Genome-wide Analysis of Transcriptional Expression Programs, Regulatory Networks and Cis-Regulatory Sequences in *Saccharomyces cerevisiae*

by

Christopher T. Harbison

A.B., Biology
Harvard College, 1996

SUBMITTED TO THE DEPARTMENT OF BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February, 2005

Signature of Author: _____
Department of Biology
November 29, 2004

Certified by: _____
Dr. Richard A. Young
Professor of Biology
Thesis Supervisor

Accepted by: _____
Dr. Stephen P. Bell
Professor of Biology
Acting Co-Chair, Biology Graduate Committee

Dedication

To my parents Frank and Liz Harbison, for their unfailing encouragement and sacrifice.

To my wife, Lauri, for her love and support over the past six years.

To my daughters, Catherine and Mary, for their abundant smiles and laughter.

Acknowledgments

Thanks to all those in the Young Lab who provided their help throughout my time here. In particular, thanks to my long-time bench neighbors Dmitry Pokholok and Nancy Hannett for daily making ours an enjoyable workplace. Thanks to Ezra Jennings, Tony Lee and John Wyrick for their sage advice as senior graduate students, and to Nicola Rinaldi for lending her consummate computational skills to the lab. Thanks to all the technicians whose hard work went into every aspect of the projects described in this thesis, to all the post-docs who provided their advice and encouragement, and to all the computational and administrative staff who made sure that things got done on time.

Thanks, too, to Ben Gordon and Ernest Fraenkel for their invaluable work on "deciphering the regulatory code."

Thanks to Prof. Steve Bell, Prof. George Church, Prof. Phil Sharp and Prof. Frank Solomon for their assistance and time, serving on my thesis committee.

Thanks to Rick Young, for challenging me to develop new skills throughout my time here, but also offering his time and knowledge to help when needed.

Thanks, again, to my family.

Genome-wide Analysis of Transcriptional Expression Programs, Regulatory Networks and Cis-Regulatory Sequences in *Saccharomyces cerevisiae*

by

Christopher T. Harbison

Submitted to the Department of Biology on October 8, 2004
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy in Biology

Abstract

Historically, knowledge of gene-specific transcription has been accumulated by the study of the individual genetic and physical interactions between transcriptional regulators and the genes they regulate, often requiring considerable time and effort. Microarray technology now enables investigation of gene expression at the level of the entire genome, allowing researchers access to rich datasets and promising new levels of depth in the understanding of transcriptional regulation. Our lab has made use of these technologies both to measure the levels of all mRNA transcripts within a population of cells, as well as to locate the regions within the genome that are bound by transcriptional regulators.

Such studies not only allow for the functional annotation of both genes and regulators, but can also provide clues about the identity of the regulatory regions within DNA, the structure of global regulatory networks and the regulation of DNA-binding proteins. These and other insights are presented here based on our genome-wide studies of transcriptional regulation in the yeast *Saccharomyces cerevisiae*.

Thesis Supervisor: Dr. Richard A. Young
Title: Professor of Biology

Table of Contents

Chapter 1

Introduction: Mechanisms Governing the Activity of Transcriptional Regulators

Introduction

The control of gene expression is critical to cell survival, proliferation and differentiation. In eukaryotes, the regulation of transcriptional initiation by RNA polymerase II is a principal means by which such control is accomplished (Gill, 2001; Hahn, 2004; Ptashne and Gann, 1997; Roeder and Rutter, 1969). Transcriptional initiation of specific genes, in turn, is mediated by transcriptional regulatory proteins, which associate in a sequence-specific manner with short regions of DNA (Dynan and Tjian, 1985; Hampsey, 1998; Latchman, 1991; Tjian, 1996). These regulators can recruit other proteins (e.g. histone-modifying and remodeling complexes, co-regulators, the RNA polymerase holoenzyme or its associated factors) required for either the activation or repression of these genes in response to signaling cues (Cosma, 2002; Orphanides and Reinberg, 2002; Tjian, 1996). As a single regulator typically regulates dozens of genes (Iyer et al., 2001; Lee et al., 2002), and some of these genes themselves encode transcriptional regulators, changes in the activity of even a few such proteins can have a profound effect on cell homeostasis, response to environmental signals and processes such as cellular differentiation.

Historically, studies of transcriptional regulators have been focused on their interactions with only a few genes. These include the earliest genetic and biochemical experiments in viral and prokaryotic systems that established the paradigm for gene regulation (Jacob and Monod, 1961; Ptashne, 1967). Only within the past decade have advances in technology led to high-throughput methods that allow for the study of coordinate gene regulation throughout an entire genome (Banerjee and Zhang, 2002;

Taverner et al., 2004). Genome-wide expression analysis (DeRisi et al., 1997; Wodicka et al., 1997), for example, enables measurement of steady-state levels of all mRNA transcripts within a population of cells. Similarly, genome-wide location analysis has been developed to identify the genomic regions occupied by DNA-binding proteins (Iyer et al., 2001; Lieb et al., 2001; Ren et al., 2000).

In recent years, it has also become apparent that the control of regulatory proteins encompasses a large spectrum of mechanisms and is effected in a highly complex fashion (Orphanides and Reinberg, 2002). The study of genomic locations of transcriptional regulators, for example, demonstrates that not all predicted DNA binding sites are always (or ever) bound (Lee et al., 2002; Lieb et al., 2001; Zeitlinger et al., 2003). It is similarly true that the act of regulator binding does not necessarily confer changes in transcriptional activity (Bar-Joseph et al., 2003; Iyer et al., 2001; Ren et al., 2000). Other advances in our understanding have come from discoveries of novel mechanisms of regulation by RNAs (Lau et al., 2001; Lee and Ambros, 2001; Novina and Sharp, 2004) as well new studies in chromatin regulation (Jenuwein and Allis, 2001; Narlikar et al., 2002).

The set of mechanisms responsible for translating the information contained within the "regulatory code" of DNA into condition-specific changes in gene expression is far from completely understood (Orphanides and Reinberg, 2002). Nevertheless, the known mechanisms that control the behavior of transcriptional regulators can be grouped into those that regulate: the genome of a cell, transcription, translation, protein modification, and higher order protein states (Table 1). In general, all of these mechanisms work by modifying the total amount of regulatory protein, its nuclear

localization, its ability to bind DNA, or its capacity to interact with other proteins necessary for transcriptional control.

It is important to realize that in many cases these mechanisms are closely linked. For example, protein modification may result in changes in localization, DNA-binding ability or the capacity to interact with other proteins. It is also true that for many regulators multiple levels of regulation may exist.

Level of regulation	Mechanism	Transcriptional Regulator	Reference
Genome	Copy number	SRY	(*1, 2*)
		Drosophila HLH proteins	(*3-5*)
	Gene rearrangement	a1, a2, α1, α2	(*6-12*)
Transcription	Silencing	Homeodomain proteins	(*13-18*)
	Transcriptional initiation	Thi2, MyoD	(*19-22*)
	Transcriptional elongation	c-Myc	(*23-25*)
	mRNA stability	Phabulosa, Phavoluta	(*26, 27*)
	mRNA localization	Ash1, Bicoid	(*28-31*)
	mRNA processing	Wt1	(*32-35*)
Translation	Translational initiation	Gcn4, C/EBPβ	(*36-43*)
	Translational elongation	LIN-14	(*44-47*)
Protein modification	Protein folding	Steroid hormone receptors	(*48-52*)
	Protein cleavage	Rim101	(*53*)
	Chemical modification	c-Jun	(*54-57*)
	Protein stability	Gcn4	(*58-60*)
Higher order states	Translocation	Msn2, NFκB	(*61-64*)
	Molecular cofactors	Leu3	(*65-69*)
	Protein-protein interactions	Gal4, Ste12, α2	(*70-74*)
	Cooperative binding	"Enhanceosomes"	(*75-80*)
	DNA accessibility	Sko1, Sum1	(*81-83*)

Table1: Methods of regulation of transcriptional regulators.

1. A. H. Sinclair *et al.*, *Nature* 346, 240-4 (Jul 19, 1990).
2. N. Nasrin *et al.*, *Nature* 354, 317-20 (Nov 28, 1991).
3. S. M. Parkhurst, D. Bopp, D. Ish-Horowicz, *Cell* 63, 1179-91 (Dec 21, 1990).
4. K. Hoshijima *et al.*, *Nucleic Acids Res* 23, 3441-8 (Sep 11, 1995).
5. C. Schutt, R. Nothiger, *Development* 127, 667-77 (Feb, 2000).
6. K. A. Nasmyth, K. Tatchell, B. D. Hall, C. Astell, M. Smith, *Nature* 289, 244-50 (Jan 22, 1981).
7. J. Rine, J. N. Strathern, J. B. Hicks, I. Herskowitz, *Genetics* 93, 877-901 (Dec, 1979).
8. A. J. Klar, J. N. Strathern, J. R. Broach, J. B. Hicks, *Nature* 289, 239-44 (Jan 22, 1981).
9. G. Ammerer, G. F. Sprague, Jr., A. Bender, *Proc Natl Acad Sci U S A* 82, 5855-9 (Sep, 1985).
10. J. Rine, I. Herskowitz, *Genetics* 116, 9-22 (May, 1987).
11. M. N. Hall, A. D. Johnson, *Science* 237, 1007-12 (Aug 28, 1987).
12. G. F. Sprague, Jr., R. Jensen, I. Herskowitz, *Cell* 32, 409-15 (Feb, 1983).
13. J. C. Hombria, B. Lovegrove, *Differentiation* 71, 461-76 (Oct, 2003).
14. M. L. Howard, E. H. Davidson, *Dev Biol* 271, 109-18 (Jul 1, 2004).
15. J. A. Simon, J. W. Tamkun, *Curr Opin Genet Dev* 12, 210-8 (Apr, 2002).
16. T. R. Breen, P. J. Harte, *Development* 117, 119-34 (Jan, 1993).
17. J. Simon, A. Chiang, W. Bender, M. J. Shimell, M. O'Connor, *Dev Biol* 158, 131-44 (Jul, 1993).
18. S. Tillib *et al.*, *Mol Cell Biol* 19, 5189-202 (Jul, 1999).
19. E. G. Jennings, Massachusetts Institute of Technology (2002).
20. D. Montarras, C. Pinset, J. Chelly, A. Kahn, F. Gros, *Embo J* 8, 2203-7 (Aug, 1989).
21. S. J. Tapscott *et al.*, *Science* 242, 405-11 (Oct 21, 1988).
22. H. Weintraub *et al.*, *Proc Natl Acad Sci U S A* 86, 5434-8 (Jul, 1989).
23. D. L. Bentley, M. Groudine, *Nature* 321, 702-6 (Jun 12-18, 1986).
24. C. Chen, A. J. Sytkowski, *J Biol Chem* 276, 38518-26 (Oct 19, 2001).
25. S. Wright, J. M. Bishop, *Proc Natl Acad Sci U S A* 86, 505-9 (Jan, 1989).
26. C. A. Kidner, R. A. Martienssen, *Nature* 428, 81-4 (Mar 4, 2004).
27. A. C. Mallory *et al.*, *Embo J* 23, 3356-64 (Aug 18, 2004).
28. X. Darzacq, E. Powrie, W. Gu, R. H. Singer, D. Zenklusen, *Curr Opin Microbiol* 6, 614-20 (Dec, 2003).
29. A. Sil, I. Herskowitz, *Cell* 84, 711-22 (Mar 8, 1996).
30. W. Driever, C. Nusslein-Volhard, *Cell* 54, 95-104 (Jul 1, 1988).
31. A. Ephrussi, D. St Johnston, *Cell* 116, 143-52 (Jan 23, 2004).
32. V. Scharnhorst, A. J. van der Eb, A. G. Jochemsen, *Gene* 273, 141-61 (Aug 8, 2001).
33. P. M. Sharma, M. Bowman, S. L. Madden, F. J. Rauscher, 3rd, S. Sukumar, *Genes Dev* 8, 720-31 (Mar 15, 1994).
34. K. D. Wagner, N. Wagner, A. Schedl, *J Cell Sci* 116, 1653-8 (May 1, 2003).
35. A. Hammes *et al.*, *Cell* 106, 319-29 (Aug 10, 2001).

36. G. Thireos, M. D. Penn, H. Greer, *Proc Natl Acad Sci U S A* 81, 5096-100. (1984).
37. A. G. Hinnebusch, *Proc Natl Acad Sci U S A* 81, 6442-6. (1984).
38. P. P. Mueller, A. G. Hinnebusch, *Cell* 45, 201-7 (Apr 25, 1986).
39. A. G. Hinnebusch, *J Biol Chem* 272, 21661-4. (1997).
40. D. P. Ramji, P. Foka, *Biochem J* 365, 561-75 (Aug 1, 2002).
41. P. Descombes, U. Schibler, *Cell* 67, 569-79 (Nov 1, 1991).
42. V. Ossipow, P. Descombes, U. Schibler, *Proc Natl Acad Sci U S A* 90, 8219-23 (Sep 1, 1993).
43. C. F. Calkhoven, C. Muller, A. Leutz, *Genes Dev* 14, 1920-32 (Aug 1, 2000).
44. L. He, G. J. Hannon, *Nat Rev Genet* 5, 522-31 (Jul, 2004).
45. N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, *Science* 294, 858-62 (Oct 26, 2001).
46. R. C. Lee, V. Ambros, *Science* 294, 862-4 (Oct 26, 2001).
47. P. H. Olsen, V. Ambros, *Dev Biol* 216, 671-80 (Dec 15, 1999).
48. D. B. DeFranco, P. Csermely, *Sci STKE* 2000, PE1 (Jul 25, 2000).
49. M. N. Arbeitman, D. S. Hogness, *Cell* 101, 67-77 (Mar 31, 2000).
50. K. J. Howard, S. J. Holley, K. R. Yamamoto, C. W. Distelhorst, *J Biol Chem* 265, 11928-35 (Jul 15, 1990).
51. W. B. Pratt, D. O. Toft, *Exp Biol Med (Maywood)* 228, 111-33 (Feb, 2003).
52. S. J. Holley, K. R. Yamamoto, *Mol Biol Cell* 6, 1833-42 (Dec, 1995).
53. W. Li, A. P. Mitchell, *Genetics* 145, 63-73 (Jan, 1997).
54. A. J. Whitmarsh, R. J. Davis, *Cell Mol Life Sci* 57, 1172-83 (Aug, 2000).
55. R. J. Davis, *Cell* 103, 239-52 (Oct 13, 2000).
56. G. L. Johnson, R. Lapadat, *Science* 298, 1911-2 (Dec 6, 2002).
57. B. Derijard *et al.*, *Cell* 76, 1025-37 (Mar 25, 1994).
58. D. Kornitzer, B. Raboy, R. G. Kulka, G. R. Fink, *Embo J* 13, 6021-30. (1994).
59. A. Meimoun *et al.*, *Mol Biol Cell* 11, 915-27. (2000).
60. Y. Chi *et al.*, *Genes Dev* 15, 1078-92. (2001).
61. W. Gorner *et al.*, *Genes Dev* 12, 586-97 (Feb 15, 1998).
62. T. Beck, M. N. Hall, *Nature* 402, 689-92 (Dec 9, 1999).
63. P. Cartwright, K. Helin, *Cell Mol Life Sci* 57, 1193-206 (Aug, 2000).
64. S. C. Sun, P. A. Ganchi, D. W. Ballard, W. C. Greene, *Science* 259, 1912-5 (Mar 26, 1993).
65. P. R. Brisco, G. B. Kohlhaw, *J Biol Chem* 265, 11667-75 (Jul 15, 1990).
66. P. Friden, C. Reynolds, P. Schimmel, *Mol Cell Biol* 9, 4056-60 (Sep, 1989).
67. C. R. Kirkpatrick, P. Schimmel, *Mol Cell Biol* 15, 4021-30 (Aug, 1995).
68. J. Y. Sze, E. Remboutsika, G. B. Kohlhaw, *Mol Cell Biol* 13, 5702-9 (Sep, 1993).
69. D. Wang, Y. Hu, F. Zheng, K. Zhou, G. B. Kohlhaw, *J Biol Chem* 272, 19383-92 (Aug 1, 1997).
70. J. Ma, M. Ptashne, *Cell* 50, 137-42 (Jul 3, 1987).
71. J. Zeitlinger *et al.*, *Cell* 113, 395-404 (May 2, 2003).
72. T. Li, M. R. Stark, A. D. Johnson, C. Wolberger, *Science* 270, 262-9 (Oct 13, 1995).
73. A. M. Miller, V. L. MacKay, K. A. Nasmyth, *Nature* 314, 598-603 (Apr 18-24, 1985).

74. S. Tan, T. J. Richmond, *Nature* 391, 660-6 (Feb 12, 1998).
75. J. Banerji, L. Olson, W. Schaffner, *Cell* 33, 729-40 (Jul, 1983).
76. D. D. Bowtell, T. Lila, W. M. Michael, D. Hackett, G. M. Rubin, *Proc Natl Acad Sci U S A* 88, 6853-7 (Aug 1, 1991).
77. D. Thanos, T. Maniatis, *Cell* 83, 1091-100 (Dec 29, 1995).
78. M. Carey, *Cell* 92, 5-8 (Jan 9, 1998).
79. M. Carey, Y. S. Lin, M. R. Green, M. Ptashne, *Nature* 345, 361-4 (May 24, 1990).
80. T. K. Kim, T. Maniatis, *Mol Cell* 1, 119-29 (Dec, 1997).
81. M. A. Garcia-Gimeno, K. Struhl, *Mol Cell Biol* 20, 4340-9 (Jun, 2000).
82. A. C. Vincent, K. Struhl, *Mol Cell Biol* 12, 5394-405 (Dec, 1992).
83. M. Pierce *et al.*, *Mol Cell Biol* 23, 4814-25 (Jul, 2003).

Genomic regulation

Controlling gene number. The most basic step at which of control of regulatory proteins can be exercised is at the level of DNA. One way in which a cell may influence the activity of regulatory protein is by controlling the presence or number of specific regulatory genes within a cell. The *SRY* gene in mammals has been identified as encoding a transcriptional regulator that is necessary for masculine development upon sexual differentiation (Nasrin et al., 1991; Sinclair et al., 1990). Inactivation of this gene in XY individuals leads to development of phenotypic females (Berta et al., 1990; Jager et al., 1990). Likewise, the presence of this gene in XX individuals leads to masculinization in mice (Koopman et al., 1991). As this gene is transmitted on the Y chromosome, it is chromosomal inheritance that represents the fundamental regulatory step of its activity.

Similarly, the copy number of transcription factor genes controls sex determination in *Drosophila melanogaster.* Specifically, the relative activity of transcription factors encoded on the sex chromosome and transcription factors encoded on autosomes serves as a chromosomal counting mechanism that determines whether or not a master regulatory gene is activated (Hoshijima et al., 1995; Parkhurst et al., 1990; Schutt and Nothiger, 2000). Other mechanisms that are known to affect the copy number of transcriptional regulatory genes (e.g. partial replication, extrachromosomal duplications, aneuploidy, viral infection) could also result in changes in transcriptional activity, and in some cases have been linked to oncogenesis (Brown et al., 1986; Schwab et al., 1983; Varmus, 1984).

Genomic rearrangements. One of the most intriguing puzzles in the early study of gene regulation was the control of cell type in yeast. In so-called "homothallic" strains of yeast, both mating types contain the information required for the differentiation into either the a or α cell type. Information from only one of the gene "cassettes" corresponding to mating type is expressed in any haploid cell (Klar et al., 1981; Rine et al., 1979). Non-expressed genes are "silenced" by a persistent alteration of the surrounding chromatin (Grewal and Moazed, 2003; Loo and Rine, 1994). Expression from these cassettes (*HMR* for a type and *HML* for α) is dependent on their location within the genome (Nasmyth et al., 1981; Rine and Herskowitz, 1987). In order to change cell type, a specific endonuclease converts a silenced version of the opposite cassette into an actively transcribed region, known as the mating type or *MAT* locus (Nasmyth et al., 1981). As it turns out, the genes encoded by the mating locus are transcription factors, namely a1 and a2 or α1 and α2 (Ammerer et al., 1985; Hall and Johnson, 1987; Sprague et al., 1983). In a single haploid cell, only either a or α types of factors are produced, resulting in a cell-type specific gene expression program. Genomic translocations that place transcriptional regulatory genes under the control of inappropriate promoters can result in disease (Boxer and Dang, 2001; Dalla-Favera et al., 1983; Hamlyn and Rabbitts, 1983; Kelly and Gilliland, 2002; Rabbitts, 1999).

Transcriptional regulation

Silencing. In the phenomenon of transcriptional silencing, gene expression is reduced in a way that is persistent and heritable. Silencing has been associated with local alterations in chromatin structure (Allfrey et al., 1964; Aparicio et al., 1991; Grewal and Moazed, 2003; Hebbes et al., 1988; Loo and Rine, 1994). Specifically, silenced regions in yeast, such as the yeast mating type loci, are occupied by histones that are less acetylated on specific lysine residues relative to those occupying the rest of the genome (Braunstein et al., 1993). Other mechanisms, including the effect of other types of chemical modification of histones (Cuthbert et al., 2004), the regulation of histone density (Wyrick et al., 1999), the role of other chromatin-associated proteins (Diffley and Stillman, 1989), the activity of regulatory RNAs (Baulcombe, 2004; Brown et al., 1992; Novina and Sharp, 2004; Schramke and Allshire, 2004), and the process of DNA methylation (Lorincz et al., 2004; Nan et al., 1998) have also been implicated in transcriptional silencing.

The regulation of homeotic genes in flies (and other animals) provides a paradigm for the role of silencing in the regulation of transcriptional regulators. During the process of development, the timing of expression of transcriptional regulators is controlled in a complex and ordered fashion (Hombria and Lovegrove, 2003). In particular, a class of transcriptional regulators called homeodomain proteins plays a critical role in establishing body plans in early development. Normal development requires that each of these proteins be expressed only during the appropriate time or in the appropriate cell type (Howard and Davidson, 2004). Control over the expression of these genes is mediated by members of the Polycomb group (PcG) and Trithorax group (trxG) of proteins (Simon and Tamkun, 2002). The former are generally responsible for

switching homeotic genes into a transcriptionally silent state (Hombria and Lovegrove, 2003; Howard and Davidson, 2004), and the latter are required for reversing the process (Breen and Harte, 1993). PcG and trxG complexes are capable of both sequence-specific recognition of DNA as well as chromatin modification and remodeling (Simon et al., 1993; Simon and Tamkun, 2002; Tillib et al., 1999). The dysregulation that results from improper silencing of their target transcriptional regulatory genes leads to severe anatomical abnormalities.

Transcriptional initiation. The control of transcriptional initiation represents the most common means of regulation of protein amounts within the cell (Latchman, 1991). Such regulation includes switching on (or off) synthesis of a given gene as well as modulating the levels of an actively transcribed gene. Although most regulators may either activate or repress gene expression, some regulators are capable of both activities depending, for example, on their association with other proteins or on specific modifications that alter their activity.

Regulators control the rate of transcription of specific genes by two general methods (Felsenfeld and Groudine, 2003; Lee and Young, 2000; Narlikar et al., 2002; Struhl, 1995). The first is to lead to alterations of the chromatin state of a gene. A number of transcriptional regulators have been found to interact with proteins that alter chromatin structure (Cosma, 2002; Struhl, 1999). The first such group of proteins are called chromatin modifiers, which covalently attach or remove chemical groups or polypeptides, through, for example, acetylation, methylation, phosphorylation, and ubiquitination of histones (Berger, 2002). Such modification results in changes in the

association of histones with DNA. The best-studied histone modifiers include histone acetyltransferases (HATs), e.g. Gcn5, CBP (Bannister and Kouzarides, 1996; Bannister and Miska, 2000; Brownell et al., 1996; Grunstein, 1997), and histone deacetylases (HDACs), e.g. Rpd3, Sir2 (Blander and Guarente, 2004; Kurdistani and Grunstein, 2003). A second category of proteins is responsible for higher-order remodeling of chromatin structures (Vignali et al., 2000). These complexes, e.g. SWI/SNF, ISWI, use ATP to mechanically reposition histones with respect to DNA. The general effect of both chromatin modifiers and remodelers is to alter the accessibility of DNA to additional transcriptional regulators, other regulatory proteins or the polymerase holoenzyme.

The second way a transcriptional regulator affects the expression of a gene is to directly or indirectly recruit components of the transcriptional machinery to the gene promoter or direct their activity. This may be accomplished by primary interactions between a regulator and transcriptional machinery or may be mediated by association with coactivators and corepressors or other intermediaries. The protein targets of regulators may include general transcription factors, so-called Mediator/SRB proteins which associate with the carboxy-terminal tail of polymerase, or polymerase itself (Hampsey, 1998; Kelleher et al., 1990; Lee and Young, 2000; Thompson et al., 1993). Binding by transcriptional regulators may recruit these elements by direct protein-protein interactions (Ptashne and Gann, 1997) or by altering the three-dimensional structure of DNA by twisting or bending in such a way that facilitates their binding (Giese et al., 1992; Lin and Green, 1991).

In the case of transcriptional repressors, control of gene expression might result from antagonism of activator function (e.g. by masking activation domains or

occluding activator binding sites). Alternatively, repression might result from a physical barrier to transcription by binding to DNA or recruiting repressive complexes, preventing the binding or elongation activities of polymerase (Hampsey, 1998; Struhl, 1999).

An example of transcriptional control of a transcriptional regulator is Thi2. The mRNA transcripts of *THI2*, a transcriptional regulator of thiamine biosynthesis, are barely detectable in cells grown in rich medium (Jennings, 2002). In a medium lacking thiamine, however, expression of *THI2* is greatly increased. Transcriptional regulation of transcriptional regulators can be understood as a vast regulatory network, consisting of distinct regulatory motifs (Horak et al., 2002; Lee et al., 2002; Luscombe et al., 2004; Milo et al., 2002), as discussed in more detail in Chapter 2. It has been shown, for example, that transcriptional regulators of the cell cycle form a complete temporal circuit in which one regulator controls the expression of the next (Simon et al., 2001).

