5-24-2012

Practical Approaches to Biological Network Discovery

Brian Haynes
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
Jeremy Buhler
Barak Cohen
Tamara Doering
Gary Stormo
Weixiong Zhang

Practical Approaches to Biological Network Discovery

by

Brian Clifton Haynes

A dissertation presented to the
Graduate School of Art and Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2012
Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Practical Approaches to Biological Network Discovery

by

Brian Clifton Haynes

Doctor of Philosophy in Science in Computer Science

Washington University in St. Louis, 2012

Research Advisor: Professor Michael Brent

This dissertation addresses a current outstanding problem in the field of systems biology, which is to identify the structure of a transcriptional network from high-throughput experimental data. Understanding of the connectivity of a transcriptional network is an important piece of the puzzle, which relates the genotype of an organism to its phenotypes. An overwhelming number of computational approaches have been proposed to perform integrative analyses on large collections of high-throughput gene expression datasets to infer the structure of transcriptional networks. I put forth a methodology by which these tools can be evaluated and compared against one another to better understand their strengths and weaknesses. Next I undertake the task of utilizing high-throughput datasets to learn new and interesting network biology in the pathogenic fungus *Cryptococcus neoformans*. Finally I propose a novel computational method for mapping out transcriptional networks that unifies two orthogonal strategies for network inference. I apply this method to map out the transcriptional network of *Saccharomyces cerevisiae* and demonstrate how network inference results can complement chromatin immunoprecipitation (ChIP) experiments, which directly probe the binding events of transcriptional regulators. Collectively, my contributions improve both the accessibility and practicality of network inference methods.

Acknowledgments

I would like to thank all of the wonderful people that mentored me throughout the course of my graduate studies, who provided advice that both shaped the direction of my research and refined me as a scientist. First I thank my advisor, Dr. Michael Brent, for providing both the guidance and resources to pursue my own research passions of applying computation to facilitate biological discovery. His appreciation for both the mathematical and biological aspects of our work made for an extremely rewarding intellectual partnership. Secondly I thank Dr. Tamara Doering for welcoming me into her lab and providing the resources to train me to become proficient at molecular biology. I am truly grateful for the experience of working in her lab. I also thank other members of my committee, Dr. Weixiong Zhang, Dr. Jeremy Buhler, Dr Gary Stormo and Dr. Barak Cohen for dedicating their time and experience to evaluate my research.

I also thank all of the members of the Brent Lab for time spent discussing science and life, especially Zeke Maier, Drew Michael and Michael Kramer. I would like to thank all of the members of the Doering Lab especially Matt Williams and Mike Skowyra for teaching me the skills required to perform molecular biology. To Mike Skowyra, Matt Williams, Dr. Stacey Gish, Alyssa Marulli and Sarah Spencer I am tremendously grateful for their direct support of my research by creating knockout strains, preparing RNA-Seq libraries, and carrying out other experiments. I also

would like to thank Dr. Meng Yang, Dr. Zhuo Wang and Mike Skowyra for countless hours spent on the water in pursuit of fish, which was therapeutic.

Finally I would like to thank my family, whom I dedicate this dissertation to. To my parents, John and Cynthia, for pushing me to always strive for excellence and my grandfather Harry for telling me to apply myself. I thank my wonderful wife, Divya, for standing by me and encouraging me throughout this process, both sharing in my successes and sacrifices. I thank my daughter Ankita for her goodbye waves and welcome home hugs, which made each day start and end with a smile.

Brian Clifton Haynes

Washington University in St. Louis
May 2012

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Living organisms respond to their environment through complex networks of molecular interactions that regulate gene transcription. Identifying the structure and logic of these networks is paramount to obtaining a systems level understanding of cellular behavior and disease. Genome wide expression profiling over hundreds of genetic backgrounds and growth conditions is now possible due to the rapidly falling cost of high-throughput sequencing.

As a result of this massive growth in the availability of genomics data, the weakest link in the biological discovery pipeline has increasingly become the process of analyzing and integrating evidence across datasets to hypothesize models of gene regulation. The development of algorithms for analyzing large compendia of gene expression data to predict the structure of transcriptional regulatory networks has been a hotbed of research for the past decade, but in spite of this intense effort these tools are not yet widely adopted. Instead, conventional analyses such as clustering and differential expression analysis continue to be applied.

This dissertation addresses the current disconnect in the field of network biology between the bench scientist and the computer scientist. I focus on bridging this divide by bringing to bear greater standards of evaluation to current tools that will enable biologists to better understand their capabilities. I experience the process of working out a biological pathway first hand and in doing so identify ways in which network inference tools can be altered to better suit the needs of bench scientists. Specifically, I unify two strategies used to work out transcriptional networks under a single framework and demonstrate the advantages of combining these complementary approaches. Each chapter of this dissertation is organized as a self-contained unit with its own background, related work, results and discussion sections.

Chapter 2 addresses a need in the field of network biology to standardize the evaluation of network inference algorithms. One of the key impediments to the adoption of network inference tools is lack of systematic evaluations on realistic benchmarks. The lack of such evaluation makes it unclear to potential users which tool is the best to apply to a given dataset. It also leaves uncertainty regarding the expected accuracy. In response to this need, I present GRENDEL, a tool for benchmarking network inference algorithms. GRENDEL is unique in that it is the first benchmarking tool to generate in-silico networks that have biologically realistic topologies and kinetic parameterizations. I evaluate several network inference algorithms using this tool and demonstrate how performance can vary greatly based on the nature of the experiments from which the gene expression data was obtained.

These findings underscore the need to understand when to apply specific inference algorithms in practice.

In chapter 3 I assume the role of a bench scientist and apply conventional analyses of high-throughput gene expression datasets to the study of *Cryptococcus neoformans*, an opportunistic fungal pathogen. I focus on characterizing a pathway that regulates the size of the cryptococcal capsule (a virulence factor) in response to its environment. I use gene expression data and quantitative phenotypic data to identify novel regulators of capsule induction. Next, using strains in which individual transcription factors have been disrupted, I contextualize a novel regulator of capsule, Ada2, in the broader pathway of capsule regulation. This was a rewarding process that not only yielded new insights into the biology of *C. neoformans* but also shaped my thinking moving forward with regard to developing network inference algorithms tailored to application to individual pathways with a limited number of gene expression measurements.

Based in part on the experience of applying conventional approaches to working out a biological pathway in chapter 3, I developed a new algorithm for network inference that I describe in chapter 4. The inspiration for this algorithm is based on the desire to more effectively utilize measurements from genetically perturbed strains, which are often available in gene expression datasets. Before large compendia of gene expression data could be produced, pathways were worked out by performing simple, two-sample comparisons to identify genes that respond significantly to

genetically perturbed regulators. To my surprise, most network inference algorithms do not make use of such information and instead rely solely on compendium-wide statistical relationships between genes identified through regression based analyses. I address this by developing a new inference algorithm, NetProphet, which combines differential expression analysis with LASSO regression. I apply this algorithm to map out the transcriptional network of *Saccharomyces cerevisiae*, demonstrating the advantages of integrating these two network inference strategies.

Contributions: I list the 6 main contributions of this dissertation and briefly summarize each.

1. **GRENDEL: a novel benchmarking suite for network inference.** I present GRENDEL, a method for generating in-silico transcriptional networks that possess biologically realistic topology and kinetic parameterizations. This contribution is presented in chapter 2.

2. **An evaluation of network inference algorithms using GRENDEL.** I apply GRENDEL to examine the effects of experimental design and technical noise on network inference algorithms. This contribution is presented in chapter 2.

3. **A transcriptional signature of capsule induction in the pathogenic fungus *Cryptococcus neoformans.*** I identify a set of genes whose expression pattern correlates with capsule radius. This signature contains many genes already phenotypically implicated in capsule formation as well as many

uncharacterized transcriptional regulators. This contribution is presented in chapter 3.

4. **An integrated model of cryptococcal capsule regulation.** I identify a new transcriptional regulator of cryptococcal capsule induction, Ada2, and contextualize it in the capsule induction pathway. This contribution is presented in chapter 3.

5. **NetProphet: a novel algorithm for mapping transcriptional networks.** I present an algorithm for inferring transcriptional network structure from gene expression data that includes measurements for strains in which transcriptional regulators have been genetically perturbed. This contribution is presented in chapter 4.

6. **An application of NetProphet to infer the transcriptional network of *Saccharomyces cerevisiae*.** I demonstrate how NetProphet can complement ChIP experiments in resolving transcriptional network structure. This contribution is presented in chapter 4.

Chapter 2

Evaluating Transcriptional Network Inference

Brian C. Haynes and Michael R. Brent
Published in Bioinformatics. (2009)

Abstract

Over the past decade, the prospect of inferring networks of gene regulation from high throughput experimental data has received a great deal of attention. In contrast to the massive effort that has gone into automated deconvolution of biological networks, relatively little effort has been invested in benchmarking the proposed algorithms. The rate at which new network inference methods are being proposed far outpaces our ability to objectively evaluate and compare them. This is largely due to a lack of fully understood biological networks to use as gold standards. We have developed the most realistic system to date that generates synthetic regulatory networks for benchmarking reconstruction algorithms. The improved biological realism of our benchmark leads to conclusions about the relative accuracies of reconstruction algorithms that are significantly different from those obtained with A-BIOCHEM, an established in-silico benchmark. The synthetic benchmark utility and

the specific benchmark networks that were used in our analyses are at:

http://mblab.wustl.edu/software/grendel/

2.1 Background

High throughput assays for mRNA expression have paved the way for computational methods that aim to reverse engineer the control architecture of gene regulation. Technologies such as spotted microarrays [1] and oligonucleotide chips [2] have allowed for genome wide expression profiling. More recently, short read sequencing has shown promise for even more precise quantification of mRNA [3,4]. Initially, analyses of high throughput expression data focused on clustering the data in order to identify coregulated genes whose products might take part in a shared biological process [5]. Shortly thereafter, algorithms were developed to reconstruct the underlying regulatory network that best accounts for the expression data. These algorithms differ in the level of detail at which they reconstruct networks. Some output an undirected graph where edges do not indicate which gene is the regulator [6]; others specify the regulator with directed edges [7], and a few even label the edges with kinetic parameters [8].

Improvement and adoption of network reconstruction algorithms has been impeded by the difficulty of objectively assessing their accuracy. Evaluation is difficult primarily because there are very few, if any, fully understood biological networks to use as gold standards. The adoption of standard benchmarks is further complicated by the fact that some inference algorithms require steady state expression data while

others require time courses, some require genetic perturbations while others do not, and so on. Currently, there is no generally accepted substrate on which to compare network reconstruction algorithms.

The most important property of network reconstruction benchmarks is sufficient biological realism to predict accuracy in practical applications. Benchmarks should also provide a sizable population of distinct networks and a range of network sizes, from small pathways to genome scale networks. Without a sufficient number of networks it is impossible to assess the statistical significance of accuracy differences. An ideal benchmark should be flexible enough to render different types of simulated expression data for the same network structure. As we will show, the accuracy of a reconstruction algorithm is strongly determined by the design of gene expression experiments from which the data were generated. A flexible benchmarking system can be used to guide both the development of reconstruction systems and the design of expression experiments aimed at generating data for them.

2.2 Related Work

Several approaches to evaluating reconstruction algorithms have been explored. One approach assumes genes that share common Gene Ontology (GO) categories [9] are more likely to be in a regulatory relationship than those that do not. However, many genes without a direct regulatory relationship also share GO terms. Predictions have also been evaluated on well studied pathways from model organisms, such as the cell

cycle pathway in *Saccharomyces cerevisiae* [10]. However, there are still uncertainties about these networks, so novel predictions could be mistaken as false positives. Another approach to benchmarking is to synthesize a small biological network through genetically engineering cells
[11]. Advantages of this approach are that the true network structure is known and gene expression is measured in a real biological system. However, this is feasible only for small networks and cannot generate enough different networks to provide the statistical power needed to conclude that one algorithm is more accurate than another.

In-silico benchmarks address the need for statistical power because they can run multiple independent trials generated from the same topological and kinetic distributions. They also provide a flexible, low cost method of comparing a wide variety of experimental designs for obtaining gene expression data. However, if in-silico benchmarks are not realistic they may provide a misleading estimate of the reconstruction accuracy in real applications.

Several in-silico regulatory networks have been proposed as benchmarks [12,13], but these are single instances of small, hand built networks and cannot provide robust estimates of expected accuracy. Systems for generating populations of artificial regulatory networks have also been developed. A-BIOCHEM [14] is a system that can generate networks according to several topological (in-degree and out-degree) distributions, such as Erdos-Renyi and power-law. However, the network generating

software is not public, and only a limited collection of networks is made available. Another limitation is that the kinetic parameters are arbitrary and the resulting networks do not conform to the timescale of a real biological system. Furthermore, translation is not modeled: mRNA acts as a surrogate for active protein product. SynTReN [15] makes the same assumptions about kinetics, but generates more realistic topologies by sampling subgraphs of known transcriptional networks. This approach has the advantage of capturing features beyond degree distribution, such as clustering coefficients, modularity and enrichment of biological network motifs. The downside of this sampling approach is that the networks generated may not be probabilistically independent, since they can contain overlapping pieces of the known networks, and this problem gets worse as the size of the benchmark networks increases. This lack of independence limits the potential for testing the statistical significance of differences between reconstruction algorithms.

To address these limitations, we have developed a publicly available, synthetic benchmarking system that is more biologically realistic than previous methods. It uses network topologies that closely reflect those of known transcriptional networks and kinetic parameters from genome wide measurements of protein and mRNA half-lives, translation rates, and transcription rates in *S. cerevisiae*. We compared our method to an established in-silico benchmark, A-BIOCHEM [14]. Using these benchmarks, we evaluated the accuracy of four network reconstruction algorithms, most of which have not been directly compared before: ARACNE [6], CLR [16], Symmetric-N [17,18] and DBmcmc [7]. Our results show that the increased realism

of our simulations leads to conclusions that are significantly different from those indicated by the more established A-BIOCHEM benchmark.

2.3 Approach

In order to provide a more realistic synthetic benchmark to users and developers of network reconstruction systems, we have built an open and extensible software toolkit, Gene REgulatory Network Decoding Evaluations tooL (GRENDEL). GRENDEL generates random gene regulatory networks according to user defined constraints on the network topology and kinetics. It then simulates the state of each regulatory network under various user defined conditions (the experimental design) and produces simulated gene expression data, including experimental noise at a user defined level. Figure 2.1 shows an overview of the workflow we use to generate and simulate regulatory networks.

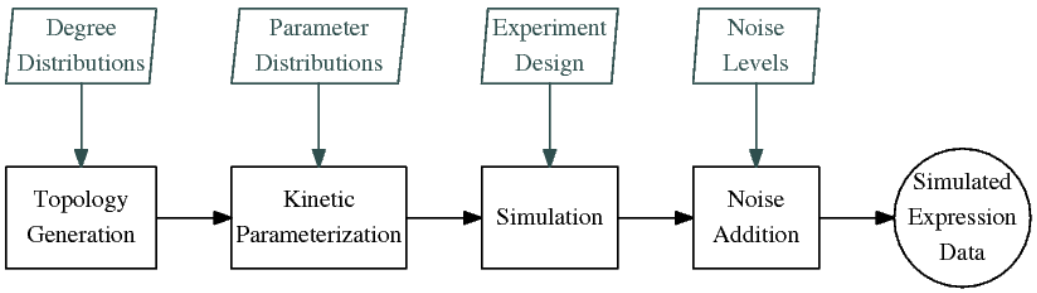

Figure 2.1. The basic workflow we are using to generate an in-silico regulatory network and produce simulated expression data from it. The user inputs are shown above each step of the process.

The artificial networks generated by GRENDEL are continuous-time dynamical systems with three independent types of molecular species: mRNAs, proteins, and environmental stimuli (e.g. extracellular glucose or iron). To our knowledge, all other in-silico benchmarks use the mRNA concentration as a proxy for active protein product. This eliminates the decorrelation of a gene's mRNA and protein concentrations that arises during condition shifts in real systems.

Figure 2.2 shows an example of mRNA-protein decorrelation in our system. In real networks, the relationship between a gene's mRNA and protein concentrations has been shown to be crucial for determining biologically relevant dynamics, as in certain oscillators [19].

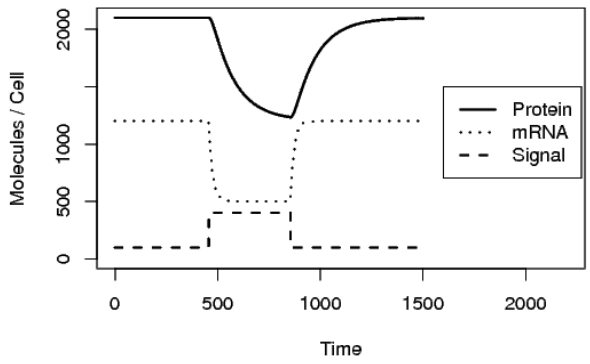

Figure 2.2. A time course plot showing the dynamics of the three molecular species in our simulation: mRNAs, proteins, and external signals. In this simulation, the signal represses transcription of a gene. Note the decorrelation of mRNA and protein following the condition shifts.

Environmental stimuli, or *signals*, were included for the purpose of supporting time courses. Signals are different than mRNAs and proteins in that they are driven by external rules and are independent of the concentrations of mRNAs and proteins. Signal transduction happens on a much faster timescale than transcription, so we can approximate it as being instantaneous. Using this approximation, the signal controls transcription in the same way a transcription factor does, simplifying the transduction cascade.

Computationally generating random biological networks involves two modular steps: topology generation and kinetic parameterization. The topology generation step defines the reagents, catalysts and products of each reaction. In our implementation the topology is represented by a directed graph with nodes representing signals and genes. An edge from node A to B in the network indicates that A regulates the transcription of B, where A is either a gene or a signal and B is a gene. After generating a graph indicating which genes regulate which other genes, GRENDEL chooses parameters for the differential equations that determine the concentration of each protein and each mRNA. These parameters allow for the simulation of both a network's responses to environmental changes and the effects of genetic interventions on those responses.

After generating a network, GRENDEL exports it in Systems Biology Markup Language (SBML) [20], a versatile representation that is becoming a standard for

communicating biochemical models. Networks specified in SBML can be simulated by using one of several SBML integration programs, including COPASI [21], CellDesigner [22], and SBML ODE Solver Library (SOSlib) [23]. Our software uses SOSlib to deterministically integrate the ODEs that define the dynamical system, resulting in noiseless expression data. Simulated experimental noise is then added to the data according to a log normal distribution, with user defined variance. Biological noise is not considered here, but the networks our method produces could be simulated with biological noise by using an SBML-based stochastic integrator [24].

2.4 Results

We set out to evaluate the utility of synthetic benchmarks for two applications: assessing the performance of network reconstruction methods relative to one another and supporting cost-benefit analysis of designs for gene expression experiments. To accomplish this, we carried out three sets of computational experiments. The first set examines how the design of a steady-state gene expression experiment affects the performance of network inference methods. The second set investigates the effects of technical noise on the quality of network inference from steady state data. The third set explores the effects of sampling frequency on network reconstruction from time course data.

Throughout, we compared the results obtained with our benchmarking suite, GRENDEL, to those obtained with A-BIOCHEM [14], a benchmark that has been used in several previous studies [25,26,6]. The reconstruction algorithms we evaluated are: ARACNE [6], CLR [16], Symmetric-N [17,18] and DBmcmc [7]. ARACNE, CLR, and Symmetric-N are applied to steady-state expression data; Symmetric-N and DBmcmc are applied to time course data. To evaluate an inference method, we compared each edge it inferred to the known network structure. To facilitate comparison among inference algorithms the gold standard network was first converted to an undirected network. For each inferred network, we calculated precision ($N_{TP} / (N_{TP} + N_{FP})$), recall ($N_{TP} / (N_{TP} + N_{FN})$), and the area under the precision-recall curve.

2.4.1 Experiment 1: Experimental design comparison

We analyzed the effects of experimental design by using a set of networks generated by GRENDEL and a set of networks (Century-SF) provided by A-BIOCHEM. We wanted to test whether the degree distributions of our networks and those of and the CenturySF networks might lead to differing conclusions about experimental design. To isolate the effects of network topology, the kinetic parameters, such as transcription and mRNA degradation rates for every gene in the system, are the same for both sets of networks.

Using these networks, we generated simulated data from five experimental designs:

- **Diverse:** 300 measurements from a diverse population
- **Knockouts:** 100 measurements knocking out each gene
- **Overexpression:** 100 measurements overexpressing each gene
- **Knockouts + overexpression:** 200 measurements knocking out and overexpressing each gene
- **Knockouts + 2 overexpressions:** 300 measurements knocking out each gene and overexpressing at two levels

The **Diverse** data set was generated for comparison to [6], who used it to model samples from a genetically and phenotypically diverse population, such as samples from tumors found in different individuals. In their model, every sample has the same network topology but completely independent, randomly chosen kinetic parameters for all genes. For each simulated measurement M_k, we set $T_i^{M'} = \sigma_{i,k}\, T_i^M$ and $D_i^{M'} = \tau_{i,k}\, D_i^M$ for each gene, where $\sigma_{i,k}$ and $\tau_{i,k}$ are random variables chosen from the uniform distribution [0.0, 2.0]. For each gold standard network topology, all of these parameters were randomly selected 300 times, creating 300 independently parameterized networks. Figure 2.3 shows the precision-recall curves for ARACNE, CLR and Symmetric-N on this data set. ARACNE is clearly the method of choice in the A-BIOCHEM network topologies, recovering close to 50% of the true edges in the network before acquiring many false edges. Using the GRENDEL network topology, the estimated accuracies of all methods were lower, but their relative accuracies were about the same as on the A-BIOCHEM topologies.

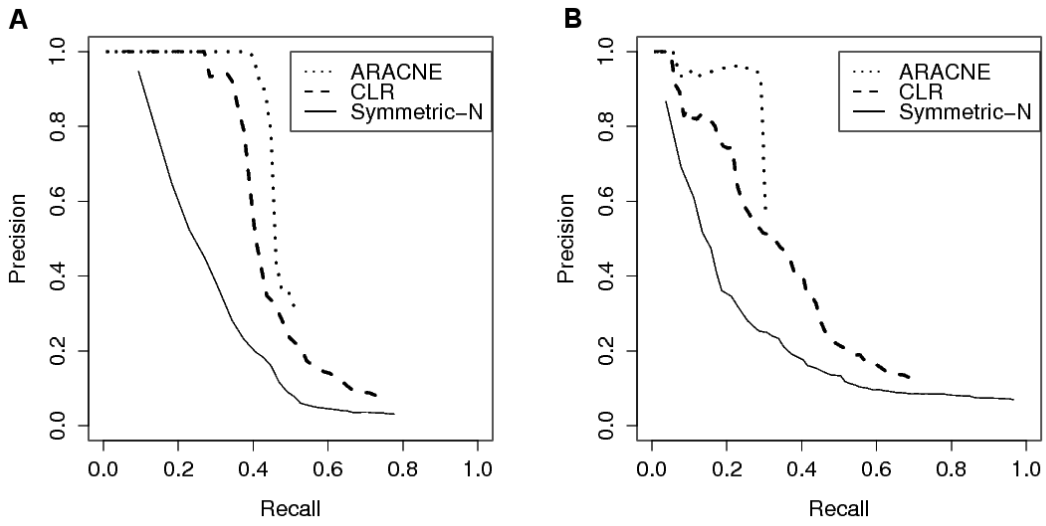

Figure 2.3. Precision recall curves for network inference from the **Diverse** design. The precision recall curves that are shown reflect the median performance, ranked according to AUC-PR. Panel A, A-BIOCHEM topology; Panel B, GRENDEL topology.

In the **Knockouts** design, for each steady state measurement M_i, a single gene was knocked out by setting $T_i^{M'}=0$. The expression level of every gene was measured 100 times, with a different gene knocked out each time. The **Overexpression** design was analogous, but each gene was overexpressed rather than being knocked out. Constitutive overexpression from a plasmid was modeled by adding to the system an additional term that produced the mRNA at a constant rate. The **Knockouts + overexpression** design combines the measurements from **Knockouts** and **Overexpression** for a total of 200 observations. **Knockouts + 2 overexpressions** augments the data from **Knockouts + overexpressions** with another 100

measurements in which each gene is expressed at twice the concentration of the first overexpression.

Figure 2.4 shows the results in terms of area under the precision recall curve (AUC-PR). The error bars represent the standard error of the mean. For the Diverse experiment, ARACNE outperforms the other methods when inferring the A-BIOCHEM networks, but for the GRENDEL networks, CLR does slightly better. Outside of the **Diverse** regime, the outcome is dramatically different: the other systems consistently outperform ARACNE.

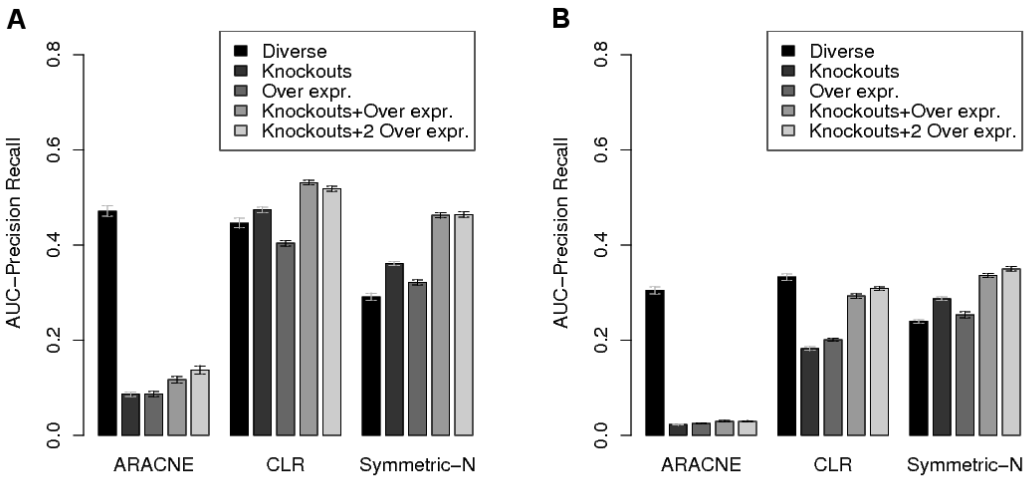

Figure 2.4. Effects of different experimental designs on reconstruction accuracy. Panel A, A-BIOCHEM topology; Panel B, GRENDEL topology.

On the A-BIOCHEM benchmark, CLR performs slightly better on Knockouts than on Diverse, but on the GRENDEL benchmark it performs much worse on Knockouts than on Diverse. Similarly, A-BIOCHEM suggests that knock-outs are more

informative to CLR than overexpressions, whereas GRENDEL shows the opposite to be true. When using the GRENDEL benchmark, the estimated accuracies of all methods were lower than with A-BIOCHEM. GRENDEL thus appears to provide a tighter upper bound on how well these methods would perform on a real biological system similar to the yeast transcriptional network.

2.4.2 Experiment 2: Effects of Technical Noise

In a follow-up experiment, we wanted to investigate the effects of experimental noise on inference accuracy. The log_2 signal intensity ratio of technical replicates in oligo and spotted arrays has been shown to follow a normal distribution whose standard deviation ranges from 0.1 to 0.5 [27]. We therefore examined three levels of simulated noise: low (s.d.=0.1), medium (s.d.=0.25), and high (s.d.=0.5), and a noise-free baseline condition. To simulate technical noise, we perturbed the noise free data for each gene by a multiplicative factor independently chosen from the specified log_2-normal distribution.

Figure 2.5 shows the impact of noise on reconstruction of networks with the A-BIOCHEM and GRENDEL topologies using simulated data from the **Knockouts + 2 overexpression** design. In both benchmarks CLR was the least sensitive to noise followed by Symmetric-N and ARACNE. For all three algorithms, the effects of noise were not as strong on the GRENDEL networks compared to the A-BIOCHEM networks. Upon further examination, we found that the effect of noise was the most

pronounced on genes with fewer than three regulators, which account for 55% of edges in A-BIOCHEM compared to 20% in GRENDEL. However, that does not account for the entire effect: the loss in accuracy in A-BIOCHEM is higher than GRENDEL even when in-degree is held constant. This suggests that global topological features may also have an effect.

