically to metabolic analyte outcomes by identifying differentially expressed genes in metabolic pathways linked to each metabolic analyte.

In addition to differential expression analysis, we also identified genes linked to metabolic outcomes by identifying genes whose expression significantly correlates with carbon flux. For each gene and each metabolic pathway we computed the Pearson correlation coefficient between the gene’s expression profile, and the computed carbon flux through the pathway. We then generated a null distribution of correlation coefficients between gene expression and pathway flux by randomly generating 10,000 expression vectors by sampling per condition from the expression of all genes within the condition with replacement. These null distributions of expression correlation with pathway flux were then used to assign false discovery rate corrected p-values to the significance of each gene’s expression correlation with flux measurements of a metabolic pathway.

Chapter 5:

Discussion

5.1 Conclusion

The central promise of Systems Biology is that a map of the cell's global connectivity will lead to the ability to understand, predict and manipulate cellular behavior in a rational fashion. With the advent of high throughput experimental approaches to assay gene expression state, and subsequent computational methods to infer gene regulatory network structure over the past 15 yeas, our ability to measure and model cellular decision making has improved greatly. However, to truly reach the goal of rational engineering of cellular state, transcriptional network models must have the capacity to predict expression and physiological phenotype state in novel conditions. This dissertation examined and applied strategies to utilize the predictive power of gene network models to guide experimental and engineering efforts.

In chapter 2 we identified the need for improved causative regulatory network models, and then integrated expression based functional interaction evidence and TF binding based physical interaction evidence, in order to utilize the full predictive power of gene network models. These causative network models are vital to understanding the information flow through the network, which allows for expression prediction and cellular engineering. In response, we developed a method to enrich the most confident edges of an inferred network model with edges that are supported by functional and physical evidence. This method combines the strengths of multiple approaches to network building by iterating between expression based network building, and de novo inference of TF binding specificity using the network interactions and protein homology.

We applied this approach to infer the *Saccharomyces cerevisiae* and *Cryptococcus neoformans* regulatory networks and TF binding motifs and, in *Cryptococcus neoformans*, we were able to identify 18 TF binding motifs, of which 15 are novel.

In Chapter 3, we inferred a model of the regulatory network in *Cryptococcus neoformans* controlling the fungal pathogen's capsule and we utilized this network to generate physiological phenotype predictions which guided further experimentation. The integrated analysis pipeline presented in this chapter demonstrates the power of regulatory networks to predict phenotypes and guide experimental efforts; this approach allowed us to identify 16 novel regulators controlling the capsule, which is a substantial improvement from the 11 previously known regulators. Also, we used the final network model to gain a better understanding of cryptococcal capsule biology through analysis of regulation of the enzymes and transporters responsible for biosynthesis of capsule sugars and modeling the dynamics of capsule induction.

In Chapter 4, we investigated the power of causative regulatory networks to guide a real cellular engineering application. To facilitate engineering efforts we presented a novel algorithm, NetSurgeon, which scores simulated overexpression and deletion interventions based on the confidence that each intervention will move the expression state towards a desired goal state. We validated our algorithm through extensive *in silico* testing using existing large publically available expression profiling datasets of deletion and overexpression regulator interventions in *S. cerevisiae*. Then, we applied our NetSurgeon method to engineer *S. cerevisiae* strains with improved biofuel production. NetSurgeon selected 8 TF deletion strains predicted to promote a fermentative transcriptional state, normally occurring in a glucose-rich medium, in an environment containing only the alternative carbon source xylose. We found that 6 of the 8 TF deletion strains successfully moved the transcriptional state of xylose-consuming cells toward a fermentative state. In addition, we observed improved industrially relevant metabolic phenotypes in several of our intervention strains including 120% higher xylose import rates and 31% increased ethanol production rates. Finally, we generated a map linking transcriptional control to metabolic flux through the central carbon metabolic pathway and highlighted several regulators critical for future metabolic engineering efforts.

Taken together, these chapters represent the first thorough examination, systematic application, and quantitative evaluation of the utilization of network models for predicting unobserved expression and phenotype state and guiding biological research. In this work we advance the network biology field by utilizing the predictive power of networks, rather than focusing on network structural accuracy. We believe that in the future quantitative novel genotype expression prediction will become viable with improvements in the accuracy and perturbation simulation of direct and functional regulatory models. In addition, we believe that many more datasets will assay cellular state in multiple dimensions and phenotypes, allowing for more complete cellular modeling and engineering. We are on the cusp of fulfilling the promise of Systems Biology to allow for prediction and manipulation of cellular behavior.

5.2 Future Directions

Chapter 2 presents a method to improve causative network model inference and *de novo* inference of TF binding specificity models by integrating the two tasks. Although this approach was successfully demonstrated, there are several limitations that could be improved with additional research effort. Currently, a relatively simple method for calculating the binding potential score is scanning a PWM over a promoter sequence and adding the significant hits. However, adopting a more biologically motivated strategy would allow for the computed binding potential to better reflect the observed binding in ChIP experiments. DNA sequence conservation is a strong predictor of TF binding sites that could be included in the binding potential scoring by up-weighting significant PWM hits that occur in conserved regions of DNA (Stormo 2013). In addition, experimentally validated binding of TFs has been observed to cluster around transcription start sites. Therefore, a weighting based on the distance of the PWM hit from the transcription start site could also improve the binding potential scoring scheme (Ouyang et al. 2009).

Besides strengthening the scoring scheme used to construct the physical network model, future work will likely improve on the methods used to combine the functional and physical transcriptional network models. In Chapter 2, the score assigned to each interaction in the combined model is the geometric mean of the interaction scores in the expression based functional model and binding based physical model. Unlike arithmetic mean, the geometric mean function requires interactions to be support by both binding and expression evidence in order to be high scoring. However, the geometric mean function does not allow for the consideration of interactions between the functional and physical evidence, which could boost the confidence in the existence of an interaction. If for example, we observe an enrichment of physical binding support for interactions in which a TF is predicted to activate its targets, then we may infer that the TF is an activator, and boost the score of activating interactions. Unfortunately, due to the lack of trustworthy labeled data, it is difficult to synthesize functional and physical regulatory models into a single causative model by applying supervised machine learning methods. There is a rich literature of unsupervised and semi-supervised methods that should be investigated, such as rank aggregation and co-training. In addition to generating better functional and physical combined models, improving and standardizing methods to integrate a variety of interaction evidence into a single network model will become more important as more types of interaction features, like chromatin marks, become widely available.

In Chapter 3, we integrated the often separated processes of computational modeling and experimentation. We did this by selecting the next set of TF deletion and expression profile experiments using a novel method, PhenoProphet, which assigns a score to each TF based on the confidence that the TF regulates a phenotype of interest. Although this method worked well, there is a high time and monetary cost associated with generating a TF deletion strain and expression profiling the strain. Therefore the PhenoProphet method could be improved by adding a complete cost benefit analysis to the scoring of TFs. An improved PhenoProphet should consider the novelty of the TF-phenotype relationship, the knowledge base of the sub-network the TF resides in, the cost associated with generating the deletion, and the amount of deletions that can be generated in total.