In higher eukaryotes, tissue-specific expression of a transcriptional regulator can serve as a mechanism for cellular differentiation. MyoD is a regulator of differentiation into muscle cells. Expression of MyoD is confined to muscle cell types (proliferating myoblasts and myotubes) (Montarras et al., 1989; Tapscott et al., 1988). The potency of MyoD as a "master regulator" of gene expression is such that it is capable of initiating myogenesis even in differentiated cells (Tapscott et al., 1988; Weintraub et al., 1989).

Transcriptional elongation. In some cases it is not the initiation of transcription but control of the elongation process that is the critical step in regulation of a transcriptional regulatory gene. In studies in a promyelocytic leukemia cell line, differentiation into granulocytes is accompanied by a drastic reduction in levels of

RNA from the proto-oncogene *c-myc*. Nuclear runoff assays quantify the incorporation of radiolabeled nucleotides, enabling the specific measurement of elongating mRNA transcripts. Such assays indicated that the fold-change of *c-myc* varied across the length of the transcript, with the first exon being 15-fold in excess of the second in differentiated cells (Bentley and Groudine, 1986). Subsequent analysis confirmed that regulation of this transcription factor is accomplished by controlling transcription elongation past the first exon in a manner that is dependent on cis-acting sequences proximal to exon 1 (Chen and Sytkowski, 2001; Wright and Bishop, 1989).

mRNA stability. The steady-state levels of mRNA are influenced not only by the accumulation of newly transcribed messages, but also by the rate at which mRNA is degraded. In some cases, mRNA stability is globally affected by RNA processing steps, such as polyadenylation or 5' capping (Albig and Decker, 2001). Recently, novel mechanisms governing the degradation of mRNAs by short segments of complementary RNA have been discovered (He and Hannon, 2004; Lau et al., 2001; Lee and Ambros, 2001; Novina and Sharp, 2004). Intriguingly, different mechanisms seem to be at play and result in different fates for mRNAs regulated by these "microRNAs" (miRNAs). In *Arabidopsis thaliana*, highly complementary microRNAs that bind to and lead to the cleavage of specific mRNAs have been identified and target many known or predicted transcription factors (Rhoades et al., 2002). For example, the plant transcriptional regulatory genes PHABULOSA (PHB) and PHAVOLUTA (PHV), whose products are required for proper leaf development, have been identified as specific targets of miRNA regulation (Kidner and Martienssen, 2004; Mallory et al., 2004). Mutations that do not

alter the protein coding sequences for these regulators but that are predicted to disrupt miRNA hybridization lead to developmental abnormalities. The existence of similar mechanisms in other organisms (Yekta et al., 2004) indicates that regulation of transcriptional regulators by short highly complementary RNA may be an important theme, particularly for the process of development in multicellular organisms.

mRNA localization. RNA transcripts may be physically sequestered in order to control differential synthesis of the corresponding protein. For example, mRNA of a transcriptional regulator may be localized to a site within the cytoplasm. The phenomenon of mating type switching in budding yeast proceeds according to generation-specific rules that require differentiation into "mother" and "daughter" during cell division (Nasmyth and Shore, 1987; Strathern and Herskowitz, 1979). The means by which this occurs is the localization of transcripts of the regulatory *ASH1* gene to the site of a developing bud and subsequently to the "daughter" cell (Darzacq et al., 2003; Sil and Herskowitz, 1996). The presence of the resulting translated Ash1 protein in the cell defines its status as a daughter, and negatively regulates transcription of products required for mating type switching. Similar processes have been found to be critical in animal development.

Localization of mRNA of transcriptional regulators to define cell identity is mirrored in *Drosophila melanogaster* (Mohr and Richter, 2001; van Eeden and St Johnston, 1999). Here, the polarity of the embryo is initially defined by the local concentration within the egg of mRNA of the *bicoid* gene, which encodes a transcription factor that directs development of the anterior body plan (Driever and Nusslein-

Volhard, 1988; Ephrussi and St Johnston, 2004). Following translation of the mRNA and segmentation of the embryo, the resulting concentration gradient of Bicoid protein contributes to differential gene regulation in each of the segments of the developing embryo.

mRNA processing and modification. In eukaryotes, mRNA transcripts usually require further processing before being translated into protein. The best known such modification is the splicing of exons (Sharp, 1994), and represents yet another step subject to regulatory control. An example of a transcription factor that undergoes complex RNA modification is that of Wilms' tumor gene, *WT1* (Scharnhorst et al., 2001). Alternative splicing alone results in four isoforms of this protein. In addition to splicing, there is evidence that this gene is also subject to RNA editing of a single base, further increasing its molecular diversity (Sharma et al., 1994). Expression of these Wt1 isoforms varies temporally and according to cell type and species, and mutations in the *WT1* gene can have wide ranging effects on the development and maintenance of organs and organ functions (Wagner et al., 2003). In some cases, splice variants have been linked to specific regulatory activity. For example, use of a specific splice site in exon 9 results in an isoform that no longer possesses DNA binding activity, but rather is associated with RNA splicing machinery. In mice, altering the specific ratios of the various isoforms of Wt1 has been shown to result in defects in heart, kidney and gonad formation (Hammes et al., 2001).

Translational regulation

Translational initiation. Translational initiation may be regulated in two ways—the rate at which initiation occurs, as well as the exact location of the start site associated with a given protein. Translational control in yeast has been studied most thoroughly for the case of the amino acid biosynthetic regulator Gcn4 (Hinnebusch, 1997). Sequences upstream of the *GCN4* transcript were found to be required for normal induction of Gcn4-dependent changes in expression (Hinnebusch, 1984; Thireos et al., 1984). These sequences were found to encode a series of four upstream open reading frames (uORFs). Under conditions of amino acid abundance, the first uORF is translated, and the ribosomal complex re-initiates at a later uORF, but dissociates prior to reaching the translational start site of the *GCN4* ORF itself (Mueller and Hinnebusch, 1986). Low intracellular levels of amino acids, however, trigger a decrease in the rate of ribosomal re-initiation, substantially increasing the likelihood that a scanning ribosomal complex will bypass the intervening uORFs before re-initiating translation at *GCN4* (Hinnebusch, 1997).

C/EBPβ is a member of a family of human transcription factors that control differentiation and proliferation in many cell types (Ramji and Foka, 2002). As with Gcn4, regulation is dependent on translational initiation. Unlike Gcn4, however, regulation of C/EBP proteins results in different isoforms of the protein (Descombes and Schibler, 1991; Ossipow et al., 1993). The utilization of different translational start sites in a single transcript results in the production of both full-length C/EBPβ as well as amino-terminally truncated versions. Both full length and truncated forms contain a DNA-binding domain and retain transcriptional regulatory activity. Interestingly,

however, the full-length version of C/EBPβ (*L*iver *A*ctivating *P*rotein) appears to serve as a transcriptional activator, whereas the smaller protein (*L*iver *I*nhibitory *P*rotein) acts a repressor (Descombes and Schibler, 1991). Mutations that eliminate the capacity for differential translational initiation lead to altered phenotypes (Calkhoven et al., 2000), and different ratios of isoforms have been observed in different tissues and at different times during differentiation (Descombes and Schibler, 1991; Ossipow et al., 1993) implicating the translational control of C/EBP-encoding genes in the determination of tissue-specificity. Given the number of growth regulatory proteins that are subject to transcriptional regulation, it has been proposed that this mechanism may serve as a common means for controlling cell proliferation (Calkhoven et al., 2000).

Translational elongation. In some cases, it is not the binding of the translational initiation factors, but downstream events that control the rate of protein synthesis. While in plants, most cases of regulation by microRNAs result in cleavage of target mRNAs, in animals, the primary mode of control appears to be translational repression (He and Hannon, 2004). One of the first and best-characterized miRNA genes is *lin-4* in *C. elegans*. This microRNA negatively regulates *lin-14*, which encodes a putative transcriptional regulator by binding to multiple complementary regions within its 3' untranslated region (Lau et al., 2001). Regulation by *lin-4* does not affect *lin-14* mRNA abundance or its ability to associate with ribosomes, but does result in decreased synthesis of LIN-14 protein (Olsen and Ambros, 1999), presumably by interfering with productive elongation during translation.

Protein modification

Proteins themselves may exist in multiple states, control over which may involve protein folding, cleavage, covalent modification and degradation. Such modifications may affect protein localization, DNA-binding ability, protein-protein interactions or protein stability, among other properties.

Protein folding. For proteins to function properly, they must be folded into the correct conformation. Steroid hormone receptors convey signals from a wide variety of cellular processes in metazoans. While unactivated "aporeceptor" is generally transcriptionally inert, activated hormone-bound receptor recognizes DNA via a zinc finger binding domain and stimulates transcription of its target genes (DeFranco and Csermely, 2000). Aporeceptor is associated with a group of proteins that includes the molecular chaperone Hsp90 and is called the molecular chaperone-containing heterocomplex (MCH) (Arbeitman and Hogness, 2000). Early studies of vertebrate steroid receptors found that dissociation from this complex was required for transcriptional activation (Howard et al., 1990; Pratt and Toft, 2003). Loss of Hsp90 activity, however, rather than resulting in constitutive activation, actually leads to impaired receptor signaling (Holley and Yamamoto, 1995). The requirement for association with the MCH results from a need for receptor to be maintained in a conformation that is conducive to hormone binding. Once this ligand binding event has taken place, association with MCH is no longer necessary and the receptor is free to bind DNA. While the above model holds for most vertebrate receptors, the role of

association with the MCH is slightly different in the *Drosophila* heterodimeric ecdysone receptor (Arbeitman and Hogness, 2000). Here, association of the receptor with ecdysone does not require the activity of Hsp90, but in vitro binding and in vivo activation does, indicating that the role of the chaperone may regulate the association with DNA as well as ligand binding.

Protein cleavage. Cleavage of a portion of a protein as a mechanism for activation was identified in the study of digestive enzymes, yet may also be employed on transcriptional regulators. Rim101 is a transcriptional regulator in yeast that regulates entry into meiosis as well as the response to changes in intracellular pH. Upon exposure to alkaline growth conditions, Rim101 undergoes C-terminal cleavage that results in its activation (Li and Mitchell, 1997). Proteolytic processing is believed to allow Rim101 to associate with co-regulatory proteins.

Chemical modification. Perhaps the most common type of control mechanism for transcriptional regulators is that of covalent modification of amino acid residues. Indeed, the modification of proteins is frequently a means to enable other types of differential regulation. Often, these modifications are the end result of signaling pathways that translate changes in the cellular environment into altered expression patterns within the nucleus. By far, the best characterized example of chemical modification of regulatory proteins is phosphorylation (Whitmarsh and Davis, 2000). Phosphorylation of serine, threonine, and tyrosine residues is the basis for regulation of the transcriptional regulators associated with MAP ("mitogen-activated" or "microtubule-associated" protein) kinase

cascades. These signaling pathways are found in yeast and higher eukaryotes. The map kinase JNK (c-*J*un *N*H_2-terminal *k*inase) is a downstream mediator of a broad spectrum of environmental signals, from the presence of cytokines to radiation exposure (Davis, 2000; Johnson and Lapadat, 2002). One target of JNK is the transcription factor encoded by the proto-oncogene c-Jun, which regulates growth-dependent genes (Derijard et al., 1994). Phosphorylation of the amino terminal activation domain of c-Jun results in the enhancement of its DNA-binding activity and concomitant increase in is transcription activity.

In addition to phosphorylation, examples of other types of chemical modification include the acetylation of the tumor-suppressor p53, which augments its DNA-binding activity (Gu and Roeder, 1997; Luo et al., 2004) and the methylation of the acute myeloid leukemia factor AML1 (Chakraborty et al., 2003).

Protein stability and degradation. A number of the mechanisms discussed here contribute to increasing the concentration of regulatory proteins. As indicated, however, over-abundance of regulators can lead to dysregulation. Therefore, the regulation of the rate of protein degradation plays a critical role in the control of many transcriptional regulators. The stability of proteins is regulated largely by the process of ubiquitination, a specialized type of modification in which the polypeptide ubiquitin is attached to the targeted protein (Pickart, 2001). The presence of ubiquitin moieties serves as a signal that targets a protein for degradation by the proteasome.

While the synthesis of Gcn4 is under translation control, its steady-state levels are also regulated by protein degradation (Kornitzer et al., 1994). Under non-inducing

conditions, Gcn4 has a very short half life (~5 minutes). Under induced conditions, however, Gcn4 half-life increases to 40 minutes. The change in Gcn4 stability is mediated by phosphorylation of Gcn4 by two different cyclin-dependent kinases (CDKs), namely Pho85 (Meimoun et al., 2000) and Srb10 (Chi et al., 2001). Srb10 was first identified as a component of the Mediator/SRB complex (Hengartner et al., 1995), which has led to the suggestion that Srb10 modifies Gcn4 during transcriptional initiation. Ubiquitination occurs via a transfer of ubiquitin from a ubiquitin-activating enzyme, E1, to a ubiquitin conjugating enzyme, E2, to the substrate protein. Specificity of interactions is conferred by a third component, the ubiquitin ligase, E3 (Pickart, 2001). Phosphorylated Gcn4 is recognized by a ubiquitin ligase called the SCF complex (Meimoun et al., 2000).

Regulation of protein stability appears to be a common mechanism for controlling transcriptional regulators. Other regulators known to be subject to degradation by the attachment of ubiquitin or ubiquitin-like proteins include c-Jun, c-Myc and p53 (Desterro et al., 2000)

Higher order protein states

Translocalization. For transcriptional regulators to exert their effect on the genome, they must be present in sufficient abundance within the nucleus. The controlled sequestration of regulatory proteins within the cytoplasm is an important regulatory mechanism, and allows for rapid changes in gene expression by obviating the time-lag associated with *de novo* synthesis of regulators. The nuclear concentration of the yeast

regulator of stress response, Msn2, increases upon exposure of cells to environmental stresses like changes in osmolarity (Gorner et al., 1998). The retention of Msn2 in the cytoplasm is mediated in part by the TOR (*T*arget *O*f *R*apamycin) signaling pathway, which modifies Msn2 and other transcriptional regulators, increasing their affinity for cytoplasmic binding partners under non-limiting nutrient conditions (Beck and Hall, 1999).

Another model protein whose activity is controlled primarily by nuclear translocation is the human immune response regulator NFκB (Cartwright and Helin, 2000). In unstimulated cells, NFκB is localized to the cytoplasm as a result of binding by an inhibitor (IκB) that masks the former's nuclear localization signal. In cells that have been exposed to immunogenic challenges or to signals like tumor necrosis factor α (TNFα), phosphorylation of IκB targets it for degradation by the proteasome. It is the loss of this inhibitor that allows NFκB to enter the nucleus and bind to its target genes. One of these genes encodes IκB itself, forming a negative feedback loop that re-establishes the cytoplasmic residence of NFκB (Sun et al., 1993).

Molecular cofactors. The binding of a transcriptional regulator does not always correspond to expression changes of its target gene. For example, Leu3, a regulator of leucine biosynthetic genes in yeast, binds to its target sites in a condition-independent manner. Just as the activities of enzymes are often regulated by interactions with small molecules involved in the same biosynthetic pathway, the transcriptional regulatory activity of Leu3 is dependent on its binding a metabolic precursor of leucine, namely alpha-isopropylmalate (Brisco and Kohlhaw, 1990; Friden et al., 1989; Kirkpatrick and

Schimmel, 1995; Sze et al., 1993; Wang et al., 1997). The accumulation of this molecular co-factor due to depletion of leucine and its subsequent binding to Leu3 leads to the conversion of Leu3 from a repressor to a transcriptional activator.

Such regulatory mechanisms are particularly useful in the control of transcriptional programs for specific biochemical pathways. In yeast, this includes regulation of the metabolism of lysine, uracil, and phosphate (Reece, 2000). Interestingly, many of the regulators subject to this type of control are members of the zinc cluster family of transcriptional regulators. In higher eukaryotes, regulation by association with small molecules forms the basis for hormone signaling (discussed in brief above).

Protein-protein interactions. Associations with other regulatory proteins can also exert control over the activity of transcriptional regulators (Remenyi et al., 2004; Wolberger, 1998). Such interactions can influence both the ability of the regulator to activate or repress gene expression, as well as the selection of binding sites themselves. Gal4, a well-studied regulator of galactose metabolism, is like Leu3 in that it can bind DNA under inducing conditions (growth in glucose medium) and non-inducing conditions (growth in galactose medium) (Ren et al., 2000). The ability of Gal4 to activate transcription is regulated by its association with Gal80. The binding of galactose to Gal80 causes it to dissociate from Gal4, alleviating the its repressive effects on Gal4 (Ma and Ptashne, 1987).

The identity of the sites bound by a regulator can be affected by interactions with other transcriptional regulators. Upon exposure of cells to pheromone of the opposite

mating type, Ste12, a regulator of yeast differentiation, binds to specific sites within the promoter regions of genes required for mating. Under conditions that induce a filamentation phenotype, however, Ste12 binds to a different set of sites. Association of Ste12 with these sites requires the DNA-binding regulator Tec1 (Zeitlinger et al., 2003).

For many transcriptional regulatory proteins DNA binding interactions are dependent on the identity of heterodimer partners. In the case of the family of basic helix-loop-helix proteins c-Jun, both homodimers and heterodimers with c-Fos bind to the same DNA sequence (Halazonetis et al., 1988). However, heterodimers bind with increased affinity. In contrast, the α2 regulator of yeast mating type is known to form a heterodimer with either of two other transcriptional regulators. Heterodimers with Mcm1 repress genes whose expression is specific to the a mating type. Heterodimers with the a1 regulator repress haploid-specific gene expression. Mcm1 and a1 help target α2 to distinct binding sites (Li et al., 1995; Miller et al., 1985; Tan and Richmond, 1998). Thus the choice of binding partner can have a profound effect both on the affinity and specificity of protein-DNA interactions.

Cooperative binding. Even in lower eukaryotes, a large fraction of cis-regulatory regions are bound by multiple transcriptional regulators (Harbison et al., 2004; Lee et al., 2002). In metazoans and viruses, the importance of multiple regulatory proteins in the control of specific genes is even more well-documented. Enhancers—cis-regulatory sequences can elevate levels of transcription—of extreme complexity have been discovered to play a role in processes such as *Drosophila* development and human immunoglobulin gene regulation (Banerji et al., 1983; Bowtell et al., 1991). Models of

enhanceosomes, the set and arrangement of proteins assembled at enhancers, hold that their individual components contribute cooperatively to gene-specific transcriptional activation (Thanos and Maniatis, 1995). This cooperativity can result from the increase in binding affinity of one regulator once another has bound. Alternatively, it has been suggested that the three-dimensional surface created by the enhanceosome results in a synergistic improvement in the ability to recruit transcriptional machinery (Carey, 1998; Carey et al., 1990). Experiments that reconstitute in vitro the activators that bind the IFN-β enhancer show that the cooperativity of multiple factors in transcriptional activation depends on the presence of the components of the enhanceosome as well as their precise positioning (Kim and Maniatis, 1997). In summary, the transcriptional activity of one regulator may be augmented as a result of the DNA binding of other regulators, even in the absence of direct interactions between the two.

DNA accessibility. The ability of transcriptional regulators to control changes in gene expression may also be controlled at the level of access to DNA binding sites. As previously discussed, such access may be profoundly affected by changes in the chromatin modification and higher order structures, as well as covalent modifications of DNA (e.g. methylation) that interfere with protein-DNA interaction. Access may also be regulated at the level of competition from other DNA-binding proteins. For example, Sko1 operates as a repressor of stress response genes in yeast. The binding site of Sko1, however, is also recognized by the transcriptional activators Aca1 and Cst6 (Garcia-Gimeno and Struhl, 2000; Vincent and Struhl, 1992). The capacity for regulatory control over the common targets of these regulators is subject to competition for binding to the

same site. Similarly, regulation of the expression of meiotic genes is governed by the antagonistic effects of binding by the Sum1 repressor and the transcriptional activator Ndt80 (Pierce et al., 2003).

Binding site occlusion may also explain the preferences for the position of binding sites relative to transcribed regions. Open reading frames have been shown to contain relatively fewer binding site sequences than intergenic regions, and these sites tend not to be occupied by regulatory proteins (Lieb et al., 2001). Data from our lab also suggest that the presence of binding sites very near the core promoter is disfavored (Harbison et al., 2004). These data are consistent with a model in which the presence of polymerase itself interferes with the binding of transcriptional regulatory proteins.

Conclusions

The importance of precise control over expression of the genome has led to the evolution of highly varied and complex mechanisms governing the control of gene-specific transcriptional regulators in eukaryotes. Although many fall under the categories listed here, there are no doubt other ways by which the activity of transcriptional regulators may be modified. The recent discovery of an apparently important and widespread mechanism like microRNA-mediated regulation indicates that much is yet to be learned about this process.

Our current level of knowledge inspires us to wonder about the roles these mechanisms play for all transcriptional regulators. Although such mechanisms may be diverse, in general they all work by modifying the total amount of regulatory protein, its nuclear localization, its ability to bind DNA, or its capacity to interact with other proteins necessary for transcriptional control. Using information about known mechanisms as a guide, and combining this with information gained about the condition-specific binding behaviors of regulators, we hope to be able to generate models that predict the regulatory mechanisms for each regulator. In the future, the high-throughput acquisition of data on protein concentration, subcellular localization, modification states and likely protein-protein interactions will both inform such models as well as refine them.

References

Albig, A. R., and Decker, C. J. (2001). The target of rapamycin signaling pathway regulates mRNA turnover in the yeast Saccharomyces cerevisiae. Mol Biol Cell *12*, 3428-3438.

Allfrey, V. G., Faulkner, R., and Mirsky, A. E. (1964). Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proc Natl Acad Sci U S A *51*, 786-794.

Ammerer, G., Sprague, G. F., Jr., and Bender, A. (1985). Control of yeast alpha-specific genes: evidence for two blocks to expression in MATa/MAT alpha diploids. Proc Natl Acad Sci U S A *82*, 5855-5859.

Aparicio, O. M., Billington, B. L., and Gottschling, D. E. (1991). Modifiers of position effect are shared between telomeric and silent mating-type loci in S. cerevisiae. Cell *66*, 1279-1287.

Arbeitman, M. N., and Hogness, D. S. (2000). Molecular chaperones activate the Drosophila ecdysone receptor, an RXR heterodimer. Cell *101*, 67-77.

Banerjee, N., and Zhang, M. Q. (2002). Functional genomics as applied to mapping transcription regulatory networks. Curr Opin Microbiol *5*, 313-317.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729-740.

Bannister, A. J., and Kouzarides, T. (1996). The CBP co-activator is a histone acetyltransferase. Nature *384*, 641-643.

Bannister, A. J., and Miska, E. A. (2000). Regulation of gene expression by transcription factor acetylation. Cell Mol Life Sci *57*, 1184-1192.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. Nat Biotechnol *21*, 1337-1342.

Baulcombe, D. (2004). RNA silencing in plants. Nature *431*, 356-363.

Beck, T., and Hall, M. N. (1999). The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. Nature *402*, 689-692.

Bentley, D. L., and Groudine, M. (1986). A block to elongation is largely responsible for decreased transcription of c-myc in differentiated HL60 cells. Nature *321*, 702-706.

Berger, S. L. (2002). Histone modifications in transcriptional regulation. Curr Opin Genet Dev *12*, 142-148.

Berta, P., Hawkins, J. R., Sinclair, A. H., Taylor, A., Griffiths, B. L., Goodfellow, P. N., and Fellous, M. (1990). Genetic evidence equating SRY and the testis-determining factor. Nature *348*, 448-450.

Blander, G., and Guarente, L. (2004). The Sir2 family of protein deacetylases. Annu Rev Biochem *73*, 417-435.

Bowtell, D. D., Lila, T., Michael, W. M., Hackett, D., and Rubin, G. M. (1991). Analysis of the enhancer element that controls expression of sevenless in the developing Drosophila eye. Proc Natl Acad Sci U S A *88*, 6853-6857.

Boxer, L. M., and Dang, C. V. (2001). Translocations involving c-myc and c-myc function. Oncogene *20*, 5595-5610.

Braunstein, M., Rose, A. B., Holmes, S. G., Allis, C. D., and Broach, J. R. (1993). Transcriptional silencing in yeast is associated with reduced nucleosome acetylation. Genes Dev *7*, 592-604.

Breen, T. R., and Harte, P. J. (1993). Trithorax regulates multiple homeotic genes in the bithorax and Antennapedia complexes and exerts different tissue-specific, parasegment-specific and promoter-specific effects on each. Development *117*, 119-134.

Brisco, P. R., and Kohlhaw, G. B. (1990). Regulation of yeast LEU2. Total deletion of regulatory gene LEU3 unmasks GCN4-dependent basal level expression of LEU2. J Biol Chem *265*, 11667-11675.

Brown, A. M., Wildin, R. S., Prendergast, T. J., and Varmus, H. E. (1986). A retrovirus vector expressing the putative mammary oncogene int-1 causes partial transformation of a mammary epithelial cell line. Cell *46*, 1001-1009.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell *71*, 527-542.

Brownell, J. E., Zhou, J., Ranalli, T., Kobayashi, R., Edmondson, D. G., Roth, S. Y., and Allis, C. D. (1996). Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. Cell *84*, 843-851.

Calkhoven, C. F., Muller, C., and Leutz, A. (2000). Translational control of C/EBPalpha and C/EBPbeta isoform expression. Genes Dev *14*, 1920-1932.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5-8.

Carey, M., Lin, Y. S., Green, M. R., and Ptashne, M. (1990). A mechanism for synergistic activation of a mammalian gene by GAL4 derivatives. Nature *345*, 361-364.

Cartwright, P., and Helin, K. (2000). Nucleocytoplasmic shuttling of transcription factors. Cell Mol Life Sci *57*, 1193-1206.