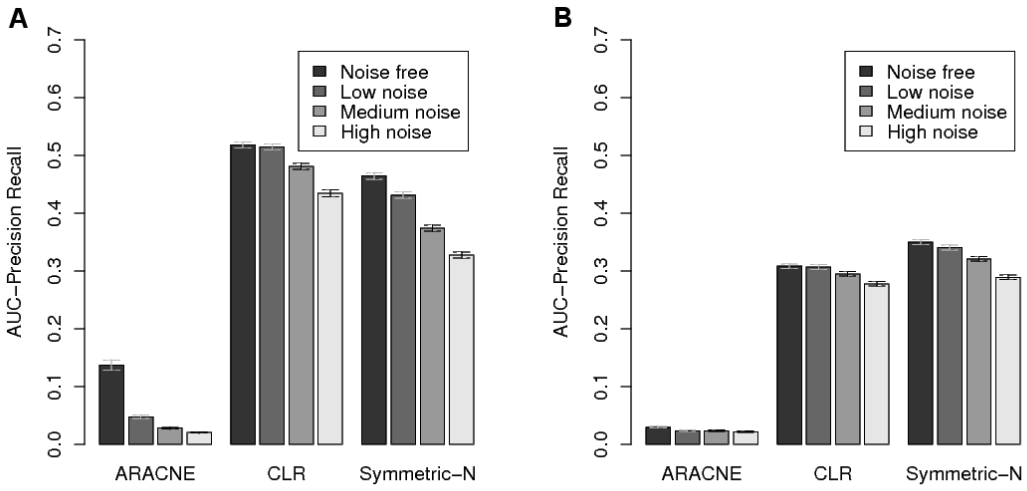

Figure 2.5. Knockouts + 2 Overexpressions revealing the effects of technical noise on network reconstruction accuracy. Panel A, A-BIOCHEM topology; Panel B, GRENDEL topology.

2.4.3 Experiment 3: Time course data

To isolate the effects of using realistic parameters for half-lives, transcription rates, and translation rates, we created two sets of networks using GRENDEL. In one set, kinetic parameters were drawn from genome wide measurements in *S. cerevisiae*. In

the second set, the kinetic parameters were as in the A-BIOCHEM benchmark -- all degradation, transcription and translation rate constants were set to 1.0. Each set contained 250 simulated networks, each with 20 genes and two external signals. For each network we simulated a time course experiment in which gene expression was measured at fixed intervals for approximately 33.3 hours. During this time each system underwent 4 condition shifts: 2 where each environmental signal was perturbed and 2 when each signal was restored to its original state.

The times at which each signal was perturbed and restored were chosen at random. We varied the sampling interval from 60 minutes to 2 minutes. For each interval, we evaluated DBmcmc and Symmetric-N on the arbitrary and realistically parameterized networks.

Figure 2.6 (Panel A) shows the accuracy of DBmcmc as a function of sampling frequency. As the sampling frequency increases, so does the accuracy, but not by very much. As the sampling interval decreases from 1 hr to 10 minutes, the modest accuracy improvement begins right away when benchmarking on networks with realistic parameters. On arbitrarily parameterized networks, however, the improvement is even smaller, and it does not begin until the sampling frequency reaches 5 minutes. A possible reason for this is that the networks with arbitrary parameters reached steady state much more quickly than those with the realistic parameters, so there is a greater chance that multiple cascading regulatory events will occur between sampling intervals. The networks with realistic parameters respond

more slowly, so they have a reduced chance of multiple regulatory events occurring between sampling intervals.

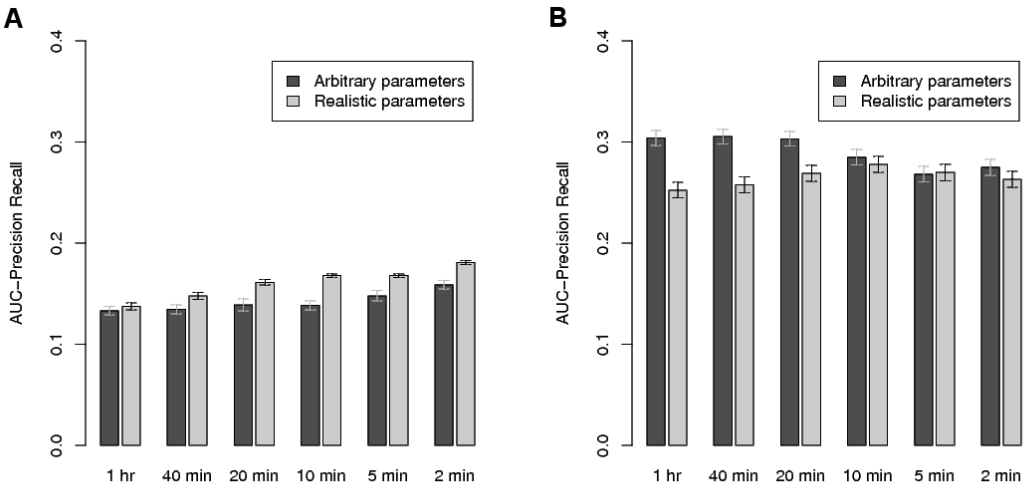

Figure 2.6. Evaluating the performance of DBmcmc and Symmetric-N comparing arbitrary kinetic parameterizations against realistic ones on a 20 gene network with 2 external signals over varying sampling frequencies (x-axis). Panel A, DBmcmc; Panel B, Symmetric-N.

For Symmetric-N, the arbitrary and realistic parameterizations cause the performance to trend quite differently than with DBmcmc, see Figure 2.6 (Panel B). For the arbitrary parameterization, performance actually benefited from sampling at longer intervals. For the realistic parameterization, performance improved as the sampling interval decreased, reaching a plateau at approximately 10 minute intervals. Symmetric-N did very well on some of the random networks and very poorly on others, with few networks yielding intermediate accuracy (data not shown). This was true for all sampling intervals and both kinetic parameterizations. The fact that the

performance distribution of Symmetric-N was bimodal underscores the need to test reconstruction algorithms over a large population of networks as opposed to a single network instance.

2.5 Discussion

One of the benefits of using simulated networks to evaluate reconstruction algorithms is the statistical power one gets from being able to generate many networks sampled from the same distribution. If an algorithm performs very poorly at reconstructing a specific subset of networks, the ability to generate large populations of networks enables developers to identify the weaknesses of their method. In-silico benchmarks also allow for properties of regulatory networks, such as degree distributions, experimental noise, biological noise, and network size, to be varied independently of one another. This helps to identify the properties that contribute most to reconstruction error.

Simulated networks also have great potential as cost effective tools for determining the optimal experimental design to use with a given network reconstruction method. We have demonstrated the use of simulated networks in determining the optimal sampling interval for a time course experiment. For steady state data, we have shown they can provide hints about how many samples should be taken to achieve the desired level of accuracy, and whether gene knockouts or overexpressions are more useful. Being able to simulate experiments will likely reduce the cost of network

reconstruction, improve its accuracy, and set expectations appropriately. However, the results obtained with simulated networks are only a first step in evaluation that must ultimately be followed by application to real biological systems. At present, simulated networks are rough approximations that omit many important aspects of biological systems, including localization and post-translation modifications.

GRENDEL is an extensible, open source toolkit that provides greater flexibility and realism than previously published synthetic benchmarks. GRENDEL's more realistic network topologies not only lead to lower accuracy estimates for all algorithms tested, but they also change estimates of which algorithms are more accurate under different experimental designs. We believe that GRENDEL will be useful both to experimentalists designing gene expression studies and algorithm developers implementing and testing new computational approaches. We hope that, through both of these avenues, it will help to advance the useful application of algorithms for reconstructions of gene regulatory networks.

2.6 Methods

2.6.1 Topology Selection

In a regulatory network, the out-degree of a gene represents the number of genes it regulates, while the in-degree represents the number of genes that regulate it. Biological networks are often described as being scale free, meaning that their

degree distributions follow a power-law [28]. However, the evidence suggests that only the out-degree distribution is scale-free. The in-degree distribution is compact (concentrated around its mean) [29,30]. To generate random networks with these characteristics, we developed a new algorithm. Our algorithm extends the preferential attachment model of [31], to support directed graphs with distinct in-degree and out-degree distributions.

The preferential attachment model starts from an empty graph and incrementally adds nodes. Newly added nodes are connected to an existing node selected according to a distribution favoring nodes that already have many connections. In our extension of this model, newly added nodes form multiple directed connections:

- Start with a graph containing signal nodes and *k* genes, but no edges. These initial nodes, which are called seeds, will have no incoming edges, so they will be unregulated. The number of seeds, *k*, is a user-selected parameter.
- For each non-seed gene g_j,
    - Assign g_j an in-degree $I[g_j]$ according to the user-specified in-degree distribution.
    - Add g_j to the network by choosing $I[g_j]$ existing network nodes as parents (regulators) according to the following distribution:

$$P(a_{i,j}=1)=\frac{B+\sum_{n=1}^{N}a_{i,n}}{Z}$$

where a_{ij} is an element of the adjacency matrix for the network under construction, B is a user-defined constant that determines the power of the power-law distribution,

and *Z* is a normalizing constant obtained by summing the numerator over all possible parents -- i.e. all nodes currently in the network. The probability of selecting each node in the network as a parent is proportional to its current out-degree plus the constant *B*. In our current implementation *k* is set to 0 if there are signals and 1 if there are not (the number of signals is a user-selectable parameter).

This algorithm produces a network in which the out degree distribution follows a power-law and the in-degree can follow any specified distribution from which sampling is possible. In an analysis of the yeast transcriptional network [32], a power-law was fit to the empirical out-degree distribution: $x^{-0.6919}$, and an exponential was fit to the empirical in-degree distribution: $e^{-0.3852x}$. GRENDEL generates networks using our extended preferential attachment algorithm with out and in-degrees that match these empirical distributions.

To get a clearer picture of the networks generated by our algorithm, we compared their degree distributions to those of the A-BIOCHEM CenturySF networks. This collection consists of 50 networks, each containing 100 genes with an average of 200 edges per network. The networks are scale free: both in and out-degree distributions can be approximated by a power-law. We generated an analogous set of 50 networks each with 100 genes, where both in and out-degree distributions were set to match the yeast network, as described above (no signals were used in this set of networks). We noted that the out-degree distributions of the GRENDEL networks have much longer tails, corresponding to the presence of larger hubs. For in-degree

distributions, the A-BIOCHEM networks follow a power-law, while GRENDEL networks are exponential. The tail lengths are the same, with the most highly regulated gene in each set of networks having 22 regulators, but the A-BIOCHEM networks have an under representation of genes with three or more regulators. When comparing two representative networks from each benchmark (see Figure 2.7) clear differences beyond degree distribution are evident. Unlike GRENDEL, the A-BIOCHEM network contains no single-input modules (SIMs) -- a network motif where a single gene exclusively regulates a set of genes [30]. A likely reason for the lack of SIMs in the A-BIOCHEM networks is that each gene has a total degree of two or more. As a result of this artifact, any gene that does not act as a transcription factor must itself be regulated at least two other genes.

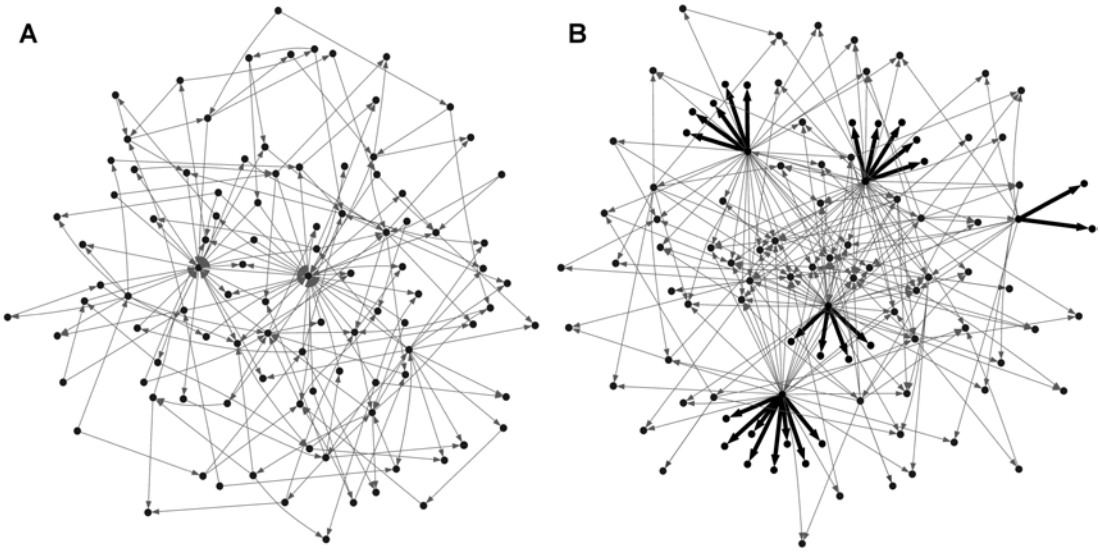

Figure 2.7. Representative 100-gene networks from the A-BIOCHEM and GRENDEL benchmarks with the SIM network-motif shown in bold. Panel A, A-BIOCHEM; Panel B, GRENDEL.

2.6.2 Kinetic Parameterization

Before the behavior of a randomly generated network can be simulated, parameters must be chosen for the differential equations that determine the concentration of each protein (p_i) and each mRNA (m_i). The equation for the change in concentration of protein *i* is

$$\frac{\delta p_i}{\delta t} = T_i m_i - D_i^P p_i$$

which requires two parameters: the protein's translation (T_i^P) and degradation (D_i^P) rate constants. The equation for the change in concentration of mRNA *i* is

$$\frac{\delta m_i}{\delta t} = S_i(R) - D_i^M m_i$$

where D_i^M is the degradation rate constant of the mRNA, *R* is a vector of regulator concentrations (signals and proteins), and S_i maps regulator concentrations to the transcription rate of gene *i*.

Similar to other approaches, we use a transcriptional rate law, $S_i(R)$, that models Hill kinetics [33,34]. We begin by defining a repression function for a single regulator:

$$F(R, K, n) = \frac{K^n}{R^n + K^n}$$

where, *R* is the concentration of the repressor, *K* is the binding affinity of the repressor, and *n* is the Hill-coefficient that controls the sigmoidicity of *F*. When the regulator concentration is zero, $F(R, K, n)$ is one (no repression). As the regulator

concentration increases without limit, *F(R, K, n)* tends toward zero (total repression). The corresponding activation function is

$$G(R,K,n) = \frac{R^n}{R^n + K^n} + 1$$

where *R* represents the activator concentration. *G(R,K,n)* is one when the activator is absent and tends toward two as activator concentration increases without limit. The effects of these activation and repression functions on the transcription rate are defined by:

$$S_i(R) = \left(\beta_i + Z\left(\prod_{R_k \in A_i} G(R_k, K_{ik}, n_{ik}) - 1\right)\right) \times \left(\prod_{R_j \in I_i} F(R_j, K_{ij}, n_{ij})\right) \times T_i^M$$

where I_i is the set of regulators acting as repressors of gene *i*, A_i is the set of regulators that act as activators of gene *i*, and *R* is a vector of regulator concentrations. T_i^M is the maximum transcription rate, β_i defines the basal transcription rate of gene *i*, and ranges from 0 to 1, *Z* is a normalization factor that forces the activation term to lie between β_i and 1.

$$Z = \frac{1 - \beta_i}{2^{|A_i|} - 1}$$

When β_i is equal to 0.5, our transcriptional regulation function is equivalent to the A-BIOCHEM transcriptional rate law described in [14]. Once a network topology has been defined, each regulator is designated as either a repressor or an activator for each gene.

The novelty of our kinetic model lies in its use of more realistic parameters. The parameter selection process begins by randomly pairing each gene in the synthetic network with a real gene from *S. cerevisiae*. The synthetic network's gene is assigned the translation rate, protein decay rate, mRNA decay rate, and mRNA transcription rate of the real gene, which are available from high throughput studies [35–38]. In this way, our synthetic networks should behave on the same timescale as a real biological system. The parameters that are not available for large numbers of real genes are the Hill coefficients n_{ik}, binding affinities K_{ik} and β_i. To facilitate direct comparisons with A-BIOCHEM, we set these parameters in order to achieve equivalence as follows: $n_{ik}=1.5$, $K_{ik}= 0.01 / max(R_k)$ where $max(R_k)$ is the saturating concentration of regulator R and $\beta_i=0.5$.

Chapter 3

Toward an Integrated Model of Capsule Regulation in *Cryptococcus neoformans*

Brian C. Haynes, Michael L. Skowyra, Sarah J. Spencer, Stacey R. Gish, Matthew Williams, Elizabeth P. Held, Michael R. Brent, and Tamara L. Doering
Published in PLoS Pathogens. (2011)

Abstract

Cryptococcus neoformans is an opportunistic fungal pathogen that causes serious human disease in immunocompromised populations. Its polysaccharide capsule is a key virulence factor which is regulated in response to growth conditions, becoming enlarged in the context of infection. We used microarray analysis of cells stimulated to form capsule over a range of growth conditions to identify a transcriptional signature associated with capsule enlargement. The signature contains 880 genes, is enriched for genes encoding known capsule regulators, and includes many uncharacterized sequences. One uncharacterized sequence encodes a novel regulator of capsule and of fungal virulence. This factor is a homolog of the yeast protein Ada2, a member of the Spt-Ada-Gcn5 Acetyltransferase (SAGA) complex that regulates transcription of stress response genes via histone acetylation. Consistent with this homology, the *C. neoformans* null mutant exhibits reduced histone H3

lysine 9 acetylation. It is also defective in response to a variety of stress conditions, demonstrating phenotypes that overlap with, but are not identical to, those of other fungi with altered SAGA complexes. The mutant also exhibits significant defects in sexual development and virulence. To establish the role of Ada2 in the broader network of capsule regulation we performed RNA-Seq on strains lacking either Ada2 or one of two other capsule regulators: Cir1 and Nrg1. Analysis of the results suggested that Ada2 functions downstream of both Cir1 and Nrg1 via components of the high osmolarity glycerol (HOG) pathway. To identify direct targets of Ada2, we performed ChIP-Seq analysis of histone acetylation in the Ada2 null mutant. These studies supported the role of Ada2 in the direct regulation of capsule and mating responses and suggested that it may also play a direct role in regulating capsule-independent antiphagocytic virulence factors. These results validate our experimental approach to dissecting capsule regulation and provide multiple targets for future investigation.

3.1 Background

Cryptococcus neoformans is an opportunistic fungal pathogen [39]. The disease it causes, cryptococcosis, is contracted by inhalation of infectious particles (spores [40] or dessicated cells), which initiate a pulmonary infection. In the setting of immune compromise the fungus disseminates, with particular predilection for the central nervous system where it can cause a fatal meningoencephalitis. In otherwise healthy hosts, the infection may remain latent for extended periods, emerging in the event of immune compromise [41]. The impact of the disease is significant, especially in populations with limited access to health care, leading to an estimated 600,000 deaths per year [42].

A variety of factors have been implicated in cryptococcal virulence. These include melanin synthesis [43]; urease and phospholipase secretion [44,45]; titan cell formation [46,47]; and the ability to survive at host body temperature. Additionally, the main feature that distinguishes *C. neoformans* from other pathogenic fungi is an extensive polysaccharide capsule that surrounds the cell wall and is required for virulence [48]. Capsule size varies tremendously with growth conditions, becoming particularly large during mammalian infection [49]. Capsule expansion can be induced *in vitro* by mimicking aspects of the host environment such as low iron availability, the presence of mammalian serum, and physiological concentrations of carbon dioxide [50–52]. Strain virulence correlates with capsule size *in vivo* [53],

implicating the regulation of capsule formation as a critical factor in the pathophysiology of cryptococcal disease.

3.2 Related Work

Our current knowledge of capsule regulation derives primarily from studies where mutations of specific genes yield cells with abnormal capsules. A variety of readily assayed phenotypes that are related to the size or nature of the capsule (including cell sedimentation behavior [54], antibody reactivity [55], India ink staining, and colony morphology) has enabled the identification of a wide array of such mutants. Most of these have reduced virulence, emphasizing the central role of the cryptococcal capsule in pathogenesis.

Capsule size is regulated by distinct and overlapping signaling pathways, including those typically associated with stress response. The best-characterized of these, the cAMP pathway, responds to amino acid starvation, low glucose, and elevated carbon dioxide [56]. Stimulation of this pathway leads to high intracellular cAMP levels, which activate the kinase Pka1 [57]. This enzyme in turn activates the C_2H_2 zinc finger transcription factor Nrg1, leading to the transcriptional induction of genes that are directly involved in capsule assembly [58]. Pka1 also activates another transcription factor, Rim101, which is necessary for capsule enlargement. Interestingly, activation of Rim101 requires elements of both the cAMP pathway and

the pH-responsive Rim signaling pathway [59]. Deletion of the genes encoding Pka1, Nrg1, or Rim101 leads to reduced capsule size.

Iron sensing mechanisms also influence capsule formation. Transcription factors Hap3 and Hap5 are involved in both iron homeostasis and capsule regulation; deletion of the corresponding genes leads to a reduction in capsule size [60]. In addition to Hap3 and Hap5, the iron responsive transcription factor Cir1 also regulates capsule [61], in part by transcriptionally regulating the cAMP pathway. Recently, ChIP-chip studies revealed that Cir1 is directly regulated by another transcription factor, Gat201 [62]. Strains lacking either Cir1 [61] or Gat201 are hypocapsular [63].

Capsule regulation is also influenced by the HOG pathway. Several proteins in this pathway (including Hog1, Pbs2, and Ssk2) negatively regulate capsule size [64]. Epistasis analysis shows that the cAMP pathway is required for this HOG-dependent influence on capsule, but the mechanism of the cross-talk between these two central signaling pathways is unknown. Normal capsule formation also requires proteins in pathways related to temperature sensing [65], sexual development [66], and cell wall integrity [67].

More broadly, chromatin remodeling has been implicated in capsule regulation, by the observation that cells lacking the histone acetyltransferase Gcn5 are hypocapsular [68]. Gcn5 is a member of the well-conserved SAGA complex, which

acts in transcriptional regulation from fungi to humans [69]. Sequence analysis suggests that other SAGA proteins are present in *C. neoformans*, but Gcn5 is the best conserved and the only one that has been characterized [68].

Over 60 genes have been identified as important players in capsule formation due to the effects of their deletion on capsule structure or morphology; we refer to such genes as 'capsule-implicated'. However, because the majority of cryptococcal transcription factors and signaling proteins are uncharacterized, it is likely that important elements of the capsule regulatory network are missing from this group. Furthermore, some capsule-implicated genes may be required for other primary functions, such as cell wall synthesis, that have incidental effects on capsule formation.

As reviewed above, components of several known signaling pathways are required for capsule formation, but there is no model that accounts for the integration of these pathways to regulate capsule growth. To begin constructing such a model, we have identified genes whose RNA levels are correlated with capsule size over a range of *in vitro* conditions. We term this set of genes the transcriptional signature of capsule. This signature includes previously capsule-implicated genes as well as multiple uncharacterized genes encoding putative regulatory factors. We chose to analyze one uncharacterized gene, *ADA2*, which encodes a putative DNA-binding protein. We now show that Ada2 is a novel regulator of capsule and of other virulence-related features of Cryptococcus. Analysis of downstream targets of Ada2 and other capsule

regulators by RNA-Seq and ChIP-Seq suggests the context of Ada2 in the capsule regulatory network and illustrates the effectiveness of this approach in unraveling complex regulatory networks.

3.3 Results

3.3.1 Identifying the Transcriptional Signature of Capsule

We reasoned that the transcript abundance of many genes involved in the regulation and synthesis of capsule would correlate with capsule size across multiple growth conditions. To test this hypothesis, and potentially identify capsule regulatory genes beyond those previously reported, we cultured the *C. neoformans* serotype A reference strain H99 in four conditions known to stimulate capsule formation to varying degrees. For each condition, we also cultured the cells in a similar medium that stimulates capsule formation to a lesser extent. The eight conditions used were low iron medium (LIM) with and without the chelating agent ethylenediaminetetraacetic acid (EDTA); phosphate-buffered saline (PBS) with and without fetal bovine serum (FBS); Dulbecco's Modified Eagle's Medium (DMEM) in room air (RA) or in 5% CO_2; and Littman's medium (LIT) with two concentrations of thiamine (LO-THI / HI-THI). After 24 h the average capsule radius in each culture was assessed by light microscopy. The remaining cells were used to isolate total

RNA for hybridization against a *C. neoformans* serotype A/D microarray (http://gtac.wustl.edu/services/microarray/rna-analysis/cryptococcus-neoformans.php). To identify genes whose transcript abundance correlated with capsule size, we compared the transcription profiles over all eight conditions to the quantitative measurements of capsule radius (Figure 3.1).

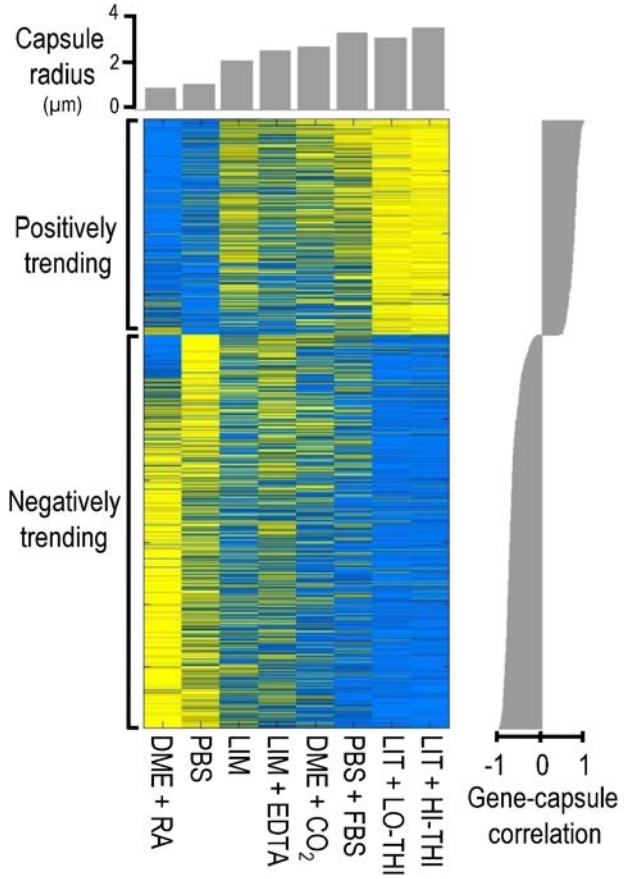

Figure 3.1. The transcriptional signature of capsule induction. Shown is a heat map of gene expression (blue, low expression; yellow, high expression) for the 880 genes whose expression, as assessed by microarray analysis, trends with capsule size. Cell growth conditions (see section 3.5) for each column are indicated below the heat

map and average capsule radius is plotted above (gray bars). The correlation of gene expression and capsule size is plotted at the right.

Our analysis revealed 880 genes whose transcript abundance correlated significantly with capsule size, which we considered the transcriptional signature of capsule induction. Within this set, we identified 316 genes whose transcription correlated positively with capsule radius and 564 genes whose transcription correlated negatively. Among the positively correlated genes, most are involved in responses to stress, including DNA damage repair, trehalose biosynthesis and sugar transport. In contrast, many of the negatively correlated genes are involved in mitochondrial function and ribosome biogenesis.

We expected that some of the genes in the transcriptional signature would specifically influence the formation of capsule (see section 3.4). Consistent with this hypothesis, the set of genes whose RNA levels correlated positively with capsule size was enriched for capsule-implicated genes ($p < 0.02$; see section 3.5); no such enrichment was observed among genes that correlated negatively. Positively correlated genes that are capsule-implicated included the genes encoding regulatory proteins Cir1, Hap5 [60], and Ste20 [70] and the phosphodiesterases Pde1 and Pde2 [71] (see section 3.4).

The transcriptional signature of capsule included previously uncharacterized genes that encode putative transcription factors, signaling proteins, and sugar transporters. It is likely that many of these genes are involved in capsule regulation and assembly. We were particularly interested in one previously uncharacterized gene, CNAG_01626, which encodes a putative DNA binding protein. Expression of CNAG_01626 correlated positively with capsule size (Figure 3.2). For comparison, Figure 3.2 also shows the correlations obtained for two cryptococcal transcriptional regulators, *CIR1* and *SSN801*, whose roles in capsule regulation have previously been demonstrated. *CIR1* showed significant positive correlation with capsule size, consistent with the hypocapsular phenotype of *cir1*Δ mutants [61], while *SSN801* exhibited a negative correlation with capsule size, consistent with the hypercapsular phenotype of the corresponding deletion mutant [63].

