Winter 1-1-2012

Improving Thermodynamic Models of Transcription by Combining ChIP and Expression Measurements of Synthetic Promoters

Robert D. Zeigler
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Computational and Systems Biology

Dissertation Examination Committee
Barak Cohen, Chair
Michael Brent
Justin Fay
James Havranek
Rohit Pappu
Gary Stormo

Improving Thermodynamic Models of Transcription by Combining ChIP and Expression Measurements of Synthetic Promoters

by

Robert David Zeigler

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

December 2012

St. Louis, Missouri

Table of Contents

Acknowledgements

I will be forever grateful to my thesis mentor, Barak Cohen. At the beginning of my PhD, I was looking for a mentor who could see the big picture, who would ask the hard questions, and who would be able to teach me to do the same. I found that and more in Barak. He has taught me how to think about science, how to be a scientist, and how to focus on the questions that really matter.

I would also like to thank my thesis committee, Gary Stormo, Michael Brent, Rohit Pappu, Jim Havranek, and Justin Fay, for their valuable insights and feedback on my project. I would like to thank Jay Gertz for laying the groundwork for this project and hooking me on combinatorial regulation; Justin Gerke for fun times and great scientific discussions; Kim Lorenz for many helpful discussions about wet lab; Ilaria Mogno for showing me how a truly efficient scientist works; Mike White for many hours of discussion and advice on western blotting, my project, and life; Priya Sudarsanam, for being a great bench-mate and entertaining my crazy ideas; Jamie Kwasnieski for her insights in my project and for valuable feedback on my thesis; Marc Sherman for helpful discussions on math and probability distributions; Linda Riles, for her infinite wisdom in all things yeast and her infectious enthusiasm for life and science; Josh Witten, Christina Chen, Chris Fiore, and Brett Maricque for helpful comments and discussions, I would like to thank Aaron Spivak for many hours of scientific discussion and shared ChIP protocol debugging, and Yue Zhao for many helpful discussions on science and computational biology, and for critical reading of the thesis manuscript.

I would like to thank my parents for always encouraging me to fulfill my potential and for instilling in me two important values for being a scientist: curiosity about the world around me, and a work ethic.

Most of all, I would like to thank my wife for supporting me through this arduous but rewarding endeavor. I am certain that she did not know what I was getting us into when I started a PhD. I will be forever grateful for her help, her encouragement during the many difficult times, and her loving support.

This work was supported by grants from the American Cancer Society (RSG-06-032-01-GMC), the National Science Foundation (0543156), and the National Institute of Health (52978).

To my wife, Amey,
for her loving support that made this possible.
And to our children,
for their understanding and enthusiasm.

Abbreviations

AAS: Amino Acid Starvation

ChIP: Chromatin immunoprecipitation

IN: ChiP input/control

IP: Precipitated material

PCR: Polymerase Chain Reaction

ReLoS: Regulatory Logic Simulator

RNAP: RNA Polymerase II

TF: Transcription Factor

TFBS: Transcription Factor Binding Sites

TSS: Transcription Start Start

ABSTRACT OF THE DISSERTATION

Improving Thermodynamic Models of Transcription by Combining ChIP and Expression Measurements of Synthetic Promoters

By Robert Zeigler

Doctor of Philosophy in Biology and Biomedical Sciences

Computational Biology

Washington University in St. Louis, 2012

Associate Professor Barak A. Cohen, Chairperson

Regulation of gene expression is a fundamental process in biology. Accurate mathematical models of the relationship between regulatory sequence and observed expression would advance our understanding of biology.

I developed ReLoS, a regulatory logic simulator, to explore mathematical frameworks for describing the relationship between regulatory sequence and observed expression and to explore methods of learning combinatorial regulatory rules from expression data. ReLoS is a flexible simulator allowing a variety of formalisms to be applied. ReLoS was used to explore the question of how complex rules of combinatorial transcriptional regulation must be to explain the complexity of transcriptional regulation observed in biology. A previously published dataset was analyzed for regulatory elements that explained the behavior of regulatory modules for 254 genes in 255 conditions. I found that ReLoS was able to recapitulate a reasonable fraction of the variation (mean gene-wise correlation of 0.7) with only twelve combinatorial rules comprising

13 *cis*-regulatory elements. This result suggested that learning the combinatorial rules of transcriptional regulation should be possible.

State ensemble statistical thermodynamic models are a class of models used to describe combinatorial transcriptional regulation. One way to parameterize these models is measuring the expression of a reporter gene driven by many similar promoters . Models parameterized in this fashion do better at explaining the sequence to expression relationship, but fail to distinguish between multiple biological mechanisms that give rise to equivalent expression results in the synthetic promoters, thus limiting the generalizability of the models. I developed a ChIP-based strategy for quantitatively measuring the relative occupancy of transcription factors on synthetic promoters. This data complements existing methods for obtaining expression data from the same promoters. Comparison of models parameterized with only expression, only occupancy, or expression and occupancy reveals specific biological details that are missed when considering only expression data. In particular, the occupancy data suggests that differential regulatory effects of Cbf1 in glucose versus amino acid are a function of how it interacts with polymerase rather than changes in concentration or binding affinity. Additionally, the occupancy data suggests that Gcn4 binds in a cooperative manner and that Gcn4 occupancy is adversely affected by the presence of a nearby Nrg1 site. Finally, the occupancy data and expression data taken together suggest that Gcn4 binds in competition with another transcription factor.

Synthesizing disparate sources of information resulted in an improved understanding of the mechanics of transcriptional regulation of the synthetic promoters and was ultimately largely successful in decoupling the DNA binding energies from the TF interactions with polymerase. However, it suggests that more sophisticated models of the relationship between occupancy and

expression may be required in at least some cases. Incorporating different sources of data into models of regulation will continue to be important for learning the biological specifics that drive expression changes.

CHAPTER 1: Introduction

Transcriptional regulation is an important biological process

Every organism must respond to changes in the environment to survive. One way these responses occur is through changes in the complement and level of genes being expressed (Gardner and Barald, 1991; Matikainen, et al., 2001; Radinsky, 1995; Owuor and Kong, 2002; Driscoll-Penn, Galgoli and Greer, 1983). The level at which a gene is expressed can be regulated at several points, but one of the major points where regulation occurs is transcription (Giniger, Varnum, and Ptashne, 1985). Transcription is regulated by the coordinated action of DNA-binding proteins called transcription factors (TFs) (Guarente, *et al.*, 1982). These proteins recognize specific DNA sequences and recruit additional factors such as protein complexes associated with RNA Polymerase II (Brent and Ptashne, 1981 and Brent and Ptashne,1985), repressive complexes (Schuller, 2003; Zhou and Winston, 2001), and chromatin remodeling complexes (Morillon A, *et al.*, 2003; Moreau, *et al*., 2003) to ultimately increase or decrease the number of mRNA transcripts being produced for a particular gene. The process of multiple TFs binding to the DNA and causing a change in expression is collectively referred to herein as combinatorial *cis*-regulation.

Although the general features of this process have long been known (Giniger, Varnum, and Ptashne, 1985 and Anderson, Ptashne, and Harrison, 1985), the ability to quantitatively model the phenomenon remains elusive. However, the need for such models has never been greater. With the advent of next generation sequencing, the genomes of many more organisms are available (Mikkelsen, *et al*., *2005;* Warren, et al. 2008; Hellsten, 2010). A complete understanding of the information in any genome will require the ability to parse the sequence and

determine which pieces of DNA function in a regulatory capacity, and what that capacity is. Such a mapping of regulatory elements has many potential applications beyond annotation, including engineering novel biological circuits and personalized medicine.

Many good models of TF site-specificity now exist

The first step to map the the regulatory landscape is to determine the location of the regulatory elements. The advent of genome-wide technologies such as ChIP-chip (Harbison, *et al.*, 2004 and Lee, Johnstone and Young, 2006), and ChIP-seq (Johnson and Mortazavi *et al*., 2007; Jothi, *et al*., 2008) for mapping *in vivo* binding events has uncovered the binding site preferences of many TFs *in vivo*. These approaches have been complemented by *in vitro* methods for learning TF binding site preferences such as protein binding microarrays (Berger, *et al*. 2006 and Mukherjee and Berger, *et al*., 2004), and SELEX (Liu and Stormo, 2005 and Tuerk and Gold, 1990). These techniques have led to rapid growth in our knowledge of the TF-specific binding preferences for many TFs. This information is extremely useful, but it does not tell us about the strength or direction of regulation of the transcription factors. For that, models which relate the TF binding site information to expression must be used.

Modeling the sequence to expression relationship is difficult using native genomic sequences

To date, attempts at modeling the sequence to expression relationship have been attempted in several organisms (Beer and Tavazoie, 2004, Segal, *et al*., 2008; Bussemaker, Li, and Siggia, 2001; Das, Banerjee, and Zhang, 2004; Vilar, 2010). These attempts have used a

variety of mathematical formalisms, from regression models (Bussemaker, Li, and Siggia, 2001; Das, Banerjee, and Zhang, 2004) to Bayesian networks (Beer and Tavazoie, 2004), to differential equations (Vu and Vohradsky, 2007). In particular, models of regulation wherein the level of gene expression driven by a particular piece of DNA is predicted directly from the DNA have generally performed poorly at a genome-wide scale (Bussemaker, Li, and Siggia, 2001; Das, Banerjee, and Zhang, 2004; Irie, *et al*., 2011 and Xiao, Segal, 2009). There are several issues that complicate the study of combinatorial *cis*-regulation in the genome.

The first issue is the presence of additional confounding variables. In the genome, each gene is subject to a different set of kinetic parameters following transcription initiation. Each gene can have its own transcription rate (Pelechano, Chávez, and Pérez-Ortín, 2010) and translation rate (Reuveni, *et al*., 2011 and Gingold and Pilpel, 2011), and genes can be regulated post-transcriptionally(Filipowicz, Bhattacharyya, Sonenberg, 2008) and post-translationally (Kuras, *et al.*, 2002). Additionally, each gene is surrounded by a different genomic context that incorporates information such as the genomic coordinates and the natural nucleosome content of the region, each of which has been shown to have an influence on the level of gene expression (Bernstein, B.E., *et al*., 2004 and Woo and Li, 2011). All of these factors make it difficult to distinguish between changes that occur as a result of differences in the composition of TF binding sites of the sequence and changes that occur for sequence-independent reasons. Ultimately, these confounding factors will need to be accounted for in a complete model of regulation, but for the purpose of understanding combinatorial *cis*-regulation, these variables complicate the problem.

The second complicating issue is the sheer size of combinatorics available to the genome relative to the available observations. For instance, *Saccharomyces cerevisiae* has approximately 200 transcription factors (Beskow and Wright, 2006). Considering every possible pairwise interaction but without regard to complications such as spacing and orientation, there are over 20,000 possible combinations of binding sites, with only about 5,800 genes. The problem is worse when considering more combinations or organisms with more transcription factors. The use of multiple related genomes may help mitigate this problem somewhat (FitzGerald, *et al*., 2006) but rapid binding site turnover (Bradley, *et al*. 2010 and Moses, *et al*. 2006) and different sets of regulators across genomes complicate the comparison.

With so many possible combinations, it is fair to ask whether it is even possible to create general-purpose models of regulation, or whether every sequence will be its own special case, requiring functional dissection by experimental methods. In addition, there are many possible mathematical formalisms for describing *cis*-regulatory interactions. Which of these formalisms is best-suited for learning the rules of regulation remains an open question. This leads directly to hypothesis one:

(H1) It will be possible to explain the complexity of biology with relatively simple, generalizable mathematical rules.

We may also pose a related question:

(Q1) Which formalism is best-suited for learning the rules of regulation?

State ensemble models of transcriptional control provide biophysically motivated parameters

State ensemble statistical thermodynamic models have been increasingly used to describe transcriptional regulation (Buchler, Gerland and Hwa, 2003; Granek and Clarke, 2005; Raveh-Sadka, Levo and Segal, 2009; Segal, *et al*., 2008; Shea and Ackers, 1985; and Wasson and Hartemink, 2009). In these models, a promoter is modeled as a series of distinct states. Each state consists of the DNA and the proteins bound to the DNA in that state. Each state is associated with a statistical mechanical weight (the Boltzmann weight), which is the exponent of the sum of the binding energies in play in the state times the concentrations of the factors bound in the state:

$$W = e^{-\Sigma\Delta G/RT} * \Pi[TF]$$

where the product is over each bound transcription factor in the state and the the sum is over all binding energies in the state. These binding energies include the affinity of TFs for the DNA as well as protein-protein interactions. The sum of the weights of all states is the partition function:

$$Z = \Sigma W$$

The probability that RNA Polymerase II is bound is then computed as the sum of all states in which polymerase is bound to the DNA divided by the partition function:

$$\Sigma W * \delta(Pol) / Z$$

where δ(Pol) is one if Polymerase is bound in the state and zero otherwise. Expression is then modeled as a function of the probability of polymerase bound. A common assumption is that the

relationship is linear (Gertz and Cohen, 2009), but other means of mapping have also been used (Segal, *et al*, 2008 and He, *et al.*, 2010).

The advantage of the thermodynamic model is its focus on the biophysical properties of the system. Since the parameters are a set of binding energies, the relative values of those parameters provide information on the relative binding strengths of the TFs to the DNA and the relative importance of interaction with RNAP and other recruited factors. Given the set of parameters, it is possible to rewrite the thermodynamic function to calculate any related quantity of interest, such as how many copies of a particular transcription factor are expected to be bound to the DNA. In contrast, models such as regression use arbitrary coefficients to describe the contribution of each sequence element to expression. These coefficients may produce a predictive model, but the parameters are agnostic as to the mechanism. The biophysical and biochemical information provided by the thermodynamic model makes it attractive, but parameterizing these models remains difficult due to computational complexity and the difficulties with genomic promoters discussed above.

Statistical thermodynamic models of transcriptional regulation can be parameterized using synthetic promoters

An alternative approach to using genomic data is the use of synthetic promoters (Cox, Surette, and Elowitz, 2007; Gertz and Cohen, 2009; Gertz, Siggia and Cohen, 2009; Kwasnieski and Mogno, *et al*., 2012; Ligr, *et al*., 2006; Melnikov, *et al*., 2012; Murphy, Balazsi, and Collins, 2007; Patwardhan, 2012; and Sharon *et*. *al*., 2012). Synthetic promoters consist of many promoter variants of a common promoter backbone. Each variant drives the expression of a

reporter gene. The reporter gene activity is assayed quantitatively through methods such as flow cytometry. Synthetic promoters simplify the parameter estimation problem by reducing the number of confounding variables and increasing the number of direct observations of closely related sequences and their effects on expression. Previously, synthetic promoters were combined with the thermodynamic description of regulation to great effect (Gertz and Cohen, 2009; Gertz, Siggia, and Cohen, 2009), explaining up to sixty percent of the total variation in expression across multiple environmental conditions. The parameters recovered were predictive of fold-changes in transcription factor concentrations in most cases (Gertz and Cohen, 2009), illustrating the benefit of using a biophysically motivated model.

The main problem with parameterizing thermodynamic models of transcriptional regulation with only expression data is that without information about the binding of proteins, the model may be missing important mechanistic details. These details matter because they determine the generalizability of the model. The point of synthetic promoters is to study the sequence to expression relationship in a simplified system so that the information can be applied to more complex problems. But if the mechanistic relationship so-derived is incorrect, then the model will fail to generalize to sequences outside of the synthetic promoters.

For example, Gertz and Cohen (2009) built predictive models of condition-specific TF effects by assuming that changes in regulation occur due to changes in TF concentrations. However, equally predictive models can be built by assuming regulatory changes are caused by differences in the interaction between the TF and RNAP. Without additional data, these two models cannot be distinguished, but they make different predictions concerning the rest of the genome. The model that assumes a change in TF concentration predicts lower TF occupancy

across the genome, leading to global regulatory changes. A model that assumes a change in TF-RNAP changes will predict the same occupancy across the genome. Moreover, the change in TF-RNAP interaction could be local, due to other interacting factors. A good example of this is Cbf1, where Cbf1-dependent activation of MET genes is mediated by the coordinated binding of Met28 and depends on the presence of an upstream RYAAT motif (Siggers, *et al*. 2011; Kuras, *et al*., 1996). In other genes, Cbf1 acts to recruit other factors (Moreau, J.L., *et al.* 2003). Thus, for Cbf1, modeling the change in regulatory affect as a change in the Cbf1-RNAP interaction is more appropriate, and would lead to a model which can be better applied to sequences other than the synthetic promoters. In order to separate these two models, we need additional information.

Although there are several possible sources of additional information, an especially appealing source is *in vivo* binding data. This data is particularly useful because it synthesizes both changes in TF concentration and changes in TF affinity and would provide a direct comparison of changes in how much TF is bound to the promoter versus how much change in regulatory potential occurred. This leads to hypothesis 2:

(H2) Given *in vivo* protein binding data, it will be possible to distinguish between models of transcriptional regulation that yield similar expression results but represent distinct biophysical mechanisms

My work in this thesis aimed to test the above hypotheses and to address the related question to hypothesis 1.

ReLos is a cis-Regulatory Logic Simulator for exploring *cis*-regulatory questions

In chapter two of this thesis, I present my work on ReLos, a *cis*-Regulatory Logic Simulator. Other simulators available at the time ReLos was published (Mendes, Sha, and Ye, 2003; Michaud, Marsh, and Dhurjati, 2003; Van den Bulcke, *et al,* 2006) were primarily designed to model the overall network of regulatory interactions and rarely considered the underlying sequence that connects regulators to genes being regulated. ReLos attempted to address that discrepancy.

With ReLos, a user is able to apply a variety of formalisms for converting sequence to expression. This directly addresses Q1 by allowing a user to simultaneously explore a particular formalism of transcriptional regulation and the effects of specific combinatorial rules on expression driven by any given sequence. By exploring different formalisms and rules, a user can gain a better appreciation for which descriptions are most appropriate for their problem. Similarly, by having a benchmark of sequence-driven regulatory rules, a user can evaluate the ability of various learning algorithms to recover the original rules.

One way to test the ability of ReLos to generate useable data was to compare the ability of the simulator to approximate biology. This was done by crafting a set of rules to mimic the behavior of 11 previously published expression modules comprising 254 genes across 255 conditions (Beer and Tavazoie, 2004). This work provided a direct test of hypothesis H1 by exploring the complexity of the regulatory rules required to reasonably approximate the underlying biology. I found that ReLos could generate data in reasonable agreement with the

expression modules (mean gene-wise correlation of 0.7) with relatively simple rules, suggesting that the underlying biology can be explained by simple, generalizable rules.

Occupancy data combined with expression data helps to deconvolve the parameters of the thermodynamic model

In chapter three of this work, I addressed hypothesis H2 by developing a ChIP-based protocol for acquiring quantitative occupancy data specific to synthetic promoters. I built libraries containing binding sites for Cbf1, Gcn4, Met31/Met32, and Nrg1. Each of these factors is known to be active in one or both of two conditions (glucose and amino acid starvation) used to test the libraries (Zhou and Winston 2001; Kuras, L, *et al.* 1996; Blaiseu, *et al* 1997; Natarajan, *et al*. 2001). I obtained occupancy data for Cbf1 and Gcn4 and expression data for the libraries in both conditions. I used the occupancy data to explore thermodynamic models of TF binding for Cbf1 and Gcn4, and the expression data to examine models of sequence to expression without regard to the occupancy data. Comparing these models revealed interesting differences, such as Gcn4 cooperativity in the binding data. Finally, I combined the occupancy and expression data to build a model that simultaneously relates sequence to expression and sequence to TF occupancy. The results of this model indicate that the occupancy data does help deconvolve the parameters and helps distinguish between different regulatory models. However, the results also indicate that improvements in the integration of the two data sources can be made, possibly by incorporating more sophisticated descriptions of the TF-RNAP interactions into the model.

References

Anderson, J.E., Ptashne, M., Harrison, S.C. “A phage repressor-operator complex at 7 A resolution.” *Nature.* 1985. 316(6029):596-601.

Beer, M.A. and Tavazoie, S. “Predicting gene expression from sequence.” *Cell.* 2004. 117:185-98.

Berger, M.F., *et al.* “Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.” *Nature Biotechnology.* 2006. 24(11):1429-1435.

Bernstein, B.E., *et al*. “Global nucleosome occupancy in yeast.” *Genome Biol.* 2004. 5:R62

Beskow A. and Wright, A. P. H. “Comparative analysis of regulatory transcription factors in Schizosaccharomyces pombe and budding yeasts.” *Yeast* .2006. 23:929–935.

Blaiseau PL, et al. “Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism.” *Mol Cell Biol.* 1997. 17(7): 3640

Bradley, R.K., *et al*. “Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species.” *PLoS Biol.* 2010. 8 (3): e1000343

Brent, R. and Ptashne, M. “A Eukaryotic Transcriptional Activator Bearing the DNA Specificity of a Prokaryotic Repressor.” *Cell.* 1985. 43:729-736.

Brent, R. and Ptashne, M. “Mechanism of action of the /exA gene product.” *Proc. Natl. Acad. Sci. USA.* 1981. 78:4204-4208.

Buchler, N.E., Gerland, U. and Hwa, T. “On schemes of combinatorial transcription logic.” *PNAS.* 2003. 100:5136.

Bussemaker, H.J., Li, H. and Siggia, E.D. “Regulatory element detection using correlation with expression.” *Nat Genet.* 2001. 27:167-174.

Cox, R. S. III, Surette, M. G. and Elowitz, M. B. “Programming gene expression with combinatorial promoters.” *Mol. Syst. Biol.* 2007. 3:145.

Das, D., Banerjee, N. and Zhang, M.Q. “Interacting models of cooperative gene regulation.” *Proc Natl Acad Sci U S A.* 2004. 101:16234-9.

Driscoll-Penn, M. D., Galgoli, B., and Greer, H. “Identification of AAS genes and their regulatory role in general control of amino acid biosynthesis in yeast.” *Proc. Natl. Acad. Sci. USA*. 1983. 80, 2704-2708.

Filipowicz, W., Bhattacharyya, S.N,. Sonenberg, N. "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" *Nat Rev Genet.* 2008. 9(2):102-14.

FitzGerald, P.C., *et al.* "Comparative genomics of Drosophila and human core promoters." *Genome Biol.* 2006. 7(7):R53

Gardner, C.A. and Barald, K.F. “The cellular environment controls the expression of engrailed-like protein in the cranial neuroepithelium of quail-chick chimeric embryos.” *Development* (Cambridge, England). 1991. 113: 1037–1048

Gertz, J.G. and Cohen, B.A. “Environment-specific combinatorial cis-regulation in synthetic promoters.” *Molecular Systems Biology.* 2009. 5:244.

Gertz, J.G., Siggia, E.D., Cohen, B.A. “Analysis of combinatorial cis-regulation in synthetic and genomic promoters.” *Nature.* 2009. 457:215.

Gingold, H and Pilpel, Y. "Determinants of translation efficiency and accuracy." *Mol Syst Biol.* 2011. 7:481.

Giniger, E., Varnum, S. M., and Ptashne, M. “Specific DNA binding of GAL4, a positive regulatory protein of yeast.” *Cell.* 1985. 40:767-774.

Granek, J. A. and Clarke, N.D. “Explicit equilibrium modeling of transcription-factor binding and gene regulation.” *Genome Biology.* 2005. 6:R87 doi:10.1186/gb-2005-6-10-r87

Guarente, L., *et al.* “A mutant lambda repressor with a specific defect in its positive control function.” *Proc. Natl. Acad. Sci. USA.* 1982. 79:2236-2239.

Harbison, C.T. *et al*. “Transcriptional regulatory code of a eukaryotic genome.” *Nature.* 2004. 431:99-104.

He, X. *et al*. “Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression.” *PLoS Comput Biol.* 2010 6(9): e1000935. doi:10.1371/journal.pcbi.1000935

Hellsten, U., *et al.* "The genome of the Western clawed frog Xenopus tropicalis." *Science.* 2010. 328(5978):633

Irie, T., *et al*. "Predicting promoter activities of primary human DNA sequences." *Nucleic Acids Res.* 2011. 39(11):e75

Johnson, D.S., Mortazavi, A. *et al*. “Genome-wide mapping of in vivo protein–DNA interactions.” *Science.* 2007. 316:1497–1502

Jothi *et al*. “Genome-wide identification of in vivo protein–DNA binding sites from ChIP-seq data.” *Nucl Acids Res.* 2008. 36(16) 5221–5231.

Kuras, L, *et al.* “Dual regulation of the met4 transcription factor by ubiquitin-dependent degradation and inhibition of promoter recruitment.” *Mol Cell.* 2002. 10(1):69-80

Kuras, L, *et al.* “A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism.” *EMBO J*. 1996. 15(10):2519.

Kwasnieski, J. C. and Mogno I. *et al*. “Complex Effects of Nucleotide Variants in a Mammalian cis-Regulatory Element.” *PNAS*. 2012. In Press.

Lee, T.I., Johnstone, S.E. and Young, R.A. “Chromatin immunoprecipitation and microarray-based analysis of protein location.” *Nat Protoc.* 2006. 1:729-748.

Ligr, M., *et al*. “Gene expression from random libraries of yeast promoters.” *Genetics*. 2006. 172:2113–2122.

Liu, J. and Stormo, G.D. “Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions.” *Nucleic Acids Res.* 2005. 33(17): e14

Matikainen T. *et al.* “Aromatic hydrocarbon receptor-driven Bax gene expression is required for premature ovarian failure caused by biohazardous environmental chemicals.” *Nat Genet.* 2001. 28: 355–360.

Melnikov, A. *et al*. “Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.” *Nature Biotechnology*. 2012. 30:271–277.

Mendes, P., Sha, W., and Ye, K. “Artificial gene networks for objective comparison of analysis algorithms” *Bioinformatics*, 2003. 19 Suppl 2:II122.

Michaud, D.J., Marsh, A.G., and Dhurjati, P.S., “eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods.” *Bioinformatics*, 2003. 19:1140.

Mikkelsen, T. S., *et al.* "Initial sequence of the chimpanzee genome and comparison with the human genome." *Nature.* 2005 437(7055): 69-87.

Moreau, J.L., *et al.* "Regulated displacement of TBP from the PHO8 promoter in vivo requires Cbf1 and the Isw1 chromatin remodeling complex." *Mol Cell*. 2003. 11(6):1609-20

Morillon A, *et al*. "Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II." *Cell* 2003. 115(4):425-35

Moses, A.M., *et al*. a'Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila.a' *PLoS Comput Biol*. 2006. 2(10): e130

Mukherjee, S. and Berger, M.F., *et al*. "Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays." *Nature Genetics.* 2004. *36(12):1331-1339. Epub 2004 Nov 14.*

Murphy, K. F., Balazsi, G. and Collins, J. J. "Combinatorial promoter design for engineering noisy gene expression." *Proc. Natl Acad. Sci. USA*. 2007. 104:12726–12731.

Natarajan K, et al. "Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast." *Mol Cell Biol.* 2001. 21(13):4347.

Patwardhan, R.P. et al. "Massively parallel functional dissection of mammalian enhancers in vivo." *Nat Biotechnol.* 2012. 30(3):265-70. doi: 10.1038/nbt.2136.

Pelechano, V., Chávez, S. and Pérez-Ortín, J.E. "A Complete Set of Nascent Transcription Rates for Yeast Genes." *PLoS ONE.* 2010. 5(11): e15442. doi:10.1371/journal.pone.0015442

Raveh-Sadka, T., Levo, M., and Segal E. "Incorporating Nucleosomes into Thermodynamic Models of Transcription Regulation." *Genome Res*. 2009. 19:1480-1496.

Reuveni S., *et al.* Genome-Scale Analysis of Translation Elongation with a Ribosome Flow Model. *PLoS Comput Biol.* 2011. 7(9): e1002127. doi:10.1371/journal.pcbi.1002127

Schuller, H.J. "Transcriptional control of nonfermentative metabolism in the yeast Saccharomyces cerevisiae." *Curr Genet*. 2003. 43(3):139-60

Segal, E. *et al*. "Predicting expression patterns from regulatory sequence in Drosophila segmentation." *Nature.* 2008. 451:535-540.

Sharon, E. *et al*. "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters." *Nature Biotechnology.* 2012. 30:521–530. doi: 10.1038/nbt.2205

Shea, M.A. and Ackers, G.K. “The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation.” *Journal of molecular biology.* 1985. 181:211.

Siggers T, Duyzend MH, Reddy J, Khan S, Bulyk ML. “Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex.” *Mol Syst Biol.* 2011. 7:555.

Tuerk C. and Gold, L. “Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.” *Science.* 1990. 249:505–510.

Van den Bulcke, T., et al. “SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms.” *BMC Bioinformatics*. 2006. 7:43.

Vilar, J.M.G. "Accurate prediction of gene expression by integration of DNA sequence statistics with detailed modeling of transcription regulation." *Biophys J.* 2010. 99:2408

Vu, T. T. and Vohradsky, J. "Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae" *Nucleic Acids Research.* 2007. 35:279-287.

Warren, W. C. *et al*. “Genome analysis of the platypus reveals unique signatures of evolution. *Nature*. 2008. 453(7192):175-83.

Wasson, T. and Hartemink, A. “An ensemble model of competitive multi-factor binding of the genome.” *Genome Research*, 2009. 19:2101–2112.

Woo, Y. H. and Li, W-H. “Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes.” *PNAS.* 2011. 108:(8)3306.

Xiao, Y. and Segal, M.R. “Identification of Yeast Transcriptional Regulation Networks Using Multivariate Random Forests.” *PLoS Comput Biol* 5(6): e1000414.

Zhou, H. and Winston, F. “NRG1 is required for glucose repression of the SUC2 and GAL genes of Saccharomyces cerevisiae.” *BMC Genet.* 2001. 2:5

CHAPTER 2: A cis-Regulatory Logic Simulator

Robert D. Zeigler, Jason Gertz and Barak A. Cohen*

Department of Genetics, Washington University School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

*Corresponding Author. Box 8510, Department of Genetics, Washington University School of Medicine, 4444 Forest Park Parkway, St. Louis, MO 63108

Email addresses:
RDZ: rdzeigle@wustl.edu
JG: jgertz@wustl.edu
BAC: cohen@wustl.edu

I designed and wrote the simulator with input from JG. BAC conceived the notion of a promoter-expression simulator. BAC and JG provided useful discussion during the project. This chapter was originally published as "A cis-Regulatory Logic Simulator" in BMC Bioinformatics 2007, **8**:272. .

Abstract

Background

A major goal of computational studies of gene regulation is to accurately predict the expression of genes based on the cis-regulatory content of their promoters. The development of computational methods to decode the interactions among cis-regulatory elements has been slow, in part, because it is difficult to know, without extensive experimental validation, whether a particular method identifies the correct cis-regulatory interactions that underlie a given set of expression data. There is an urgent need for test expression data in which the interactions among cis-regulatory sites that produce the data are known. The ability to rapidly generate such data sets would facilitate the development and comparison of computational methods that predict gene expression patterns from promoter sequence.

Results

We developed a gene expression simulator which generates expression data using user-defined interactions between cis-regulatory sites. The simulator can incorporate additive, cooperative, competitive, and synergistic interactions between regulatory elements. Constraints on the spacing, distance, and orientation of regulatory elements and their interactions may also be defined and Gaussian noise can be added to the expression values. The simulator allows for a data transformation that simulates the sigmoid shape of expression levels from real promoters. We found good agreement between sets of simulated promoters and predicted regulatory modules from real expression data. We present several data sets that may be useful for testing new methodologies for predicting gene expression from promoter sequence.

Conclusions

We developed a flexible gene expression simulator that rapidly generates large numbers of simulated promoters and their corresponding transcriptional output based on specified interactions between cis-regulatory sites. When appropriate rule sets are used, the data generated by our simulator faithfully reproduces experimentally derived data sets. We anticipate that using simulated gene expression data sets will facilitate the direct comparison of computational strategies to predict gene expression from promoter sequence. The source code is available online from http://www.genetics.wustl.edu/bclab/relos/ and as supplementary material. The test sets are available as supplementary material.

Background

Transcriptional regulation of genes is controlled largely through the concerted action of combinations of cis-regulatory sites in the promoters and surrounding regulatory DNA of genes. The interactions between cis-regulatory sites can be complex and may include synergistic [1], competitive [2], and amplifying [3] interactions, and are often influenced by the spacing and orientation of the sites relative to each other and to the transcriptional start site[4, 5]. The complexity of the "cis-regulatory code" makes predicting gene expression from promoter sequence a challenging problem.

Computational approaches for determining the cis-regulatory code include multiple regression models [6], Bayesian networks [7], logic operators [8], and machine learning methods [9]. Though their mathematical frameworks differ, all of these approaches use large-scale transcriptional data (usually microarray-based expression profiling data) and attempt to correlate expression patterns with the presence or absence of computationally predicted cis-regulatory motifs. Currently, we do not have good ways to compare the performance of these different approaches to each other or to new approaches being developed. A serious problem in comparing these methods is the lack of robust test data in which the cis-regulatory interactions underlying the expression data are accurately known. We need data in which the "true" answer is known if we are to compare methodologies. To address this limitation, we built a rule based simulator to create test data sets.

Simulators are playing a useful role in reconstructing gene regulatory networks (GRN). A GRN models the regulatory connections between genes, as opposed to the interactions between cis-regulatory sites in a promoter. Because the true GRN of a cell is not known,

artificially created GRNs are used to evaluate the accuracy of algorithms that attempt to determine network architecture and dynamics[10]. GRN simulators provide test datasets [11, 12], which in turn are used to assess the performance of network reconstruction techniques [13]. We anticipate that gene expression simulators will play a similar role in the development of computational approaches to decipher the interactions between cis-regulatory sites.

We present a regulatory rule simulator that generates random promoters and produces expression data based on user-defined interactions between cis-elements. Whereas a GRN simulator attempts to create a web of genes connected in a biologically relevant manner, our simulator generates promoter regions and predicts the expression from those promoters. We also present test datasets, created by the simulator, which can be used to assess the performance of algorithms that attempt to determine underlying regulatory rules. The promoter generator and simulator, named ReLoS (cis-**Re**gulatory **Lo**gic **S**imulator), are available for download at http://www.genetics.wustl.edu/bclab/relos/relos-dist.zip. A web interface is also available and can be accessed from http://www.genetics.wustl.edu/bclab/relos/. The test data sets are available in supplementary file 1.

Results and Discussion

Simulating Regulatory Rules

Gene expression simulations using Relos are divided into discrete steps (Figure 2.1). The user first specifies the number of cis-regulatory sites that will be part of the simulation. Next, the user creates a rule set that defines the interactions between cis-regulatory sites and their effects on gene expression. Relos then generates a set of promoters consisting of random combinations of these cis-regulatory sites. Finally, the expression of each promoter is determined by applying the rule set to each promoter sequence. The simulator outputs a list of promoter sequences with their corresponding expression values. At every step, the user may specify parameters to customize the simulations.

With Relos a user can encode a wide variety of cis-regulatory rules. The rules are defined in an XML simulation file to make the attributes of the simulation, including the rules, legible to the user. A single rule in a rule-set is defined by the cis-regulatory sites involved, the conditions required by the rule, conditions excluded by the rule, context dependencies for each condition, and the output expression generated by that rule. Logical relationships such as OR, NOT and AND can be expressed in describing interactions between sites. Constraints on the spacing, orientation, and distance of sites from each other can be incorporated into any rule. Rule outputs may be combined in linear and non-linear ways (see Methods). A rule may simply specify the additive contribution of a particular regulatory element, or it may determine the parameters of an epistatic (eg: cooperative, competitive, synergistic, etc.) interaction between elements. Promoters are parsed by each rule in the order in which the rules are specified. When a rule

matches a promoter, however, that rule may specify a set of rules which should be skipped in the analysis of the matched promoter.

Promoter processing by rules is delegated to the "analyzer". The analyzer is responsible for determining whether a rule will affect a promoter, based on the constraints specified for the rule. The analyzer is also responsible for specifying the effect of a rule on the expression of a promoter. Analyzers serve as the central point of extensibility in Relos. For each rule, it is possible to specify a custom analyzer. Relos comes with a regular expression analyzer, which modifies promoter expression if the regular expression is matched. Another analyzer allows user-defined mathematical functions to be used to determine rule outputs. For example, a Hill function [14, 15] might be used to describe cooperativity between sites. The flexibility inherent in the design of Relos allows users to simulate virtually any mode of regulation among cis-regulatory sites.