Our focus in Chapter 3 is on identifying the key transcription factors that regulate capsule induction. However, we can also use the network to identify the non-regulator end-effector genes that encode the enzymes and transporters necessary for capsule synthesis. In preliminary investigations, we trained a random forest model to classify genes as capsule synthesis end-effectors based on their regulation patterns. This machine learning approach was promising, but investigation of other machine learning classifiers, and the inclusion of additional predictive features, such as temporal expression patterns, could improve classification accuracy.

Conservation is an important sign of functionality in DNA sequence analysis. In recent years, with the increased availability of expression profiles and network inference methods, transcriptional regulatory networks have been inferred for many well-studied organisms. Although transcriptional networks exist for many organisms, these networks are rarely compared to identify conserved interactions among orthologous genes. A promising area of future work is multi-species network conservation and alignment. One potential benefit of network conservation research is strengthening network structural accuracy by modifying the confidence in each inferred interaction based on the level of conservation of the interaction. In addition, networks could be aligned and those alignments could then be compared to identify important similarities and differences related to organism phenotype similarities and differences. Currently it is possible to apply these principles of network conservation to better infer and understand the regulatory network controlling capsule induction in *Cryptococcus neoformans*. In Gene Expression Omnibus there are at least 100 expression profiles of five fungi, including the Ascomycota phylum fungi *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Candida albicans*, and the Basidiomycota phylum fungi *Cryptococcus neoformans* and *Ustilago maydis* (Barrett et al. 2013). Transcriptional networks could be inferred for each of the five fungi, and the network of *Cryptococcus neoformans* could be modified based on interaction conservation. In addition, non-conserved interactions unique to *Cryptococcus neoformans* could be studied in depth due to their potential importance for cryptococcal specific virulence.

In addition to using the network to guide experimentation, in Chapter 3, we also overlay temporal expression patterns on the network hierarchy to generate an initial model of the dynamics of cryptococcal capsule regulation. However, to truly understand the dynamics of capsule regulation, we must observe the cryptococcal expression state over the course of capsule induction and in a variety of different host-like conditions. With expression profiles in a variety of partial and complete capsule inducing conditions, the conditions and molecular responses required to generate capsule could be identified. Also, with large-scale expression profiling over multiple time-points of capsule induction, the transcriptional network models specific to each time-point could be inferred and the time-course of specific processes required to generate capsule could be understood.

The ultimate goal of this thesis was to develop methods which utilized the predictive power of transcriptional regulatory network models to aid in experimental design and cellular engineering efforts. In Chapter 4 we demonstrated that transcriptional network models are capable of making accurate qualitative predictions of the gene expression and physiological phenotype state in novel genotypes, which can be used to guide future research. However, the accuracy of quantitative predictions necessary to truly understand and manipulate cellular decision making is still lacking with current network models.

In uncompleted research work, we have developed a workflow for generating accurate quantitative phenotype predictions for an organism in a novel genotype. We have proposed to first generate a causative network model by integrating direct and functional evidence using the methods presented in Chapter 2. Then we would simulate the novel genotype using the causative network's underlying parameterized mathematical model to generate the predicted expression state. Next, we would take the intersection of two rounds of feature selection to limit the set of explanatory genes used to predict the physiological phenotype. In one round of feature selection, we would test expression prediction accuracy through cross validation on observed expression states, and choose only the genes whose expression can be predicted reliably by the network model. In the second round of feature selection, we would identify gene biomarkers of the phenotype as genes whose expression pattern correlates with the measured levels of the phenotype. After feature selection, we would use the observed gene expression profiles and matching phenotype levels to train a supervised machine learning model, such as a random forest, to link the expression state of the selected genes to physiological phenotype state. Finally, we would generate the predicted phenotype level in the novel genotype by supplying the trained machine learning model with the predicted expression state of the selected genes in the novel genotype. We have attempted to apply this approach to modeling and predicting the response of *Drosophila melanogaster* to high and low sugar feeding, but several problems emerged that limited the accuracy of hemolymph glucose levels, fat body triacylglyceride levels (TAG) and animal weight phenotype predictions.

We found that the gene expression predictions were quite noisy, resulting in only a very small fraction of genes passing both feature selection steps. In an effort to improve the novel genotype expression predictions, we have started to improve the causative network model. Previously, we have observed that better network models can be constructed by combining multiple network models, each inferred from a different expression profiling dataset[CITE HAYNES ET AL 2012]. We will improve our sugar response network model by combining network models inferred using 1,551 expression profiles from 14 large publically available *Drosophila melanogaster* expression datasets. In addition, we will integrate direct regulatory evidence, in the form of binding specificities, for an additional 72 *Drosophila melanogaster* TFs. When integrated into the causative network model, 372 of the 701 modelled TFs will have both direct and functional regulation evidence. Also, we will apply the methods discussed in Chapter 2 to discover the binding specificity of addition TFs.

We also found that further work was required to improve the prediction of physiological phenotypes from gene expression predictions. Due to the expected noise of gene expression predictions it is necessary to utilize noise tolerant prediction methods, even after improvements to the causative network model. One approach to reduce the noise inherent in the independent variable genes used to predict phenotype would be to combine genes into pathways of genes, and use the expression of each pathway, summarized by the mean expression of genes within each pathway, as the independent variables. Another difficulty we encountered when predicting the

Drosophila melanogaster phenotypes was that the gene expression and phenotype states differ greatly between the observed high and control sugar feeding conditions. Unfortunately, neither sugar feeding condition had enough observations of expression and phenotype level to generate a sugar environment specific model. Therefore, a sugar independent model of phenotype level was required, but we encountered difficulties training an accurate model of both conditions. Specifically, due to the large differences between the sugar feeding conditions, error reduction during model training focused on the effect of sugar feeding, rather than the more nuanced effect of genotype. Further exploratory work is required to fix this issue, and viable methods may include normalizing the data to remove the effect of the sugar feeding condition, or utilizing additional feature selection to identify predictors important in both sugar feeding conditions.

References

Abdulrehman, D., Monteiro, P. T., Teixeira, M. C., Mira, N. P., Lourenço, A. B., dos, S. C., . . . others. 2010. YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic acids research* gkq964.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**(17): 3389-3402.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., ... & Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* **324**(5935): 1720-1723.

Bahn YS, Kojima K, Cox GM, Heitman J. 2005. Specialization of the HOG pathway and its impact on differentiation and virulence of *Cryptococcus neoformans*. *Mol Biol Cell* **16**(5): 2285-2300.

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. 2006. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic acids research* **34**(suppl 2): W369-W373.

Balaji, S., Babu, M. M., Iyer, L. M., Luscombe, N. M., & Aravind, L. 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *Journal of molecular biology* **360**(1): 213-227.

Bar-Peled M, Griffith CL, Doering TL. 2001. Functional cloning and characterization of a UDP glucuronic acid decarboxylase: the pathogenic fungus *Cryptococcus neoformans* elucidates UDP-xylose synthesis. *Proc Natl Acad Sci U S A* **98**(21): 12003-12008.