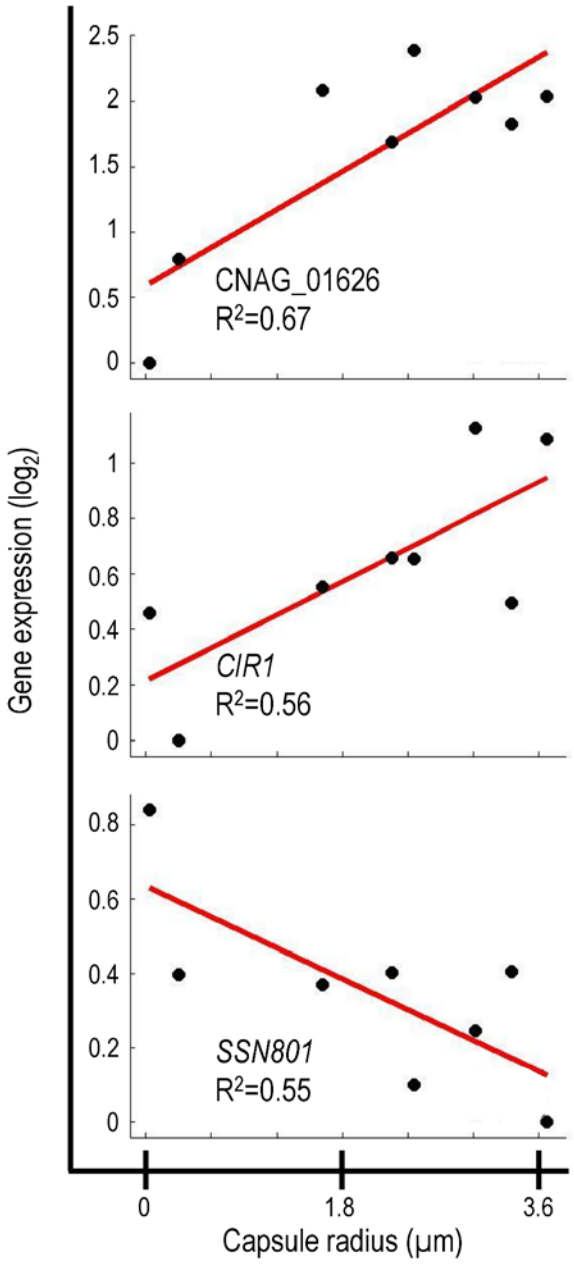

Figure 3.2. Correlation of gene expression and capsule size for selected genes. Data is shown for three genes that demonstrate correlation between gene expression and capsule size.

Given the strong correlation of CNAG_01626 transcription with capsule size, we suspected that the corresponding gene product was a regulator of capsule formation. Because this gene encodes multiple putative DNA-binding domains (Myb-like and SWIRM), we further expected that it might act at a transcriptional level. This hypothesis was supported by the 33% homology we noted between the amino acid sequence predicted for CNAG_01626 and that of the *Saccharomyces cerevisiae* Ada2 protein. In *S. cerevisiae*, Ada2 is a member of the Spt-Ada-Gcn5 Acetyltransferase (SAGA) complex that mediates histone acetylation [72]. Within SAGA, Ada2 is required for proper catalytic activity of the acetyltransferase Gcn5 [73]. Based on the homology between the cryptococcal gene and *S. cerevisiae ADA2*, we decided to refer to CNAG_01626 as *ADA2*.

3.3.2 Cryptococcal Ada2 Influences Capsule Formation

Since transcription of the cryptococcal *ADA2* gene positively correlates with capsule size, we hypothesized that deleting *ADA2* would yield hypocapsular cells. To assess the role of this putative transcriptional regulator in capsule formation, we replaced the *ADA2* genomic coding sequence with a nourseothricin-resistance marker (*NAT*) in the serotype A strain KN99α, derived from the serotype A reference strain H99 [74]. We then incubated *ada2*Δ mutant cells under capsule-inducing conditions and examined capsule size by negative staining with India ink. Consistent with the microarray analysis (Figure 3.2), *ada2*Δ mutant cells had dramatically reduced capsule compared to wild type (Figure 3.3). This phenotype was reversed by

complementation with the *ADA2* genomic coding sequence (*ada2*Δ::*ADA2*; see section 3.5).

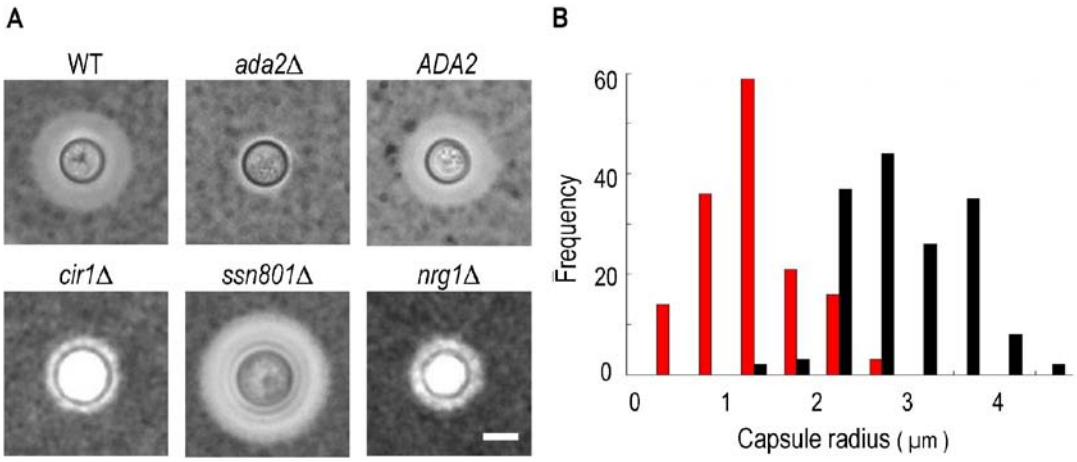

Figure 3.3. Cells lacking *ADA2* display reduced capsule size under inducing conditions. Panel A, negative staining with India ink of KN99α cells (WT), the indicated deletion strains, and the complemented *ada2*Δ mutant (*ADA2*). All images are at the same magnification. Scale bar, 5 μm. Panel B, histogram of capsule size for the *ada2*Δ mutant (red) and WT (black) populations. Capsule radius is represented in microns.

To facilitate comparison of *ada2*Δ to strains lacking other capsule regulators, we also deleted *CIR1*, *NRG1*, and *SSN801* in KN99α (see section 3.5). Consistent with earlier reports, the *ssn801*Δ capsule was enlarged, while the *cir1*Δ and *nrg1*Δ capsules were reduced, similar to the capsule produced by *ada2*Δ (Figure 3.3, panel A).

3.3.3 Cryptococcal Ada2 Is Localized to the Nucleus and Is Involved in Histone Acetylation

Having demonstrated that cryptococcal Ada2 influences capsule expansion, we proceeded to further investigate its role. Given the function of the SAGA complex in histone acetylation in *S. cerevisiae* [72], we expected that cryptococcal Ada2 would reside in the nucleus. To test our hypothesis, we integrated a hemagglutinin (HA) epitope-tag sequence at the 3' end of the *ADA2* genomic coding sequence and examined the localization of the tagged protein (Ada2-HA) by immunofluorescence microscopy. Consistent with the nuclear role of Ada2 in *S. cerevisiae*, the tagged cryptococcal protein colocalizes with nuclear DNA (Figure 3.4).

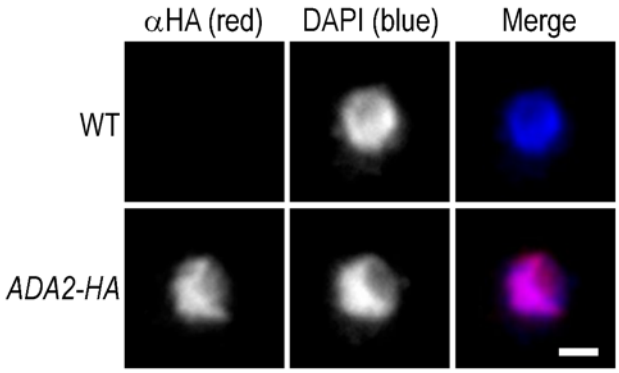

Figure 3.4. Cryptococcal Ada2 is localized to the nucleus. Wild type cells (WT) and cells modified to express HA epitope-tagged Ada2 from the native locus (*ADA2-HA*) were labeled with an antibody against HA (αHA, red), and counter-stained with DAPI (blue) to show the location of chromatin. All images were acquired at the same settings and are shown at the same magnification. Scale bar, 1 μm.

In *S. cerevisiae*, the SAGA complex activates transcription of stress-responsive genes by acetylating specific lysine residues at the N-terminal tails of histones H2B and H3 [72,75]. One of these modifications is the acetylation of lysine 9 of histone H3 (H3K9). To assess whether Ada2 is involved in similar histone acetylation in *C. neoformans*, we analyzed the abundance of acetylated H3K9 in the *ada2*Δ mutant by immunofluorescence microscopy using an antibody specific for this modification. We found that the fluorescence intensity of mutant cell nuclei was reduced by at least 50% compared to nuclei of both wild type and complemented cells (Figure 3.5), a result we confirmed on the population level by immunoblotting with the same antibody (not shown). In contrast, H4 acetylation, which is not SAGA specific [76,77], showed no difference between the *ada2*Δ mutant and either the wild type or complemented strains (not shown). These results demonstrate the role of CNAG_01626 in histone acetylation, likely in the context of *C. neoformans* SAGA, and strongly support our identification of this novel capsule regulator as the cryptococcal homolog of the *S. cerevisiae ADA2*.

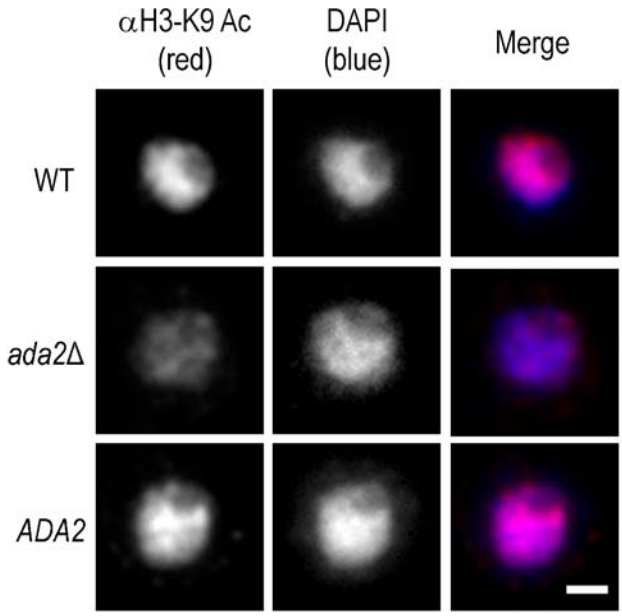

Figure 3.5. Histone acetylation is markedly reduced in the absence of Ada2. Shown are immunofluorescence micrographs of wild type (WT), *ada2*Δ, and complemented *ada2*Δ (*ADA2*) cells grown in capsule inducing conditions for 90 min and then probed with antibody to H3K9 (αH3-K9). All images were acquired at the same settings and are shown at the same magnification; scale bar, 1 μm.

3.3.4 Ada2 Functions in a Subset of Stress Response Pathways and in Mating

In *S. cerevisiae* and other fungi, the SAGA complex regulates the response to stress conditions such as elevated temperature, high salt concentration, and oxidative damage [78,79]. We found that the *ada2*Δ mutant grew normally compared to wild type on rich medium (YPD) at 30 °C (Figure 3.6). However, the mutant exhibited a

subtle growth impairment at 37 °C, and a moderate attenuation of growth at 39 °C. In all cases, the complemented strain behaved like wild type (Figure 3.6).

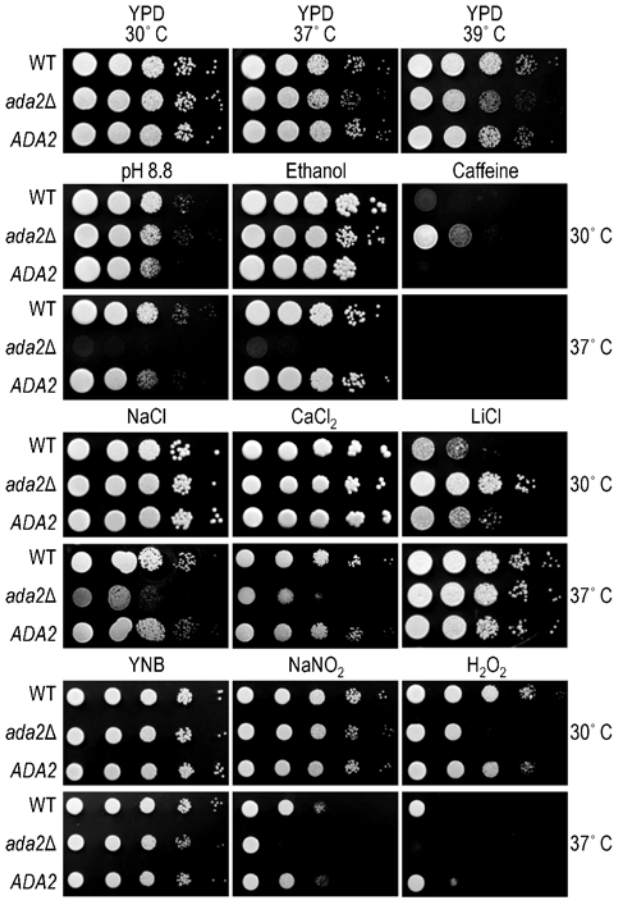

Figure 3.6. Ada2 is required for growth under certain stress conditions. Ten-fold serial dilutions of the indicated strains were grown in the conditions shown (see section 3.5 for details). Top panel, growth on rich medium (YPD) at the temperatures indicated above the images; middle panels (four rows of images), growth on YPD with the indicated stressor at the temperatures shown at the right; bottom panel (two rows of images), growth on minimal medium (YNB) or YNB with the indicated stressor at the temperatures shown at the right.

To further compare the phenotype of *ada2*Δ to known fungal SAGA mutants, we next tested a panel of stress conditions for their effect on growth of the wild type, mutant, and complemented strains at both 30 and 37 °C (Figure 3.6). We found that mutant cells were highly sensitive to alkaline pH, with no growth at pH 8.8 at 37 °C, while growth at physiological or acidic (5.5) pH was like that of wild type (not shown). Growth of *ada2*Δ at 37 °C was also abolished when 6% ethanol was included in the medium, in notable contrast to the growth of wild type cells under this condition, and was impaired at 0.4 M $CaCl_2$. Conditions that challenge cell integrity, including media containing calcofluor white (0.2%), congo red (0.5%), low levels of SDS (0.01%), or high sorbitol (2 M), had no effect on mutant growth (not shown). Similarly, KCl (1.2 M) and NaCl (0.4 M) did not alter growth (not shown), although high NaCl concentrations (1.2 M) did reduce growth at 37 °C compared to wild type (Figure 3.6). The *ada2*Δ mutant also showed enhanced growth on caffeine and LiCl at 30 °C, although this difference was not observed at the higher temperature tested (see section 3.4).

The ability of *C. neoformans* to withstand nitrosative and oxidative stress is required for the virulence of this yeast [80,81]. We therefore tested the effect of Ada2 absence on cryptococcal sensitivity to compounds that induce such stress. Growth of the *ada2*Δ mutant was not affected by $NaNO_2$ (0.5 mM) at 30 °C but exhibited a significant defect at 37 °C. The mutant was highly sensitive to oxidative stress (0.5 mM H_2O_2), with growth attenuated at 30 °C and absent at 37 °C. (Figure 3.6). We also examined the ability of this mutant to produce melanin, a feature of *C.*

neoformans that is associated with virulence [43]. We observed no difference in melanin production on medium containing L-3,4-dihydroxyphenylalanine (L-DOPA; not shown).

Finally, we tested the sensitivity of the *ada2*Δ strain to several pharmacological agents. These included fluconazole, amphotericin B, and flucytosine, all antifungal compounds used to treat cryptococcal infections. Growth in all cases was comparable to that of wild type, in contrast to the increased fluconazole sensitivity observed upon deletion of *ADA2* in *Candida albicans* [77]. We also tested the sensitivity of *ada2*Δ to FK506, a compound that inhibits calcineurin signaling. A *C. neoformans gcn5*Δ strain has been shown to be FK506 sensitive, suggesting a defect in this pathway [68] (see section 3.4); *ada2*Δ cells were even more sensitive to this compound (data not shown).

While back-crossing the *ada2*Δ mutant, we noticed that this strain was slow to filament. To investigate the potential role of Ada2 in cryptococcal sexual development, we crossed mating type a and α cells bearing the *ada2*Δ mutation to KN99a and KN99α cells and to each other (Figure 3.7). Deletion of *ADA2* in either mating type dramatically impaired the formation of dikaryotic filaments in unilateral crosses between the mutant and wild type. A bilateral cross between two *ada2*Δ mutants of opposite mating type showed no visible hyphal development even after 13 days, while the complemented strain behaved identically to wild type.

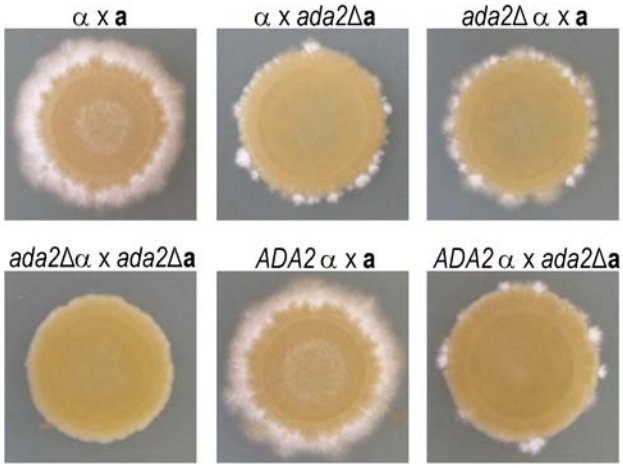

Figure 3.7. Ada2 is required for normal hyphal development. Wild type (no strain designation), *ada2*Δ, and complemented *ada2*Δ (*ADA2*) strains of opposite mating type were mixed and grown under conditions that induce mating (see section 3.5). Patches were imaged after 13 days.

3.3.5 Cryptococcal Ada2 is Essential for *C. neoformans* Virulence

The *ada2*Δ mutant displays a smaller capsule, demonstrates reduced resistance to oxidative and nitrosative stress, and grows more slowly at 37ºC compared to wild type. Based on these characteristics, we hypothesized that the mutant would also be attenuated for virulence. Indeed, we found that pulmonary growth of the *ada2*Δ mutant was impaired by almost 100-fold compared to the wild type and complemented strains in an inhalational mouse model of cryptococcosis (Figure 3.8, panel A), although it did grow slightly better than a completely acapsular mutant (*cap59*Δ). To pursue this observation, we conducted a survival study with the same

four strains. By three weeks post-inoculation, all mice infected with the wild type and complemented strains had succumbed to the infection (Figure 3.8, panel B). In contrast, mice infected with the *ada2*Δ or *cap59*Δ mutants remained healthy throughout the study, confirming the requirement for Ada2 in the virulence of this yeast.

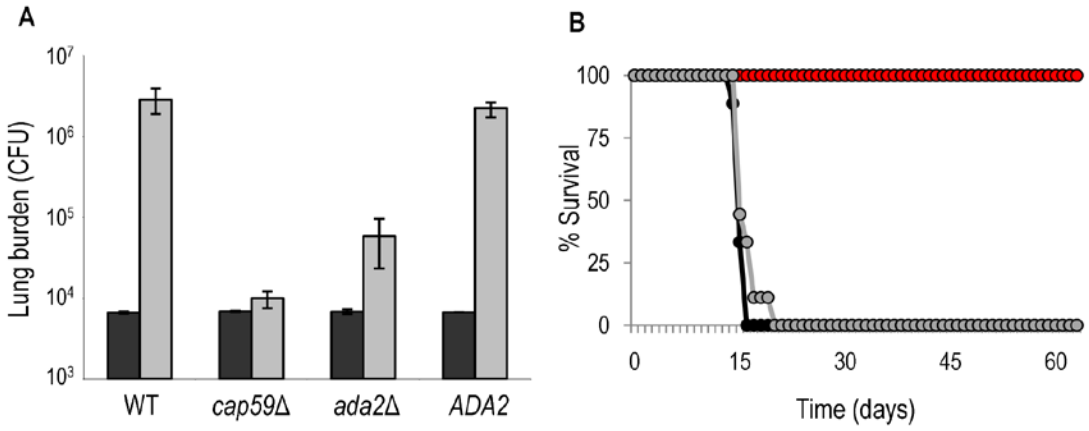

Figure 3.8. Ada2 is required for growth and virulence in mice. Panel A, C57Bl/6 mice were intranasally inoculated with 1.25×10^4 cells of the indicated strains, and total colony forming units (CFU) were isolated from the lungs after one hour (black bars) or one week (gray bars). The mean ± maximum and minimum is shown. Panel B, survival curve of A/Jcr mice that were similarly inoculated with 10^5 cells of wild type (black), *ada2*Δ (red), or complemented *ada2*Δ (gray). Like those infected with *ada2* Δ, all mice that were infected in the same study with *cap59* survived the entire period (not shown).

3.3.6 Cryptococcal Ada2 Transcriptionally Regulates Genes Required for Host Adaptation

To identify the genes and processes regulated by Ada2, we used RNA-Seq to perform transcriptome analysis of the *ada2*Δ mutant and wild type cells cultured in either capsule-inducing or capsule non-inducing conditions (see section 3.5). The majority (92%) of the resulting short reads mapped to the *C. neoformans* serotype A reference sequence [82], indicating the excellent quality of the data. The average 300-fold coverage of the cryptococcal transcripts we obtained in these studies allowed confident sequence identification, and will help improve annotation of the *C. neoformans* genome.

Gene expression analysis revealed 460 genes that were differentially expressed in the *ada2*Δ mutant compared to wild type under the capsule inducing condition; 675 genes were differentially expressed between the two strains under the capsule non-inducing condition. We examined the genes whose expression was significantly affected in one or both conditions. Most of these (73%) were regulated in a sign consistent manner in the two conditions (*e.g.* if gene expression was reduced in the *ada2*Δ mutant in non-inducing conditions it was also reduced in the mutant in inducing conditions), although the magnitude of changes did vary. Gene ontology (GO) analysis (see section 3.5) indicated that processes significantly enriched in the response to loss of Ada2 included ribosomal protein synthesis, sugar transport, and carbohydrate metabolism.

Consistent with the filamentation defect we observed in the *ada2*Δ mutant (Figure 3.7), we noted several genes downstream of Ada2 that are involved in cryptococcal sexual development (Table 3.1). Two mating type-specific genes (encoding the homeodomain regulator, Sxi1α, and the pheromone receptor, Ste3α) showed decreased expression in the *ada2*Δ mutant. A variety of genes that are independent of mating type but are implicated in the pheromone response pathway were also found to respond to loss of Ada2 (Table 3.1).

Our initial interest in Ada2 was stimulated by its importance in capsule synthesis. In the *ada2*Δ mutant, we observed a reduction in transcript abundance for a number of genes that, when deleted, yield small capsules (Table 3.1). These observations are consistent with the hypocapsular and avirulent phenotypes of the *ada2*Δ mutant. The *ada2*Δ mutant also showed reduced expression for genes involved in oxidative stress; this agrees with the hypersensitivity to oxidative stress observed in the mutant and may also contribute to the avirulent phenotype. Expression of two genes (*BLP1* and *GAT204*), which have recently been implicated in capsule-independent mechanisms of cryptococcal virulence [62], was also reduced in the *ada2*Δ mutant (see section 3.4).

		Gene Name			Fold change (log2) *ada2*Δ vs WT	
	Broad ID	*C. neo*	*S. cer*	Description	uninduced	induced
Mating	CNAG_06808	*STE3*	*STE3*	pheromone receptor Ste3	**-5.35**	**-7.71**
	CNAG_03137		*SGV1*	Ste11 protein kinase	3.63	**-0.86**
	CNAG_04323		*PRM10*	DUF1212 family protein	**-0.89**	-0.73
	CNAG_06814	*SXI1*		Sxi1	**-0.79**	-0.49
	CNAG_04755		*BCK1*	Ste/Ste11 protein kinase	**-0.59**	-0.23
	CNAG_03706		*GLC7*	phosphatase PP1	**-0.32**	-0.14
	CNAG_02981		*SIN3*	Sin3 protein	**0.47**	0.46
	CNAG_02375		*FIG4*	phosphatase	**0.84**	0.46
	CNAG_05752		*KAR3*	kinesin	**0.81**	0.56
Capsule	CNAG_03644	*CAS3*		Cas3	-4.21	**-7.69**
	CNAG_05264	*NSTA*	*YJL216C*	alpha-amylase AmyA	**-2.06**	**-1.97**
	CNAG_03438	*HXT1*	*HXT2*	hexose transporter	**-1.15**	**-1.35**
	CNAG_02797	*CPL1*		pria protein	**-1.01**	-0.65
	CNAG_07937	*CAS1*		O-acetyltransferase	**-0.60**	-0.41
	CNAG_04312	*MAN1*	*PMI40*	mannose-6-phosphate isomerase	**-0.57**	-0.28
	CNAG_07554	*CAP10*		capsule associated protein	**0.43**	-0.15
	CNAG_00124	*CAS32*		Cas32	**-0.50**	0.19
	CNAG_05581	*CHS3*	*CHS3*	chitin synthase 4	0.15	**0.61**
	CNAG_05139	*UGT1*		Ugt1	**0.36**	0.73
	CNAG_02138	*CAS4*	*DNA2*	DNA replication helicase dna2	**0.64**	**1.40**
Oxidation	CNAG_05265		*RCK1*	hypothetical protein	2.28	**-5.91**
	CNAG_04415		*YJR096W*	oxidoreductase	**-0.76**	**-3.82**
	CNAG_05027		*FMS1*	amine oxidase	-0.64	**-3.81**
	CNAG_04508		*GRX4*	conserved hypothetical protein	**-2.21**	-3.64
	CNAG_03848		*GRX7*	glutathione transferase	**-3.21**	**-2.18**
	CNAG_03199		*GRX3*	oxidoreductase superfamily	-0.14	**-1.26**
	CNAG_03936		*PST2*	cytoplasmic protein	**-0.71**	-0.74
	CNAG_01005		*GRX1*	glutathione transferase	**-0.80**	-0.46
	CNAG_00581		*PEP4*	endopeptidase	**-0.30**	-0.44
	CNAG_02859		*POS5*	NADH kinase	**-4.08**	-0.13
Antiphagocytosis	CNAG_06762	*GAT204*	*GAT2*	conserved hypothetical protein	-1.28	**-1.44**
	CNAG_06346	*BLP1*		conserved hypothetical protein	-2.70	**-0.86**

Table 3.1. Genes downstream of Ada2 implicated in processes related to mating or virulence. Genes listed were identified by differential expression analysis of mutant

versus wild type and found to be significantly changed in either capsule non-inducing or inducing conditions. Fold change of mutant versus wild type is indicated in the rightmost columns; bold font indicates statistically significant change. Yellow indicates a positive fold change relative to wild type; blue indicates a negative fold change.

3.3.7 RNA-Seq Analysis Suggests New Relationships in Capsule Regulation

To place Ada2 in the context of the broader capsule regulation network, we performed RNA-Seq analysis on mutants that lack the transcriptional regulators Cir1 and Nrg1. We chose these transcription factors because, like the *ada2*Δ mutant, both the *cir1*Δ and the *nrg1*Δ mutants are hypocapsular, demonstrate attenuated avirulence, and exhibit defects in mating. We identified 1265 genes that were differentially expressed in the *nrg1*Δ mutant compared to wild type under the capsule inducing condition and 1084 under the non-inducing condition. For the *cir1*Δ mutant these values were 1257 and 529, respectively.

Cryptococcal sexual development is regulated by Cir1 and Nrg1 [61,58], as well as by Ada2 (Figure 3.7). To identify common regulatory targets shared by these three transcription factors, we examined the gene expression data from the *ada2*Δ, *cir1*Δ, and *nrg1*Δ mutants. Among genes previously implicated in cryptococcal sexual development, we found that only *SXI1*α was downstream of all three regulators, with

its transcription reduced in *nrg1*Δ and *ada2*Δ but increased in *cir1*Δ. Transcription of the pheromone receptor *STE3*α was similarly reduced in *ada2*Δ and elevated in *cir1*Δ although it was not significantly changed in *nrg1*Δ. Genes regulated by Nrg1 included the cell type-specific p21-activated protein kinase *STE20*α, as well as other mating type-independent genes that are involved in sexual development, but these were not regulated by Ada2 or Cir1.

By comparing mutants generated in the same strain background and grown in the same conditions, we were able to confidently identify capsule-implicated genes that are downstream of Cir1 or Nrg1, some of which are also regulated by Ada2. For example, Nrg1 and Ada2 share downstream targets that include *CAS4*, *CAS32*, *CPL1*, *MAN1*, *NSTA*, and *CHS3*. Similarly, *CAP10*, *CAS1*, *CAS4*, and *CPL1* are all downstream of both Cir1 and Ada2. Notably, *CAS4* and *CPL1* are shared targets of all three regulators (Ada2, Nrg1 and Cir1).