Real expression data are bounded. At the lower bound, a cell cannot express less than zero copies of a gene. There is also an upper limit of detection in any experimental setup and to the levels of RNA that can be produced when a promoter is fully occupied by the transcriptional machinery and transcribing at the maximum rate. These constraints produce sigmoid expression patterns. For this reason, Relos allows users to sigmoidally transform the output data. Users may explicitly tell Relos to transform the data. In this case, Relos uses a sigmoid transformation centered on the average expression for the simulation (see methods). Using the simulation expression mean to center the transformation allows rule-sets to be compared in terms of the variation present in the parsed promoters. Simulations with large variance will show a spread of values between zero and one. Simulations with little variance will, when transformed, cluster

around the value of 0.5. One consequence of the mean-dependent transformation is that it is impossible to generate a transformed dataset in which all expression is either "on" or "off" since datasets with very little variation will result in midline expression when transformed. Users may therefore specify a rule at the end of the pipeline employing a custom analyzer to transform the data. Relos comes with a SigmoidalTransform analyzer (see Methods) that can be used for this purpose, but users may also provide their own transformations. The SigmoidalTransform analyzer uses four parameters (see Methods) to adjust the shape and scale of the transformation. These parameters are independent of the simulation dataset and determine an absolute scale of expression onto which all rule-sets are mapped. By using a consistent set of parameters, users can compare rule-sets with regard to their strength of expression and compare variances according to where the mean lies in the absolute expression scale. Since this transformation does not depend on the dataset, the absolute scale is arbitrarily determined by the choice of parameters and users should be careful to use rules consistent with the scale determined by the parameters.

In addition to rules, their analyzers and constraints, and transformation parameters, the XML simulation file contains other adjustable attributes for the simulation. For example, after the promoters have been interpreted using the current rule set, Gaussian noise is added by the simulator with a user defined standard deviation. Relos is also capable of generating random promoters based on user-defined properties, such as promoter length, cis-regulatory elements and their frequencies and outputting promoters in either fasta or Relos format. These synthetic promoters can be used directly by the simulator. For more details, see Methods.

Examples of simulated datasets are shown in Figure 2.2. As a visual aid to interpret the output of the simulations, histograms illustrating the distribution of expression values are shown. Figure 2.2a shows the distribution of expression values for 5000 fixed-length random promoters consisting of variable numbers of a single type of cis-regulatory activator site and neutral spacer elements, where all elements are equally probable. The expression is therefore a reflection of the distribution of the activator element. Relos outputs the expected Poisson distribution for expression. Figure 2.2b shows the results from an activator-repressor combination. Because expression is now a function of two inputs, it follows the expected Gaussian distribution. Figure 2.2c shows the results from a synergistic rule set, with noise at 5% of the expression level. In this simulation, each element has a small additive effect on expression individually, but when both regulatory elements are present in the same promoter, a large expression effect is observed. As expected, the result of the simulation is a bimodal distribution, where the second peak represents promoters containing both regulatory elements. Figure 2.2d shows the output of a cooperative interaction, modeled by a Hill function. A Hill function is a transition function of the form:

$$y = \frac{x^n}{(\varphi^n + x^n)}$$

Where x is the input and φ and n are parameters used to adjust the location and steepness of the transition. Hill functions have been used to model biological cooperativity in proteins such as Hemoglobin [14] and in cis-regulatory interactions [15]. In Figure 2.2d, x is the number of cooperative elements, n is 3, and φ is 5. Since the expression is a function of the number of A-elements, and the number of A-elements is distributed according to the Poisson distribution, the

expression pattern should be a function of a Poisson distribution. As expected, the simulator output in Figure 2.2d follows a Poisson distribution with an elongated right tail. This tail represents the high expression of promoters with multiple cooperative sites. See supplementary file 2 for the rule-sets used to create figure 2.2.

Test Datasets

The main motivation for creating the simulator was to synthesize expression datasets for which we know the underlying regulatory rules. These datasets will be necessary to compare the accuracy of different methods that infer cis-regulatory rules because there are no experimental datasets for which the true underlying relationships between cis-regulatory sites are known. We therefore created ten test datasets using different rule-sets. The test datasets vary in the number and types of rules and in the complexity of the rule-set. We have made the datasets and rule sets used to generate them (see supplementary file 1) available in both Relos format and fasta format. We anticipate that the availability of test datasets will allow researchers to evaluate their own methods and compare their methods against commonly used algorithms that deduce regulatory rules from expression data. While the test data we provide will be useful for researchers who want to get started right away testing their rule-finding algorithms, we emphasize that the real power of Relos is the capability it provides to quickly produce custom data sets for algorithm testing. Researchers can now rapidly create their own test datasets to compare the dependency of any method on any particular parameter (number or sites, types of interactions, noisy data).

Comparison to Experimental Data

We simulated the expression of five different regulatory modules comprised of 254 yeast genes described in Beer and Tavazoie [7]. A classification tree was constructed to place each gene into its correct module based on the presence or absence of different regulatory elements. Overall, 80% (204/254) of the promoters were placed into their original module. We then created a rule set based on the classification tree which incorporated "AND", "OR", and "NOT" logic. This rule set was used to simulate expression values for each gene in each of the 255 conditions reported in Beer and Tavazoie (see Methods). The results of the simulation and the observed expression values are shown in figure 2.3. The median gene-wise correlation coefficient between the simulated and experimental expression was 0.78, illustrating that simulated data closely matching observed data can be produced with Relos. These results show that Relos can discriminate between promoters and create biologically relevant data sets.

One noticeable discrepancy between the Relos data and the Beer and Tavazoie data was the noise function. Relos uses Gaussian noise, scaled by the noise-less expression value. This results in a smaller absolute level of noise around expression values close to zero. The Beer and Tavazoie data does not appear to follow this trend; the absolute level of noise around zero is still quite large. Accordingly, we wrote an unscaled noise analyzer that applies unscaled Gaussian noise to simulated data.

We also used the same rule sets defined above to analyze Relos-generated promoters. Completely synthetic promoters were created based on the frequency distributions of the cis-regulatory sites that comprised the five modules we simulated. When the rule set was applied to these computationally derived promoters the five expression patterns from Beer and Tavazoie

were again recapitulated. (see additional file 6) Randomly generated promoters, filtered through Relos, faithfully replicate the observed expression patterns in real data

Conclusions

We sought to create a tool that simulates expression from promoters based on cis-regulatory logic. Because there are examples of additivity, synergism, cooperativity, and competition between regulatory sites we created ways to simulate these interactions in a straightforward manner. The full spectrum of interactions between regulatory sites is not known. We recognize that our knowledge of cellular regulation is still relatively limited and that new types of interactions may appear. We therefore did not want to be limited by preconceived models. With its rule-pipeline and analyzer plug-in architecture, Relos allows for virtually any regulatory model to be implemented.

The ease of specifying regulatory models and the speed with which data can be generated will allow algorithms that predict gene expression from promoter sequence to be comprehensively tested. Algorithms that attempt to determine regulatory logic rules from expression and sequence data can be analyzed for their performance with respect to noise, the number of underlying rules, and the complexity of the interactions between the rules. Furthermore, researchers can study the size of the dataset required for an algorithm to recapitulate the rules and the ability of the algorithm to recapitulate the specified rules, as opposed to alternate rule sets which also correlate with the data. We have used Relos to generate a test dataset for use in such studies. We anticipate that the ability to rapidly generate unlimited quantities of simulated expression data will speed the design and comparison of algorithms to decode the cis-regulatory logic that underlies real patterns of gene expression.

The final arbiter of the performance of cis-regulatory rule-finding algorithms will be how well they capture the trends in real data. Algorithms that perform well on synthetic data sets,

such as those produced by Relos, will not necessarily perform well on biological data. Because experimentally derived data is still of limited quantity and variable quality, extensive testing on synthetic data is the best way to understand the strengths and limitations of specific rule-finding methods. Testing and training on synthetic data avoids over fitting rule finders on the limited quantities of real data that are now available. Testing rule-finding methods on synthetic data sets will clearly be one of the paths forward on the way to decoding the interactions between cis-regulatory sites.

Methods

Promoter Generation

Relos generates a promoter as a set of elements. Each promoter element is associated with a "cis-element" and an orientation. Each cis-element has an identifier (eg: A, Oct4, etc), a sequence, and a frequency (expected occurrence). The sequence is only used for output purposes; all built-in rule processing is done on promoter elements.

Relos supports two modes of promoter generation: exact length and expected length. In exact length mode, a cis-element is selected from the user-specified list of elements by a roulette wheel selection process. The selected element is added to the promoter starting from the position furthest upstream of the transcription start site. The element is added in a sense or anti-sense orientation with equal probability. Element selection and addition continues until the number of elements added equals the user-specified length. Relos does not insert spacer elements between cis-elements. Rather, all cis-elements are treated as spacer elements unless a rule is defined which uses the cis-element in a manner inconsistent with a spacer element (see Rule Specification below).

In expected-length mode, the element frequencies are transformed by:

$$D_i = \frac{d_i}{(1 + \frac{1}{E}) * \sum_{i=1}^{n} d_i}$$

Where D_i is the transformed frequency of the i-th element, d_i is the non-transformed frequency for i-th element, E is the expected promoter length, and n is the number of elements. This results in a distribution of cis-elements that includes a "stop" pseudo-element with probably 1/E. The

distribution sums to one and preserves the relative probabilities of the user-specified elements. Promoter elements are added as in the exact length procedure until the stop element is selected.

Rule Specification

Rules are specified in an XML-based format defined by the expression_rules.dtd document type definition file. Each rule is defined in terms of the cis-elements the rule uses, an optional custom analyzer to use in place of the default Relos analyzer, the "output" (the amount by which the rule will affect the current expression level for the promoter), and the "operation" (the way in which the output will affect the current expression). Rules may also define precluded rules. Precluded rules are those that are prevented from operating on a promoter should the precluding rule match. Rules using the default analyzer, or custom analyzers that rely on the default analyzer, may specify one or more conditions that determine whether a particular element on the promoter "matches" the rule. Conversely, these conditions may "exclude" elements on the promoter that should not match the rule.

Conditions are comprised of the cis-element(s) to consider, the allowed position(s) and required orientation of the element(s), and zero or more contexts. Each context defines a cis-element that must appear in the promoter with the element under consideration in the condition. Contexts may include specification of the spacing between the two elements and the orientation of the "context" element.

More details on rule specification can be found in supplementary file 3.

Promoter Analysis

Relos uses a pipeline to perform rule by rule analysis of the promoters. Typically, promoters are moved through the pipeline in the order in which the rules appear in the simulation

XML file. However, when a precluding rule matches a promoter, Relos prevents the precluded rules from operating on the matched promoter. Rules which define a custom analyzer delegate promoter analysis to the custom module. All other rules delegate promoter analysis to the default analyzer. The default analyzer determines the number of elements in a promoter that match the rule and multiplies the number of matches by the output amount to determine the magnitude of the effect on the current promoter expression. Promoter expression is then affected by this amount according to the operation defined for the current rule. Valid operations include add (new expression equals the current expression plus the output); multiply (new expression equals the current expression times the output); exponentiate (new expression equals the old expression raised to the power of the output); and replace (new expression equals the output). Matching is performed on a promoter element-wise basis. If the attributes and contexts of at least one condition and no exclusions match, an element will be considered a match. When no conditions or exclusions are specified, the element only needs to match one of the cis-elements specified by the rule.

Once all promoters have been through the rule pipeline, a user-specified amount of noise is added to each promoter by replacing the current expression value with a random value X, where the probability of replacing the current expression value with X is given by the Gaussian distribution,

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/(2\sigma^2)}$$

Where μ is the current expression value, $\sigma = \mu * \eta$, and η is the user defined level of noise. The Relos default sets the noise to be 5% of the current expression level.

Relos will also transform the data to fit a sigmoidal curve if specified by the user. For each promoter, the transformed expression value is given by:

$$V_T = \frac{1}{1 + e^{-(V_0 - \mu_T)}}$$

Where V_T is the transformed expression value of a particular promoter, V_0 is the original expression for that promoter, and μ_T is the mean of the untransformed expression of all promoters in the simulation. An alternative method of transformation is provided by adding a transformation rule with a custom analyzer to the end of the pipeline. Relos provides an example of a transforming analyzer in the SigmoidalTransformAnalyzer, which transforms the data according to:

$$V_T = \frac{\phi}{(1 + e^{-\alpha * (V_0 - \gamma)})^{\beta}}$$

Where V_T is the transformed expression value, V_0 is the original expression, α adjusts the slope of the curve at the inflection point, β adjusts the position of the inflection point, γ determines the expected midline expression, and φ scales the resulting transformation.

More details on promoter analysis can be found in supplementary file 3.

Creating Test Dataset

Ten test-set simulations were run. Two hundred promoters, comprised of eight cis-elements selected from a pool of four possible elements (A-D), were generated for each simulation, except for test-set simulation ten. A noise level of 5% of the expression level was used. None of the datasets were subjected to upper or lower bound constraints. The first nine test-set simulation rule sets were comprised of: an additive activator, an activator with spacing and ordering constraints, two synergistic rule sets with spacing constraints, two cooperative rule

sets, a dominant-negative competitive rule set, a dominant positive rule set, and a rule set with constraints on many elements and an enhancer. In the final test-set simulation, two hundred promoters were generated, each comprised of eight cis-elements selected from a pool of eight possible elements (A-H). The final simulation rule set consisted of multiple additive and non-additive effects, incorporating many of the non-additive effects encountered separately in other rule sets. For more details, see supplementary file 1.

Comparison to Experimental Data

Beer and Tavazoie [7] classified 49 transcriptional modules in *S. cerevisiae*. We simulated modules 1, 11, 41, 45, and 49. These modules were chosen because they vary in size, expression outputs, and regulatory complexity. Promoters with no regulatory motifs were removed from the dataset, leaving 254 promoters. Tree regression [16] was performed to determine the best classification tree for separating the promoters into the five transcriptional modules. Input to the classification for each promoter was the presence or absence of each of the 666 proposed motifs and their assigned module. Based on the structure of the classification tree, a general rule set was constructed (additional file 7). The ruleset was then duplicated for each microarray experiment, except the output for each rule was changed to match the average expression for that module. All 254 promoters were used as input sequences for each of the 255 simulations. Tree regression and statistical calculations were performed in R.

We used Relos to generate synthetic promoters based on the frequency of the motifs used in the above rule set. The frequency of each motif was determined in the 254 biological promoters as the number of times each motif occurred divided by the total number of motifs in these promoters. The frequencies of the remaining biological motifs not considered by the

ruleset were conglomerated into a single “Spacer” motif (see additional file 8). Relos was used to generate 1000 promoters which were then analyzed by the same rule set described above, with the addition of an “all spacer” rule.

Authors’ contributions

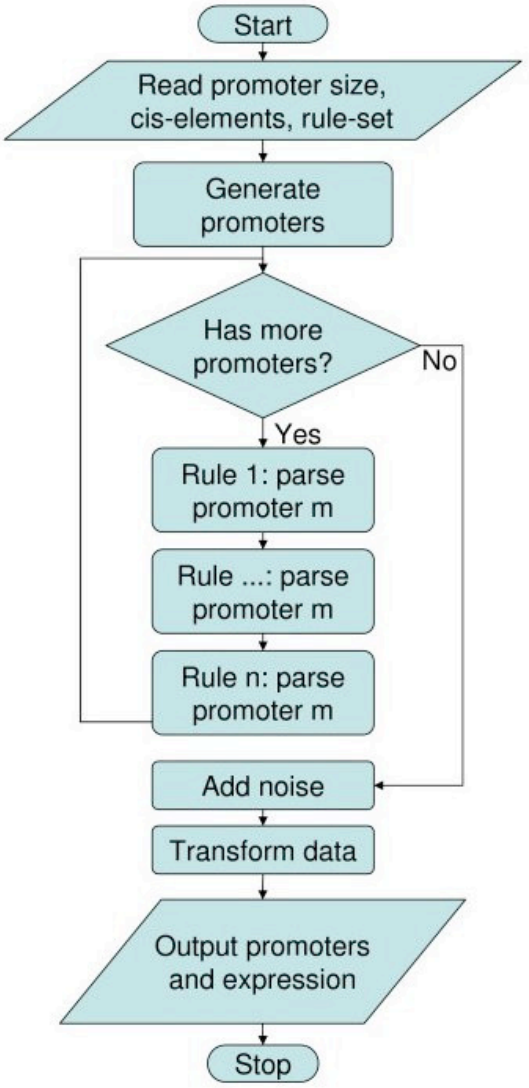
RDZ designed and wrote the simulator with input from JG. BAC conceived the notion of a promoter-expression simulator. BAC and JG provided guidance during the project. RDZ drafted the manuscript with help from BAC and JG. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the Cohen Lab for discussions and critical readings of the manuscript. JG is funded by a National Science Foundation Graduate Fellowship. The project was supported by grants from the American Cancer Society (RSG-06-039-01-GMC) and the National Science Foundation (0543156).

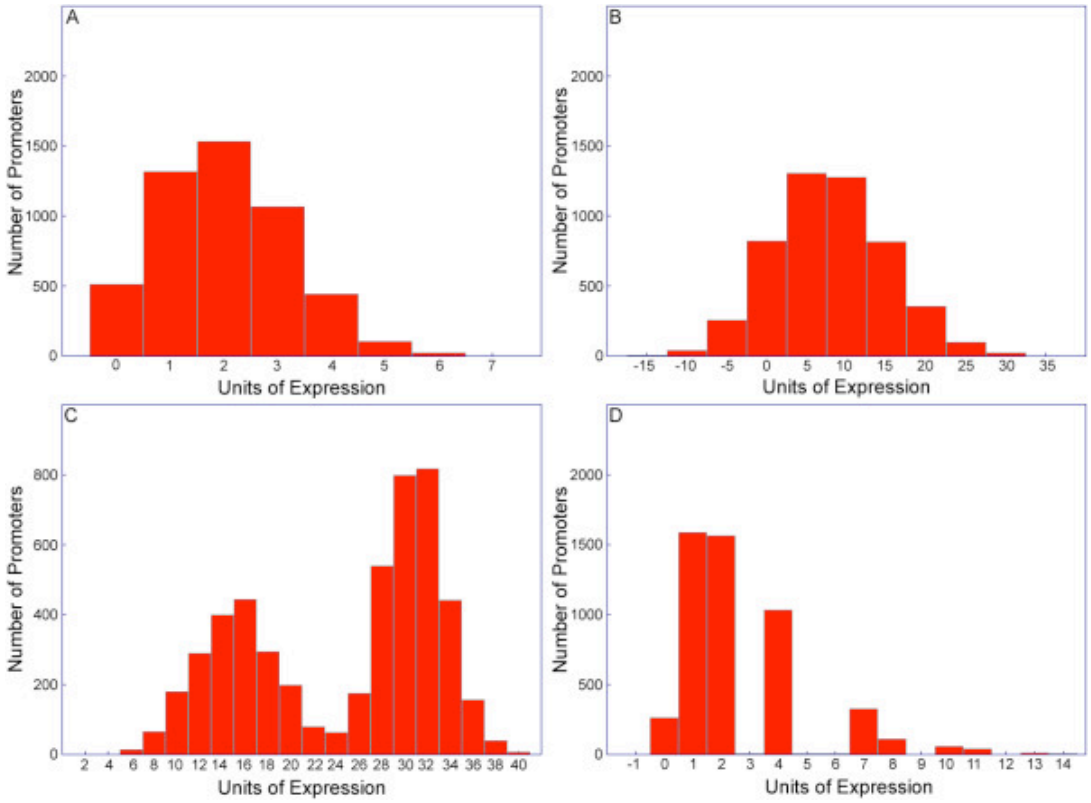
Figures

Figure 2.1: Flow of Relos.



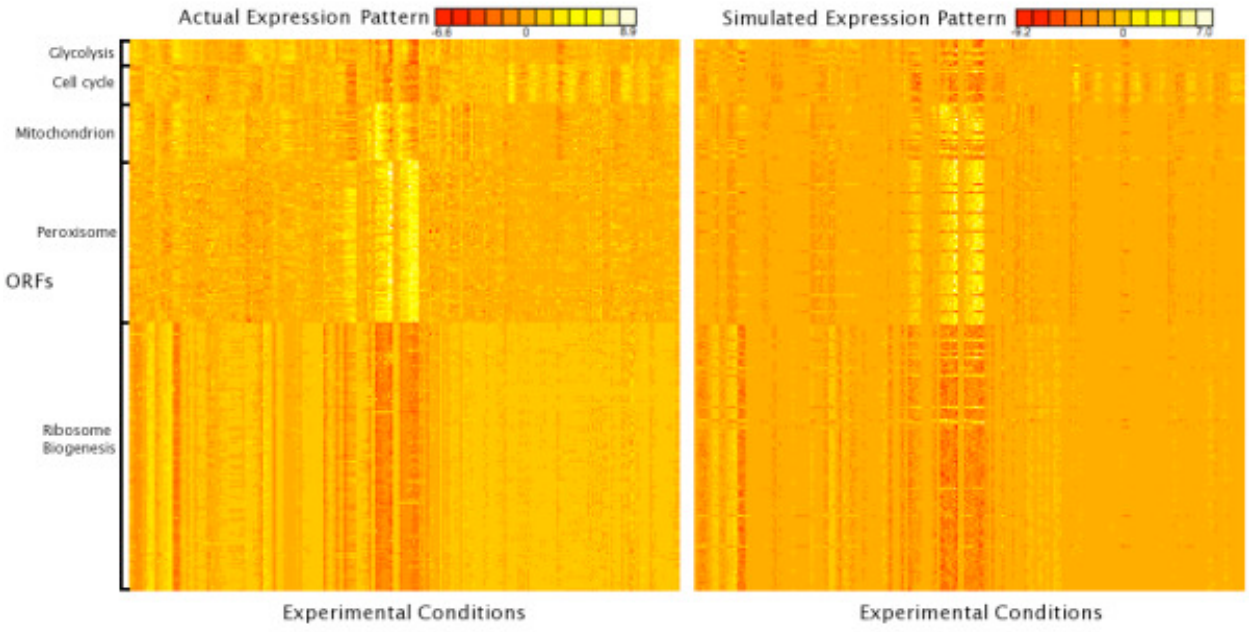
Users supply Relos with cis-elements to use, the number and size of promoters to generate, and the rules used to analyze the promoters. Relos generates the promoters then analyzes the rules by passing the promoters through a rule-pipeline of the user-defined rules. Noise is then added, and the data is optionally transformed via a sigmoidal transform to ensure upper and lower limits of expression.

Figure 2.2: Sample Relos outputs



Relos was used to generate and analyze promoters using four different models. Five thousand promoters were generated in all Figure 2.2 simulations. ***A.*** A simulation that depicts a single activator, modeled as an additive rule. ***B.*** A simulation that depicts an activator and a repressor modeled as additive rules. ***C.*** A simulation that depicts a synergistic rule between two regulatory elements. Each element has a small additive contribution to expression, but promoters with at least one of each element have enhanced expression. Gaussian noise was added to the output of the simulation at 5% of the level of expression of individual promoters. ***D.*** A simulation that depicts a cooperative interaction between two regulatory elements modeled with a hill function. Noise was added to the simulation as in *C*.

Figure 2.3: Comparison of Relos vs. Biologically Generated Data



Tree regression was performed on five modules (1,11,41,45,49) from Beer and Tavazoie[7]. The tree was converted to a ruleset and the ruleset used to generate expression values for each promoter in the modules. The median gene-wise correlation is 0.78. The real microarray expression values are depicted on the left and the Relos-generated expression values are on the right.

Additional Files

(Available at http://www.biomedcentral.com/1471-2105/8/272/additional)

Additional File 1 – Supplementary File 1

File name: testsets.zip
File format: compressed archive (ZIP)/ASCII text/PNG
Title: Test Datasets
Description: A compressed archive (zip) containing: the rulesets used to generate the test-set datasets (ASCII/xml); the datasets in both Relos and fasta format (ASCII); and histograms of each test-set to provide an overview of the data (PNG).

Additional File 2 – Supplementary File 2

File name: figure2_rulesets.zip
File format: compressed archive (zip)/ASCII text
Title: Figure 2 Rule-sets
Description: A compressed archive (zip) file containing the simulation files (ASCII/xml). used in the generation of figure 2.

Additional File 3 – Supplementary File 3

File name: rulespecification_and_promoteranalysis.txt
File format: ASCII text
Title: Rule Specification and Promoter Analysis
Description: Detailed information on how to specify rules and how promoters are analyzed.

Additional File 4 – Supplementary Table 1: Relos Dependencies

File name: relos_dependencies.pdf
Fil format: PDF
Title: Supplementary Table 1: Relos Dependencies
Description: A listing of modules needed for Relos to run and where they can be obtained.

Additional File 5 – Relos Source

File name: relos-src.zip
File format: compressed archive (ZIP)/ASCII text
Title: Relos Source Code
Description: A compressed archive (zip) containing the perl source for running Relos, the xml document-type definitions (DTD) which define simulation files, example simulation files, the README, and the source license (GPL).

Additional File 6 – Generated Promoter Modules

File name: microarray_generatedpromoters.png

File format: Portable Network Graphics (PNG)
Title: Image of modules from generated promoters
Description: A “heat map” image showing the expression from the generated promoters. Promoters with only “Spacer” elements are not depicted.

Additional File 7 – Sample Ruleset for Biological Expression Comparison

File name: rules1.xml
File format: XML/ASCII text
Title: Sample Ruleset for Biological Comparison
Description: Ruleset used in one of the 255 “microarray” simulations.

Additional File 8 – Ruleset For Promoter Generation

File name: genprom.xml
File format: Title: Ruleset Used for Promoter Generation
Description: The ruleset used to generate the 1000 promoters used in testing the biological relevance of Relos-generated promoters.

References

1. Uemura, H., et al., *The role of Gcr1p in the transcriptional activation of glycolytic genes in yeast Saccharomyces cerevisiae.* Genetics, 1997. **147**(2): p. 521-32.
2. Pierce, M., et al., *Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression.* Mol Cell Biol, 2003. **23**(14): p. 4814-25.
3. Yuh, C.H., H. Bolouri, and E.H. Davidson, *Cis-regulatory logic in the endo16 gene: switching from a specification to a differentiation mode of control.* Development, 2001. **128**(5): p. 617-29.
4. Makeev, V.J., et al., *Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information.* Nucleic Acids Res, 2003. **31**(20): p. 6016-26.
5. Anholt, R.R., et al., *The genetic architecture of odor-guided behavior in Drosophila: epistasis and the transcriptome.* Nat Genet, 2003. **35**(2): p. 180-4.
6. Roven, C. and H.J. Bussemaker, *REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data.* Nucleic Acids Res, 2003. **31**(13): p. 3487-90.
7. Beer, M.A. and S. Tavazoie, *Predicting gene expression from sequence.* Cell, 2004. **117** (2): p. 185-98.
8. Istrail, S. and E.H. Davidson, *Logic functions of the genomic cis-regulatory code.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4954-9.
9. Ligr, M., et al., *Gene expression from random libraries of yeast promoters.* Genetics, 2006. **172**(4): p. 2113-22.
10. Michaud, D.J., A.G. Marsh, and P.S. Dhurjati, *eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods.* Bioinformatics, 2003. **19**(9): p. 1140-6.
11. Mendes, P., W. Sha, and K. Ye, *Artificial gene networks for objective comparison of analysis algorithms.* Bioinformatics, 2003. **19 Suppl 2**: p. II122-II129.
12. Van den Bulcke, T., et al., *SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms.* BMC Bioinformatics, 2006. **7**: p. 43.
13. Laubenbacher, R. and B. Stigler, *A computational algebra approach to the reverse engineering of gene regulatory networks.* J Theor Biol, 2004. **229**(4): p. 523-37.
14. Hill, A.V., *The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves.* J. Physiol., 1910. **40**: p. iv - vii.
15. Granek, J.A. and N.D. Clarke, *Explicit equilibrium modeling of transcription-factor binding and gene regulation.* Genome Biol, 2005. **6**(10): p. R87.
16. Breiman, L., et al., *Classification and Regression Trees*. 1998, Boca Raton, Florida: CRC Press LLC.

Chapter 3: Improving thermodynamic models of transcriptional regulation through the combination of ChIP and expression data

Robert Zeigler[1] and Barak Cohen[1,2]

[1]Computational and Systems Biology Program, Washington University, St. Louis, MO

This work was performed in collaboration with Barak Cohen. My contribution was the experimental design and execution and the data analysis. Barak provided funding and valuable discussions.

Abstract

Transcription factor-mediated differential gene expression is important for many biological processes. Although many studies have identified binding preferences for transcription factors, few have studied how combinations of sites give rise to specific gene expression levels. Synthetic promoters have emerged as a means to simplify the combinatorial problem presented by the genome and parameterize models of expression based on promoter binding site composition. We sought to improve the biological accuracy of statistical thermodynamics models of transcription by developing a ChIP-based assay to quantitatively measure the occupancy of transcription factors on synthetic promoters *in vivo* and incorporating these data into the model. We applied our ChIP assay to Gcn4 and Cbf1 using libraries with binding sites for Gcn4, Cbf1, Met31/Met32, and Cbf1 in SC-Ura + 2% glucose and amino acid starvation (AAS) conditions. We found that the expression-only approach, despite being predictive of expression, misses several biological phenomenon, including a negative interaction between Gcn4 and Nrg1, and Gcn4 self-cooperativity. We also found that the ChIP data allow us to distinguish between competing mechanisms of regulatory change for the factor Cbf1.

Introduction

Differential gene expression lies at the heart of many biological processes including development (Istrail, De-Leon and Davidson, 2007; Prud'homme, Gompel and Carroll, 2007), differentiation (Gardner and Barald, 1991), and environmental response (Matikainen, *et al.*, 2001; Radinsky, 1995; Owuor and Kong, 2002). Often, changes in gene expression occur by one or more transcription factors (TFs) binding to specific DNA sequences (TFBS) and recruiting or inhibiting recruitment of RNA polymerase II (Johnson *et al.*, 2007; Manke, Roider and Vingron, 2008; Matys, V. *et al*., 2003; Morozov, 2005). The ability to quantitatively and accurately model changes in gene expression as a function of changes in TFBS composition is desirable to increase our understanding of and ability to engineer biology.

To date, many studies have been conducted attempting to learn the binding site specificities of TFs through a variety of methods, including analysis of promoters of suspected targets (Bussemaker, Li, and Siggia, 2001; Hughes, *et al.,* 2000*;* Hertz, Hartzell, and Stormo, 1990; Wang and Stormo, 2003), analysis of sequences bound by the TF in ChIP-ChIP, ChIP-seq, or differential expression studies experiments (Foat, *et al.* 2006; Harbison, *et al.* 2004; Lee, Johnstone, and Young, 2006; MacIsaac, *et al.,* 2006; Ren, *et al*., 2000; Roider, 2007; Valouev, *et al.*, 2008), and through *in vitro* binding studies (Liu and Stormo, 2005; Berger, 2006 and Mukherjee and Berger, 2004). These studies contribute the important first step of predicting which sequences *in vivo* are likely to be bound by a particular transcription factor. However, they fail to predict the transcriptional effect of binding and typically do not consider combinatorial binding effects such as multiple factors competing for the same site.

Recently, attempts have been made to correlate whole-genome expression profiles and ChIP occupancy data to the DNA content of regulatory sequences using models based on statistical thermodynamic models (Buchler, Gerland and Hwa, 2003; Granek and Clarke, 2005; Raveh-Sadka, Levo and Segal, 2009; Segal, E., *et al*., 2008; Shea and Ackers, 1985; and Wasson and Hartemink, 2009). Although these efforts are hampered by data insufficiency, the results have been promising and show that statistical thermodynamic models of transcription are a reasonable solution for producing predictive models that also help explain the underlying mechanism. The main difficulty is parameterizing the models given genomic data. The number of possible molecular events is immense compared to the number of gene expression observations available. An alternative approach to parameterizing the models is the use of synthetic promoters (Cox, Surette, and Elowitz, 2007; Gertz and Cohen, 2009; Gertz, Siggia and Cohen, 2009; Kwasnieski and Mogno, *et al*., 2012; Ligr, *et al.*, 2006; Melnikov, *et al.,* 2012; Murphy, Balazsi, and Collins, 2007; Patwardhan, 2012; and Sharon *et. al*., 2012.) In this approach, multiple promoter variants in the same promoter backbone are used to drive the expression of a single reporter gene, such as YFP. The reporter gene's expression level is assayed and used to estimate model parameters. This approach has the advantage of simplifying the system to focus on the effect on expression of various combinations of binding sites independent of other regulatory parameters.

Previous synthetic promoter approaches used only the sequence and expression data to infer relationships between the sequence content and the gene expression. The model from Gertz and Cohen (2009) performs well on the given data, explaining approximately 60% of the gene expression variable. However, the degree to which the model accurately describes the

underlying biophysical mechanisms responsible for the observed remains an open question. This is an important question since the degree to which the model can be confidently applied to contexts outside of the synthetic promoter system is directly related to the degree to which the model reflects the actual biophysical mechanism of the system.

I sought to extend the synthetic promoter approach by developing a ChIP-based metric of transcription factor occupancy on synthetic promoters. I applied this approach to a library of binding sites for transcription factors responsive to amino acid starvation (Blaiseau, *et al.,* 1997; Blaiseau and Thomas, 1998; Arndt and Fink, 1986) or glucose (Park, *et al.*, 1999) and tested the behavior of the sites in both glucose and amino acid starvation conditions. I used both the occupancy and expression data to attempt to separate the TF-DNA binding energy from the protein-protein interactions and to determine how different combinations of binding sites alter naive TF-DNA sequence binding preferences. This knowledge will help us better understand and model the relationship between TF occupancy and gene expression.

Methods

Construction of strains

Strain BC905 (Mat alpha, his3Δ1 leu2Δ0 lys2Δ::BirA ura3Δ0) was created by integrating BirA into the genome of strain BY4742 (Mat alpha, his3Δ1 leu2Δ0 lysΔ ura3Δ0) at the *lys2* locus via PCORE (Storici and Resnick, 2006). Briefly, a cassette containing KAN and *URA3* (PCORE) was inserted into the *lys2* locus using primers RZ131 and RZ132 (Table 3.6) and standard transformation protocols (Gietz and Woods, 2002) with selection on G418. BirA was inserted into this strain by transforming with BirA plus homology to the *lys2* region generated by primers RZ133 and RZ134 (Table 3.6) amplifying from plasmid prs313-BirA-NLS (van Werven and Timmers, 2006) with counter-selection on 5-FOA. Insertion was verified by PCR around the upstream and downstream regions of integration (primers RZ147-RZ149, Table 3.6) and by sequencing.

CBF1, *GCN4*, *MET31*, and *NRG1* were C-terminally tagged with the myc-C-avi tag by amplifying myc-C-avi with KAN from plasmid PUG6-myc-C-avi (van Werven and Timmers, 2006) using primer pairs referred to in Table 3.6 RZ129 and RZ130 (*CBF1*), RZ137 and RZ138 (*GCN4*), RZ135 and RZ136 (*MET31*), and RZ127 and RZ128 (*NRG1*) and transforming the resulting PCR product into BC905 using G418 selection to create strains BC906 (BC905 + *CBF1*::*myc-C-Avitag*, BC907 (BC905 + *GCN4*::*myc-C-Avitag)*, BC908 (BC905 + *MET31*::*myc-C-Avitag*, and BC909(BC905 + *NRG1*::*myc-C-Avitag*. Insertion was verified by PCR (Table 3.6, primers RZ92-RZ99, RZ143, RZ144) and by Sanger sequencing. The resulting strains were backcrossed to BY4741, sporulated, and offspring selected which matched the appropriate

genotype (MAT alpha his3Δ1 leu2Δ0 lys2Δ::BirA ura3Δ0 *CBF1*::myc-C-avi KAN). Retention of the tag and BirA was verified by PCR post-mating.