Chakraborty, S., Sinha, K. K., Senyuk, V., and Nucifora, G. (2003). SUV39H1 interacts with AML1 and abrogates AML1 transactivity. AML1 is methylated in vivo. Oncogene *22*, 5229-5237.

Chen, C., and Sytkowski, A. J. (2001). Erythropoietin activates two distinct signaling pathways required for the initiation and the elongation of c-myc. J Biol Chem *276*, 38518-38526.

Chi, Y., Huddleston, M. J., Zhang, X., Young, R. A., Annan, R. S., Carr, S. A., and Deshaies, R. J. (2001). Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. Genes Dev *15*, 1078-1092.

Cosma, M. P. (2002). Ordered recruitment: gene-specific mechanism of transcription activation. Mol Cell *10*, 227-236.

Cuthbert, G. L., Daujat, S., Snowden, A. W., Erdjument-Bromage, H., Hagiwara, T., Yamada, M., Schneider, R., Gregory, P. D., Tempst, P., Bannister, A. J., and Kouzarides, T. (2004). Histone deimination antagonizes arginine methylation. Cell *118*, 545-553.

Dalla-Favera, R., Martinotti, S., Gallo, R. C., Erikson, J., and Croce, C. M. (1983). Translocation and rearrangements of the c-myc oncogene locus in human undifferentiated B-cell lymphomas. Science *219*, 963-967.

Darzacq, X., Powrie, E., Gu, W., Singer, R. H., and Zenklusen, D. (2003). RNA asymmetric distribution and daughter/mother differentiation in yeast. Curr Opin Microbiol *6*, 614-620.

Davis, R. J. (2000). Signal transduction by the JNK group of MAP kinases. Cell *103*, 239-252.

DeFranco, D. B., and Csermely, P. (2000). Steroid receptor and molecular chaperone encounters in the nucleus. Sci STKE *2000*, PE1.

Derijard, B., Hibi, M., Wu, I. H., Barrett, T., Su, B., Deng, T., Karin, M., and Davis, R. J. (1994). JNK1: a protein kinase stimulated by UV light and Ha-Ras that binds and phosphorylates the c-Jun activation domain. Cell *76*, 1025-1037.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

Descombes, P., and Schibler, U. (1991). A liver-enriched transcriptional activator

protein, LAP, and a transcriptional inhibitory protein, LIP, are translated from the same mRNA. Cell *67*, 569-579.

Desterro, J. M., Rodriguez, M. S., and Hay, R. T. (2000). Regulation of transcription factors by protein degradation. Cell Mol Life Sci *57*, 1207-1219.

Diffley, J. F., and Stillman, B. (1989). Similarity between the transcriptional silencer binding proteins ABF1 and RAP1. Science *246*, 1034-1038.

Driever, W., and Nusslein-Volhard, C. (1988). The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. Cell *54*, 95-104.

Dynan, W. S., and Tjian, R. (1985). Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. Nature *316*, 774-778.

Ephrussi, A., and St Johnston, D. (2004). Seeing is believing: the bicoid morphogen gradient matures. Cell *116*, 143-152.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448-453.

Friden, P., Reynolds, C., and Schimmel, P. (1989). A large internal deletion converts yeast LEU3 to a constitutive transcriptional activator. Mol Cell Biol *9*, 4056-4060.

Garcia-Gimeno, M. A., and Struhl, K. (2000). Aca1 and Aca2, ATF/CREB activators in Saccharomyces cerevisiae, are important for carbon source utilization but not the response to stress. Mol Cell Biol *20*, 4340-4349.

Giese, K., Cox, J., and Grosschedl, R. (1992). The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. Cell *69*, 185-195.

Gill, G. (2001). Regulation of the initiation of eukaryotic transcription. Essays Biochem *37*, 33-43.

Gorner, W., Durchschlag, E., Martinez-Pastor, M. T., Estruch, F., Ammerer, G., Hamilton, B., Ruis, H., and Schuller, C. (1998). Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. Genes Dev *12*, 586-597.

Grewal, S. I., and Moazed, D. (2003). Heterochromatin and epigenetic control of gene expression. Science *301*, 798-802.

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature *389*, 349-352.

Gu, W., and Roeder, R. G. (1997). Activation of p53 sequence-specific DNA binding by

acetylation of the p53 C-terminal domain. Cell *90*, 595-606.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nat Struct Mol Biol *11*, 394-403.

Halazonetis, T. D., Georgopoulos, K., Greenberg, M. E., and Leder, P. (1988). c-Jun dimerizes with itself and with c-Fos, forming complexes of different DNA binding affinities. Cell *55*, 917-924.

Hall, M. N., and Johnson, A. D. (1987). Homeo domain of the yeast repressor alpha 2 is a sequence-specific DNA-binding domain but is not sufficient for repression. Science *237*, 1007-1012.

Hamlyn, P. H., and Rabbitts, T. H. (1983). Translocation joins c-myc and immunoglobulin gamma 1 genes in a Burkitt lymphoma revealing a third exon in the c-myc oncogene. Nature *304*, 135-139.

Hammes, A., Guo, J. K., Lutsch, G., Leheste, J. R., Landrock, D., Ziegler, U., Gubler, M. C., and Schedl, A. (2001). Two splice variants of the Wilms' tumor 1 gene have distinct functions during sex determination and nephron formation. Cell *106*, 319-329.

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev *62*, 465-503.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

He, L., and Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet *5*, 522-531.

Hebbes, T. R., Thorne, A. W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. Embo J *7*, 1395-1402.

Hengartner, C. J., Thompson, C. M., Zhang, J., Chao, D. M., Liao, S. M., Koleske, A. J., Okamura, S., and Young, R. A. (1995). Association of an activator with an RNA polymerase II holoenzyme. Genes Dev *9*, 897-910.

Hinnebusch, A. G. (1984). Evidence for translational regulation of the activator of general amino acid control in yeast. Proc Natl Acad Sci U S A *81*, 6442-6446.

Hinnebusch, A. G. (1997). Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. J Biol Chem *272*, 21661-21664.

Holley, S. J., and Yamamoto, K. R. (1995). A role for Hsp90 in retinoid receptor signal transduction. Mol Biol Cell *6*, 1833-1842.

Hombria, J. C., and Lovegrove, B. (2003). Beyond homeosis--HOX function in morphogenesis and organogenesis. Differentiation *71*, 461-476.

Horak, C. E., Luscombe, N. M., Qian, J., Bertone, P., Piccirrillo, S., Gerstein, M., and Snyder, M. (2002). Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. Genes Dev *16*, 3017-3033.

Hoshijima, K., Kohyama, A., Watakabe, I., Inoue, K., Sakamoto, H., and Shimura, Y. (1995). Transcriptional regulation of the Sex-lethal gene by helix-loop-helix proteins. Nucleic Acids Res *23*, 3441-3448.

Howard, K. J., Holley, S. J., Yamamoto, K. R., and Distelhorst, C. W. (1990). Mapping the HSP90 binding region of the glucocorticoid receptor. J Biol Chem *265*, 11928-11935.

Howard, M. L., and Davidson, E. H. (2004). cis-Regulatory control circuits in development. Dev Biol *271*, 109-118.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol *3*, 318-356.

Jager, R. J., Anvret, M., Hall, K., and Scherer, G. (1990). A human XY female with a frame shift mutation in the candidate testis-determining gene SRY. Nature *348*, 452-454.

Jennings, E. G. (2002) Genome-wide expression and location profiling in Saccharomyces cerevisiae: Experimental and graphical analysis, Massachusetts Institute of Technology, Cambridge.

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. Science *293*, 1074-1080.

Johnson, G. L., and Lapadat, R. (2002). Mitogen-activated protein kinase pathways mediated by ERK, JNK, and p38 protein kinases. Science *298*, 1911-1912.

Kelleher, R. J., 3rd, Flanagan, P. M., and Kornberg, R. D. (1990). A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. Cell *61*, 1209-1215.

Kelly, L. M., and Gilliland, D. G. (2002). Genetics of myeloid leukemias. Annu Rev Genomics Hum Genet *3*, 179-198.

Kidner, C. A., and Martienssen, R. A. (2004). Spatially restricted microRNA directs leaf polarity through ARGONAUTE1. Nature *428*, 81-84.

Kim, T. K., and Maniatis, T. (1997). The mechanism of transcriptional synergy of an in vitro assembled interferon-beta enhanceosome. Mol Cell *1*, 119-129.

Kirkpatrick, C. R., and Schimmel, P. (1995). Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. Mol Cell Biol *15*, 4021-4030.

Klar, A. J., Strathern, J. N., Broach, J. R., and Hicks, J. B. (1981). Regulation of transcription in expressed and unexpressed mating type cassettes of yeast. Nature *289*, 239-244.

Koopman, P., Gubbay, J., Vivian, N., Goodfellow, P., and Lovell-Badge, R. (1991). Male development of chromosomally female mice transgenic for Sry. Nature *351*, 117-121.

Kornitzer, D., Raboy, B., Kulka, R. G., and Fink, G. R. (1994). Regulated degradation of the transcription factor Gcn4. Embo J *13*, 6021-6030.

Kurdistani, S. K., and Grunstein, M. (2003). Histone acetylation and deacetylation in yeast. Nat Rev Mol Cell Biol *4*, 276-284.

Latchman, D. S. (1991). Eukaryotic Transcription Factors (San Diego, CA, University Press, Cambridge).

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science *294*, 858-862.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862-864.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lee, T. I., and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet *34*, 77-137.

Li, T., Stark, M. R., Johnson, A. D., and Wolberger, C. (1995). Crystal structure of the MATa1/MAT alpha 2 homeodomain heterodimer bound to DNA. Science *270*, 262-269.

Li, W., and Mitchell, A. P. (1997). Proteolytic activation of Rim1p, a positive regulator of yeast sporulation and invasive growth. Genetics *145*, 63-73.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet *28*, 327-334.

Lin, Y. S., and Green, M. R. (1991). Mechanism of action of an acidic transcriptional activator in vitro. Cell *64*, 971-981.

Loo, S., and Rine, J. (1994). Silencers and domains of generalized repression. Science *264*, 1768-1771.

Lorincz, M. C., Dickerson, D. R., Schmitt, M., and Groudine, M. (2004). Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol.

Luo, J., Li, M., Tang, Y., Laszkowska, M., Roeder, R. G., and Gu, W. (2004). Acetylation of p53 augments its site-specific DNA binding both in vitro and in vivo. Proc Natl Acad Sci U S A *101*, 2259-2264.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308-312.

Ma, J., and Ptashne, M. (1987). The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. Cell *50*, 137-142.

Mallory, A. C., Reinhart, B. J., Jones-Rhoades, M. W., Tang, G., Zamore, P. D., Barton, M. K., and Bartel, D. P. (2004). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. Embo J *23*, 3356-3364.

Meimoun, A., Holtzman, T., Weissman, Z., McBride, H. J., Stillman, D. J., Fink, G. R., and Kornitzer, D. (2000). Degradation of the transcription factor Gcn4 requires the kinase Pho85 and the SCF(CDC4) ubiquitin-ligase complex. Mol Biol Cell *11*, 915-927.

Miller, A. M., MacKay, V. L., and Nasmyth, K. A. (1985). Identification and comparison of two sequence elements that confer cell-type specific transcription in yeast. Nature *314*, 598-603.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824-827.

Mohr, E., and Richter, D. (2001). Messenger RNA on the move: implications for cell polarity. Int J Biochem Cell Biol *33*, 669-679.

Montarras, D., Pinset, C., Chelly, J., Kahn, A., and Gros, F. (1989). Expression of MyoD1 coincides with terminal differentiation in determined but inducible muscle cells. Embo J *8*, 2203-2207.

Mueller, P. P., and Hinnebusch, A. G. (1986). Multiple upstream AUG codons mediate translational control of GCN4. Cell *45*, 201-207.

Nan, X., Ng, H. H., Johnson, C. A., Laherty, C. D., Turner, B. M., Eisenman, R. N., and

Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386-389.

Narlikar, G. J., Fan, H. Y., and Kingston, R. E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell *108*, 475-487.

Nasmyth, K., and Shore, D. (1987). Transcriptional regulation in the yeast life cycle. Science *237*, 1162-1170.

Nasmyth, K. A., Tatchell, K., Hall, B. D., Astell, C., and Smith, M. (1981). A position effect in the control of transcription at yeast mating type loci. Nature *289*, 244-250.

Nasrin, N., Buggs, C., Kong, X. F., Carnazza, J., Goebl, M., and Alexander-Bridges, M. (1991). DNA-binding properties of the product of the testis-determining gene and a related protein. Nature *354*, 317-320.

Novina, C. D., and Sharp, P. A. (2004). The RNAi revolution. Nature *430*, 161-164.

Olsen, P. H., and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol *216*, 671-680.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. Cell *108*, 439-451.

Ossipow, V., Descombes, P., and Schibler, U. (1993). CCAAT/enhancer-binding protein mRNA is translated into multiple proteins with different transcription activation potentials. Proc Natl Acad Sci U S A *90*, 8219-8223.

Parkhurst, S. M., Bopp, D., and Ish-Horowicz, D. (1990). X:A ratio, the primary sex-determining signal in Drosophila, is transduced by helix-loop-helix proteins. Cell *63*, 1179-1191.

Pickart, C. M. (2001). Mechanisms underlying ubiquitination. Annu Rev Biochem *70*, 503-533.

Pierce, M., Benjamin, K. R., Montano, S. P., Georgiadis, M. M., Winter, E., and Vershon, A. K. (2003). Sum1 and Ndt80 proteins compete for binding to middle sporulation element sequences that control meiotic gene expression. Mol Cell Biol *23*, 4814-4825.

Pratt, W. B., and Toft, D. O. (2003). Regulation of signaling protein function and trafficking by the hsp90/hsp70-based chaperone machinery. Exp Biol Med (Maywood) *228*, 111-133.

Ptashne, M. (1967). Specific binding of the lambda phage repressor to lambda DNA. Nature *214*, 232-234.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. Nature *386*, 569-577.

Rabbitts, T. H. (1999). Perspective: chromosomal translocations can affect genes controlling gene expression and differentiation--why are these functions targeted? J Pathol *187*, 39-42.

Ramji, D. P., and Foka, P. (2002). CCAAT/enhancer-binding proteins: structure, function and regulation. Biochem J *365*, 561-575.

Reece, R. J. (2000). Molecular basis of nutrient-controlled gene expression in Saccharomyces cerevisiae. Cell Mol Life Sci *57*, 1161-1171.

Remenyi, A., Scholer, H. R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat Struct Mol Biol *11*, 812-815.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell *110*, 513-520.

Rine, J., and Herskowitz, I. (1987). Four genes responsible for a position effect on expression from HML and HMR in Saccharomyces cerevisiae. Genetics *116*, 9-22.

Rine, J., Strathern, J. N., Hicks, J. B., and Herskowitz, I. (1979). A suppressor of mating-type locus mutations in Saccharomyces cerevisiae: evidence for and identification of cryptic mating-type loci. Genetics *93*, 877-901.

Roeder, R. G., and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Scharnhorst, V., van der Eb, A. J., and Jochemsen, A. G. (2001). WT1 proteins: functions in growth and differentiation. Gene *273*, 141-161.

Schramke, V., and Allshire, R. (2004). Those interfering little RNAs! Silencing and eliminating chromatin. Curr Opin Genet Dev *14*, 174-180.

Schutt, C., and Nothiger, R. (2000). Structure, function and evolution of sex-determining systems in Dipteran insects. Development *127*, 667-677.

Schwab, M., Alitalo, K., Varmus, H. E., Bishop, J. M., and George, D. (1983). A cellular oncogene (c-Ki-ras) is amplified, overexpressed, and located within karyotypic abnormalities in mouse adrenocortical tumour cells. Nature *303*, 497-501.

Sharma, P. M., Bowman, M., Madden, S. L., Rauscher, F. J., 3rd, and Sukumar, S.

(1994). RNA editing in the Wilms' tumor susceptibility gene, WT1. Genes Dev *8*, 720-731.

Sharp, P. A. (1994). Split genes and RNA splicing. Cell *77*, 805-815.

Sil, A., and Herskowitz, I. (1996). Identification of asymmetrically localized determinant, Ash1p, required for lineage-specific transcription of the yeast HO gene. Cell *84*, 711-722.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106*, 697-708.

Simon, J., Chiang, A., Bender, W., Shimell, M. J., and O'Connor, M. (1993). Elements of the Drosophila bithorax complex that mediate repression by Polycomb group products. Dev Biol *158*, 131-144.

Simon, J. A., and Tamkun, J. W. (2002). Programming off and on states in chromatin: mechanisms of Polycomb and trithorax group complexes. Curr Opin Genet Dev *12*, 210-218.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell-Badge, R., and Goodfellow, P. N. (1990). A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature *346*, 240-244.

Sprague, G. F., Jr., Jensen, R., and Herskowitz, I. (1983). Control of yeast cell type by the mating type locus: positive regulation of the alpha-specific STE3 gene by the MAT alpha 1 product. Cell *32*, 409-415.

Strathern, J. N., and Herskowitz, I. (1979). Asymmetry and directionality in production of new cell types during clonal growth: the switching pattern of homothallic yeast. Cell *17*, 371-381.

Struhl, K. (1995). Yeast transcriptional regulatory mechanisms. Annu Rev Genet *29*, 651-674.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell *98*, 1-4.

Sun, S. C., Ganchi, P. A., Ballard, D. W., and Greene, W. C. (1993). NF-kappa B controls expression of inhibitor I kappa B alpha: evidence for an inducible autoregulatory pathway. Science *259*, 1912-1915.

Sze, J. Y., Remboutsika, E., and Kohlhaw, G. B. (1993). Transcriptional regulator Leu3 of Saccharomyces cerevisiae: separation of activator and repressor functions. Mol Cell Biol *13*, 5702-5709.

Tan, S., and Richmond, T. J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature *391*, 660-666.

Tapscott, S. J., Davis, R. L., Thayer, M. J., Cheng, P. F., Weintraub, H., and Lassar, A. B. (1988). MyoD1: a nuclear phosphoprotein requiring a Myc homology region to convert fibroblasts to myoblasts. Science *242*, 405-411.

Taverner, N. V., Smith, J. C., and Wardle, F. C. (2004). Identifying transcriptional targets. Genome Biol *5*, 210.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell *83*, 1091-1100.

Thireos, G., Penn, M. D., and Greer, H. (1984). 5' untranslated sequences are required for the translational control of a yeast regulatory gene. Proc Natl Acad Sci U S A *81*, 5096-5100.

Thompson, C. M., Koleske, A. J., Chao, D. M., and Young, R. A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. Cell *73*, 1361-1375.

Tillib, S., Petruk, S., Sedkov, Y., Kuzin, A., Fujioka, M., Goto, T., and Mazo, A. (1999). Trithorax- and Polycomb-group response elements within an Ultrabithorax transcription maintenance unit consist of closely situated but separable sequences. Mol Cell Biol *19*, 5189-5202.

Tjian, R. (1996). The biochemistry of transcription in eukaryotes: a paradigm for multisubunit regulatory complexes. Philos Trans R Soc Lond B Biol Sci *351*, 491-499.

van Eeden, F., and St Johnston, D. (1999). The polarisation of the anterior-posterior and dorsal-ventral axes during Drosophila oogenesis. Curr Opin Genet Dev *9*, 396-404.

Varmus, H. E. (1984). The molecular genetics of cellular oncogenes. Annu Rev Genet *18*, 553-612.

Vignali, M., Hassan, A. H., Neely, K. E., and Workman, J. L. (2000). ATP-dependent chromatin-remodeling complexes. Mol Cell Biol *20*, 1899-1910.

Vincent, A. C., and Struhl, K. (1992). ACR1, a yeast ATF/CREB repressor. Mol Cell Biol *12*, 5394-5405.

Wagner, K. D., Wagner, N., and Schedl, A. (2003). The complex life of WT1. J Cell Sci *116*, 1653-1658.

Wang, D., Hu, Y., Zheng, F., Zhou, K., and Kohlhaw, G. B. (1997). Evidence that intramolecular interactions are involved in masking the activation domain of transcriptional activator Leu3p. J Biol Chem *272*, 19383-19392.

Weintraub, H., Tapscott, S. J., Davis, R. L., Thayer, M. J., Adam, M. A., Lassar, A. B., and Miller, A. D. (1989). Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of MyoD. Proc Natl Acad Sci U S A *86*, 5434-5438.

Whitmarsh, A. J., and Davis, R. J. (2000). Regulation of transcription factor function by phosphorylation. Cell Mol Life Sci *57*, 1172-1183.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol *15*, 1359-1367.

Wolberger, C. (1998). Combinatorial transcription factors. Curr Opin Genet Dev *8*, 552-559.

Wright, S., and Bishop, J. M. (1989). DNA sequences that mediate attenuation of transcription from the mouse protooncogene myc. Proc Natl Acad Sci U S A *86*, 505-509.

Wyrick, J. J., Holstege, F. C., Jennings, E. G., Causton, H. C., Shore, D., Grunstein, M., Lander, E. S., and Young, R. A. (1999). Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast. Nature *402*, 418-421.

Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science *304*, 594-596.

Zeitlinger, J., Simon, I., Harbison, C. T., Hannett, N. M., Volkert, T. L., Fink, G. R., and Young, R. A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell *113*, 395-404.

Chapter 2

Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*

Published as: Lee, T.-I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., Young, R. A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*. **298**:799-804.

My contribution to this project

The effort to profile the genomic binding locations for over 100 yeast transcriptional regulators grown in nutrient-rich medium was led by Tony Lee, François Robert and Nicola Rinaldi. I was responsible for overseeing production of the microarrays for the project, I contributed binding data for some of these regulators and I performed confirmation of certain protein-DNA interactions. In addition, I conducted analysis on the combined datasets, identified known interactions from the literature and contributed to the content of the resulting publication.

Summary

We have determined how most of the transcriptional regulators encoded in the eukaryote *Saccharomyces cerevisiae* associate with genes across the genome in living cells. Just as maps of metabolic networks describe the potential pathways that may be used by a cell to accomplish metabolic processes, this network of regulator-gene interactions describes potential pathways yeast cells can use to regulate global gene expression programs. We use this information to identify network motifs, the simplest units of network architecture, and demonstrate that an automated process can use motifs to assemble a transcriptional regulatory network structure. Our results reveal that eukaryotic cellular functions are highly connected through networks of transcriptional regulators that regulate other transcriptional regulators.

Introduction

Genome sequences specify the gene expression programs that produce living cells, but how the cell controls global gene expression programs is far from understood. Each cell is the product of specific gene expression programs involving regulated transcription of thousands of genes. These transcriptional programs are modified as cells progress through the cell cycle, in response to changes in environment, and during organismal development (Causton et al., 2001; Cho et al., 1998; DeRisi et al., 1997; Gasch et al., 2000; Spellman et al., 1998).

Gene expression programs depend on recognition of specific promoter sequences by transcriptional regulatory proteins (Garvie and Wolberger, 2001; Lee and Young, 2000; Orphanides and Reinberg, 2002). Because these regulatory proteins recruit and regulate chromatin modifying complexes and components of the transcription apparatus, knowledge of the sites bound by all the transcriptional regulators encoded in a genome can provide the necessary information to nucleate models for transcriptional regulatory networks. With the availability of complete genome sequences and development of a method for genome-wide binding analysis (also known as genome-wide location analysis), investigators can identify the set of target genes bound *in vivo* by each of the transcriptional regulators that are encoded in a cell's genome. This approach has been used to identify the genomic sites bound by nearly a dozen regulators of transcription (Bar-Joseph et al., 2002; Iyer et al., 2001; Lieb et al., 2001; Ren et al., 2000) and several regulators of DNA synthesis (Wyrick et al., 2001) in yeast.

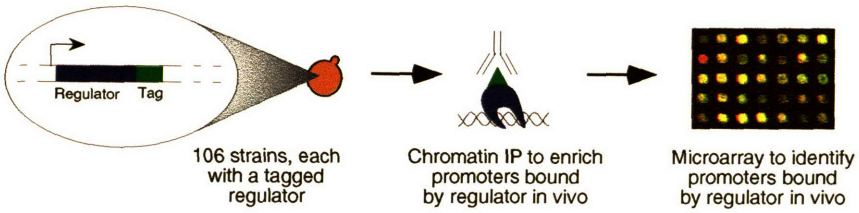
Results

Experimental Design

We have used genome-wide location analysis to investigate how yeast transcriptional regulators bind to promoter sequences across the genome (Fig. 1A). All 141 transcription factors listed in the Yeast Proteome Database (Costanzo et al., 2000) and reported to have DNA-binding and transcriptional activity were selected for study. Yeast strains were constructed such that each of the transcription factors contained a myc epitope tag. To increase the likelihood that tagged factors were expressed at physiologic levels, we introduced epitope tag coding sequences into the genomic sequences encoding the C-terminus of each regulator as described (Knop et al., 1999). The appropriate insertion of the tag and expression of the tagged protein were confirmed by PCR and Western analysis. The introduction of an epitope tag might be expected to affect the function of some transcriptional regulators, and for 17 of the 141 factors, we were not able to obtain viable tagged cells, despite three attempts at tagging each regulator. Not all of the transcriptional regulators were expected to be expressed at detectable levels when yeast cells are grown in rich media, but Western blot analysis showed that 106 of the 124 tagged regulator proteins could be detected under these conditions.

We performed a genome-wide location analysis experiment (Ren et al., 2000) for each of the 106 yeast strains that expressed epitope-tagged regulators. Each tagged strain was grown in three independent cultures in rich medium (the most common experimental condition used with yeast). Genome-wide location data were subjected to quality control filters, normalized, and the ratio of immunoprecipitated to control DNA was determined

A. Systematic Genome-wide Location Analysis of Regulators



B. Influence of P-value Cutoff

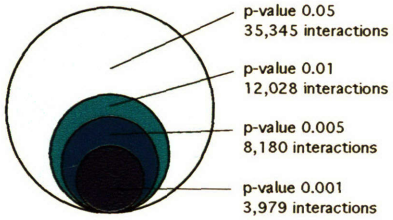


Figure 1. Systematic genome-wide location analysis for yeast transcription regulators.