In addition to genes that are likely to be directly involved in capsule biosynthesis, we found many genes whose expression was affected by the loss of Cir1 or Nrg1 that are involved in regulating capsule formation. For example, the *nrg1*Δ mutant showed altered transcription of genes in the cAMP pathway, including increased transcription of *RIM101* and decreased transcription of *PKA2* and *PDE2*. Consistent with previous reports [60], we also observed altered transcript levels in the *cir1*Δ mutant that correspond to a number of pH-specific pathway genes, including *RIM9*

and *RIM20*. The latter gene product is involved in proteolytic activation of Rim101 [59].

Finally, we discovered that Cir1 and Nrg1 regulate the expression of two HOG pathway genes: absence of either protein led to reduced transcription of *HOG1* and increased transcription of *PBS2*. Additionally, both Cir1 and Nrg1 appeared to enhance the expression of *TUP1* [83], which encodes a regulator that may operate in the HOG pathway. Interestingly, data from a previous microarray study indicated that *ADA2* (at that time uncharacterized) increased in expression upon deletion of HOG pathway members (*HOG1* or *SSK1*) [84] (see section 3.4).

3.3.8 ChIP-Seq Indicates Genes Directly Regulated by Ada2-dependent Histone Acetylation

Ada2 is required for the majority of H3K9 acetylation in *C. neoformans* (Figure 3.5 and immunoblotting data not shown). We reasoned that localizing Ada2-dependent occurrences of this modification would lead us to genes that are directly regulated by Ada2. We therefore used chromatin immunoprecipitation (ChIP) to isolate DNA directly associated with acetylated H3K9 in *ada2*Δ and wild type cells that we could analyze by short read sequencing (ChIP-Seq).

We obtained 84 million short reads from our ChIP-Seq studies, which we aligned to the serotype A reference sequence and analyzed to identify genomic regions with

statistically significant coverage (“peaks”) in IP samples compared to input DNA. From triplicate experiments, we identified an average of 2014 peaks in wild type cells, compared to only 364 in *ada2*Δ. This 82% reduction is consistent with our earlier observations on the Ada2-dependence of most H3K9 modification (Figure 3.5). Consistent with H3K9 acetylation in *S. cerevisiae* [85,86], the majority of the peaks identified in wild type (75%) were within 500 bp of at least one transcription start site (TSS) as annotated [82]. Most peaks in wild type were also located in the 5’ region immediately downstream of the TSS, with a strong depletion near the TSS and a modest enrichment upstream of the TSS (Figure 3.9, black bars in panel A). In contrast, only 28% of peaks in *ada2*Δ were within 500 bp of a TSS and almost none of these were downstream of the TSS (Figure 3.9, red bars in panel A). Thus, not only is histone acetylation in this mutant depleted throughout the genome, the pattern of acetylation is also changed, with the most dramatic depletion occurring in the region immediately downstream of the transcription start site.

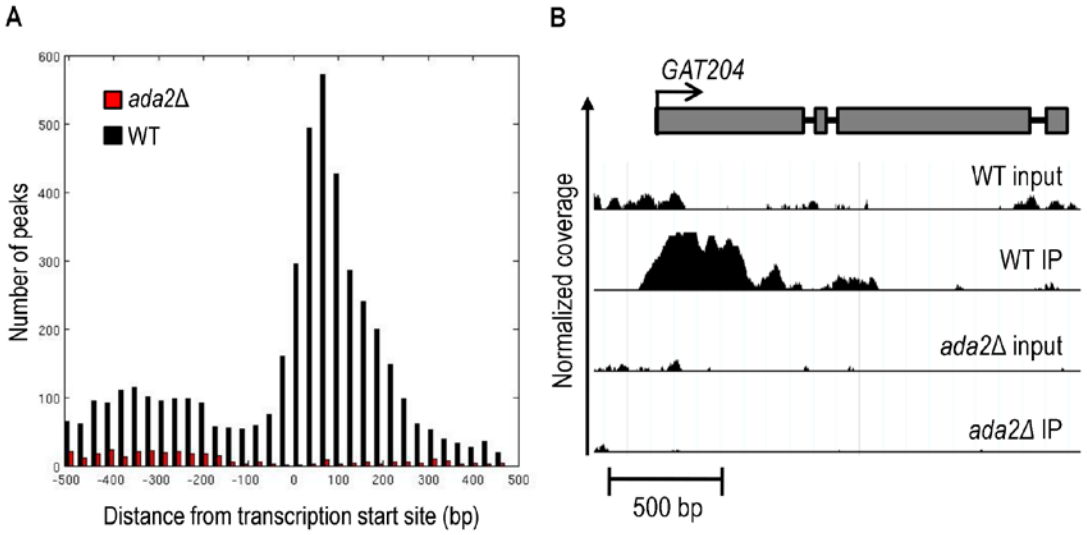

Figure 3.9. Ada2-dependent acetylation of H3K9 is enriched near gene transcription start sites. ChIP-Seq was performed on wild type (WT) and the *ada2*Δ mutant to identify genes located in the proximity of acetylated H3K9. Panel A, a histogram of peaks that occur within 500 bp of the transcription start site of all identified genes [82]. Panel B, an example of ChIP-Seq data aligned to a gene model of *GAT204*, which was identified as Ada2-dependent by both ChIP-Seq and RNA-Seq. The y-axis represents normalized coverage (reads per million mapped) for samples defined in the text. Coverage is shown for 2 standard deviations above the mean input sample coverage and above. Note that the input DNA profiles are similar for WT and mutant cells, while specific H3K9 associated sequences show TSS-associated peaks only in the WT.

The loss of histone acetylation in *ada2*Δ cells suggested Ada2-dependent transcriptional activation at specific loci (see example in Figure 3.9, panel B). We anticipated that some of these genes would also show reduced transcription by RNA-

Seq in the *ada2*Δ mutant; this was indeed the case ($p < 0.003$). In contrast, we found no such relationship for genes with increased transcription in *ada2*Δ (i.e., genes that are directly or indirectly repressed by Ada2; $p > 0.99$), consistent with the generally activating function of the SAGA complex. Overall, we found that genes differentially expressed in the *ada2*Δ deletion strain that also lost histone acetylation near the TSS were twice as likely to exhibit reduced transcriptional abundance as genes that did not lose histone acetylation (Figure 3.10).

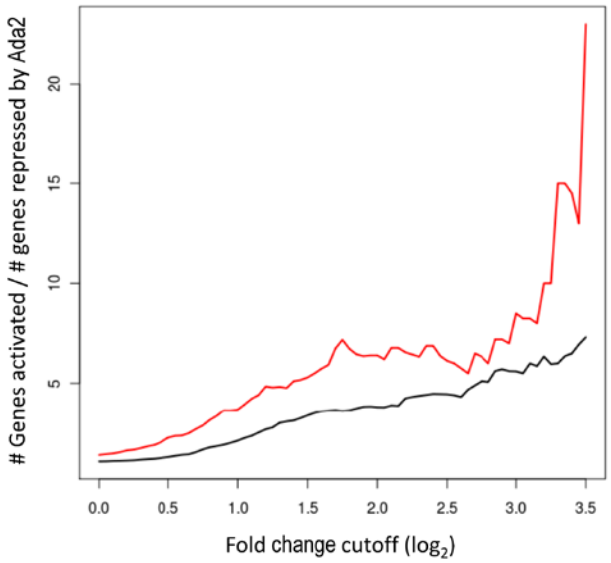

Figure 3.10. Ada2-dependent loss of H3-K9 acetylation is associated with activation. The ratio of Ada2 activated to Ada2 repressed genes (y-axis) is determined by an analysis of differential gene expression from RNA-Seq data comparing *ada2*Δ and wild type strains. The cutoff to be counted as differentially expressed is varied from 0 fold to ~12 fold (x-axis). Ada2 activated genes exhibit a negative fold change greater than the cutoff and Ada2 repressed genes exhibit a positive fold change greater than the cutoff. Genes that lose neighboring H3-K9 acetylation near their

TSS in the *ada2*Δ mutant are shown in red, genes that shown unchanged H3-K9 acetylation are shown in black.

We were particularly interested in genes that were activated by Ada2 according to our RNA-Seq analysis and also showed Ada2 dependent H3K9 acetylation in our ChIP-seq analysis. This set is significantly enriched for genes that are directly regulated by Gat201 ($p < 0.0001$), including *BLP1* and *GAT204* [62]. The genes implicated by both RNA-seq and ChIP-seq also include a number with known capsule phenotypes, such as *CPL1*, *HXT1*, *STE3*α, and *UGT1* (see section 3.4).

3.4 Discussion

We analyzed gene expression in *C. neoformans* yeast cells cultured over a diverse set of growth conditions that stimulate capsule production to varying degrees and identified a transcriptional signature of capsule formation. Gene ontology (GO) analysis shows that this signature is enriched for genes involved in stress response, as expected from the conditions we used to induce capsule formation. The signature also contains a significant number of genes that have previously been implicated in capsule regulation; the expression of most of these correlates with capsule in a manner consistent with the null phenotype. The phosphodiesterases Pde1 and Pde2 are exceptions to this pattern: their transcript levels correlated positively with capsule size, while their disruption increases capsule size [71]. Pde1 and Pde2 hydrolyze cAMP to AMP and thereby inhibit the cAMP-dependent activation of regulators

known to stimulate capsule formation. Elevated levels of cAMP occurring under capsule inducing conditions may lead to elevated transcription of *PDE*1 and *PDE*2, which would ultimately attenuate the cAMP signal. Feedback inhibition of cAMP signaling via post-translational activation of phosphodiesterases has been documented in both *S. cerevisiae* and *C. neoformans* [87,71].

One sequence in the transcriptional signature that correlated significantly with capsule size (Figure 3.2) encoded the putative transcriptional regulator, Ada2. This protein has been characterized most extensively for its role within the SAGA complex, which broadly regulates the transcription of genes involved in stress response and development in multiple organisms [69]. This pattern holds true for *C. neoformans*, based on the increased sensitivity of mutants that lack either *ADA2* (this work) or *GCN5* [68] to reactive oxygen species, ethanol, alkaline pH, elevated temperature, and $CaCl_2$ (Figure 3.6). All of these sensitivities are shared by *S. cerevisiae* SAGA mutants [79,88], and the last two also are shared by SAGA mutants in other fungi including *C. albicans*, S. *pombe* and *S. kluyveri* [77,79].

Despite many conserved functions of the SAGA complex across fungal species, several phenotypes of *ada2* mutants in *C. neoformans* differ markedly from those observed in other fungi, perhaps reflecting the specific evolutionary pressures of the cryptococcal niche. Whereas *ada2* mutants in *C. neoformans* display increased caffeine resistance (Figure 3.6), for example, disruption of SAGA components in *S. cerevisiae*, *S. pombe* and *S. kluyveri* has the opposite effect. Also, *ada2* mutants in *C.*

neoformans show an increase in LiCl resistance but no change in KCl resistance (Figure 3.6), while other fungi defective in SAGA typically exhibit normal growth in LiCl but are KCl sensitive relative to wild type [79]. Interestingly, *ada2* mutants in *C. neoformans* have wild type sensitivity to fluconazole in contrast to *ada2* mutants in *C. albicans*, which have increased sensitivity [77]. Finally, *C. neoformans ada2*Δ differs from other fungi in its regulation of sexual development. In *S. pombe*, the *ada2*Δ mutant is enhanced for mating, probably through a mechanism that does not directly involve histone acetylation [89]. In the *C. neoformans ada2*Δ strain, we instead found dramatically decreased sexual development (Figure 3.7), reduced transcript abundance of the pheromone receptor *STE3α*, and loss of H3K9 acetylation at the *STE3α* promoter. These results suggest that these two fungi differ in both the direction and the mechanism of Ada2's influence on sexual development.

Recently, another component of the SAGA complex, Gcn5, was shown to play a role in capsule formation and virulence in *C. neoformans* [68]. H99 cells lacking Gcn5, like our mutant lacking Ada2, are hypocapsular and hypovirulent. To compare the roles of these proteins, we examined genes that are differentially expressed by *ada2*Δ and *gcn5*Δ upon growth in DMEM, using our RNA-Seq data for *ada2*Δ and published microarray data sets for *gcn5*Δ [68]. We found a significant overlap in the sets of genes whose expression is affected by each mutation ($p < 1e\text{-}5$), supporting the idea that some genes are jointly regulated by Gcn5 and Ada2, probably due to the coordinated role of these proteins in SAGA-mediated histone acetylation.

In addition to shared characteristics, we observed important differences between the *ada2*Δ and *gcn5*Δ mutants at both the phenotypic and transcriptional levels. The *ada2*Δ mutant is more resistant to high temperature, showing ~10-fold growth inhibition on rich medium at 39 °C compared to wild type (Figure 3.6), a condition where *gcn5*Δ does not grow at all [68]. In contrast, *ada2*Δ is more sensitive than *gcn5*Δ to the calcineurin inhibitor FK506. (The minimal inhibitory concentration (MIC) for *gcn5*Δ is 10-fold below that of its H99 parent [68], while the MIC for *ada2*Δ (performed as in [68]) is at least 67-fold below that of KN99α; data not shown.). We also found that expression of both *STE3*α and *SXI1*α responds to the loss of Ada2 (Table 3.1), whereas no sexual development genes have been reported to be downstream of Gcn5 [68]. Consistent with this difference, *ada2*Δ is severely defective in filamentation (Figure 3.7) while *gcn5*Δ filaments normally (T. R. O'Meara and J. A. Alspaugh, personal communication). Furthermore, two genes involved in the recently described 'antiphagocytic response' [62], *GAT204* and *BLP1*, showed a loss of both H3K9 acetylation (Figure 3.9, panel B) and expression (Table 3.1) in *ada2*Δ but no change in expression in *gcn5*Δ [68]. It will be interesting to determine whether these transcriptional differences manifest phenotypically.

The phenotypic differences between *ada2*Δ and *gcn5*Δ may be due to Gcn5-independent functions of Ada2 in *C. neoformans*. Acetylation at some loci may rely on Ada2 partnering with a histone acetyltransferase (HAT) other than Gcn5, or it may be that the regulation of these loci is independent of acetylation altogether. For example, in *S. cerevisiae* Ada2 regulates gene silencing by preventing the spread of

repressive chromatin [90]. Such mechanisms remain to be investigated in *C. neoformans*. Given the importance of SAGA in virulence, the roles of Ada2, Gcn5 and other SAGA subunits in *C. neoformans* biology are worthy of further investigation.

After identifying Ada2 as a novel regulator of capsule, we sought to identify elements downstream of it in the capsule regulatory network. To do this, we performed RNA-Seq on the *ada2*Δ mutant and wild type strains, considering genes differentially expressed between these two strains to be downstream of Ada2. To identify probable direct targets of Ada2, we performed ChIP-Seq using antibodies specific for H3K9 acetylation, comparing the *ada2*Δ mutant and wild type strains. We reasoned that genes that lose histone acetylation near their transcription start sites in the *ada2*Δ mutant are likely direct targets of Ada2 via the SAGA complex or another histone acetyltransferase (HAT) complex involving Ada2.

The *ada2*Δ mutant strain revealed a dramatically altered landscape of H3K9 acetylation compared to the wild type, with more than an 80% reduction in acetylated sites across the genome and even greater reduction around transcription start sites (Figure 3.9). This nearly total loss of H3K9 acetylation in the *ada2*Δ mutant is consistent with the established global HAT activity of SAGA in *S. cerevisiae* [91,92]. In contrast to its broad histone modification activity, SAGA only influences expression of 10% of *S. cerevisiae* genes [76]. RNA-Seq analysis of the *ada2*Δ mutant strain revealed that Ada2 influences transcription of 14% of the genes

in *C. neoformans*, indicating that the transcriptional regulatory role of SAGA in *C. neoformans* is also locus specific. ChIP-Seq data further suggest that Ada2 exerts the minority of its influence through direct regulation: only 3% of cryptococcal genes exhibit both altered H3K9 acetylation and expression in *ada2*Δ cells, while 11% exhibit altered expression only. This large indirect response could be mediated in part via the 8 putative transcription factors that Ada2 directly regulates as evidenced by our studies.

Consistent with the activating role of SAGA, the set of genes with reduced expression in *ada2*Δ was significantly enriched for those that lost H3K9 acetylation. (In contrast, genes with increased expression in *ada2* Δ showed no significant overlap with those that lost H3K9 acetylation.) Some genes, including the capsule-implicated gene *UGT1*, showed increased expression together with loss of H3K9 acetylation in the *ada2*Δ mutant, perhaps because H3K9 acetylation at certain loci makes repressor binding sites more accessible. Alternatively, these genes may be directly activated by Ada2 through H3K9 acetylation yet also indirectly repressed by Ada2, which could yield net repression.

We observed phenotypic changes in the *ada2*Δ mutant in sexual development, capsule formation, stress response, and virulence; we also found genes with known roles in these processes to be directly regulated by Ada2 as evidenced by ChIP-Seq and RNA-Seq. For example, we found that Ada2 directly regulates genes encoding proteins implicated in capsule formation, including *HXT1* [93], *CPL1* [63], and

UGT1 [94], consistent with the capsule defect of the *ada2*Δ mutant (Fig 3). We also identified the gene encoding pheromone receptor Ste3 as a direct target of Ada2 in the mating type α (*MAT*α) cells used in these studies, consistent with the observed filamentation defect in *ada2*Δ (Figure 3.7). Ste3 has also been implicated in mating in *MAT*a [66]. Ste3a has further been shown to regulate virulence factors including titan cell [46] and capsule formation [66], although no such relation has been reported for Ste3α. If Ada2 also regulates Ste3a then it may additionally influence capsule via this pathway in *MAT*a cells. Future studies of *MAT*a *ada2*Δ mutants will be needed to address this possibility.

Gat201 is a GATA family transcription factor reported to act as a positive regulator of capsule [63]. Interestingly, we observe a significant overlap in the genes that are directly activated by Ada2 (as shown by ChIP-Seq) and those that are direct targets of Gat201 (by ChIP-chip [62]), including the antiphagocytic genes *BLP1* and *GAT204*. Since the SAGA complex typically works in concert with other transcription factors, this suggests that Ada2 may work with Gat201 to activate transcription. It may be that Gat201 recruits Ada2 in the context of SAGA for these purposes. Alternatively, another factor may recruit the SAGA complex, which then enables Gat201 to bind.

To explore the interplay between regulatory pathways we considered two transcription factors, Cir1 [61] and Nrg1 [58], which like Ada2 enhance both capsule and mating responses. In the set of genes regulated by Cir1 and Nrg1, we identified

two that encode proteins in the HOG pathway, Hog1 and Pbs2; both Cir1 and Nrg1 transcriptionally repress Pbs2 and activate Hog1. Cells lacking either Pbs2 or Hog1 show increased capsule formation and sexual development [64]. Furthermore, both *ADA2* and *GCN5* were shown in earlier work to be transcriptionally repressed by Hog1 under nutrient rich conditions [84]. This observation, in conjunction with our data, suggests that the HOG pathway may regulate capsule and mating via Ada2 (Figure 3.11). Our transcriptional analysis suggests that Nrg1 and Cir1 operate on the HOG pathway through a shared incoherent feed-forward loop, by transcriptionally activating Hog1 and simultaneously repressing Pbs2. In nutrient rich conditions, Hog1 is constitutively phosphorylated by Pbs2 and represses mating and capsule. The logic of this circuit implies that in capsule inducing conditions Cir1 and Nrg1 repress transcription of *PBS2*; this leads to reduced levels of phosphorylated Hog1, thus derepressing *ADA2* transcription and enhancing capsule formation. Simultaneously transcription of *HOG1* is increased, leading to an even greater abundance of unphosphorylated Hog1. This increase in Pbs2 substrate may allow rapid restoration of the transcriptional repression of Ada2 once the environmental cues for capsule induction are no longer present. Although transcript levels of *ADA2* were not significantly altered in the *nrg1*Δ and *cir1*Δ mutants at the 90-minute time point that we tested, *ADA2* expression may be affected by these mutations at later time points. Future studies will also be needed to determine whether the influences of Cir1 and Nrg1 on *PBS2* and *HOG1* result from direct or indirect regulation, and to better characterize the exact structure and function of this

hypothesized regulatory circuit. This model, rich in testable hypotheses, illustrates the power of combining RNA-Seq and ChIP-Seq data in an integrated analysis.

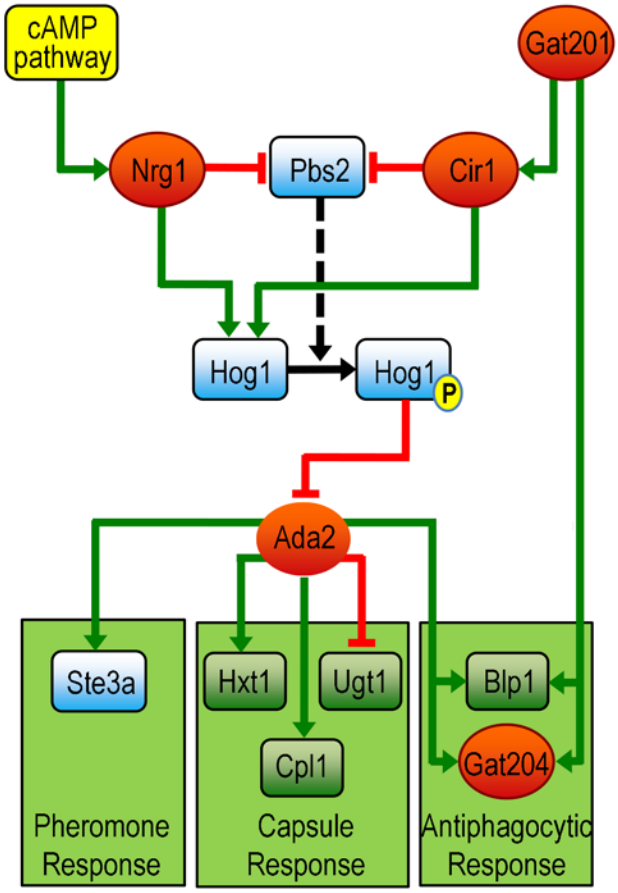

Figure 3.11. A model of Ada2 within the broader network of capsule, mating, and antiphagocytic responses. Links from Cir1 and Nrg1 are supported by RNA-Seq data presented here. Links from Hog1 and Pbs2 are supported by published microarray data [84]; links from Gat201 are supported by published data from microarrays and ChIP-chip [62]; and links from Ada2 are supported by RNA-Seq and ChIP-chip data presented here. Red ovals, transcription factors; blue rounded rectangles, signaling proteins; green rounded rectangles, other proteins; green lines, stimulation of

transcription; red lines, inhibition of transcription; solid black arrow, phosphorylation; P, phosphate; dashed arrow, catalysis by Pbs2.

Our identification of Ada2 in the capsule transcriptional network validates our strategy for probing capsule regulation and suggests that it may be valuable in studying the regulation of other processes that are important in microbial pathogenesis. These studies also lead in numerous exciting directions for the future. Our parallel comparison of multiple mutants in the same strain background and growth conditions has allowed us to identify previously unobserved relationships among capsule regulators, which we look forward to testing. Our analysis of the transcriptional signature of capsule induction also suggests multiple potential transcription factors that can be pursued to further probe the complex confluence of pathways that lead to capsule synthesis, and our implementation of ChIP-Seq in *C. neoformans* demonstrates a high-resolution way for differentiating direct from indirect regulatory relationships. Overall, our work highlights the power of integrative transcriptome analysis to dissect regulatory networks in *C. neoformans* and beyond.

3.5 Materials and Methods

3.5.1 Ethics Statement

All animal studies were reviewed and approved by the Animal Studies Committee of Washington University School of Medicine and conducted according to the National Institutes of Health guidelines for housing and care of laboratory animals.

3.5.2 Materials

All chemicals were from Sigma, primers were from Invitrogen, and restriction enzymes were from New England Biolabs unless otherwise noted. All kits and enzymes were used according to manufacturer recommendations unless otherwise specified.

3.5.3 Strains and Growth Conditions

All strains used in this study are capsule serotype A, which causes the majority of illness in immunocompromised patients [95]. Microarray experiments to identify the transcriptional signature of capsule were performed with *C. neoformans* H99 and mutants were constructed in *C. neoformans* KN99. All cells were grown with continuous shaking (230 rpm) at 30°C in YPD medium (1% w/v yeast extract, 2% w/v peptone, 2% w/v glucose), or at 30ºC on agar plates (YPD medium with 2% w/v

agar). As appropriate, media were supplemented with either 100 µg/ml of nourseothricin (from Werner BioAgents) or 100 µg/ml of Geneticin (G418; from Invitrogen). Genetic crosses were performed at room temperature (RT) in the dark on V8 agar plates (5% v/v V8 juice, 0.05% w/v KH_2PO_4 pH 5, 4% w/v agar) as described [74]. To induce expression of genes involved in capsule formation, cells cultured overnight in YPD were collected by centrifugation, washed in DMEM, and adjusted to 4×10^7 cells/ml in DMEM. This cell suspension was first incubated at 30ºC in room air for 2 hr, then shifted to 37ºC with 5% CO_2 for 1.5 hr. Conditions used for phenotypic testing of mutants are detailed in Text S1.

3.5.4 RNA Isolation

Approximately 2×10^8 cells were collected by centrifugation, suspended in TRIzol reagent (from Invitrogen), and subjected to mechanical lysis by bead beating at 4ºC with 0.5-mm glass beads for 1 min, followed by a 2-min rest, for a total of 5 cycles. Following lysis, total RNA was extracted according to the manufacturer's instructions. Residual DNA was removed from the RNA preparation by treatment with the Turbo DNA-free kit (from Ambion) according to the manufacturer's instructions.

3.5.5 Microarray Experiments

H99 cells were cultured overnight at 37ºC in the following eight conditions: low iron medium with or without both 500 mM ethylenediaminetetraacetic acid (EDTA) and 10 mM bathophenanthroline disulfonate (BPDS); phosphate-buffered saline (PBS) with or without 10% v/v fetal bovine serum; Dulbecco's Modified Eagle's Medium (from Sigma) in room air or 5% CO_2; and Littman's medium [51] with either 0.01 µg/ml or 1 µg/ml thiamine. All experiments were performed in triplicate. Total RNA was isolated from each culture and hybridized to a *C. neoformans* serotype A/D microarray against a shared reference pool of RNA as described [96]. Slides were scanned on a Perkin-Elmer ScanArray Express HT scanner to measure Cy3 and Cy5 fluorescence as described [96]. Normalization of the raw spot intensities was performed using LIMMA [97]. Normalization was performed using normexp with an offset of 50 followed by Loess and values for replicate probes on the array were averaged to represent expression of the associated gene. The correlation between gene expression and capsule radius (which was measured for each sample at the time of RNA isolation) was assessed using SAM [98] and statistical significance was calculated using a false discovery threshold of 5%. A hypergeometric test was applied to determine the enrichment of capsule-implicated genes (genes whose mutation yields an alteration in capsule size or morphology) in the positively and negatively correlating sets of genes. The complete array data set is available at GEO accession number GSE31911.

3.5.6 Strain Construction

The *C. neoformans* H99 reference sequence was accessed through the Fungal Genome Initiative database at the Broad Institute of MIT and Harvard available at http://www.broadinstitute.org/science/projects/projects. Cryptococcal genomic DNA was isolated as described [99] and a split-marker approach [100] was used to replace each genomic coding sequence of interest with a nourseothricin resistance marker (*NAT*) by homologous recombination. Each mutant was also labeled with a unique signature tag by incorporating a 13-bp tag sequence and an 18-bp priming site (5' - AGAGACCTCGTGGACATC - 3') immediately downstream of *NAT*. We also used the split-marker gene replacement approach to introduce a single copy of the hemagglutinin (HA) epitope-tag sequence at the 3' end of the *ADA2* genomic coding sequence.

3.5.7 Capsule Induction and Quantitation of Capsule Size

Cells cultured in YPD were washed extensively in DMEM, then adjusted to 10^6 cells/ml in DMEM and incubated for 24 hours at 37ºC with 5% CO_2. Capsules were visualized by negative staining with India ink, and a minimum of 100 randomly chosen cells were imaged with identical acquisition settings on a Zeiss Axioskop 2 MOT Plus wide-field fluorescence microscope. Capsule radius was calculated as half the difference between the capsule diameter and the diameter of the cell body.