Media

All strain growth was done in YPD; synthetic complete with 2% glucose (SC); synthetic complete lacking uracil with 2% glucose (SC-Ura); synthetic complete lacking Trp with 2% glucose (SC-Trp); minimal media + 2% glucose with 300 uM his, 1 mM lys, 2 mM leu, 400 uM Trp (Min); minimal media + 2% glucose with 300 uM his, 1 mM lys, 2 mM leo, 200 uM Ura (Min+Ura-Trp); or in the same media with 0.9 uM biotin (YPDB, SCB, SCB-Ura (glucose), MinB, MinB+Ura-Trp).

Synthetic Promoter Library Creation

Libraries of synthetic promoters were created as described previously (Gertz, Siggia, and Cohen, 2009; Gertz and Cohen, 2009). Briefly, oligos with recognition sites for Cbf1 (Table 3.6 RZ84 and RZ85), Gcn4 (Table 3.6 RZ86 and RZ87), Met31 (Table 3.6 RZ88 and RZ89), and Nrg1 (Table 3.6 RZ90 and RZ91) were annealed, then mixed in ratios equal to the Tms of the annealed products, and ligated together. The ligation products were size selected with YM100 Microcon columns and cloned into plasmid pJG102 (Gertz, Siggia, and Cohen, 2009) and maxiprepped. The resulting plasmid was digested to produce a linear product with flanking homology to TRP1. The linear product was integrated into the avi-tagged strains following standard large-scale transformation protocols (Gietz and Schiestl, 2007). Ten 96-well plates of colonies were picked for each tagged strain, which were subjected to three rounds of dilution

purification consisting of growing the strains overnight in SC-URA, then pinning them onto SC-URA agar plates and allowing them to grow for two days. The final strains were replica-plated onto SC-Trp and the strains which grew were noted and excluded from the expression analysis.

Library Sequencing

Synthetic promoters were sequenced on the Illumina MiSeq platform using a double barcoding strategy. One of 96-well-specific barcoded primers was used with one of forty plate-specific barcoded primers to colony PCR the synthetic promoters such that each promoter was amplified with a unique combination of well and plate primers. The well primers included a SalI restriction site and the plate primers included an Mfe1 restriction site at the 3' end. Five uL of each PCR reaction was pooled together and ethanol precipitated, resuspended in 10 mLs of water, phenol/chloroform extracted, then ethanol precipitated and resuspended in 1 mL of H2O. About one third of this material was run on a 1.5% TAE agarose gel and a band from ~150 bp to 800 bp was purified (Qiagen #28704) to remove primer-dimers and the remaining material frozen down as stock. Approximately 500 ng of gel-purified material was combined in each of four tubes with 5 uL of 1 uM pre-annealed custom sequencing adapters RZ231 and RZ233 (Table 3.6), 1 uL each of EcoR1-HF (NEB R3101) and Mfe1 (NEB R0589S), 4 uL of 10 mM ATP, 4 uL of NEB Buffer 4 (NEB B7004S) and water to 39 uL. This mix was digested for 20' at 37C, at which point 1 uL of T4 DNA Ligase (NEB M0202S) was added to the mix, which was cycled three times between 16C and 25C for 10 minutes at each temperature, followed by 10 minutes at 16C, 10 minutes at 65C, then 20 minutes at 37C accompanied by 1 uL fresh EcoR1-HF and Mfe1. The resulting material was PCR-cleaned (Qiagen #28014) using a single column

for all four tubes. Ten uL of the eluted material was combined in each of three tubes with 1 uL each of Xho1 (NEB R0146S) and Sal1-HF (NEB R3138), 2 uL of NEB Buffer 4, 2 uL of 10 mM ATP, 2 uL of 10X BSA (B9001S), and 2 uL of 1uM pre-annealed custom sequencing adapters RZ230 and RZ232 (Table 3.6). The mix was digested/ligated as for the other adapters. The final solution was run on a 1.5% TAE Agarose gel and size selected for 150-700 bp then run with 25% PhiX on the MiSeq platform using the 2x150 bp chemistry, but run for 250 cycles forward and 50 cycles reverse. The resulting sequence data was analyzed by custom python scripts that used a minimum hamming distance approach to determine the TFBS composition of the promoters allowing for up to one mismatch per binding site.

Growth Conditions

For expression measurements, strains were grown in glucose and amino acid starvation (AAS) conditions as described previously (Gertz and Cohen, 2009) with the addition of 0.9 uM Biotin to all media. For the ChIP measurements, strains were grown as for expression in 96-well format overnight. For the glucose condition, 30 uL of overnight culture from each well for a given tagged factor was pooled together, and 20 mLs of this pooled culture was added to 980 mLs of SCB-Ura (see media) and grown for approximately four and a half hours to a final optical density (OD660) of 0.6-1.0. For the AAS condition, growth was carried out as for expression measurements except that after growth to mid-log phase in glucose, 30 uL of each strain for a given tagged factor was pooled together and 20 mLs of the pooled culture was spun down briefly (two minutes at 1000G) and the supernatant decanted. The pellet was resuspended

in 10 mLs of MinB (see media) and added to 990 mLs of MinB media. Final OD660 after six hours of growth was between 0.8 and 1.2.

YFP Expression Measurements

Strains were grown as described then fixed by adding 4% paraformaldehyde solution (4% formaldehyde, 100 mM sucrose) to a final concentration of 1%. YFP intensities were measured by flow cytometry on a Beckman Coulter Cell Lab Quanta SC. The final expression measurement was the ratio of raw fluorescence to volume of the cell, as reported as the "electronic volume" by the instrument, normalized to the mean expression of three to four no-insert control promoters on the same plate. Samples with fewer than 80% of counts with a fluorescence intensity between 10 and 900 raw fluorescence units were discarded from further analysis.

Biotin-ChIP

Pooled strains grown in 1 L of glucose or AAS media for ChIP were crosslinked with a final concentration of 1% formaldehyde for 15 minutes at room temperature. Crosslinking was quenched by adding 150 mLs of 2.5 M glycine and mixing at room temperature for five minutes followed by centrifugation and three 50 mL chilled TBS washes. The final cell pellet was transferred to three microcentrifuge tubes using the supernatant remaining after decanting and spun in a microcentrifuge for three minutes at 3000rcf and 4C. Each tube represents a single ChIP technical replicate. The supernatant was removed and the pellets frozen at least overnight at -80C. Frozen pellets were thawed on ice and resuspended in 2 mLs of Lysis Buffer (50 mM

HEPES, 150 mM NaCl, 1 mM EDTA, 1% v/v Triton X-100, 0.1% w/v sodium deoxycholate, 0.1% w/v SDS) and protease inhibitor (Roche #11836170001). Each replicate was distributed into two 2 mL tubes pre-filled three-quarters full with zirconium silicate beads. The tubes were bead beat six times on the highest setting for three minutes each time with a one minute pause in an ice bath between each beating using a Biospec Products Mini Bead Beater. The lysed cell matter was extracted by centrifugation, and the pellets resuspended in a final volume of 5 mLs of Lysis Buffer in a 15 mL centrifuge tube. The resuspended pellet was sonicated two times for 30 seconds each time with a Branson Sonifier 250 tip sonicator at power level six, duty 75% followed by four times for 30 seconds at power level five, duty 75%, with at least two minutes on ice between each sonication. The conical tubes were spun for two minutes at 3200G at 4C in a benchtop centrifuge. The supernatant was transferred to microcentrifuge tubes and centrifuged for 30 minutes at 4C at >16,000rcf. During the spin, 500 uL (per technical replicate) of Dynal M280 streptavidin-coated magnetic beads (Life Technologies, 112-05D) were washed three times with PBS and distributed into two 2 mL tubes per replicate. Four mLs of the supernatant was added to washed beads, 2 mLs per each 2 mL tube and incubated at room temperature for one hour. After incubation, the beads were bound to magnets and the supernatant removed and set aside for use as “input” (IN) material. The two tubes of beads per replicate (IP) were combined and washed were washed with 1.8 mL of solution for two by five minutes in each of Lysis Buffer, High Salt Lysis Buffer (50 mM HEPES, 0.5M NaCl, 1 mM EDTA, 1% v/v Triton X-100, 0.1% w/v sodium deoxycholate), LiCl Wash Buffer (500 mM LiCl, 1% NP-40 alternative, 10 mM Tris pH 8.0, 1 mM EDTA), SDS Wash Buffer (10 mM Tris pH 8.0, 1 mM EDTA, 3% SDS), and TE (10 mM Tris pH 8.0, 1 mM EDTA). The beads were resuspended in

250 uL TE + 0.5% SDS + 10 uL of 20 mg/mL Proteinase-K (NEB P8102S) and distributed into three 250 uL PCR tubes per replicate. Then 72.5 uL of IN material was combined with 72.5 uL of TE + 1% SDS to which 10 uL of 20 mg/mL Proteinase-K was added and distributed into three 250 uL PCR tubes per replicate. The tubes were incubated for four hours at 42C followed by two hours at 72C followed by six hours at 65C. The material from each replicate was recombined and purified via ChIP cleanup columns (Zymo D5205), eluting in 35 uL of elution buffer.

qPCR of ChIP Samples

Factor-specific qPCR primers were chosen by selecting the most highly enriched probes for the factor from Harbison, *et al.* (2004), scanning the probe sequence with Patser (Hertz and Stormo, 1999) for motif matches using motifs from Zhao and Stormo (2011) for Cbf1 and from Spivak and Stormo (2012) for the remaining factors, then using Primer3 to design qPCR primers that flanked the best motif matches. Three independent dilutions of two ChIP replicates were diluted to approximately 0.01-0.1 ng/uL final concentration for both the IN and IP. For each dilution of each sample, 3 uL of each dilution were added to 30 uL of water. Eleven uL from each 33 uL of water were added to each of two wells of a 96-well plate. To half the wells, 12.5 uL of SYBR Green QPCR Master Mix (Thermo Scientific AB-1158/A) and 0.75 uL each of 10 uM primers (Table 3.6 RZ169 and RZ170), amplifying a region in the SUC2 promoter were added. To the other half, 0.75 uL each of factor-specific target primers from Table 3.6 RZ158 and RZ159 (Cfb1), RZ177 and RZ178 (Gcn4), RZ183 and RZ184 (Met31), RZ193 and RZ194 (Nrg1) were added along with 12.5 uL of SYBR Green QPCR Master Mix. The resulting plate

was sealed and run on a Stratagene Mx3000p qPCR thermocycler. The replicates were averaged and analyzed using the delta-delta Ct method.

Sequencing of ChIP Synthetic Promoters

Sequencing of the ChIPed synthetic promoters was done by adding adapter sequences to synthetic promoters in the IN and IP samples via PCR amplification using 23 uL of IP material with 1 uL each of 10 uM primers which were barcoded in the forward read based on sample identity and in the reverse read based on the identity of the tagged transcription factor (see Table 3.5 for the list of barcoded primers used) and three different starting concentration of input material. The resulting products were gel-purified on a 1.5% TAE agarose gel, size selecting for approximately 150bp to 600bp. Input samples were retained on the basis of similar gel-intensities to the corresponding IP sample as an approximate concentration measure. The resulting samples were combined, ethanol precipitated and reconstituted in 30 uL of water. The forward sequencing adapter was added by digestion/ligation exactly as for library sequencing. The final concentration of sequence-able fragments was determined by qPCR using SYBR Green QPCR master mix, primers RZ259 and RZ260, and eight synthetic promoter standards, diluted across five orders of magnitude. The material was sequenced on the Illumina HiSeq 2000 platform using one lane of a paired-end 101bp run.

Occupancy of Synthetic Promoters

The relative occupancy of synthetic promoters was determined by mapping each sequencing read back to the promoter of origin. First, the read was parsed to determine which binding sites were present. This information was used to map the read back to the originating

promoter. The read counts were normalized by the total number of reads that mapped to a given sample type and avi-tagged transcription factor. The ratio of normalized IP counts to normalized input counts for a particular promoter was divided by the median normalized IP/input ratio of all promoters lacking a binding site for the ChIPed factor to give the normalized relative occupancy. Scaling to the median background occupancy effectively scales the occupancy values relative to the non-specific binding of the factor. This places all occupancy values from all factors and conditions on the same relative scale, assuming that the non-specific binding distribution is the same for all factors. For demonstrating technical replicate variance, the occupancy was calculated separately for each replicate. For modeling purposes, the replicates were generally combined by summing the promoter coverage across replicates and computing occupancy from the summed values. The exception was Gcn4 in AAS where a single ChIP replicate was used due to substantial depletion of promoters with four or more binding sites in the input of two of the replicates. Promoters with fewer than fifty reads in the inputs were excluded from the analysis.

Thermodynamic Model of Transcription

Modeling of expression and occupancy used the thermodyamic model of transcription described previously (Gertz, Siggia, and Cohen, 2009; Gertz and Cohen, 2009, Buchler, Gerland, and Hwa, 2003). The model considers unbound DNA as a reference state and computes the statistical weight of each possible configuration *k* of transcription factors and proteins bound to the DNA as:

$$W_k = e^{-\Delta G_k}$$

Where ΔG_k is given as:

$$\Delta G_k = \sum_{i=1}^{L} (\Delta G_{tf\,i DNA} + \Delta G_{tf\,i,RNAP} \cdot \delta(RNAP)) \cdot \delta(TF_i) + \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} \Delta G_{ixn\ tf\,i,j} \cdot \delta(TF_i) \cdot \delta(TF_j) \cdot \varepsilon(i,j)$$

where L is the number of TF binding sites in the synthetic promoter, $\Delta G_{tf\ i\ DNA}$ is the binding energy of the TF at site i, reflecting its concentration and affinity for the site, $\Delta G_{tf\ i\ RNAP}$ is the binding energy between the TF at site i and RNAP, $\delta(RNAP)$ is one if RNAP is bound in the current state and zero otherwise, $\Delta G_{ixn\ tf\ i,j}$ is the binding energy between the TF at site i and the TF at site j, $\delta(TF_i)$ is one if the TF at site i is bound in the current state and zero otherwise, and $\varepsilon(i,j)$ is one if there are no other TFs bound between sites i and j in the current state, and zero otherwise. The probability of polymerase bound is then given as:

$$P(RNAP_{bound}) = \frac{\sum_{k=1}^{N} W_k \delta_k(RNAP)}{\sum_{k=1}^{N} W_k}$$

Where N is the total number of states (typically 2^L for non-competitive binding), and $\delta_k(RNAP)$ is one if RNAP is bound in state k and zero otherwise. The probability of occupancy for a particular TF is computed as:

$$P(\geq 1\ TF_{bound}) = \frac{\sum_{k=1}^{N} W_k \delta_k(TF)}{\sum_{k=1}^{N} W_k}$$

Where N is the total number of states (2^L for non-competitive binding), and $\delta_k(TF)$ is one if the TF is bound to one or more of its sites in the state and zero otherwise. The observed occupancy and expression values were assumed to linearly related to the predicted probabilities, respectively:

$$Occupancy = \alpha \cdot P(\geq 1\ TF_{bound})$$

and

$$\mathrm{E}xp\mathrm{res}sion = \beta \cdot P(RNAP_{bound})$$

Where α and β are the least-squares estimates. The current model does not account for non-specific TF-DNA interactions and always predicts an occupancy of zero for promoters with no specific binding site for the factor of interest. Therefore, promoters lacking a specific binding site for the factor whose occupancy was being calculated were excluded from model estimation and validation. Model parameters were recovered in several ways. First, by performing simultaneous optimization with only expression data. Second by fitting with all useable occupancy data. Third, by fitting to occupancy and expression data. When fitting only to occupancy data, TF-DNA and TF-TF binding energies were explored with a simultaneous fit to all environments. When fitting to expression data, the optimization was carried out simultaneously for multiple environments and factors largely as previously described (Gertz and Cohen, 2009) with modifications as follows. First, expression values for multiple biological replicates within a particular strain background were collapsed into a single promoter using the median expression of all biological replicates. Second, no down-weighting of short-promoter residuals was used. Finally, the optimization was done in R using nlminb with default parameters. When fitting with both expression and occupancy data, the occupancy data was first re-scaled by the ratio of the means of the occupancy and expression data to put it on a similar quantitative scale to the expression data to ensure that neither the occupancy nor the expression would dominate the residual sum of squares for fitting. Optimization was performed as for fitting with expression

data, with the probability of both occupancy and expression scaled to the mean of all observed values.

Competitive Binding Model

The competitive binding model functioned exactly as the standard model except that each Gcn4 site had three possible states: unbound, bound by Gcn4, and bound by the unidentified competitor. No direct interaction between Gcn4 and the competitor was modeled. The competitor was assumed to have the same concentration and the same effect on polymerase in both conditions. The Gcn4 effect on polymerase was held constant in both conditions, but its concentration in both conditions was allowed to vary. All other parameters were fit as for the non-competitive model.

Cross Validation of Models

All models were subjected to 5-fold cross validation. The promoters and associated expression or occupancy values were randomly partitioned into five equally sized sets. In each round of cross validation, training was performed on four out of the five sets of data and validation was performed on the left-out set of data. Each partition was used once and only once for validation.

Results

Promoter Libraries with tagged TFs show similar expression behavior

Four libraries were constructed as detailed in Gertz and Cohen (2009) using previously published TF binding sites of Cbf1, Gcn4, Met31/Met32, and Nrg1. (See Table 3.6, oligos RZ84-RZ91). The libraries were built in strains where the bacterial biotin ligase BirA was integrated into the yeast genome at the *LYS2* locus and in each library one of the four transcription factors thought to bind the sites was tagged with the myc-C-Avi tag (van Werven and Timmers, 2006). The number of total and unique promoters for each library is reported in Table 3.1. The strains were grown as outlined in methods for either ChIP or expression analysis in both the glucose and AAS conditions.

Expression driven by the synthetic promoters was measured by flow cytometry as detailed in the methods. In general, the libraries showed similar expression distributions to each other in both conditions (Figures 3.1 and 3.2), indicating the tag does not alter protein function. The exception was the library in the Cbf1-tagged strain which showed a bimodal distribution of expression in glucose compared to the unimodal distribution observed for the other tagged-strains (Figure 3.1). The difference in expression can be accounted for entirely by the tag on Cbf1 as indicated in Figure 3.3 which shows the mean expression for matched promoters in the Cbf1-tagged vs. Gcn4-tagged backgrounds. Promoters with Cbf1 sites in the Cbf1-tagged strain are universally expressed higher than in the other strains, whereas promoters with no Cbf1 sites are expressed the same between the libraries. A list of all promoters and their expression data is available in Table 3.7.

Library ChIP shows enrichment for Cbf1 and Gcn4 but not Met31 or Nrg1

Strains were grown and ChIPed in triplicate as detailed in the methods. The resulting samples of enriched (IP) and unenriched (IN) material were used for qPCR-based validation of the ChIP. QPCR was carried out by diluting the IP and input for two of three ChIP replicates per strain to between 0.01-0.1ng/uL and setting up triplicate qPCR reactions for each sample, amplifying both an unbound control region and a bound target region. In general, Cbf1 is expected to be actively transcribed in both conditions, whereas Gcn4 is predominantly active in the AAS condition. The ChIP for Cbf1 and Gcn4 showed significant enrichment of the bound region relative to the unbound regions (Figure 3.4). As expected, Cbf1 binding is similar in both conditions, whereas Gcn4 binding is much higher in AAS. Nrg1 should be active in glucose, and Met31 should be active in AAS. However, ChIP of these factors showed little if any enrichment or even depletion (Figure 3.5), suggesting that these factors cannot be ChIPed in these conditions or that the target regions are not bound *in vivo*.

ChIP of synthetic promoters is highly reproducible

The IP and IN samples were also used for PCR-selected sequencing of the ChIPed synthetic promoters. The ratio of normalized reads in the IP to normalized reads in the IN was used as a measure of relative occupancy. The occupancy replicates were first examined for technical reproducibility. Figure 3.6 compares ChIP replicates for ChIP of Cbf1 in both glucose and AAS conditions and clearly shows that the ChIP results are very consistent across multiple technical replicates, with an average R^2 of 0.94.

ChIP of synthetic promoters shows quantitative differences driven by the TFBS

For Cbf1 and Gcn4, there is clear signal in the ChIP in both conditions (Figures 3.7 and 3.8). The signal is similar for Cbf1 in both glucose and AAS conditions and shows a nearly linear increase in the mean relative occupancy with respect to the number of Cbf1 sites that are present in the promoter. Variation around the mean is a combination of assay noise and real biological variation, where different combinations of binding sites serve to increase or decrease the occupancy of the factor. In the case of Cbf1, the relatively tight peaks around the mean occupancy for one, two, and three Cbf1 sites suggests that Cbf1 occupancy is dependent primarily on the number of Cbf1 sites.

Gcn4 occupancy also increases with the number of Gcn4 sites. However, in glucose, the total occupancy is much lower than in AAS and the increase in occupancy appears to be a non-linear response to the number of Gcn4 binding sites. In AAS, Gcn4 occupancy responds more strongly to the number of Gcn4 binding sites. However, there is much wider dispersion of occupancy for Gcn4 in AAS than for Cbf1 in either condition. This suggests that in the context of the binding sites used for the libraries, Gcn4 occupancy is affected by other transcription factor binding events to a greater extent than Cbf1.

No occupancy information for Met31 or Nrg1

Consistent with the qPCR results, Met31 and Nrg1 ChIP showed no signal in the synthetic promoters (Figures 3.9 and 3.10). Met31 and Met32 both recognize the same binding motif, but previous *in vivo* studies have shown that Met32 binds in preference to Met31 (Carrillo,

et al., 2012). Thus the lack of ChIP signal on the synthetic promoters is likely all or partly due to the preferential binding of Met32.

The Nrg1 site used contains the Nrg1 consensus motif GGACCCTT (Spivak and Stormo, 2012) and functions as a repressor even in the Nrg1-tagged strain (Figure 3.11 and Table 3.3). Thus, Nrg1 appears to be functional even when tagged, but there is no observable ChIP signal. This may be due to epitope masking. Nrg1 is known to recruit the ssn6-tup1 repressive complex (Park, *et al.*, 1999) and this complex may render the biotin tag inaccessible for pull down after crosslinking. Due to lack of ChIP signal, promoters in the avi-tagged Met31 and avi-tagged Nrg1 strains were excluded from all remaining analyses. Therefore, no attempt was made at separating the TF-DNA binding energy from the interaction with RNAP for these factors.

Thermodynamic modeling of expression shows good agreement between predicted and observed expression

All expression data from both conditions were used to fit a model of sequence and expression without regard to occupancy. The initial model fit the TF-RNAP interactions and the TF-DNA binding energy in glucose relative to AAS. Additional parameters were attempted to test their improvement to the fit, including allowing for orientation-specific effects for the TF-RNAP interactions and testing all pairwise TF-TF interactions in turn. Five-fold cross validation was performed on the final model (Table 3.4). The final model parameters and values are listed in Table 3.3. The overall fit had an R^2 of 0.53 (Figure 3.12). This is comparable to the previously obtained fit of 0.6 (Gertz and Cohen, 2009), despite using two fewer parameters to model the data, having two different tagged transcription factors, and having a greater diversity

of promoters (212 unique promoter in glucose versus 131 unique promoters published previously.) Modeling the interaction between Cbf1 and RNAP separately for the tagged and untagged versions of Cbf1 resulted in a significant improvement to the fit of the model (R^2 0.53 vs. 0.34, $P<10^{-16}$, *F*-test). This is consistent with the bulk analysis showing that of Cbf1-containing promoters in the Cbf1-tagged background express higher than the same promoter in the other strains. Notably, when fitting solely with expression data, no TF-TF interactions were found to significantly improve the fit.

Thermodynamic modeling of occupancy predicts occupancy and interactions not observed in expression modeling

The normalized relative occupancy data in glucose and AAS conditions from the Cbf1 and Gcn4-tagged strains were used to fit a model relating sequence to occupancy of Cbf1 and Gcn4 on the synthetic promoters. The current implementation of the model does not consider non-specific DNA interactions, so promoters with no specific binding sites for the ChIPed factor were excluded from the fit. The initial model consisted only of the DNA-binding energy of Cbf1 and Gcn4 in glucose and AAS (R^2 0.54). Each transcription factor interaction between Cbf1 and all other factors and Gcn4 and all other factors was added to the model, one at a time, to determine if any interactions had a significant impact on the model performance. All p-values were corrected to account for the number of tests being performed.

Surprisingly, two interactions were significant by themselves despite no interactions being found significant in the expression-only model. The $\Delta G_{Gcn4\text{-}Gcn4}$ and the $\Delta G_{Gcn4\text{-}Nrg1}$ interactions made significant improvements to the fit of the model (R^2 0.56, $P=1.11e^{-04}$ and R^2

0.56, P=5.20e^{-05} respectively, *F*-test with Bonferroni correction). Adding the $\Delta G_{Gcn4\text{-}Gcn4}$ interaction to a model that includes the $\Delta G_{Gcn4\text{-}Nrg1}$ interaction also led to a significant improvement in the performance and stability of the model (R^2 0.57, P=1.43e^{-05}, *F*-test with Bonferroni correction). The final model, which includes the DNA binding energies, the $\Delta G_{Gcn4\text{-}Nrg1}$ and the $\Delta G_{Gcn4\text{-}Gcn4}$, predicts virtually no change in the DNA binding energy of Cbf1 between the two conditions ($\Delta\Delta G$: -0.08), versus a large change in the DNA binding energy of Gcn4 when moving from glucose to AAS ($\Delta\Delta G$: -2.74). The final model resulted in a fit with explanatory power on par with the thermodynamic modeling of expression (R^2 0.57 for occupancy versus R^2 0.53 for expression), suggesting that the model can describe the variation in both data sets equally well (Figure 3.13).

Overall, thermodynamic model simultaneously fits occupancy and expression data well, but suggests specific model improvements

All expression and occupancy data were combined to fit both the TF-DNA binding energy and the polymerase interaction terms simultaneously. The parameters to include were chosen based on the expression-only and occupancy-only fits (Table 3.3). In general, the fit-model was able to converge on reasonable predictions of both expression and occupancy (Figure 3.14). In particular, the model overall did better on predicting both categories of data than the model fit separately to either source of data. Whereas the model fit only on occupancy was incapable of predicting expression and the model fit only on expression predicted occupancy with an R^2 of 0.36, the model fit on both data sets predicted expression with an R^2 of 0.425 and occupancy with an R^2 of 0.556 (Table 3.4). Thus, the model fit on both data sets has good

predictive power across a broader range of biological questions of interest than models fit with either data set alone. However, the question of why the model fit on both sets of data performs significantly worse when predicting expression than the model fit on only expression deserves exploration.

One possibility is that the scales of the data may not be equivalent, despite effort to avoid that issue. This could be addressed by weighted regression or allowing the occupancy and expression data to scale separately from each other. Alternatively, the occupancy data may be noisier than the expression data. This is somewhat difficult to resolve with the good correlations observed between ChIP technical replicates (Figure 3.6), but it is formally possible that the ChIP assay is giving consistently incorrect results for some promoters. Here, the appropriate solution is to down-weight the residuals of the occupancy data relative to the expression data. Finally, there may be a mismatch between the model description of direct interaction between a TF and RNAP and the actual mode of effect. The model currently assumes a direct recruitment/inhibition model for how TFs interact with polymerase. This seems reasonable in many cases, but may not always be appropriate. There is some support for this idea. Gcn4, known to directly recruit the Mediator complex which directly associates with RNAP, was well-modeled across a variety of model architectures as long as they included competitive binding (see below) whereas Cbf1, which can affect expression indirectly (Moreau, *et al.* 2003), tended to be more problematic, as though the tight binding implied by the occupancy did not correlate well with overall effect on expression caused by Cbf1. This would require a more sophisticated description of the interaction between Cbf1 and RNAP to resolve. This may be an interesting avenue to explore in future work.

Due to the discrepancy in the two data sets regarding the behavior of the Gcn4 site in glucose (activation versus repression), the Gcn4-RNAP term behaved poorly in models which used both expression and occupancy and did not incorporate competitive binding. The fitting procedure consistently drove the Gcn4-glucose interaction to a highly unfavorable value which resulted in numerical instabilities in the fit. This occurred every time a fit was attempted with this model (>10). This suggests that the two data sets provide conflicting information regarding the effect of Gcn4 in glucose which can only be resolved with a more sophisticated model.

Occupancy distinguishes between distinct hypotheses of mechanism of Cbf1 effect on expression

The expression data suggest a change in the effect on expression due to Cbf1 in the AAS versus glucose conditions. This change can be modeled by allowing the binding energy of Cbf1 to change between the conditions, or by allowing the Cbf1-RNAP interaction to change between the two conditions. When fitting with only expression data, these two models produce equally good fits (R^2 of 0.53 in both cases), despite representing distinct biophysical mechanisms. The occupancy model distinguishes between the two mechanisms, suggesting that there is virtually no change in the binding energy of Cbf1. This is consistent with previous results using GFP-fused Cbf1 that showed no concentration difference between the two conditions (Gertz and Cohen, 2009). At that time, the condition-specific effect of Cbf1 was assumed to be the result of differential binding due to protein-protein interactions. The occupancy data clearly shows, however, that any Cbf1 binding differences between the conditions are negligible. This argues for a shift in the activating potential of Cbf1 rather than a change in the binding energy.

Gcn4 Binds Cooperatively

Since the $\Delta G_{Gcn4\text{-}Gcn4}$ term made a significant improvement in the fit of the occupancy data, I sought to explore the cooperativity further by applying a Hill function to the Gcn4 AAS binding data. The function fit was:

$$Occupancy = a + nsites^n/(x^n + nsites^n)$$

which is the standard Hill function with an intercept to model non-specific binding. The function was fit twice, once with n constrained to 1 (no cooperativity) and once with n allowed to vary. Figure 3.15 shows the mean fractional occupancy (normalized occupancy / max occupancy) of Gcn4 by number of Gcn4 binding sites present in the promoter and the standard error of the mean together with the two fits (red: n=1, x=3.23, a=0.011; blue: n=2.56, x=2.27, a=0.051). All parameters of the cooperative hill function were deemed significant by t-test, with the hill coefficient deemed highly significant ($P<2e^{-16}$). This suggests that Gcn4 binds the promoters in a cooperative manner. This is consistent with recent results showing that Gcn4 binds to multiple sites in the mediator complex (Jedidi, *et al*. 2010; Brzovic, *et al*. 2011). Although there is no allostery in the binding of Gcn4 to mediator, the effect of Gcn4 making multiple contacts with mediator could result in cooperativity of the sort previously postulated for transcription factors generally (Tanaka, 1996) and observed for Mig1 (Gertz, Siggia, and Cohen, 2009).

Gcn4 occupancy is negatively impacted by Nrg1

Modeling of only binding data revealed a negative interaction between Gcn4 and Nrg1. Although the interaction was significant, it had a small effect on the overall fit (R^2 0.54 without, R^2 0.56 with, $P=5.20e^{-05}$). To investigate whether the parameter was meaningful or simply fitting

noise, the interaction was examined outside the thermodynamic model through a linear model. If the interaction is truly meaningful, it should should show up in both models. Figure 3.16 shows a clear effect of adding the Gcn4:Nrg1 interaction term to the model. The top graph is the model without the interaction (adjusted R^2 0.49) and the bottom is the model with the interaction (adjusted R^2 of 0.57, P<2e-16, *F*-test). This strongly suggests that the effect of the Nrg1-Gcn4 interaction in the model is real, and not just modeling noise. Nrg1 is known to recruit the ssn6-tup1 repressive complex (Park *et al.,* 1999), and it is plausible that this recruitment interferes with the ability of Gcn4 to bind and recruit mediator.

One question that arises is why the Gcn4 cooperativity and the Nrg1-Gcn4 interaction are observed only in the binding data and not in the expression data. There are two non-exclusive possibilities. One is that the expression-only model already incorporates a form of synergism in the interaction terms between the TFs and polymerase which may mask the cooperative binding by Gcn4 and competition between Nrg1 and Gcn4 in the fitting procedure. The second possibility is data insufficiency. In AAS conditions, promoters with many Gcn4 binding sites are highly active. However, there is a limit in the dynamic range of the flow cytometer and many of the highly active Gcn4-containing promoters exceed the upper limit of the dynamic range so their expression measurements have to be discarded. In AAS, comparing the total set of promoters versus those for which there are reliable expression measurements in AAS, there is a significant depletion of promoters with more than one Gcn4 binding site ($P<2e^{-16}$, hypergeometric test), so observing cooperativity in the expression data is difficult. Unlike the expression data, the dynamic range of the occupancy data is largely limited only by sequencing depth, allowing a larger variety of Gcn4-containing promoters to be sampled (145 Gcn4-

containing promoters in AAS occupancy vs. 54 Gcn4-containing promoters in AAS expression). This suggests that expression measurements with larger dynamic ranges might be better able to capture TF-TF interactions, but doesn't rule out the possibility that the Gcn4 cooperativity and Gcn4-Nrg1 interactions are masked in the model by the TF-polymerase interactions.

Gcn4 site shows switching behavior

Contrary to prior results (Gertz and Cohen, 2009), the Gcn4 binding site showed different behavior between glucose and AAS conditions. In AAS, it was a strong activating sequence (Figure 3.17A), consistent with the known role of Gcn4 in recruiting mediator and other transcriptional complexes (Jedidi, *et al*. 2010; Herbig, *et, al.* 2010) whereas in glucose, it functioned as a weak repressor (Figure 3.17B). The switching behavior occurred regardless of which factor was tagged (data not shown), indicating that the repressive effect is independent of the tagging. When modeling only expression, allowing the Gcn4-RNAP interaction to differ between conditions revealed the same trend: the site activates in AAS conditions but represses weakly in Glucose (Table 3.3). Forcing the model to use same polymerase interaction term for the Gcn4 site in both conditions resulted in a significantly worse fit (R^2 of 0.53 vs. 0.43, $P < 10^{-16}$, *F*-test). Attempting to fit the model by constraining the $\Delta G_{Gcn4\text{-}RNAP}$ term while allowing the binding energy of Gcn4 in glucose to vary as was done previously (Gertz and Cohen, 2009) resulted in a good fit (R^2 0.50) but the resulting change in binding energy equates to 8.12×10^{-14}-fold lower apparent Ka. This drastic of a change in binding energy is not biologically reasonable and is an artifact of fitting the data to an inappropriate model.

There are two possibilities for why the Gcn4 site switches behavior in different contexts. The first is that Gcn4 somehow changes from an activator to a repressor, presumably through a postranslational modification or interactions with other proteins. The second is that Gcn4 competes with another factor for the same binding site. The occupancy data can distinguish between these two hypothesis. If expression is negatively correlated with Gcn4 occupancy in glucose, it strongly suggests that Gcn4 is switching behavior. However, if expression is positively correlated with occupancy, it suggests that another repressive factor is competing with Gcn4. Figure 3.18 shows expression vs. occupancy in both glucose and AAS. As expected, there is a strong positive correlation between Gcn4 occupancy and expression in AAS. However, there is also a positive correlation between Gcn4 occupancy and expression in glucose. This argues that Gcn4 is not switching behavior in glucose, but that another factor is binding in competition to the Gcn4 site. In fact, competitive binding with Gcn4 competition with at least one other factor (Bas1) has been previously reported (Arndt and Fink, 1986; Springer, *et al.*, 1996). Thus, both the data and the literature support the idea of competitive binding occurring in the synthetic promoters.

Competitive model of binding better explains Gcn4 expression and occupancy

The thermodynamic model was extended to incorporate competitive binding with Gcn4 by the addition of promoter states where the competing protein is bound to the site instead of Gcn4. The model assumed that the effect on polymerase of the two competitors was consistent across conditions. The TF-DNA binding energy of one competitor was fixed in both conditions and the other (Gcn4) was allowed to vary between conditions. When this model was fit only

with expression data, it performed exactly the same as the model fit without competition where Gcn4 is assumed to switch behavior in the two conditions. With only expression data, competitive binding cannot be distinguished from Gcn4 having different effects on RNAP in the two conditions. Without additional data to constrain the fit, these two mechanisms cannot be separated.