Bar-Peled M, Griffith CL, Ory JJ, Doering TL. 2004. Biosynthesis of UDP-GlcA, a key metabolite for capsular polysaccharide synthesis in the pathogenic fungus *Cryptococcus neoformans*. *Biochem J* **381**(Pt 1): 131-136.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., ... & Soboleva, A. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research* **41**(D1): D991-D995.

Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F., & Cullin, C. 1993. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic acids research* **21**(14): 3329.

Bembom, O., Keles, S., & van der Laan, M. J. 2007. Supervised detection of conserved motifs in DNA sequences with cosmo. *Statistical applications in genetics and molecular biology* **6**(1).

Berger, M. F., & Bulyk, M. L. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature protocols* **4**(3): 393-411.

Blaiseau, P.-L., Lesuisse, E., & Camadro, J.-M. 2001. Aft2p, a novel iron-regulated transcription activator that modulates, with Aft1p, intracellular iron use and resistance to oxidative stress in yeast. *Journal of Biological Chemistry* **276**(36): 34221-34226.

Bonneau, R., Reiss, D. J., Shannon, P., Facciotti, M., Hood, L., Baliga, N. S., & Thorsson, V. 2006. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome biology* **7**(5): R36.

Bryne, J., Valen, E., Tang, M., Marstrand, T., Winther, O., da Piedade, I., . . . Sandelin, A. 2008. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Research* **36**(suppl 1): D102-D106.

Cahan, P., Li, H., Morris, S. A., da Rocha, E. L., Daley, G. Q., & Collins, J. J. 2014. CellNet: network biology applied to stem cell engineering. *Cell* **158**(4): 903-915.

Cameron, D. E., Bashor, C. J., & Collins, J. J. 2014. A brief history of synthetic biology. *Nature Reviews Microbiology* **12**(5): 381-390.

Cardinale, S., & Arkin, A. P. 2012. Contextualizing context for synthetic biology—identifying causes of failure of synthetic biological systems. *Biotechnology journal* **7**(7): 856-866.

Chandel, A. K., & Singh, O. V. 2011. Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of 'Biofuel'. *Applied microbiology and biotechnology* **89**(5): 1289-1303.

Chikamori M, Fukushima K. 2005. A new hexose transporter from *Cryptococcus neoformans*: molecular cloning and structural and functional characterization. *Fungal Genet Biol* **42**(7): 646-655.

Christensen, R. G., Enuameh, M. S., Noyes, M. B., Brodsky, M. H., Wolfe, S. A., & Stormo, G. D. 2012. Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics* **28**(12): i84-i89.

Chua, G., Morris, Q. D., Sopko, R., Robinson, M. D., Ryan, O., Chan, E. T., . . . Hughes, T. R. 2006. Identifying transcription factor functions and targets by phenotypic activation. *Proceedings of the National Academy of Sciences* **103**(32): 12045-12050.

Chuang, H.-Y., Hofree, M., & Ideker, T. 2010. A decade of systems biology. *Annual review of cell and developmental biology* **26**: 721.

Chun CD, Brown JC, Madhani HD. 2011. A major role for capsule-independent phagocytosis-inhibitory mechanisms in mammalian infection by *Cryptococcus neoformans*. *Cell Host Microbe* **9**(3): 243-251.

Coelho C, Bocca AL, Casadevall A. 2014. The tools for virulence of *Cryptococcus neoformans*. *Advances in applied microbiology* **87**: 1-41.

Conway, T. 2003. Microarray expression profiling: capturing a genome-wide portrait of the transcriptome. *Molecular microbiology* **47**(4): 879-889.

Cottrell TR, Griffith CL, Liu H, Nenninger AA, Doering TL. 2007. The pathogenic fungus *Cryptococcus neoformans* expresses two functional GDP-mannose transporters with distinct expression patterns and roles in capsule synthesis. *Eukaryot Cell* **6**(5): 776-785.

Cramer KL, Gerrald QD, Nichols CB, Price MS, Alspaugh JA. 2006. Transcription factor Nrg1 mediates capsule formation, stress response, and pathogenesis in *Cryptococcus neoformans*. *Eukaryot Cell* **5**(7): 1147-1156.

Cryptococcus neoformans var. grubii H99 Sequencing Project, (n.d.) Broad Institute of Harvard and MIT. Available: http://www.broadinstitute.org/.

Das, M. K., & Dai, H. K. 2007. A survey of DNA motif finding algorithms. *BMC bioinformatics* **8**(Suppl 7): S21.

De Smet, R., & Marchal, K. 2010. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* **8**(10): 717-729.

D'Souza CA, Alspaugh JA, Yue C, Harashima T, Cox GM, Perfect JR, Heitman J. 2001. Cyclic AMP-dependent protein kinase controls virulence of the fungal pathogen *Cryptococcus neoformans*. *Mol Cell Biol* **21**(9): 3179-3191.

Eisenman HC, Casadevall A. 2012. Synthesis and assembly of fungal melanin. *Applied microbiology and biotechnology* **93**(3): 931-940.

Elemento O, Tavazoie S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol* **6**(2): R18.

Elemento, O., Slonim, N., & Tavazoie, S. 2007. A universal framework for regulatory element discovery across all genomes and data types. *Molecular cell* **28**(2): 337-350.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., ... & Cherry, J. M. 2014. The reference genome sequence of *saccharomyces cerevisiae*: then and now. *G3: Genes| Genomes| Genetics* **4**(3): 389-398.

Erecinska M, Silver IA. 2001. Tissue oxygen tension and brain sensitivity to hypoxia. *Respir Physiol* **128**(3): 263-276.

Ernst, J., Beg, Q., Kay, K., Balazsi, G., Oltvai, Z., & Bar-Joseph, Z. 2008. A semi-supervised method for predicting transcription factor-gene interactions in *Escherichia coli*. *PLoS Computational Biology* **4**(3): e1000044.

Faith, J. J., Hayete, B., Thaden, J. T., Mogno, I., Wierzbowski, J., Cottarel, G., ... & Gardner, T. S. 2007. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS biology* **5**(1): e8.

Feng, R., Desbordes, S. C., Xie, H., Tillo, E. S., Pixley, F., Stanley, E. R., & Graf, T. 2008. PU. 1 and C/EBPα/β convert fibroblasts into macrophage-like cells. *Proceedings of the National Academy of Sciences* **105**(16): 6057-6062.

Fu J, Hettler E, Wickes BL. 2006. Split marker transformation increases homologous integration frequency in *Cryptococcus neoformans*. *Fungal genetics and biology : FG & B* **43**(3): 200-212.

Gardner, T. S., & Faith, J. J. 2005. Reverse-engineering transcription control networks. *Physics of life reviews* **2**(1): 65-88.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., . . . Brown, P. O. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell* **11**(12): 4241-4257.

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., . . . others. 2012. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**(7414): 91-100.

Gonçalves, D. L., Matsushika, A., Belisa, B., Goshima, T., Bon, E. P., & Stambuk, B. U. 2014. Xylose and xylose/glucose co-fermentation by recombinant *Saccharomyces cerevisiae* strains expressing individual hexose transporters. *Enzyme and microbial technology* **63**: 13-20.

Grant, C. E., Bailey, T. L., & Noble, W. S. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**(7): 1017-1018.