3.5.8 Immunofluorescence Microscopy

Cells were cultured overnight in YPD, and the expression of genes involved in capsule formation was induced as described above. Cells were then collected by centrifugation, washed in PBS, adjusted to 3×10^8 cells/ml in 4% w/v formaldehyde buffered in PBS, and incubated for 1 hr with rotation. Fixed cells were collected by centrifugation (1 min, 400 × g), washed extensively in PBS, adjusted to 3×10^8 cells/ml in Lysis Buffer (50 mM sodium citrate pH 6.0, 1 M sorbitol, 35 mM β-mercaptoethanol) plus 25 mg/ml Lysing Enzymes (from Trichoderma harzianum), and incubated for 1 hr at 30°C. Digested cells were collected by centrifugation (3 min, 400 × g), washed with HS Buffer (100 mM HEPES pH 7.5, 1 M sorbitol), and resuspended in 100-200 µl of HS Buffer. The cell suspension was spotted in 20-µl aliquots on a glass microscope slide coated with 0.1% w/v poly-L-lysine and incubated for 20 min at RT.

All subsequent treatments and washes were performed by the application of 20-µl volumes and incubation at RT, and were followed by aspiration. The slides were first treated with HS Buffer containing 1% v/v Triton X-100 and incubated for 10 min; they were then washed with PBS and treated with Blocking Buffer (5% v/v goat serum, 0.02% v/v Tween-20 in PBS) for 1 hr. Cells were next labeled with either a high-affinity rat anti-HA monoclonal antibody (0.2 µg/ml in Blocking Buffer; from Roche), a rabbit anti-acetyl-Histone H3 polyclonal antibody (0.5 µg/ml in Blocking Buffer; from Millipore), a rabbit anti-acetyl-Histone H4 polyclonal antibody (1

µg/ml in Blocking Buffer; from Millipore), or Blocking Buffer alone overnight in a moist chamber at 4ºC. Cells were then washed with Blocking Buffer, and treated for 1 hr in the dark with either Alexa Fluor 594 goat anti-rat IgG or Alexa Fluor 594 goat anti-rabbit IgG (2 µg/ml in Blocking Buffer; from Invitrogen). Next, cells were again washed with Blocking Buffer, counterstained with 4',6-diamidino-2-phenylindole (DAPI; 5 µg/ml in PBS) for 20 min in the dark, washed with PBS, allowed to air-dry, and mounted in Prolong Gold (from Invitrogen). Brightfield and fluorescence images were acquired simultaneously on a Zeiss Axioskop 2 MOT Plus wide-field fluorescence microscope. All samples were imaged with identical acquisition settings.

3.5.9 Growth and Virulence in Mice

Two types of animal studies were performed, both in compliance with all institutional guidelines for animal experimentation. For a short term model of fungal survival in the mouse lung, strains to be tested were cultured overnight in YPD medium, collected by centrifugation, washed in PBS, and diluted to 2.5×10^5 cells/ml in PBS. For each strain, eight 4-6 week-old female C57Bl/6 mice (from Jackson Laboratories) were anesthetized with a combination of ketaset-HCl and xylazine, and inoculated intranasally with 50 µl of the prepared yeast suspension. Three animals from each cohort were sacrificed at 1 hr post-inoculation; the remaining five were sacrificed after 7 days. Lungs were harvested following sacrifice, and homogenized in PBS. Serial dilutions of the homogenate were plated

on YPD agar for determination of colony-forming units (CFU). Initial inocula were also plated to confirm CFU.

To assess longer-term affects of cryptococcal infection, each strain was cultured and prepared as above, with the exception that the cells were diluted to 2×10^6 cells/ml in PBS. Ten 4-6 week-old female A/Jcr mice (from the National Cancer Institute) were anesthetized as described above and inoculated intranasally with 100 µl of the prepared cell suspension. The animals were weighed within 1 hr post-inoculation, and subsequently on every other day. Mice were sacrificed if weight decreased to a value less than 80% of peak weight (an outcome which in this protocol precedes any signs of disease) or upon completion of the study. Initial inocula were plated to confirm CFUs.

3.5.10 RNA-Seq

Cells were cultured overnight in YPD, and grown for 90 minutes in either capsule-inducing (DMEM, 37 °C, 5% CO_2) or capsule non-inducing (DMEM, 30 °C, room air) conditions prior to isolation of total RNA. A minimum of two biological replicates were performed for each mutant (*ada2*Δ, *nrg1*Δ and *cir1*Δ) and four for wild type. PolyA+ RNA was purified from total RNA using the Dynabeads mRNA Purification Kit according to the manufacturer's instructions (from Invitrogen). Each sample was resuspended in 2 µl of 100 mM zinc acetate and heated at 60°C for 3 minutes to fragment the RNA by hydrolysis. The reaction was quenched by the

addition of 2 µl volumes of 200 mM EDTA and purified with an Illustra Microspin G25 column (from GE Healthcare). First strand cDNA was made using hexameric random primers and SuperScript III Reverse Transcriptase (from Invitrogen) according to the manufacturer recommendations, and the product was treated with *E. coli* DNA ligase, DNA polymerase I, and RNase H to prepare double stranded cDNA using standard methods. The cDNA libraries were end-repaired with a Quick Blunting kit (from New England BioLabs) and A-tailed using Klenow exo- and dATP. Illumina adapters with four base barcodes were ligated to the cDNA and fragments ranging from 150-250 bp in size were selected using gel electrophoresis as recommended by the manufacturer. The libraries were enriched in a 10-cycle PCR with Phusion Hot Start II High-Fidelity DNA Polymerase (from Finnzymes Reagents) and pooled in equimolar ratios for multiplex sequencing. Single read, 36-cycle runs were completed on the Illumina Genome Analyzer IIx.

Sequenced reads were aligned to the *C. neoformans* H99 reference sequence [82] using Tophat [101]. Reads that aligned uniquely to the reference sequence were considered for gene expression quantification with Cufflinks [102] using the current genome annotation provided by the Broad institute. Gene expression was normalized using the Cufflinks provided option for quartile normalization. Differential expression analysis comparing mutant to wild type was performed with LIMMA [97] and ELNN [103] using a 5% false discovery rate. Genes whose expression was found to be significantly changed by either analysis method were counted as

differentially expressed. RNA-Seq data is available at GEO accession number GSE32049.

3.5.11 Gene Ontology (GO) Enrichment

GO enrichment analysis was performed by assigning GO categories to each gene according to the Broad Institute's PFAM annotations using the mapping provided by the Gene Ontology project (http://www.geneontology.org/external2go/pfam2go). A hypergeometric test was applied for each GO category, the resulting p-values were corrected for multiple hypothesis testing, and a cutoff of 0.05 was used to determine significance.

3.5.12 Chromatin Immunoprecipitation (ChIP)

Wild type and *ada2*Δ cells were cultured in triplicate overnight in YPD, and grown in capsule-inducing conditions (DMEM, 37 ºC, 5% CO_2) for 90 minutes. Cells were then fixed for 5 min in 1% (v/v) formaldehyde, and the reaction quenched with a final concentration of 125 mM glycine. Fixed cells were collected by centrifugation, washed with PBS, and resuspended in Buffer A (50 mM HEPES pH 7.5, 140 mM NaCl, 1 mM EDTA, 1% v/v Triton X-100, 0.1% w/v sodium deoxycholate) supplemented with protease inhibitors and 20 mM sodium butyrate (a histone deacetylase inhibitor). The cell suspension was subjected to mechanical bead-beating with 0.5-mm zirconium silicate beads for 2 min at 4ºC, followed by a 2-min rest, for

a total of 10 cycles. Chromatin was then sheared by sonicating the lysate for 30 sec at 40% power output, followed by a 1-min rest on ice, for a total of 40 cycles, and the lysate clarified by centrifugation. A fraction of the sheared chromatin was reserved as an input sample and the remainder was used for immunoprecipitation. Acetylated histone H3 was immunoprecipitated overnight with anti-acetyl-H3 (K9) antibody (from Millipore) tethered to protein-A sepharose (10 ml in a total volume of 700 ml). The beads were next washed sequentially in Buffer A, Buffer B (50 mM HEPES pH 7.5, 500 mM NaCl, 1 mM EDTA, 1% (v/v) Triton X-100, 0.1% (w/v) sodium deoxycholate), Buffer C (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% (v/v) NP-40, 0.5% w/v sodium deoxycholate), and Buffer D (10 mM Tris, 1 mM EDTA), and immunoprecipitated protein was eluted with Buffer E (50 mM Tris pH 8.0, 10 mM EDTA, 1% (w/v) SDS). Crosslinked DNA from input and IP samples was released by incubating the eluate at 65ºC overnight, and extracted with a solution of phenol/chloroform/isoamyl alcohol (25:24:1) prior to ethanol precipitation and resuspension in water. Mock IP reactions with no antibody yielded no measurable product (not shown) and were not quantified further.

ChIP-DNA for input and IP samples was end-repaired with Klenow DNA Polymerase and the DNA was purified with AMPure XP System beads (Beckman Coulter Genomics) and modified with A-tails using Klenow exo- before ligation to adapters to incorporate 7-base index sequences using T4 DNA ligase (Enzymatics). Adapter addition was confirmed on an Agilent 2100 bioanalyzer, and the DNA was PCR-amplified and then gel purified to remove adapter dimers and select sizes

optimal for high-throughput sequencing (150 to 300 bp). Libraries were 12-way multiplexed on an individual lane of an Illumina Hi-Seq 2000 flow cell, resulting in approximately 7 million 42-bp single ended reads per sample.

Reads generated from the input and IP samples were aligned to the *C. neoformans* serotype A reference sequence [82] using Bowtie [104]. Reads that mapped to multiple genomic loci were discarded. Peak calling was performed using MACS [105] with a significance threshold of 1×10^{-10}. To assess gross differences between the mutant and wild type, the average number of peaks over the three biological replicates of each strain was compared. Peaks were associated with specific genes if the peak center fell within 500 bp of the gene transcription start site according to the current annotation by the Broad Institute [82]. (For genes with unannotated 5'-UTRs this may correspond to the translation start site.) Ada2-dependent peak loss was identified by cases where a gene in two of the three wild type biological replicates possessed a neighboring peak and no peak was found to neighbor the gene in any of the three *ada2*Δ mutant replicates. ChIP-Seq data is available at GEO accession number GSE32075.

Chapter 4

NetProphet: A practical method for mapping transcriptional regulatory networks

Abstract

Deriving models of transcriptional networks from compendia of gene expression data has been a long standing goal of systems biology. Here we present a new algorithm, NetProphet, which integrates two network inference strategies: regression and differential expression. Regression analysis identifies predictive relationships between transcription factors and their potential targets. Differential expression analysis of transcription factor deletions relative to a wild type strain identifies genes downstream of the transcription factor in the network. By integrating these two analyses we achieve inference accuracy superior to that of either analysis alone. We compare our approach to other network inference algorithms on the DREAM4 in-silico benchmarks and on the complete transcriptional network of *Saccharomyces cerevisiae*. The latter comparison uses gene expression profiles of strains carrying transcription factor deletions and evaluates predicted networks by comparison to binding evidence from hundreds of chromatin immunoprecipitation on chip (ChIP-chip) experiments. NetProphet is substantially more accurate than alternative

algorithms in both evaluations. Finally we use the transcriptional network inferred by our algorithm to complement the ChIP-chip evidence by identifying novel protein-DNA interactions that are supported by both functional annotation and sequence affinity evidence.

4.1 Background

A major goal of systems biology is to develop predictive network models of the emergent behaviors of living cells. Such models can serve as hypothesis generators, guiding experimentation, and as a basis for engineering new functions. Microarray technology and high throughput sequencing of cDNA (RNA-seq) have given us an increasingly lucid window into the transcriptional states of cells. Analyses of these data have begun to shed light on the transcriptional networks that coordinate gene regulatory responses. Cells respond to environmental stimuli through signaling cascades that affect transcription via the network of transcription factors and their direct targets. By profiling cells over a range of growth conditions, a large compendium of gene expression data can be collected which represents the potential states of the transcriptional network. The response of the transcriptional network to precise molecular manipulations, such as gene disruption, RNA interference, and overexpression can also be measured. Network inference analyses that integrate a large collection of transcriptional profiles have the potential to predict the structure of transcriptional networks, thereby complementing evidence of protein-DNA interactions from experiments (e.g. ChIP-chip [106] or ChIP-seq [107]).

4.2 Related Work

4.2.1 Regression based approaches to network inference

One general strategy used to infer a transcriptional network from a collection of gene expression profiles is to use some form of regression analysis, wherein each gene's expression is predicted as a function of the expression of other genes that encode transcriptional regulators. The idea behind this strategy is that a regulator's expression profile will be predictive of its targets' expression profiles. Implementations of this strategy can be as simple as measuring the correlation or mutual information [108,16] between the expression profiles of regulatory genes and potential target genes. More complex implementations employ a multivariate predictor function, where multiple regulators' gene expression profiles are taken together to predict a target gene's expression [109,110]. To account for uncertainty in model selection and noise in the measurements, probabilistic approaches such as Bayesian networks have been applied [111].

In regression-based network inference strategies, each gene's potential regulators are ranked based on their ability to predict the target gene's expression. A specific network can be derived from such a ranking by including regulators whose rank exceeds a certain threshold. The predictive relationship that regression reveals suggests the potential for a direct interaction, in which the transcription factor binds

the promoter of the target and by doing so impacts the target's transcription. However, the regression approach has encountered several problems. First, expression of one gene may be predictive of another gene's expression for reasons other than direct regulatory interaction. For instance, the two may be co-regulated by a third factor. Second, in most regression based analyses measurements from genetically perturbed samples are inappropriately treated as equivalent to measurements from varying growth conditions. In fact, measurements from genetically perturbed backgrounds (e.g. single gene disruptions) are profiles of an altered system in which an individual gene's expression has been forced to zero. Causal Bayesian networks [112–114] offer one way of addressing this type of measurement, however their treatment of these measurements primarily assists with edge orientation rather than edge inclusion.

4.2.2 Inferring Network Structure with Differential Expression Evidence

A complementary strategy based on differential expression (DE) analysis relies on single-gene perturbations for both edge orientation and edge inclusion. In this approach, DE analysis is applied to transcription factor deletion strains, thereby identifying the genes downstream of the deleted factor. Hu et al. (2007) and Pinna et al. (2010) constructed networks with edges from each TF to all the genes that were differentially expressed (above some significance threshold) when the TF was deleted [115,116]. These networks, which contain a mixture of direct and indirect

interactions, are then refined by pruning edges that are less likely to be direct. This pruning step removes the lower confidence legs of feed-forward loops (FLLs) in the network, essentially retaining the most confident path between two genes to explain the effect of a perturbation. While these refinement approaches make explicit use of single-gene perturbations, they cannot use other sources of data such as environmental perturbations and time-course measurements. They can only identify targets of regulators that have been individually perturbed.

Surprisingly, there have been few comparisons of the accuracy of the two major approaches to network inference from gene expression data -- differential expression and regression. As a result, it is unclear which strategy is better in practice.

4.2.3 Inferring network structure by integrating analyses

It has been observed that combining multiple inference algorithms over the same compendia of expression data can yield more accurate results than relying on a single algorithm [110,117]. This improvement relies in part on the individual algorithms making different types of systematic errors. Given that regression and differential expression based strategies are such orthogonal approaches to network inference, they make prime candidates for integration. Combining regression and differential expression for network inference has been proposed [118], where regression is applied to time-course data and DE analysis is applied to gene knockout data. However this approach was intentionally biased toward the DE analysis.

Furthermore, the combined approach showed little or no benefit over the DE alone [118]. Regression and DE analyses were also combined in the context of the Inferelator algorithm [109,119,120]. Inferelator combines three analyses (LASSO regression, mutual information and differential expression) using an arithmetic averaging scheme that gives each analysis equal influence. Combining regression and DE in this way did show a benefit over the individual analyses. This result inspired us to search for even better ways of combining LASSO regression with DE analysis.

4.2.4 Initial Approaches for Integrating Differential Expression Evidence with Regression

Before developing our current algorithm, NetProphet, we investigated other approaches to network inference that we will discuss briefly. The overarching theme of these earlier attempts was the same: to make better use of measurements from genetically perturbed strains.

Initially, we took a Bayesian approach in which we modeled a transcriptional network as a probability distribution over a system of ordinary differential equations (ODEs) such as those described in section 2.6.2. The probability distribution was defined as $P(M|D)$ where M was the model (network structure and parameters) and D contained the gene expression measurements. The model was the structure and

parameterization of the network, which we encoded using ODEs which expressed the rate of mRNA changes as a function of TF concentrations. *P(M|D)* can be expressed as the product of the likelihood function *P(D|M)* and a prior *P(M)*. In our implementation, the likelihood function encoded the agreement between the system of ODEs and the measurements, D according to a normal error distribution. The model prior *P(M)* encoded our prior beliefs about the structure of the model based on DE analysis of single gene perturbations in the data. The probability distribution on the model structure and parameters was learned using a Markov chain Monte Carlo approach to sample the model structure. We derived our confidence of edges being present in the model structure based on their expected probabilities obtained during sampling. In addition to assessing the expectation that a particular edge will be present in the model, we also obtained the expected outcomes for genetic perturbations on the system. We achieve this by integrating the genetically perturbed form of the system over the course of sampling using the wild type expression state as the initial condition.

We tested two forms of the prior *P(M)*, both of which followed a Laplace distribution. The first form encoded prior knowledge about reachability between two genes in the directed graph structure that defined the transcriptional network. This prior knowledge was based on differential expression evidence which we interpreted as follows: if gene A responds significantly to a gene B deletion, then there exists a path from B to A in the transcriptional network. There were several problems with

incorporating differential expression evidence in this way. While interpreting that differential expression is indicative of a path in the network is perhaps the most conservative treatment of the evidence, we found that it does not sufficiently constrain the space of potential networks. We also encountered extreme computational overhead in implementing this prior, which required computing the transitive closure of the network structure over the course of sampling. Overall we found that the benefits in accuracy when using this approach did not warrant the effort.

In the second form of the model prior *P(M)* that we examined, we made a stronger assertion regarding the differential expression evidence: that more significantly differentially expressed genes are more likely to be direct targets of the perturbed gene in the transcriptional network. The accuracy results we obtained with this implementation were encouraging, but there were a few issues that eventually made us abandon this strategy. First, the cost of sampling networks using MCMC made predicting genome scale network structure prohibitive. The limit in terms of network size that we could reasonably approach was roughly 500 genes. A second problem we found with this algorithm was that based on the influence of the prior, *P(M)*, edges would be included in the model but the parameters associated with those edges would be very close to zero resulting in no regulatory influence. Often times these edges would be correct based on the known network structure, but not predictive in the of the target's expression.

Now, we propose our current approach, NetProphet, which combines the regression and differential expression strategies in a novel way. A comparison of NetProphet to alternative algorithms such as GENIE3 [121] and Inferelator [120] reveals substantial accuracy improvements both on the DREAM4 in-silico benchmark and on reconstruction of the *Saccharomyces cerevisiae* transcriptional network from empirical gene expression data. Finally, we demonstrate how network inference can complement ChIP experimental studies by identifying many novel protein-DNA interactions in yeast that are supported by functional annotation and sequence affinity evidence.

4.3 Approach

Our approach for inferring a transcriptional network from a compendium of gene expression data is to combine the predictions from two independent analyses. The first analysis uses all of the expression data to learn a sparse linear model that predicts the expression of each gene as a function of the expression of regulatory genes (transcription factors and cofactors). The second analysis assesses differential expression on each expression profile in the compendium in which a specific regulatory gene has been perturbed (via knockout, knockdown, or overexpression) compared to wild type control in the same growth condition. Each analysis scores the potential regulatory interactions of the system (regulator, target gene pairs). The scores of the two analyses are combined through a model averaging scheme. Simple

averaging can be extended with a weighting scheme that favors predicted interactions for which the regression and differential expression evidence agree on the sign of regulation (activation or repression).

4.4 Results

4.4.1 In-silico evaluation

In order to compare NetProphet to other methods, we started with in-silico gene networks, which allow network inference algorithms to be evaluated in a controlled manner [122,123]. Widely used benchmark networks are provided by the DREAM (Dialog for Reverse Engineering Assessments) consortium [124].

We compared NetProphet's accuracy on the DREAM4 in-silico networks to that of other network inference algorithms, including GENIE3 [121], Inferelator [109,125,120], and 19 anonymous algorithms that were included in the DREAM4 inference evaluation. Inferelator was a top performer for DREAM3 and DREAM4 network inference challenges. We used the version described in [120] that integrates three different analysis types (differential expression, LASSO regression and mutual information) to infer regulatory network structure. GENIE3 was the best performer for DREAM4 for the multi-factorial network inference challenge and was the best performer overall for DREAM5. In all of the in-silico comparisons we used the unweighted version of NetProphet (see section 4.6.4 for data handling details).

The DREAM4 in-silico network benchmark consists of 5 independent networks, each containing 100 simulated genes, generated by the GeneNetWeaver tool [126]. The network structures are graphs with edges directed from regulators to their direct targets; post-transcriptional regulation is not modeled. The structure of each network is a subgraph of the transcriptional network of either *S. cerevisiae* or *E. coli*. Gene expression datasets are generated from each network using stochastic differential equations that reflect the regulatory interactions in the network. Simulations of both steady state and time course measurements are included. The steady state measurements are of the wild-type network, each single-gene knockout, and each single-gene knockdown.

First, we compared the ranking accuracy of NetProphet to that of the two baseline methods that it integrates: LASSO regression and differential expression analysis (DE Rank). Figure 4.1 shows the precision-recall curves [127] for a single, representative DREAM4 network. These curves reveal that NetProphet produced a ranking superior to either of the individual analyses it integrates. NetProphet recovered over 40% of the true network structure before making a single error while each baseline method recovered less than 10%. This suggests that the baseline methods are making errors on distinct sets of edges and that averaging their scores tends to neutralize these errors.

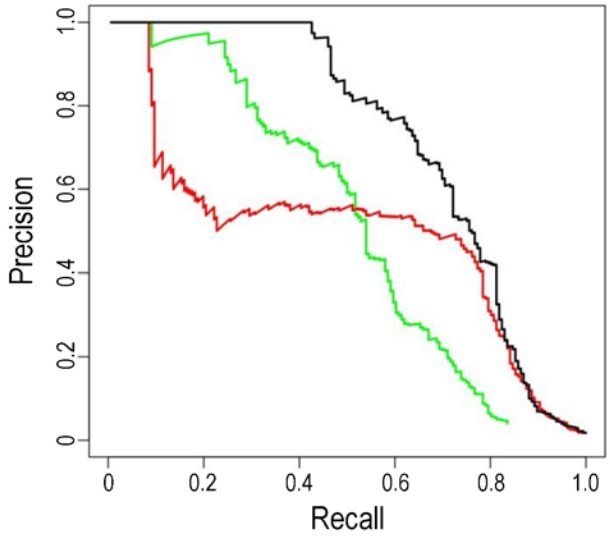

Figure 4.1. Precision-recall curves for DREAM 4 network 1; Netprophet (black); LASSO (green); DE Rank (red).

Averaging over all five DREAM4 networks, we calculated the area under the precision recall curves for NetProphet, DE Rank, LASSO, Inferelator, GENIE3 and 19 anonymous entrants in DREAM4 (Figure 4.2). NetProphet yielded an average area under the precision recall curve (AUC-PR) of 0.54 while the baseline analyses LASSO and DE Rank yielded AUC-PR of 0.36 and 0.35, respectively. NetProphet was also more accurate than Inferelator, GENIE3 and the 19 anonymous methods. Inferelator showed the second best performance amongst all methods. GENIE3's performance was much lower than we had expected, given its dominance in the DREAM4 multifactorial network challenge and later success in DREAM5. The performance of GENIE3 relative to Inferelator and NetProphet on these networks may in part be explained by the lack of multifactorial perturbations (broad treatments that influence the basal transcription rates of many genes) in these datasets. GENIE3

may rely on this measurement type for more accurate inference. These clear results encouraged us to perform an evaluation in the model organism *Saccaromyces cerevisiae*.

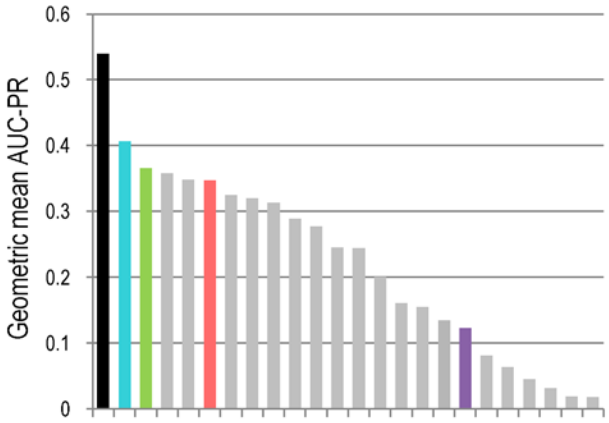

Figure 4.2. Average area under the precision recall curves for all 5 DREAM4 networks. NetProphet (black); Inferelator (teal); LASSO (green); DE Rank (salmon); GENIE3 (purple); anonymous methods (grey).

4.4.2 Evaluation in S. cerevisiae

To evaluate inference algorithms on the yeast transcriptional network, we applied them to a microarray dataset that includes profiles of 269 transcription factor deletion strains grown in rich medium [115]. The microarray data were normalized by the method described in [128] with minor differences described in Supplemental Methods. The network structure inferred by each algorithm was compared to a gold standard for protein-DNA interactions that we created by combining ChIP-chip

evidence from Tnet [129,32] and Yeastract [130–132]. Yeastract contains many other interactions that are not supported by ChIP evidence but these were omitted. Our gold standard for the yeast protein-DNA interaction network consists of 30094 protein-DNA interactions for 188 transcription factors (TFs).

We conducted DE analysis on the normalized gene expression data using LIMMA and converted the resulting log-odds scores into signed significance scores D_{ij} (see section 4.6.2). To validate this analysis, we compared the p-values assigned by LIMMA [97] for each gene to the p-values published by [128] and found the ranking orders to be nearly identical (data not shown). We then performed LASSO regression (see section 4.6.1) on the normalized gene expression data to learn coefficients B_{ij} for a sparse linear model. DE and LASSO displayed a striking concordance on the regulatory influence (activating or repressing) of each transcription factor on its predicted gene target (Figure 4.3). The regulatory relationships that received a positive score in both analyses (activation, Figure 4.3 region I) or a negative score in both (repression, region III) outnumbered those that received a positive score in one analysis and a negative score in the other (regions II and IV). Furthermore, the scores in regions I and III were significantly greater (p<1e-15) than the scores in regions II and IV (LASSO mean scores: 0.058 versus 0.045; DE mean scores 0.026 versus 0.013). Finally, there was substantial enrichment for interactions supported by ChIP-chip in region I (2-fold enriched; p < 1e-154) and region III (1.5 fold enriched; p < 1e-68). Enrichment in other regions was slight.

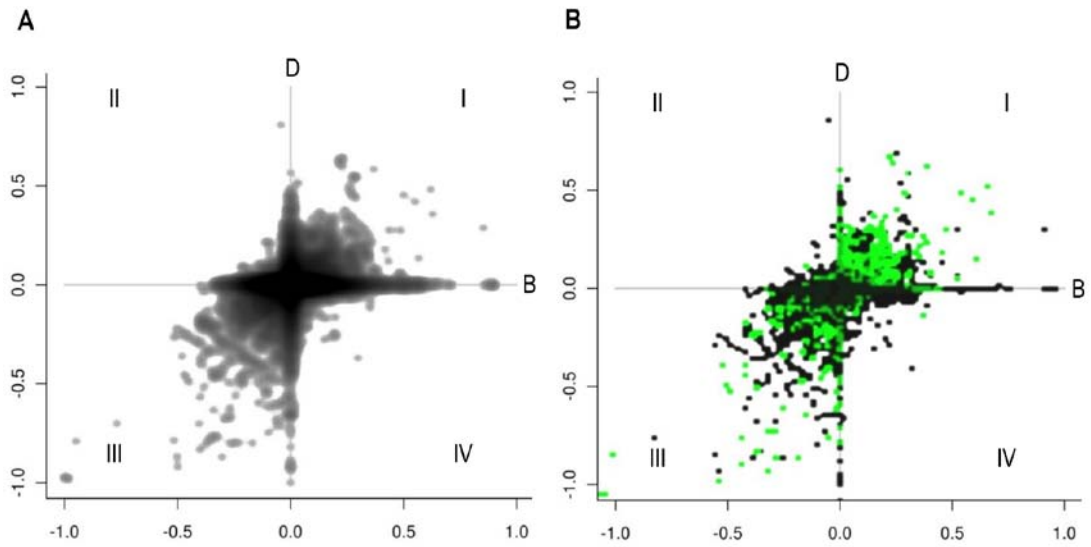

Figure 4.3. Comparison of LASSO and DE concordance and ChIP enrichment by corcordance in *S. cerevisiae*. Panel A, Density plot of the normalized LASSO coefficients: B_{ij} (x-axis) plotted against signed DE scores: D_{ij} (y-axis); Darker color indicates greater score density. Panel B, Scatterplot of ChIP enrichment by LASSO and DE concordance; black regions are not enriched for ChIP supported interactions, and green regions are enriched relative to background.