The combined expression and occupancy data were also fit with the competitive model. Both models resulted in similar fits (see Table 3.4, Figure 3.14, and Figure 3.19), although the competitive model is marginally better at predicting expression. However, the non-competitive model consistently set the glucose Gcn4-RNAP term as highly unfavorable and resulted in numerically questionable fits, whereas the competitive model resulted in the same numerically stable fit 40% of the time. In this best fit, the difference in Gcn4-DNA binding energies between the two conditions equates to fold change in the apparent Ka of approximately 25-fold, This seems large given previously published data (Albrecht, *et al.* 1998) but is still within the realm of possibility. Thus, incorporating competition in the model resulted in a more stable, biologically accurate fit.

Discussion

I sought to improve our quantitative understanding of the biophysical mechanisms underlying transcriptional regulation. In particular, I incorporated ChIP data into existing statistical thermodynamic models of regulation and compared models parameterized with only expression, only ChIP data, and with both. Comparing the results of these modeling procedures revealed several interesting features.

First, Gcn4 occupancy seems to be more sensitive to the particular configuration of binding sites present in the promoter, whereas no context-dependent binding of Cbf1 was captured by the model. This suggests that some transcription factors are more sensitive to binding site context than others. It is interesting to note that Cbf1 is known to recruit chromatin remodeling complexes (Moreau, *et al*., 2003; Kent, *et al.,* 2004), whereas Gcn4 directly recruits the mediator complex (Herbig, *et al.,* 2010 ; Jedidi, *et al*., 2010). It may be necessary for proper Cbf1 function for it to be able to bind to DNA regardless of what other factors are binding nearby, including nucleosomes. On the other hand, it may be desirable for Gcn4 occupancy to be more easily influenced by the presence of other TF binding sites to guard against inappropriate activation. This line of reasoning suggests the hypothesis that transcription factors which directly recruit polymerase and related subunits will be more heavily influenced by binding site context than factors which are involved in earlier processes such as chromatin remodeling.

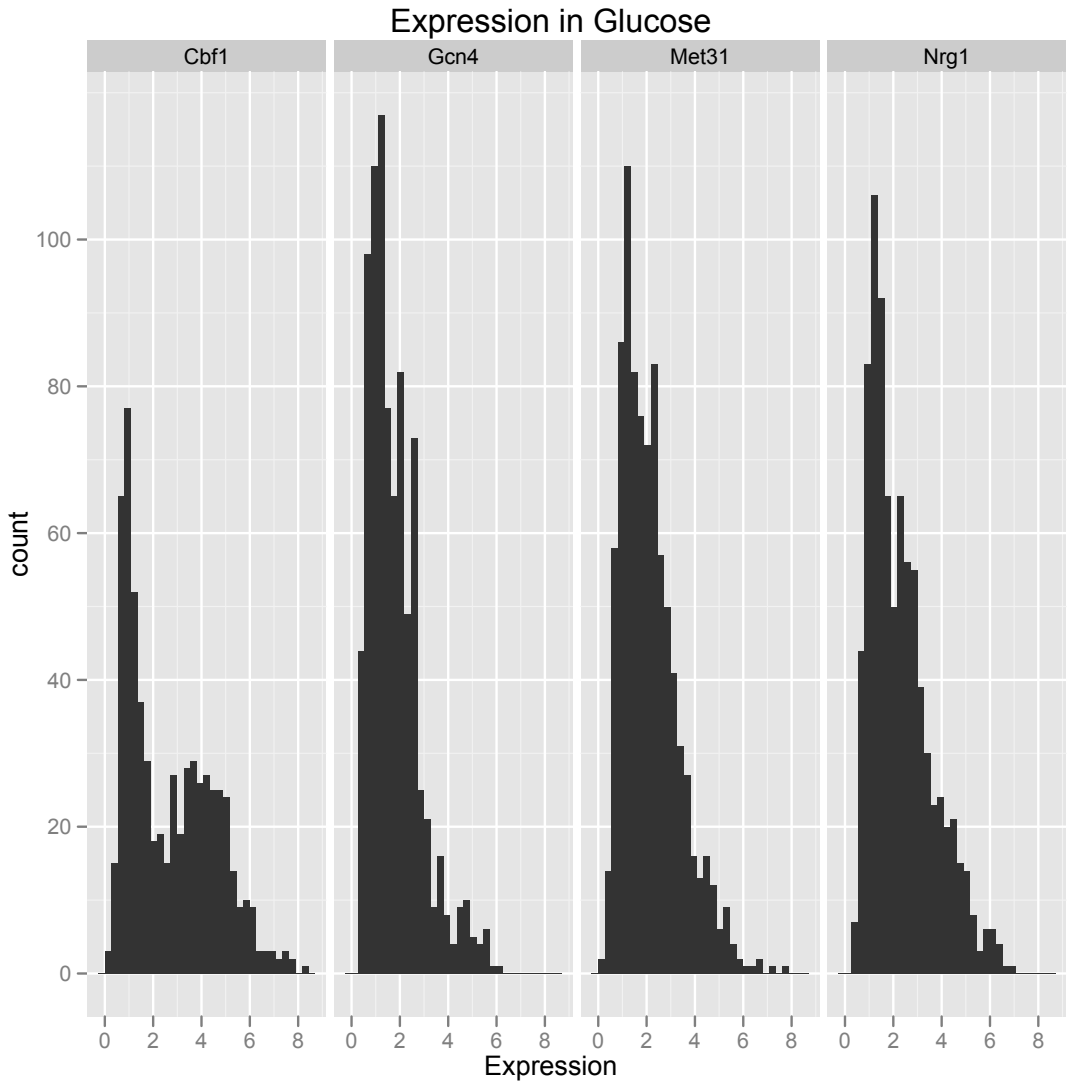
Second, the occupancy data reveals additional information that is masked by having just the expression data. In the expression data, the Gcn4 site activates in AAS and represses in Glucose. With only the expression data, this effect can be reasonably modeled as a switch in the Gcn4-RNAP interaction. However, the occupancy data suggest that Gcn4 still activates in

glucose but is being largely outcompeted by another factor. Gcn4 competition with other factors has been previously documented (Springer, *et al*., 1996; Arndt and Fink, 1986), and the incorporation of both occupancy data with expression data allows that process to be both discovered and better modeled.

Finally, combining occupancy and expression data for parameterizing thermodynamic models of transcription eliminates alternative hypothesis that cannot be be discarded with only expression data. In the case of Cbf1, two of three competing hypotheses for explaining differential regulation by Cbf1 in AAS versus glucose were eliminated. The binding of Cbf1 is the same in both conditions, strongly suggesting that the differential regulation occurs by altering the activation potential of Cbf1, possibly through Cbf1-dependent Met4 (Blaiseau and Thomas, 1998; Thomas, *et al.,* 1992). In all, we find that incorporating protein binding information in the form of ChIP data provides the ability to quantitatively reason about the biophysical mechanisms that underly observed expression data and to distinguish between different biological mechanisms that give rise to the same expression data.

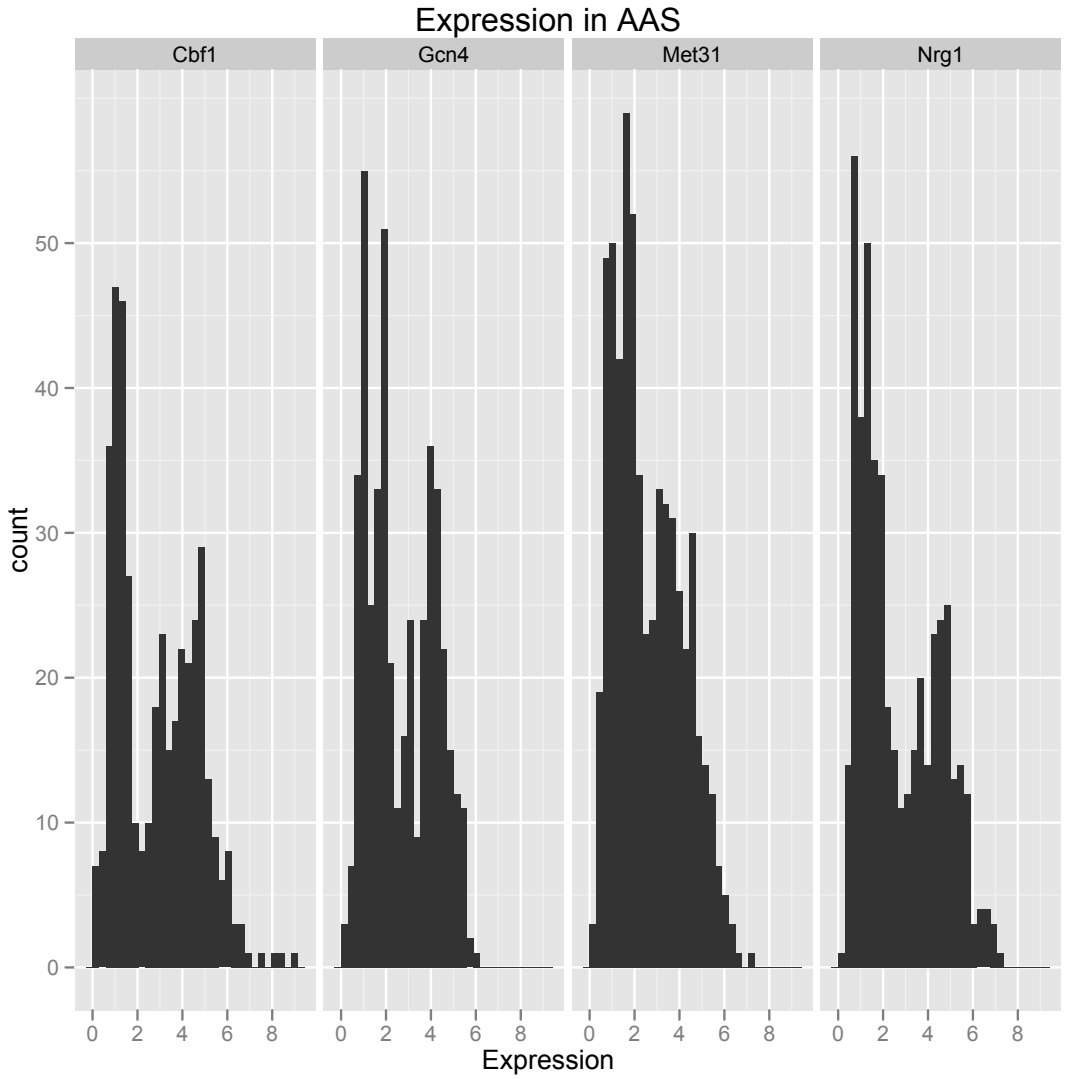
Figures

Figure 3.1: Expression distributions in glucose similar across all libraries except Cbf1



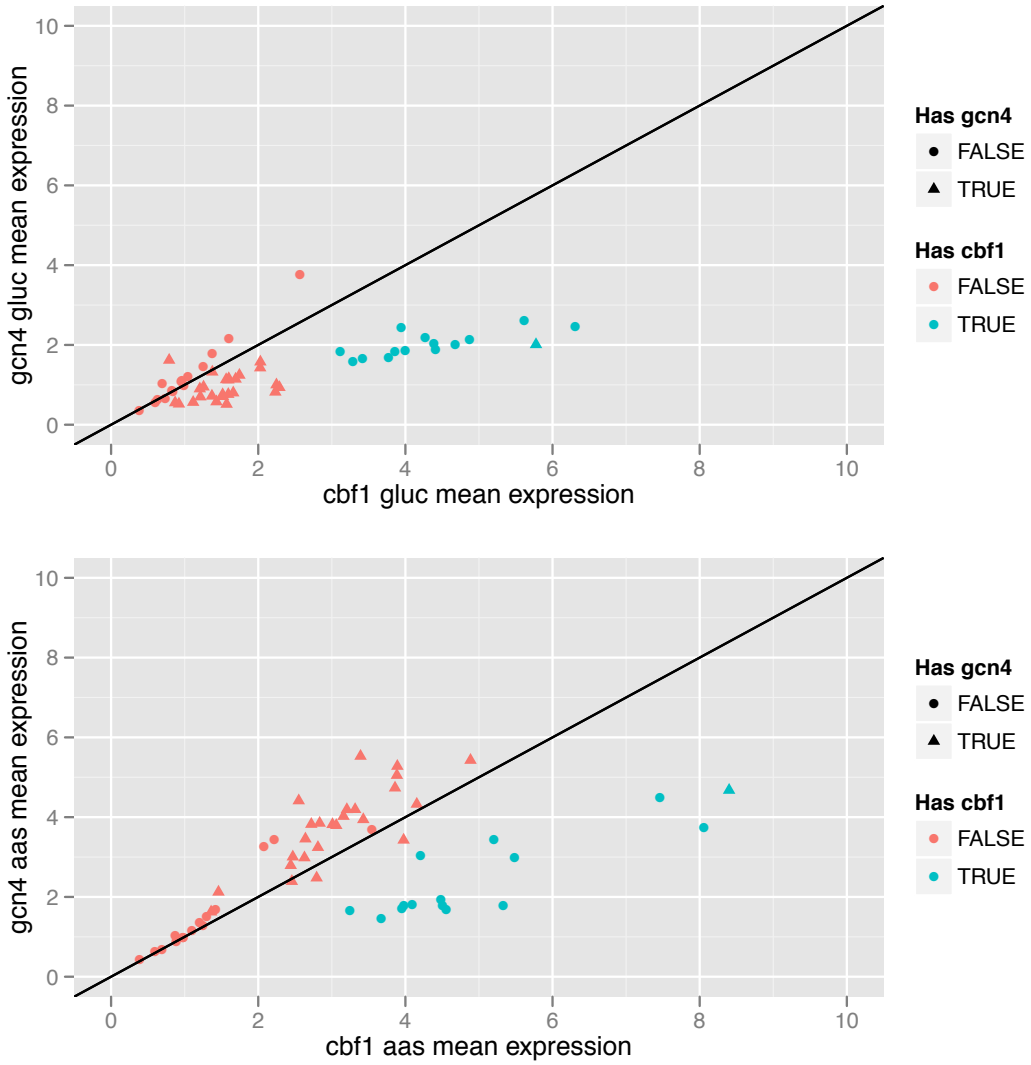
Libraries were grown to mid-log phase in SCB-Ura, fixed with a final concentration of 1% formaldehyde, and the fluorescence intensities measured by flow cytometry. The distribution of all libraries is similar except for Cbf1 (after multiple hypothesis correction, $P < 10^{-16}$ for Cbf1-Gcn4, Cbf1-Met31, Cbf1-Nrg1; P=0.18, Gcn4-Met31; P=0.06, Gcn4-Nrg1; P=0.45, Met31-Nrg1, Kolmogorov-Smirnov test).

Figure 3.2: Expression distributions in AAS similar across all libraries



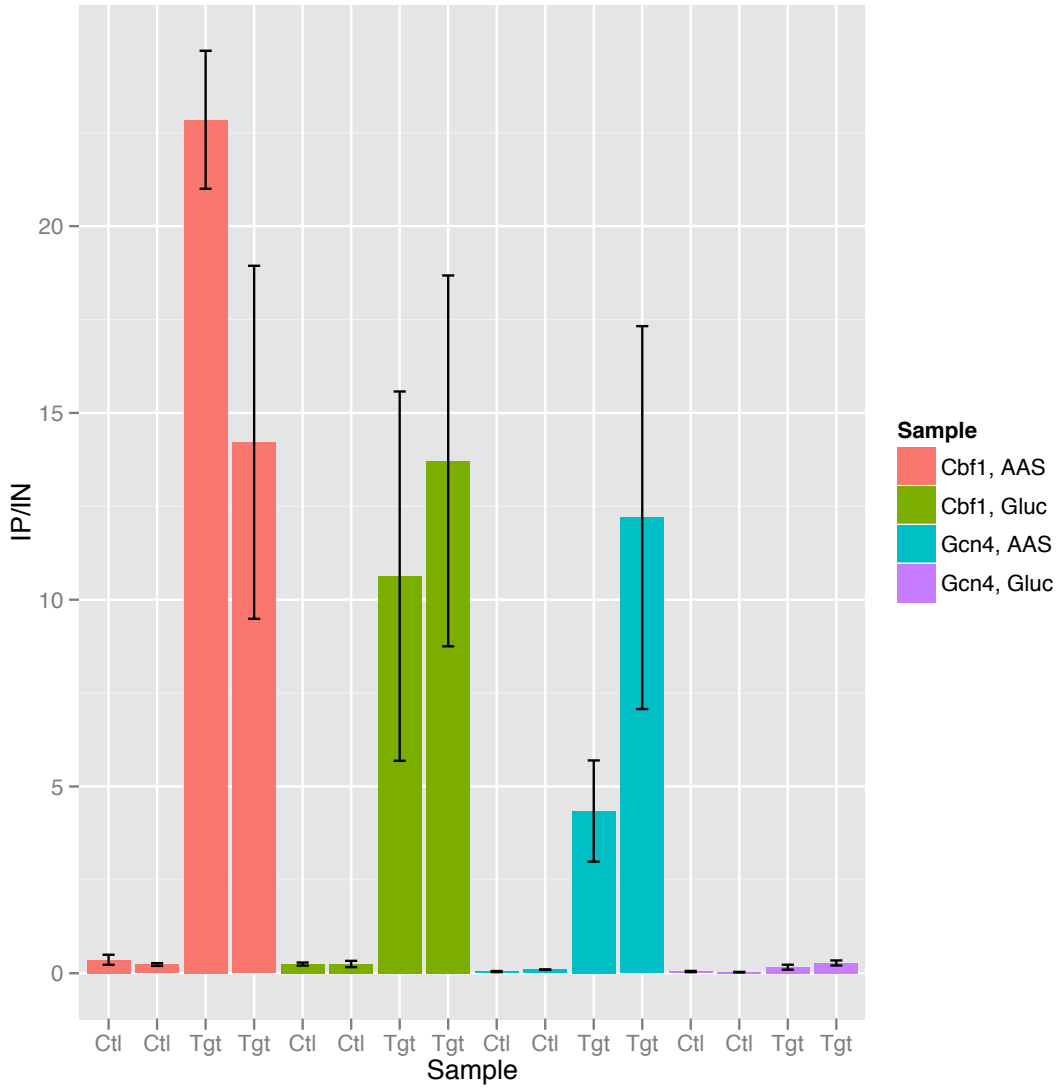
Libraries were grown to mid-log phase in SCB-Ura, fixed with a final concentration of 1% formaldehyde, and the fluorescence intensities measured by flow cytometry. The distribution of all libraries is similar but with some significant differences by Kolmogorov-Smirnov testing, suggesting some strain-specific effects in AAS, probably due to the protein tag (after multiple hypothesis correction, P = 0.0096, Cbf1-Gcn4; P=0.20, Cbf1-Met31; P=0.00017, Cbf1-Nrg1; P=0.41 Gcn4-Met31; P=$6.3e^{-10}$, Gcn4-Nrg1; P=$1.3e^{-5}$ Met31-Nrg1).

Figure 3.3: Avi-tagging increases Cbf1 activation potential, but not Gcn4



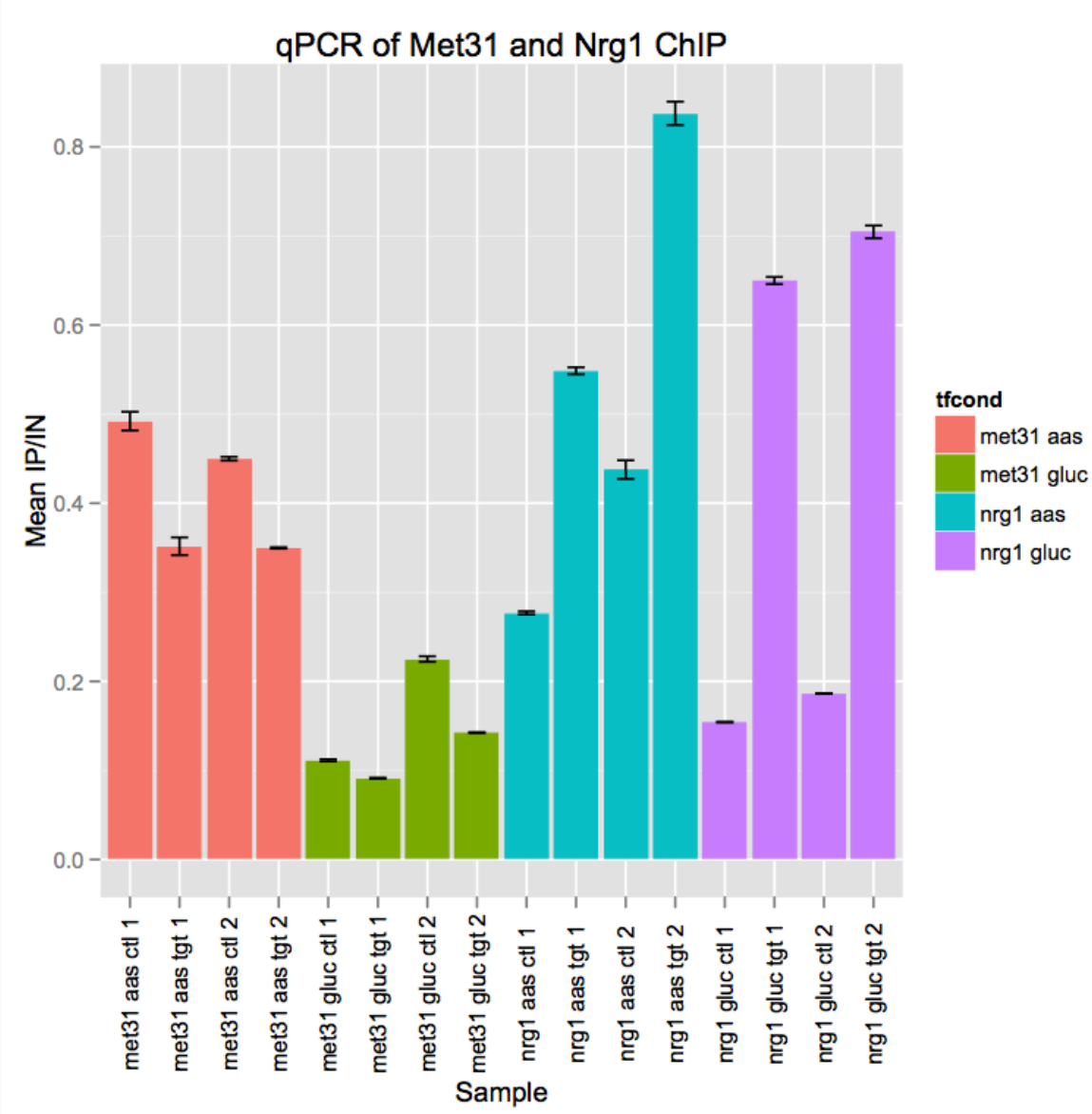
Mean expression of identical promoters in the Cbf1-tagged background and Gcn4-tagged background were compared. Promoters without cbf1 in them (red) generally show good agreement between the two libraries, falling along the black line. Promoters with Cbf1 sites in them (blue) consistently show higher expression in the Cbf1-tagged background than in the Gcn4-tagged background. Promoters with Gcn4 sites in them (triangles) do not appear to differ between the two libraries, indicating that the tag on Gcn4 has little if any effect on expression.

Figure 3.4: Specific enrichment of bound regions for Cbf1 and Gcn4 in glucose and AAS



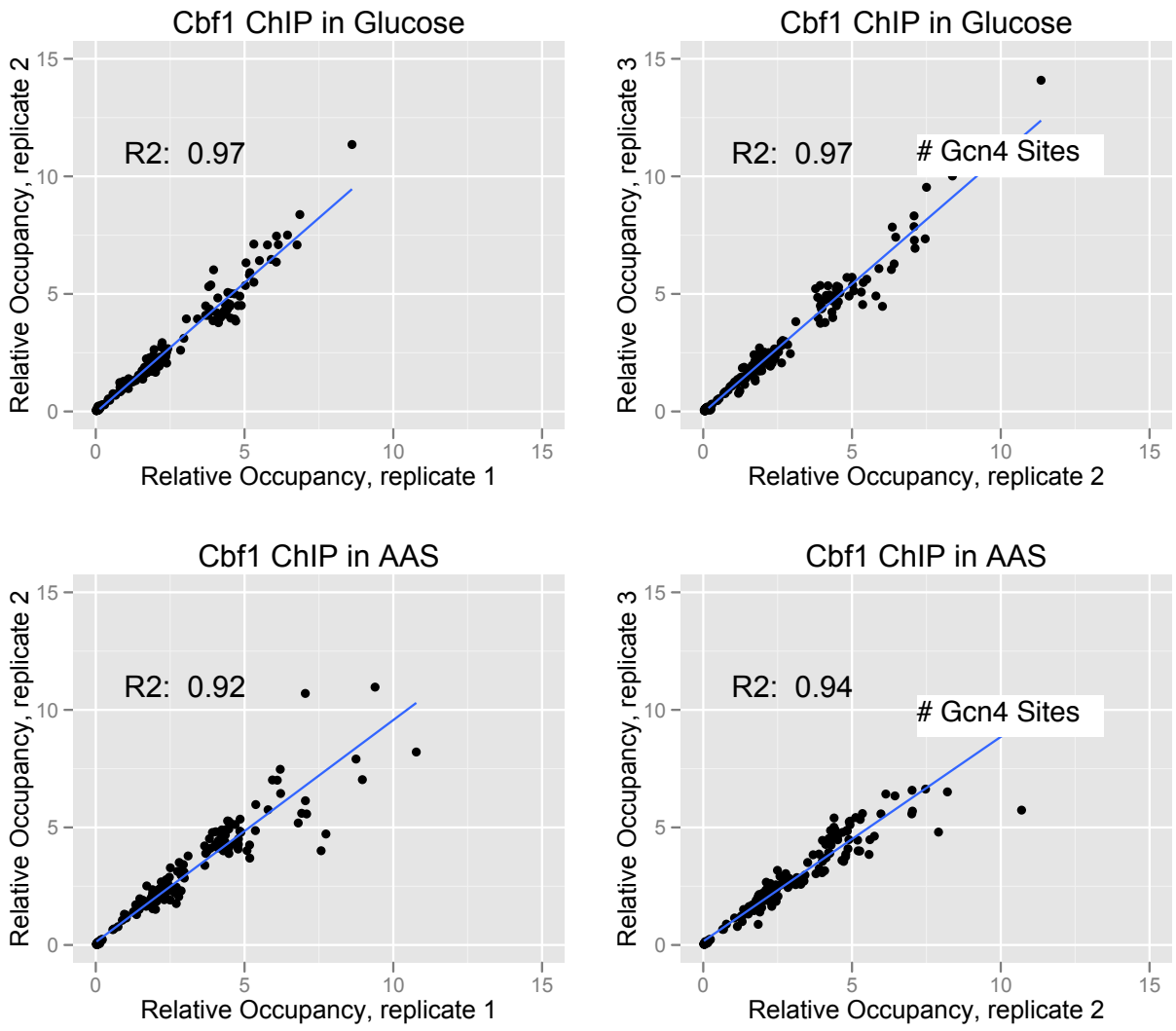
ChIP was performed on synthetic promoter-bearing strains with Avi-tagged Cbf1 and avi-tagged Gcn4 as outlined in methods. Enrichment was gauged by qPCR of a control (Ctl) region (SUC2) versus a target (Tgt) region (Ade 3 for Cbf1, CPA2 for Gcn4). Gluc is in Glucose, AAS is in Amino Acid Starvation. The target region is differentially bound with respect to the control region for both factors in all conditions, though only marginally so for Gcn4 in Glucose.

Figure 3.5: Limited or no specific enrichment of Met31 and Nrg1 ChIP



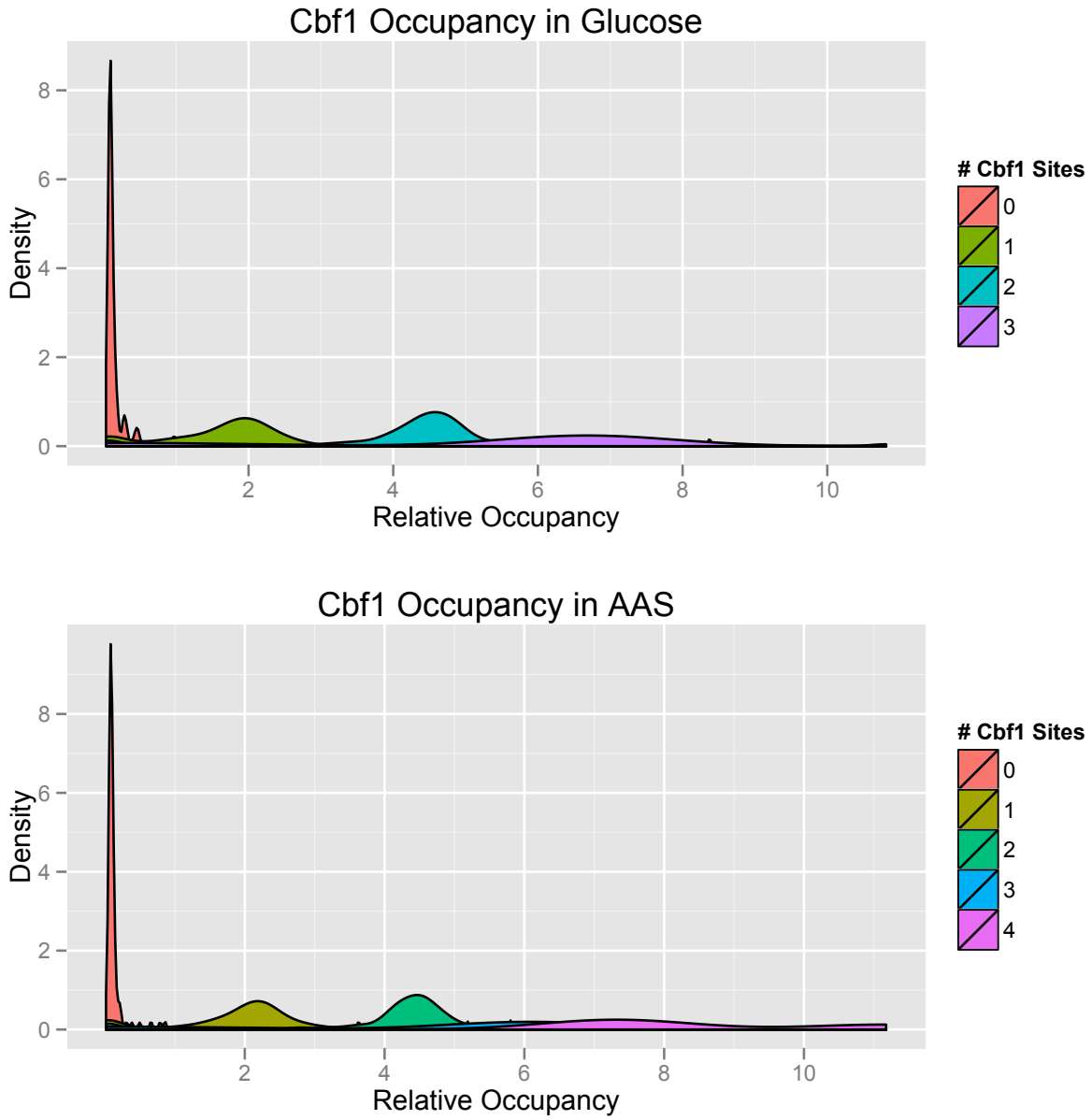
ChIP was performed on synthetic promoter-bearing strains with Avi-tagged Met31 and avi-tagged Nrg1 as outlined in methods. Enrichment was gauged by qPCR of a control (Ctl) region (SUC2) versus a target (Tgt) region (CAF120 for Met31, NRG1 for Nrg1). Gluc is in Glucose, AAS is in Amino Acid Starvation. The target region is not differentially bound with respect to the control region for Met31. Although Nrg1 looks significantly enriched, the level of enrichment is similar to levels obtained from ChIP done in strains with only BirA (data not shown).

Figure 3.6 ChIP with synthetic-promoter sequencing shows good technical replication



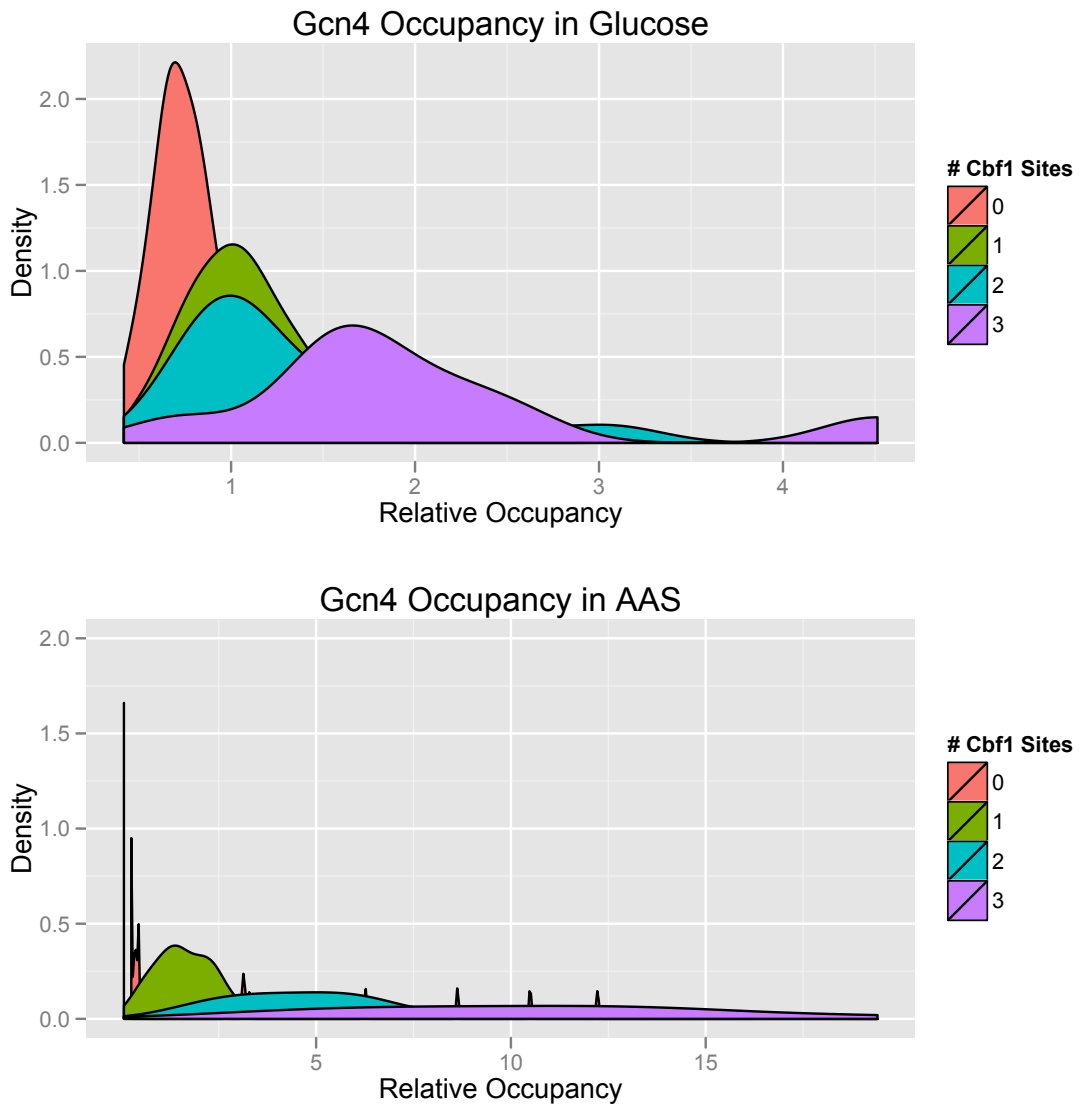
ChIP was performed in triplicate on synthetic promoter-bearing strains with an avi-tagged transcription factor, followed by specific sequencing of the ChIPed synthetic promoters. The relative occupancy for one replicate is plotted against the relative occupancy for another replicate for avi-tagged Cbf1 in both glucose and AAS conditions. The replicates show very good agreement, indicating that the relative occupancy values are highly reproducible from replicate-to-replicate.