(A) Methodology. Yeast transcriptional regulators were tagged by introducing the coding sequence for a c-myc epitope tag into the normal genomic locus for each regulator. Of the yeast strains constructed in this fashion, 106 contained a single epitope-tagged regulator whose expression could be detected in rich growth conditions. Chromatin immunoprecipitation (ChIP) was performed on each of these 106 strains. Promoter regions enriched through the ChIP procedure were identified by hybridization to microarrays containing a genome-wide set of yeast promoter regions.

(B) Effect of *P* value threshold. The sum of all regulator-promoter region interactions is displayed as a function of varying *P* value thresholds applied to the entire location dataset for the 106 regulators. More stringent *P* values reduce the number of interactions reported, but decrease the likelihood of false positive results.

for each spot. We calculated a confidence value (*P* value) for each spot from each array using an error model (Hughes et al., 2000). The data for each of the three samples comprising an experiment were combined using a weighted average method (Hughes et al., 2000); each ratio was weighted by *P* value, then averaged. Final *P* values for these combined ratios were then calculated.

Given the properties of the biological system studied here (cell populations, DNA-binding factors capable of binding to both specific and non-specific sequences) and the expectation of noise in microarray-based data, it was important to employ error models to obtain a probabilistic assessment of regulator location data. The total number of protein-DNA interactions in the location analysis dataset, using a range of *P* value thresholds, is shown in Fig. 1B. We selected specific *P* value thresholds to facilitate discussion of a subset of the data at a high confidence level, but note that this artificially imposes a "bound or not bound" binary decision for each protein-DNA interaction.

We generally describe results obtained at a *P* value threshold of 0.001 because our analysis indicates that this threshold maximizes inclusion of legitimate regulator-DNA interactions while minimizing false positives. Various experimental and analytical methods indicate that the frequency of false positives in the genome-wide location data at the 0.001 threshold is 6-10%. For example, conventional, gene-specific chromatin immunoprecipitation experiments have confirmed 93 of 99 binding interactions (involving 29 different regulators) that were identified by location analysis data at a threshold *P* value of 0.001. The use of a high confidence threshold should underestimate the regulator-DNA interactions that actually occur in these cells. We estimate that approximately one-third of the actual regulator-DNA interactions in cells are not

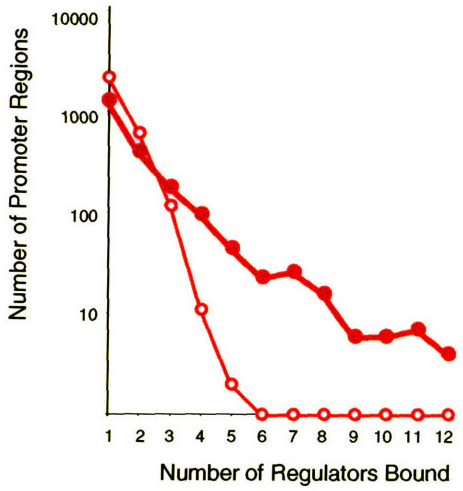
reported at the 0.001 threshold.

Regulator Density

We observed nearly 4000 interactions between regulators and promoter regions at a *P* value threshold of 0.001. The promoter regions of 2343 of 6270 yeast genes (37%) were bound by one or more of the 106 transcriptional regulators in yeast cells grown in rich medium (YPD). Many yeast promoters were bound by multiple transcriptional regulators (Fig. 2A), a feature previously associated with gene regulation in higher eukaryotes (Lemon and Tjian, 2000; Merika and Thanos, 2001), suggesting that yeast genes are also frequently regulated through combinations of regulators. More than one-third of the genes that are bound by regulators were bound by two or more regulators (0.001 *P* value threshold), and a disproportionately high number of promoter regions were bound by four or more regulators when compared to the expected distribution from randomized data. Due to the stringency of the *P* value threshold, we expect that this represents an underestimate.

The number of different promoter regions bound by each regulator in cells grown in rich media ranged from 0 to 181 (0.001 *P* value threshold), with an average of 38 promoter regions per regulator (Fig. 2B). The regulator Abf1 bound the greatest number of promoter regions (181). Regulators that should be active under growth conditions other than YPD were typically found, as expected, to bind the smallest number of promoter regions. For example, Thi2, which activates transcription of thiamine biosynthesis genes under conditions of thiamine starvation (Kawasaki et al., 1990; Nishimura et al., 1992), was among the regulators that bound the smallest number of

A. Number of Regulators Bound Per Promoter Region



B. Number of Promoter Regions Bound Per Regulator

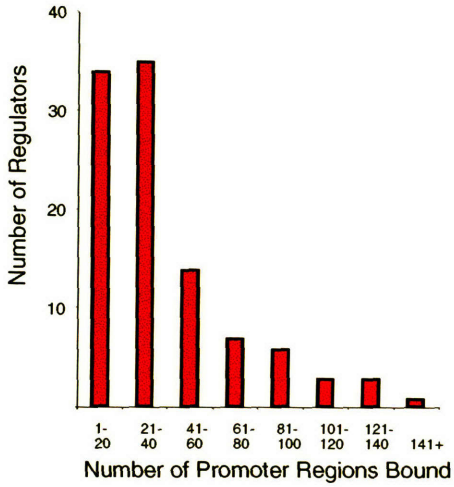


Figure 2. Genome-wide distribution of transcriptional regulators.

(A) A plot of the number of regulators bound per promoter region. The distribution for the actual location data (red circles) is shown alongside the distribution expected from the same set of *P* values randomly assigned among regulators and intergenic regions (white circles). At a *P* value threshold of 0.001, significantly more intergenic regions bind 4 or more regulators than expected by chance.

(B) Distribution of the number of promoter regions bound per regulator.

promoters (3). The identification of a set of promoter regions that are bound by specific regulators allowed us to predict sequence motifs that are bound by these regulators.

Network Motifs

The simplest units of commonly used transcriptional regulatory network architecture, or network motifs, provide specific regulatory capacities such as positive and negative feedback loops. We used the genome-wide location data to identify six different regulatory network motifs: autoregulation, multi-component loops, feedforward loops, single input, multi-input and regulator chains (Fig. 3). These motifs suggest models for regulatory mechanisms that can be tested. Descriptions of the algorithms used to identify motifs and a complete compilation of motifs can be obtained at http://web.wi.mit.edu/young/regulator_network.

Autoregulation motifs. An autoregulation motif consists of a regulator that binds to the promoter region of its own gene. Ten autoregulation motifs were identified using genome-wide location data for the 106 regulators (0.001 *P* value threshold), suggesting that approximately 10% of yeast genes encoding regulators are autoregulated. This percentage does not change significantly at less stringent *P* value thresholds. In contrast, studies of *E. coli* genetic regulatory networks indicate that the majority (52-74%) of prokaryotic genes encoding transcriptional regulators are autoregulated (Shen-Orr et al., 2002; Thieffry et al., 1998).

Autoregulation is thought to provide several selective growth advantages, including reduced response time to environmental stimuli, decreased biosynthetic cost of regulation, and increased stability of gene expression (Becskei and Serrano, 2000;

Guelzim et al., 2002; McAdams and Arkin, 1997; Shen-Orr et al., 2002; Thieffry et al., 1998). For example, upon exposure to mating pheromone, the levels of the pheromone responsive Ste12 transcriptional regulator rapidly increase because Ste12 binds to and upregulates its own gene (Dolan and Fields, 1990; Ren et al., 2000) (Fig. 3). The consequent increase in Ste12 protein leads to the binding of other genes required for the mating process (Ren et al., 2000).

Multi-component loop motifs. A multi-component loop motif consists of a regulatory circuit whose closure involves two or more factors (Fig. 3). Three multi-component loop motifs were observed in the location data for 106 regulators (0.001 *P* value threshold). The closed loop structure provides the capacity for feedback control and offers the potential to produce bistable systems that can switch between two alternative states (Ferrell, 2002). The multi-component loop motif has yet to be identified in bacterial genetic networks (Shen-Orr et al., 2002; Thieffry et al., 1998).

Feedforward loop motifs. Feedforward loop motifs contain a regulator that controls a second regulator, and have the additional feature that both regulators bind a common target gene (Fig. 3). The regulator location data reveal that feedforward loop architecture has been highly favored during the evolution of transcriptional regulatory networks in yeast. We found that 36 regulators are involved in 45 feedforward loops potentially controlling 536 genes in the yeast network (approximately 25% of genes that are bound in the genome-wide location dataset).

In principle, a feedforward loop can provide several features to a regulatory circuit. The feedforward loop may act as a switch that is designed to be sensitive to sustained rather than transient inputs (Shen-Orr et al., 2002). Feedforward loops have

Examples of Network Motifs in the Yeast Regulatory Network

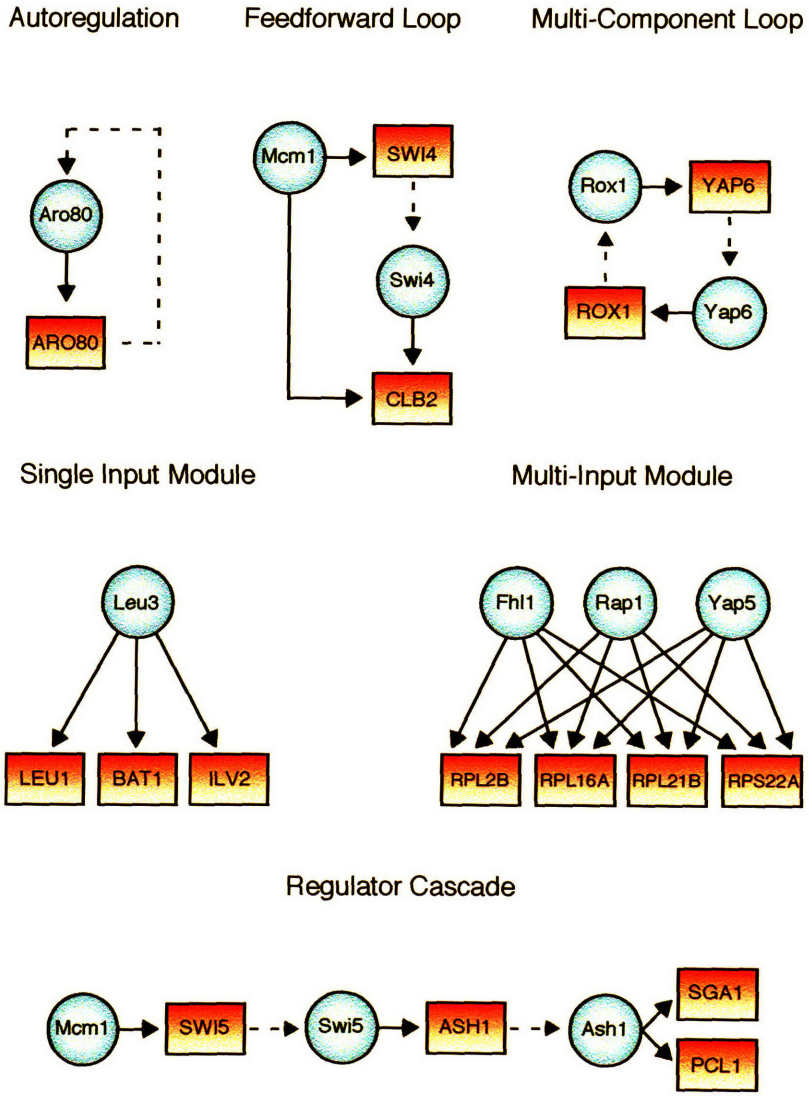


Figure 3. Examples of network motifs in the yeast regulatory network.

Regulators are represented by blue circles, and gene promoters are represented by red rectangles. Binding of a regulator to a promoter is indicated by a solid arrow. Genes encoding regulators are linked to their respective regulators by dashed arrows. For example, in the autoregulation motif, the Ste12 protein binds to the *STE12* gene, which is transcribed and translated into Ste12 protein. These network motifs were uncovered by searching binding data with various algorithms. For details on the algorithms used, and a full list of motifs found, see http://web.wi.mit.edu/young/regulator_network.

the potential to provide temporal control of a process because expression of the ultimate target gene may depend on the accumulation of adequate levels of the master and secondary regulators. Feedforward loops may provide a form of multistep ultrasensitivity (Goldbeter and Koshland, 1984) as small changes in the level/activity of the master regulator at the top of the loop might be amplified at the ultimate target gene due to the combined action of the master regulator and a second regulator that is under the control of the master regulator.

Single input motifs. Single input motifs contain a single regulator that binds a set of genes under a specific condition. Single input motifs are potentially useful for coordinating a discrete unit of biological function such as a set of genes that code for the subunits of a biosynthetic apparatus or enzymes of a metabolic pathway. For example, several genes of the leucine biosynthetic pathway are controlled by the Leu3 transcriptional regulator (Fig. 3).

Multi-input motifs. Multi-input motifs consist of a set of regulators that bind together to a set of genes. We found 181 combinations of two or more regulators that could bind to a common set of promoter regions. This motif offers the potential for coordination of gene expression across a wide variety of growth conditions. For example, each of the regulators bound to a set of genes can be responsible for regulating those genes in response to a unique input. In this manner, two different regulators responding to two different inputs would allow coordinate expression of the set of genes under these two different conditions.

Regulator chain motifs. Regulator chain motifs consist of chains of three or more regulators in which one regulator binds the promoter for a second regulator, the second

binds to the promoter for a third regulator, and so forth (Fig. 3). This network motif is observed frequently in the location data for yeast regulators; we found 188 regulator chain motifs, which varied in size from 3 to 10 regulators. The chain represents the simplest circuit logic for ordering transcriptional events in a temporal sequence. The most straightforward form of this appears in the regulatory circuit of the cell cycle where regulators functioning at one stage of the cell cycle regulate the expression of factors required for entry into the next stage of the cell cycle (Simon et al., 2001).

Motifs suggest models for regulation. The regulatory motifs described above suggest models for gene regulatory mechanisms whose predictions can be tested with experimental data. One regulatory motif that caught our attention involved ribosomal protein genes; ribosomes are important protein biosynthetic machines, but transcriptional regulation of ribosomal protein genes is not well understood. Fhl1, a protein whose function was not previously known, forms a single-input regulatory motif consisting of essentially all ribosomal protein genes, but little else. No other regulator studied here exhibited this behavior. This predicts that loss of Fhl1 function should have a profound effect on ribosome biosynthesis if no other regulators are capable of taking its place. Indeed, a mutation in Fhl1 causes severe defects in ribosome biosynthesis (Hermann-Le Denmat et al., 1994), an observation that was difficult to interpret previously in the absence of the genome-wide location data. Many ribosomal protein genes are also components of a multi-input motif involving Fhl1 and additional regulators (Fig. 3), suggesting that expression of these genes may be coordinated by multiple regulators under various growth conditions. This model and others suggested by regulatory motifs can be addressed with future experiments.

Assembling Motifs into Network Structures

We assume that regulatory network motifs form building blocks that can be combined into larger network structures. An algorithm was developed that explores all the genome-wide location data together with the expression data from over 500 expression experiments to identify groups of genes that are both coordinately bound and coordinately expressed. In brief, the algorithm begins by defining a set of genes, *G*, that are bound by a set of regulators *S*, using the 0.001 *P* value threshold. We find a large subset of genes in *G* that are similarly expressed over the entire set of expression data, and use those genes to establish a core expression profile. Genes are then dropped from *G* if their expression profile is significantly different from this core profile. The remainder of the genome is scanned for genes with expression profiles that are similar to the core profile. Genes with a significant match in expression profiles are then examined to see if the set of regulators *S* are bound. At this step, the probability of a gene being bound by the set of regulators is used, rather than the individual probabilities of that gene being bound by each of the individual regulators. Since we are assaying the combined probability of the set of regulators being bound, and are relying on similarity of expression patterns, we can relax the *P* value for individual binding events and thus recapture information that is lost due to the use of an arbitrary *P* value threshold. The process is repeated until all combinations of genes bound by regulators have been considered. Additional details of the algorithm are available upon request. The resulting sets of regulators and genes are essentially multi-input motifs refined for common expression (MIM-CE). We expect these to be robust examples of coordinate binding

and expression and therefore useful for nucleating network models.

The refined motifs were used to construct a network structure for the yeast cell cycle using an automatic process that requires no prior knowledge of the regulators that control transcription during the cell cycle. The cell cycle regulatory network was selected because of the importance of this biological process, the availability of extensive genome-wide expression data for the cell cycle (Cho et al., 1998; Spellman et al., 1998) and the extensive literature that can be used to explore features of a network model. Our goal was to determine whether the computational approach would construct the regulatory logic of cell cycle from the location and expression data without previous knowledge of the regulators involved. We reasoned that MIM-CEs that are significantly enriched in genes whose expression oscillates through the cell cycle (Spellman et al., 1998) would identify the regulators that control these genes. Eleven regulators were identified by this approach. To construct the cell cycle network, a new set of MIM-CEs was generated using only the eleven regulators and the cell cycle expression data (Spellman et al., 1998).

To produce a cell cycle transcriptional regulatory network model, the MIM-CEs were aligned around the cell cycle on the basis of peak expression of the genes in the group by means of an algorithm described previously (Bar-Joseph et al., 2002) (Fig. 4). Three features of the resulting network model are notable. First, the computational approach correctly assigned all of the regulators to stages of the cell cycle where they were shown to function in previous studies (Simon et al., 2001). Second, two regulators that have been implicated in cell cycle control but whose functions were ill-defined

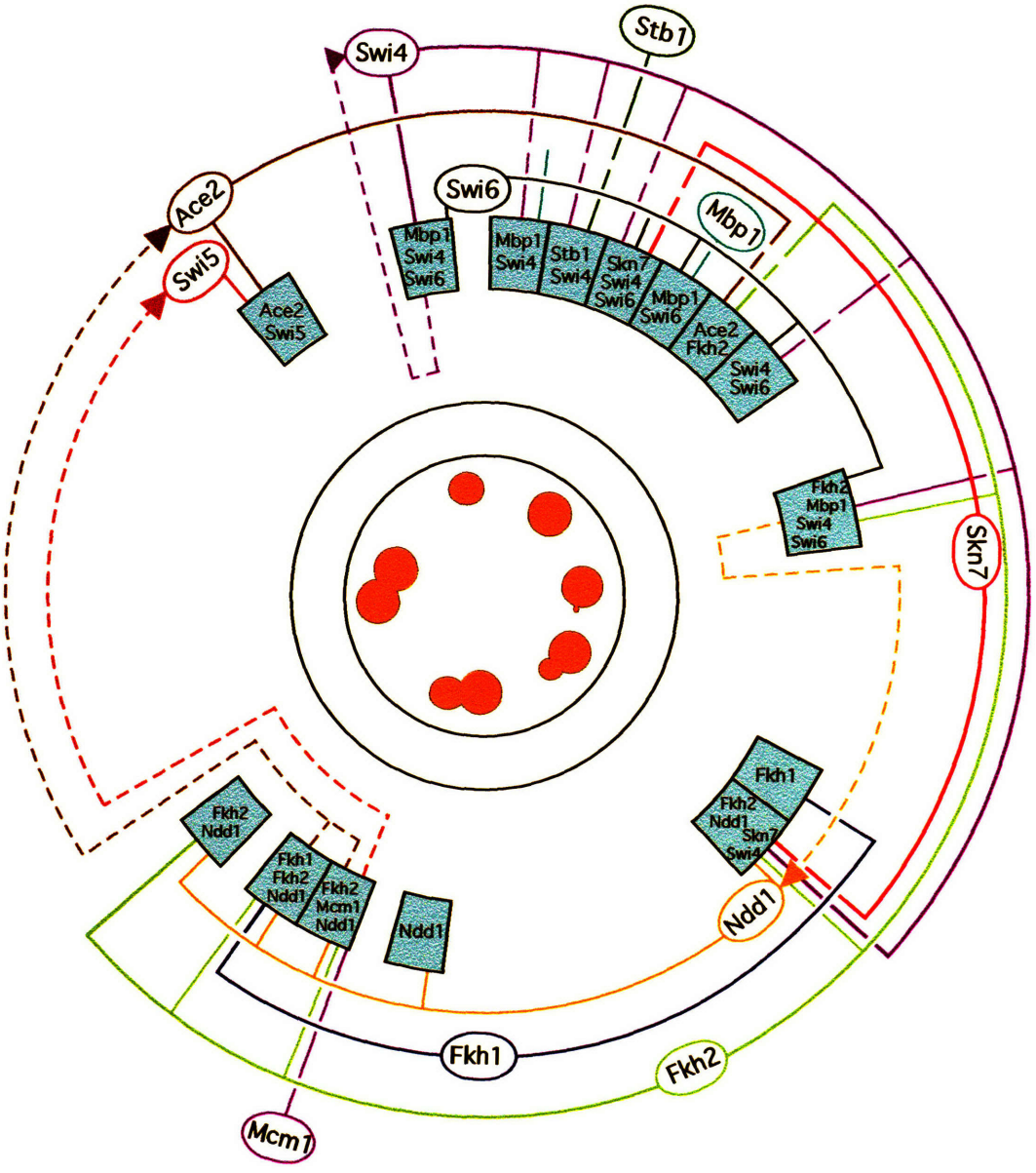

Swi4
Stb1
Ace2
Swi6
Mbp1
Swi5
Ace2
Swi5
Mbp1
Swi4
Swi6
Mbp1
Swi4
Stb1
Swi4
Skn7
Swi4
Swi6
Mbp1
Swi6
Ace2
Fkh2
Swi4
Swi6
Fkh2
Mbp1
Swi4
Swi6
Skn7
Fkh1
Fkh2
Ndd1
Skn7
Swi4
Ndd1
Fkh2
Ndd1
Fkh1
Fkh2
Ndd1
Fkh2
Mcm1
Ndd1
Ndd1
Fkh1
Fkh2
Mcm1

Figure 4. Model for the yeast cell cycle transcriptional regulatory network.

A transcriptional regulatory network for the yeast cell cycle was derived from a combination of binding and expression data as described in the text. Yeast cell morphologies are depicted during the various stages of the cell cycle. Each blue box represents a set of genes that are bound by a common set of regulators and co-expressed throughout the cell cycle. The text inside each blue box identifies the common set of regulators that bind to the set of genes represented by the box. Each box is positioned in the cell cycle according to the time of peak expression levels for the genes represented by the box. Regulators, represented by ovals, are connected to the sets of genes they regulate by solid lines. The arc associated with each regulator effectively defines the period of activity for the regulator. Dashed lines indicate that a gene in the box encodes a regulator found in the outer rings.

(Bouquin et al., 1999; Morgan et al., 1995), could be assigned within the network on the basis of direct binding data. Third, and most importantly, the reconstruction of the regulatory architecture was automatic and required no prior knowledge of the regulators that control transcription during the cell cycle. This approach should represent a general method for constructing other regulatory networks.

Coordination of Cellular Processes

Transcriptional regulators were often bound to genes encoding other transcriptional regulators (Fig. 5). For example, there were many instances in which transcriptional regulators within a functional category (e.g., cell cycle) bound to genes encoding regulators within the same category. We have noted that cell cycle regulators bound to other cell cycle regulators (Simon et al., 2001), and this phenomenon was also apparent among transcriptional regulators that fall into the metabolism and environmental response categories. For example, the metabolic regulator Gcn4 bound to promoters for *PUT3* and *UGA3*, genes that encode transcriptional regulators for amino acid and other metabolic functions. The stress response activator Yap6 bound to the gene encoding the Rox1 repressor, and vice versa, suggesting positive and negative feedback loops.