Next, we made predictions for the structure of the yeast transcriptional network using both weighted and unweighted versions of NetProphet. For the unweighted version we combined the confidence scores from LASSO regression and DE analysis as described (see section 4.6.3). For the weighted version, we estimated optimal region weights by performing cross-validation over the training data (interactions labeled as true or false by ChIP evidence), selecting weights that maximized the average AUC-PR (see section 4.6.5).

Of interactions ranked in the top 1000 by the unweighted model, 52% fell into region III followed by 40% in region I; less than 8% of the interactions were found outside of regions I and III. In the ranking produced by the weighted model, approximately 60% of the top 1000 interactions were in region I and 20% in region III. The remaining 20% were those that showed evidence of differential expression but were not included in the LASSO model (Figure. 4.4, region D). Thus, the weight estimation algorithm determined that the most accurate predictions were those in which both analyses agreed on the sign of regulation (activating or repressing). Furthermore, activating interactions were found to be even more accurate than repressing interactions (see section 4.5). Finally, differential expression alone was more reliable than either LASSO alone or LASSO with a sign opposite that of differential expression.

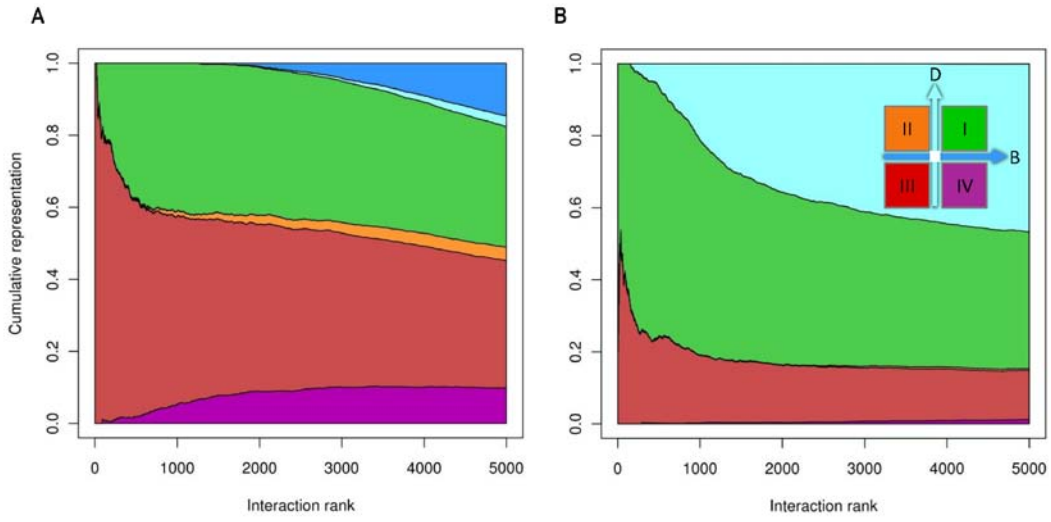

Figure 4.4. Cumulative region representation by interaction rank; Panel A: NetProphet (unweighted). Panel B: NetProphet weighted.

We compared the precision-recall curves of unweighted NetProphet against weighted NetProphet, LASSO, DE Rank, GENIE3, and Inferelator (Figure 4.5). As in the in-silico comparison, NetProphet was more accurate than the base-line methods (LASSO and DE Rank). Interestingly, DE Rank revealed a distinct advantage over LASSO. This advantage which was not seen in the in-silico network evaluation, suggests that LASSO regression is not as informative for inferring real regulatory networks as it is for artificial networks. Weighted NetProphet was the most accurate algorithm, followed by unweighted NetProphet. For the 5 top ranked predictions, Inferelator demonstrated near equal performance with NetProphet, but accuracy rapidly diverged for less highly ranked interactions.

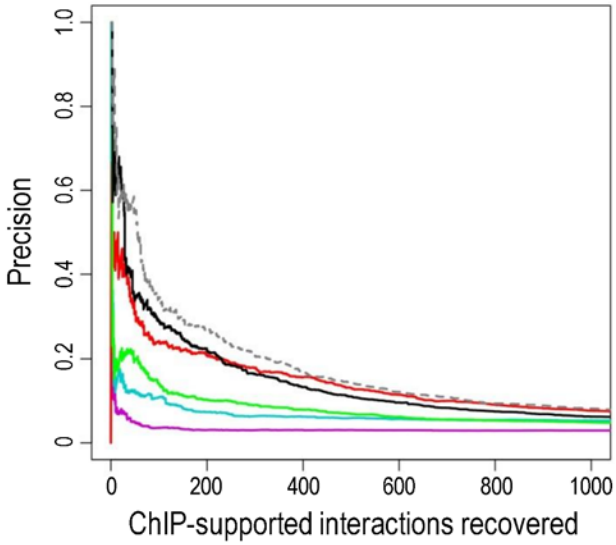

Figure 4.5. Precision-recall for the transcriptional network of *S. cerevisiae.* Unweighted NetProphet (black); weighted NetProphet (grey dash); LASSO (green);

DE Rank (red); GENIE3 (purple); Inferelator (teal). The X-axis represents the number of ChIP supported interactions recovered.

Next, we performed a regulator-centric evaluation that considers the accuracy with which the targets of a randomly selected regulator can be predicted. This metric more closely reflects the typical application of a network prediction algorithm, in which investigators seek novel targets of regulators relevant to a specific biological process. Global accuracy is less predictive of an algorithm's usefulness in this application, since it can be dominated by a small number of regulators that have many targets. Here we calculate the fraction of transcriptional regulators in *S. cerevisiae* for which each algorithm identifies at least one true target in the top k predictions, for k from 1 to 10 (Figure 4.6). Only the 187 transcriptional regulators for which we have ChIP evidence are included. For 17% of those regulators, the target ranked most highly by unweighted NetProphet was correct; for weighted NetProphet this fraction increased to 20%. The baseline analyses DE Rank and LASSO identified the correct target for 15% and 13% of regulators respectively, with Inferelator identifying correct targets for 9% and GENIE3 for 6%. This analysis demonstrates that NetProphet is not only inferring the transcriptional network with greater accuracy, it is also accurately identifying the targets of more regulators. Thus, if one is interested in a specific transcriptional regulator, NetProphet is more likely to identify its targets than the alternative algorithms.

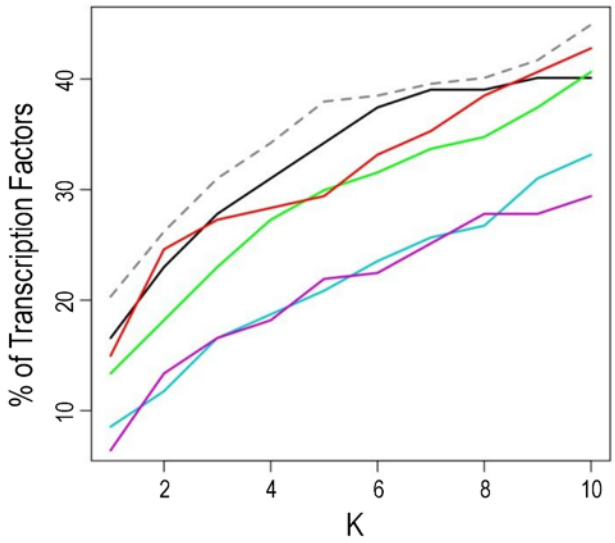

Figure 4.6. Evaluation of methods against the global transcriptional network of *S. cerevisiae*. Percentage of transcription factors (y-axis) for which a ChIP-supported interaction was identified in top K predictions (x-axis). NetProphet (black); NetProphet weighted (grey dash); LASSO (green); DE Rank (red); GENIE3 (purple); Inferelator (teal).

4.4.3 Refining the transcriptional network of *S. cerevisiae*

Although ChIP evidence may be the best available standard by which to evaluate network inference algorithms, it does not definitively identify the true transcriptional network of an organism. Detection of DNA-binding events by ChIP-chip or ChIP-seq is subject to statistical error and experimental bias, which will admit a certain fraction of false positive interactions. Furthermore, some true binding events are non-functional or spurious in nature [133], leading to an inflated number of interactions many of which do not play significant role in the regulation of gene

expression. Conversely, ChIP may fail to detect true interactions for the same statistical and experimental reasons. True interactions can also be missed because ChIP of cells grown in one condition may not reflect the binding events of a different growth condition.

Gene-expression based network-inference algorithms have the potential to complement ChIP studies in refining transcriptional network models. Here, we explore this potential by examining the network inferred by weighted NetProphet. We focus on the interactions ranked above a threshold that we set so that 30% of the predictions were ChIP positive. Only TFs for which there was ChIP data were considered when setting this threshold, but the threshold did admit a number of predictions involving TFs that had not been chipped. This threshold yielded 955 predicted interactions including 106 different regulators and 561 different target genes. ChIP experiments were available for 64 of the 106 regulators, and of the 64 regulators tested by ChIP-chip, 60 had known position weight matrices that specified their sequence affinity. We obtained these position weight matrices (PWMs) from ScerTF, which is a compilation of the most trusted position weight matrices for yeast transcription factors [134].

We were interested in identifying protein-DNA interactions predicted by NetProphet that were not supported by ChIP-chip but were supported by the presence of binding sites in the promoter region of the target gene. To establish binding site evidence for an interaction we scanned the PWMs over the yeast promoters using FIMO [135].

We defined a gene's promoter region to be 800 bases upstream of the transcription start site (TSS) excluding sequence from neighboring open reading frames (ORFs) and 200 bases downstream of the TSS. For a given transcription factor we considered two models of binding. We considered a protein-DNA interaction to be supported by the strong-site model if the target promoter's most significant binding site is in the 90th percentile of strength among all promoters that contain a significant (p< 0.005) binding site for the given regulator. Thus, by definition, only 10% of the promoters with significant hits could be considered to have strong sites. We consider a protein-DNA interaction to be supported by the weak-site model if the sum of the negative log p-values for all significant sites in the promoter is in the 90th percentile among all promoters that contain a significant (p< 0.005) binding site for the given factor. This threshold could admit up to another 10% of promoters with significant sites, although in practice many of these are the same promoters that satisfy the strong site model. We say a protein-DNA interaction has evidence of binding potential if it is supported by either the strong- or the weak-site model.

Next, we investigated the overlap among interactions predicted by NetProphet, interactions supported by ChIP-chip, and interactions supported by the strong-site or weak-site binding models. Only predicted interactions whose regulator has been ChIPed are considered in this analysis. Predicted interactions that were ChIP-positive overlapped significantly with those that were supported by both the strong-site model (p < 1e-53) and the weak-site model (p < 1e-15). Predicted interactions that were ChIP-negative also overlapped significantly with those that were supported by both

the strong-site model (p < 1e-12) and the weak-site model (p < 1e-6). Of the predicted interactions that were ChIP-negative, 36% were supported by evidence of binding potential (Figure 4.7); of those that were ChIP-positive 74% were. If the fraction of correct predictions supported by binding evidence is independent of ChIP support then 49% of the ChIP-negative predictions are correct.

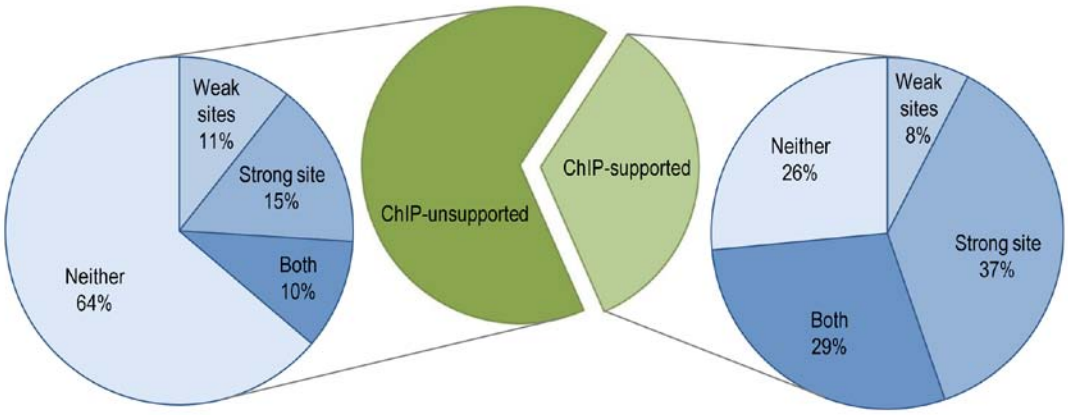

Figure 4.7. Classification of confidently identified interactions. Interactions identified at a confidence level that yielded 30% precision (omitting pseudo genes and interactions originating from transcription factors whose targets are unknown by ChIP).

The complete network of predicted interactions that are ranked above threshold and supported by either ChIP or binding potential is shown in Figure 4.8. We found several interactions that were not supported by the ChIP studies we used but were confirmed by other sources. For example, we predicted Gal80 directly represses *GAL2* expression, which is known to be true by a well studied mechanism in which Gal80 forms a complex with transcriptional activator Gal4, which inhibits Gal4 from

activating the transcription of many genes involved in galactose uptake and metabolism including *GAL2* [136].

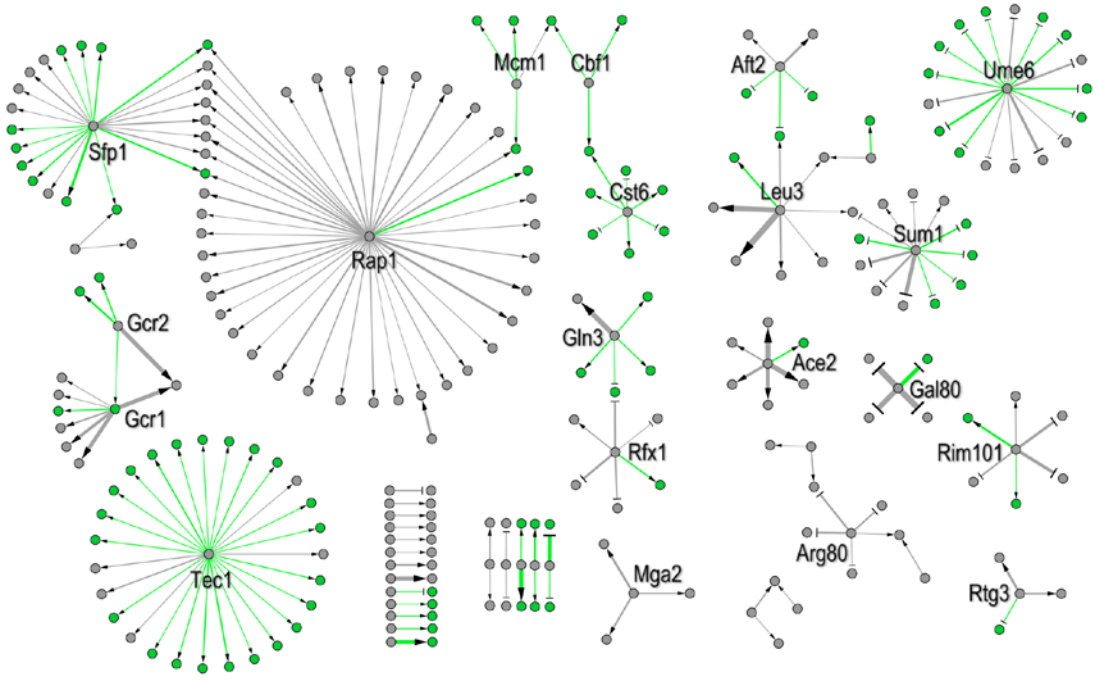

Figure 4.8. Interactions predicted by NetProphet that are supported by ChIP (grey) or unsupported by ChIP but supported by sequence affinity evidence (green); interactions unsupported by either are not shown. Genes which are the targets of novel interactions are also colored green (grey otherwise). Regulators with more than three predicted interactions are labeled. All predictions are ranked above a threshold at which 30% are ChIP-positive. Edge width indicates the magnitude of the score assigned by NetProphet.

NetProphet also predicted that Leu3, the master regulator of leucine biosynthesis, directly activates expression of *DIC1*, which encodes a mitochondrial dicarboxylate

carrier. This interaction was not detected by the original ChIP studies performed on Leu3 [137]. However, a later ChIP-chip experiment [138] hinted that Leu3 does bind the *DIC1* promoter and this was later confirmed by Calling Card-Seq [139]. Most of the novel NetProphet predictions that showed significant binding potential were consistent with the function of the regulator and its known targets. For example, NetProphet predicted that Ace2, which is known to regulate genes involved in daughter cell fate determination, directly activates *DSE3*, a gene which encodes a daughter cell specific protein. In addition to *DSE3*, Ace2 was predicted to regulate daughter cell specific genes *DSE1*, *DSE2* and *DSE4*, all of which are supported by ChIP evidence. A later ChIP study confirmed Ace2 does in fact bind the *DSE3* promoter and is required for its expression [140]. NetProphet also predicts several novel targets for two key transcriptional regulators of sporulation: Ume6 and Sum1. Ume6 represses genes involved in early meiosis in the presence of a fermentable carbon source. NetProphet predicts that Ume6 represses *SPO1*, *HOP2*, *CTF19*, and *ADY2* all of which are required for sporulation but remain undetected by ChIP studies of Ume6. Additionally NetProphet predicts that Ume6 represses 5 of the 13 genes in the DUP380 gene family: *COS1*, *COS2*, *COS3*, *COS4* and *COS8*. This family contains subtelomerically encoded proteins of unknown function, but our prediction suggests that they may be involved in early meiosis. We also identified many novel targets of Sum1, a transcription factor that represses genes involved in mid-phase meiosis. We predict Sum1 directly represses *GAS4*, *SPR1* and *SPO21*, all of which are known to be involved in ascospore wall formation. We also predict

Sum1 represses two genes of unknown function, *YAL018C* and *YAL047C*, suggesting that they may also be involved in sporulation.

Many predictions involve regulators for which no sequence preference is known and no ChIP studies have been performed. We found the majority of these predictions to be for proteins involved in chromatin remodeling, such as members of complexes including Tup1/Cyc8, Swi/Snf, SAGA and Rpd3/Sin3. Tup1 is inferred to directly regulate 126 genes. Tup1 is predicted to repress roughly half of its targets and activate the other half, consistent with Tup1's known role of acting as both a repressor and activator of gene expression [141,142]. Of the genes Tup1 is predicted to repress, the PAU gene family is significantly represented with 18 of the 23 members accounted for (hypergeometric $p < 1e\text{-}30$). These genes are subtelomerically encoded and expressed during fermentative growth, although their function is not well understood. Tup1's role in the regulation of this gene family has been previously reported [143,144]. In addition to its repressing role, we predict Tup1 activates expression of 71 genes, including 47 retrotransposable elements. This is consistent with the phenotype of the *TUP1* null mutant, which exhibits decreased transposition of these elements [145].

We predict 129 genes to be transcriptionally regulated by the Swi/Snf complex (subunits Swi3, Snf2, Snf5, and Snf6). Consistent with the general role of this complex as a positive regulator of transcription [146], 91% of the predicted interactions are activating. A large fraction of the target genes are functionally

involved in amino acid biosynthesis and transport and glucose metabolism. For SAGA, a different complex involved in transcriptional activation, we identified 17 targets (for subunits Spt3, Spt20, and Hfi1). Consistent with the general activating role of this complex, all targets were predicted to be activated. There was good overlap amongst the predicted target genes of different SAGA subunits, with 7 of the 17 targets sharing two of the three regulatory subunits. This was not the case for the Rpd3/Sin3 deacetylase complex -- we predicted a total of 89 targets for two subunits, Sin3 and Sds3, with only three targets in common: *STE2*, *SST2* and *PHO12*. The non-overlapping target sets for Sin3 and Sds2 may indicate divergent roles for these two proteins.

4.5 Discussion

We developed a novel algorithm, which we call NetProphet, to predict regulatory protein-DNA interactions from gene expression profiles of strains in which specific transcription factors have been perturbed. Evaluations using in silico networks from DREAM4 and genomic data from *S. cerevisiae* both indicated that previous algorithms did not exploit deletion-strain profiles as effectively as NetProphet. One likely reason is that many previous algorithms, exemplified by GENIE3, treat expression data from TF deletion strains as generic samples of possible cellular states; no special treatment is applied to the perturbed TF gene. As a result, the perturbed gene is ignored as a cause of observed changes and, worse, the algorithms

attempt to manufacture a regulatory explanation for its externally perturbed expression.

A second likely reason for NetProphet's greater accuracy is that previous algorithms have combined differential expression (DE) analysis with regression (or regression-like) analysis in a suboptimal way. For example, Inferelator ranks potential interactions by DE significance, but it combines this ranking with a regression-based ranking in a way that does not penalize disagreement as much as NetProphet does. Specifically, NetProphet combines scores by applying a geometric-like mean, where Inferelator uses an arithmetic-like mean. Another approach to combining DE and regression is to discard interactions that are not supported by DE; those that are supported by DE are then ranked only by regression [119]. Our findings reveal that ranking by DE significance is at least as good as ranking by regression and ranking by both is better still.

The idea for a weighted version of NetProphet arose from examining the plot in Figure 4.3, which revealed that predicted interactions for which both DE and regression agree are the most likely to be ChIP supported. Further analysis of plot regions I-IV and axes D and B revealed that the predictions identified as activating by both analyses are most likely to be correct (region I). The next best groups, with roughly equal accuracy, are those that are called repressing by both analyses (region III) and those that are detected by DE but not by regression (axis D, excluding the origin). The least ChIP-supported groups are those that are called activating by one

analysis and repressing by the other (regions II and IV) along with those that show no evidence of differential expression (axis B, including the origin). Although these observations were based on microarray data on *S. cerevisiae*, we see no reason to doubt that they would apply generally to other types of cells and expression profiling technologies. Thus, while a slight advantage might be gained by re-estimating weights for a new data set, using the weights derived here is likely to provide an improvement over the unweighted algorithm when re-estimation is not practical.

One property that we observed in the yeast transcriptional network inferred by NetProphet was the striking consistency in the sign (activating or repressing) of all interactions predicted for a given regulator (Figure 4.8). No formal constraint was made to bias the inferred network to have this property. Moreover, the sign of regulation for a given regulator was generally consistent with the known role of the regulator as a repressor or activator (or both). For example, all inferred targets of Ume6 are predicted to be repressed which is consistent with the role of Ume6 as a repressor of early meiosis genes in the presence of glucose. For regulators that are known to function as both activators and repressors, such as Tup1, we predicted a mixture of repressing and activating interactions. This property may be useful in post-analysis of inferred networks to hypothesize the function of an uncharacterized regulator or to identify the most trusted novel targets for a regulator whose function is known.

NetProphet's ChIP-supported predictions were not limited to a few regulators with many targets. For 20% of the regulators, the top ranked prediction was ChIP-

supported, and for more than 40% of the regulators at least one the top 10 predictions was ChIP-supported (Figure 4.6). We believe this to be quite good, considering that many transcriptional regulators may not even be active in the growth conditions used for these experiments (rich media). Furthermore, it is known that many *S. cerevisiae* regulators have paralogs that can compensate for their absence, thus masking the effect of gene deletions [147]. In spite of this, NetProphet was able to identify genuine targets for a broad swath of the yeast transcriptional regulators using expression data from single-gene deletion strains.

In addition to the predicted interactions that are supported by ChIP evidence, many interactions are not detected by ChIP but are likely to be real. Evidence for this includes high functional coherence among the predicted targets of most regulators. In addition, predicted targets of a regulator that are not supported by ChIP are nonetheless hugely enriched for genes whose promoters have exceptional potential for binding the regulator. Several of these novel interactions were confirmed by experimental datasets not included in the major ChIP-chip compendium.

To date, network reconstruction algorithms have been applied in only a limited way, typically in projects that are not focused on specific regulators or biological phenomena. We developed NetProphet in response to a perceived need for algorithms that can be applied to the investigation of specific biological phenomena. For example, we are investigating the network by which *Cryptococcus neoformans*, an opportunistic fungal pathogen, regulates the polysaccharide capsule it must deploy to grow in a human host [148]. We have found it much easier to generate

high quality expression data on tens of TF deletion strains than to generate high quality ChIP-seq data for the same TFs. Even when a TF has been successfully ChIPed, many questions remain. For example, ChIP hits are based on identification of clear peak signatures in a promoter that stand out from the background tag density in the surrounding region [105]. If a TF regulates a gene by binding a large number weak sites scattered throughout the promoter, the ChIP signal may be indistinguishable from background noise. It has been reported that ChIP-Seq tag density is a good predictor of a transcription factor's sequence affinity for a given site [149], suggesting that strong binding sites would be easier to distinguish from background than weaker sites. Indeed, our analysis of promoter sequences suggests that promoters with an exceptional number of weak sites for a TF are less likely to show a ChIP hit, even if one considers only targets that show differential expression when the TF is deleted. NetProphet predicts many of these functional targets with an exceptional number of weak sites, even though it does not consider sequence at all in making its predictions. Successfully carrying out ChIP is more difficult than perturbing TFs and profiling their expression, but even for TFs that have been ChIPed, NetProphet appears to provide additional functional targets that are not detected by ChIP.

The gold standard for validating predicted functional TF binding sites is comparison of expression from the wild-type promoter to expression from the same promoter carrying a point mutation in the TF binding site. The wild-type and mutated promoters should show differential expression in wild type cells but not in cells

lacking the TF that is thought to bind the site (and it's backups, if any). One can imagine validating a large number of binding sites that lack ChIP support in this way, but only if sites can be predicted with sufficient accuracy to warrant the effort of validation. Based on the results presented here, it appears that NetProphet has approached, perhaps even crossed, this accuracy threshold. We can therefore see a path to a future where we will no longer have to accept the false negatives and false positives of ChIP; instead, we will have recourse to predictions that are accurate enough to warrant the definitive test for functional regulatory binding.

4.6 Methods

4.6.1 Sparse regression for network inference

Sparse linear models have been applied to the network inference problem for some time now [150,109,151]. The network inference problem requires fitting a large number of parameters (one for each possible interaction) to relatively few measurements. Thus over fitting, in which parameter values are heavily influenced by non-representative characteristics of the small data sample, is a major challenge. Regression using sparse linear models controls over fitting by producing minimally complex networks in which most parameters are set to zero. LASSO is an L1 constrained regression technique for parameterizing sparse linear models [152]. Computationally efficient implementations of LASSO are available, enabling genome scale network inference problems to be approached.

We apply LASSO regression to learn a sparse linear model that encodes the transcriptional network. We formulate our model as:

$$\underset{B}{\operatorname{argmin}} \sum_{jk} \left(Y_{jk} - \left(\sum_{i} B_{ij} \cdot X_{ik} \right) \right)^2 + \theta \sum_{ij} \left| B_{ij} \right|$$

Where Y_{jk} is the value of the jth gene's expression in measurement k and X_{ik} is the expression value of the ith regulator in the kth measurement. In other words, the X_{ik} are the subset of the Y_{jk} for which gene i encodes a TF. B_{ij} is the coefficient that the LASSO procedure learns to describe the influence of regulator i on gene j. Θ is a weight that scales the L1 parameter penalty relative to the sum of squared prediction errors. When Θ is 0, the optimization of B is equivalent to ordinary least squares regression. As Θ grows, components of B are forced to zero, yielding a sparser solution. In this application we disallow auto regulation by prohibiting B_{ij} from becoming non-zero when regulator i is encoded by gene j. We handle gene perturbations in the regression by omitting measurements in which gene j has been perturbed when fitting the coefficients B_{ij}.

To select a model, optimal parameters are learned over a range of Θ values, each representing a different degree of model complexity. Each model is assessed for its predictive error by performing 10-fold cross-validation on the gene expression data. The value of Θ which minimizes predictive error is selected as defining the appropriate model complexity. The final model is selected by learning a solution using all of the expression data (no longer cross-validating) using the optimal value of Θ, which was learned during cross validation. We use a least angle regression

[153] implementation to efficiently learn the values of B that minimize the predictive error. To produce a ranking of potential regulator-target interactions, the B_{ij} values are ranked according to their magnitude, where the largest $|B_{ij}|$ value indidates regulator i has the greatest influence on gene j.