Figure 3.7: Cbf1 enrichment is specific to Cbf1 sites



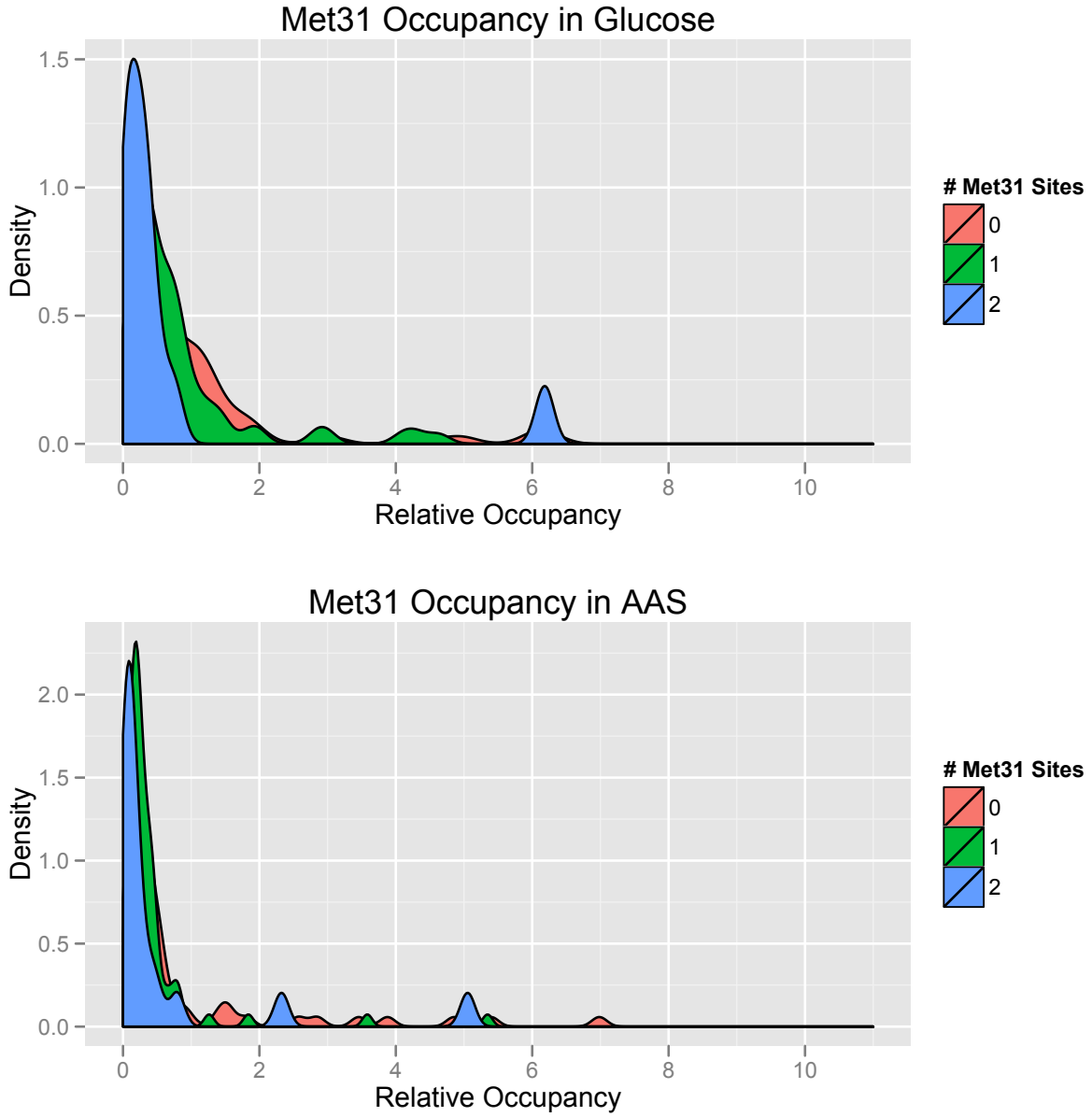
ChIP was performed on synthetic promoter-bearing strains with Avi-tagged Cbf1, followed by specific sequencing of the ChIPed synthetic promoters. The relative occupancy of Cbf1 is plotted here according to the number of Cbf1 sites in the promoter. There is a clear and nearly linear shift in the mean occupancy according to the number of Cbf1 sites present in both glucose and AAS conditions.

Figure 3.8: Gcn4 enrichment is specific to Gcn4 sites



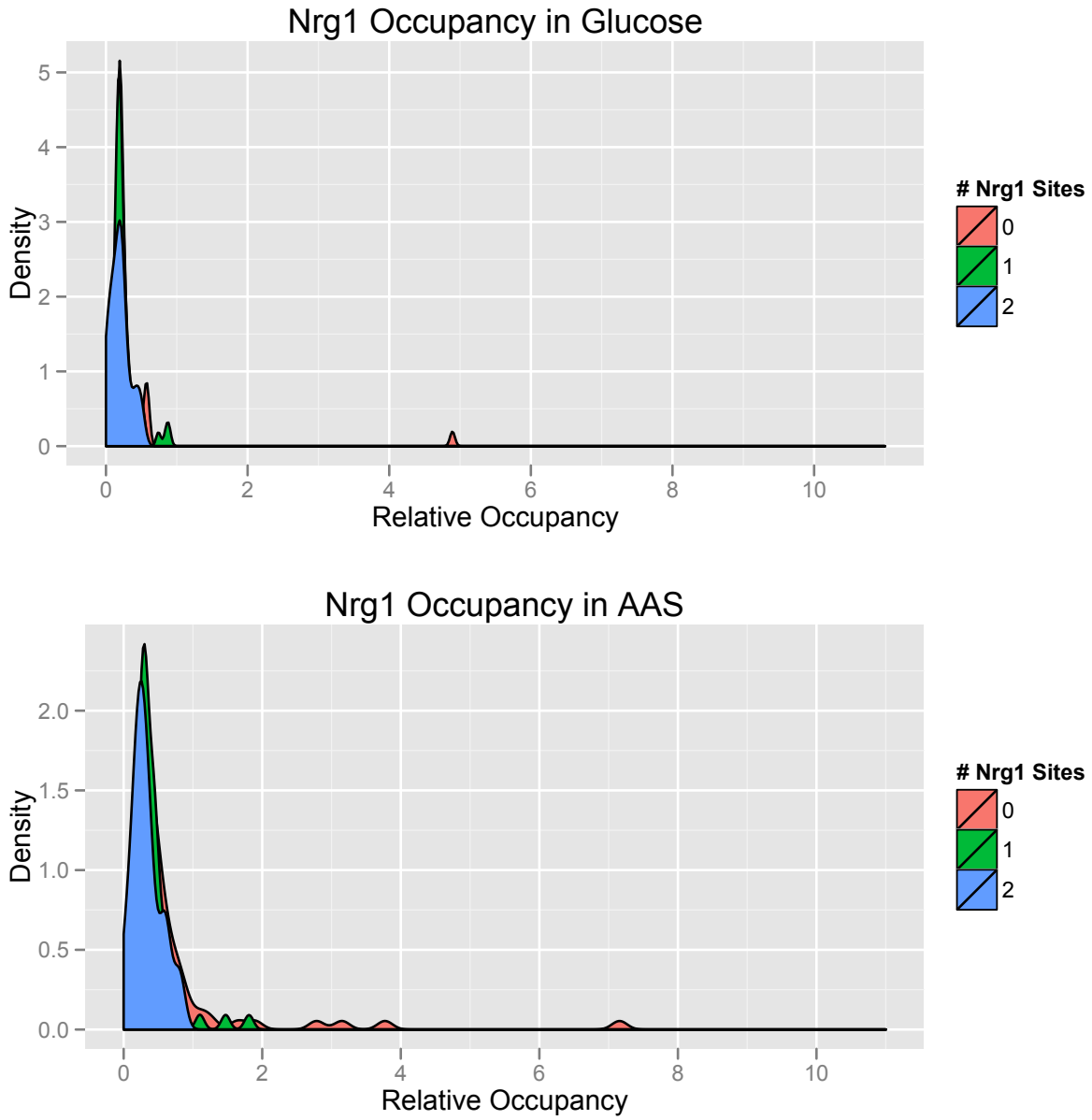
ChIP with synthetic promoter sequencing was performed in the Avi-tagged Gcn4 strain. The distribution of the relative occupancy of Gcn4 by the number of Gcn4 sites in the promoter is shown. The y-axis for the AAS has been truncated to 2. There is a clear shift in the mean occupancy with increasing Gcn4 site content, but the distribution does not shift linearly with the number of sites in Gcn4, and is much broader than the Cbf1 occupancy distributions, suggesting that other TF binding events have a larger impact on Gcn4 occupancy than for Cbf1.

Figure 3.9: Met31 ChIP shows no Met31/Met32 site-specific enrichment



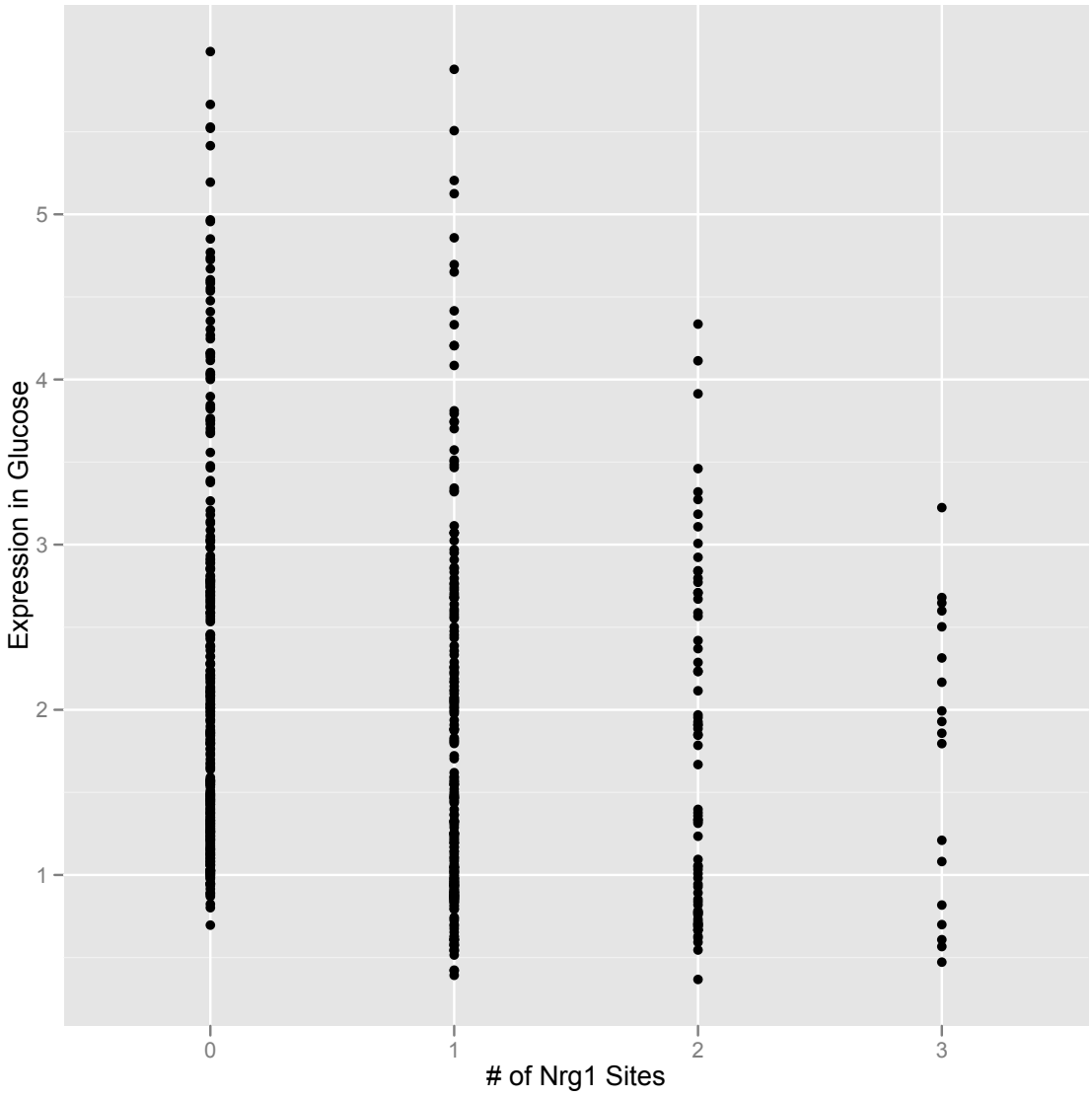
ChIP with synthetic promoter sequencing was performed in the avi-tagged Met31 strain. The distribution of the relative occupancy of Met31 by the number of Met31/Met32 sites in the promoter is shown. On average, there is no site-specific enrichment for either condition.

Figure 3.10: Nrg1 ChIP shows no Nrg1 site-specific enrichment



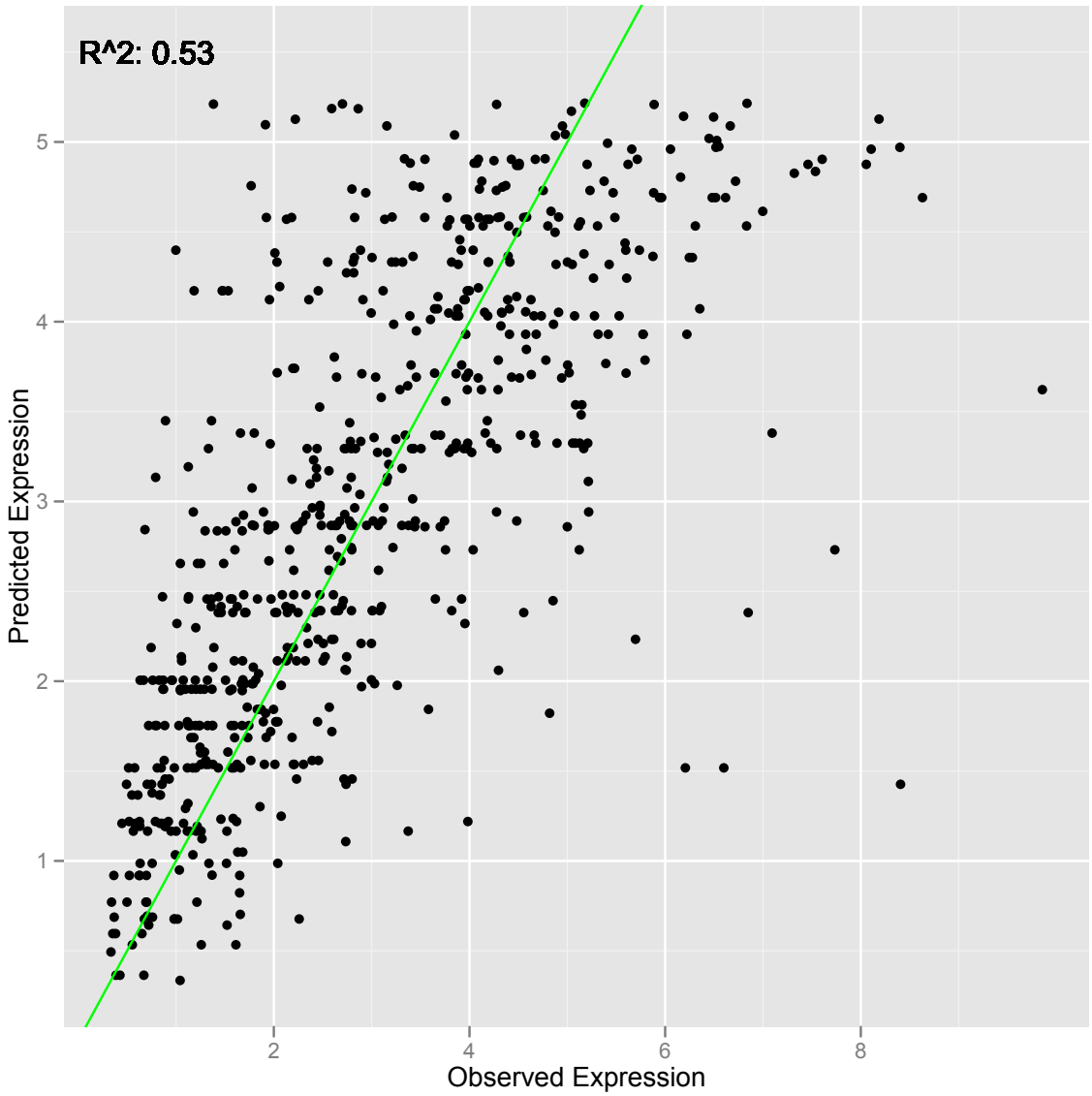
ChIP with synthetic promoter sequencing was performed in the avi-tagged Nrg1 strain. The distribution of the relative occupancy of Nrg1 by the number of Nrg1/Nrg1 sites in the promoter is shown. On average, there is no site-specific enrichment for either condition.

Figure 3.11: The Nrg1 site functions as a repressor in the tagged-Nrg1 strain



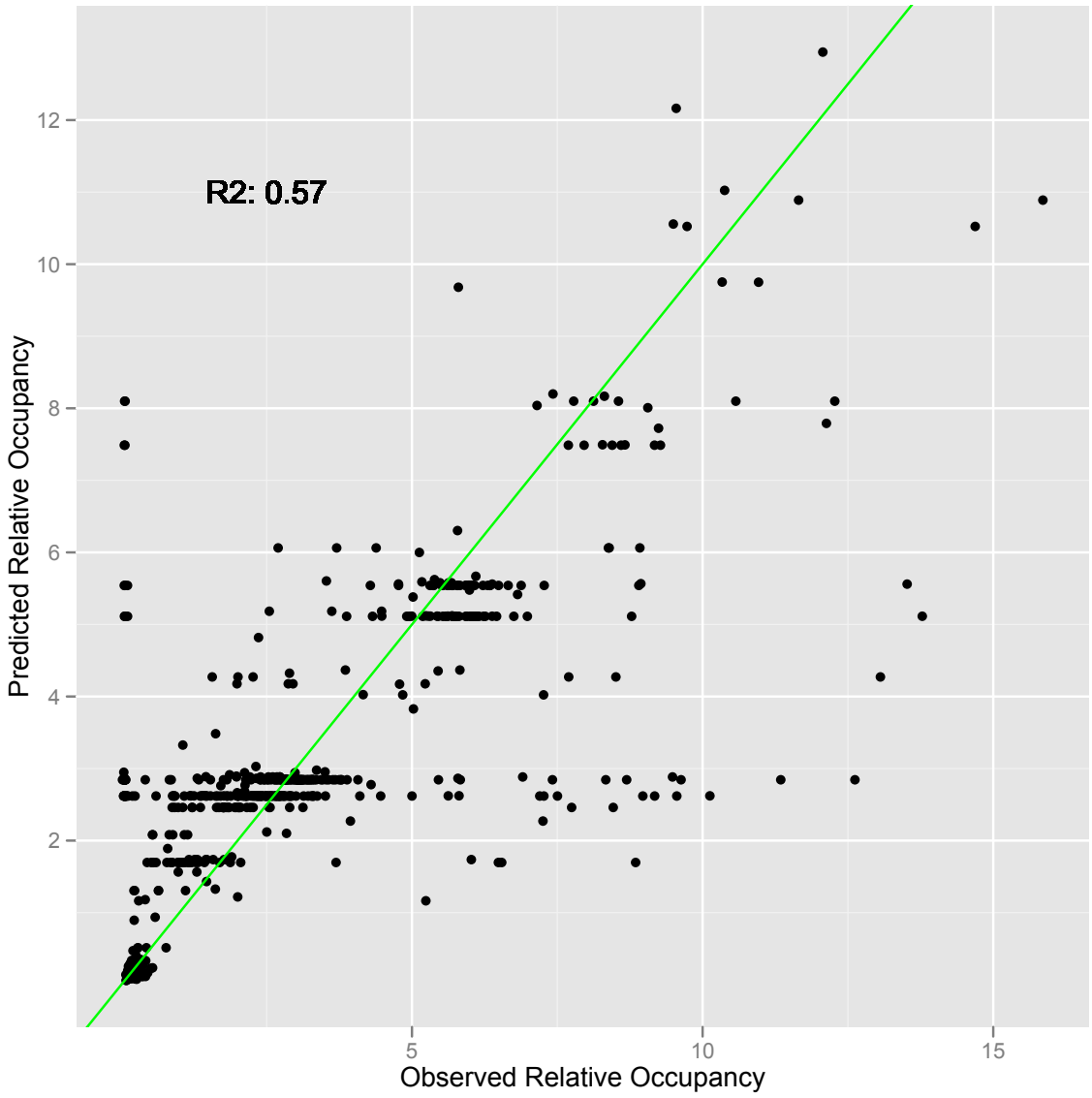
Strains bearing synthetic promoters with avi-tagged Nrg1 were grown to mid-log phase in glucose then fixed with formaldehyde (1% final concentration). Expression was measured via flow cytometry. The expression of promoters is plotted versus the total number of Nrg1 sites in the promoter. There is a clear decreasing trend in expression as a function of the number of Nrg1 sites showing that the site functions as a repressor.

Figure 3.12: Fit of expression by thermodynamic model



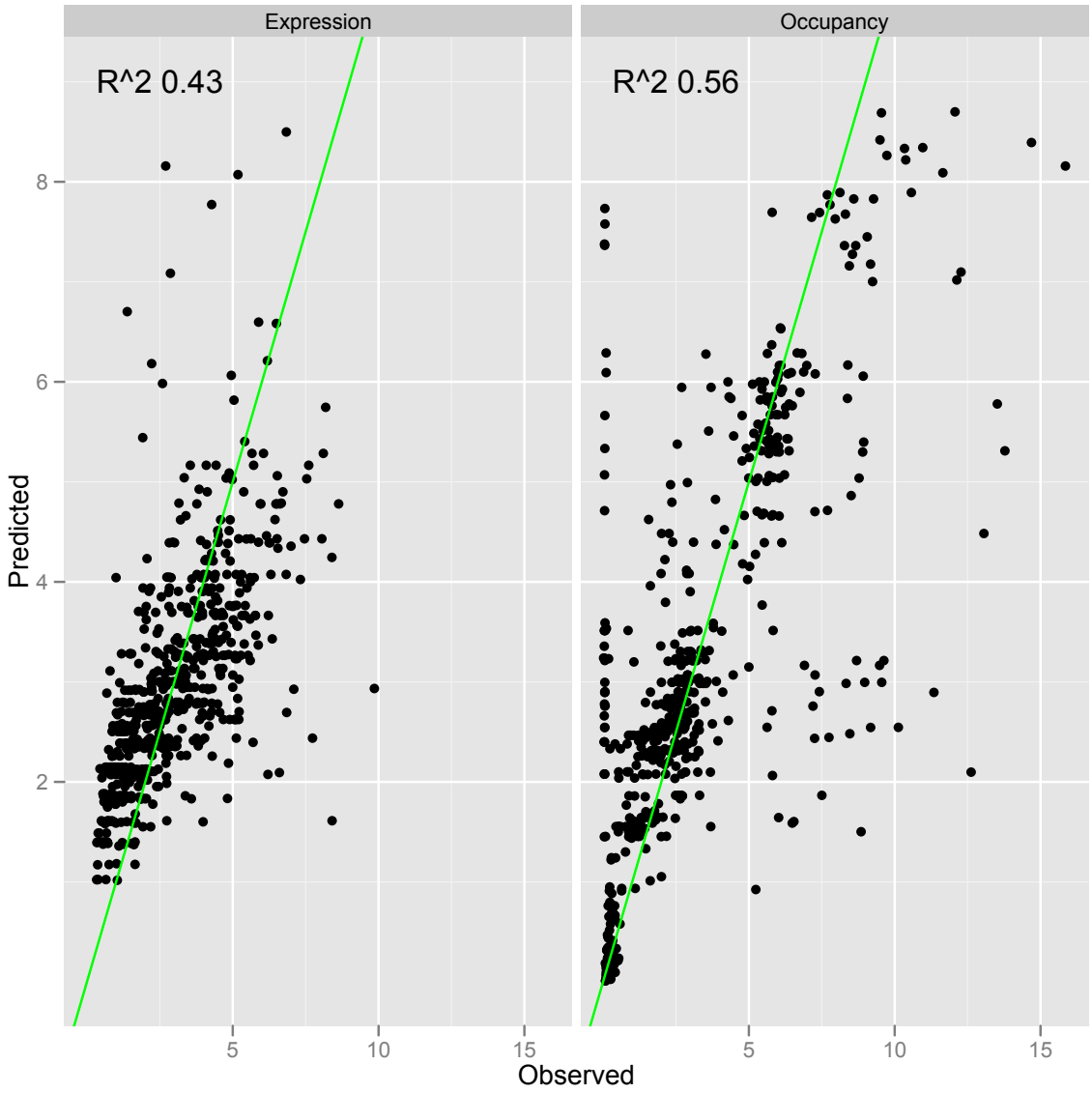
Synthetic bearing promoters with myc-C-avi-tagged Cbf1 or myc-C-avi-tagged Gcn4 were grown in glucose and AAS conditions and formaldehyde-fixed. YFP levels were measured via flow cytometry and normalized by cell volume and plate controls. The resulting data was used with the synthetic promoter sequence to parameterize a thermodynamic model of expression (Table 3.3). x-axis: the observed expression. y-axis: predicted expression. Green line: best fit line.

Figure 3.13: Fit of occupancy by thermodynamic model



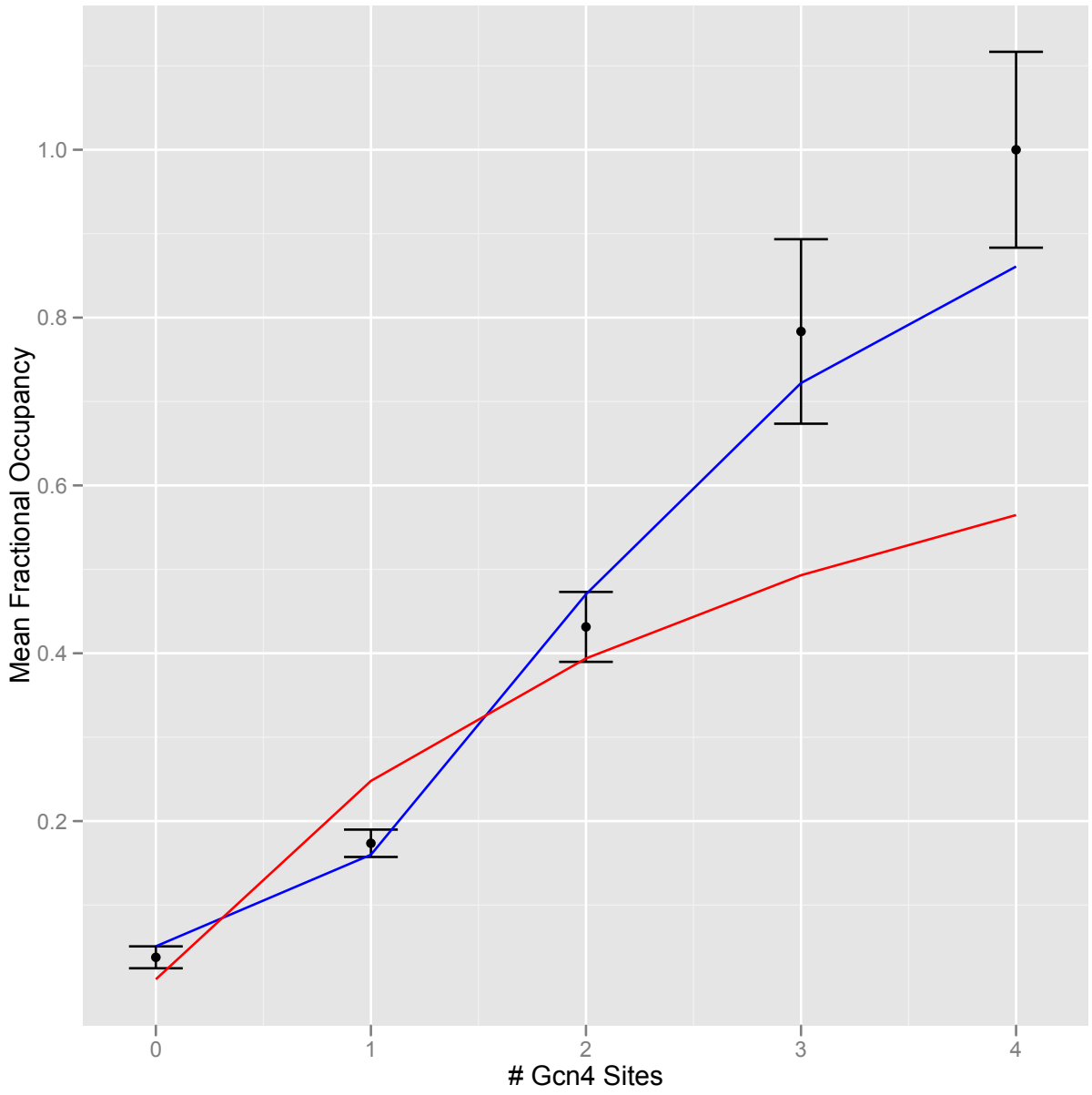
Synthetic bearing promoters with myc-C-avi-tagged Cbf1 or myc-C-avi-tagged Gcn4 were grown in glucose and AAS conditions and ChIPed as describe in Methods. The synthetic promoters in the IN and IP samples were sequenced and the normalized ratio of IP to IN counts used as the relative occupancy. A thermodynamic model was fit to the data (Table 3.3). x-axis: the observed relative occupancy. y-axis: the occupancy predicted by the model. Green line: the best-fit line.

Figure 3.14: Fit of occupancy and expression by thermodynamic model, no competitive binding



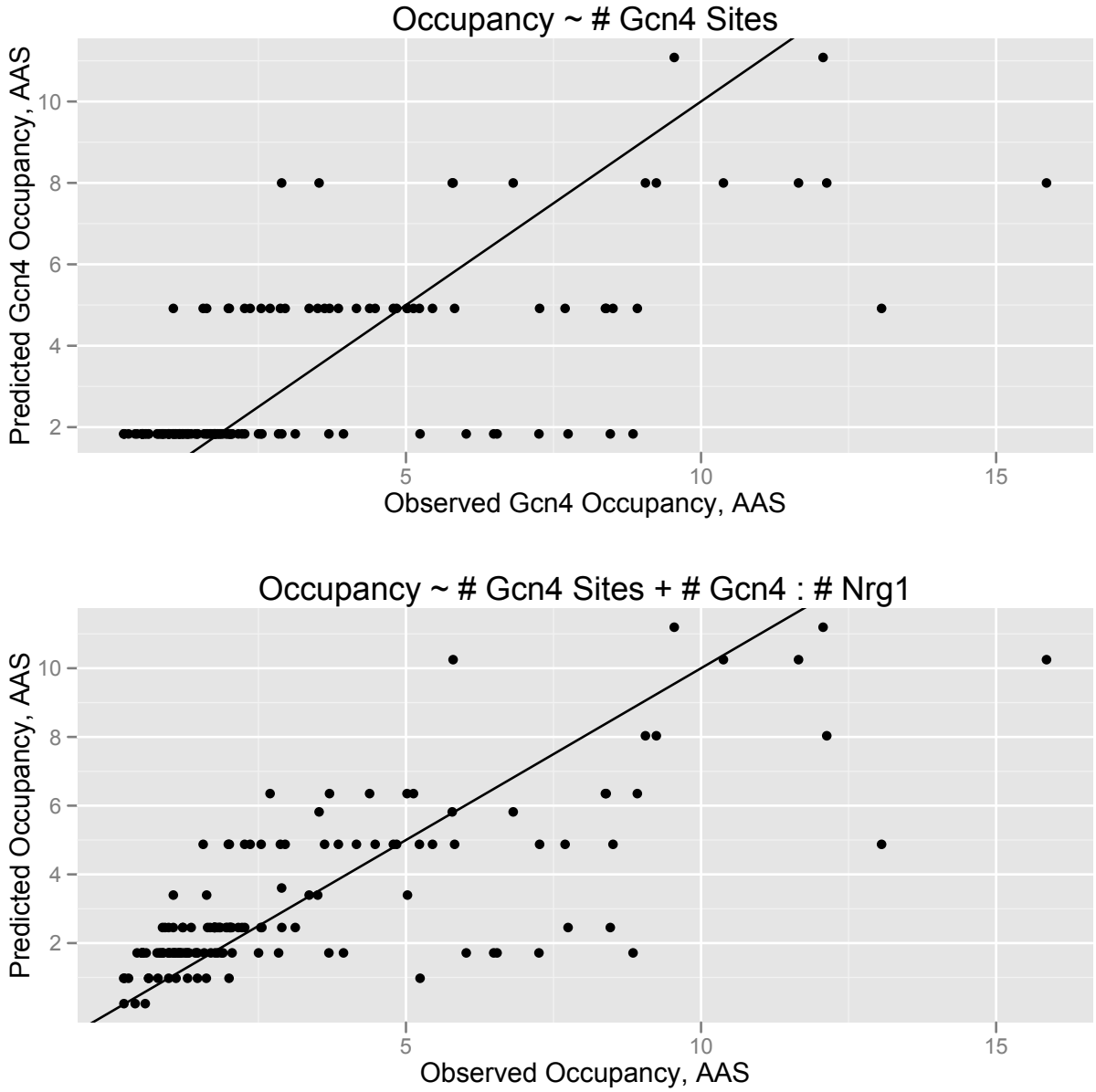
Synthetic bearing promoters with myc-C-avi-tagged Cbf1 or myc-C-avi-tagged Gcn4 were grown in glucose and AAS conditions and assayed for occupancy and expression as described in Methods. Both sets of data were used to fit a thermodynamic model of transcriptional regulation. x-axis: Observed values, either expression (left) or relative occupancy (right). y-axis: values predicted by the thermodynamic model.

Figure 3.15: Gcn4 binding in AAS acts cooperatively



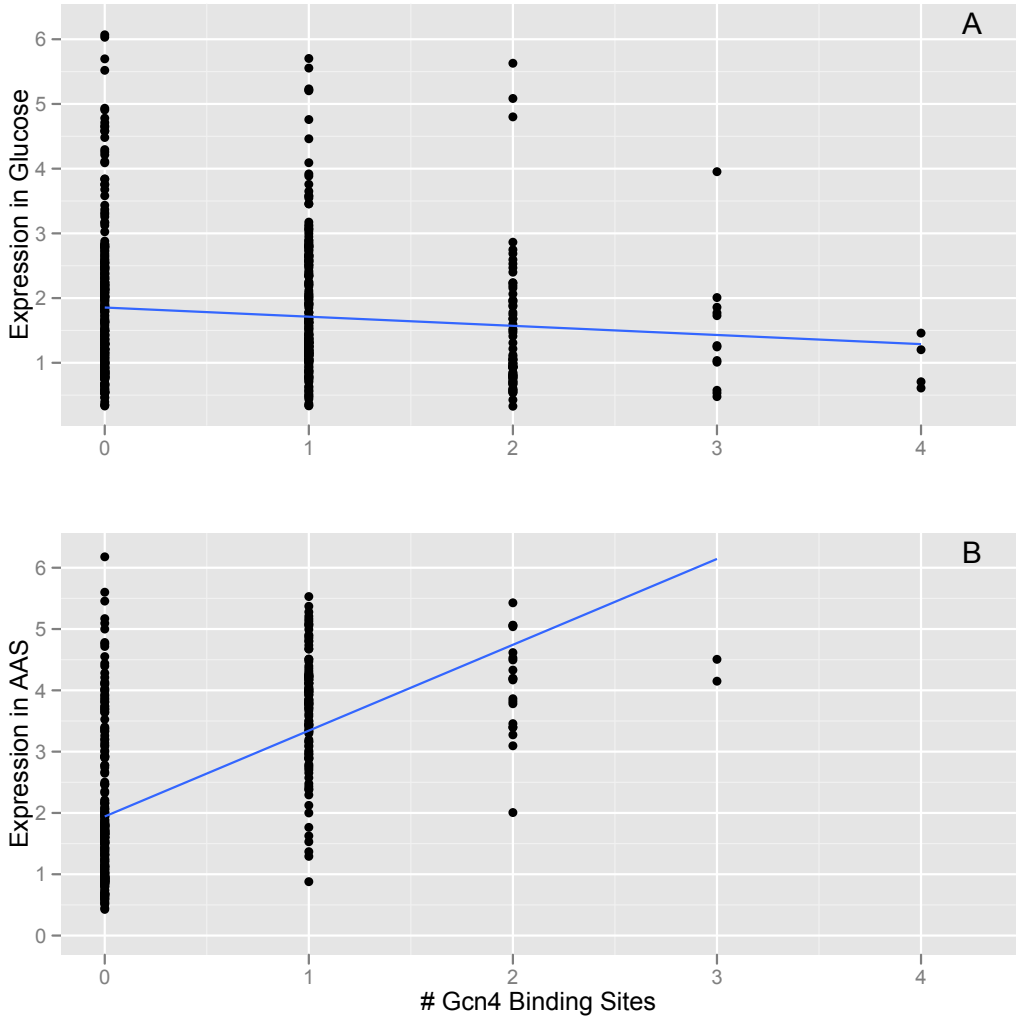
Synthetic promoters in a yeast strain bearing avi-tagged Gcn4 were grown in AAS conditions, ChIPed, and specifically sequenced. The resulting occupancy scores were normalized to the max of the mean occupancy by number of Gcn4 sites and fit with a Hill function with the Hill coefficient constrained to 1 (red) or allowed to vary (blue, Hill coefficient = 2.56; $P<e^{-16}$, t-test).