We also found that multiple transcriptional regulators within each category were able to bind to genes encoding regulators that are responsible for control of other cellular processes. For example, the cell cycle activators bind to genes for transcriptional regulators that play key roles in metabolism (*GAT1*, *GAT3, NRG1, SFL1*), environmental responses (*ROX1*, *YAP1*, *ZMS1*), development (*ASH1*, *SOK2, MOT3*), and DNA, RNA and protein biosynthesis *(ABF1*). These observations are likely to explain, in part, how

Diverse Cellular Functions are Connected Through Transcriptional Networks

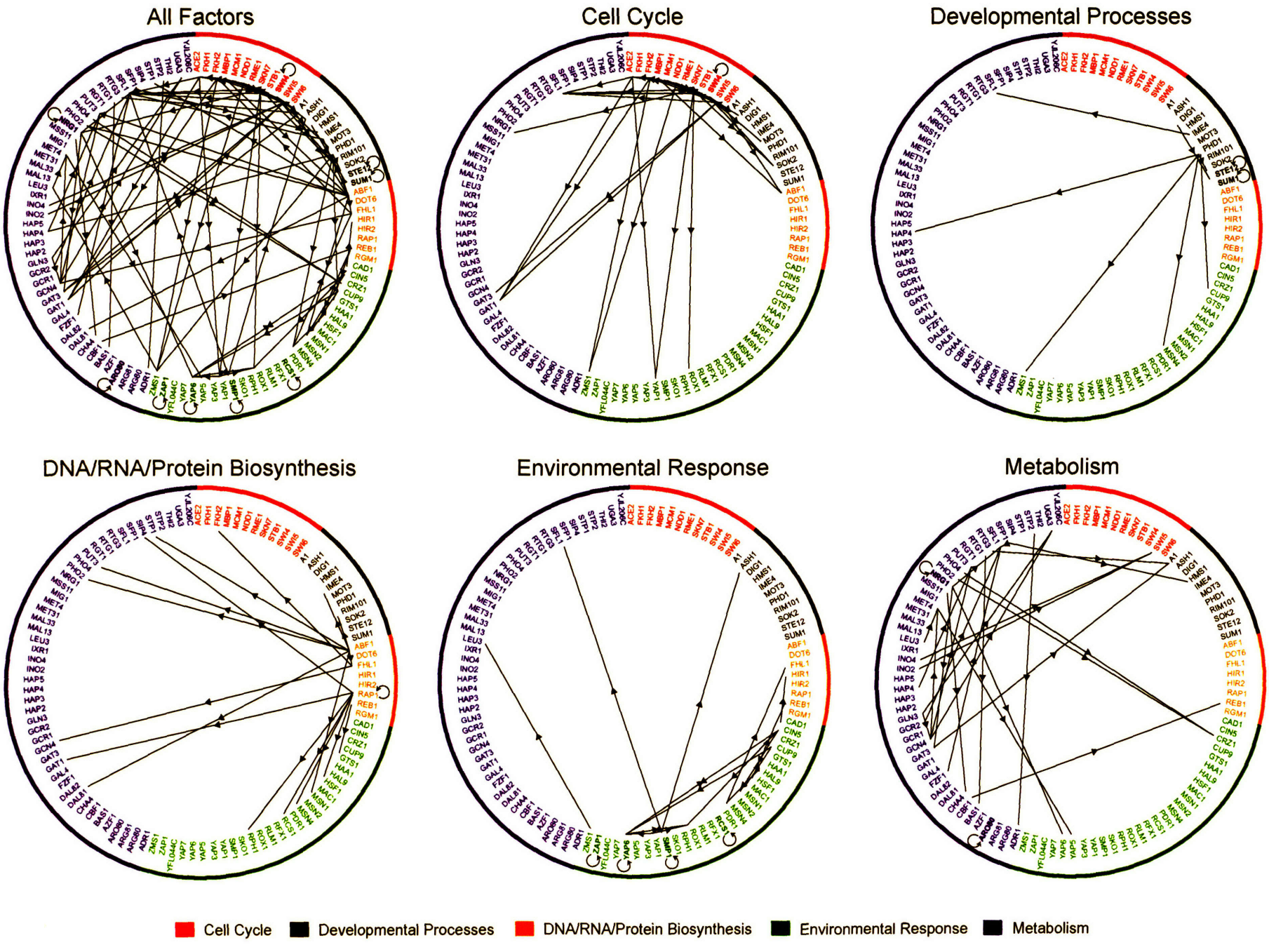


Figure 5. Network of transcriptional regulators binding to genes encoding other transcriptional regulators.

All 106 transcriptional regulators that were subjected to location analysis in rich media are displayed in a circle, segregated into functional categories based on the primary functions of their target genes (Cell Cycle in red, Development in black, DNA/RNA/Protein Biosynthesis in tan, Environmental Response in green, and Metabolism in blue). Lines with arrows depict binding of a regulator (0.001 *P* value threshold) to the gene encoding another regulator. Circles with arrows depict binding of a regulator to the promoter region of its own gene.

cells coordinate transcriptional regulation of the cell cycle with other cellular processes. These connections are generally consistent with previous experimental information regarding the relationships between cellular processes. For example, the developmental regulator Phd1 has been shown to regulate genes involved in pseudohyphal growth during certain nutrient stress conditions; we found that Phd1 also binds to genes that are key to regulation of general stress responses (*MSN4*, *CUP9* and *ZMS1*) and metabolism (*HAP4*).

These observations have several important implications. The control of most, if not all, cellular processes is characterized by networks of transcriptional regulators that regulate other regulators. It is also evident that the effects of transcriptional regulator mutations on global gene expression as measured by expression profiling (Causton et al., 2001; Chu et al., 1998; DeRisi et al., 1997; Devaux et al., 2002; Epstein et al., 2001; Gasch et al., 2000; Ho et al., 1999; Hughes et al., 2000; Jelinsky and Samson, 1999; Lopez and Baker, 2000; Lyons et al., 2000; Madhani et al., 1999; Natarajan et al., 2001; Roberts et al., 2000; Shamji et al., 2000; Travers et al., 2000) are as likely to reflect the effects of the network of regulators as they are to identify the direct targets of a single regulator.

Discussion

This study identified network motifs that provide specific regulatory capacities for yeast, revealing the regulatory strategies that were selected during evolution for this eukaryote. These motifs can be used as building blocks to construct large network structures through an automated approach that combines genome-wide location and expression data in the absence of prior knowledge of regulator functions. The network of transcriptional regulators that control other transcriptional regulators is highly connected, suggesting that the network substructures for cellular functions such as cell cycle and development are themselves coordinated at a transcriptional level.

It is possible to envision mapping the regulatory networks that control gene expression programs in considerable depth in yeast and in other living cells. More complete understanding of transcriptional regulatory networks in yeast will require knowledge of regulator binding sites under various growth conditions and experimental testing of models that emerge from computational analysis of regulator binding, gene expression and other information. The approach described here can also be used to discover transcriptional regulatory networks in higher eukaryotes. Knowledge of these networks will be important for understanding human health and designing new strategies to combat disease.

Methods

Additional information about the methods used as well as supporting online material is available at the authors' website: http://web.wi.mit.edu/young/regulator_network.

Acknowledgments

We thank J. Terragni, D. Pokholok and E. Kanin for experimental support. We thank J. Liang for assistance with the supporting website. We thank R. DeShaies, T. Ideker, and T. Jaakkola for helpful discussions. Supported by fellowships from National Cancer Institute of Canada (F.R.), Human Frontier Science Program (J.Z.), Sloan/DOE Program for Computational Molecular Biology (D.O.), Program in Mathematics and Molecular Biology at Florida State University and Burroughs Wellcome Fund Interfaces Program (Z.B.J.), Helen Hay Whitney foundation (B.R.). Supported by grants from the National Institutes of Health and Corning, Inc.

References

Bar-Joseph, Z., Gerber, G. K., Gifford, D. K., Jaakkola, T. S., and Simon, I. (2002). A new approach to analyzing gene expression time series data. Paper presented at: Sixth Annual International Conference on Research in Computational Molecular Biology.

Becskei, A., and Serrano, L. (2000). Engineering stability in gene networks by autoregulation. Nature *405*, 590-593.

Bouquin, N., Johnson, A. L., Morgan, B. A., and Johnston, L. H. (1999). Association of the cell cycle transcription factor Mbp1 with the Skn7 response regulator in budding yeast. Mol Biol Cell *10*, 3389-3400.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell *12*, 323-337.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell *2*, 65-73.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. Science *282*, 699-705.

Costanzo, M. C., Hogan, J. D., Cusick, M. E., Davis, B. P., Fancher, A. M., Hodges, P. E., Kondu, P., Lengieza, C., Lew-Smith, J. E., Lingner, C., *et al.* (2000). The yeast proteome database (YPD) and Caenorhabditis elegans proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information. Nucleic Acids Res *28*, 73-76.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

Devaux, F., Carvajal, E., Moye-Rowley, S., and Jacq, C. (2002). Genome-wide studies on the nuclear PDR3-controlled response to mitochondrial dysfunction in yeast. FEBS Lett *515*, 25-28.

Dolan, J. W., and Fields, S. (1990). Overproduction of the yeast STE12 protein leads to constitutive transcriptional induction. Genes Dev *4*, 492-502.

Epstein, C. B., Waddle, J. A., Hale, W. t., Dave, V., Thornton, J., Macatee, T. L., Garner, H. R., and Butow, R. A. (2001). Genome-wide responses to mitochondrial dysfunction. Mol Biol Cell *12*, 297-308.

Ferrell, J. E., Jr. (2002). Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. Curr Opin Cell Biol *14*, 140-148.

Garvie, C. W., and Wolberger, C. (2001). Recognition of specific DNA sequences. Mol Cell *8*, 937-946.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell *11*, 4241-4257.

Goldbeter, A., and Koshland, D. E., Jr. (1984). Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. J Biol Chem *259*, 14441-14447.

Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet *31*, 60-63.

Hermann-Le Denmat, S., Werner, M., Sentenac, A., and Thuriaux, P. (1994). Suppression of yeast RNA polymerase III mutations by FHL1, a gene coding for a fork head protein involved in rRNA processing. Mol Cell Biol *14*, 2905-2913.

Ho, Y., Costanzo, M., Moore, L., Kobayashi, R., and Andrews, B. J. (1999). Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1, a Swi6-binding protein. Mol Cell Biol *19*, 5267-5278.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., *et al.* (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109-126.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.

Jelinsky, S. A., and Samson, L. D. (1999). Global response of Saccharomyces cerevisiae to an alkylating agent. Proc Natl Acad Sci U S A *96*, 1486-1491.

Kawasaki, Y., Nosaka, K., Kaneko, Y., Nishimura, H., and Iwashima, A. (1990). Regulation of thiamine biosynthesis in Saccharomyces cerevisiae. J Bacteriol *172*, 6145-6147.

Knop, M., Siegers, K., Pereira, G., Zachariae, W., Winsor, B., Nasmyth, K., and Schiebel, E. (1999). Epitope tagging of yeast genes using a PCR-based strategy: more tags and improved practical routines. Yeast *15*, 963-972.

Lee, T. I., and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. Annu Rev Genet *34*, 77-137.

Lemon, B., and Tjian, R. (2000). Orchestrated response: a symphony of transcription

factors for gene control. Genes Dev *14*, 2551-2569.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet *28*, 327-334.

Lopez, M. C., and Baker, H. V. (2000). Understanding the growth phenotype of the yeast gcr1 mutant in terms of global genomic expression patterns. J Bacteriol *182*, 4970-4978.

Lyons, T. J., Gasch, A. P., Gaither, L. A., Botstein, D., Brown, P. O., and Eide, D. J. (2000). Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. Proc Natl Acad Sci U S A *97*, 7957-7962.

Madhani, H. D., Galitski, T., Lander, E. S., and Fink, G. R. (1999). Effectors of a developmental mitogen-activated protein kinase cascade revealed by expression signatures of signaling mutants. Proc Natl Acad Sci U S A *96*, 12530-12535.

McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression. Proc Natl Acad Sci U S A *94*, 814-819.

Merika, M., and Thanos, D. (2001). Enhanceosomes. Curr Opin Genet Dev *11*, 205-208.

Morgan, B. A., Bouquin, N., and Johnston, L. H. (1995). Two-component signal-transduction systems in budding yeast MAP a different pathway? Trends Cell Biol *5*, 453-457.

Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G., and Marton, M. J. (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol Cell Biol *21*, 4347-4368.

Nishimura, H., Kawasaki, Y., Kaneko, Y., Nosaka, K., and Iwashima, A. (1992). Cloning and characteristics of a positive regulatory gene, THI2 (PHO6), of thiamin biosynthesis in Saccharomyces cerevisiae. FEBS Lett *297*, 155-158.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. Cell *108*, 439-451.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Roberts, C. J., Nelson, B., Marton, M. J., Stoughton, R., Meyer, M. R., Bennett, H. A., He, Y. D., Dai, H., Walker, W. L., Hughes, T. R., *et al.* (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science *287*, 873-880.

Shamji, A. F., Kuruvilla, F. G., and Schreiber, S. L. (2000). Partitioning the

transcriptional program induced by rapamycin among the effectors of the Tor proteins. Curr Biol *10*, 1574-1581.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet *31*, 64-68.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106*, 697-708.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273-3297.

Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998). From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays *20*, 433-440.

Travers, K. J., Patil, C. K., Wodicka, L., Lockhart, D. J., Weissman, J. S., and Walter, P. (2000). Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. Cell *101*, 249-258.

Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P., and Aparicio, O. M. (2001). Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. Science *294*, 2357-2360.

Chapter 3

Transcriptional Regulatory Code of a Eukaryotic Genome

Published as: Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., Jennings, E. G., Zeitlinger, J., Pokholok, D. K., Kellis, M., Rolfe, P. A., Takusagawa, K. T., Lander, E. S., Gifford, D. K., Fraenkel, E., Young, R. A. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*. **431**:99-104.

My contributions to this project

The following work includes data from a number of projects in the lab aimed at understanding the effect of changes in environmental growth conditions on the genomic binding of transcriptional regulators. It also builds on our previous study by including data on as many known and putative transcriptional regulators as possible (for a total of 203). The data for all of these experiments were generated by myself and by eight other members of the lab. To identify the most likely binding specificities of these regulators, we collaborated with Ernest Fraenkel and Ben Gordon in the computational work of combining our data with information on phylogenetic conservation and performing motif discovery. I coordinated this project with assistance from Tony Lee in our lab, overseeing data generation, analysis and the publication of the results.

Supplementary Material for this work is presented as Appendix A.

Summary

DNA-binding transcriptional regulators interpret the genome's regulatory code by binding to specific sequences to induce or repress gene expression (Jacob and Monod, 1961). Comparative genomics has recently been used to identify potential cis-regulatory sequences within the yeast genome on the basis of phylogenetic conservation (Blanchette and Tompa, 2003; Cliften et al., 2003; Kellis et al., 2003; Pritsker et al., 2004; Wang and Stormo, 2003), but this information alone does not reveal if or when transcriptional regulators occupy these binding sites. We have constructed an initial version of yeast's transcriptional regulatory code by mapping the sequence elements that are bound by regulators under various conditions and that are conserved among *Saccharomyces* species. The organization of regulatory elements in promoters and the environment-dependent use of these elements by regulators are discussed. We find that environment-specific use of regulatory elements predicts mechanistic models for the function of a large population of yeast's transcriptional regulators.

Results and Discussion

We used genome-wide location analysis (Iyer et al., 2001; Lee et al., 2002; Lieb et al., 2001; Ren et al., 2000) to determine the genomic occupancy of 203 DNA-binding transcriptional regulators in rich media conditions and, for 84 of these regulators, in at least one of twelve other environmental conditions (Supplementary Table 1, Supplementary Figure 1, http://web.wi.mit.edu/young/regulatory_code). These 203 proteins are likely to include nearly all of the DNA-binding transcriptional regulators encoded in the yeast genome. Regulators were selected for profiling in an additional environment if they were essential for growth in that environment or if there was other evidence implicating them in regulation of gene expression in that environment. The genome-wide location data identified 11,000 unique interactions between regulators and promoter regions at high confidence ($P \leq 0.001$).

To identify the cis-regulatory sequences that likely serve as recognition sites for transcriptional regulators, we merged information from genome-wide location data, phylogenetically conserved sequences, and prior knowledge (Figure 1a). We used six motif discovery programs (Bailey and Elkan, 1995; Liu et al., 2002; Roth et al., 1998) to discover 68,279 DNA sequence motifs for the 147 regulators that bound more than ten probes (Supplementary Methods; Supplementary Figure 2). From these motifs we derived the most likely specificity for each regulator through clustering and stringent statistical tests. This motif discovery process identified highly significant ($P \leq 0.001$) motifs for each of 116 regulators. We determined a single high-confidence motif for 65 of these regulators using additional criteria including the requirement for conservation

across three of four related yeast species. Examples of novel and "re-discovered" motifs are depicted in Figure 1b, and comparisons of the discovered motifs to those described previously are shown in Supplementary Table 2. The discovered motifs provide significantly more information than was previously available; for 21 of the regulators there was no prior specificity information in the literature, and detailed probability matrices had previously been determined for only 17 regulators for which we report motifs (Knuppel et al., 1994). In the case of Cin5, which showed the largest difference between the computationally derived motif (TTACRTAA) and the previously reported site (TTACTAA; Supplementary Table 2), we found that the motif that we report is also the preferred in vitro target (Supplementary Figure 3). We supplemented the discovered motifs with additional motifs from the literature that also passed conservation tests, and we used this compendium of sequence motifs for 102 regulators (Supplementary Table 3) in all subsequent analysis.

We constructed an initial version of the transcriptional regulatory code by mapping on the yeast genome sequence the motifs that are bound by regulators at high confidence ($P \leq 0.001$) and that are conserved among *sensu stricto Saccharomyces* species (Figure 2; http://web.wi.mit.edu/fraenkel/regulatory_map/). This map includes 3,353 interactions within 1,296 promoter regions. Maps of regulatory sites encompassing larger numbers of promoters, constructed with lower confidence information, can also be viewed on the authors' website. Because the information used to construct the map includes binding data from multiple growth environments, the map describes transcriptional regulatory potential within the genome. During growth in any one environment, only subsets of the binding sites identified in the map are occupied by

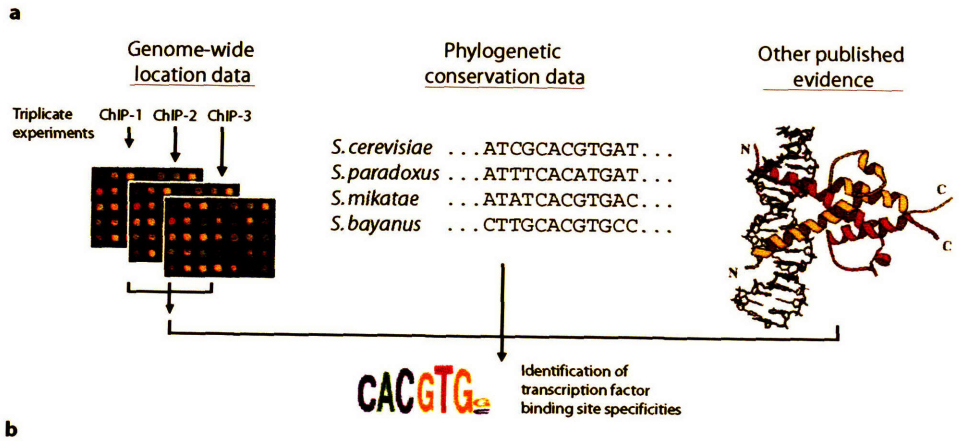

a
Genome-wide location data
Phylogenetic conservation data
Other published evidence
Triplicate experiments
ChIP-1 ChIP-2 ChIP-3
S. cerevisiae ...ATCGCACGTGAT...
S. paradoxus ...ATTTCACATGAT...
S. mikatae ...ATATCACGTGAC...
S. bayanus ...CTTGCACGTGCC...
N
C
C
N
Identification of transcription factor binding site specificities
b
"Rediscovered" sequence specificities
"Discovered" sequence specificities
Abf1
Bas1
Pho4
Rpn4
Ste12
Phd1
Rds1
Snt2
Stb4
YDR026C

Figure 1. Discovering binding site specificities for yeast transcriptional regulators.

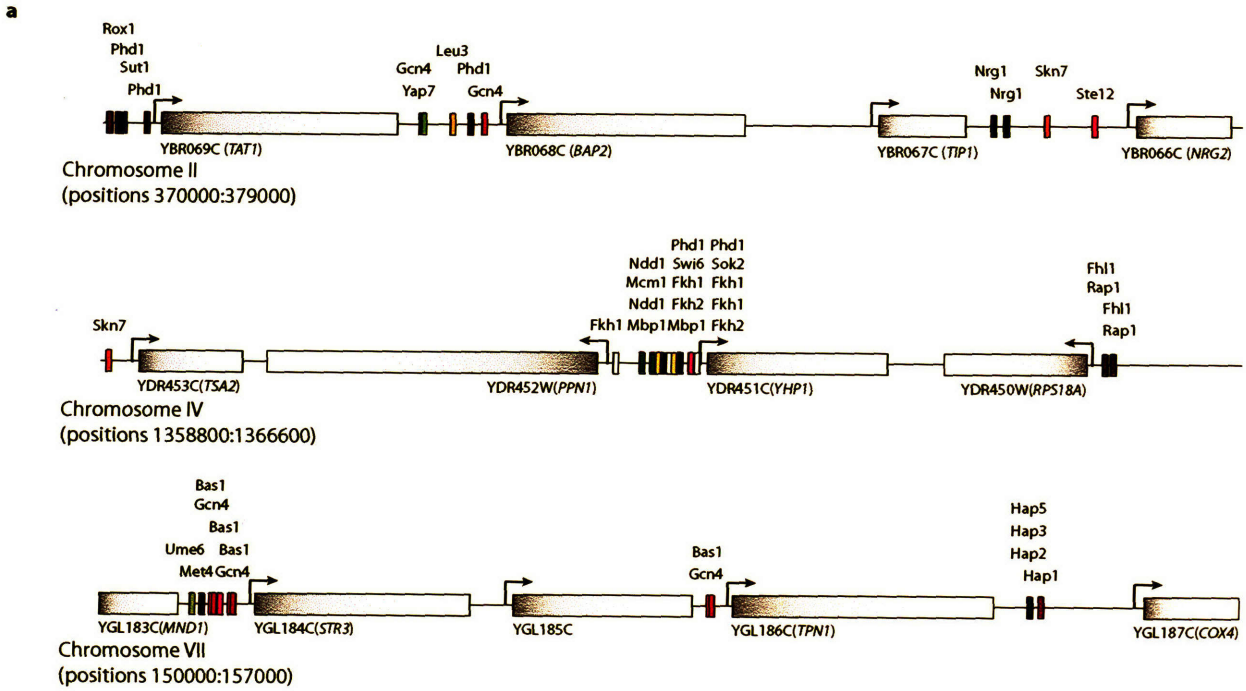
(A). Cis-regulatory sequences that likely serve as recognition sites for transcriptional regulators were identified by combining information from genome-wide location data, phylogenetically conserved sequences, and previously published evidence, as described in Supplementary Methods. The compendium of regulatory sequence motifs can be found in Supplementary Table 3.

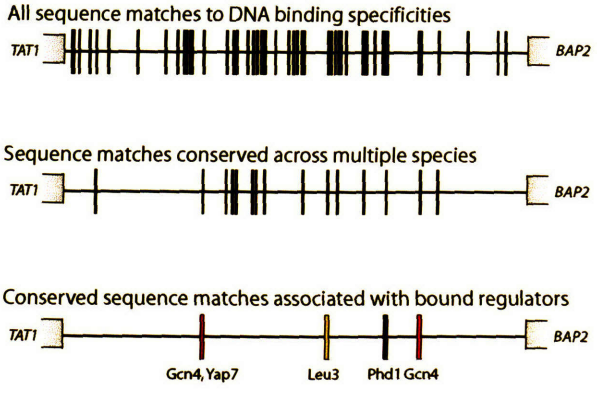
(B). Selected sequence specificities that were "rediscovered" and were newly discovered are displayed. The total height of the column is proportional to the information content of the position, and the individual letters have height proportional to the product of their frequency and the information content (Schneider and Stephens, 1990).

transcriptional regulators, as we describe in more detail below.

Where the functions of specific transcriptional regulators were established previously, the functions of the genes they bind in the regulatory map are highly consistent with this prior information. For example, the amino acid biosynthetic regulators Gcn4 and Leu3 bind to sites in the promoter of *BAP2* (chromosome II), which encodes an amino acid transporter (Figure 2a). Six well-studied cell cycle transcriptional regulators bind to the promoter for *YHP1* (chromosome IV), which has been implicated in regulation of the G1 phase of the cell cycle. The regulator of respiration Hap5, binds upstream of *COX4* (chromosome VII), which encodes a component of the respiratory electron transport chain. Where regulators with established functions bind to genes of unknown function, these target genes are newly implicated in such functional processes.

The utility of combining regulator binding data and sequence conservation data is illustrated in Figure 2b. All sequences matching the regulator DNA binding specificities described in this study (Supplementary Table 2) that occur within the 884 base-pair intergenic region upstream of the gene *BAP2* are shown in the upper panel. The subset of these sequences that have been conserved in multiple yeast species, and are thus likely candidates for regulator interactions, are shown in the middle panel. The presence of these conserved regulatory sites indicates the potential for regulation via this sequence, but does not indicate whether the site is actually bound by a regulator under some growth condition. The incorporation of binding information (bottom panel) identifies those conserved sequences that are utilized by regulators in cells grown under the conditions examined.


a
Rox1
Phd1
Sut1
Phd1
Leu3
Gcn4
Phd1
Yap7
Gcn4
Nrg1
Skn7
Nrg1
Ste12
YBR069C (TAT1)
YBR068C (BAP2)
YBR067C (TIP1)
YBR066C (NRG2)
Chromosome II
(positions 370000:379000)
Phd1 Phd1
Ndd1 Swi6 Sok2
Mcm1 Fkh1 Fkh1
Ndd1 Fkh2 Fkh1
Fkh1 Mbp1Mbp1 Fkh2
Fhl1
Rap1
Fhl1
Rap1
Skn7
YDR453C(TSA2)
YDR452W(PPN1)
YDR451C(YHP1)
YDR450W(RPS18A)
Chromosome IV
(positions 1358800:1366600)
Bas1
Gcn4
Bas1
Ume6
Bas1
Met4
Gcn4
Bas1
Gcn4
Hap5
Hap3
Hap2
Hap1
YGL183C(MND1)
YGL184C(STR3)
YGL185C
YGL186C(TPN1)
YGL187C(COX4)
Chromosome VII
(positions 150000:157000)


b
All sequence matches to DNA binding specificities
TAT1
BAP2
Sequence matches conserved across multiple species
TAT1
BAP2
Conserved sequence matches associated with bound regulators
TAT1
BAP2
Gcn4, Yap7
Leu3
Phd1 Gcn4

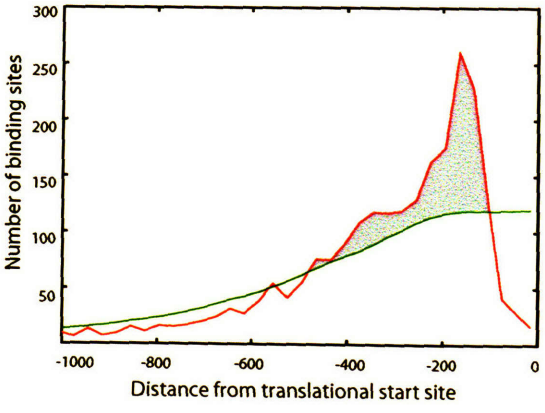

c
Number of binding sites
300
250
200
150
100
50
-1000
-800
-600
-400
-200
0
Distance from translational start site

Figure 2. Drafting the yeast transcriptional regulatory map.

(A). Portions of chromosomes illustrating locations of genes (grey rectangles) and conserved DNA sequences (coloured boxes) bound in vivo by transcriptional regulators.