Greenfield, A., Madar, A., Ostrer, H., & Bonneau, R. 2010. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PloS one* **5**(10): e13397.

Gupta, A., Christensen, R. G., Bell, H. A., Goodwin, M., Patel, R. Y., Pandey, M., ... & Stormo, G. D. 2014. An improved predictive recognition model for Cys2-His2 zinc finger proteins. *Nucleic acids research* **42**(8): 4800-4812.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. 2007. Quantifying similarity between motifs. *Genome biology* **8**(2): R24.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**(7004): 99-104.

Haynes, B. C., Maier, E. J., Kramer, M. H., Wang, P. I., Brown, H., & Brent, M. R. 2013. Mapping functional transcription factor networks from gene expression data. *Genome research* **23**(8): 1319-1328.

Haynes, B. C., Skowyra, M. L., Spencer, S. J., Gish, S. R., Williams, M., Held, E. P., Brent, M. R., Doering, T. L. 2011. Toward an integrated model of capsule regulation in *Cryptococcus neoformans*. *PLoS pathogens* **7**(12): e1002411.

Hecker, M., Goertsches, R. H., Engelmann, R., Thiesen, H. J., & Guthke, R. 2009. Integrative modeling of transcriptional regulation in response to antirheumatic therapy. *BMC bioinformatics* **10**(1): 262.

Hector, R. E., Qureshi, N., Hughes, S. R., & Cotta, M. A. 2008. Expression of a heterologous xylose transporter in a *Saccharomyces cerevisiae* strain engineered to utilize xylose improves aerobic xylose consumption. *Applied microbiology and biotechnology* **80**(4): 675-684.

Hedges SB, Blair JE, Venturi ML, Shoe JL. 2004. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC evolutionary biology* **4**: 2.

Hicks JK, Bahn YS, Heitman J. 2005. Pde1 phosphodiesterase modulates cyclic AMP levels through a protein kinase A-mediated negative feedback loop in *Cryptococcus neoformans*. *Eukaryot Cell* **4**(12): 1971-1981.

Hlynialuk, C., Schierholtz, R., Vernooy, A., & der, G. v. 2008. Nsf1/Ypl230w participates in transcriptional activation during non-fermentative growth and in response to salt stress in *Saccharomyces cerevisiae*. *Microbiology* **154**(8): 2482-2491.

Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* **39**(5): 683-687.

Irrthum, A., Wehenkel, L., & Geurts, P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PloS one* **5**(9): e12776.

Janbon G, Ormerod KL, Paulet D, Byrnes EJ, 3rd, Yadav V, Chatterjee G, Mullapudi N, Hon CC, Billmyre RB, Brunel F et al. 2014. Analysis of the genome and transcriptome of *Cryptococcus neoformans var. grubii* reveals complex RNA expression and microevolution leading to virulence attenuation. *PLoS Genet* **10**(4): e1004261.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., ... & Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell* **152**(1): 327-339.

Jung WH, Sham A, White R, Kronstad JW. 2006. Iron regulation of the major virulence factors in the AIDS-associated pathogen *Cryptococcus neoformans*. *PLoS biology* **4**(12): e410.

Kemmeren P, Sameith K, van de Pasch LA, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW et al. 2014. Large- scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**(3): 740-752.

Kitano, H. 2002. Systems biology: a brief overview. *Science* **295**(5560): 1662-1664.

Kronstad JW, Hu G, Choi J. 2011. The cAMP/Protein Kinase A Pathway and Virulence in *Cryptococcus neoformans*. *Mycobiology* **39**(3): 143-150.

Kwon-Chung KJ, Fraser JA, Doering TL, Wang Z, Janbon G, Idnurm A, Bahn YS. 2014. *Cryptococcus neoformans* and *Cryptococcus gattii*, the Etiologic Agents of Cryptococcosis. *Cold Spring Harbor perspectives in medicine* **4**(7).

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**(9): 1813-1831.

Lee H, Chang YC, Varma A, Kwon-Chung KJ. 2009. Regulatory diversity of *TUP1* in *Cryptococcus neoformans*. *Eukaryot Cell* **8**(12): 1901-1908.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., . . . others. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**(5594): 799-804.

Lee, Y., & Zhou, Q. 2013. Co-regulation in embryonic stem cells via context-dependent binding of transcription factors. *Bioinformatics* **29**(17): 2162-2168.

Lin, Y., Chomvong, K., Acosta-Sampson, L., Estrela, R., Galazka, J. M., Kim, S. R., . . . Cate, J. H. 2014. Leveraging transcription factors to speed cellobiose fermentation by *Saccharomyces cerevisiae*. *Biotechnology for biofuels* **7**(1): 126.

Litcofsky, K. D., Afeyan, R. B., Krom, R. J., Khalil, A. S., & Collins, J. J. 2012. Iterative plug-and-play methodology for constructing and modifying synthetic gene networks. *Nature methods* **9**(11): 1077-1080.

Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, Noble SM. 2008. Systematic genetic analysis of virulence in the human fungal pathogen *Cryptococcus neoformans*. *Cell* **135**(1): 174-188.

Liu, X., Brutlag, D. L., & Liu, J. S. 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific symposium on Biocomputing* **6**(2001): 127-138.

Ma H, Hagen F, Stekel DJ, Johnston SA, Sionov E, Falk R, Polacheck I, Boekhout T, May RC. 2009. The fatal fungal outbreak on Vancouver Island is characterized by enhanced intracellular parasitism driven by mitochondrial regulation. *Proc Natl Acad Sci U S A* **106**(31): 12980-12985.

Madar, A., Greenfield, A., Vanden-Eijnden, E., & Bonneau, R. 2010. DREAM3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one* **5**(3): e9803.

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., ... & DREAM5 Consortium. 2012. Wisdom of crowds for robust gene network inference. *Nature methods* **9**(8): 796-804.

Marbach, D., Roy, S., Ay, F., Meyer, P. E., Candeias, R., Kahveci, T., ... & Kellis, M. 2012. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome research* **22**(7): 1334-1349.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., ... & Bryant, S. H. 2012. CDD: conserved domains and protein three-dimensional structure. *Nucleic acids research* gks1243.

Marro, S., Pang, Z. P., Yang, N., Tsai, M. C., Qu, K., Chang, H. Y., ... & Wernig, M. 2011. Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. *Cell stem cell* **9**(4): 374-382.

Matsushika, A., Goshima, T., Hoshino, T., & others. 2014. Transcription analysis of recombinant industrial and laboratory *Saccharomyces cerevisiae* strains reveals the molecular basis for fermentation of glucose and xylose. *Microbial cell factories* **13**(1): 16.

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., . . . Kloos, D. 2003. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Research* **31**(1): 374-378

McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A* **107**(14): 6544-6549.

Morin, R. D., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T. J., ... & Marra, M. A. 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**(1): 81.

Morris, S. A., & Daley, G. Q. 2013. A blueprint for engineering cell fate: current technologies to reprogram cell identity. *Cell research* **23**(1): 33-48.

Morris, S. A., Cahan, P., Li, H., Zhao, A. M., San Roman, A. K., Shivdasani, R. A., ... & Daley, G. Q. 2014. Dissecting engineered cell types and enhancing cell fate conversion via CellNet. *Cell* **158**(4): 889-902.