Figure 3.16: Nrg1-Gcn4 interaction negatively affects Gcn4 occupancy



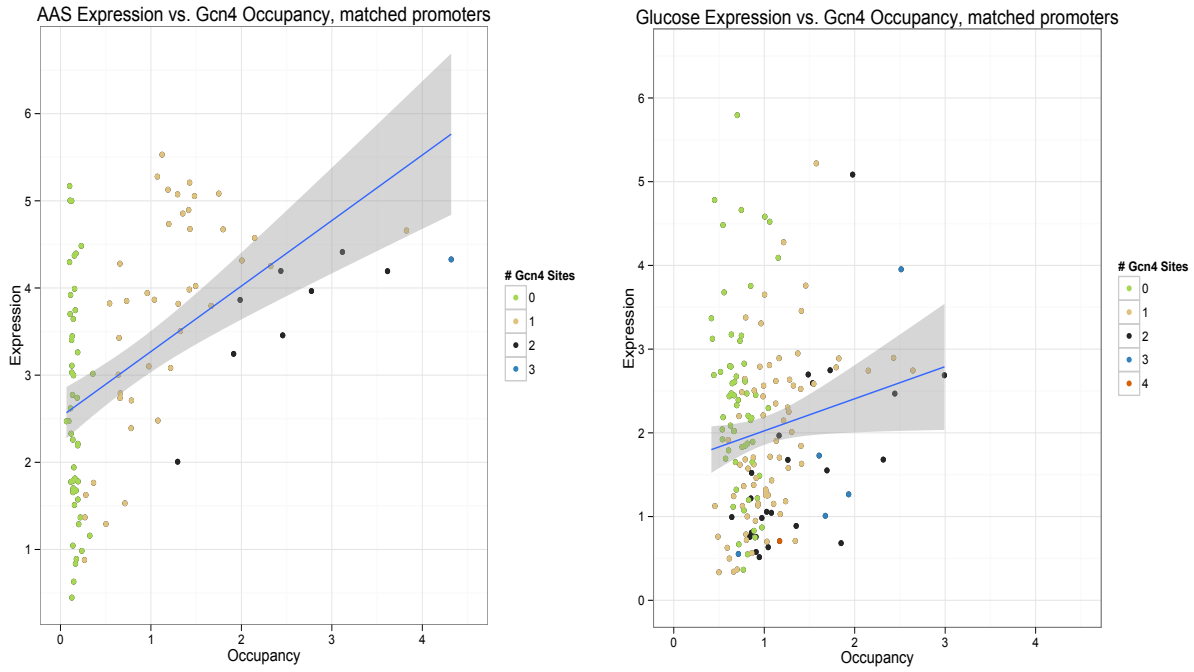
Avi-tagged Gcn4 strains carrying synthetic promoters were grown in AAS media and ChIPed as per methods. Top: a linear model built only using the number of Gcn4 sites as a predictor. Bottom: a linear model built using the number of Gcn4 sites and the interaction between the number of Gcn4 sites and the number of Nrg1 sites. All parameters are significant by t-test ($P < 10^{-6}$).

Figure 3.17: The Gcn4 site functions as a weak repressor in glucose and a strong activator in AAS



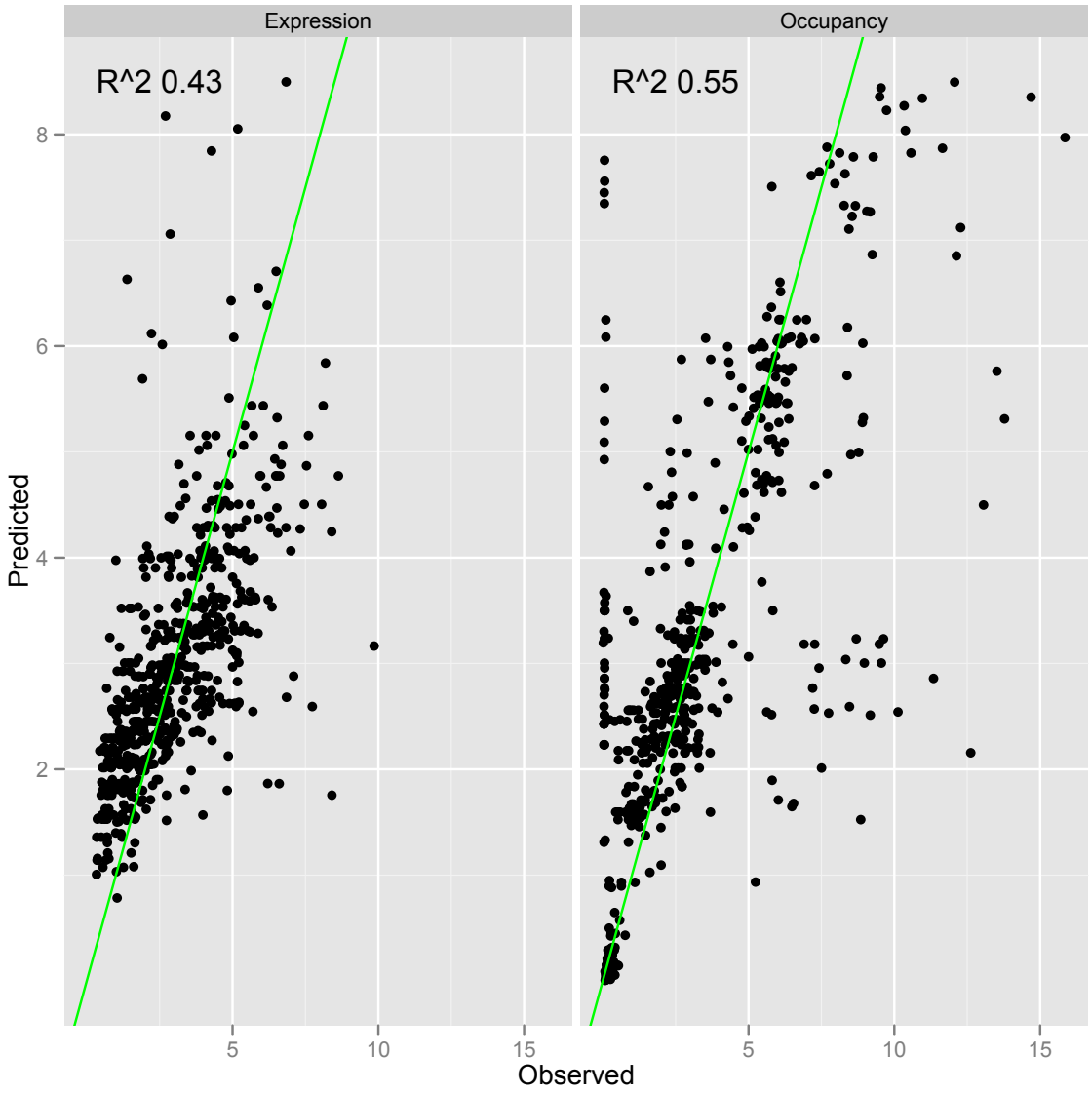
Strains bearing synthetic promoters with avi-tagged Gcn4 were grown to mid-log phase in glucose and formaldehyde-fixed or switched to AAS media, grown for six hours and formaldehyde-fixed. Expression was measured via flow cytometry. The expression of promoters is plotted versus the total number of Gcn4 sites. The blue line is a regression of expression on the number of Gcn4 sties. In glucose (A), the Gcn4 site is weakly repressing but in AAS (B) it is strongly activating.

Figure 3.18: Occupancy of Gcn4 is positively correlated with expression



Expression is plotted as a function of relative occupancy for synthetic promoters in strains bearing avi-tagged Gcn4. There is a clear positive correlation between occupancy and expression in AAS. There is also a positive correlation in glucose, despite the Gcn4 site behaving like a repressor in that condition. This suggests that another factor is competing with Gcn4 for binding in glucose.

Figure 3.19: Fit of occupancy and expression by thermodynamic model, with competitive binding



Synthetic bearing promoters with myc-C-avi-tagged Cbf1 or myc-C-avi-tagged Gcn4 were grown in glucose and AAS conditions and assayed for occupancy and expression as described in Methods. Both sets of data were used to fit a thermodynamic model of transcriptional regulation that incorporated competitive binding with Gcn4. x-axis: Observed values, either expression (left) or relative occupancy (right). y-axis: values predicted by the thermodynamic model.

Tables

Table 3.1: Summary of usable promoters for expression analysis

Tagged TF	Total, Glucose	Unique, Glucose	Total, AAS	Unique, AAS
Cbf1	529	218	374	125
Gcn4	614	213	396	114
Met31[a]	643	271	475	170
Nrg1[a]	634	271	393	139

[a]Omitted from expression analysis due to lack of ChIP signal for occupancy analysis.

Cbf1, Gcn4, Met31, and Nrg1 were tagged with the myc-C-avi tag (van Werven and Timmers, 2006) in a strain harboring the bacterial biotin ligase BirA. Synthetic promoters containing sites for all four factors were constructed in each strain. 960 colonies were picked for each library, purified, sequenced, then grown in glucose and AAS. The library members were crosslinked, then run on a Beckman Coulter Cell Lab Quanta SC flow cytometer. The numbers shown are the number of strains for which sequence information was determined and for which a reliable fluorescence value was obtained.

Table 3.2: Summary of usable promoters for occupancy analysis

Tagged TF	Glucose	AAS
Cbf1	290	291
Gcn4	199	229
Met31[a]	0	0
Nrg1[a]	0	0

[a]No observable ChIP signal.

ChIP was performed on the libraries of synthetic promoters and the promoters specifically sequenced as described in Methods. Promoters with fewer than 50 reads in the input replicates were discarded. Met31 and Nrg1 showed no specific enrichment, so all promoters were discarded. The table summarizes the total number of promoters used for analysis for each factor and condition.

Table 3.3: Parameter values from thermodynamic model fitting to be finished

Fit Type	Parameter	Value (+/- 95% CI)
Expression only	$\Delta G_{Cbf1\text{-}DNA,glucose}$	1.32±0.71
Expression only	$\Delta G_{Met31/Met32\text{-}DNA,glucose}$	0.53±0.78
Expression only	$\Delta G_{Nrg1\text{-}DNA,glucose}$	0.41±0.47
Expression only	$\Delta G_{Cbf1,tagged-RNAP}$	-3.84±0.71
Expression only	$\Delta G_{Cbf1,untagged-RNAP}$	-1.14±0.72
Expression only	$\Delta G_{Gcn4,aas-RNAP}$	-1.55±0.32
Expression only	$\Delta G_{Gcn4,gluc-RNAP}$	0.48±0.33
Expression only	$\Delta G_{Met31/Met32-RNAP}$	-1.11±0.32
Expression only	$\Delta G_{Nrg1-RNAP}$	5.08±35.7
Expression only	$\Delta G_{RNAP\text{-}DNA}$	-0.53±.28
Occupancy only	$\Delta G_{Cbf1\text{-}DNA,glucose}$	3.00±1.86
Occupancy only	$\Delta G_{Cbf1\text{-}DNA,AAS}$	2.91±1.87
Occupancy only	$\Delta G_{Gcn4\text{-}DNA,glucose}$	5.80± 2.17
Occupancy only	$\Delta G_{Gcn4\text{-}DNA,AAS}$	3.06± 1.82
Occupancy only	$\Delta G_{Gcn4\text{-}Gcn4}$	-2.62±1.24
Occupancy only	$\Delta G_{Gcn4\text{-}Nrg1}$	1.02±0.65
Expression and Occupancy	$\Delta G_{Cbf1\text{-}DNA,glucose}$	4.87*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}DNA,glucose}$	2.92*
Expression and Occupancy	$\Delta G_{Met31/Met32\text{-}DNA,glucose}$	0.6*
Expression and Occupancy	$\Delta G_{Nrg1\text{-}DNA,glucose}$	-1.26*
Expression and Occupancy	$\Delta G_{Cbf1\text{-}DNA,AAS}$	4.95*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}DNA,AAS}$	1.23*
Expression and Occupancy	ΔG_{RNAP}	1.00*

Fit Type	Parameter	Value (+/- 95% CI)
Expression and Occupancy	$\Delta G_{Cbf1,tagged\text{-}RNAP}$	-5.36*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}RNAP,glucose}$	17.65*
Expression and Occupancy	$\Delta G_{Met31/32\text{-}RNAP}$	-0.42*
Expression and Occupancy	$\Delta G_{Nrg1\text{-}RNAP}$	0.49*
Expression and Occupancy	$\Delta G_{Cbf1,untagged\text{-}RNAP}$	-3.15*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}RNAP,\ AAS}$	-0.60*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}Gcn4}$	-2.01*
Expression and Occupancy	$\Delta G_{Gcn4\text{-}Nrg1}$	-3.00*
Expression and Occupancy, Competitive	$\Delta G_{Cbf1\text{-}DNA,Glucose}$	5.04±9.89
Expression and Occupancy, Competitive	$\Delta G_{Gcn4,Glucose}$	3.82±1.77
Expression and Occupancy, Competitive	$\Delta G_{Met31/Met32\text{-}DNA,glucose}$	-1.30±-0.41
Expression and Occupancy, Competitive	$\Delta G_{Nrg1\text{-}DNA,glucose}$	-0.93±1.49
Expression and Occupancy, Competitive	$\Delta G_{Cbf1\text{-}DNA,AAS}$	5.13±9.89
Expression and Occupancy, Competitive	$\Delta G_{Gcn4\text{-}DNA,AAS}$	0.64±0.34
Expression and Occupancy, Competitive	ΔG_{RNAP}	0.92±0.22
Expression and Occupancy, Competitive	$\Delta G_{Cbf1,tagged\text{-}RNAP}$	-5.54±10.0
Expression and Occupancy, Competitive	$\Delta G_{Gcn4\text{-}RNAP,glucose}$	-0.82±0.34
Expression and Occupancy, Competitive	$\Delta G_{Met31/32\text{-}RNAP}$	-0.38±0.21
Expression and Occupancy, Competitive	$\Delta G_{Nrg1\text{-}RNAP}$	0.52±0.29
Expression and Occupancy, Competitive	$\Delta G_{Cbf1,untagged\text{-}RNAP}$	-3.3±9.63

Fit Type	Parameter	Value (+/- 95% CI)
Expression and Occupancy, Competitive	$\Delta G_{Gcn4\text{-}RNAP,\ AAS}$	0.31±0.22
Expression and Occupancy, Competitive	$\Delta G_{Gcn4\text{-}Gcn4}$	-1.78±0.56
Expression and Occupancy, Competitive	$\Delta G_{Gcn4\text{-}Nrg1}$	3.6±12.90

* Confidence interval estimates could not be calculated due to numerical instabilities introduced by the $\Delta G_{Gcn4\text{-}RNAP,}$ glucose parameter.

Table 3.4: Overall fits and cross validation results

Data Used	# Parameters	Expression R^2	Occupancy R^2	Cross Validation
Expression-only	10	0.53	0.36	0.53
Occupancy-only	6	NA	0.57	0.57
Expression and Occupancy, noncompetitive	15	0.425	0.556	0.42 (expression) 0.56(occupancy)
Expression and Occupancy, competitive	15	0.431	0.554	0.43 (expression) 0.56 (occupancy)

Expression, occupancy, or expression and occupancy were modeled using the thermodynamic model described in Methods. Each model was fit with the number of parameters indicated (see Table 3.3 for specific parameter details). The model fit with only expression was also used to predict occupancy. The occupancy-only model cannot be used to predict expression since fitting of RNAP interaction terms was not attempted with only occupancy data. When fitting with expression and occupancy, the Gcn4 site was modeled without and with competitive binding (noncompetitive and competitive, respectively). Five-fold cross validation was performed on all models and and the mean R^2 across the validations is reported in the Cross Validation column.

Table 3.5: Barcoded "well" and "plate" PCR primers used for library sequencing.

Sequence	Barcode	Use
CTACGTCGACACACACTTAATCGTTCTTCCACACGGATC	ACACACT	WA1[a]
CTACGTCGACACACTACTAATCGTTCTTCCACACGGATC	ACACTAC	WA2[a]
CTACGTCGACACAGATGTAATCGTTCTTCCACACGGATC	ACAGATG	WA3[a]
CTACGTCGACACATCGTTAATCGTTCTTCCACACGGATC	ACATCGT	WA4[a], CLP1[b]
CTACGTCGACACATGAGTAATCGTTCTTCCACACGGATC	ACATGAG	WA5[a], CLP2[b]
CTACGTCGACACGAGACTAATCGTTCTTCCACACGGATC	ACGAGAC	WA6[a], CLP3[b]
CTACGTCGACACGTCTGTAATCGTTCTTCCACACGGATC	ACGTCTG	WA7[a], CAP1[b]
CTACGTCGACACTACTATAATCGTTCTTCCACACGGATC	ACTACTA	WA8[a], CAP2[b]
CTACGTCGACACTAGCTTAATCGTTCTTCCACACGGATC	ACTAGCT	WA9[a], CAP3[b]
CTACGTCGACACTATGCTAATCGTTCTTCCACACGGATC	ACTATGC	WA10[a], GLP1[b]
CTACGTCGACACTGAGATAATCGTTCTTCCACACGGATC	ACTGAGA	WA11[a], GLP2[b]
CTACGTCGACACTGCATTAATCGTTCTTCCACACGGATC	ACTGCAT	WA12[a]
CTACGTCGACAGACAGCTAATCGTTCTTCCACACGGATC	AGACAGC	WB1[a]
CTACGTCGACAGAGCACTAATCGTTCTTCCACACGGATC	AGAGCAC	WB2[a]
CTACGTCGACAGATGCATAATCGTTCTTCCACACGGATC	AGATGCA	WB3[a]
CTACGTCGACAGCAGCGTAATCGTTCTTCCACACGGATC	AGCAGCG	WB4[a], GLP3[b]
CTACGTCGACAGCATGATAATCGTTCTTCCACACGGATC	AGCATGA	WB5[a], GAP1

Sequence	Barcode	Use
CTACGTCGACAGCGAGTTAATCGTTCTTCCACACGGATC	AGCGAGT	WB6[a], GAP2[b]
CTACGTCGACAGCTATCTAATCGTTCTTCCACACGGATC	AGCTATC	WB7[a], GAP3[b]
CTACGTCGACAGCTCATTAATCGTTCTTCCACACGGATC	AGCTCAT	WB8[a], MLP1[b]
CTACGTCGACAGTACAGTAATCGTTCTTCCACACGGATC	AGTACAG	WB9[a], MLP2[b]
CTACGTCGACATAGCGATAATCGTTCTTCCACACGGATC	ATAGCGA	WB10[a], MLP3[b]
CTACGTCGACATCACACTAATCGTTCTTCCACACGGATC	ATCACAC	WB11[a], MAP1[b]
CTACGTCGACATCTACATAATCGTTCTTCCACACGGATC	ATCTACA	WB12[a], MAP2[b]
CTACGTCGACATGCAGACTAATCGTTCTTCCACACGGATC	ATGCAGAC	WC1[a]
CTACGTCGACATGCTCGCTAATCGTTCTTCCACACGGATC	ATGCTCGC	WC2[a]
CTACGTCGACCACACATCTAATCGTTCTTCCACACGGATC	CACACATC	WC3[a]
CTACGTCGACCACTACGCTAATCGTTCTTCCACACGGATC	CACTACGC	WC4[a], GAK3[b]
CTACGTCGACCACTCTCCTAATCGTTCTTCCACACGGATC	CACTCTCC	WC5[a]
CTACGTCGACCAGATAGCTAATCGTTCTTCCACACGGATC	CAGATAGC	WC6[a]
CTACGTCGACCAGCGCTCTAATCGTTCTTCCACACGGATC	CAGCGCTC	WC7[a]
CTACGTCGACCATATCACTAATCGTTCTTCCACACGGATC	CATATCAC	WC8[a]
CTACGTCGACCATGATCCTAATCGTTCTTCCACACGGATC	CATGATCC	WC9[a]
CTACGTCGACCGACGAGCTAATCGTTCTTCCACACGGATC	CGACGAGC	WC10[a], MAP3[b]
CTACGTCGACCGAGACGCTAATCGTTCTTCCACACGGATC	CGAGACGC	WC11[a], NLP1[b]

Sequence	Barcode	Use
CTACGTCGACCGATAGACTAATCGTTCTTCCACACGGATC	CGATAGAC	WC12[a]
CTACGTCGACCGCGCTGCTAATCGTTCTTCCACACGGATC	CGCGCTGC	WD1[a]
CTACGTCGACCGCTGACCTAATCGTTCTTCCACACGGATC	CGCTGACC	WD2[a]
CTACGTCGACCGTCACACTAATCGTTCTTCCACACGGATC	CGTCACAC	WD3[a]
CTACGTCGACCGTGTATCTAATCGTTCTTCCACACGGATC	CGTGTATC	WD4[a]
CTACGTCGACCTACAGTCTAATCGTTCTTCCACACGGATC	CTACAGTC	WD5[a]
CTACGTCGACCTAGCATCTAATCGTTCTTCCACACGGATC	CTAGCATC	WD6[a]
CTACGTCGACCTATATGCTAATCGTTCTTCCACACGGATC	CTATATGC	WD7[a]
CTACGTCGACCTCAGCACTAATCGTTCTTCCACACGGATC	CTCAGCAC	WD8[a]
CTACGTCGACCTCGAGCCTAATCGTTCTTCCACACGGATC	CTCGAGCC	WD9[a]
CTACGTCGACCTCGTAGCTAATCGTTCTTCCACACGGATC	CTCGTAGC	WD10[a], NLP2[b]
CTACGTCGACCTCTCGTCTAATCGTTCTTCCACACGGATC	CTCTCGTC	WD11[a], NLP3[b]
CTACGTCGACCTGACGCCTAATCGTTCTTCCACACGGATC	CTGACGCC	WD12[a], NAP1[b]
CTACGTCGACCTGCGACGCTAATCGTTCTTCCACACGGATC	CTGCGACGC	WE1[a]
CTACGTCGACCTGTCAGGCTAATCGTTCTTCCACACGGATC	CTGTCAGGC	WE2[a]
CTACGTCGACGACATCTGCTAATCGTTCTTCCACACGGATC	GACATCTGC	WE3[a]
CTACGTCGACGACGCGAGCTAATCGTTCTTCCACACGGATC	GACGCGAGC	WE4[a], NAP2[b]
CTACGTCGACGAGACACGCTAATCGTTCTTCCACACGGATC	GAGACACGC	WE5[a]
CTACGTCGACGAGCACGGCTAATCGTTCTTCCACACGGATC	GAGCACGGC	WE6[a]
CTACGTCGACGAGTAGCGCTAATCGTTCTTCCACACGGATC	GAGTAGCGC	WE7[a], NAP3[b]
CTACGTCGACGAGTGTAGCTAATCGTTCTTCCACACGGATC	GAGTGTAGC	WE8[a]

Sequence	Barcode	Use
CTACGTCGACGATAGATGCTAATCGTTCTTCCACACGGATC	GATAGATGC	WE9[a]
CTACGTCGACGATCAGAGCTAATCGTTCTTCCACACGGATC	GATCAGAGC	WE10[a], CLK1[b]
CTACGTCGACGATGTAGGCTAATCGTTCTTCCACACGGATC	GATGTAGGC	WE11[a], CLK2[b]
CTACGTCGACGCACTCAGCTAATCGTTCTTCCACACGGATC	GCACTCAGC	WE12[a], CLK3[b]
CTACGTCGACGCAGAGTGCTAATCGTTCTTCCACACGGATC	GCAGAGTGC	WF1[a]
CTACGTCGACGCAGCAGGCTAATCGTTCTTCCACACGGATC	GCAGCAGGC	WF2[a]
CTACGTCGACGCGACGAGCTAATCGTTCTTCCACACGGATC	GCGACGAGC	WF3[a], CAK1[b]
CTACGTCGACGCTCATGGCTAATCGTTCTTCCACACGGATC	GCTCATGGC	WF4[a], CAK2[b]
CTACGTCGACGCTCGACGCTAATCGTTCTTCCACACGGATC	GCTCGACGC	WF5[a], CAK3[b]
CTACGTCGACGTACATCGCTAATCGTTCTTCCACACGGATC	GTACATCGC	WF6[a]
CTACGTCGACGTAGACAGCTAATCGTTCTTCCACACGGATC	GTAGACAGC	WF7[a]
CTACGTCGACGTATCACGCTAATCGTTCTTCCACACGGATC	GTATCACGC	WF8[a]
CTACGTCGACGTCACTGGCTAATCGTTCTTCCACACGGATC	GTCACTGGC	WF9[a], GLK1[b]
CTACGTCGACGTCTGATGCTAATCGTTCTTCCACACGGATC	GTCTGATGC	WF10[a]
CTACGTCGACGTGAGCGGCTAATCGTTCTTCCACACGGATC	GTGAGCGGC	WF11[a], GLK2[b]
CTACGTCGACGTGCTATGCTAATCGTTCTTCCACACGGATC	GTGCTATGC	WF12[a], GLK3[b]
CTACGTCGACGTGTACTAGCTAATCGTTCTTCCACACGGATC	GTGTACTAGC	WG1[a]
CTACGTCGACTACACTAAGCTAATCGTTCTTCCACACGGATC	TACACTAAGC	WG2[a]

Sequence	Barcode	Use
CTACGTCGACTACAGAGAGCTAATCGTTCTTCCACACGGATC	TACAGAGAGC	WG3[a], GAK1[b]
CTACGTCGACTACGACTAGCTAATCGTTCTTCCACACGGATC	TACGACTAGC	WG4[a], GAK2[b]
CTACGTCGACTAGAGCAAGCTAATCGTTCTTCCACACGGATC	TAGAGCAAGC	WG5[a]
CTACGTCGACTAGCTACAGCTAATCGTTCTTCCACACGGATC	TAGCTACAGC	WG6[a], CAK3[b]
CTACGTCGACTAGTCGTAGCTAATCGTTCTTCCACACGGATC	TAGTCGTAGC	WG7[a]
CTACGTCGACTATATGTAGCTAATCGTTCTTCCACACGGATC	TATATGTAGC	WG8[a]
CTACGTCGACTATCGCGAGCTAATCGTTCTTCCACACGGATC	TATCGCGAGC	WG9[a], MLK1[b]
CTACGTCGACTATGCACAGCTAATCGTTCTTCCACACGGATC	TATGCACAGC	WG10[a], MLK2[b]
CTACGTCGACTCACGATAGCTAATCGTTCTTCCACACGGATC	TCACGATAGC	WG11[a], MLK3[b]
CTACGTCGACTCATAGCAGCTAATCGTTCTTCCACACGGATC	TCATAGCAGC	WG12[a], MAK1
CTACGTCGACTCATGTAAGCTAATCGTTCTTCCACACGGATC	TCATGTAAGC	WH1[a]
CTACGTCGACTCGACATAGCTAATCGTTCTTCCACACGGATC	TCGACATAGC	WH2[a]
CTACGTCGACTCGCACAAGCTAATCGTTCTTCCACACGGATC	TCGCACAAGC	WH3[a]
CTACGTCGACTCTATAGAGCTAATCGTTCTTCCACACGGATC	TCTATAGAGC	WH4[a], MAK2[b]
CTACGTCGACTCTGACGAGCTAATCGTTCTTCCACACGGATC	TCTGACGAGC	WH5[a], MAK3[b]
CTACGTCGACTGAGTAGAGCTAATCGTTCTTCCACACGGATC	TGAGTAGAGC	WH6[a], NLK1[b]
CTACGTCGACTGCATACAGCTAATCGTTCTTCCACACGGATC	TGCATACAGC	WH7[a], NLK2[b]

Sequence	Barcode	Use
CTACGTCGACTGCGTCAAGCTAATCGTTCTTCCACACGGATC	TGCGTCAAGC	WH8[a], NLK3[b]
CTACGTCGACTGCTCGAAGCTAATCGTTCTTCCACACGGATC	TGCTCGAAGC	WH9[a], NAK1[b]
CTACGTCGACTGCTGTGAGCTAATCGTTCTTCCACACGGATC	TGCTGTGAGC	WH10[a], NAK2[b]
CTACGTCGACTGTAGTCAGCTAATCGTTCTTCCACACGGATC	TGTAGTCAGC	WH11[a], NAK3[b]
CTACGTCGACTGTCAGTAGCTAATCGTTCTTCCACACGGATC	TGTCAGTAGC	WH12[a]
ACGTACAATTGACGATGTTGAGAACGGTTCGGCATTG	ACGAT	Cbf1 P1[a]
ACGTACAATTGACGCAGTTGAGAACGGTTCGGCATTG	ACGCA	Cbf1 P2[a]
ACGTACAATTGACGTGGTTGAGAACGGTTCGGCATTG	ACGTG	Cbf1 P3[a]
ACGTACAATTGAGCGCGTTGAGAACGGTTCGGCATTG	AGCGC	Cbf1 P4[a]
ACGTACAATTGAGCTGGTTGAGAACGGTTCGGCATTG	AGCTG	Cbf1 P5[a]
ACGTACAATTGAGTCGGTTGAGAACGGTTCGGCATTG	AGTCG	Cbf1 P6[a]
ACGTACAATTGATATGGTTGAGAACGGTTCGGCATTG	ATATG	Cbf1 P7[a]
ACGTACAATTGATGACGTTGAGAACGGTTCGGCATTG	ATGAC	Cbf1 P8[a]
ACGTACAATTGATGTAGTTGAGAACGGTTCGGCATTG	ATGTA	Cbf1 P9[a]
ACGTACAATTGCACGAGTTGAGAACGGTTCGGCATTG	CACGA	Cbf1 P10[a]
ACGTACAATTGCAGATTGTTGAGAACGGTTCGGCATTG	CAGATT	Met31 P1[a]
ACGTACAATTGCAGTGTGTTGAGAACGGTTCGGCATTG	CAGTGT	Met31 P2[a]
ACGTACAATTGCGACATGTTGAGAACGGTTCGGCATTG	CGACAT	Met31 P3[a]
ACGTACAATTGCGAGCTGTTGAGAACGGTTCGGCATTG	CGAGCT	Met31 P4[a]
ACGTACAATTGCGTAGTGTTGAGAACGGTTCGGCATTG	CGTAGT	Met31 P5[a]
ACGTACAATTGCGTCTTGTTGAGAACGGTTCGGCATTG	CGTCTT	Met31 P6[a]
ACGTACAATTGCGTGATGTTGAGAACGGTTCGGCATTG	CGTGAT	Met31 P7[a]

Sequence	Barcode	Use
ACGTACAATTGCTCTATGTTGAGAACGGTTCGGCATTG	CTCTAT	Met31 P8[a]
ACGTACAATTGCTGAGTGTTGAGAACGGTTCGGCATTG	CTGAGT	Met31 P9[a]
ACGTACAATTGCTGTCTGTTGAGAACGGTTCGGCATTG	CTGTCT	Met31 P10[a]
ACGTACAATTGGACATACGTTGAGAACGGTTCGGCATTG	GACATAC	Nrg1 P1[a]
ACGTACAATTGGACTGACGTTGAGAACGGTTCGGCATTG	GACTGAC	Nrg1 P2[a]
ACGTACAATTGGATCGACGTTGAGAACGGTTCGGCATTG	GATCGAC	Nrg1 P3[a]
ACGTACAATTGGATGTACGTTGAGAACGGTTCGGCATTG	GATGTAC	Nrg1 P4[a]
ACGTACAATTGGCACAACGTTGAGAACGGTTCGGCATTG	GCACAAC	Nrg1 P5[a]
ACGTACAATTGGCGACACGTTGAGAACGGTTCGGCATTG	GCGACAC	Nrg1 P6[a]
ACGTACAATTGGCGCGACGTTGAGAACGGTTCGGCATTG	GCGCGAC	Nrg1 P7[a]
ACGTACAATTGGCGTAACGTTGAGAACGGTTCGGCATTG	GCGTAAC	Nrg1 P8[a]
ACGTACAATTGGTACGACGTTGAGAACGGTTCGGCATTG	GTACGAC	Nrg1 P9[a]
ACGTACAATTGGTGATACGTTGAGAACGGTTCGGCATTG	GTGATAC	Nrg1 P10[a]
ACGTACAATTGGTGTGCACGTTGAGAACGGTTCGGCATTG	GTGTGCAC	Gcn4 P1[a]
ACGTACAATTGTAGCGCACGTTGAGAACGGTTCGGCATTG	TAGCGCAC	Gcn4 P2[a]
ACGTACAATTGTATAGCACGTTGAGAACGGTTCGGCATTG	TATAGCAC	Gcn4 P3[a]
ACGTACAATTGTATCTCACGTTGAGAACGGTTCGGCATTG	TATCTCAC	Gcn4 P4[a]
ACGTACAATTGTATGACACGTTGAGAACGGTTCGGCATTG	TATGACAC	Gcn4 P5[a]
ACGTACAATTGTCGAGCACGTTGAGAACGGTTCGGCATTG	TCGAGCAC	Gcn4 P6[a]
ACGTACAATTGTCTATCACGTTGAGAACGGTTCGGCATTG	TCTATCAC	Gcn4 P7[a]
ACGTACAATTGTCTGCCACGTTGAGAACGGTTCGGCATTG	TCTGCCAC	Gcn4 P8[a]
ACGTACAATTGTGACGCACGTTGAGAACGGTTCGGCATTG	TGACGCAC	Gcn4 P9[a]
ACGTACAATTGTGAGTCACGTTGAGAACGGTTCGGCATTG	TGAGTCAC	Gcn4 P10[a]

[a]W = Well and P = Plate, so WA1 is Well A1 and P1 is Plate 1.
[b]ChIPed sample barcode: C=Cbf1-tagged; G=Gcn4-tagged; A=AAS; L=Glucose; K=Input; P=IP; 1-3=Sample replicate. So CLP1=Cbf1-tagged IP in glucose, replicate 1.

Synthetic promoters were amplified by using one well-specific PCR primer and one plate-specific PCR primer. Custom adapters were ligated on to the products and sequenced on an Illumina MiSEQ machine. A subset of well-specific primers were reused to barcode CHiP samples for multiplexing on an Illumina HiSEQ 2000. All primers are listed in 5'-3' order.