4.6.2 Differential expression analysis

Differential expression (DE) analysis is used to characterize the responsiveness of genes to a perturbation relative to an unperturbed control condition. For the purpose of resolving transcriptional network structure, we are interested in identifying genes that show significantly different expression in strains where a particular transcriptional regulator has been disrupted. We use DE analysis to rank potential regulator-target interactions based on the estimated probability that each gene changes expression in response to the regulator deletion. The observed difference for gene j in the wildtype relative to a strain that lacks regulator i is defined as follows:

$$F_{\Delta i}(j) = \log_2(E_{WT}(j)) - \log_2(E_{\Delta i}(j))$$

$F_{\Delta i}(j)$ is the $\log_2$-fold change of gene j in the wildtype relative to a Δi background. $E_{\Delta i}(j)$ is the mean expression of gene j in a Δi background, and $E_{WT}(j)$ is the mean expression of gene j in the wild type background. We define the log-odds that gene j is differentially expressed in the Δi background as:

$$L_{\Delta i}(j) = \log\left(\frac{\Pr(F_{\Delta i}(j) \neq 0)}{\Pr(F_{\Delta i}(j) = 0)}\right)$$

We compute $L_{\Delta i}(j)$ using LIMMA, which estimates the posterior log odds score by way of a moderated t-statistic in which gene specific variances are shrunk toward a common value (see [97] for details). A signed confidence score is assigned to each potential regulator-target interaction as follows:

$$D_{ij} = \begin{cases} L_{\Delta i}(j) \cdot \text{sgn}(F_{\Delta i}(j)) & L_{\Delta i}(j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

D_{ij} represents the signed confidence score that regulator i directly regulates gene j. The sign of D_{ij} indicates whether regulator i is repressing or activating gene j. When it is more likely that gene j's expression is unchanged in the Δi background (i.e. $L_{\Delta i}(j) < 0$), the interaction is assigned a confidence score of 0. If the gene expression compendium contains no measurements for the Δi strain, D_{ij} is set to zero for all j. The D_{ij} values are ordered according to their magnitude thus defining a ranking of hypothesized interactions from the ith regulator to the jth gene.

4.6.3 Model integration

We integrate the sparse regression and DE analyses using a model averaging scheme. Before combing the score matrices B (from regression) and D (from differential expression), each matrix is normalized such that its values lie on the interval [-1,1]. This is done by dividing each element in B by $max(|B_{ij}|)$ and similarly for D. After normalization the combined scores are computed as follows:

$$M_{ij} = \left(\left|B_{ij}\right| + c_b\right) \cdot \left(\left|D_{ij}\right| + c_d\right)$$

Where c_b and c_d are offset constants to prevent M_{ij} from becoming zero when only one of the two individual scores are zero; both c_b and c_d are set to 0.01. M_{ij} is the combined confidence score for the interaction from regulator *i* to target gene *j*.

We also report on a weighted extension of this model, which can be used with the weights reported here or trained for each new organism for which known protein-DNA interactions are available. The learning algorithm weights the combined confidences scores M_{ij} according to the sign of the two scores B_{ij} and D_{ij}. We define 6 regions based on the paired signs of the two scores as follows:

$$R(B_{ij}, D_{ij}) = \begin{cases} I & B_{ij} > 0; D_{ij} > 0 \\ II & B_{ij} < 0; D_{ij} > 0 \\ III & B_{ij} < 0; D_{ij} < 0 \\ IV & B_{ij} > 0; D_{ij} < 0 \\ B & B_{ij} \neq 0; D_{ij} = 0 \\ D & B_{ij} = 0; D_{ij} \neq 0 \end{cases}$$

The regions define the consensus (or non-consensus) of the signs of the two scores. Regions I and III represent consensus between the two scores for activation and repression respectively. Regions II and IV represent disagreement about the sign of regulation, and regions B and D represent cases where one analysis detects evidence of the interaction and the other does not. Possible interactions that are not detected by either analysis are excluded from further consideration. The weights applied to these regions are represented by a vector of positive numbers denoted ω. In the weighted model, the combined score for the regulation of gene *j* by regulator *i* is:

$$M_{ij} = \left(\left|B_{ij}\right| + c_b\right) \cdot \left(\left|D_{ij}\right| + c_d\right) \cdot \omega_{R(B_{ij}, D_{ij})}$$

where $R(B_{ij}, D_{ij})$ is the subscript for the region corresponding to the regression coefficient B_{ij} and the DE weight D_{ij}. In the weighted model, the offset coefficients c_b and c_d and the weight vector ω are learned by cross-validation on the training data (labeled interactions), maximizing the area under the precision recall curve.

4.6.4 Analysis of DREAM4 expression data

Inference of the DREAM4 networks was performed using our method, NetProphet, and publicly available versions of Inferelator (http://err.bio.nyu.edu/inferelator/) and GENIE3 (http://www.montefiore.ulg.ac.be/~huynh-thu/software.html). Datasets for the DREAM4 in-silico 100 networks were obtained from http://wiki.c2b2.columbia.edu/dream/index.php/D4c2. These datasets covered 5 networks each containing 100 genes. The individual datasets for each network contained single measurements for wild type, all single knockouts, all single knockdowns (50% expression) and 10 time courses (each with 21 time points), for a total of 411 measurements. All expression data is provided in a normalized format such that each gene's expression lies on the interval [0,1]. We applied Inferelator to this dataset as described in [120], using the CLR + Inferelator + MCZ pipeline . Similarly for GENIE3 we inferred the network structure using the GENIE3 functions: *read.expr.matrix* and *get.weight.matrix* and their default parameterizations.

To properly handle the time course data, a spline was fit to the expression values for each gene. For each time point the derivative of the spline was used to estimate the rate of change for each gene, and the transcription rate of each gene was estimated by adding the gene's concentration to the rate of change at each time point (assuming a unit degradation rate constant). These estimated transcription rates were used instead of the expression measurements as the response matrix, *Y*, for LASSO regression. Note that a mixture of steady state and time course measurements in the response matrix is compatible under this formulation, because steady state concentrations are equal to transcription rates assuming a unit degradation rate constant. In addition to modifying the response variable to allow for a mixing of steady state and time course data, the covariate matrix, *X*, is also modified. Gene expression measurements for time course measurements in the covariate matrix are replaced with protein concentration estimates for each time point (which are effectively lagged expression measurements). We estimated a gene's protein concentrations using the spline fit to mRNA measurements, and integrated an ODE which defines a protein's rate of change as a function of the mRNA concentration, minus the protein concentration times a degradation rate constant. The degradation rate constant which we set to 0.01 for all genes defines the lag between the mRNA and protein species to be roughly one time point. Finally, a $\log_2$ transformation was performed on both the response matrix *Y* and covariate matrix *X* before applying LASSO regression.

Differential expression analysis of the knockout measurements was used to compute the DE rank scores D_{ij}. LIMMA was used to compute these scores by comparing each knockout (which consisted of a single measurement) to 11 measurements of wild type (one for each time 0 point of the 10 time courses, which was a steady state wild type measurement, and one which was provided separate of the time courses).

4.6.5 Analysis of *S. cerevisiae* Microarray Data

The microarray data used for the inference of the yeast transcriptional network was originally published in Hu et al., 2007, and later reanalyzed in Reimand et al., 2010. Our normalization of this data largely follows the scheme described in Reimand et al., 2010. Briefly, we downloaded the raw GenePix files for each of the 588 microarrays from the Longhorn Microarray Database [154]. Normalization of the raw spot intensities was performed using Linear Models for Microarray Data, LIMMA [97]. Three different array platforms were used in this study. To ensure for meaningful print-tip correction, samples were imported in three batches according to their platform membership. For each batch, background correction was performed using the LIMMA function *backgroundCorrect* with the method type *normexp* and an offset of 50. Print-tip loess normalization was performed using the LIMMA function *normalizeWithinArrays*. After print-tip correction, all batches were merged into a single *MAList* object, and treated in a platform independent manner. For each array, M and A values for duplicate probes were averaged. Individual mutants were hybridized against one of three wild type RNA samples (2 from BY4741 and 1 from

S288C). Each mutant had a minimum of two biological replicates, although the replicates were often hybridized against different wild type samples. We handle this differently than Reimand et al 2010 by constructing a linear model which relates all mutants to one of the BY4741 samples using the LIMMA functions *modelMatrix* and *lmFit*. The coefficients returned by *lmFit*, which correspond to the log-ratio of each gene's expression in each mutant relative to the wild type strain, were saved as the table of gene expression used by all subsequent network inference analyses. Differential expression of each deletion strain was assessed relative to the wild type strain (BY4741) using an empirical Bayesian moderated t-test [97] implemented with the LIMMA function *eBayes*. Log-odds scores returned by this function correspond to the values $L_{\Delta i}(j)$ described in section 4.5.2.

Subsequent inference of the yeast transcriptional network was performed using Inferelator, GENIE3 and our method NetProphet. We used the same Inferelator pipeline (CLR + Inferelator +MCZ) that was applied to the DREAM4 data to infer the yeast transcriptional network. This pipeline was modified to restrict the set of allowed regulators to be transcription factors. Additionally, given the absence of time course data in the yeast dataset, regular CLR was used instead of mixed CLR. For GENIE3 and Inferelator, the table of expression data for all mutants was moved out of $\log_2$-space by raising each value to the power of 2, and each gene's expression was normalized to lie in the interval [0,1] by dividing by its maximum expression value over all measurements. The data was treated this way to maintain conformance with the DREAM4 data standards on which Inferelator and GENIE3 were originally

tested. For NetProphet, the values were left as log ratios and the data was normalized such that the standard deviation of each gene over all measurements was clamped (so as not to exceed 1 standard deviation from the mean of all genes' original standard deviations). The normalization was performed in this way for the LASSO regression so that genes with low variance would be given less priority but genes with exceptionally high variance would not dominate the solution. We originally tested GENIE3 and Inferelator using log-ratio expression values but found the accuracy to be better when applying the DREAM4 normalization standard (unlogged expression values on the interval [0,1]).

For the weighted version of NetProphet, we estimated optimal region weights using five rounds of two-fold cross validation, partitioning the training data (interactions labeled as true or false by ChIP evidence) by regulator, and an additional five rounds of two-fold cross validation partitioning the training data by target. The region weights $\boldsymbol{\omega}$ were allowed to take on the following values: [1e-3, 1e-2, 1e-1, 1, 2, 3]. The offset coefficients c_b and c_p were allowed to take on the values: [1e-2, 1e-1, 1]. Weights and offsets were selected so as to maximize the average AU-PRC (area under the precision recall curve) over all rounds of cross validation. The weights selected by cross-validation were: $\boldsymbol{\omega}_I = 3$, $\boldsymbol{\omega}_{II} = 1$, $\boldsymbol{\omega}_{III} = 1$, $\boldsymbol{\omega}_{IV} = 1$, $\boldsymbol{\omega}_B = 2$, $\boldsymbol{\omega}_D = 2$; and the offset coefficients were $c_b = 0.1$ and $c_d = 0.01$.

Chapter 5

Discussion

5.1 Future Directions

An inherent limitation of inferring transcriptional networks from gene expression (transcriptomics) data alone is that the activity levels of transcription factors do not necessarily correlate with the factor's mRNA abundance. The prospect of measuring transcription factor activity on a genome scale is currently infeasible, and thus any method that wishes to model transcription factor activity independently from transcription factor mRNA concentration must resort to some scheme that attempts to infer the activity levels from the gene expression data potentially combined with prior knowledge, such as protein-protein interaction data. Often, when working out the details of a specific pathway, qualitative knowledge of relevant signaling pathways and protein-protein complexes is known. For example, if two transcription factors are known to form a stable complex it may prove more advantageous to combine their expression levels into a single covariate which represents the complex. In addition to incorporating prior knowledge of protein-protein interactions it may also prove useful to allow the activity levels of the transcription factors themselves be treated as latent variables to be inferred or marginalized over. An interesting task,

which turns the network inference problem on its head, is: given the structure of a transcriptional network and a compendium of gene expression data, infer the activity levels of the transcription factors in the network for each measurement. Future approaches to network inference will need to identify better ways of handling the latent activity levels of regulators than approximating them with the mRNA levels of the respective regulator.

5.2 Conclusion

Identifying the structure transcriptional networks and the cis-regulatory logic that governs their behavior is a key step in linking the genotype of an organism to its phenotypes. Our ability to experimentally interrogate the molecular state of a cell's regulatory network has improved dramatically in the past decade with new experimental approaches such as RNA-Seq and ChIP-Seq, which take advantage of the falling cost of high-throughput sequencing. A key determinant of the future success of the network biology enterprise lies in its ability to bring forth new computational approaches that integrate multiple experimental lines of evidence into testable models of biological systems. This dissertation examines and applies strategies for mapping out biological pathways and ways in which these strategies can be improved.

In chapter 2, I identified a lack of rigorous standardized testing of the computational methods that were being proposed for predicting network structure from gene expression measurements. To address this I proposed GRENDEL, a tool for

generating in-silico networks that possess biologically realistic network topologies and kinetic parameterizations. I evaluated several widely used network inference algorithms using GRENDEL and revealed that the relative accuracy of the algorithms was dataset dependent.

In chapter 3, I focused on elucidating a pathway in the fungal pathogen *Cryptococcus neoformans* that regulates capsule size in response to its surrounding environment. I applied conventional analyses to high throughput datasets to identify a missing regulator of this response, Ada2. Ada2 was discovered by identifying a transcriptional signature that contains genes whose expression significantly correlated with the capsule radius over a range of growth conditions. I then used phenotypic analysis, RNA sequencing, and chromatin-immunoprecipitation sequencing (ChIP-Seq) to situate Ada2 in the complex network of genes that regulate capsule and other cryptococcal virulence factors.

In chapter 4 I presented a novel computational method, NetProphet, which was designed to analyze a large compendium of gene expression measurements from cells in which individual transcription factors have been genetically removed. I demonstrated this computational approach by applying it to map out the transcriptional network of *Saccharomyces cerevisiae*. By comparing against hundreds of ChIP-chip experiments that directly probe the structure of the transcriptional network we found that NetProphet achieves accuracy superior to alternative network inference algorithms. In addition, NetProphet identified hundreds

of novel interactions that were not detected in previous experiments. We presented compelling evidence that many of these novel interactions are likely to be real due to their functional context and support by evidence based on the sequence specificity of the transcription factor.

Taken together, these three chapters represent a major step towards making network inference algorithms both accessible and practical for widespread use. I believe future work in this area will continue to adopt the theme of integrating analyses within and between experimental data types to synthesize more accurate and comprehensive renderings of biological networks.

Appendix A

Effects of translational inhibition on capsule synthesis

The cryptococcal capsule becomes enlarged during infection. This response is mediated by signaling pathways that sense the host environment and detect nutrient limitation, increases in temperature and CO_2. These signaling pathways post translationally regulate the activity levels of transcription factors that initiate a transcriptional program, which culminates in capsule synthesis and exocytosis. This program likely has multiple layers forming a transcriptional cascade. The genes that are direct targets of transcription factors whose activity is post-translationally regulated by the upstream signaling pathways we refer to as the *first responders*. Altered expression levels of genes in the set of first responders that code for transcription factors would in turn impact their direct targets (the second responders) and so on.

The goal of this experiment was to identify the *first responders* by measuring the transcriptional response to capsule induction while blocking the translational machinery. By blocking the translational machinery, all of the observed responses to capsule induction would be caused by proteins which were produced prior to translational inhibition. Thus, the cascade of responses which occur as a result of

capsule induction would effectively be halted at the first stage in which expression of *first responders* would be altered, but not second or third responders.

First we conducted 24 hour growth curves of H99 cells incubated with varying concentrations of cycloheximide (CHX), a translational inhibitor, to identify concentrations at which cells did not grow. Cells were inoculated from an overnight culture in 50 ml of YPD at 0.1 OD/ml. The following concentrations of CHX were tested: 10, 12.5, 25, 50, 100, 200 µg/ml. It was found that concentrations of 100 to 200 ug/ml were completely fungistatic and even concentrations down to 10 ug/ml of CHX cells showed less than 1 division over 24 hours. To assess viability, cells were washed twice in H_2O and spotted on YPD agar. Cells were viable at all tested concentrations of CHX.

Next we examined the effect of CHX on capsule formation. Overnight H99 cells were inoculated at 1 OD/ml in 10 ml of DMEM + 20 ul of CHX stock (10mg/ml in H_2O) in aerated tissue culture flasks. 4 cultures were prepared in this experiment: CHX (-/-), CHX (-/+), CHX (+/-) and CHX(+/+), where x/y indicates the treatment for the first 9 hours of growth (x) and last 17 hours (y); + indicates with CHX and – indicates without. After 9 hours of growth at 37ºC + 5% CO_2 all cultures were washed twice and resuspended in their respective treatment. Figure A.1 reveals capsule size of each culture after 26 hours. Cells grown continuously in the presence of CHX (+/+) formed no detectable capsule whereas cells grown continuously without CHX (-/-) presented normal capsule sizes, confirming that inhibition of

translation is sufficient to block capsule formation. Cells grown without CHX then with (-/+) revealed a capsule size intermediate to that of (-/-) and (+/+). Most interestingly, cells which were grown for the first 9 hours in the presence of CHX and then transferred to CHX free DMEM (+/-) produced extremely large capsules, nearly twice the size of cells grown continuously without CHX.

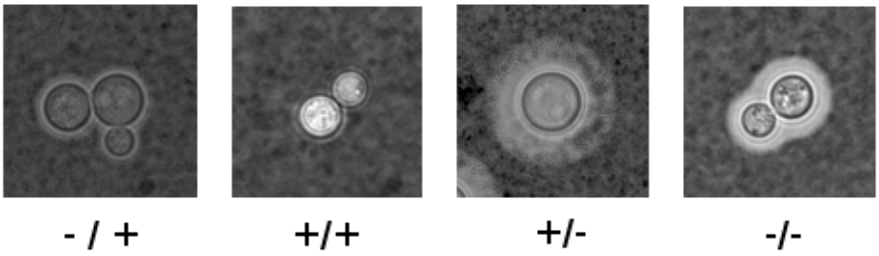

Figure A.1 H99 cells grown for 9 hours with or without cycloheximide (CHX) then washed and grown for another 17 hours with or without CHX.

Next, we transcriptionally profiled cells before and after capsule induction with or without CHX by RNA-Seq. Cells were taken from an overnight culture and inoculated in 4 separate aerated tissue culture flasks with 80 ml DMEM at 1 OD/ml. Cultures were grown at 30ºC for two hours. After two hours, 40 ul of CHX at 10mg/ml was added to two cultures (CHX+). All cultures were incubated at 30ºC for another 30 minutes. Next, one CHX- and one CHX+ culture were grown at 37ºC, 5% CO_2 and the other two were grown at 30ºC for 90 minutes. Samples were spun down for 5 minutes and cell pellets were frozen in a methanol dry ice bath. Total RNA was isolated by lyophilyzation followed by vortexing with glass beads and

Trizol extraction. Total RNA was prepared into an Illumina library and sequenced resulting in 4 million reads per sample. Sequenced reads were aligned to the H99 reference sequence and differential expression was assessed by fold change in transcript coverage (after normalizing by sample sequencing depth).

The differential expression analysis comparing non-induced versus induced cells identified 442 genes upregulated and 311 down-regulated. The same comparison for cells treated with CHX revealed 230 genes up-regulated and 271 down-regulated. The intersection of these two differentially expressed sets (with and without CHX inhibition), which represents the *first responders*, was 33 genes up-regulated and 70 genes down-regulated. No previously capsule implicated genes were identified as *first responders*. Putative transcriptional regulators identified as *first responders* included CNAG_04093, CNAG_04345 and CNAG_04837, all of which were down regulated. Examining the effects of CHX in both inducing and non-inducing conditions we found large similar (77% overlap) transcriptional responses with 3141 genes differentially expressed at 30ºC and 2585 at 37ºC + 5% CO_2. Genes repressed in the presence of CHX were functionally representative of proteolytic processes and genes that were activated were heavily involved in ribosome biogenesis. This response suggests a reaction to the presence of CHX that attempts to extend the life of existing proteins and produce more ribosomes to cope with inability to synthesize new proteins.

Overall we found the set of first responders to be quite small and contained no genes which were previously implicated in capsule formation. It is possible that the treatment by cycloheximide produces such a dramatic response relative to the cells reaction to growth in capsule inducing conditions that the *first responders* are not adequately detected.

Preliminary evaluation of RNA-Seq in *Cryptococcus neoformans*

To evaluate the ability of RNA-Seq to accurately and reproducibly quantify gene expression in the fungal pathogen *Cryptococcus neoformans*, we performed RNA-Seq on KN99α cells grown in capsule non-inducing (DMEM, 30ºC) and inducing (DMEM, 37ºC, 5% CO_2) conditions on three separate days. Cells for each day of induction were prepared independently starting from freezer stocks and grown on YPD agar for approximately 7 days. For each day, cells were inoculated in triplicate from an overnight culture into 20 ml DMEM in an aerated tissue culture flask. 6 cultures grown for 90 minutes in non-inducing and inducing conditions and total RNA was isolated using the Trizol extraction protocol. Illumina libraries were prepared from total RNA in two separate batches. In library 1, days 1 and 2 were prepared and in library 2, days 2 and 3 were prepared. Thus, the same total RNA from day 2 (6 samples) was prepared in two separate libraries. The 24 samples were pooled and sequenced yielding approximately 12 +/- 6 million reads per sample.

Sequences were aligned against the H99 reference sequence and quantified using Tophat and Cufflinks.

An analysis of the coefficient of variation for each of the cryptococcal genes was performed for each set of 3 biological replicates revealing a median coefficient of variation of 0.12 for most sets of biological replicates. We performed a clustergram analysis using Matlab, and determined that the top-level grouping neatly separates the samples by treatment (inducing vs non-inducing). Beyond this top-level grouping there was no preference to group samples by library prep or day of growth.

Next, we performed a comparison of RNA-Seq expression levels to those of Nanostring when quantifying the same samples. 25 genes were selected to be measured by Nanostring according to estimates from initial RNA-Seq experiments of their CoV. The set of 25 genes was selected to span the range from high to low CoV, with the genes of highest CoV being genes with little to no expression. Within this set of 25 genes, two housekeeping genes were selected: ACT1 and PDA1. Nanostring quantifications were taken from the same lysates as the RNA-Seq samples by reserving a fraction of the upper aqueous phase of each Trizol extraction. The expression values of the Nanostring quantifications were compared against RNA-Seq after normalizing by the arithmetic mean of the two housekeeping genes PDA1 and ACT1. This normalization seemed to be necessary for Nanostring but not RNA-Seq, when using the quartile normalization option for Cufflinks. Figure A.2 shows that that overall the Nanostring and RNA-Seq are in agreement about the

relative abundances levels of all 25 genes over these samples. However for individual genes, only 5 genes revealed an R^2 greater than 0.75 between the two technologies. These 5 genes were also clearly differentially expressed between the inducing and non-inducing conditions, showing the largest fold changes of all genes identified as differentially expressed by either technology. The remaining 20 genes include the housekeeping genes and other genes which showed little variation, genes which were not reliably detected by Nanostring and genes which appear to be differentially expressed according to Nanostring but not by RNA-Seq (about 5 genes). One gene which is quantified well above the negative controls for Nanostring is clearly differentially expressed by RNA-Seq, but not Nanostring. This suggests that that for large changes, both technologies are able to detect the differences but for smaller differences in gene expression, one technology may be able to detect better than the other (for reasons unknown) and this does not appear to be one sided (ie each technology can reliably detect differences that the other cannot).

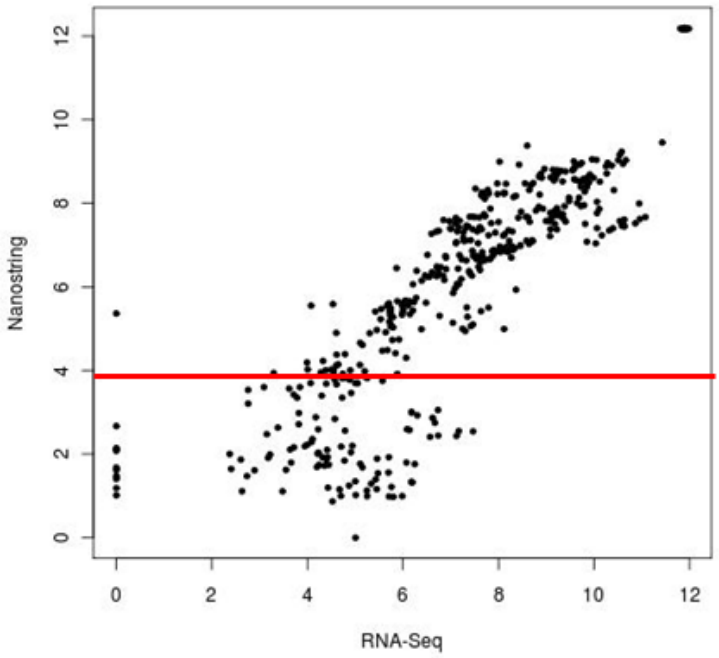

Figure A.2 A scatter plot of normalized RNA-Seq expression values against Nanostring expression values for 25 measured genes over 18 samples. The red line denotes the coverage level of negative control probes for Nanostring, measurements below this line are considered undetected by Nanostring.

Comparing the coefficient of variation for the 25 genes quantified by both measurement technologies, we found that on average, Nanostring produces a slightly smaller CoV (Figure A.3). Additionally, comparing CoV for genes from across the two conditions they are somewhat preserved within a technology, suggesting there is something intrinsic about the gene, independent of these two conditions that causes its CoV to be systematically higher (or lower).

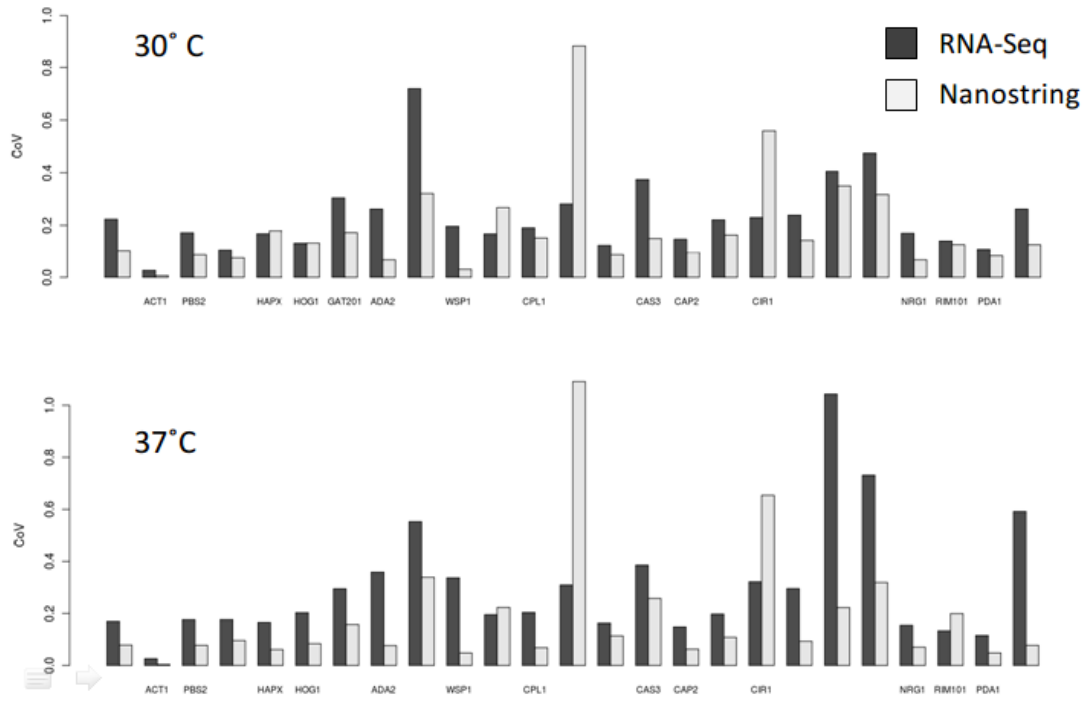

Figure A.3 CoV comparison of RNA-Seq and Nanostring over 25 measured genes in capsule non-inducing (top) and inducing (bottom) conditions. Gene names appear at the bottom of the bars with blanks for unnamed genes.

Finally we performed differential expression analysis using Cuffdiff of inducing versus non-inducing conditions within 3 independent growth days and the technical replicate of the library prep for growth day 2, resulting in 4 sets of differentially expressed genes (Figure A.4). Interestingly, sets 1_1 and 2_2 have roughly the same number of genes in common as sets 2_1 and 2_2. Given that sets 2_1 and 2_2 were prepared from the same total RNA this suggests that the principle source of noise is library prep specific and not due to biological differences between days of induction or technical variation in the RNA isolation.