(B). Combining binding data and sequence conservation data. The diagram depicts all sequences matching a motif from our compendium (top), all such conserved sequences (middle) and all such conserved sequences bound by a regulator (bottom).

(C). Regulator binding site distribution. The red line shows the distribution of distances from the start codon of open reading frames to binding sites in the adjacent upstream region. The green line represents a randomized distribution.

The distribution of binding sites for transcriptional regulators reveals there are constraints on the organization of these sites in yeast promoters (Figure 2c). Binding sites are not uniformly distributed over the promoter regions, but rather show a sharply peaked distribution. Very few sites are located in the region 100 base pairs (bp) upstream of protein coding sequences. This region typically includes the transcription start site and is bound by the transcription initiation apparatus. The vast majority (74%) of the transcriptional regulator binding sites lie between 100 and 500 bp upstream of the protein coding sequence, far more than would be expected at random (53%). Regions further than 500 bp contain fewer binding sites than would be expected at random. It appears that yeast transcriptional regulators function at short distances along the linear DNA, a property that reduces the potential for inappropriate activation of nearby genes.

We note that specific arrangements of DNA binding site sequences occur within promoters, and suggest that these promoter architectures provide clues to regulatory mechanisms (Figure 3). For example, the presence of a DNA binding site for a single regulator is the simplest promoter architecture and, as might be expected, we found that sets of genes with this feature are often involved in a single, common biological function (Supplementary Table 4). A second type of promoter architecture consists of repeats of a particular binding site sequence. Repeated binding sites have been shown to be necessary for stable binding by the regulator Dal80 (Cunningham and Cooper, 1993). This repetitive promoter architecture can also allow for a graded transcriptional response, as has been observed for the *HIS4* gene (Donahue et al., 1983). A number of regulators, including Dig1, Mbp1, and Swi6 show a statistically significant preference for repetitive motifs (Supplementary Table 5). A third class of promoter contains binding sites for

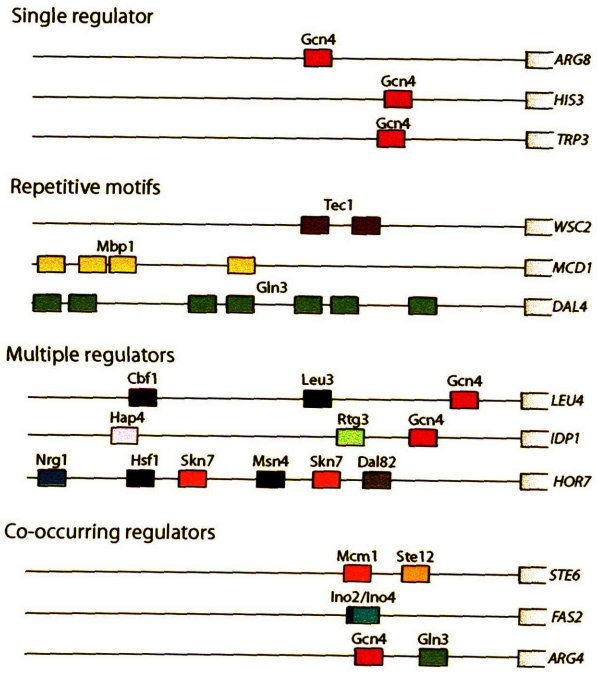

Single regulator
Gcn4
ARG8
Gcn4
HIS3
Gcn4
TRP3
Repetitive motifs
Tec1
WSC2
Mbp1
MCD1
Gln3
DAL4
Multiple regulators
Cbf1
Leu3
Gcn4
LEU4
Hap4
Rtg3
Gcn4
IDP1
Nrg1
Hsf1
Skn7
Msn4
Skn7
Dal82
HOR7
Co-occurring regulators
Mcm1
Ste12
STE6
Ino2/Ino4
FAS2
Gcn4
Gln3
ARG4

Figure 3. Yeast promoter architectures.

Single regulator architecture: promoter regions that contain one or more copies of the binding site sequence for a single regulator. Repetitive motif architecture: promoter regions that contain multiple copies of a binding site sequence of a regulator. Multiple regulator architecture: promoter regions that contain one or more copies of the binding site sequences for more than one regulator. Co-occurring regulator architecture: promoters that contain binding site sequences for recurrent pairs of regulators. Additional information can be found in Supplementary Tables 4-6.

multiple different regulators. This promoter arrangement implies that the gene may be subject to combinatorial regulation, and we expect that in many cases the various regulators can be used to execute differential responses to varied growth conditions. Indeed, we note that many of the genes in this category encode products that are required for multiple metabolic pathways and are regulated in an environment-specific fashion. In the fourth type of promoter architecture we discuss here, binding sites for specific pairs of regulators occur more frequently within the same promoter regions than would be expected by chance (Supplementary Table 6). This "co-occurring" motif architecture implies that the two regulators physically interact or have shared functions at multiple genes.

By conducting genome-wide binding experiments for some regulators under multiple cell growth conditions, we learned that regulator binding to a subset of the regulatory sequences is highly dependent on the environmental conditions of the cell (Supplementary Figure 4). We observed four common patterns of regulator binding behaviour (Figure 4, Supplementary Table 7). Prior information about the regulatory mechanisms employed by well-studied regulators in each of the four groups suggests hypotheses to account for the environment-dependent binding behaviour of the other regulators.

"Condition invariant" regulators bind essentially the same set of promoters (within the limitations of noise) in two different growth environments (Figure 4). Leu3, which is known to regulate genes involved in amino acid biosynthesis, is among the best studied of the regulators in this group. Binding of Leu3 in vivo has been shown to be necessary, but not sufficient for activation of Leu3-regulated genes

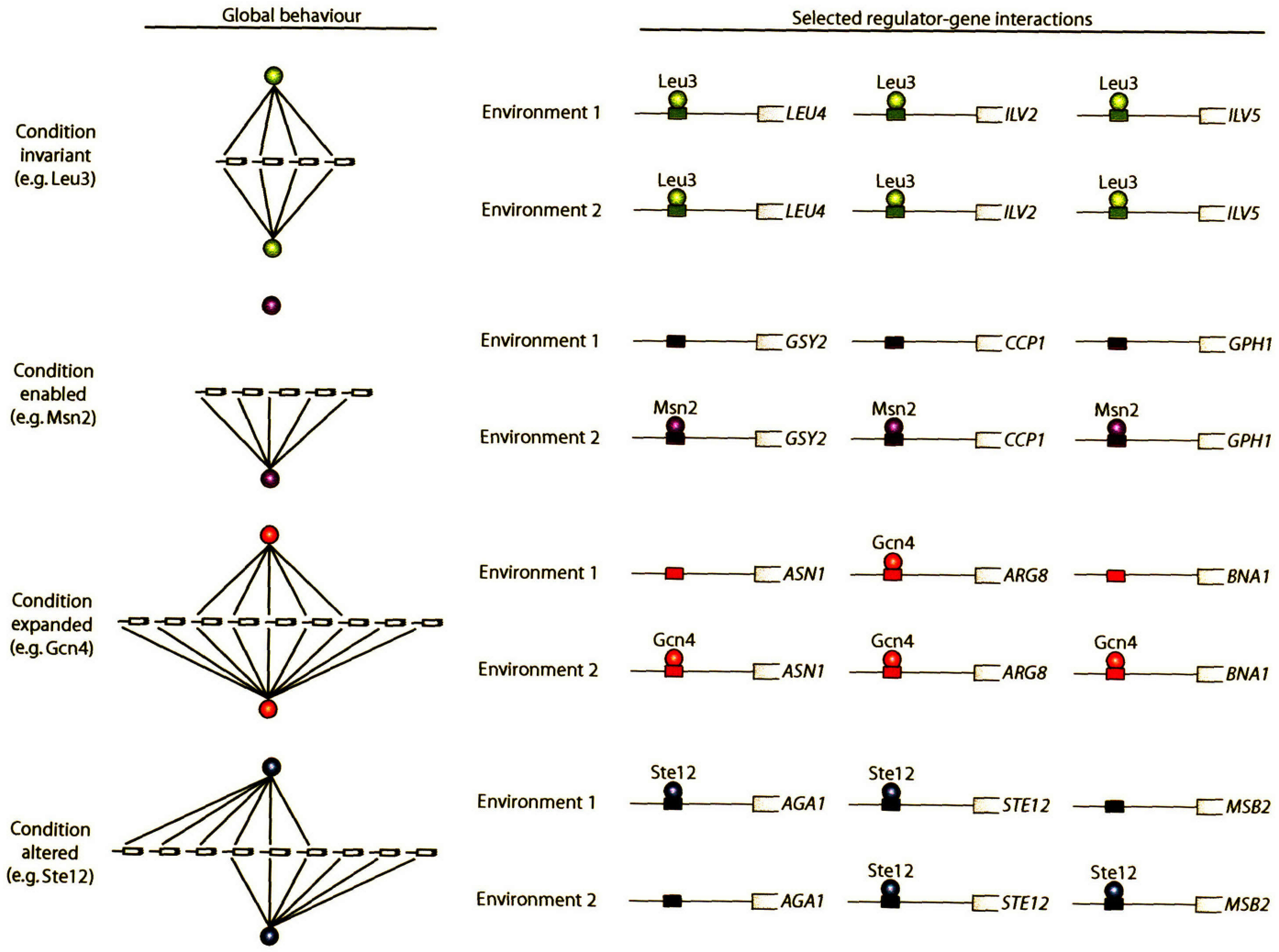

Global behaviour
Selected regulator-gene interactions
Condition invariant (e.g. Leu3)
Condition enabled (e.g. Msn2)
Condition expanded (e.g. Gcn4)
Condition altered (e.g. Ste12)
Environment 1
Environment 2
Leu3
LEU4
ILV2
ILV5
GSY2
CCP1
GPH1
Msn2
ASN1
Gcn4
ARG8
BNA1
Ste12
AGA1
STE12
MSB2

Figure 4. Environment-specific utilization of the transcriptional regulatory code.

Four patterns of genome-wide binding behaviour are depicted in a graphic representation on the left, where transcriptional regulators are represented by coloured circles and are placed above and below a set of target genes/promoters. The lines between the regulators and the target genes/promoters represent binding events. Specific examples of the environment-dependent behaviours are depicted on the right. Coloured circles represent regulators and coloured boxes represent their DNA binding sequences within specific promoter regions. We note that regulators may exhibit different behaviours when different pairs of conditions are compared.

(Kirkpatrick and Schimmel, 1995). Rather, regulatory control of these genes requires association of a leucine metabolic precursor with Leu3 to convert it from a negative to positive regulator. We note that other zinc cluster type regulators that show "condition invariant" behaviour are known to be regulated in a similar manner (Axelrod et al., 1991; Ma and Ptashne, 1987). Thus, it is reasonable to propose that the activation or repression functions of some of the other regulators in this class have requirements in addition to DNA binding.

"Condition enabled" regulators do not bind the genome detectably under one condition, but bind a substantial number of promoters with a change in environment. Msn2 is among the best-studied regulators in this class, and the mechanisms involved in Msn2-dependent transcription provide clues to how the other regulators in that class may operate. Msn2 is known to be excluded from the nucleus when cells grow in the absence of stresses, but accumulates rapidly in the nucleus when cells are subjected to stress (Beck and Hall, 1999; Chi et al., 2001). This condition-enabled behaviour was also observed for the thiamine biosynthetic regulator Thi2, the nitrogen regulator Gat1, and the developmental regulator Rim101. We suggest that many of these transcriptional regulators are regulated by nuclear exclusion or by another mechanism that would cause this extreme version of condition-specific binding.

"Condition expanded" regulators bind to a core set of target promoters under one condition, but bind an expanded set of promoters under another condition. Gcn4 is the best-studied of the regulators that fall into this "expanded" class. The levels of Gcn4 are reported to increase 6-fold when yeast are introduced into media with limiting nutrients (Albrecht et al., 1998), due largely to increased nuclear protein stability (Chi et al.,

2001; Kornitzer et al., 1994), and under this condition we find Gcn4 binds to an expanded set of genes. Interestingly, the probes bound when Gcn4 levels are low contain better matches to the known Gcn4 binding site than probes that are bound exclusively at higher protein concentrations, consistent with a simple model for specificity based on intrinsic protein affinity and protein concentration (Supplementary Figure 5). The expansion of binding sites by many of the regulators in this class may reflect increased levels of the regulator available for DNA binding.

"Condition altered" regulators exhibit altered preference for the set of promoters bound in two different conditions. Ste12 is the best studied of the regulators whose binding behaviour falls into this "altered" class. Depending on the interactions with other regulators, the specificity of Ste12 can change and alter its cellular function (Zeitlinger et al., 2003). For example, under filamentous growth conditions, Ste12 interacts with Tec1, which has its own DNA-binding specificity (Baur et al., 1997). This condition-altered behaviour was also observed for the transcriptional regulators Aft2, Skn7, and Ume6. We propose that the binding specificity of many of the transcriptional regulators may be altered through interactions with other regulators or through modifications (e.g., chemical) that are environment-dependent.

Substantial portions of eukaryotic genome sequence are believed to be regulatory (Cliften et al., 2003; Kellis et al., 2003; Waterston et al., 2002), but the DNA sequences that actually contribute to regulation of genome expression have been ill-defined. By mapping the DNA sequences bound by specific regulators in various environments, we identify the regulatory potential embedded in the genome and provide a framework for modeling the mechanisms that contribute to global gene expression. We anticipate that

the approaches used here to map regulatory sequences in yeast can also be used to map the sequences that control genome expression in higher eukaryotes.

Methods

Strain Information

For each of the 203 regulators, strains were generated in which a repeated Myc epitope coding sequence was integrated into the endogenous gene encoding the regulator. Polymerase chain reaction (PCR) constructs containing the Myc epitope coding sequence and a selectable marker flanked by regions of homology to either the 5' or 3' end of the targeted gene were transformed into the W303 yeast strain Z1256 (Lee et al., 2002; Ren et al., 2000). Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and Western blotting, respectively.

Genome wide location analysis

Genome-wide location analysis was performed as previously described (Lee et al., 2002; Ren et al., 2000). Bound proteins were formaldehyde-crosslinked to DNA in vivo, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-myc antibody, followed by reversal of the crosslinks to separate DNA from protein. Immunoprecipitated DNA and DNA from an unenriched sample were amplified and differentially fluorescently labelled by ligation-mediated PCR. These samples were hybridized to a microarrray consisting of spotted PCR products representing the intergenic regions of the *S. cerevisiae* genome. Relative intensities of spots were used as the basis for an error model that assigns a probability score (*P* value) to binding interactions. All microarray data is available from ArrayExpress (accession number: E-WMIT-10) as well as from the authors' web site.

Growth environments

We profiled all 203 regulators in rich medium. In addition, we profiled 84 regulators in at least one other environmental condition. The list of regulators is given in Supplementary Table 1.

Regulator Binding Specificity

The putative specificities of regulators were identified by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance using uniform metrics and then clustered to yield representative motifs (Supplementary Figure 2).

We used six methods to identify the specific sequences bound by regulators: AlignACE (Roth et al., 1998), MEME (Bailey and Elkan, 1995), Mdscan (Liu et al., 2002), the method of Kellis et al. (Kellis et al., 2003) and two additional new methods that incorporate conservation data: MEME_c and CONVERGE. MEME_c uses the existing MEME program without change, but applies it to a modified set of sequences in which bases that are not conserved in the *sensu stricto Saccharomyces* species were replaced with the letter "N". CONVERGE is a novel expectation-maximization (EM)-based algorithm for discovering specificities using sequence information from multiple genomes. Rather than searching for sites that are identical across the *sensu stricto* species, as is the case for MEME_c, CONVERGE searches for loci where all aligned sequences are consistent with the same specificity model. See Supplementary Methods for runtime parameters and additional details for all of these methods.

Each of the programs we used attempts to measure the significance of its results with one or more statistical scores. However, we observed that these programs report results with high scores even when applied to random selections of intergenic regions. To distinguish the true motifs, we chose a set of statistical measures that are described in the Supplementary Methods, and we converted these scores into the empirical probability that a motif with a similar score could be found by the same program in randomly selected sequences. To estimate these P values, we ran each program 50 times on randomly selected sets of sequences of various sizes. We accepted only those motifs that were judged to be significant by these scores ($P \leq 0.001$).

Significant motifs from all programs were pooled together and clustered using a k-medoids algorithm. Aligned motifs within each cluster were averaged together to produce consensus motifs and filtered according to their conservation. This procedure typically produced several distinct consensus motifs for each regulator. To choose a single specificity for each regulator, we compared the results with information in the TRANSFAC (Matys et al., 2003), YPD (Hodges et al., 1999), and SCPD (Zhu and Zhang, 1999) databases. When no prior information was available, we chose the specificity with the most significant statistical score.

Regulatory Code

Potential binding sites were included in the map of the regulatory code if they satisfied two criteria. First, a locus had to match the specificity model for a regulator in the *Saccharomyces cerevisiae* genome and at least two other *sensu stricto cerevisiae* genomes with a score $\geq 60\%$ of the maximum possible. Second, the locus had to lie

in an intergenic region that also contained a probe bound by the corresponding regulator in any condition ($P \leq 0.001$). All analyses of promoter architecture and environment-specific binding were based on this map, and can be found in Supplementary Information.

Supplementary Methods

More detailed information concerning all the methods used in this paper can be found in at http://web.wi.mit.edu/young/regulatory_code and in Supplementary Information.

Acknowledgements

We thank T. Ideker and S. McCuine for help in selecting regulators to study in environmental conditions. We thank E. Herbolsheimer, G. Bell, R. Latek and F. Lewitter for computational assistance, and E. McReynolds for technical assistance. E. F. is a Whitehead Fellow and was funded in part by Pfizer. D.B.G. was supported by NIH/NIGMS NRSA Award GM068278.

References

Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U., and Braus, G. H. (1998). Monitoring the Gcn4 protein-mediated response in the yeast Saccharomyces cerevisiae. J Biol Chem *273*, 12696-12702.

Axelrod, J. D., Majors, J., and Brandriss, M. C. (1991). Proline-independent binding of PUT3 transcriptional activator protein detected by footprinting in vivo. Mol Cell Biol *11*, 564-567.

Bailey, T. L., and Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. Proc Int Conf Intell Syst Mol Biol *3*, 21-29.

Baur, M., Esch, R. K., and Errede, B. (1997). Cooperative binding interactions required for function of the Ty1 sterile responsive element. Mol Cell Biol *17*, 4330-4337.

Beck, T., and Hall, M. N. (1999). The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. Nature *402*, 689-692.

Blanchette, M., and Tompa, M. (2003). FootPrinter: A program designed for phylogenetic footprinting. Nucleic Acids Res *31*, 3840-3842.

Chi, Y., Huddleston, M. J., Zhang, X., Young, R. A., Annan, R. S., Carr, S. A., and Deshaies, R. J. (2001). Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. Genes Dev *15*, 1078-1092.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., Waterston, R., Cohen, B. A., and Johnston, M. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science *301*, 71-76.

Cunningham, T. S., and Cooper, T. G. (1993). The Saccharomyces cerevisiae DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA). J Bacteriol *175*, 5851-5861.

Donahue, T. F., Daves, R. S., Lucchini, G., and Fink, G. R. (1983). A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. Cell *32*, 89-98.

Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E., and Garrels, J. I. (1999). The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. Nucleic Acids Res *27*, 69-73.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J Mol Biol *3*, 318-356.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241-254.

Kirkpatrick, C. R., and Schimmel, P. (1995). Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. Mol Cell Biol *15*, 4021-4030.

Knuppel, R., Dietze, P., Lehnberg, W., Frech, K., and Wingender, E. (1994). TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J Comput Biol *1*, 191-198.

Kornitzer, D., Raboy, B., Kulka, R. G., and Fink, G. R. (1994). Regulated degradation of the transcription factor Gcn4. Embo J *13*, 6021-6030.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet *28*, 327-334.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. Nat Biotechnol *20*, 835-839.

Ma, J., and Ptashne, M. (1987). The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. Cell *50*, 137-142.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., *et al.* (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res *31*, 374-378.

Pritsker, M., Liu, Y. C., Beer, M. A., and Tavazoie, S. (2004). Whole-genome discovery of transcription factor binding sites by network-level conservation. Genome Res *14*, 99-108.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J.,

Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol *16*, 939-945.

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097-6100.

Wang, T., and Stormo, G. D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. Bioinformatics *19*, 2369-2380.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

Zeitlinger, J., Simon, I., Harbison, C. T., Hannett, N. M., Volkert, T. L., Fink, G. R., and Young, R. A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell *113*, 395-404.

Zhu, J., and Zhang, M. Q. (1999). SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics *15*, 607-611.

Chapter 4

Analysis of the Transcriptional Regulation of Amino Acid Metabolism in *S. cerevisiae* Using Genome-Wide Binding Data

My contributions to this project

The following work represents a project I undertook to elucidate in depth the response of transcriptional regulators to a specific change in environmental growth conditions. I was responsible for the design of the experiments and performed genome-wide location analysis with major contributions from technicians in our lab—Jean-Bosco Tagne, Jane Yoo and Dave Reynolds. I conducted all subsequent analysis with the exception of module discovery, which was performed in conjunction with Ziv Bar-Joseph and Georg Gerber of David Gifford's lab.

Summary

The ability to synthesize protein plays a fundamental role in the capacity for cellular growth, and is limited, in part, by the availability of amino acids. We have used genome-wide location analysis to profile 34 transcription factors implicated in the transcriptional regulation of the cellular response to amino acid starvation. The results confirm what is known from the literature, but also extend our understanding of the complexity of this response, which integrates genes associated with many metabolic pathways and appears to be governed by an interconnected network of transcription factors. We define a regulatory network that allows for control of specific pathways as well as large-scale coordinated responses, and identify Cbf1 as a key regulator in the latter process. Surprisingly, we also find new evidence for multiple levels of control of Gcn4, a well-studied and essential regulator of this response. Finally, we have combined our location data with expression data to generate regulatory modules consisting of sets of genes whose expression is likely controlled by a given factor or set of factors.

Introduction

The utility of yeast as a model system in molecular biology was demonstrated by the early insights gained into the basic cellular functions of eukaryotic cells including cell cycle, cell division and metabolism. For example, the study of amino acid auxotrophs has led to a wealth of information on the mechanisms by which cells regulate the production and consumption of these "building blocks of protein." Nevertheless, most of this work is the accumulation of studies of single regulator/gene interactions. Advances in the use of DNA microarrays have allowed for investigations into changes of entire cellular expression programs (DeRisi et al., 1997; Holstege et al., 1998; Natarajan et al., 2001; Roth et al., 1998; Wodicka et al., 1997), as well identification of the genomic binding sites of transcriptional regulators (Iyer et al., 2001; Lee et al., 2002; Ren et al., 2000). We have made use of this latter technology, genome-wide location analysis, which combines chromatin immunoprecipitation with microarray technology, to study the genomic regulators implicated in the regulation of amino acid biosynthesis. Regulators were selected for profiling if they met one of four conditions: they were previously characterized as such in the literature; their deletions resulted in abnormal growth under amino acid starvation conditions; they were previously found to be physically associated with amino acid genes in location analysis experiments; or the gene expression of the regulators changed during growth under amino acid starvation conditions. We then used location analysis to examine the binding of these regulators both in growth under rich medium as well as in amino acid starvation medium.

Results and Discussion

Network architecture of amino acid biosynthesis regulation

Most of the factors chosen to be profiled under amino acid starvation conditions have a primary role in regulation of amino acid metabolism. We examined the extent to which factors were dedicated to regulating specific biosynthetic pathways. Genes encoding proteins involved in amino acid biosynthesis were segregated according to the pathway in which they functioned (Fig. 1), with factors binding promoter regions of three or more genes within a certain pathway being assigned to that pathway. Generally, all factors fell into one of two categories. Specific regulators bound only to promoter regions of genes primarily associated with a single amino acid biosynthetic pathway. An example of such a regulator is Leu3, which binds upstream of a relatively small number of targets under either condition, but whose targets include the leucine biosynthetic genes *ILV2*, *BAT1*, *LEU1*, *LEU4* and *LEU9*. In contrast, some regulators, namely Gcn4 and Cbf1, appear to regulate multiple biosynthetic pathways. The general regulatory nature of Gcn4 is well documented (Hinnebusch and Fink, 1983), but that of Cbf1 is unexpected. This factor has been previously implicated in maintaining centromere function, but also in the regulation of methionine biosynthetic genes. We find that Cbf1 not only binds to the promoter regions of genes associated with this pathway, but also to genes required to synthesize aromatic amino acids, proline, and aspartate, among others, indicating that this factor may play a central role in coordinating the transcriptional response to amino acid starvation.