Moyrand F, Fontaine T, Janbon G. 2007. Systematic capsule gene disruption reveals the central role of galactose metabolism on *Cryptococcus neoformans* virulence. *Mol Microbiol* **64**(3): 771-781.

Moyrand F, Janbon G. 2004. *UGD1*, encoding the *Cryptococcus neoformans* UDP-glucose dehydrogenase, is essential for growth at 37 degrees C and for capsule biosynthesis. *Eukaryot Cell* **3**(6): 1601-1608.

Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., ... & White, K. P. 2011. A cis-regulatory map of the Drosophila genome. *Nature* **471**(7339): 527-531.

Nielsen K, Cox GM, Litvintseva AP, Mylonakis E, Malliaris SD, Benjamin DK, Jr., Giles SS, Mitchell TG, Casadevall A, Perfect JR et al. 2005. *Cryptococcus neoformans* {alpha} strains preferentially disseminate to the central nervous system during coinfection. *Infect Immun* **73**(8): 4922-4933.

O'Meara TR, Alspaugh JA. 2012. The *Cryptococcus neoformans* capsule: a sword and a shield. *Clin Microbiol Rev* **25**(3): 387-408.

O'Meara TR, Xu W, Selvig KM, O'Meara MJ, Mitchell AP, Alspaugh JA. 2014. The *Cryptococcus neoformans* Rim101 transcription factor directly regulates genes required for adaptation to the host. *Mol Cell Biol* **34**(4): 673-684.

Ouyang, Z., Zhou, Q., & Wong, W. H. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences* **106**(51): 21521-21526.

Pampulha, M. E., & Loureiro-Dias, M. C. 1990. Activity of glycolytic enzymes of *Saccharomyces cerevisiae* in the presence of acetic acid. *Applied Microbiology and Biotechnology* **34**(3): 375-380.

Park BJ, Wannemuehler KA, Marston BJ, Govender N, Pappas PG, Chiller TM. 2009. Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. *Aids* **23**(4): 525-530.

Pedruzzi, I., Bürckert, N., Egger, P., & Virgilio, C. D. 2000. *Saccharomyces cerevisiae* Ras/cAMP pathway controls post-diauxic shift element-dependent transcription through the zinc finger protein Gis1. *The EMBO Journal*, **19**(11), 2569-2579.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* gkv007.

Rodrigues ML, Casadevall A, Zaragoza O. 2011. The architecture and antigenic composition of the polysaccharide capsule. In *Cryptococcus, from human pathogen to model yeast*, (ed. J Heitman, TR Kozel, J Kwon-Chung, J Perfect, A Casadevall). ASM Press, Washington, D.C.

Sabiiti W, Robertson E, Beale MA, Johnston SA, Brouwer AE, Loyse A, Jarvis JN, Gilbert AS, Fisher MC, Harrison TS et al. 2014. Efficient phagocytosis and laccase activity affect the outcome of HIV-associated cryptococcosis. *J Clin Invest* **124**(5): 2000-2008.

Salusjärvi, L., Kankainen, M., Soliymani, R., Pitkänen, J.-P., Penttilä, M., & Ruohonen, L. 2008. Regulation of xylose metabolism in recombinant *Saccharomyces cerevisiae*. *Microbial cell factories* **7**(1): 18.

Sedlak, M., & Ho, N. W. 2004. Characterization of the effectiveness of hexose transporters for transporting xylose during glucose and xylose co-fermentation by a recombinant Saccharomyces yeast. *Yeast* **21**(8): 671-684.

Shazman, S., Lee, H., Socol, Y., Mann, R. S., & Honig, B. 2013. OnTheFly: a database of Drosophila melanogaster transcription factors and their binding sites. *Nucleic acids research*, gkt1165.

Simcha, D., Price, N. D., & Geman, D. 2012. The limits of de novo DNA motif discovery. *PloS one* **7**(11): e47836.

Song MH, Lee JW, Kim MS, Yoon JK, White TC, Floyd A, Heitman J, Strain AK, Nielsen JN, Nielsen K et al. 2012. A flucytosine-responsive Mbp1/Swi4-like protein, Mbs1, plays pleiotropic roles in antifungal drug resistance, stress response, and virulence of *Cryptococcus neoformans*. *Eukaryot Cell* **11**(1): 53-67.

Spivak AT, Stormo GD. 2012. ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res* **40**(Database issue): D162-168.

Srikanta D, Santiago-Tirado FH, Doering TL. 2014. *Cryptococcus neoformans*: historical curiosity to modern pathogen. *Yeast* **31**(2): 47-60.

Stolovitzky, G., Monroe, D., & Califano, A. 2007. Dialogue on Reverse-Engineering Assessment and Methods. *Annals of the New York Academy of Sciences* **1115**(1): 1-22.

Stormo, G. 2013. Introduction to protein-DNA interactions: Structure, thermodynamics, and bioinformatics. Cold Spring Harbor Laboratory Press.

Takahashi, K., & Yamanaka, S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**(4): 663-676.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., ... & Sidow, A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods* **5**(9): 829-834.

Verfaillie, A., Imrichová, H., Van de Sande, B., Standaert, L., Christiaens, V., Hulselmans, G., ... & Aerts, S. 2014. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. *PLoS computational biology* **10**(7): e1003731.

Wang ZA, Griffith CL, Skowyra ML, Salinas N, Williams M, Maier EJ, Gish SR, Liu H, Brent MR, Doering TL. 2014. *Cryptococcus neoformans* dual GDP- mannose transporters and their role in biology and virulence. *Eukaryot Cell* **13**(6): 832-842.

Wang, H. H., Isaacs, F. J., Carr, P. A., Sun, Z. Z., Xu, G., Forest, C. R., & Church, G. M. 2009. Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**(7257): 894-898.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., . . . others. 2014. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**(6): 1431-1443.

Zaman S, Lippman SI, Zhao X, Broach JR. 2008. How Saccharomyces responds to nutrients. *Annual review of genetics* **42**: 27-81.

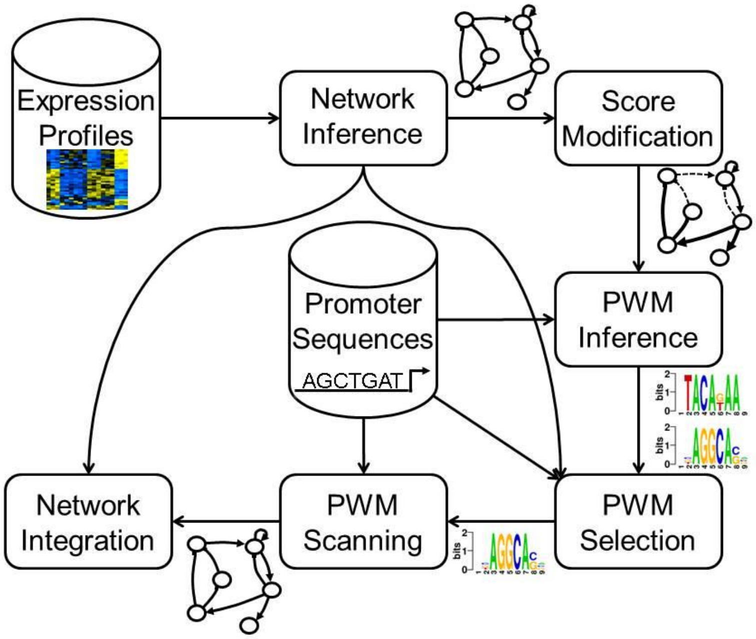
Zaragoza O, Casadevall A. 2004. Experimental modulation of capsule size in *Cryptococcus neoformans*. *Biol Proced Online* **6**: 10-15.