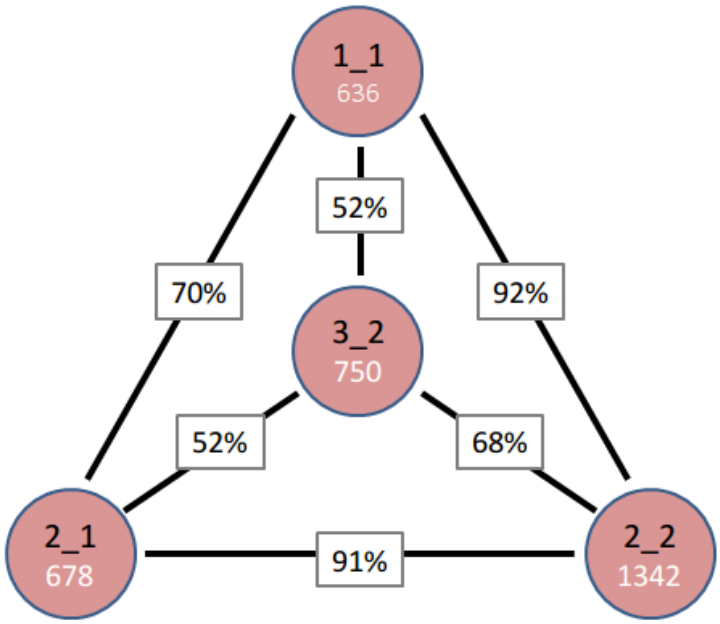

Figure A.4 Differentially expressed gene set comparison. Differentially expressed gene sets are denoted as salmon circles labeled at X_Y, with X representing the day of growth and Y representing the library prep. The number of genes within each set is represented at the bottom of each circle and the boxes contain the percent overlap between each set pair as a fraction of the smaller set.

References

1. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (New York, N.Y.) 270: 467-70.

2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature biotechnology 14: 1675-80. doi:10.1038/nbt1296-1675

3. Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, et al. (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature methods 5: 613-9. doi:10.1038/nmeth.1223

4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628. doi:nmeth.1226 [pii] 10.1038/nmeth.1226

5. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 95: 14863-14868.

6. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, et al. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in

a mammalian cellular context. BMC Bioinformatics 7 Suppl 1: S7. doi:1471-2105-7-S1-S7 [pii] 10.1186/1471-2105-7-S1-S7

7. Husmeier D (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics (Oxford, England) 19: 2271-82.

8. Goutsias J, Lee NH (2007) Computational and experimental approaches for modeling gene regulatory networks. Current pharmaceutical design 13: 1415-36.

9. Braunstein A, Paganani A, Weigt M, Zecchina R (2008) Gene-network inference by message passing. Journal of Physics 95.

10. Kim SY, Imoto S, Miyano S (2003) Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in bioinformatics 4: 228-35.

11. Stolovitzky G, Monroe D, Califano A (2007) Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Annals of the New York Academy of Sciences 1115: 1-22. doi:10.1196/annals.1407.021

12. Smith VA, Jarvis ED, Hartemink AJ (2003) Influence of network topology and data collection on network inference. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 164-75.

13. Zak D, Doyle F, Gonye G, Schwaber J (2001) Simulation studies for the identification of genetic networks from cDNA array and regulatory activity data. In: Proceedings of the Second International Conference on Systems Biology. pp. 231-238.

14. Mendes P, Sha W, Ye K (2003) Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics (Oxford, England) 19 Suppl 2: ii122-9.

15. Bulcke T Van den, Leemput K Van, Naudts B, Remortel P van, Ma H, et al. (2006) SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC bioinformatics 7: 43. doi:10.1186/1471-2105-7-43

16. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, et al. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 5: e8. doi:06-PLBI-RA-0740R3 [pii] 10.1371/journal.pbio.0050008

17. Agrawal H (2002) Extreme self-organization in networks constructed from gene expression data. Phys Rev Lett 89: 268702.

18. Chen G, Larsen P, Almasri E, Dai Y (2008) Rank-based edge reconstruction for scale-free genetic regulatory networks. BMC Bioinformatics 9: 75. doi:1471-2105-9-75 [pii] 10.1186/1471-2105-9-75

19. Hatzimanikatis V, Lee KH (1999) Dynamical analysis of gene networks requires both mRNA and protein expression information. Metabolic engineering 1: 275-81. doi:10.1006/mben.1999.0115

20. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics (Oxford, England) 19: 524-31.

21. Hoops S, Sahle S, Gauges R, Lee C, Pahle J, et al. (2006) COPASI--a COmplex PAthway SImulator. Bioinformatics (Oxford, England) 22: 3067-74. doi:10.1093/bioinformatics/btl485

22. Funahashi A, Morohashi M, Kitano H, Tanimura N (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. Biosilico 4: 159-162.

23. Machné R, Finney A, Müller S, Lu J, Widder S, et al. (2006) The SBML ODE Solver Library: a native API for symbolic and fast numerical analysis of reaction networks. Bioinformatics (Oxford, England) 22: 1406-7. doi:10.1093/bioinformatics/btl086

24. Ramsey S, Orrell D, Bolouri H (2005) Dizzy: stochastic simulation of large-scale genetic regulatory networks. Journal of bioinformatics and computational biology 3: 415-36.

25. Fuente A de la, Bing N, Hoeschele I, Mendes P (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. Bioinformatics (Oxford, England) 20: 3565-74. doi:10.1093/bioinformatics/bth445

26. Laubenbacher R, Stigler B (2004) A computational algebra approach to the reverse engineering of gene regulatory networks. Journal of theoretical biology 229: 523-37. doi:10.1016/j.jtbi.2004.04.037

27. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, et al. (2005) Multiple-laboratory comparison of microarray platforms. Nature methods 2: 345-50. doi:10.1038/nmeth756

28. Barabási A-L, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5: 101-13. doi:10.1038/nrg1272

29. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. BioEssays□: news and reviews in molecular, cellular and developmental biology 20: 433-40. doi:10.1002/(SICI)1521-1878(199805)20:5<433::AID-BIES10>3.0.CO;2-2

30. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of Escherichia coli. Nature genetics 31: 64-8. doi:10.1038/ng881

31. Barabasi A, Albert R (1999) Emergence of scaling in random networks. Science (New York, N.Y.) 286: 509-12.

32. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. Journal of molecular biology 360: 213-27. doi:10.1016/j.jmb.2006.04.029

33. Hill A (1910) The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. J Physiol 40: iv-vii.

34. Hofmeyr JH, Cornish-Bowden A (1997) The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. Computer applications in the biosciences□: CABIOS 13: 377-85.

35. Belle A, Tanay A, Bitincka L, Shamir R, O'Shea EK (2006) Quantification of protein half-lives in the budding yeast proteome. Proceedings of the National Academy of Sciences of the United States of America 103: 13004-9. doi:10.1073/pnas.0605420103

36. García-Martínez J, Aranda A, Pérez-Ortín JE (2004) Genomic run-on evaluates transcription rates for all yeast genes and identifies gene regulatory mechanisms. Molecular cell 15: 303-13. doi:10.1016/j.molcel.2004.06.004

37. Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-41. doi:10.1038/nature02046

38. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717-28.

39. Heitman J, Kozel TR, Kwon-Chung J, Perfect J, Casadevall A (2011) Cryptococcus, from human pathogen to model yeast. Washington, D.C.: ASM Press. 620 p.

40. Giles SS, Dagenais TR, Botts MR, Keller NP, Hull CM (2009) Elucidating the pathogenesis of spores from the human fungal pathogen Cryptococcus neoformans. Infect Immun 77: 3491-3500. doi:IAI.00334-09 [pii] 10.1128/IAI.00334-09

41. Garcia-Hermoso D, Janbon G, Dromer F (1999) Epidemiological evidence for dormant Cryptococcus neoformans infection. J Clin Microbiol 37: 3204-9.

42. Park BJ, Wannemuehler KA, Marston BJ, Govender N, Pappas PG, et al. (2009) Estimation of the current global burden of cryptococcal meningitis among persons living with HIV/AIDS. AIDS (London, England) 23: 525-30.

43. Gomez BL, Nosanchuk JD (2003) Melanin and fungi. Curr Opin Infect Dis 16: 91-96.

44. Cox GM, Mukherjee J, Cole GT, Casadevall A, Perfect JR (2000) Urease as a virulence factor in experimental cryptococcosis. Infect Immun 68: 443-8.

45. Cox GM, McDade HC, Chen SC, Tucker SC, Gottfredsson M, et al. (2001) Extracellular phospholipase activity is a virulence factor for Cryptococcus neoformans. Mol Microbiol 39: 166-175.

46. Okagaki LH, Strain AK, Nielsen JN, Charlier C, Baltes NJ, et al. (2010) Cryptococcal cell morphology affects host cell interactions and pathogenicity. PLoS Pathog 6: e1000953.

47. Zaragoza O, García-Rodas R, Nosanchuk JD, Cuenca-Estrella M, Rodríguez-Tudela JL, et al. (2010) Fungal cell gigantism during mammalian infection. PLoS Pathog 6: e1000945.

48. Doering TL (2009) How sweet it is! Capsule formation and cell wall biogenesis in Cryptococcus neoformans. Annu Rev Microbiol 63: 223-247.

49. Rivera J, Feldmesser M, Cammer M, Casadevall A (1998) Organ-dependent variation of capsule thickness in Cryptococcus neoformans during experimental murine infection. Infect Immun 66: 5027-30.

50. Vartivarian SE, Anaissie EJ, Cowart RE, Sprigg HA, Tingler MJ, et al. (1993) Regulation of cryptococcal capsular polysaccharide by iron. J Infect Dis 167: 186-90.

51. Littman ML (1958) Capsule synthesis by Cryptococcus neoformans. Trans N Y Acad Sci 20: 623-48.

52. Granger DL, Perfect JR, Durack DT (1985) Virulence of Cryptococcus neoformans. Regulation of capsule synthesis by carbon dioxide. J Clin Invest 76: 508-16.

53. Clancy CJ, Nguyen MH, Alandoerffer R, Cheng S, Iczkowski K, et al. (2006) *Cryptococcus neoformans var. grubii* isolates recovered from persons with AIDS demonstrate a wide range of virulence during murine meningoencephalitis that correlates with the expression of certain virulence factors. Microbiology (Reading, England) 152: 2247-55.

54. Chang YC, Kwon-Chung KJ (1994) Complementation of a capsule-deficient mutation of Cryptococcus neoformans restores its virulence. Mol Cell Biol 14: 4912-9.

55. Janbon G, Himmelreich U, Moyrand F, Improvisi L, Dromer F (2001) Cas1p is a membrane protein necessary for the O-acetylation of the Cryptococcus neoformans capsular polysaccharide. Mol Microbiol 42: 453-67.

56. Pukkila-Worley R, Alspaugh JA (2004) Cyclic AMP signaling in Cryptococcus neoformans. FEMS Yeast Res 4: 361-7.

57. D'Souza CA, Heitman J (2001) Conserved cAMP signaling cascades regulate fungal development and virulence. FEMS Microbiol Rev. 25: 349-364.

58. 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: 1147-56.

59. O`Meara TR, Norton D, Price MS, Hay C, Clements MF, et al. (2010) Interaction of Cryptococcus neoformans Rim101 and protein kinase A regulates capsule. PLoS Pathog 6: e1000776.

60. Jung WH, Saikia S, Hu G, Wang J, Fung CK-Y, et al. (2010) HapX positively and negatively regulates the transcriptional response to iron deprivation in Cryptococcus neoformans. PLoS Pathog 6: e1001209.

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

62. Chun CD, Brown JCS, Madhani HD (2011) A major role for capsule-independent phagocytosis-inhibitory mechanisms in mammalian infection by Cryptococcus neoformans. Cell Host Microbe 9: 243-51.

63. Liu OW, Chun CD, Chow ED, Chen C, Madhani HD, et al. (2008) Systematic genetic analysis of virulence in the human fungal pathogen *Cryptococcus neoformans*. Cell 135: 174-88.

64. Bahn Y-S, 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: 2285-300.

65. Zhang S, Hacham M, Panepinto J, Hu G, Shin S, et al. (2006) The Hsp70 member, Ssa1, acts as a DNA-binding transcriptional co-activator of laccase in Cryptococcus neoformans. Mol Microbiol 62: 1090-101.

66. Chang YC, Miller GF, Kwon-Chung KJ (2003) Importance of a developmentally regulated pheromone receptor of Cryptococcus neoformans for virulence. Infect Immun 71: 4953-60.

67. Gerik KJ, Bhimireddy SR, Ryerse JS, Specht CA, Lodge JK (2008) PKC1 is essential for protection against both oxidative and nitrosative stresses, cell integrity, and normal manifestation of virulence factors in the pathogenic fungus Cryptococcus neoformans. Eukaryot Cell 7: 1685-1698.

68. O`Meara TR, Hay C, Price MS, Giles S, Alspaugh JA (2010) Cryptococcus neoformans histone acetyltransferase Gcn5 regulates fungal adaptation to the host. Eukaryot Cell 9: 1193-202.

69. Koutelou E, Hirsch CL, Dent SYR (2010) Multiple faces of the SAGA complex. Curr Opin Cell Biol 22: 374-82.

70. Wang P, Nichols CB, Lengeler KB, Cardenas ME, Cox GM, et al. (2002) Mating-type-specific and nonspecific PAK kinases play shared and divergent roles in Cryptococcus neoformans. Eukaryot Cell 1: 257-72.

71. Hicks JK, Bahn Y-S, Heitman J (2005) Pde1 phosphodiesterase modulates cyclic AMP levels through a protein kinase A-mediated negative feedback loop in Cryptococcus neoformans. Eukaryot Cell 4: 1971-81.

72. Grant PA, Duggan L, Côté J, Roberts SM, Brownell JE, et al. (1997) Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. Genes Dev 11: 1640-50.

73. Balasubramanian R, Pray-Grant MG, Selleck W, Grant PA, Tan S (2002) Role of the Ada2 and Ada3 transcriptional coactivators in histone acetylation. J Biol Chem 277: 7989-95.

74. Nielsen K, Cox GM, Wang P, Toffaletti DL, Perfect JR, et al. (2003) Sexual cycle of Cryptococcus neoformans var. grubii and virulence of congenic a and alpha isolates. Infect Immun 71: 4831-41.

75. Grant PA, Eberharter A, John S, Cook RG, Turner BM, et al. (1999) Expanded lysine acetylation specificity of Gcn5 in native complexes. J Biol Chem 274: 5895-900.

76. Huisinga KL, Pugh BF (2004) A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13: 573-85.

77. Sellam A, Askew C, Epp E, Lavoie H, Whiteway M, et al. (2009) Genome-wide mapping of the coactivator Ada2p yields insight into the functional roles of SAGA/ADA complex in Candida albicans. Mol Biol Cell 20: 2389-400.

78. Johnsson A, Xue-Franzén Y, Lundin M, Wright APH (2006) Stress-specific role of fission yeast Gcn5 histone acetyltransferase in programming a subset of stress response genes. Eukaryot Cell 5: 1337-46.

79. Xue-Franzén Y, Johnsson A, Brodin D, Henriksson J, Bürglin TR, et al. (2010) Genome-wide characterisation of the Gcn5 histone acetyltransferase in budding yeast during stress adaptation reveals evolutionarily conserved and diverged roles. BMC Genomics 11: 200.

80. Missall TA, Pusateri ME, Donlin MJ, Chambers KT, Corbett JA, et al. (2006) Posttranslational, translational, and transcriptional responses to nitric oxide stress in Cryptococcus neoformans: implications for virulence. Eukaryot Cell 5: 518-529.

81. Brown SM, Campbell LT, Lodge JK (2007) Cryptococcus neoformans, a fungus under stress. Curr Opin Microbiol 10: 320-325.

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

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

84. Ko Y-J, Yu YM, Kim G-B, Lee G-W, Maeng PJ, et al. (2009) Remodeling of global transcription patterns of Cryptococcus neoformans genes mediated by the stress-activated HOG signaling pathways. Eukaryot Cell 8: 1197-217.

85. Roh T-young, Ngau WC, Cui K, Landsman D, Zhao K (2004) High-resolution genome-wide mapping of histone modifications. Nat Biotechnol 22: 1013-6.

86. Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in S. cerevisiae. PLoS Biol 3: e328.

87. Ma P, Wera S, Dijck P Van, Thevelein JM (1999) The PDE1-encoded low-affinity phosphodiesterase in the yeast Saccharomyces cerevisiae has a specific function in controlling agonist-induced cAMP signaling. Mol Biol Cell 10: 91-104.

88. Serrano R, Bernal D, Simón E, Ariño J (2004) Copper and iron are the limiting factors for growth of the yeast Saccharomyces cerevisiae in an alkaline environment. J Biol Chem 279: 19698-704.

89. Helmlinger D, Marguerat S, Villén J, Gygi SP, Bähler J, et al. (2008) The S. pombe SAGA complex controls the switch from proliferation to sexual differentiation through the opposing roles of its subunits Gcn5 and Spt8. Genes Dev 22: 3184-95.

90. Jacobson S, Pillus L (2009) The SAGA subunit Ada2 functions in transcriptional silencing. Mol Cell Biol 29: 6033-45. doi:10.1128/MCB.00542-09

91. Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M, et al. (2004) Global position and recruitment of HATs and HDACs in the yeast genome. Mol Cell 16: 199-209.

92. Pokholok DK, Harbison CT, Levine S, Cole M, Hannett NM, et al. (2005) Genome-wide map of nucleosome acetylation and methylation in yeast. Cell 122: 517-27.

93. Chikamori M, Fukushima K (2005) A new hexose transporter from Cryptococcus neoformans: molecular cloning and structural and functional characterization. Fungal Genet Biol 42: 646-55.

94. 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: 771-81.

95. Dromer F, Mathoulin S, Dupont B, Brugiere O, Letenneur L (1996) Comparison of the efficacy of amphotericin B and fluconazole in the treatment of cryptococcosis in human immunodeficiency virus-negative patients: retrospective analysis of 83 cases. French Cryptococcosis Study Group. Clin Infect Dis 22 Suppl 2: S154-60.

96. 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: 776-785.

97. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3: Article3.

98. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116-21.

99. Nelson RT, Hua J, Pryor B, Lodge JK (2001) Identification of virulence mutants of the fungal pathogen *Cryptococcus neoformans* using signature-tagged mutagenesis. Genetics 157: 935-947.

100. Fu J, Hettler E, Wickes BL (2006) Split marker transformation increases homologous integration frequency in Cryptococcus neoformans. Fungal Genet Biol 43: 200-212. doi:S1087-1845(06)00023-5 [pii] 10.1016/j.fgb.2005.09.007

101. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England) 25: 1105-11.

102. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-5.

103. Lo K, Gottardo R (2007) Flexible empirical Bayes models for differential gene expression. Bioinformatics (Oxford, England) 23: 328-35.

104. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10: R25.

105. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

106. Aparicio O, Geisberg JV, Struhl K (2004) Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.] Chapter 17: Unit 17.7. doi:10.1002/0471143030.cb1707s23

107. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. Science 316: 1497-1502. doi:1141319 [pii] 10.1126/science.1141319

108. Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 418-29.

109. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, et al. (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol 7: R36. doi:gb-2006-7-5-r36 [pii] 10.1186/gb-2006-7-5-r36

110. Marbach D, Mattiussi C, Floreano D (2009) Combining multiple results of a reverse-engineering algorithm: application to the DREAM five-gene network challenge. Annals of the New York Academy of Sciences 1158: 102-13. doi:10.1111/j.1749-6632.2008.03945.x

111. Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian networks to analyze expression data. Journal of computational biology□: a journal of

computational molecular cell biology 7: 601-20. doi:10.1089/106652700750050961

112. Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press, London. p.

113. Pe'er D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. Bioinformatics (Oxford, England) 17 Suppl 1: S215-24.

114. Yoo C, Thorsson V, Cooper GF (2002) Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 498-509.

115. Hu Z, Killion PJ, Iyer VR (2007) Genetic reconstruction of a functional transcriptional regulatory network. Nature genetics 39: 683-7. doi:10.1038/ng2012

116. Pinna A, Soranzo N, Fuente A de la (2010) From knockouts to networks: establishing direct cause-effect relationships through graph analysis. PloS one 5: e12912. doi:10.1371/journal.pone.0012912

117. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, et al. (2010) Revealing strengths and weaknesses of methods for gene network inference.

Proceedings of the National Academy of Sciences of the United States of America 107: 6286-91. doi:10.1073/pnas.0913357107

118. Yip KY, Alexander RP, Yan K-K, Gerstein M (2010) Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. PloS one 5: e8121. doi:10.1371/journal.pone.0008121

119. 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: e9803. doi:10.1371/journal.pone.0009803

120. 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: e13397. doi:10.1371/journal.pone.0013397

121. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods. PloS one 5. doi:10.1371/journal.pone.0012776

122. Haynes BC, Brent MR (2009) Benchmarking regulatory network reconstruction with GRENDEL. Bioinformatics 25: 801-807. doi:btp068 [pii] 10.1093/bioinformatics/btp068

123. Marbach D, Schaffter T, Mattiussi C, Floreano D (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering

methods. Journal of computational biology : a journal of computational molecular cell biology 16: 229-39. doi:10.1089/cmb.2008.09TT

124. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, et al. (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PloS one 5: e9202. doi:10.1371/journal.pone.0009202

125. Madar A, Greenfield A, Ostrer H, Vanden-Eijnden E, Bonneau R (2009) The Inferelator 2.0: a scalable framework for reconstruction of dynamic regulatory network models. Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference 2009: 5448-51. doi:10.1109/IEMBS.2009.5334018

126. Schaffter T, Marbach D, Floreano D (2011) GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics (Oxford, England) 27: 2263-70. doi:10.1093/bioinformatics/btr373

127. Olson D, Delen D (2008) Advanced Data Mining Techniques. Springer. 138 p.

128. Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM (2010) Comprehensive reanalysis of transcription factor knockout expression data in

Saccharomyces cerevisiae reveals many new targets. Nucleic acids research 38: 4768-77. doi:10.1093/nar/gkq232

129. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Current opinion in structural biology 14: 283-91. doi:10.1016/j.sbi.2004.05.004

130. Monteiro PT, Mendes ND, Teixeira MC, D'Orey S, Tenreiro S, et al. (2008) YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. Nucleic acids research 36: D132-6. doi:10.1093/nar/gkm976

131. Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, et al. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucleic acids research 34: D446-51. doi:10.1093/nar/gkj013

132. Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, et al. (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. Nucleic acids research 39: D136-40. doi:10.1093/nar/gkq964

133. Li X-Y, Thomas S, Sabo PJ, Eisen MB, Stamatoyannopoulos JA, et al. (2011) The role of chromatin accessibility in directing the widespread, overlapping

patterns of Drosophila transcription factor binding. Genome biology 12: R34. doi:10.1186/gb-2011-12-4-r34

134. Spivak AT, Stormo GD (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. Nucleic acids research 40: D162-8. doi:10.1093/nar/gkr1180

135. Grant CE, Bailey TL, Noble WS (2011) FIMO: scanning for occurrences of a given motif. Bioinformatics (Oxford, England) 27: 1017-8. doi:10.1093/bioinformatics/btr064

136. Egriboz O, Jiang F, Hopper JE (2011) Rapid GAL gene switch of Saccharomyces cerevisiae depends on nuclear Gal3, not nucleocytoplasmic trafficking of Gal3 and Gal80. Genetics 189: 825-36. doi:10.1534/genetics.111.131839

137. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99-104. doi:10.1038/nature02800

138. Tang L, Liu X, Clarke ND (2006) Inferring direct regulatory targets from expression and genome location analyses: a comparison of transcription factor deletion and overexpression. BMC genomics 7: 215. doi:10.1186/1471-2164-7-215

139. Wang H, Mayhew D, Chen X, Johnston M, Mitra RD (2011) Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. Genome research 21: 748-55. doi:10.1101/gr.114850.110

140. Voth WP, Yu Y, Takahata S, Kretschmann KL, Lieb JD, et al. (2007) Forkhead proteins control the outcome of transcription factor binding by antiactivation. The EMBO journal 26: 4324-34. doi:10.1038/sj.emboj.7601859

141. Zhang L, Guarente L (1994) Evidence that TUP1/SSN6 has a positive effect on the activity of the yeast activator HAP1. Genetics 136: 813-7.

142. Conlan RS, Gounalaki N, Hatzis P, Tzamarias D (1999) The Tup1-Cyc8 protein complex can shift from a transcriptional co-repressor to a transcriptional co-activator. The Journal of biological chemistry 274: 205-10.

143. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423: 241-54. doi:10.1038/nature01644

144. Green SR, Johnson AD (2004) Promoter-dependent roles for the Srb10 cyclin-dependent kinase and the Hda1 deacetylase in Tup1-mediated repression in Saccharomyces cerevisiae. Molecular biology of the cell 15: 4191-202. doi:10.1091/mbc.E04-05-0412

145. Dakshinamurthy A, Nyswaner KM, Farabaugh PJ, Garfinkel DJ (2010) BUD22 affects Ty1 retrotransposition and ribosome biogenesis in Saccharomyces cerevisiae. Genetics 185: 1193-205. doi:10.1534/genetics.110.119115

146. Biggar SR, Crabtree GR (1999) Continuous and widespread roles for the Swi-Snf complex in transcription. The EMBO journal 18: 2254-64. doi:10.1093/emboj/18.8.2254

147. Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, et al. (2009) Backup in gene regulatory networks explains differences between binding and knockout results. Molecular systems biology 5: 276. doi:10.1038/msb.2009.33

148. Haynes BC, Skowyra ML, Spencer SJ, Gish SR, Williams M, et al. (2011) Toward an integrated model of capsule regulation in Cryptococcus neoformans. PLoS pathogens 7: e1002411. doi:10.1371/journal.ppat.1002411

149. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic acids research 36: 5221-31. doi:10.1093/nar/gkn488

150. Peeters R, Westra R (2004) On the identification of sparse gene regulatory networks. In: In Proc. 16th Int. Symp. on Math. Theory of Networks.

151. Steinke F, Seeger M, Tsuda K (2007) Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models. BMC systems biology 1: 51. doi:10.1186/1752-0509-1-51

152. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of Roy. Stat. Soc. B 58: 267–288.

153. Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann. Statist. 32: 407–499.

154. Killion PJ, Sherlock G, Iyer VR (2003) The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). BMC bioinformatics 4: 32. doi:10.1186/1471-2105-4-32

Vita

Brian Clifton Haynes

Date of Birth	July 8, 1981
Place of Birth	Memphis, Tennessee
Degrees	Ph.D. Computer Science, Washington University in St Louis May 2012
	M.S. Computer Science, Washington University in St Louis May 2007
	B.S. Cum Laude, Computer Science, University of Memphis May 2003
Honors	Spencer T. and Ann W. Olin Award (for superior achievement in biomedical research), Washington University School of Medicine (2012)
	Genome Analysis Training Program Fellowship, National Human Genome Research Institute (2008-2010)
	Distinguished Masters Fellowship, Washington University (2004)
Publications	Haynes B.C., Kramer M.H., and Brent M.R. (2012) NetProphet: A practical method for mapping transcriptional regulatory networks. (in review)
	Chiappinelli K.B., Haynes B.C., Brent M.R., Goodfellow P.J. (2012) Reduced DICER1 elicits an interferon response in endometrial cancer cell lines. *Molecular Cancer Research*.
	Haynes B.C., Skowyra M.L., Spencer S.J., Gish S.R., Williams M., Held E.P. Brent M.R., and Doering, T.L. (2011) Toward an integrated model of capsule regulation in *Cryptococcus neoformans*. *PLoS Pathogens*. 7(12):e1002411. (F1000 article factor: 6; top 2% of published articles in biology and medicine)

Kumar P., Yang M., Haynes B.C., Skowyra M.L., and Doering T.L. (2011) Emerging themes in cryptococcal capsule synthesis. *Current Opinions in Structural Biology* (5):597-602.

Haynes B.C. and Brent M.R. (2009) Benchmarking regulatory network reconstruction with GRENDEL. *Bioinformatics* 25(6):801-807.

Graesser A.C., Olney A., Haynes B.C., & Chipman P. (2005) AutoTutor: A cognitive system that simulates a tutor that facilitates learning through mixed-initiative dialogue. In C. Forsythe, M.L. Bernard, & T.E. Goldsmith (Eds.), Cognitive systems: Human cognitive models in systems design. Mahwah, NJ: Erlbaum.

Graesser A.C., Chipman P., Haynes B.C., & Olney A. (2005) AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions in Education*, 48, 612-618.

Graesser A.C., Person N., Haynes B.C., VanEck R., Adcock A. (2004) AutoTutor has Tutorial Dialog in Natural Language, Interactive Simulation, and Lesson Authoring Tools. Conference Proceedings of 2004 American Educational Research Association Annual Meeting (AERA).

May 2012