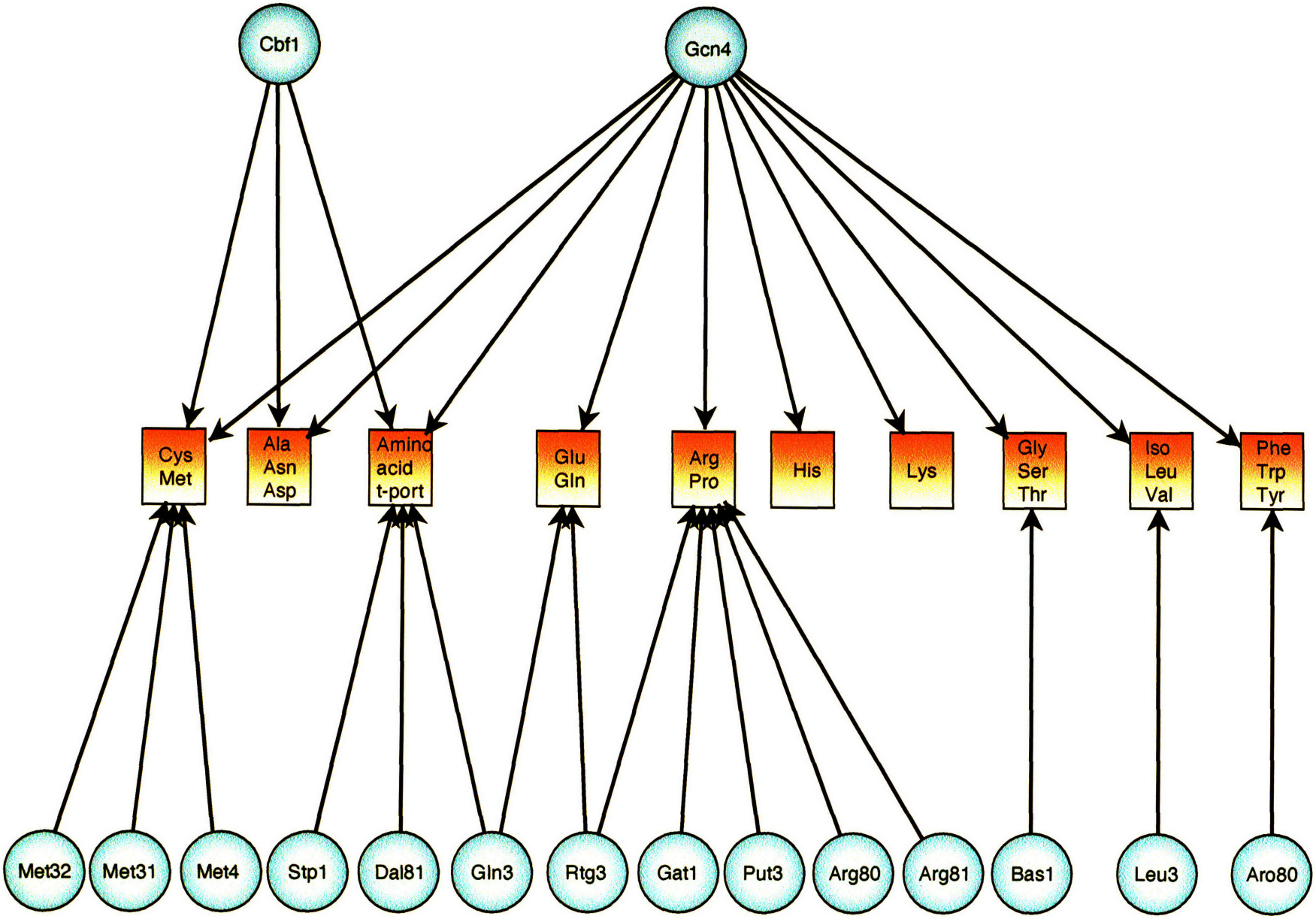

Cbf1
Gcn4
Cys
Met
Ala
Asn
Asp
Amino
acid
t-port
Glu
Gln
Arg
Pro
His
Lys
Gly
Ser
Thr
Iso
Leu
Val
Phe
Trp
Tyr
Met32
Met31
Met4
Stp1
Dal81
Gln3
Rtg3
Gat1
Put3
Arg80
Arg81
Bas1
Leu3
Aro80

Figure 1. Target pathways of transcriptional regulators.

Amino acid metabolic pathways are represented by the central row of circles. Arrows indicate that a factor (outer rows of circles) binds to at least three of the upstream intergenic regions of genes in a given pathway with $P \leq 0.001$. "General" factors are on top, "specific" factors are below.

We were also able to use location data to assign new functions to some transcription factors. For example, we find that the factor Bas1 binds upstream of genes involved in purine synthesis, but also upstream of genes in the serine biosynthetic pathway (including *SHM2*, *GCV1*, *GCV2* and *GCV3*). This confirms previous evidence that Bas1 might regulate multiple pathways (Denis et al., 1998), especially those upstream of purine biosynthesis. Likewise, Rtg3, which is responsible for regulation of the TCA cycle (Jia et al., 1997; Liu and Butow, 1999), and Gat1, involved in nitrogen regulation (Coffman et al., 1996), also appear to regulate genes involved in amino acid biosynthesis, particularly in the arginine and proline biosynthetic pathways.

Transcriptional regulation of transcriptional regulators

One of the advantages of genome-wide location analysis is its ability to identify regulatory interactions among transcriptional regulators themselves. In analyzing cell cycle, for example, we found that it is characterized by a regulatory architecture in which one regulator or set of regulators activates transcription of a regulatory gene required for control of a subsequent phase of the cell cycle (Simon et al., 2001). This motif extends in a continuous loop throughout the cell cycle.

We find similar evidence for the importance of the regulation of regulators in the response to amino acid starvation (Fig 2). The most obvious is the extent to which Gcn4 binds upstream of other regulatory genes, including Met4, Leu3, Lys14, Put3 and Uga3. While there exists evidence for Gcn4 regulation of Met4 (Mountain et al., 1993) and Leu3 (Zhou et al., 1987) the finding that Gcn4 directly regulates so many other regulators

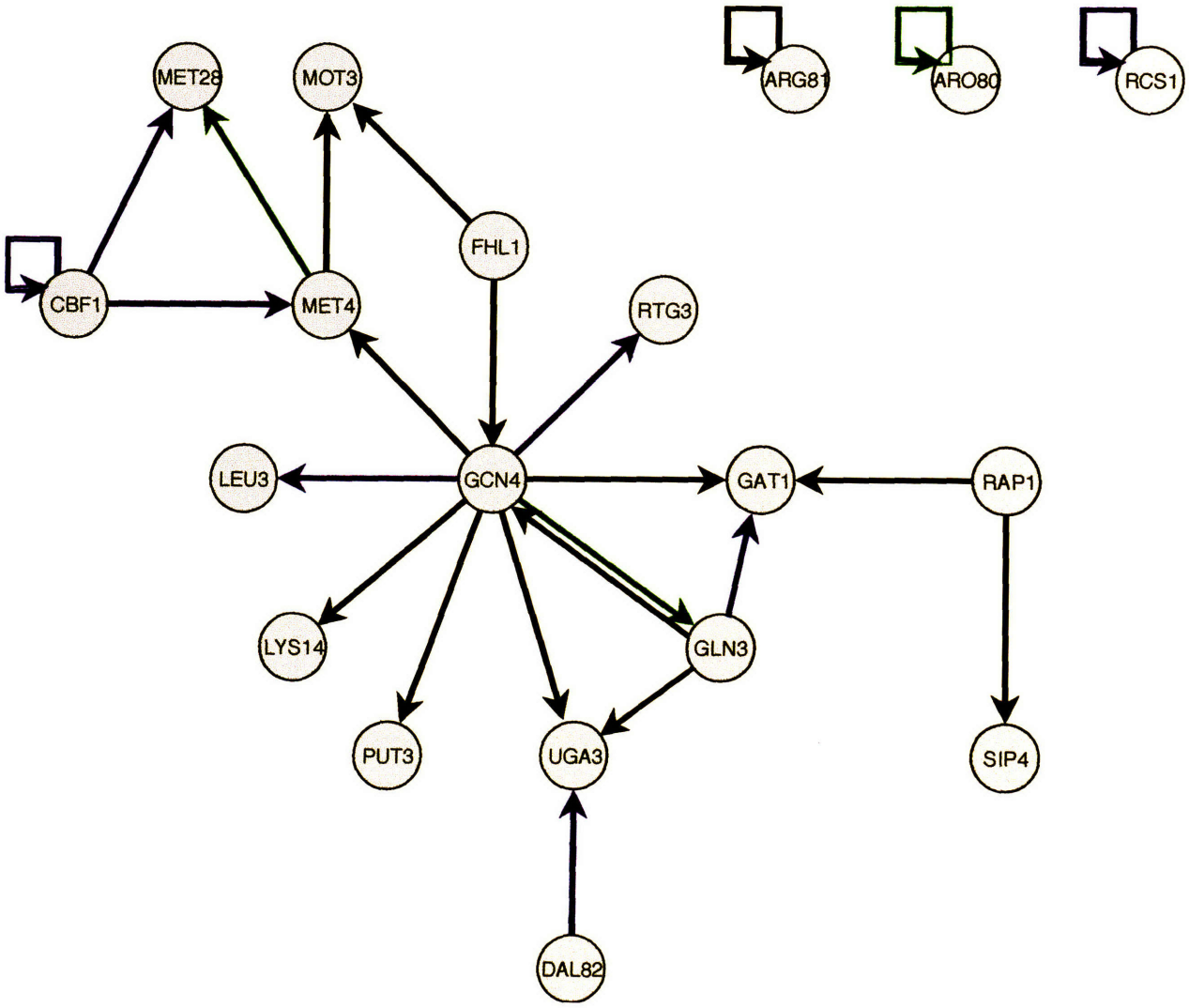

ARG81
ARO80
RCS1
MET28
MOT3
FHL1
CBF1
MET4
RTG3
LEU3
GCN4
GAT1
RAP1
LYS14
GLN3
PUT3
UGA3
SIP4
DAL82

Fig. 2. Regulator-regulator network.

Arrows indicate that a given factor binds to the upstream intergenic region of a corresponding factor with $P \leq 0.001$. Black arrows indicate that a binding event is observed under both rich and starvation growth conditions; blue arrows indicate that a binding event is observed under the starvation growth condition only; green arrows indicate that a binding event is observed under the rich growth condition only.

is novel. Such architecture, however, explains both the expression changes of these regulatory genes upon amino acid starvation (Natarajan et al., 2001), and suggests a mechanism whereby part of the general control response may be mediated through secondary regulators.

We find other interactions that likely play an important role in coordinating regulatory responses. One such set of interactions exists among the genes which encode regulators of methionine biosynthesis. Cbf1, Met4 and Met28 are all members of a complex that regulates methionine and cysteine (Cherest et al., 1997; Kuras et al., 1997; Kuras et al., 1996; Masselot and De Robichon-Szulmajster, 1975; Mountain et al., 1993; Thomas et al., 1992). Consistent with previous genetic and in vitro biochemical data (Kuras et al., 1997) we find that the promoter of the *MET28* gene is bound by both Met4 and Cbf1. Additionally, we find that Cbf1 binds to the promoter region of the *MET4* gene as well as to that of *CBF1* itself. Thus a model emerges in which three genes encoding members of a single transcriptional activation complex are themselves regulated by elements of that complex. Presumably such a mechanism could allow for feedback regulation as well as help control production of stoichiometric levels of complex components.

We also note that a similar network appears to exist for factors involved in the regulation of nitrogen and nitrogenous compounds. Gln3, a primary regulator of genes involved in nitrogen metabolism (Courchesne and Magasanik, 1988; Mitchell and Magasanik, 1984), binds upstream of a related regulator, *GAT1*, as well as to the amino acid regulators *GCN4* and *UGA3*. Gcn4, in turn, also binds upstream of *GLN3*, *GAT1* and *UGA3*. Finally, Dal82, another regulator of nitrogen metabolism, as well as Leu3

and Arg81 are connected to this subnetwork. Transcriptional control of nitrogen metabolism has previously been shown to exhibit complex cross-regulatory properties (Coffman et al., 1997), with Gln3 required both for maximal induction of the *UGA* (*U*tilization of *GA*BA) genes (Talibi et al., 1995) and Gat1 activity (Coffman et al., 1996). As amino acid metabolism is inextricably tied to the type and quantity of nitrogen sources in the cell, such regulatory connections provide a means whereby these two metabolic processes may be coordinately regulated.

In addition to coordinating closely linked metabolic processes, regulation of transcriptional regulators is also a mechanism in which crosstalk between metabolic pathways can occur. For example, we note that the promoter of the transcription factor Rtg3, implicated in regulation of citric acid cycle genes, is bound by Gcn4. As carbon metabolism, like nitrogen metabolism, represents a major metabolic input for amino acid metabolism, Gcn4's regulation of *RTG3* may be a means to ensure adequate sources of the carbon compound precursors for amino acid biosynthesis. Similarly, we find that Fhl1, a key regulator of ribosomal genes (Lee et al., 2002), binds upstream of *GCN4* under both conditions. Rap1, another major regulator of ribosomal genes (Shore and Nasmyth, 1987) binds upstream of *GCN4* under both conditions with a slightly less restrictive *P* value. This connection between regulators of protein synthesis (Fhl1 and Rap1) and the major regulator of amino acid synthesis may represent a mechanism whereby the cell coordinates these interrelated processes. Interestingly, Rap1 is also required for the full induction of certain targets of Gcn4 (Devlin et al., 1991; Yu et al., 2001).

Analysis of Gcn4 Regulation

Gcn4, the major regulator of the "general control" response to amino acid starvation, is itself known to be regulated at many levels. A close look at binding data for Gcn4 reveals that location analysis can reveal multiple mechanisms by which a transcription factor itself may be regulated. We find that Gcn4 binding data confirms a known mechanism of Gcn4 regulation, extends another, and suggests a third (Fig. 3).

Kornitzer et al. have shown that levels of Gcn4 are controlled, in part, at the level of protein stability (Kornitzer et al., 1994). Two cyclin-dependent kinases Pho85 and Srb10 have been shown to phosphorylate Gcn4 under non-starvation conditions, leading to its rapid degradation by the proteasome (Chi et al., 2001). Our binding data support the idea suggested by Shemer (Shemer et al., 2002) that Gcn4 regulates levels of Pcl5, the cyclin partner of Pho85 in a negative feedback loop.

Gcn4 has been a well-studied model for translational regulation. Levels of Gcn4 protein increase upon a switch to conditions of amino acid starvation as a result of increased translation of Gcn4 mRNA transcripts. This translational control is mediated by the rate of reinitiation of ribosomal tertiary complexes whose activity is modulated by the levels of aminoacylated tRNAs (Hinnebusch, 1984; Hinnebusch, 1997; Thireos et al., 1984). Targets of Gcn4 in *S. cerevisiae* include the tRNA synthetase genes *ILS1*, *MES1* and *KRS1* (Lanker et al., 1992; Meussdoerffer and Fink, 1983; Mirande and Waller, 1988). We find that a number of other tRNA synthetase gene promoters are also bound by Gcn4, namely, *VAS1*, *DED81*, *YDR341C*, *YHR020W*, *FRS2* (all $P \leq 0.001$) and *THS1* ($P \leq 0.005$). Lanker et al. have suggested a model in which the lysyl tRNA synthetase, Krs1, forms an autoregulatory feedback loop with Gcn4. As the genes listed above

Translation
Gcn4
ILS1
YDR341C
VAS1
DED81
YHR020W
tRNA synthetases

Protein stability
Gcn4
PCL5
Pcl5

Transcription
Gcn4
GLN3
GCN4
Gln3

Fig. 3. Targets of Gcn4 binding whose products may modulate Gcn4 activity.

Gcn4 binds upstream of genes encoding products that regulate Gcn4 activity. These negative feedback loops are predicted to affect Gcn4 translation, stability and transcription.

together represent at least nine tRNA synthetases (associated with every class of amino acid) and, as depletion of any amino acid leads to increased translation of Gcn4 (Wek et al., 1995), we suggest a more general model in which transcription of tRNA synthetases as a group is regulated by Gcn4. In this model, the depletion of amino acids results in a lower concentration of charged tRNAs, indirectly stimulating translation of *GCN4*. Higher levels of Gcn4, in turn, activate transcription of tRNA synthetase genes, leading eventually to restored levels of charged tRNAs and turning off translation of *GCN4*.

An additional level of regulation of Gcn4 is postulated to occur at the transcriptional level. Mutants of *GCN4* that are insensitive to translational regulation nevertheless show an increase in protein levels upon amino acid starvation (Albrecht et al., 1998). The identity of a transcriptional activator of Gcn4, however, has proved elusive. We find that the nitrogen utilization regulator Gln3 binds upstream of *GCN4* under both rich and amino acid starvation conditions. We also find that Gcn4 binds to the promoter region of the *GLN3* gene (*P* value 0.0011). These binding data suggest that a positive autoregulatory feedback loop may exist between these two genes. This result is intriguing because Gcn4 has been suggested to be responsible for part of the response to rapamycin (Valenzuela et al., 2001), which is known to be mediated in part by Gln3.

Comparison with expression data

Expression analysis of Gcn4 (Natarajan et al., 2001) has demonstrated that this factor plays an important role in controlling the expression of at least 500 genes in response to amino acid starvation. It is not known, however, to what extent these genes are regulated directly by Gcn4 or indirectly, for example by a factor itself regulated

by Gcn4. We attempted to shed light on this question by comparing the results of expression analysis with those of our location analysis. We find that 153 (28%) of the 540 targets identified by expression analysis are confirmed by location analysis. Similar overlaps between location and expression data are common, with a number of factors (divergent promoters, *P* value stringency, fold cut-offs, experimental noise, secondary effects) contributing to any discrepancies.

We then looked to see if we could identify genes whose expression was dependent on secondary effects of Gcn4 activity, that is, genes whose promoters were not bound by Gcn4, but were bound by regulators that are themselves transcriptionally regulated targets of Gcn4. A number of regulators of nitrogen and amino acid metabolic pathways form an integrated network (Fig. 2). Gcn4, for example, binds to the upstream regions of *LYS14*, *UGA3*, *GAT1*, *PUT3*, *MET4*, *LEU3* and *RTG3* ($P \leq 0.001$), and that of *GLN3* is bound with a slightly less restrictive *P* value. Of those genes whose expression changes, but which are not targeted by Gcn4, regulation of a few can be accounted for by secondary effects. The genes include *MET2*, *MET28*, *MET14*, *MET17* and *SUL2* which are bound by Met4 ($P \leq 0.005$). Similarly, Gln3, Rtg3, Gat1, Leu3, Put3 and Uga3 bind upstream of genes not bound by Gcn4. In total, at least 45 genes may be regulated in this fashion.

The above results indicate that either some expression-derived targets of Gcn4 are spurious or that Gcn4 location data is not able to account for all Gcn4-regulated genes. To investigate this further, we applied more stringent criteria to the interpretation of expression data. A total of 316 genes were induced in all four experiments by Natarajan. Of these, Gcn4 binding is associated with more than one-third (109) at *P* value ≤

0.005. Of the remainder, binding by other factors, particularly Aro80, Bas1, Dal82, Cad1, Cbf1 and Rap1, can account for changes in expression for 73 genes. Nevertheless, a significant number of genes showing consistent changes in expression are not associated with binding by our factors. We surmise that some of the differences result from the different conditions used to induce starvation, different strains used, and the contribution of regulators not profiled (many "unbound" genes are involved in stress response), among other factors.

Interestingly, we find a number of cases in which factors and genes form "feed forward" loops (Fig. 4). Such motifs consist of a primary regulator that binds to a promoter of a secondary regulatory gene, and both the primary and secondary regulator bind the promoter of a common target gene. It appears as if many of these target genes are controlled by a secondary regulator in rich medium, and controlled by a primary regulator under amino acid starvation conditions (data not shown). The fact that the secondary factor is regulated by the primary factor may be the result of the need to activate transcription of some genes not regulated directly by Gcn4, but by the secondary regulator (for example, the set of genes regulated by Met4). Alternatively, such a motif could provide a means for modulating the transcriptional output of the target gene. Recent work in network analysis supports this latter hypothesis, suggesting that feed forward loops help to buffer responses to mild environmental perturbations (Mangan et al., 2003; Shen-Orr et al., 2002; Yekta et al., 2004).

Finally, we were surprised to note that, while most previously identified classes of targets of Gcn4 were confirmed as such by our binding data, we did not observe Gcn4 binding upstream of genes encoding purine biosynthetic enzymes. We note, however,

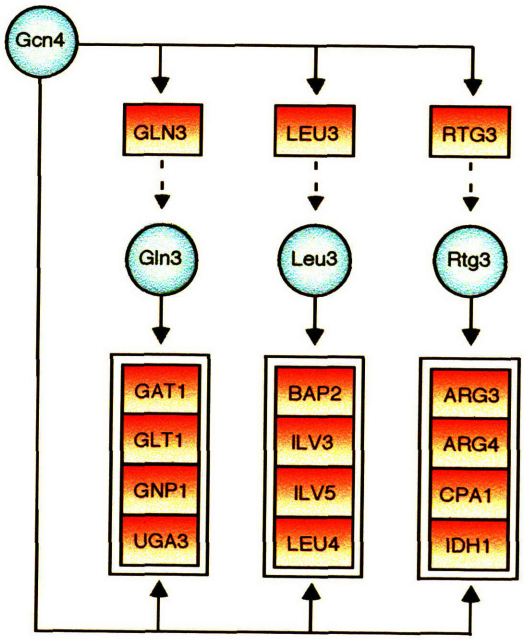

Gcn4
GLN3
LEU3
RTG3
Gln3
Leu3
Rtg3
GAT1
GLT1
GNP1
UGA3
BAP2
ILV3
ILV5
LEU4
ARG3
ARG4
CPA1
IDH1

Figure 4. Feed forward loops.

Gcn4 is implicated in a number of feed forward loops. These regulatory motifs consist of transcriptional regulators that control the activity of genes that encode other transcriptional regulators. In addition, both regulators regulate a common set of genes.

that we did observe binding of Bas1 at these genes (*ADE2*, *ADE4*, *ADE6*, *ADE8*, *ADE13*, *ADE17*) as has been suggested by the literature (Daignan-Fornier and Fink, 1992; Denis and Daignan-Fornier, 1998). We believe that the previous suggestion that these genes are Gcn4-regulated (Mosch et al., 1991; Rolfes and Hinnebusch, 1993) is confounded by two factors. The first is that the consensus binding sequence for Gcn4 is the same as that proposed for Bas1, so that mutations in a promoter that eliminate the binding of one will eliminate binding of the other. The second is that purine metabolism shares a common metabolic intermediate (AICAR) with histidine metabolism (Arndt et al., 1989; Daignan-Fornier and Fink, 1992; Springer et al., 1996), suggesting that an imbalance in one pathway might affect the other.

Modules

In an effort to combine independent lines of evidence, we have fused our location data with genome-wide expression data. We used an algorithm that identifies sets of genes whose expression is highly correlated and then determines the likelihood that such a set shows upstream binding by a factor or set of factors base on location data (Fig. 5). This method helps both to boost confidence in stringently interpreted binding data as well as to identify likely target genes that might have been excluded based on a strict *P* value threshold (Bar-Joseph et al., 2003). It also identifies candidate factors for combinational regulation of target genes.

One surprising result from the module analysis is the fact that Dal81 and Dal82, which are believed to work together to activate transcription (Talibi et al., 1995), appear to regulate separate modules. Dal82 is associated with a set of *DAL* (*D*egradation of

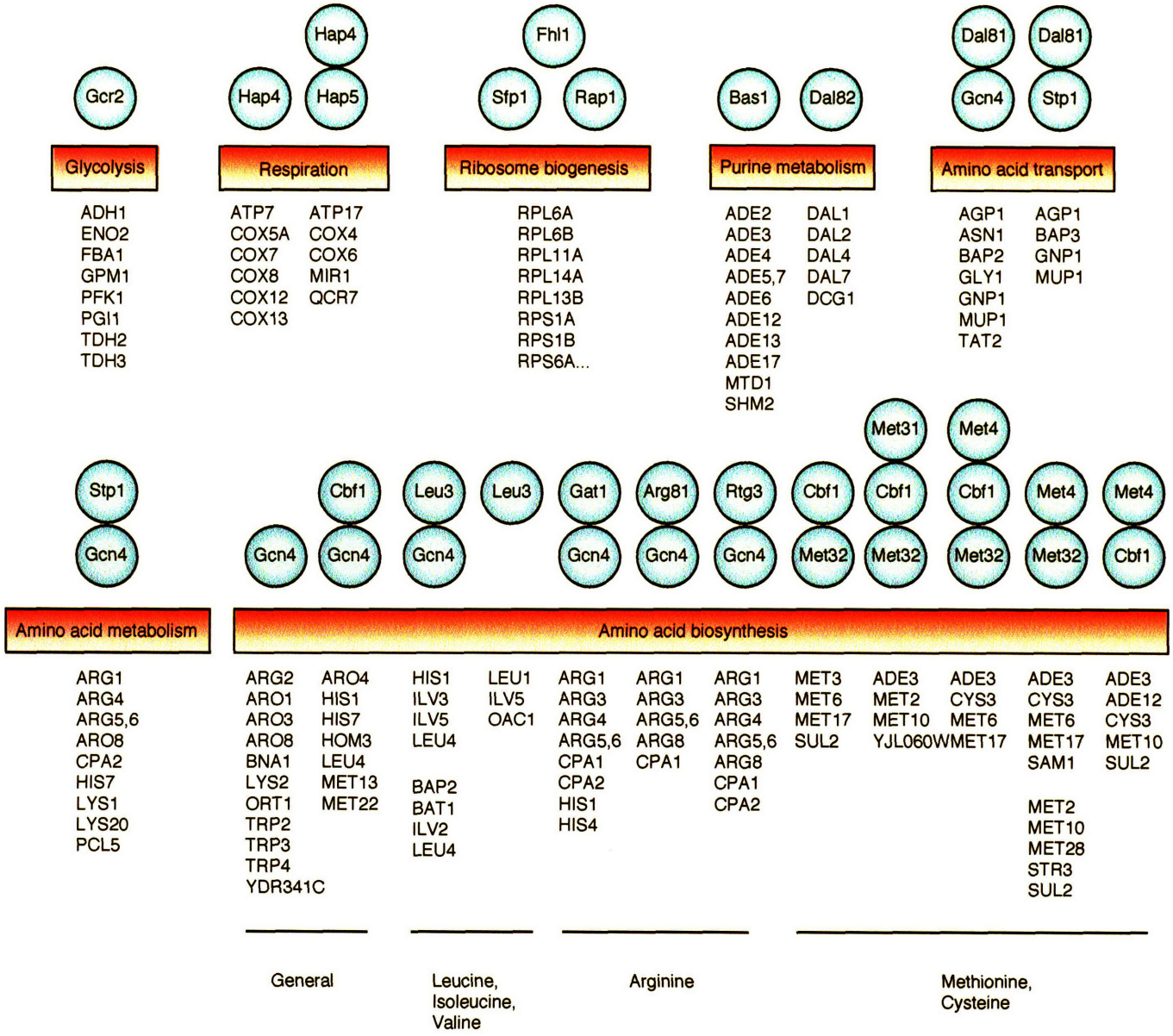

Gcr2
Glycolysis
ADH1
ENO2
FBA1
GPM1
PFK1
PGI1
TDH2
TDH3
Hap4
Hap4
Hap5
Respiration
ATP7
COX5A
COX7
COX8
COX12
COX13
ATP17
COX4
COX6
MIR1
QCR7
Fhl1
Sfp1
Rap1
Ribosome biogenesis
RPL6A
RPL6B
RPL11A
RPL14A
RPL13B
RPS1A
RPS1B
RPS6A...
Bas1
Dal82
Purine metabolism
ADE2
ADE3
ADE4
ADE5,7
ADE6
ADE12
ADE13
ADE17
MTD1
SHM2
DAL1
DAL2
DAL4
DAL7
DCG1
Dal81
Gcn4
Dal81
Stp1
Amino acid transport
AGP1
ASN1
BAP2
GLY1
GNP1
MUP1
TAT2
AGP1
BAP3
GNP1
MUP1
Stp1
Gcn4
Amino acid metabolism
ARG1
ARG4
ARG5,6
ARO8
CPA2
HIS7
LYS1
LYS20
PCL5
Gcn4
Cbf1
Gcn4
Leu3
Gcn4
Leu3
Gat1
Gcn4
Arg81
Gcn4
Rtg3
Gcn4
Cbf1
Met32
Met31
Cbf1
Met32
Met4
Cbf1
Met32
Met4
Met32
Met4
Cbf1
Amino acid biosynthesis
ARG2
ARO1
ARO3
ARO8
BNA1
LYS2
ORT1
TRP2
TRP3
TRP4
YDR341C
ARO4
HIS1
HIS7
HOM3
LEU4
MET13
MET22
HIS1
ILV3
ILV5
LEU4
BAP2
BAT1
ILV2
LEU4
LEU1
ILV5
OAC1
ARG1
ARG3
ARG4
ARG5,6
CPA1
CPA2
HIS1
HIS4
ARG1
ARG3
ARG5,6
ARG8
CPA1
ARG1
ARG3
ARG4
ARG5,6
ARG8
CPA1
CPA2
MET3
MET6
MET17
SUL2
ADE3
MET2
MET10
YJL060W
ADE3
CYS3
MET6
MET17
ADE3
CYS3
MET6
MET17
SAM1
MET2
MET10
MET28
STR3
SUL2
ADE3
ADE12
CYS3
MET10
SUL2
General
Leucine,
Isoleucine,
Valine
Arginine
Methionine,
Cysteine

Fig. 5. Fusion of expression and location data into functional modules.