Table 3.6: Oligonucleotides used for strain manipulation, validation, PCR, and sequencing

Name	Sequence	Purpose
RZ84	5’-GATCGTATCACGTGCTTTAC-3’	Cbf1 site, forward
RZ85	3’-CATAGTGCACGAAATGCTAG-5’	Cbf1 site, reverse
RZ86	5’-GATCGTAATGACTCATTTAC-3’	Gcn4 site, forward
RZ87	3’-CATTACTGAGTAAATGCTAG-5’	Gcn4 site, reverse
RZ88	5’-GATCGTAGCCACAGTTTTAC-3’	Met 31/32 site, forward
RZ89	3’-CATCGGTGTCAAAATGCTAG-5’	Met 31/32 site, reverse
RZ90	5’-GATCGTATGAGGACCCTTAC-3’	Nrg1 site, forward
RZ91	3’-CATACTCCTGGGAATGCTAG-5’	Nrg1 site, reverse
RZ92	5’-CATTCTTACCCACTCCTGTTCTAG -3’	Gcn4 Avi-tagging check, upstream PCR primer
RZ93	5’-CGCGTCTGACTTCTAATCAGAAG-3’	Gcn4 Avi-tagging check, downstream PCR primer
RZ94	5’-CCGATGAAGCAAACATCGAAAAG -3’	Cbf1 Avi-tagging check, upstream PCR primer
RZ95	5’-TCCGTCCCGTCCTCTTTTAC -3’	Cbf1 Avi-tagging check, downstream PCR primer
RZ96	5-’CCGGAAAATATGGCTAGAGGTC -3’	Met31 Avi-tagging check, upstream PCR primer
RZ97	5’-GTACGTCACCACTTTGTGCG -3’	Met31 Avi-tagging check, downstream PCR primer

Name	Sequence	Purpose
RZ98	5’-CGGAAGCAAAGAACAGATCCA -3’	Nrg1 Avi-tagging check, upstream PCR primer
RZ99	5’-CCAGACATGATCTTAAGCGGAAG -3’	Nrg1 Avi-tagging check, downstream PCR primer
RZ127	5’-GACATGATAATTGCTTGCAACACTATAGAACACATTTGAAAAAGGGACAAGGGATCGAGCAGAAGCTGAT -3’	myc-C-Avi tagging primer, Nrg1, upstream
RZ128	5’-AGTGCGGAATAGTAGTACTGCTAATGAGAAAAACACGGGTATACCGTCAACTGCAGGTCGACAACCCTTAAT -3’	myc-C-Avi tagging primer, Nrg1, downstream
RZ129	5’-AACAAGAGAACGAAAGAAAAAGCACTAGGAGCGATAATCCACATGAGGCTGGGATCGAGCAGAAGCTGAT -3’	myc-C-Avi tagging primer, Cbf1, upstream
RZ130	5’-GTGCTATGGGGCAGAGACGCAGATACATAGGGAGACTCGAAATACATTTACTGCAGGTCGACAACCCTTAAT -3’	myc-C-Avi tagging primer, Cbf1, downstream
RZ131	5’-CTTTTTTGTGCCTTTGTTACGTCTATATTCTATTGAAACTGGAGCTCGTTTTCGACACTGG -3’	Insert PCORE into lys-2, upstream PCR primer
RZ132	5’-TATTATATATTATTCTCGGAGTTTTTAAGTGACATCACCCTCCTTACCATTAAGTTGATC -3’	Insert PCORE into lys-2, downstream PCR primer
RZ133	5’-CTTTTTTGTGCCTTTCTTACGTCTATATTCATTGAAACTGGACTGGGTCATGGCTGCG -3’	Insert BirA into lys-2, upstream PCR primer
RZ134	5’-TATTATATATTATTCTCGGAGTTTTTAAGTGACATCACCCAAGCTTGCAAATTAAAGCCTTCGAG -3’	Insert BirA into lys-2, downstream PCR primer

Name	Sequence	Purpose
RZ135	5’-GCTCATCAAGGATGCGATAAAGAATGGTACCGGCCTGTTGGGGATCGAGCAGAAGCTGAT -3’	myc-C-Avi tagging primer, Met31, upstream
RZ136	5’-ATTCTACTTATCTCAATGGCTAAAGTATATATCTATCTATCTGCAGGTCGACAACCCTTAAT -3’	myc-C-Avi tagging primer, Met31, downstream
RZ137	5’-AAATGAGGTTGCCAGATTAAAGAAATTAGTTGGCGAACGCGGGATCGAGCAGAAGCTGAT -3’	myc-C-Avi tagging primer, Gcn4, upstream
RZ138	5’-GCGTGGTGTAAAATTCTACTTAAGAAAATTGGCATAAAAACTGCAGGTCGACAACCCTTAAT -3’	myc-C-Avi tagging primer, Gcn4, downstream
RZ143	5’-CGACCTCATGCTATACCTGAGAAAG -3’	myc-C-avi integration check PCR primer, upstream, internal to tag
RZ144	5’-TGGGGATGTATGGGCTAAATGTAC -3’	myc-C-Avi integration check PCR primer, downstream, internal to Kan
RZ147	5’-GCAGTTGCTTTCTCCTATGGGAAG -3’	PCORE and BirA integration check PCR primer, upstream
RZ148	5’-GAATTGGTCAGTATCGACCTGTGAA -3’	PCORE and Bira integration check PCR primer, downstream
RZ149	5’-GTTAGAAGAAAAGAGTCGGGATCTCTG -3’	BirA integration check PCR primer, upstream, internal to BirA

Name	Sequence	Purpose
RZ150	5’-CTGTACAGACGCGTGTACGC -3’	BirA integration check PCR primer, downstream, internal to BirA
RZ151	5’-TTAAGTCCGGGGATCCCCAG -3’	Universal myc-C-Avi-tag sequencing primer, internal to Avi tag.
RZ158	5’-GGGAGGAGTCATGGCAAATA -3’	Cbf1 ChIP check qPCR primer: ADE765.
RZ159	5’-CGTATACGGTGACGACGAGA -3’	5’ PRIMER AROUND ADE756 SET 2 SET 4
RZ169	5’-TAGGGGCTTAGCATCCACAC -3’	SUC2 qPCR Primer
RZ170	5’-TGGATACCTTCGACAGCTCA -3’	SUC2 qPCR Primer
RZ177	5’-CCCCTAAACATTCAGATTGTAAAC -3’	Gcn4 ChIP check qPCR primer (YJR109C)
RZ178	5’-TCTCGATGCTTACTCAAGGTG -3’	Gcn4 ChiP check qPCR primer (YJR109C)
RZ183	5’-GCCGCCACAGAAAACTTAC -3’	Met31 ChIP check qPCR primer (YNL278W)
RZ184	5’-GAGCTATGGGCAATTGTACG -3’	Met31 ChiP check qPCR primer (YNL278W)
RZ193	5’-CCGGAAAAGAAGGGAAAAAT -3’	Nrg1 ChIP check qPCR primer (YDR043C)

Name	Sequence	Purpose
RZ194	5’-CCTGCAGCCAGACTGTAGAA -3’	Nrg1 ChIP check qPCR primer (YDR043C)
RZ226	5’-CCCTCGTTCAATTGCTCACCTCGAC -3’	Custom read 1 sequencing primer for sequencing synthetic promoters.
RZ227	5’-GCTCCCCATTTCACGAATTG-3’	Custom read 2 sequencing primer for synthetic promoters
RZ230	/5Phos/ TCGAGGTGAGCAATTGAACGAGGGGTGTAGATCTC GGTGGTCGCCGTATCATT -3’	Read 1 flow cell adapter and sequencing primer
RZ231	/5Phos/ AATTCGTGAAATGGGGAGCATCTCGTATGCCGTCTT CTGCTTG -3’	Read 2 flow cell adapter and sequencing primer
RZ232	5’- AATGATACGGCGACCACCGAGATCTACACCCCTCG TTCAATTGCTCACC -3’	Read 1 flow cell adapter and sequencing primer (reverse complement)
RZ233	5’- CAAGCAGAAGACGGCATACGAGATGCTCCCCATTT CACG -3’	Read 2 flow cell adapter and sequencing primer (reverse complement)
RZ257.1	5’- CAAGCAGAAGACGGCATACGAGATCGGTCTCGGC ATTCCTGCTGAACCGCTCTTCCGATCTGTTGAGAAC GGTTCGGCATTG -3’	Downstream pcr primer for synthetic promoter amplification for sequencing post-ChIP (1/4)

Name	Sequence	Purpose
RZ257.2	5’-CAAGCAGAAGACGGCATACGACGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTCGTTGAGAACGGTTCGGCATTG -3’	Downstream pcr primer for synthetic promoter amplification for sequencing post-ChIP (2/4)
RZ257.3	5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTTAGTTGAGAACGGTTCGGCATTG -3’	Downstream pcr primer for synthetic promoter amplification for sequencing post-ChIP (3/4)
RZ257.4	5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCTACAGTTGAGAACGGTTCGGCATTG -3’	Downstream pcr primer for synthetic promoter amplification for sequencing post-ChIP (4/4)
RZ259	5’- TGTAATCGTTCTTCCACACGGATC -3’	qPCR Primer for library concentration check, post prep.
RZ260	5’- TTCCTGCTGAACCGCTCTTC-3’	qPCR Primer for library concentration check, post prep.

Table 3.7: List of all promoters and condition-specific expression and occupancy values

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
c	4.48	1.96	4.48	1.92	Cbf1
C	4.87	2.15	4.51	2.11	Cbf1
cC	6.50	4.71	NA	4.27	Cbf1
cCcNC	6.84	8.60	NA	7.40	Cbf1
cCnGGn	3.34	3.04	5.89	4.04	Cbf1
cCNm	6.19	4.07	NA	4.09	Cbf1
cg	2.82	1.69	3.16	1.96	Cbf1
Cg	6.28	1.85	NA	2.14	Cbf1
cgg	2.06	0.03	5.04	0.03	Cbf1
cgGm	3.21	1.28	NA	1.63	Cbf1
CGGn	2.03	1.34	NA	1.70	Cbf1
cgm	6.62	1.57	NA	1.87	Cbf1
CgM	3.77	0.05	NA	0.05	Cbf1
cGN	3.96	1.30	NA	1.65	Cbf1
CgnM	4.04	NA	NA	NA	Cbf1
cGnmM	5.47	1.58	NA	2.14	Cbf1
cgnN	1.80	0.85	4.34	1.38	Cbf1
cm	4.13	3.50	NA	2.68	Cbf1
CMcC	2.70	0.05	5.18	0.05	Cbf1
cmG	6.48	1.85	NA	2.18	Cbf1
cMgN	1.00	0.04	NA	0.05	Cbf1
cmM	6.05	2.28	NA	2.55	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
cmn	3.77	1.64	5.20	2.01	Cbf1
CMn	4.40	2.00	5.62	2.38	Cbf1
CMNgC	2.22	0.05	NA	0.05	Cbf1
CMnmg	5.88	1.78	NA	2.31	Cbf1
CMNMn	3.80	1.84	4.51	2.30	Cbf1
cN	3.94	1.66	4.20	1.91	Cbf1
Cn	2.91	1.97	3.13	2.08	Cbf1
CN	3.96	1.89	4.17	2.09	Cbf1
cNCCM	4.28	0.04	NA	0.04	Cbf1
Cng	4.68	1.48	NA	1.83	Cbf1
cngCnG	4.77	3.52	NA	3.63	Cbf1
cnGN	1.66	0.70	3.49	1.17	Cbf1
CNGNmn	0.89	0.72	1.77	1.01	Cbf1
cnm	6.83	NA	NA	NA	Cbf1
cNM	4.00	1.68	NA	2.09	Cbf1
cnNm	2.45	1.06	3.54	1.34	Cbf1
g	1.73	0.10	3.16	0.10	Cbf1
G	1.60	0.10	3.06	0.09	Cbf1
gc	3.00	0.05	6.67	0.05	Cbf1
Gc	6.25	2.08	NA	2.26	Cbf1
GcgG	3.60	1.58	NA	1.61	Cbf1
GCgggn	2.41	1.78	NA	1.52	Cbf1
gcmm	7.61	1.96	NA	2.30	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
gCN	4.57	1.76	6.52	1.98	Cbf1
Gcn	4.41	1.61	NA	1.82	Cbf1
gg	2.80	0.99	NA	0.79	Cbf1
Gg	2.72	0.29	3.88	0.66	Cbf1
GG	2.23	0.10	4.88	0.10	Cbf1
gGCm	4.91	NA	NA	NA	Cbf1
gGCn	5.02	1.89	8.19	2.27	Cbf1
gGm	2.50	0.08	4.75	0.09	Cbf1
gGM	2.04	0.05	NA	0.05	Cbf1
ggMMM	5.15	0.14	NA	0.15	Cbf1
GGmn	1.57	0.09	2.55	0.08	Cbf1
ggNm	1.17	0.04	3.82	0.03	Cbf1
gGnm	1.12	0.10	2.03	0.10	Cbf1
GGNmc	5.27	2.59	NA	3.10	Cbf1
ggnN	0.72	0.07	1.80	0.08	Cbf1
gM	4.55	0.07	NA	0.07	Cbf1
Gm	2.03	0.09	3.39	0.08	Cbf1
GM	2.42	0.45	4.19	0.23	Cbf1
GMC	8.63	2.14	NA	2.35	Cbf1
gMG	2.32	0.33	5.23	0.86	Cbf1
GmGg	2.57	NA	4.88	NA	Cbf1
GmGggN	2.74	NA	4.95	NA	Cbf1
Gmgm	1.95	NA	NA	NA	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
GmgmGG	2.33	NA	NA	NA	Cbf1
gMgn	1.66	0.08	3.32	0.09	Cbf1
gMGn	1.59	0.10	3.20	0.13	Cbf1
Gmgnm	1.39	0.13	2.80	0.08	Cbf1
gmMc	5.72	NA	NA	NA	Cbf1
GmmgC	7.54	3.92	NA	4.15	Cbf1
gn	3.37	1.15	NA	1.65	Cbf1
GN	1.26	0.08	2.47	0.09	Cbf1
gNC	6.22	1.90	NA	2.04	Cbf1
gng	2.04	0.07	3.04	0.09	Cbf1
gNGGNn	1.04	0.10	1.96	0.07	Cbf1
gNM	1.59	0.08	3.98	0.06	Cbf1
GNm	1.20	0.08	2.34	0.08	Cbf1
GNM	1.74	0.09	2.72	0.08	Cbf1
gnmG	6.21	NA	NA	NA	Cbf1
GNmN	3.98	NA	NA	NA	Cbf1
GnnCn	2.79	1.57	3.42	1.87	Cbf1
GnnMCn	4.18	1.95	NA	2.50	Cbf1
m	1.04	0.09	1.30	0.06	Cbf1
M	1.25	0.13	1.42	0.12	Cbf1
mc	5.38	2.28	NA	2.61	Cbf1
mC	6.72	2.27	3.85	2.52	Cbf1
MCG	5.94	2.04	NA	2.32	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
Mcggg	3.90	1.43	NA	1.82	Cbf1
mcm	8.11	2.08	NA	2.31	Cbf1
mCm	5.66	2.36	NA	2.64	Cbf1
MCMgMg	5.41	1.93	NA	2.18	Cbf1
mcn	4.14	1.83	NA	2.32	Cbf1
MCN	5.11	1.88	NA	2.12	Cbf1
MCnCn	1.91	0.04	2.59	0.04	Cbf1
MCnn	3.12	1.68	4.10	2.13	Cbf1
MCNN	1.18	0.05	1.92	0.03	Cbf1
mG	6.85	1.96	NA	2.14	Cbf1
Mg	2.25	0.27	3.86	0.13	Cbf1
MG	2.14	0.16	3.89	0.13	Cbf1
mgc	6.52	2.22	NA	2.44	Cbf1
Mgc	5.96	2.30	NA	2.42	Cbf1
MGCG	4.58	3.22	NA	3.27	Cbf1
MgcM	4.09	0.15	NA	0.65	Cbf1
MGCn	5.74	1.69	NA	2.04	Cbf1
MgG	2.13	0.13	NA	0.10	Cbf1
MGg	2.23	0.10	4.27	0.11	Cbf1
MggCNM	4.83	NA	NA	NA	Cbf1
mggN	1.43	0.09	2.81	0.09	Cbf1
MGGNm	2.14	NA	NA	NA	Cbf1
MGm	3.15	NA	5.13	NA	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
mGmC	3.54	0.06	NA	0.05	Cbf1
MGmCc	2.86	0.08	NA	0.08	Cbf1
MGmM	5.39	0.13	NA	0.11	Cbf1
mgMNc	2.94	0.06	NA	0.01	Cbf1
mGn	0.88	0.04	3.41	0.03	Cbf1
mGNc	5.60	2.20	NA	2.37	Cbf1
MGnGC	5.61	1.99	NA	2.27	Cbf1
mGNNG	1.17	0.09	2.90	0.09	Cbf1
mm	3.70	0.09	4.09	0.08	Cbf1
mM	3.34	0.20	4.94	NA	Cbf1
MM	3.65	0.10	4.51	0.09	Cbf1
MMgc	4.43	0.06	NA	0.04	Cbf1
MMGN	1.56	0.09	3.79	0.10	Cbf1
mmm	4.32	0.10	NA	0.10	Cbf1
mmn	7.73	2.35	NA	2.70	Cbf1
MmnCMN	7.32	2.06	NA	2.05	Cbf1
MmNm	2.78	0.06	4.63	NA	Cbf1
mn	0.63	0.08	0.87	0.07	Cbf1
Mn	0.76	0.09	0.87	0.07	Cbf1
MN	0.95	0.30	1.25	0.21	Cbf1
MNcg	3.92	4.41	NA	5.25	Cbf1
mnCn	1.47	0.03	2.83	0.04	Cbf1
MnG	0.80	0.06	1.33	0.08	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
MngG	1.20	0.04	5.00	0.03	Cbf1
mNGmm	3.31	0.09	NA	0.11	Cbf1
MNm	1.60	NA	2.22	NA	Cbf1
MnMG	1.97	0.49	4.33	0.19	Cbf1
MNMG	1.32	0.12	NA	0.10	Cbf1
mnMGG	2.20	0.10	4.10	0.09	Cbf1
mNnc	1.48	0.04	2.18	0.03	Cbf1
MNnc	4.00	2.11	4.55	2.34	Cbf1
mnng	0.79	0.07	1.36	0.07	Cbf1
MnNgnG	1.02	0.08	1.61	0.09	Cbf1
n	0.61	0.11	0.59	0.06	Cbf1
N	0.84	0.10	0.89	0.07	Cbf1
nc	4.39	2.01	3.98	2.22	Cbf1
nC	4.63	1.90	3.96	2.14	Cbf1
Nc	2.36	0.19	NA	0.31	Cbf1
NC	1.96	0.47	2.13	0.63	Cbf1
NcG	5.31	1.81	NA	2.08	Cbf1
NCg	5.77	1.75	NA	1.92	Cbf1
NcGnGN	1.43	0.69	4.07	1.11	Cbf1
ncGNNm	1.36	0.72	3.43	0.99	Cbf1
ncM	4.80	1.88	NA	2.30	Cbf1
Ncm	5.31	1.81	7.46	2.17	Cbf1
NCmn	1.53	NA	4.29	NA	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
nCmnGm	5.59	1.14	NA	2.16	Cbf1
ncN	3.29	5.89	3.67	9.60	Cbf1
NCngG	2.19	1.13	4.98	1.51	Cbf1
nCnGm	3.22	1.08	6.55	1.77	Cbf1
nCNGNG	0.86	0.03	4.05	0.04	Cbf1
nCnm	3.98	1.42	5.49	1.87	Cbf1
nCNMCn	6.45	3.89	NA	4.19	Cbf1
ng	1.12	0.08	2.79	0.07	Cbf1
nG	1.14	0.08	2.66	0.07	Cbf1
Ng	1.52	0.11	3.01	0.11	Cbf1
NG	1.21	0.16	2.63	0.21	Cbf1
ngc	5.78	2.02	8.40	2.15	Cbf1
NGc	5.42	2.18	NA	2.50	Cbf1
NgCgg	5.14	NA	NA	NA	Cbf1
nGcM	2.89	0.05	NA	0.05	Cbf1
nGcNm	4.86	1.20	NA	2.00	Cbf1
ngg	1.34	0.05	4.43	0.03	Cbf1
Ngg	1.52	0.16	2.64	0.18	Cbf1
nGgCNn	1.13	0.07	3.39	0.03	Cbf1
ngGm	6.60	0.09	NA	0.07	Cbf1
ngM	1.56	0.09	3.43	0.08	Cbf1
nGm	1.38	0.08	2.83	0.09	Cbf1
nGMGMC	7.00	2.46	NA	2.88	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
NGn	0.69	NA	1.30	NA	Cbf1
NGnc	7.09	NA	NA	NA	Cbf1
NgNGG	1.26	0.08	4.15	0.08	Cbf1
nm	0.83	0.11	1.10	0.11	Cbf1
Nm	0.96	0.08	1.20	0.08	Cbf1
NM	1.51	0.14	NA	0.12	Cbf1
nMC	6.31	2.15	8.06	2.36	Cbf1
nMcm	6.16	2.25	NA	2.71	Cbf1
nmg	1.37	0.09	2.44	0.11	Cbf1
nMG	1.23	0.03	5.17	0.03	Cbf1
nMgm	1.83	0.09	2.99	0.09	Cbf1
nmGN	0.92	0.16	1.46	0.50	Cbf1
nmM	4.04	0.06	NA	0.04	Cbf1
nMm	2.57	0.09	3.54	0.09	Cbf1
NMM	5.12	1.62	NA	1.70	Cbf1
NMMgn	4.82	NA	NA	NA	Cbf1
nmn	0.49	0.07	0.58	0.07	Cbf1
Nmn	8.41	NA	NA	NA	Cbf1
nmnM	1.38	0.11	2.08	0.06	Cbf1
nN	0.63	0.42	0.69	0.38	Cbf1
NN	0.70	0.08	0.98	0.07	Cbf1
nnC	3.98	1.95	4.41	2.14	Cbf1
nNC	9.86	2.60	6.35	2.80	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
NNc	4.12	2.11	3.88	2.47	Cbf1
NNC	4.29	2.03	3.64	2.15	Cbf1
NNcc	6.53	4.81	NA	4.59	Cbf1
NncgN	3.22	0.98	5.88	1.51	Cbf1
NNCN	3.42	1.71	3.25	2.01	Cbf1
NNG	1.21	0.13	2.46	0.10	Cbf1
Nngc	4.16	2.13	NA	2.40	Cbf1
nnGGg	1.61	0.13	4.91	0.15	Cbf1
NnM	0.86	0.08	1.08	0.06	Cbf1
NNm	0.71	0.07	0.86	0.07	Cbf1
NNn	0.38	NA	0.39	0.31	Cbf1
Basal	NA	0.17	0.98	0.07	Cbf1
cc	NA	4.50	NA	4.33	Cbf1
CC	NA	4.89	NA	4.51	Cbf1
Cccgcn	NA	8.11	NA	7.22	Cbf1
CccMM	NA	6.04	NA	5.92	Cbf1
ccm	NA	4.28	NA	4.41	Cbf1
ccNgc	NA	6.63	NA	6.51	Cbf1
ccnGM	NA	3.85	1.38	3.63	Cbf1
cGcGmm	NA	4.65	NA	4.56	Cbf1
cgcM	NA	4.40	NA	4.21	Cbf1
cGmCCg	NA	6.25	NA	5.44	Cbf1
cM	NA	5.70	NA	6.34	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
CMCGCM	NA	7.28	NA	8.04	Cbf1
CMCm	NA	4.79	NA	4.82	Cbf1
CMCn	NA	4.47	NA	4.52	Cbf1
cMgM	NA	2.10	NA	2.47	Cbf1
CMMm	NA	0.09	NA	0.05	Cbf1
cmNc	NA	4.51	NA	4.80	Cbf1
cnCC	NA	7.20	NA	9.34	Cbf1
cncg	NA	3.92	NA	4.17	Cbf1
cNcG	NA	4.17	NA	4.32	Cbf1
CNCN	NA	4.34	NA	4.39	Cbf1
CNm	NA	2.19	NA	2.20	Cbf1
GC	NA	5.65	NA	5.64	Cbf1
GcC	NA	4.38	NA	4.51	Cbf1
gcgcN	NA	4.11	NA	4.10	Cbf1
gCGGN	NA	0.94	NA	1.30	Cbf1
gcgmGn	NA	1.16	NA	1.62	Cbf1
GCM	NA	7.04	NA	6.61	Cbf1
GcMNc	NA	5.01	NA	4.85	Cbf1
gCnGc	NA	4.41	NA	4.55	Cbf1
gCNnn	NA	0.93	NA	1.42	Cbf1
ggCMc	NA	4.91	NA	4.62	Cbf1
gGMGN	NA	1.93	NA	3.63	Cbf1
gGnccm	NA	4.67	NA	4.29	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
GMCn	NA	2.17	NA	1.95	Cbf1
GmGcCN	NA	4.47	NA	4.64	Cbf1
gmN	NA	0.95	NA	2.07	Cbf1
GmN	NA	2.01	NA	2.03	Cbf1
GNCC	NA	6.89	NA	6.77	Cbf1
GNNc	NA	2.14	NA	2.23	Cbf1
mcC	NA	4.83	NA	4.94	Cbf1
MCcG	NA	3.39	NA	3.26	Cbf1
MCCG	NA	4.56	NA	4.08	Cbf1
MccN	NA	4.52	NA	4.34	Cbf1
McG	NA	1.87	NA	2.34	Cbf1
mCmC	NA	5.48	NA	5.53	Cbf1
MCMC	NA	4.74	NA	4.58	Cbf1
MCMccg	NA	6.75	NA	6.18	Cbf1
MCMMNc	NA	5.30	NA	5.23	Cbf1
MGc	NA	7.50	NA	7.33	Cbf1
MGCCcN	NA	6.50	NA	5.65	Cbf1
MGccnc	NA	6.80	NA	6.32	Cbf1
mgcmC	NA	5.07	NA	5.06	Cbf1
mGCN	NA	7.95	NA	7.21	Cbf1
mGggC	NA	2.09	NA	2.36	Cbf1
mGMn	NA	0.07	NA	0.11	Cbf1
MgN	NA	8.37	NA	5.80	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
MGnnNM	NA	0.09	NA	0.08	Cbf1
Mm	NA	0.46	NA	NA	Cbf1
MMgC	NA	0.07	NA	0.06	Cbf1
mmnmC	NA	2.58	NA	2.77	Cbf1
MnccMM	NA	4.79	NA	4.55	Cbf1
MNgMmn	NA	0.08	NA	0.07	Cbf1
mNm	NA	0.17	NA	0.16	Cbf1
mnmC	NA	2.76	NA	2.95	Cbf1
MNmmCN	NA	2.52	NA	2.44	Cbf1
ncc	NA	4.47	NA	4.43	Cbf1
nCCGnm	NA	4.15	NA	3.93	Cbf1
NCcmg	NA	4.26	NA	4.28	Cbf1
ncg	NA	1.38	NA	1.68	Cbf1
NCgGg	NA	1.55	NA	1.76	Cbf1
ncGgm	NA	1.10	NA	1.61	Cbf1
ncGmm	NA	1.83	NA	2.07	Cbf1
ngC	NA	2.56	NA	2.85	Cbf1
ngCc	NA	4.75	NA	4.60	Cbf1
Ngm	NA	0.06	NA	0.09	Cbf1
nGMC	NA	7.20	NA	8.63	Cbf1
NgNm	NA	3.94	NA	5.05	Cbf1
NMcc	NA	4.73	NA	4.83	Cbf1
nmCGC	NA	10.81	NA	10.28	Cbf1

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
Nmg	NA	4.27	NA	5.18	Cbf1
nmgG	NA	0.20	NA	0.14	Cbf1
NmmgMC	NA	2.64	NA	2.88	Cbf1
nMNMm	NA	0.10	NA	0.13	Cbf1
nn	NA	0.29	NA	0.10	Cbf1
Nn	NA	0.27	NA	NA	Cbf1
NNgCgM	NA	4.56	NA	4.41	Cbf1
NNGn	NA	0.10	NA	0.07	Cbf1
cCCC	NA	NA	NA	11.18	Cbf1
mGCM	NA	NA	NA	4.44	Cbf1
NccmN	NA	NA	NA	4.73	Cbf1
NGccmN	NA	NA	NA	6.80	Cbf1
c	2.18	0.86	1.94	0.15	Gcn4
C	2.15	0.85	1.78	0.20	Gcn4
cc	2.67	0.69	2.21	0.19	Gcn4
CC	2.30	1.05	2.19	0.19	Gcn4
Ccc	3.02	NA	2.01	NA	Gcn4
Cccgcn	2.72	NA	4.67	1.80	Gcn4
cCcNC	1.13	NA	NA	NA	Gcn4
CcG	3.07	NA	NA	3.11	Gcn4
CCG	2.20	NA	NA	10.36	Gcn4
ccgC	2.37	NA	NA	NA	Gcn4
Ccm	3.10	0.73	4.40	0.17	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
ccN	2.59	0.65	2.47	0.07	Gcn4
cCN	2.45	0.91	1.69	0.14	Gcn4
cCnGGn	0.59	NA	NA	4.11	Gcn4
cCNm	2.83	0.78	5.00	0.11	Gcn4
cg	1.05	NA	NA	NA	Gcn4
Cg	2.52	1.40	NA	2.27	Gcn4
Cgcm	2.78	1.80	NA	2.16	Gcn4
cGmCCg	2.47	2.44	NA	6.14	Gcn4
cGN	1.30	1.02	5.05	1.48	Gcn4
cGnmM	4.28	1.21	NA	2.29	Gcn4
cm	2.79	0.63	3.99	0.16	Gcn4
cM	0.79	3.20	5.60	6.27	Gcn4
CM	3.16	0.75	NA	0.39	Gcn4
CMCGCM	2.74	2.15	NA	3.13	Gcn4
CMCm	4.48	0.54	NA	0.13	Gcn4
CMCn	2.47	0.64	3.92	0.11	Gcn4
cmG	2.25	1.27	NA	2.39	Gcn4
cmM	5.79	0.70	NA	0.16	Gcn4
cmn	1.69	0.57	3.45	0.13	Gcn4
cmNc	2.39	0.71	NA	0.16	Gcn4
cmnmc	3.37	0.42	5.17	0.10	Gcn4
CMNn	3.58	NA	NA	NA	Gcn4
cn	1.12	0.66	NA	0.08	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
cN	2.45	0.67	3.03	0.13	Gcn4
CN	1.89	0.87	1.78	0.12	Gcn4
CnC	5.70	NA	NA	NA	Gcn4
cNcG	2.90	2.43	NA	2.21	Gcn4
Cncgm	2.65	NA	NA	1.94	Gcn4
cng	1.38	0.88	5.13	1.19	Gcn4
Cng	1.26	1.02	NA	1.62	Gcn4
CnG	1.26	1.04	5.21	1.43	Gcn4
cngCnG	1.22	0.85	NA	4.29	Gcn4
cnMc	3.12	0.43	3.40	0.12	Gcn4
CNMgg	1.05	1.08	NA	5.48	Gcn4
cnncMN	2.19	0.55	4.30	0.10	Gcn4
g	1.15	1.10	4.02	1.49	Gcn4
G	1.18	1.24	3.79	1.67	Gcn4
Gc	2.15	1.21	4.57	2.15	Gcn4
GC	2.74	2.64	NA	9.48	Gcn4
GcC	2.56	1.32	4.32	2.01	Gcn4
gcgCM	2.75	1.73	NA	6.28	Gcn4
gcgcN	1.97	1.16	NA	6.67	Gcn4
GcgG	1.25	4.48	NA	14.26	Gcn4
GCgggn	1.03	NA	NA	14.77	Gcn4
gCM	3.38	0.79	NA	2.15	Gcn4
Gcm	2.79	0.95	NA	2.14	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
GCm	2.49	0.75	NA	2.15	Gcn4
GCM	3.07	NA	NA	NA	Gcn4
Gcn	1.62	0.89	3.98	1.42	Gcn4
gCnGc	2.59	1.53	NA	5.09	Gcn4
gCNnn	0.37	0.70	1.29	0.50	Gcn4
Gg	0.93	NA	5.05	10.26	Gcn4
GG	0.89	1.35	5.43	5.37	Gcn4
gGCn	1.12	3.13	NA	7.13	Gcn4
GggnG	0.71	1.17	NA	11.69	Gcn4
gGm	1.68	2.32	NA	4.54	Gcn4
gGMGN	1.86	NA	NA	14.85	Gcn4
ggMMM	5.08	1.98	NA	10.28	Gcn4
GGmn	0.52	0.95	4.41	3.12	Gcn4
gM	1.71	0.88	5.07	1.29	Gcn4
Gm	1.70	1.17	5.53	1.12	Gcn4
GM	1.58	1.27	NA	1.49	Gcn4
gmc	2.81	1.06	NA	2.72	Gcn4
gmC	2.64	0.78	NA	2.47	Gcn4
GMC	2.62	1.13	NA	2.49	Gcn4
gMG	1.60	NA	NA	NA	Gcn4
GmGg	1.73	1.61	NA	7.10	Gcn4
Gmgm	1.94	NA	NA	NA	Gcn4
GmgmGg	1.20	NA	NA	NA	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
gMgn	0.81	0.86	4.20	2.44	Gcn4
gMGn	0.85	NA	4.19	3.62	Gcn4
GmM	5.21	NA	NA	NA	Gcn4
GMMg	2.24	NA	NA	NA	Gcn4
GN	0.95	0.90	3.01	0.64	Gcn4
gNC	2.20	0.72	3.87	1.04	Gcn4
gNM	1.13	0.76	3.43	0.65	Gcn4
GNM	1.24	0.66	3.82	0.54	Gcn4
GnMm	3.92	NA	NA	1.21	Gcn4
Gnn	0.34	0.66	0.88	0.27	Gcn4
GNNc	1.68	0.79	2.71	0.78	Gcn4
m	1.22	0.92	1.51	0.16	Gcn4
M	1.49	0.95	1.68	0.17	Gcn4
mc	2.44	0.61	3.64	0.14	Gcn4
McG	2.64	1.28	NA	2.06	Gcn4
mCGC	2.89	1.82	NA	2.18	Gcn4
Mcggg	1.01	1.68	NA	12.71	Gcn4
mcgm	3.76	1.46	NA	2.51	Gcn4
mCm	4.78	0.45	NA	0.17	Gcn4
Mcm	4.29	NA	NA	NA	Gcn4
mCmC	3.68	0.56	4.37	0.15	Gcn4
Mcmgn	1.89	NA	NA	2.51	Gcn4
MCMMNc	4.09	1.15	NA	0.13	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
MCN	2.09	0.63	2.77	0.13	Gcn4
McNcnm	2.88	NA	NA	NA	Gcn4
MCnn	1.83	0.76	1.81	0.16	Gcn4
mg	2.01	NA	NA	NA	Gcn4
mG	1.43	1.08	4.66	3.83	Gcn4
Mg	1.46	0.91	4.73	1.20	Gcn4
MG	1.71	1.06	5.28	1.07	Gcn4
mgc	2.59	1.55	NA	2.42	Gcn4
Mgc	2.95	1.37	NA	2.26	Gcn4
MGc	3.31	0.96	NA	2.22	Gcn4
mgCCnc	2.56	0.99	4.25	2.32	Gcn4
mgcmC	3.46	1.41	NA	3.55	Gcn4
MGg	1.06	1.03	NA	3.30	Gcn4
MggCc	1.78	NA	NA	NA	Gcn4
MGGcG	3.95	2.51	NA	19.41	Gcn4
MggCNM	1.95	NA	NA	NA	Gcn4
mggN	0.58	0.91	3.24	1.91	Gcn4
mgN	0.79	0.80	2.74	0.66	Gcn4
MGnGC	1.68	1.26	NA	5.93	Gcn4
mm	4.66	0.75	NA	0.18	Gcn4
MM	4.52	1.06	NA	0.16	Gcn4
mMccCG	2.82	NA	NA	2.78	Gcn4
MMGg	0.68	1.85	NA	10.92	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
MMgmN	2.44	0.99	NA	1.40	Gcn4
mmn	2.80	0.62	5.00	0.12	Gcn4
MMNcg	1.18	NA	NA	NA	Gcn4
mmnmC	4.58	1.01	NA	0.20	Gcn4
Mn	0.67	0.72	1.04	0.14	Gcn4
MN	1.07	0.78	1.29	0.20	Gcn4
mNC	2.20	0.82	3.02	0.36	Gcn4
Mng	1.67	NA	NA	4.52	Gcn4
MNm	2.16	NA	3.44	NA	Gcn4
mNmC	3.17	0.64	NA	0.09	Gcn4
Mnmg	1.13	0.45	NA	0.65	Gcn4
MnMG	1.36	0.77	NA	1.08	Gcn4
MNMG	1.57	0.82	NA	1.01	Gcn4
mnMGG	0.75	NA	NA	4.43	Gcn4
MNnc	1.87	0.81	1.68	0.15	Gcn4
mnng	1.62	0.73	1.63	0.28	Gcn4
n	0.55	0.82	0.63	0.15	Gcn4
N	0.83	0.89	0.89	0.18	Gcn4
nc	2.02	0.67	1.79	0.14	Gcn4
nC	2.02	0.66	1.70	0.14	Gcn4
Nc	2.04	0.54	1.67	0.14	Gcn4
ncc	2.61	0.79	2.33	0.12	Gcn4
nCCGnm	1.84	1.41	NA	2.45	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
ncg	1.33	NA	4.22	NA	Gcn4
NcG	2.30	1.26	NA	1.54	Gcn4
ncGgm	1.55	1.69	NA	4.72	Gcn4
nCGm	2.89	1.16	NA	1.78	Gcn4
NcGm	2.51	0.89	NA	1.60	Gcn4
ncGmm	5.22	1.58	NA	2.16	Gcn4
ncm	2.33	0.70	3.11	0.13	Gcn4
Ncm	2.61	0.67	4.48	0.23	Gcn4
ncmN	2.00	NA	NA	NA	Gcn4
nCmnGG	0.76	0.84	NA	6.15	Gcn4
ncN	1.58	NA	1.46	NA	Gcn4
nCnGm	1.25	1.03	NA	1.98	Gcn4
nCnm	1.85	0.79	3.00	0.14	Gcn4
nCNMCn	1.92	0.54	2.74	0.18	Gcn4
ng	0.56	0.86	2.48	1.08	Gcn4
Ng	0.71	1.34	3.82	1.30	Gcn4
NG	1.00	0.81	3.08	1.22	Gcn4
ngc	2.01	1.30	4.68	1.43	Gcn4
NGc	1.90	1.13	5.08	1.75	Gcn4
NGC	2.21	0.99	4.89	1.42	Gcn4
ngCMGC	2.70	1.49	NA	5.86	Gcn4
NGcNn	0.76	0.49	1.53	0.71	Gcn4
Ngg	0.76	0.91	3.46	2.46	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
NGg	0.63	1.04	3.96	2.77	Gcn4
NggGc	1.27	1.93	NA	11.09	Gcn4
NgGMCm	2.69	2.99	NA	10.42	Gcn4
ngM	1.13	0.93	3.94	0.96	Gcn4
nGm	1.32	1.02	3.85	0.73	Gcn4
nGM	1.15	0.93	4.28	0.66	Gcn4
ngMc	2.35	1.13	NA	2.07	Gcn4
nGMC	3.00	NA	NA	7.94	Gcn4
NgMG	0.98	0.97	NA	3.52	Gcn4
ngMm	3.65	1.00	NA	1.28	Gcn4
ngMMn	1.92	0.60	NA	1.20	Gcn4
NgN	0.50	0.61	1.76	0.37	Gcn4
NgNGG	0.55	0.71	4.33	4.32	Gcn4
nm	0.87	0.97	1.16	0.32	Gcn4
Nm	1.20	0.83	1.37	0.22	Gcn4
NM	1.32	0.69	1.57	0.19	Gcn4
nMC	2.47	0.81	3.75	0.16	Gcn4
NMccNc	2.69	0.44	2.62	0.11	Gcn4
nmCNN	1.10	NA	2.08	NA	Gcn4
nmg	0.72	0.80	2.79	0.66	Gcn4
Nmg	1.03	1.18	3.51	1.33	Gcn4
nmGN	0.52	NA	2.12	NA	Gcn4
nmGng	0.99	0.64	3.86	1.99	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
nMm	3.75	0.85	3.70	0.11	Gcn4
Nmn	2.74	NA	NA	NA	Gcn4
nmnM	1.79	0.61	3.26	0.19	Gcn4
nn	0.36	0.77	2.26	0.14	Gcn4
nN	0.62	NA	0.68	NA	Gcn4
Nn	0.52	NA	0.68	NA	Gcn4
NN	1.65	0.87	0.98	0.24	Gcn4
NNccc	2.73	0.52	2.47	0.09	Gcn4
NNCN	1.65	0.68	1.66	0.14	Gcn4
NNG	0.70	1.03	2.39	0.78	Gcn4
Nngc	1.63	1.41	4.85	1.35	Gcn4
nnGg	1.52	0.86	2.01	1.30	Gcn4
NNGn	0.34	0.50	1.37	0.27	Gcn4
Nnm	0.75	0.90	0.84	0.17	Gcn4
NnmG	0.63	0.59	3.10	0.98	Gcn4
nNN	0.65	NA	0.67	NA	Gcn4
NNn	0.35	NA	0.43	NA	Gcn4
cC	NA	0.71	NA	0.11	Gcn4
ccMmM	NA	0.83	NA	0.16	Gcn4
ccNgc	NA	2.11	NA	8.01	Gcn4
CMM	NA	0.65	NA	0.25	Gcn4
CNm	NA	0.97	NA	0.16	Gcn4
CNMM	NA	0.53	NA	0.09	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
gCGGNN	NA	2.19	NA	7.08	Gcn4
gcgmGn	NA	1.49	NA	11.32	Gcn4
GGNmc	NA	1.07	NA	9.42	Gcn4
gMcNMM	NA	1.41	NA	3.48	Gcn4
GMGmn	NA	0.88	NA	2.89	Gcn4
gMM	NA	1.07	NA	2.64	Gcn4
gn	NA	1.56	NA	10.83	Gcn4
McNg	NA	1.08	NA	1.57	Gcn4
MGmM	NA	1.16	NA	2.46	Gcn4
MNgMmn	NA	1.06	NA	1.59	Gcn4
NCcmG	NA	1.99	NA	3.06	Gcn4
nG	NA	0.93	NA	1.35	Gcn4
NGCGNG	NA	1.52	NA	8.35	Gcn4
nGMGMC	NA	1.08	NA	6.40	Gcn4
nGMnCM	NA	0.68	NA	1.79	Gcn4
nGnm	NA	1.54	NA	6.41	Gcn4
NmmgMC	NA	1.88	NA	4.83	Gcn4
NNm	NA	0.75	0.45	0.13	Gcn4
Cn	NA	NA	NA	10.50	Gcn4
ggn	NA	NA	NA	15.99	Gcn4
GNgcCc	NA	NA	NA	8.89	Gcn4
gNGGNn	NA	NA	NA	3.54	Gcn4
mgmMNm	NA	NA	NA	8.88	Gcn4