Zhu X, Williamson PR. 2003. A CLC-type chloride channel gene is required for laccase activity and virulence in *Cryptococcus neoformans*. *Molecular microbiology* **50**(4): 1271-1281

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101**(476): 1418-1429.

Appendix

Appendix Figure 1. PWM Inference and Integration Steps.



The steps necessary to infer and integrate PWMs into a network model are outlined above. First, we utilize NetProphet to infer a gene regulatory network model from gene expression profiles. Then we modify the TF-target interaction scores using each TF's regulatory sign and DBD homology (see 2.5.4 Improving DNA Binding Specificity Inference). FIRE is used to infer PWMs for each TF by identifying 7-mers that occur often in the promoter sequences of targets with high modified TF-target interaction scores. Good PWMs are selected by comparing the PWM ranking of target genes to the expression-only network ranking of target genes. Finally, selected PWMs are used to build a binding-based network model which is combined with the expression-based network model.

Appendix Table 1. Mutant strains generated in KN99α.

Gene name	Gene identifier	Selection criterion	Capsule thickness[a]
ADA2	CNAG_01626	Correlation	-15
ARO8001	CNAG_04345	Correlation	1
ASG1	CNAG_03849	Correlation	-2
ASG101	CNAG_03018	Other[b]	NA
BIK1	CNAG_06352	Phenoprophet	-1
CAC1	CNAG_03202	Literature	-29
CCD3	CNAG_00732	Correlation	1
CCD4	CNAG_03279	Correlation	-1
CCD6	CNAG_06252	Correlation	0
CEP3	CNAG_06276	Correlation	3
CIR1	CNAG_04864	Literature	-18
CLR1	CNAG_04353	Phenoprophet	3
CLR2	CNAG_03378	Phenoprophet	0[c]
CLR3	CNAG_00871	Correlation	4
CLR4	CNAG_04908	Correlation	-3
CLR5	CNAG_05067	Correlation	-3
CLR6	CNAG_07797	Other[d]	-4
ECM2201	CNAG_00883	Correlation	-10
FAP1	CNAG_07506	Correlation	-9
FHL1	CNAG_05535	Phenoprophet	-12
FKH101	CNAG_05861	Phenoprophet	3
FKH2	CNAG_02566	Phenoprophet	3
GAT201	CNAG_01551	Literature	-21
HAP2	CNAG_07435	Phenoprophet	0
HAP3	CNAG_02215	Literature	-7
HAP5	CNAG_07680	Literature	-9
HOG1	CNAG_01523	Literature	5
MAL13	CNAG_02774	Correlation	-2
MBS1	CNAG_07464	Phenoprophet	-12
MCM1	CNAG_07924	Other[d]	-3
MLR1	CNAG_00031	Correlation	-2
NRG1	CNAG_05222	Literature	-9
PDR802	CNAG_03894	Phenoprophet	3
PKR1	CNAG_00570	Literature	11
RDS2	CNAG_03902	Correlation	0
RIM101	CNAG_05431	Literature	-16
SSN801	CNAG_00440	Literature[e]	0[c]
SWI6	CNAG_01438	Phenoprophet	-3
TUP1	CNAG_02153	Literature	7
USV101	CNAG_05420	Correlation	3
YRM103	CNAG_04093	Correlation	-1

[a] Difference from wild type, in pixels.

[b] *ASG101* was deleted for an unrelated project; it is included here because its expression profile was included as input for our NetProphet network.

[c] Strains lacking *CLR2* and *SSN801* have normal capsule thickness but increased variance; see text.

[d] *CLR6* and *MCM1* were selected for deletion based on manual inspection of an early version of the network.

[e] *ssn801* cells were previously reported to be hypercapsular, but we find that the distribution of capsule thicknesses for these cells has the same mean, but significantly increased variance, compared to wild type.

Appendix Table 2.A. Capsule-implicated genes not shown in Appendix Table 1 that were used for PhenoProphet analysis.

LOCUS	NAME	YEAST ORF	YEAST NAME	Cn Sc PROTEIN ALIGNMENT EVAL	RATIONALE	CITATION	PMID
CNAG_00124	*CAS32*	YJL010C	NOP9	5E-1	Hypocapsular in Double	Moyrand et al., 2004	15590825
CNAG_00268	*ILV2*	YMR108W	ILV2	0E+0	Hypocapsular	Kingsbury et al.,2004	15133116
CNAG_00375	*GCN5*	YGR252W	GCN5	4E-112	Hypocapsular	O'Meara et al.,2010	20581290
CNAG_00396	*PKA1*	YKL166C	TPK3	5E-125	Hypocapsular	Hicks et al.,2004	14871933
CNAG_00531	*ENA1*	YDR039C	ENA2	0E+0	Hypercapsular	Jung et al., 2012	22343280
CNAG_00600	*CAP60*	YPL058C	PDR12	2E+0	Hypocapsular	Chang and Kwon-Chung,1998; Moyrand and Janbon,2004	9573112, 15590833
CNAG_00623	*EGCrP1*	YIR007W	YIR007W	1E-54	Hypocapsular	Ishibashi et al 2012	22072709
CNAG_00697	*UGE1*	YBR019C	GAL10	6E-83	Hypercapsular	Moyrand et al.,2007	17462022
CNAG_00701	*CAS31*	YGR004W	PEX31	2E+0	Hypocapsular in double	Moyrand et al., 2004	15590825
CNAG_00721	*CAP59*	YBR274W	CHK1	1E+0	Hypocapsular	Chang and Kwon-Chung,1994; Moyrand and Janbon,2004	8007987, 15590833
CNAG_00746	*CAS35*	YHR165C	PRP8	4E+0	Hypocapsular	Moyrand et al.,2004; Moyrand et al.,2007	15590825, 17462022
CNAG_00769	*PBS2*	YJL128C	PBS2	9E-104	Hypercapsular	Bahn et al.,2005	15728721
CNAG_00996	*PMT4*	YJR143C	PMT4	2E-176	Hypocapsular	Willger et al.,2009	19633715
CNAG_01106	*VPH1*	YOR270C	VPH1	0E+0	Hypocapsular	Erickson et al.,2001	11737651

CNAG_01156	*CAP2*	YNL183C	NPR1	1E+0	Homology	Moyrand et al.,2007	17462022
CNAG_01172	*PBX1*	YGR099W	TEL2	1E+0	Hypocapsular	Liu et al.,2007	17337638
CNAG_01371	*CRG2*	YOR301W	RAX1	3E-10	Hypercapsular	Shen et al,2008	18658258
CNAG_01654	*CAS34*	YOR377W	ATF1	2E+0	GXM defect	Moyrand et al.,2007	17462022
CNAG_01678	*NHA1*	YLR138W	NHA1	1E-122	Hypercapsular	Jung et al., 2012	22343280
CNAG_01727	*SSA1*	YER103W	SSA4	0E+0	Hypercapsular	Zhang et al.,2006	17040492
CNAG_01845	*PKC1*	YBL105C	PKC1	2E-128	Hypocapsular	Heung et al.,2005	15946943
CNAG_01890	*MET6*	YER091C	MET6	0E+0	Hypocapsular	Pascon et al.,2004	15347759
CNAG_02029	*WSP1*	YOR181W	LAS17	1E-26	Hypocapsular	Shen et al,2011	21357479
CNAG_02036	*CAS4*	YML038C	YMD8	9E-5	Hypocapsular in double	Moyrand et al.,2007	17462022
CNAG_02236	*PPG1*	YNR032W	PPG1	4E-104	Hypocapsular	Gerik et al.,2005	16194228
CNAG_02581	*CAS33*	YGL173C	KEM1	5E-1	Hypocapsular in double	Moyrand et al. 2004	15590825
CNAG_02702	*CLC-A*	YJR040W	GEF1	3E-122	Hypocapsular	Zhu and Williamson,2003	14622414
CNAG_02797	*CPL1*	YML029W	USA1	1E+0	Hypocapsular	Liu et al.,2008	18854164
CNAG_02885	*CAP64*	YGR191W	HIP1	8E+0	Hypocapsular	Chang et al.,1996; Moyrand and Janbon,2004	8675296, 15590825
CNAG_03120	*AGS1*	YGR292W	MAL12	4E-5	Hypocapsular	Reese et al.,2007	17244196
CNAG_03322	*UXS1*	YBR019C	GAL10	1E-10	Hypocapsular	Moyrand et al.,2002	12139628
CNAG_03426	*GMT2*	YGL225W	VRG4	1E-89	Hypocapsular in double	Cottrell et al., 2007	17351078
CNAG_03438	*HXT1*	YMR011W	HXT2	6E-87	Hypercapsular	Chikamori and Fukushima, 2005	15907385
CNAG_03582	*RIM20*	YOR275C	RIM20	8E-26	Hypocapsular	O'Meara et al.,2010	20174553
CNAG_03644	*CAS3*	YPL164C	MLH3	4E+0	Hypocapsular in double	Moyrand et al. 2004	15590825
CNAG_03670	*IRE1*	YHR079C	IRE1	1E-104	Hypocapsular	Cheon et al. 2011	21852949
CNAG_03735	*CAP4*	YGR286C	BIO2	1E-1	Homology	Moyrand et al. 2007	17462022
CNAG_03818	*SSK1*	YLR006C	SSK1	9E-50	Hypercapsular	Bahn et al,2007	17951522
CNAG_04312	*MAN1*	YER003C	PMI40	2E-52	Hypocapsular	Wills et al.,2001	11359567
CNAG_04320	*CPS1*	YBR023C	CHS3	1E-2	Hypocapsular	Chang et al.,2006	16790766
CNAG_04	*GPA1*	YER020W	GPA2	9E-112	Hypocapsular	Alspaugh et	9389652

505						al.,1997	
CNAG_04730	*GPR4*	YMR172W	HOT1	2E-2	Hypocapsular	Xue et al.,2006	16291861
CNAG_04969	*UGD1*	YDR109C	YDR109C	6E-1	Hypocapsular	Moyrand and Janbon,2004; Griffith et al.,2004	15590833, 15383535
CNAG_05081	*PDE1*	YGL248W	PDE1	1E-5	Hypercapsular	Hicks et al.,2005	16339715
CNAG_05139	*UGT1*	YER077C	YER077C	6E-1	Hypercapsular	Moyrand et al.,2007	17462022
CNAG_05148	*CXT1*	YLR288C	MEC3	4E+0	GXM defect	Moyrand et al.,2007; Klutts et al.2007	17462022 ,17430900
CNAG_05218	*ACA1*	YNL138W	SRV2	6E-75	Hypocapsular	Bahn et al.,2004	15590822
CNAG_05254	*NSTX*	YPL244C	HUT1	6E-41	Hypocapsular	Doering et al. unpublished	
CNAG_05465	*GIB2*	YMR116C	ASC1	1E-92	overexpression suppresses gpa1-capsule defect	Palmer,et al 2006	16950773
CNAG_05562	*PBX2*	YGR125W	YGR125W	2E-1	Hypocapsular	Liu et al.,2007	17337638
CNAG_05563	*HOS2*	YGL194C	HOS2	3E-137	Hypercapsular	Liu et al.,2008	18854164
CNAG_05581	*CHS3*	YBR023C	CHS3	0E+0	Hypercapsular	Baker et al.,2007,Banks et al.,2005	17400891, 16278457
CNAG_05650	*UBP5*	YMR304W	UBP15	8E-160	Hypocapsular	Fang et al.,2012	22719877
CNAG_05703	*LRG1*	YDL240W	LRG1	7E-61	Hypocapsular	Gerik et al.,2005	16194228
CNAG_05721	*MFE2*	YKR009C	FOX2	0E+0	Hypocapsular	Kretschmer et al.,2012	22707485
CNAG_05817	*GMT1*	YGL225W	VRG4	2E-92	Hypocapsular	Cottrell et al.,2007	17351078
CNAG_06016	*CAP6*	YIR021W	MRS1	6E-1	Homology	Moyrand et al.,2007	17462022
CNAG_06301	*SCH9*	YHR205W	SCH9	2E-142	Hypercapsular	Wang et al.,2004	15503029
CNAG_06591	*SET302*	YKR029C	SET3	7E-21	Hypercapsular	Liu et al.,2008	18854164
CNAG_06808	*CPRa*	YKL178C	STE3	2E-24	Hypocapsular	Chang et al.,2003	12933837
CNAG_06813	*CAP1*	YML128C	MSC1	2E+0	Homology	Lengeler et al.,2002	12455690
CNAG_07408	*STE20*	YNL298W	CLA4	1E-113	Hypocapsular	Wang et al.,2002	12455960
CNAG_07470	*PDE2*	YOR360C	PDE2	2E-21	Hypercapsular	Hicks et al.,2005	16339715
CNAG_07554	*CAP10*	YOR172W	YRM1	4E+0	Hypocapsular	Chang and Kwon-Chung,1999; Moyrand and Janbon,2004	10482503, 15590825

CNAG_07636	*CSR2*	YBL061C	SKT5	3E-73	Hypercapsular	Baker et al.,2007,Banks et al.,2005	17400891, 16278457
CNAG_07701	*CTR2*	YLR411W	CTR3	5E-1	Hypocapsular	Chun and Madhani, 2010	20824073
CNAG_07718	*CIN1*	YIR006C	PAN1	2E-33	Hypocapsular	Shen et al,2010	20345666
CNAG_07937	*CAS1*	YGL139W	FLC3	5E+0	O-acetylation defect	Janbon et al.,2001	17462022

Appendix Table 2.B. Regulator genes used for methods analysis in Fig. 3.4, with closest yeast homolog, protein alignment Eval, TF status (DNA-binding TF or not), capsule phenotype, and source for the phenotype; blue shading indicates genes that were not part of our uniform deletion set in KN99.