Promoter	Expression	Glucose Expression	Glucose Occupancy	AAS Expression	AAS Occupancy
mmCN	NA	NA	NA	0.44	Gcn4
MmnNC	NA	NA	NA	0.09	Gcn4
MNcg	NA	NA	NA	7.37	Gcn4
mNm	NA	NA	NA	3.12	Gcn4
NcMM	NA	NA	NA	3.29	Gcn4
nMNMm	NA	NA	NA	2.66	Gcn4
NNmCNm	NA	NA	NA	0.08	Gcn4

Synthetic promoters were constructed and expression and occupancy values obtained as detailed in Methods. For promoters, C=Cbf1, fwd; c=Cbf1, rev; G=Gcn4, fwd; g=Gcn4, rev; M=Met31/Met32, fwd; m=Met31/Met32, rev; N=Nrg1, fwd; n=Nrg1, rev, where “fwd” and “rev” refer to the corresponding sequences in Table 3.6. Promoter sequences are listed from most distal to most proximal to the TSS of YFP. In the “Glucose Expression”, “Glucose Occupancy”, “AAS Expression”, and “AAS Occupancy” columns, NA means the value is not available. For Expression columns, this is due to the expression being out of the dynamic range of the cytometer. For Occupancy columns, this is due to their being too few reads in the IN sample to reliably estimate the input distribution.

References

Albrecht, G, *et. al.* "Monitoring the Gcn4 Protein-mediated Response in the Yeast Saccharomyces cerevisiae." *The Journal of Biological Chemistry*. 1998. 273:12696.

Arndt, K. and Fink, G. R. "GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC 3' sequences." *Proc Natl Acad Sci U S A*. 1986. 83(22): 8516-20.

Berger, M.F., *et al.* "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities." *Nature Biotechnology.* 2006. 24(11):1429-1435.

Blaiseau, P.L. *et al*. "Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism." *Mol Cell Biol.* 1997. 17:3640-8.

Blaiseau, P.L. and Thomas, D. "Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA." *EMBO J.* 1998. 17(21):6327-36.

Brzovic, P.S., *et al*. "The acidic transcription activator Gcn4 binds the mediator subunit Gal11/ Med15 using a simple protein interface forming a fuzzy complex." *Molecular Cell*. 2011. 44(6): 942.

Buchler, N.E., Gerland, U. and Hwa, T. "On schemes of combinatorial transcription logic." *PNAS.* 2003. 100:5136.

Bussemaker, H.J., Li, H. and Siggia, E.D. "Regulatory element detection using correlation with expression." *Nat Genet.* 2001. 27:167-174.

Carrillo, E. *et al.* "Characterizing the roles of Met31 and Met32 in coordinating Met4-activated transcription in the absence of Met30." *Molecular Biology of the Cell*. 2012. 23(10):1928.

Cox, R. S. III, Surette, M. G. and Elowitz, M. B. "Programming gene expression with combinatorial promoters." *Mol. Syst. Biol.* 2007. 3:145.

Foat, B.C., Morozov, A.V. and Bussemaker, H.J. "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE." *Bioinformatics.* 2006. 22.

Gardner, C.A. and Barald, K.F. "The cellular environment controls the expression of engrailed-like protein in the cranial neuroepithelium of quail-chick chimeric embryos." *Development* (Cambridge, England). 1991. 113: 1037–1048.

Gertz, J.G. and Cohen, B.A. “Environment-specific combinatorial cis-regulation in synthetic promoters.” *Molecular Systems Biology.* 2009. 5:244.

Gertz, J.G., Siggia, E.D., Cohen, B.A. “Analysis of combinatorial cis-regulation in synthetic and genomic promoters.” *Nature.* 2009. 457:215.

Gietz, R.D. and Schiestl, R.H. “Large-scale High Efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method.” *Nature Protocols.* 2007. 2(1):38.

Gietz, R.D. and Woods, R.A. “Transformation of yeast by LiAc/SS carrier DNA/PEG method.” *Methods in Enzymol*ogy. 2002. 350: 87.

Granek, J. A. and Clarke, N.D. “Explicit equilibrium modeling of transcription-factor binding and gene regulation.” *Genome Biology.* 2005. 6:R87 doi:10.1186/gb-2005-6-10-r87

Harbison, C.T. *et al*. “Transcriptional regulatory code of a eukaryotic genome.” *Nature.* 2004. 431:99-104.

Herbig, E, *et al*. “Mechanism of Mediator Recruitment by Tandem Gcn4 Activation Domains and Three Gal11 Activator-Binding Domains.” *Molecular and Cellular Biology.* 2010. 30(10): 2376.

Hertz, G.Z., Hartzell, G.W. and Stormo, G.D. “Identification of consensus patterns in unaligned DNA sequences known to be functionally related.” *Comput. Appl. Biosci.* 1990. 6:81-92.

Hertz, G. Z. and Stormo, G. D. “Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.” *Bioinformatics* (1999) 15(7): 563-577 doi: 10.1093/bioinformatics/15.7.563

Hughes, J.D. *et al*. “Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.” *Journal of Molecular Biology.* 2000. 296:1205-14.

Istrail, S., De-Leon, S.B. and Davidson, E.H. “The regulatory genome and the computer.” *Developmental Biology.* 2007. 310:187.

Jedidi, I, *et al*. “Activator Gcn4 employs multiple segments of Med15/Gal11, including the KIX domain, to recruit mediator to target genes *in vivo*.” *Journal of Biological Chemistry.* 2010. 285 (4):2438.

Johnson, D.S. *et al*. “Genome-Wide Mapping of *in Vivo* Protein-DNA Interactions.” *Science.* 2007. 316:1497-1502.

Kent, N.A., *et al*. "Cbf1p is required for chromatin remodeling at promoter-proximal CACGTG motifs in yeast." *J Biol Chem.* 2004. 279(26):27116-23.

Kuras, L., Barbey R. and Thomas, D. "Assembly of a bZIP-bHLH transcription activation complex: formation of the yeast Cbf1-Met4- Met28 complex is regulated through Met28 stimulation of Cbf1 DNA binding." *EMBO J.* 1997. 16:2441–2451.

Kwasnieski, J. C. and Mogno I. *et al.* "Complex Effects of Nucleotide Variants in a Mammalian cis-Regulatory Element." *PNAS.* 2012. In Press.

Lee, T.I., Johnstone, S.E. and Young, R.A. "Chromatin immunoprecipitation and microarray-based analysis of protein location." *Nat Protoc.* 2006. 1:729-748.

Ligr, M., *et al.* "Gene expression from random libraries of yeast promoters." *Genetics.* 2006. 172:2113–2122.

Liu, J. and Stormo, G.D. "Combining SELEX with quantitative assays to rapidly obtain accurate models of protein-DNA interactions." *Nucleic Acids Res.* 2005. 33(17): e141

MacIsaac, K.D. *et al*. "An Improved Map of Conserved Regulatory Sites for Saccharomyces cerevisiae." *BMC Bioinformatics.* 2006. 7:113

Manke, T., Roider, H.G. and Vingron, M. "Statistical modeling of transcription factor binding affinities predicts regulatory interactions." *PLoS Comput Biol* 2008. 4:e1000039.

Matikainen T. *et al.* "Aromatic hydrocarbon receptor-driven Bax gene expression is required for premature ovarian failure caused by biohazardous environmental chemicals." *Nat Genet.* 2001. 28: 355–360.

Matys, V. *et al*. "TRANSFAC: transcriptional regulation, from patterns to profiles." *Nucleic Acids Res.* 2003. 31:374-8.

Melnikov, A. *et al*. "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay." *Nature Biotechnology.* 2012 30:271–277. doi: 10.1038/nbt.2137

Moreau, J.L., *et al.* "Regulated displacement of TBP from the PHO8 promoter *in vivo* requires Cbf1 and the Isw1 chromatin remodeling complex." *Mol Cell.* 2003. 11(6):1609-20.

Morozov, A.V. "Protein-DNA binding specificity predictions with structural models." *Nucleic Acids Research.* 2005. 33:5781-5798.

Mukherjee, S. and Berger, M.F., *et al*. "Rapid analysis of the DNA binding specificities of transcription factors with DNA microarrays." *Nature Genetics.* 2004. *36(12):1331-1339. Epub 2004 Nov 14.*

Murphy, K. F., Balazsi, G. and Collins, J. J. "Combinatorial promoter design for engineering noisy gene expression." *Proc. Natl Acad. Sci. USA*. 2007. 104:12726–12731.

Owuor, E.D. and Kong, A.N. "Antioxidants and oxidants regulated signal transduction pathways." *Biochem Pharmacol*. 2002. 64: 765–770.

Park, S.H., *et al.* "Nrg1 is a transcriptional repressor for glucose repression of STA1 gene expression in Saccharomyces cerevisiae." *Molecular and Cellular Biology*. 1999. 19(3):2044.

Patwardhan, R.P. *et al. "*Massively parallel functional dissection of mammalian enhancers *in vivo*." *Nat Biotechnol*. 2012. 30(3):265-70. doi: 10.1038/nbt.2136.

Prud'homme, B., Gompel, N. and Carroll, S.B. "Emerging principles of regulatory revolution."*PNAS.* 2007. 104:8605.

Radinsky, R. "Modulation of tumor cell gene expression and phenotype by the organ-specific metastatic environment." *Cancer Metastasis Rev.* 1995. 14: 323–338.

Raveh-Sadka, T., Levo, M., and Segal E. "Incorporating Nucleosomes into Thermodynamic Models of Transcription Regulation." *Genome Res*. 2009. 19:1480-1496.

Ren, B. *et al.* "Genome-Wide Location and Function of DNA Binding Proteins." *Science Signaling.*2000. 290:2306.

Roider, H. *et al*. "Predicting transcription factor affinities to DNA from a biophysical model." *Bioinformatics.* 2007. 23:134.

Segal, E. *et al*. "Predicting expression patterns from regulatory sequence in Drosophila segmentation." *Nature.* 2008. 451:535-540.

Sharon, E. *et al.* "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters." *Nature Biotechnology.* 2012. 30:521–530. doi: 10.1038/nbt.2205

Shea, M.A. and Ackers, G.K. "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation." *Journal of molecular biology.* 1985. 181:211.

Spivak, A.T. and Stormo, G. D. "ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species." *Nucleic Acids Research*. 2012. 40:1, D162.

Springer, C., *et al.* "Amino acid and adenine cross-pathway regulation act through the same 5'-TGACTC-3' motif in the yeast HIS7 promoter."*J Biol Chem.* 1996. 271(47):29637-43.

Storici, F. and Resnick, M. "The Delitto Perfetto Approach to *In Vivo* Site-Directed Mutagenesis and Chromosome Rearrangements with Synthetic Oligonucleotides in Yeast." *Methods in Enzymology.* 2006. 409:329.

Tanaka, M. "Modulation of promoter occupancy by cooperative DNA binding and activation-domain function is a major determinant of transcriptional regulation by activators *in vivo*." *PNAS.* 1996. 93(9):4311.

Thomas, D. *et al*. "MET4, a leucine zipper protein, and centromere-binding factor 1 are both required for transcriptional activation of sulfur metabolism in Saccharomyces cerevisiae." *Mol Cell Biol.* 1992. 12(4):1719-27.

Valouev, A. *et al.* "Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data." *Nat Methods.* 2008.doi:nmeth.1246

van Werven, F.J. and Timmers, H.T.M. "The use of biotin tagging in Saccharomyces cerevisiae improves the sensitivity of chromatin immunoprecipitation." *Nucleic Acids Research*. 2006. 34

Wang, T. and Stormo, G.D. "Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics.* 2003. 19:2369-80.

Ward, L. and Bussemaker, H. "Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences." *Bioinformatics.* 2008. 24:i165.

Wasson, T. and Hartemink, A. "An ensemble model of competitive multi-factor binding of the genome." *Genome Research*, 2009. 19:2101–2112.

Zhao, Y. and Stormo, G.D. "Quantitative analysis demonstrates most transcription factors require only simple models of specificity." *Nat Biotech.* 2011. 29(6):480

CHAPTER 4: Discussion

The ability to accurately model the biology of transcriptional regulation is a challenging problem. The number of moving parts is enormous and the computational complexity rapidly increases. However, this ability is a desirable one since the amount of regulatory sequence is much larger than coding sequence(Thurman, *et al.*, 2012 and Neph, 2012). To make sense of the many genomes now available, the ability to examine a sequence and predict its functional consequences is critical. One problem that arises in trying to understand the function of regulatory DNA is whether it is possible to capture the complexity of transcriptional regulation using relatively simple, generalizable mathematical rules. If this can be done, then principles learned in a simple system or one part of a system can be generalized and applied across the system. If it cannot be done, then all regulatory sequence is a special case requiring time-consuming and expensive experimental procedures to understand. This was formally expressed as hypothesis H1 of this thesis:

(H1) It will be possible to explain the complexity of biology with relatively simple, generalizable mathematical rules.

In chapter two of this thesis, I developed ReLos, a flexible framework for exploring the functional consequences of simple mathematical regulatory rules and used the tool to address hypothesis 1. I performed tree regression on a network of eleven expression modules comprising 254 genes across 255 environmental conditions (Beer and Tavazoie, 2004) to recover a proposed

set of combinatorial rules. ReLos was used to determine the functional consequences of those rules. It was shown that the module behavior could be reasonably recapitulated (mean gene-wise correlation of 0.7). Importantly, this level of correlation was achieved by considering only 13 *cis*-elements and the mean effects per condition of twelve combinatorial conditions describing the interactions of those elements. These results support hypothesis H1 and suggest that much of *cis*-regulation can be explained by a reasonably simple set of combinatorial interactions. The precise mathematical details of those interactions may be complicated, but the total number and type of interactions that must be considered to approximate the biology were surprisingly few. It is also very likely the case that as more genes and regulatory modules are added, more regulatory elements and rules would need to be considered. However, the basic result remains the same: a handful of cis-regulatory elements using a small set simple set of rules were all that were required to distinguish the various patterns of expression, rather than dozens of elements and scores of rules.

One area where ReLos could be improved is its pre-configured support for additional frameworks for modeling *cis*-regulation. Since multiple different mathematical formalisms have been applied to the problem of mapping sequence to expression (Beer and Tavazoie, 2004; Bussemaker, Li and Siggia, 2001; Conlon, 2003; Das, Banerjee and Zhang, 2004; Keles, van der Laan and Eisen, 2002; Wang, *et al.,* 2002). ReLos does not pre-specify any particular mathematical framework for determining the functional consequences of a given sequence, but allows the user to choose between several sets of pre-defined formalisms and to "plug in" their own formalism if the existing abilities were unsuitable to the user's need. One notable exception to ReLos-supplied frameworks is a statistical thermodynamic description of transcription (Shea

and Akers, 1995; Buchler, Gerland and Hwa, 2003). These models have gained popularity in recent years (Cohen, Siggia, and Gertz, 2009; Djordjevic, Sengupta and Shraiman, 2003; He, et al, 2009; Roider, et al. 2007; Segal, *et al*. 2008 and Wasson and Hartemink, 2009) due in large part due to their ability to describe complex systems with biophysically motivated parameters. Whereas other formalisms predict gene expression, the thermodynamic model provides biological hypotheses about the binding of transcription factors and polymerase. Extending ReLos to include this formalism would allow researchers to easily explore the biophysical consequences of a variety of parameter choices on both transcription factor binding and gene expression, and to compare the results with other mathematical models of expression.

One problem with any model that relates sequence to expression is learning the model parameters. Synthetic promoters (Gertz and Cohen, 2009; Gertz, Siggia and Cohen, 2009; Ligr *et al*., 2006; Kwasnieski, and Mogno *et al.*, 2012; Melnikov, *et al*., 2012 and Patwarden, *et al*., 2012; Sharon, *et al.,* 2012) have emerged as an *in vivo* method to characterize models of expression in a controlled and systematic manner. To date, all such methods have relied directly or indirectly on expression driven by the synthetic promoter. However, models built with only expression data cannot distinguish between multiple biologically distinct hypothesis that produce equivalent expression results and risk missing important biological features such as transcription factor cooperativity that may be masked in the expression data. Even in the absence of complex interactions, having only expression data means that the model cannot deconvolve the effects of TF-RNAP interactions from TF-DNA affinity. This results in being unable to distinguish between models wherein TF-DNA binding is relatively strong and TF-RNAP interactions are relatively weak from models where TF-DNA binding is relatively weak and TF-RNAP

interactions are relatively strong. In theory, an additional independent source of data would allow for these parameters to be separated, resulting in more meaningful mechanistic descriptions of biology. This is expressed broadly in hypothesis H2:

(H2) Given *in vivo* protein binding data, it will be possible to distinguish between models of transcriptional regulation that yield similar expression results but represent distinct biophysical mechanisms

In chapter three of this thesis, a ChIP-based approach for measuring occupancy on synthetic promoters was developed to address this hypothesis. This approach was used to obtain occupancy information for Cbf1 and Gcn4 in libraries of promoters containing sites for Cbf1, Gcn4, Met31/Met32, and Nrg1 in glucose and Amino Acid Starvation (AAS) conditions and to the best of my knowledge, provides the first data set where quantitative occupancy information and accurate expression measurements are available for a large set of synthetic promoters. The uniqueness of this dataset is in its ability to directly ask what the effect on expression of binding is for a particular transcription factor. This question cannot be readily addressed by genomic methods. The genomic ChIP signal is complicated by many factors, such as different shearing efficiencies, that make direct comparison of binding at one location to another difficult. In this case, all promoters were integrated at the same genomic locus and all drive the expression of the same gene. This allows a direct examination of the effect of binding on expression.

The principal aim of generating occupancy data was to distinguish between different biophysical hypotheses in the model that produce the same expression results. To address this

question, the parameters of the thermodynamic model were fit to each data source separately and then to both data sources simultaneously. There were several cases where having both data sets made distinguishing between alternative hypotheses possible. One case was to distinguish between a change in the apparent Ka of Cbf1 versus a change in the Cbf1-RNAP interaction between Glucose and AAS conditions. The effect on expression of the two models was equivalent, but the occupancy data strongly argue that Cbf1 binding is the same between the two conditions (Table 3.3 and Figure 3.7) The occupancy data also revealed apparent Gcn4 cooperativity (Figure 3.15), and a negative interaction between Gcn4 and Nrg1 (Figure 3.16). One question is why an effect such as the Nrg1 site interaction was not found when fitting only to expression data. There are two non-exclusive possibilities, one technical and one mathematical.

The expression data for these libraries was obtained via flow cytometry. However, the cytometer has an upper limit of detection and many promoters with multiple Gcn4 sites in them are gated out due to expressing beyond the limits of detection. This means that the expression data is being trained on a subset of data where there are fewer promoters with Gcn4 but no Nrg1. Without adequate examples of Gcn4 behavior in the presence and absence of Nrg1, it is difficult to fit the interaction terms. In theory, this limitation could be resolved by adjusting the voltage down to allow accurate quantification of highly expressed promoters. In practice, however, the voltage is already low enough that to lower it further would risk losing the ability to normalize the expression to the control promoters present on each plate. An alternative would be a new method of quantification. Recently, Kwasnieski and Mogno, *et al.* (2012) developed such a method using next generation sequencing. This allows for the quantification of a much larger

dynamic range of expression, at the expense of losing information about the population variance. It will be interesting to build new libraries using these techniques and to observe whether the interactions can be recovered with expression data with a larger dynamic range.

The second reason the interactions may not be recovered when fitting to expression data is the model formalism. When fitting to occupancy, there is little to no information regarding polymerase binding, so a smaller subset of the model terms are used. In particular, the interaction term between the factors and RNAP was not modeled when fitting only to occupancy data. It is formally possible for an interaction to be expressed through the polymerase term by destabilizing all states in which both interacting proteins are bound to the promoter. This effect is necessarily weaker in impact than a direct interaction, but could be sufficient to mask the interaction in the expression data. Indeed, a fit to both data sets that includes the Gcn4-Nrg1 interaction term results in a substantially weaker Nrg1-RNAP interaction (Table 3.3). Whatever the case, the interaction is clearly present in the occupancy data, which argues for a mode of direct interaction.

One of the most interesting phenomenon observed in the data was the switching of the Gcn4 site from a repressive role in glucose to an activating role in AAS (Figure 3.17). With only the expression data, this effect can be modeled by Gcn4 switching behavior between conditions. Inclusion of the occupancy data results in numeric constraints that are only cleanly resolved by accounting for competitive binding, thus suggesting competitive binding by Gcn4 and another factor. This behavior is consistent with earlier reports of Gcn4 competing with other factors such as Bas1 (Arndt and Fink, 1986 and Springer, 1996).

Competition between factors, especially between activators and repressors, has emerged recently as a recurring theme in transcriptional regulation (Zhou and O'Shea, 2011; White, *et al.*, 2012 and Wasson and Hartemink, 2009). In one sense, competition should not be surprising. There are many factors being expressed in the cell at the same time, all of which have varying affinity for a given sequence of DNA. What does seem surprising is the degree to which competitive binding seems to be a feature of transcriptional regulation rather than a side-effect. The recurrence of this effect suggests that more theoretical work should be done to fully explore the functional consequences of competitive binding.

There are several additional points of model improvements that should be considered for future research. First, the current version of the model does not take into account non-specific binding. This means that promoters without a pre-identified binding site for a factor are always predicted to have an occupancy of zero. Hence, those promoters were excluded from the fit, although they may contain useful information for setting the relative scale of occupancy values.

Another area for future research is in directly integrating both sources of data into the thermodynamic model, discussed more thoroughly in the chapter three results. When incorporating both sources of data, non-linear fitting routines seemed to prefer parameters sets that favored fitting occupancy well over fitting expression well. While this is most likely due to fitting artifacts such as the relative biological noise in the two data sets, the idea that some factors such as Cbf1 may require more sophisticated descriptions of their effect on expression cannot be entirely ruled out, and should be researched further.

In modeling the occupancy data, one choice that had to be made was whether to model the data as the expected number of proteins bound to the promoter (the average occupancy), or as

the probability of at least one transcription factor being bound to the promoter (the probability of occupancy). The average occupancy is expected to increase monotonically with the number of binding sites whereas the probability of occupancy is expected to saturate. For this work, the probability of occupancy was chosen as the appropriate model for several reasons. Empirically, using the probability of occupancy resulted in better fits to the data than using the average occupancy. Second, the occupancy appeared to start saturating (for instance, see Figure 3.15) with increasing number of binding sites. However, there were not many promoters with more than three binding sites for a particular factor, so it is difficult to determine from this data whether the observed saturation is real or an artifact of small numbers. To address this issue, another library could be built that focused on binding sites for one or two factors, thus increasing the chance of observing many binding sites for a single factor in a single promoter.

An interesting corollary to the probability of occupancy question is what the cell actually reads out. Does the cell engage in a molecular form of counting how many proteins are bound to a site, or does the cell only care that at least one protein is bound to the site? This question could be addressed by combining the occupancy approach outlined in this work with the next-generation-sequencing approach to synthetic promoters developed by Kwasnieski and Mogno, *et al.* (2012) to build libraries with fewer types of binding sites, resulting in more promoters with many binding sites. The next-generation-sequencing approach expands the dynamic range of the expression assay. In theory, this would allow us to observe the point at which expression saturates and combine the information with the occupancy data to learn whether saturation occurs at a point equal to or greater than the probability of at least one TF bound being one.

In the absence of the perfect experiment, it is interesting to note the results of Sharon, *et al.* (2012) They systematically varied thousands of promoters, including promoters with multiple GCN4 sites (up to seven). In their hands, they observed a logistic binding curve for GCN4 with expression saturating at four binding sites. This coincidentally agrees well with the binding data observed in my work, where the transition point of Gcn4 occurred between two and three binding sites, and the probability of binding appears to start saturating at four binding sites. Although these data are consistent, a direct measurement using the same promoter backbone and promoter construction are necessary to validate this intriguing possibility.

Tagging a transcription factor, building libraries for each tagged factor, and performing ChIP is a time and labor-intensive process and it is worth considering the costs and benefits of this approach. On the one hand, BirA is now integrated into the genome of the yeast strains we use for constructing synthetic promoter libraries, so study of additional factors is a one-step rather than a multi-step process. Furthermore, recent advances in library construction made by Kwasnieski and Mogno, *et al.* (2012) reduce the cost and effort required to build multiple libraries, and the new pooled strategy of library creation works well with the pooled strategy for ChIP employed in this study. On the other hand, each ChIP experiment is costly in terms of reagents and there may be alternative sources of information that can be more readily acquired, such as the TF concentration through GFP fusions. With that said, the occupancy does provide information that cannot be readily obtained any other way since it is the synthesis of the Ka and concentration of the TF and its interactions with other proteins. For instance, a study of only concentration of the TF would not have revealed the Gcn4 cooperativity, and a previous analysis of Cbf1 concentration resulted in the erroneous conclusion that the effective Ka of Cbf1 is

considerably less in glucose than in AAS. The key moving forward will be to increase the amount of information obtained from any given experiment. One way to do that would be to focus on the binding of a single transcription factor in a variety of different libraries, including libraries with different strengths of binding sites. This would maximize the information learned for the effort required to tag and ChIP the factor.

In all, the occupancy data complemented the expression data and provided new avenues for questioning and model improvement. Hypothesis (H2) claimed that having occupancy data will make it possible to distinguish between different biological models that give rise to the same expression results. This was demonstrated in several cases and provides encouraging results for additional study of ways to incorporate multiple sources of data into models of expression.

References

Arndt, K. and Fink, G. R. "GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC 3' sequences." *Proc Natl Acad Sci U S A*. 1986. 83(22): 8516-20.

Beer, M.A. and Tavazoie, S. "Predicting gene expression from sequence." *Cell.* 2004. 117:185-98.

Brzovic, P.S., *et al*. "The acidic transcription activator Gcn4 binds the mediator subunit Gal11/ Med15 using a simple protein interface forming a fuzzy complex." *Molecular Cell*. 2011. 44(6): 942.

Buchler, N.E., Gerland, U. and Hwa, T. "On schemes of combinatorial transcription logic." *PNAS.* 2003. 100:5136.

Bussemaker, H.J., Li, H., Siggia, E.D. "Regulatory element detection using correlation with expression." *Nat Genet*. 2001. 27:167–171. doi: 10.1038/84792.

Conlon, E.M., *et al*. "Integrating regulatory motif discovery and genome-wide expression analysis." *Proc Natl Acad Sci U S A*. 2003. 100:3339–3344. doi:

Das, D., Banerjee, N. and Zhang, M.Q. "Interacting models of cooperative gene regulation." *Proc Natl Acad Sci U S A.* 2004. 101:16234-9.

Djordjevic, M., Sengupta, A. and Shraiman, B. "A Biophysical Approach to Transcription Factor Binding Site Discovery." *Genome Res.* 2003. 13: 2381-2390.

He, X., *et al*. "A Biophysical Model for Analysis of Transcription Factor Interaction and Binding Site Arrangement from Genome-Wide Binding Data." *PLoS ONE.* 2009. 4(12): e8155. doi: 10.1371/journal.pone.0008155

Jedidi, I, *et al*. "Activator Gcn4 employs multiple segments of Med15/Gal11, including the KIX domain, to recruit mediator to target genes in vivo." *Journal of Biological Chemistry.* 2010. 285 (4):2438.

Keles, S., van der Laan M., Eisen, M.B. "Identification of regulatory elements using a feature selection method." *Bioinformatics.* 2002. 18:1167–1175. doi: 10.1093/bioinformatics/18.9.1167.

Kwasnieski, J. C. and Mogno, I., *et al.* "Complex Effects of Nucleotide Variants in a Mammalian cis-Regulatory Element." *PNAS.* 2012. In Press.

Ligr, M., *et al.* "Gene expression from random libraries of yeast promoters." *Genetics.* 2006. 172:2113–2122.

Melnikov, A. *et al*. "Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay." *Nature Biotechnology.* 2012 30:271–277. doi: 10.1038/nbt.2137

Neph, S. "An expansive human regulatory lexicon encoded in transcription factor footprints." *Nature.* 2012. 489:83–90.

Patwardhan, R.P. *et al. "*Massively parallel functional dissection of mammalian enhancers in vivo." *Nat Biotechnol*. 2012. 30(3):265-70. doi: 10.1038/nbt.2136.

Roider, H. *et al*. "Predicting transcription factor affinities to DNA from a biophysical model." *Bioinformatics.* 2007. 23:134.

Segal, E. *et al*. "Predicting expression patterns from regulatory sequence in Drosophila segmentation." *Nature.* 2008. 451:535-540.

Shea, M.A. and Ackers, G.K. "The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation." *Journal of molecular biology.* 1985. 181:211.

Sharon, E. *et al*. "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters." *Nat Biotech.* 2012. 30:521.

Springer, C., *et al.* "Amino acid and adenine cross-pathway regulation act through the same 5'-TGACTC-3' motif in the yeast HIS7 promoter." *J Biol Chem.* 1996. 271(47):29637-43.

Thurman, R., *et al.* "The accessible chromatin landscape of the human genome." *Nature. 2012.* 489:75–82.

Wang, W., *et al.* "A systematic approach to reconstructing transcription networks in Saccharomycescerevisiae." *Proc Natl Acad Sci U S A*. 2002. 99:16893–16898. doi: 10.1073/pnas.252638199.

Wasson, T. and Hartemink, A. "An ensemble model of competitive multi-factor binding of the genome." *Genome Research*, 2009. 19:2101–2112.

White, M., *et al*. "A model of spatially restricted transcription in opposing gradients of activators and repressors." *Molecular Systems Biology.* 2012. 8 Article number: 614

Zhou, X. and O'Shea, E.K. “Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4.” *Molecular Cell*. 2011. 42(6):826-36.