LOCUS	NAME	YEAST ORF	YEAST NAME	Cn Sc PROTEIN ALIGNMENT EVAL	IS TF	CAPSULE EVIDENCE	SOURCE	PMID
CNAG_00031	*MLR1*	YLR014C	*PPR1*	2E-27	TRUE	Normal	This study	
CNAG_00156	*SP1*	YNL027W	*CRZ1*	2E-23	TRUE	Hypercapsular	Adler et al, 2011	21487010
CNAG_00193	*GAT1*	YFL021W	*GAT1*	2E-21	TRUE	DefectInGXM	Kmetzsch et al. 2011	20673806
CNAG_00440	*SSN801*	YNL025C	*SSN8*	5E-24	FALSE	Hypervariable	Liu et al.,2008; found to be hypervarible in this study	18854164
CNAG_00570	*PKR1*	YIL033C	*BCY1*	5E-61	FALSE	Hypercapsular	D'Souza et al.,2001	11287622
CNAG_00732	*CCD3*	YML076C	*WAR1*	1E-1	TRUE	Normal	This study	
CNAG_00871	*CLR3*	YFL031W	*HAC1*	2E+0	TRUE	Hypercapsular	This study	
CNAG_00883	*ECM2201*	YLR228C	*ECM22*	9E-5	TRUE	Hypocapsular	This study	
CNAG_01242	*HAPX*	YDR259C	*YAP6*	3E-4	TRUE	Normal	Jung et al. 2010	21124817
CNAG_01438	*SWI6*	YLR182W	*SWI6*	1E-40	TRUE	Hypocapsular	This study	
CNAG_01454	*STE12α*	YHR084W	*STE12*	7E-31	TRUE	Hypocapsular	Yue et al. 1999	10581270
CNAG_01523	*HOG1*	YLR113W	*HOG1*	4E-172	FALSE	Hypercapsular	Bahn et al.,2005	15728721
CNAG_0	*GAT201*	YMR136W	*GAT2*	4E-13	TRUE	Hypocapsular	Liu et	18854164

1551							al.,2008	
CNAG_01626	*ADA2*	YDR448W	*ADA2*	2E-74	TRUE	Hypocapsular	Haynes et al. 2011	22174677
CNAG_02153	*TUP1*	YCR084C	*TUP1*	1E-102	FALSE	Hypercapsular	Lee et al,2009	19820119
CNAG_02215	*HAP3*	YBL021C	*HAP3*	5E-41	TRUE	Hypocapsular	Jung et. al 2010	21124817
CNAG_02435	*CWC2/ BWC2*	YMR136W	*GAT2*	2E-11	TRUE	Normal	Idnurm et al. 2005; Lu et al. 2005	15760278, 15813738
CNAG_02566	*FKH2*	YNL068C	*FKH2*	2E-15	TRUE	Hypercapsular	This study	
CNAG_02774	*MAL13*	YKL038W	*RGT1*	6E-8	TRUE	Normal	This study	
CNAG_03202	*CAC1*	YJL005W	*CYR1*	0E+0	FALSE	Hypocapsular	Alspaugh et al.,2002	12455973
CNAG_03279	*CCD4*	YDR213W	*UPC2*	7E-7	TRUE	Normal	This study	
CNAG_03366	*ZNF2*	YNL027W	*CRZ1*	1E-14	TRUE	Normal	Lin et al. 2010	20485569
CNAG_03378	*CLR2*	YLR399C	*BDF1*	1E+0	FALSE	Hypervariable	This study	
CNAG_03409	*SKN7*	YHR206W	*SKN7*	6E-30	TRUE	Normal	Coenjaerts et al. 2006	16696662
CNAG_03849	*ASG1*	YIL130W	*ASG1*	3E-18	TRUE	Normal	This study	
CNAG_03894	*PDR802*	YLR256W	*HAP1*	4E-8	TRUE	Hypercapsular	This study; called normal capsular by Liu et al, 2008	
CNAG_03902	*RDS2*	YPL133C	*RDS2*	6E-49	TRUE	Normal	This study	
CNAG_04093	*YRM103*	YLR014C	*PPR1*	5E-7	TRUE	Normal	This study	
CNAG_04345	*ARO8001*	YDR421W	*ARO80*	6E-6	TRUE	Normal	This study	
CNAG_04353	*CLR1*	YJR127C	*RSF2*	2E-2	TRUE	Hypercapsular	This study	
CNAG_04864	*CIR1*	YJL110C	*GZF3*	8E-18	TRUE	Hypocapsular	Jung et al.,2006	17121456
CNAG_04908	*CLR4*	YGR089W	*NNF2*	9E-1	TRUE	Hypocapsular	This study	
CNAG_05067	*CLR5*	YLR399C	*BDF1*	6E-1	FALSE	Hypocapsular	This study	
CNAG_05222	*NRG1*	YDR043C	*NRG1*	2E-17	TRUE	Hypocapsular	Cramer et al.,2006	16835458
CNAG_05392	*ZAP104*	YJL056C	*ZAP1*	7E-39	TRUE	Normal	Schneider et al. 2012	22916306
CNAG_05420	*USV101*	YPL230W	*USV1*	5E-19	TRUE	Hypercapsular	This study	
CNAG_05431	*RIM101*	YHL027W	*RIM101*	5E-32	TRUE	Hypocapsular	O'Meara et al.,2010	20174553
CNAG_05535	*FHL1*	YPR104C	*FHL1*	4E-16	TRUE	Hypocapsular	This study	

CNAG_05861	*FKH101*	YIL131C	*FKH1*	2E-9	TRUE	Hypercapsular	This study	
CNAG_06134	*BZP1*	YFL031W	*HAC1*	8E-2	TRUE	Normal	Idnurm et al. 2009	19151325
CNAG_06252	*CCD6*	YDR213W	*UPC2*	8E-8	TRUE	Normal	This study	
CNAG_06276	*CEP3*	YMR168C	*CEP3*	1E-6	TRUE	Hypercapsular	This study	
CNAG_06352	*BIK1*	YCL029C	*BIK1*	1E-7	FALSE	Normal	This study	
CNAG_06762	*GAT204*	YMR136W	*GAT2*	3E-6	TRUE	Normal	Chun et al. 2011	21402362
CNAG_07435	*HAP2*	YGL237C	*HAP2*	2E-24	FALSE	Normal	This study	
CNAG_07464	*MBS1*	YDL056W	*MBP1*	7E-36	TRUE	Hypocapsular	Song et al, 2012	22080454
CNAG_07506	*FAP1*	YNL023C	*FAP1*	8E-23	TRUE	Hypocapsular	This study	
CNAG_07680	*HAP5*	YOR358W	*HAP5*	2E-35	FALSE	Hypocapsular	Jung et al,2010	21124817
CNAG_07797	*CLR6*	YKL070W	*YKL070W*	3E-8	FALSE	Hypocapsular	This study	
CNAG_07924	*MCM1*	YMR043W	*MCM1*	6E-34	TRUE	Hypocapsular	This study	