An algorithm was used to identify sets of genes with correlated expression and the transcription factors which likely bind them. Factors are listed in blue circles, functional categories of genes whose promoter regions are bound are in red rectangles and genes themselves are listed below.

*AL*lantoin) genes in agreement with its role per the literature (Dorrington and Cooper, 1993; Olive et al., 1991). Dal81, however, is similar to Gcn4 and Stp1 in its ability to regulate amino acid permeases, as it binds upstream of no fewer than eight such genes, indicating that Dal81 may have a more general role in regulating transport of amino acids and peptides than was previously thought (Bernard and Andre, 2001; Iraqui et al., 1999).

Finally, we note that a number of different modules were generated for combinations of the factors Cbf1, Met4, Met31 and Met32. As stated above, these factors are known to act together in the regulation of methionine biosynthesis. There is evidence, however, that the exact composition of this complex can vary at different genes (Blaiseau and Thomas, 1998), indicating that the different modules may in fact represent bona fide sets of genes whose expression is governed by combinatorial control.

Conclusion

We have analyzed the genomic binding locations for 34 factors under rich and amino acid starvation conditions. We have shown how location analysis can be used to map the transcriptional regulatory networks that underlie the cellular response to changing environmental conditions. We have discovered novel functions for transcriptional regulators. We have identified interactions between transcriptional regulators that coordinate cellular responses to amino acid starvation. Finally, we have combined location data with that of expression data to help understand both the direct sets of target genes for factors and which factors may act together to achieve combinatorial regulation of the genome.

Methods

Strain Information

For each of the 34 regulators, strains were generated in which a repeated Myc epitope coding sequence was integrated into the endogenous gene encoding the regulator. Polymerase chain reaction (PCR) constructs containing the Myc epitope coding sequence and a selectable marker flanked by regions of homology to either the 5' or 3' end of the targeted gene were transformed into the W303 yeast strain Z1256. Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and Western blotting, respectively.

Growth environments

Cells profiled in rich medium were grown in YPD (1% yeast extract/2% peptone/2% glucose) to an OD600 of ~0.8. Cells profiled under amino acid starvation conditions were grown to an OD600 of ~0.6 in synthetic complete medium followed by treatment with the inhibitor of amino acid biosynthesis sulfometuron methyl (0.2 μg/ml final) for two hours.

Genome-wide location analysis

Genome-wide location analysis was performed as previously described (Lee et al., 2002; Ren et al., 2000). Bound proteins were formaldehyde-crosslinked to DNA in vivo, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-Myc antibody, followed by reversal of the crosslinks to separate DNA from protein. Immunoprecipitated DNA and DNA from an unenriched

sample were amplified and differentially fluorescently labelled by ligation-mediated PCR. These samples were hybridized to a microarrray consisting of spotted PCR products representing the intergenic regions of the *S. cerevisiae* genome. Relative intensities of spots were used as the basis for an error model that assigns a probability score (*P* value) to binding interactions.

References

Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U., and Braus, G. H. (1998). Monitoring the Gcn4 protein-mediated response in the yeast Saccharomyces cerevisiae. J Biol Chem *273*, 12696-12702.

Arndt, K. T., Styles, C. A., and Fink, G. R. (1989). A suppressor of a HIS4 transcriptional defect encodes a protein with homology to the catalytic subunit of protein phosphatases. Cell *56*, 527-537.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003). Computational discovery of gene modules and regulatory networks. Nat Biotechnol *21*, 1337-1342.

Bernard, F., and Andre, B. (2001). Genetic analysis of the signalling pathway activated by external amino acids in Saccharomyces cerevisiae. Mol Microbiol *41*, 489-502.

Blaiseau, P. L., and Thomas, D. (1998). Multiple transcriptional activation complexes tether the yeast activator Met4 to DNA. Embo J *17*, 6327-6336.

Cherest, H., Davidian, J. C., Thomas, D., Benes, V., Ansorge, W., and Surdin-Kerjan, Y. (1997). Molecular characterization of two high affinity sulfate transporters in Saccharomyces cerevisiae. Genetics *145*, 627-635.

Chi, Y., Huddleston, M. J., Zhang, X., Young, R. A., Annan, R. S., Carr, S. A., and Deshaies, R. J. (2001). Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. Genes Dev *15*, 1078-1092.

Coffman, J. A., Rai, R., Cunningham, T., Svetlov, V., and Cooper, T. G. (1996). Gat1p, a GATA family protein whose production is sensitive to nitrogen catabolite repression, participates in transcriptional activation of nitrogen-catabolic genes in Saccharomyces cerevisiae. Mol Cell Biol *16*, 847-858.

Coffman, J. A., Rai, R., Loprete, D. M., Cunningham, T., Svetlov, V., and Cooper, T. G. (1997). Cross regulation of four GATA factors that control nitrogen catabolic gene expression in Saccharomyces cerevisiae. J Bacteriol *179*, 3416-3429.

Courchesne, W. E., and Magasanik, B. (1988). Regulation of nitrogen assimilation in Saccharomyces cerevisiae: roles of the URE2 and GLN3 genes. J Bacteriol *170*, 708-713.

Daignan-Fornier, B., and Fink, G. R. (1992). Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. Proc Natl Acad Sci U S A *89*, 6746-6750.

Denis, V., Boucherie, H., Monribot, C., and Daignan-Fornier, B. (1998). Role of the myb-like protein bas1p in Saccharomyces cerevisiae: a proteome analysis. Mol Microbiol *30*, 557-566.

Denis, V., and Daignan-Fornier, B. (1998). Synthesis of glutamine, glycine and 10-formyl tetrahydrofolate is coregulated with purine biosynthesis in Saccharomyces cerevisiae. Mol Gen Genet *259*, 246-255.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

Devlin, C., Tice-Baldwin, K., Shore, D., and Arndt, K. T. (1991). RAP1 is required for BAS1/BAS2- and GCN4-dependent transcription of the yeast HIS4 gene. Mol Cell Biol *11*, 3642-3651.

Dorrington, R. A., and Cooper, T. G. (1993). The DAL82 protein of Saccharomyces cerevisiae binds to the DAL upstream induction sequence (UIS). Nucleic Acids Res *21*, 3777-3784.

Hinnebusch, A. G. (1984). Evidence for translational regulation of the activator of general amino acid control in yeast. Proc Natl Acad Sci U S A *81*, 6442-6446.

Hinnebusch, A. G. (1997). Translational regulation of yeast GCN4. A window on factors that control initiator-trna binding to the ribosome. J Biol Chem *272*, 21661-21664.

Hinnebusch, A. G., and Fink, G. R. (1983). Positive regulation in the general amino acid control of Saccharomyces cerevisiae. Proc Natl Acad Sci U S A *80*, 5374-5378.

Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S., and Young, R. A. (1998). Dissecting the regulatory circuitry of a eukaryotic genome. Cell *95*, 717-728.

Iraqui, I., Vissers, S., Bernard, F., de Craene, J. O., Boles, E., Urrestarazu, A., and Andre, B. (1999). Amino acid signaling in Saccharomyces cerevisiae: a permease-like sensor of external amino acids and F-Box protein Grr1p are required for transcriptional induction of the AGP1 gene, which encodes a broad-specificity amino acid permease. Mol Cell Biol *19*, 989-1001.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.

Jia, Y., Rothermel, B., Thornton, J., and Butow, R. A. (1997). A basic helix-loop-helix-leucine zipper transcription complex in yeast functions in a signaling pathway from mitochondria to the nucleus. Mol Cell Biol *17*, 1110-1117.

Kornitzer, D., Raboy, B., Kulka, R. G., and Fink, G. R. (1994). Regulated degradation of

the transcription factor Gcn4. Embo J *13*, 6021-6030.

Kuras, L., Barbey, R., and Thomas, D. (1997). 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 *16*, 2441-2451.

Kuras, L., Cherest, H., Surdin-Kerjan, Y., and Thomas, D. (1996). 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 *15*, 2519-2529.

Lanker, S., Bushman, J. L., Hinnebusch, A. G., Trachsel, H., and Mueller, P. P. (1992). Autoregulation of the yeast lysyl-tRNA synthetase gene GCD5/KRS1 by translational and transcriptional control mechanisms. Cell *70*, 647-657.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Liu, Z., and Butow, R. A. (1999). A transcriptional switch in the expression of yeast tricarboxylic acid cycle genes in response to a reduction or loss of respiratory function. Mol Cell Biol *19*, 6720-6728.

Mangan, S., Zaslaver, A., and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol *334*, 197-204.

Masselot, M., and De Robichon-Szulmajster, H. (1975). Methionine biosynthesis in Saccharomyces cerevisiae. I. Genetical analysis of auxotrophic mutants. Mol Gen Genet *139*, 121-132.

Meussdoerffer, F., and Fink, G. R. (1983). Structure and expression of two aminoacyl-tRNA synthetase genes from Saccharomyces cerevisiae. J Biol Chem *258*, 6293-6299.

Mirande, M., and Waller, J. P. (1988). The yeast lysyl-tRNA synthetase gene. Evidence for general amino acid control of its expression and domain structure of the encoded protein. J Biol Chem *263*, 18443-18451.

Mitchell, A. P., and Magasanik, B. (1984). Regulation of glutamine-repressible gene products by the GLN3 function in Saccharomyces cerevisiae. Mol Cell Biol *4*, 2758-2766.

Mosch, H. U., Scheier, B., Lahti, R., Mantsala, P., and Braus, G. H. (1991). Transcriptional activation of yeast nucleotide biosynthetic gene ADE4 by GCN4. J Biol Chem *266*, 20453-20456.

Mountain, H. A., Bystrom, A. S., and Korch, C. (1993). The general amino acid control regulates MET4, which encodes a methionine-pathway-specific transcriptional activator

of Saccharomyces cerevisiae. Mol Microbiol *7*, 215-228.

Natarajan, K., Meyer, M. R., Jackson, B. M., Slade, D., Roberts, C., Hinnebusch, A. G., and Marton, M. J. (2001). Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol Cell Biol *21*, 4347-4368.

Olive, M. G., Daugherty, J. R., and Cooper, T. G. (1991). DAL82, a second gene required for induction of allantoin system gene transcription in Saccharomyces cerevisiae. J Bacteriol *173*, 255-261.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., *et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Rolfes, R. J., and Hinnebusch, A. G. (1993). Translation of the yeast transcriptional activator GCN4 is stimulated by purine limitation: implications for activation of the protein kinase GCN2. Mol Cell Biol *13*, 5099-5111.

Roth, F. P., Hughes, J. D., Estep, P. W., and Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol *16*, 939-945.

Shemer, R., Meimoun, A., Holtzman, T., and Kornitzer, D. (2002). Regulation of the transcription factor Gcn4 by Pho85 cyclin PCL5. Mol Cell Biol *22*, 5395-5404.

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. Nat Genet *31*, 64-68.

Shore, D., and Nasmyth, K. (1987). Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. Cell *51*, 721-732.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106*, 697-708.

Springer, C., Kunzler, M., Balmelli, T., and Braus, G. H. (1996). Amino acid and adenine cross-pathway regulation act through the same 5'-TGACTC-3' motif in the yeast HIS7 promoter. J Biol Chem *271*, 29637-29643.

Talibi, D., Grenson, M., and Andre, B. (1995). Cis- and trans-acting elements determining induction of the genes of the gamma-aminobutyrate (GABA) utilization pathway in Saccharomyces cerevisiae. Nucleic Acids Res *23*, 550-557.

Thireos, G., Penn, M. D., and Greer, H. (1984). 5' untranslated sequences are required for the translational control of a yeast regulatory gene. Proc Natl Acad Sci U S A *81*, 5096-5100.

Thomas, D., Jacquemin, I., and Surdin-Kerjan, Y. (1992). 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 *12*, 1719-1727.

Valenzuela, L., Aranda, C., and Gonzalez, A. (2001). TOR modulates GCN4-dependent expression of genes turned on by nitrogen limitation. J Bacteriol *183*, 2331-2334.

Wek, S. A., Zhu, S., and Wek, R. C. (1995). The histidyl-tRNA synthetase-related sequence in the eIF-2 alpha protein kinase GCN2 interacts with tRNA and is required for activation in response to starvation for different amino acids. Mol Cell Biol *15*, 4497-4506.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997). Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol *15*, 1359-1367.

Yekta, S., Shih, I. H., and Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science *304*, 594-596.

Yu, L., Sabet, N., Chambers, A., and Morse, R. H. (2001). The N-terminal and C-terminal domains of RAP1 are dispensable for chromatin opening and GCN4-mediated HIS4 activation in budding yeast. J Biol Chem *276*, 33257-33264.

Zhou, K., Brisco, P. R., Hinkkanen, A. E., and Kohlhaw, G. B. (1987). Structure of yeast regulatory gene LEU3 and evidence that LEU3 itself is under general amino acid control. Nucleic Acids Res *15*, 5261-5273.

Chapter 5

Future Challenges for Interpreting the Transcriptional Regulatory Code

Introduction

While work using genome-wide location analysis to identify the cis-regulatory elements responsible for enabling gene-specific regulation in yeast provides an important resource for the study of genomics, in many ways it merely provides a rough foundation for other fundamental questions of molecular biology. Three questions in particular arise immediately from this work, the answers to which will be pursued in further experimentation and analysis.

Part I.

The first question is, "What factors contribute to the binding of some sequences in the genome in the absence of binding to apparently identical elements elsewhere?" For most regulators only a subset of the total sites matching their binding specificity are ever bound. For example, the binding site for Gcn4 occurs 3,323 times in the intergenic regions examined here. The number of these sites that coincide with binding ($P \leq 0.001$), however, is 295. Indeed, in the "post-genomic era," in which the DNA sequences of organisms including yeast and humans are now available, the power of genome-wide location analysis lies largely in its ability to identify which sites are actually bound in vivo, and as such, represent bona fide regulatory elements.

One explanation for this discrepancy lies in our understanding of what constitutes the specificity of a DNA-binding regulator. Discovery of regulatory motifs is a complex process subject to computational, biological and empirical constraints. Hence, what is believed to be a sequence sufficient for specifying protein binding may, in fact, not be. The subtleties of a stretch of regulatory DNA may not always be captured by current computational approaches. Continuing advances in computational methods for motif identification will no doubt contribute to our appreciation of such cryptic elements.

Many sites (even those bear little resemblance to the consensus) are capable of binding in the presence of sufficient levels of protein. Gcn4 preferentially binds the palindromic sequence TGASTCA (Arndt and Fink, 1986; Harbison et al., 2004). Under rich growth conditions in which Gcn4 levels are low, approximately 70% of targets have a close match to this consensus. As Gcn4 levels accumulate under conditions of

amino acid starvation, however, the additional targets bound tend to contain slightly weaker matches to this sequence. Hence, given a limiting amount of protein, it seems that only the highest affinity sites throughout the genome will be consistently bound.

Similarly, many predicted binding specificities are highly degenerate, that is, the bound protein tolerates a substantial degree of sequence variation. Such low levels of specificity may be due to the intrinsic DNA-binding ability of the protein. For example, the binding specificity of the regulator Mcm1 is CCWWWWWWGG (Jarvis et al., 1989). As discussed in Chapter 1, interactions between Mcm1 with α2 (or alternatively with regulators Arg80, Ste12 or Fkh2) lead to changes in the binding behavior of the resulting heterodimer (El Bakkoury et al., 2000; Errede and Ammerer, 1989; Kumar et al., 2000; Primig et al., 1991). In these cases, the specificity of binding is not conferred solely by the content of the bound DNA sequence, but rather by interactions with other binding proteins.

We have found that there appears to be a preference within yeast for regulatory motifs to occur within a certain distance of the translational start site of an associated ORF. This may be contrasted with transcription in higher eukaryotes, which is characterized by regulatory elements that frequently act at large distances. Nevertheless, it is likely that sites that are sufficiently distant (in either linear or three-dimensional terms) from core promoters fail to maintain strong interactions with DNA-binding proteins as a result of the loss of the reciprocal stabilization found in association with an appropriately assembled transcriptional complex.

Recently, a powerful technique that helps to identify the minimal DNA sequence content required for protein binding in vitro has been developed (Bulyk et al., 1999;

Bulyk et al., 2001). Protein binding microarrays (PBM's) consist of microarrays containing double-stranded DNA of known sequence. Purified DNA-binding protein is added to the arrays under conditions which allow for specific binding. These proteins can be detected by either introducing a reporter epitope (e.g. GFP) or through the recognition of fluorophore-conjugated antibodies. Such methods can be used to calculate the binding affinities for proteins, can identify the sequences that serve as the best targets of recognition, and, in general, could help determine whether a sequence has the inherent *potential* for binding even if it is not identified as being bound in vivo.

A second technical limitation lies in the detection of protein binding in vivo. For genome-wide location analysis, as for other microarray-based technologies, one major challenge is the discrimination of signal from a noisy background. We have attempted to overcome this through the use of triplicate experiments that are then fed into an error model that assigns a probability that a DNA-protein interaction is due to chance. Because we generally use stringent thresholds ($P \leq 0.001$) in analyzing our data, we are necessarily excluding "real" interactions that fail to meet this somewhat arbitrary threshold. Better understanding of the systematic noise associated with this technique, analyses that make use of rank-ordered metrics, and improvements in our error model offer hope for reducing this rate of false positives and better capturing genuine binding events.

Even with perfect knowledge of regulator binding in cells grown under a particular condition, we have seen that knowledge of environment-dependent binding is critical to the identification of the entire set of targets bound by a transcriptional regulator. While we have attempted to select conditions in which a subset of our 203

regulators are believed to be biologically active, the number of possible environmental conditions to be tested is limitless. Additional protein-DNA interactions will be found as more regulators are profiled under conditions other than growth in rich medium.

As mentioned in Chapter 1, a major contributing factor in constraining the activity of DNA-binding proteins is control over the accessibility of DNA. The most obvious candidate for preventing the binding of regulators to DNA is alteration of chromatin. The positioning of nucleosomes is known to serve as a mechanism for regulator binding (Adams and Workman, 1993; Han and Grunstein, 1988; Lee et al., 2004). Some fraction of unbound sites matching a binding specificity are likely excluded from interacting with regulators due to occupancy by these histone complexes. Other sites are made inaccessible through higher-order structures that lead to highly condensed regions of the genome (heterochromatin). Recent studies have characterized the global role of histone modifying and remodeling proteins, acetylated and methylated histones, and proteins associated with chromatin (Kurdistani et al., 2002; Kurdistani et al., 2004; Lee et al., 2004; Lieb et al., 2001; Ng et al., 2002; Robert et al., 2004). Much work remains to be done, however, in synthesizing these data with our understanding of transcriptional regulator binding.

Finally, the role of both direct and indirect interactions among DNA-binding proteins in controlling their association with DNA is of extreme importance (Remenyi et al., 2004; Wolberger, 1998). The binding of some proteins to neighboring regions of DNA may be required for enabling another protein to bind. Conversely, the binding of one regulator may be prevented by the binding of a competing protein to an identical or overlapping site. For example, we have observed that the regulators Cbf1, Pho4 and

Ino2, all members of the helix-loop-helix family of transcriptional regulators, recognize a nearly identical consensus binding sequence (CACGTG). Furthermore, all of these regulators apparently bind to the same site upstream of the *PHO86* gene, which encodes a phosphate transporter. This binding, however, is contingent upon environmental conditions. Under rich growth conditions, Ino2, but not Pho4 or Cbf1, is bound. However, under phosphate-depleted conditions, Pho4 is bound; and under amino acid starvation conditions Cbf1 binding is observed. One model that explains these results is that these proteins are engaged in competition for this binding site and that changes in their relative abundance allow for binding of one to the exclusion of the others.

In order to identify physical or regulatory interactions that might contribute to binding behavior, we have identified a set of "co-occurring" motifs and corresponding factors. Other recent work has also been directed at finding such "word pairs" (Bulyk et al., 2004; Chiang et al., 2003; GuhaThakurta and Stormo, 2001), in at least one case even predicting the identity of a key regulatory partner on the basis of neighboring sequence (Mootha et al., 2004). While some motifs may co-occur because their cognate binding proteins are evolutionarily conserved and functionally redundant (e.g. stress response regulators Msn2 and Msn4), an alternate explanation for their co-occurrence is that protein-protein interactions occur between the regulators that bind them. Another possibility is that cooperative binding effects occur between such proteins in the absence of contact between them. A systematic analysis of these pairs of regulators might reveal the nature of any interactions between them. For example, deletion of one partner could result in the complete ablation of binding by the other, a reduction in its number of targets or its relocation to a different set entirely. At least one study reports just such

effects for pairs of combinatorial regulators (Zeitlinger et al., 2003).

Part II

A second fundamental question that will serve as the basis for future investigations is, "How does the binding of transcriptional regulators translate into changes in expression?" The simplest model of transcriptional control of the genome is one in which binding by a regulator to a promoter region correlates with a change in expression with the corresponding gene. For example, Bas1 is a regulator of adenine biosynthesis. In cells grown in minimal medium, Bas1 occupies the promoter regions of *ADE2*, *ADE3*, *ADE5,7*, and *ADE8*, all enzymes involved in purine metabolism. These genes also increase in expression in cells grown in the presence of limiting amounts of adenine.

Such simple models of regulatory control, however, are insufficient to explain entire expression programs. In some cases, the activity of a regulator does not uniformly correspond with activation or repression. The Arg80/81 complex, for example, operates under conditions of arginine abundance to induce expression of arginine catabolic enzymes, but also to repress expression of genes encoding arginine biosynthetic genes (De Rijcke et al., 1992). Similarly, Abf1 has a role in both transcriptional activation as well as transcriptional silencing (Buchman and Kornberg, 1990; Diffley and Stillman, 1989).

A second limitation of this simple model is that binding by some regulators is necessary, but not sufficient, for changes in gene expression. In such cases, an additional level of regulation is required. As discussed in Chapter 1, for example, the activation potential of genomically bound Leu3 is achieved only upon its association with a leucine

metabolite. Conversely, it can be imagined that even a transient (and undetected) association of some regulators with a binding site, could lead to a long lasting effect on gene expression. Generally, however, we find that promoter regions of highly expressed genes are more likely to be bound by a regulator than those of genes expressed at low levels.

We have observed that nearly half of all genes in yeast are associated with the binding of multiple regulators. Combinatorial binding of regulators is thought to be a mechanism for maximizing the flexibility of regulatory control with a minimal number of regulators. The exact complement of bound proteins in proximity to a gene can profoundly affect its transcriptional activity. Some steps have been taken to identify "modules" of genes that exhibit coherent expression patterns and are bound by a common set of transcriptional regulators (Bar-Joseph et al., 2003b; Lee et al., 2002). Briefly, commonly bound gene promoters are identified in location data using strict thresholds. The corresponding genes are then analyzed to determine, for a subset, if a close correlation of their expression exists. Finally, on the basis of a shared correlation of expression, additional genes are qualified for inclusion in the module by relaxing the significance threshold for binding. One limitation of such approaches, however, is that they rely on correlation of genes across collections of hundreds (or even thousands) of expression experiments. Consequently, genes whose expression is correlated only under a limited number of biologically relevant experiments may not be selected for inclusion. An alternative approach is to build up networks using a carefully chosen selection of high quality expression data matched to location experiments for individual conditions.

As analysis of protein partnerships may determine genomic binding locations,

so too would it help to define the individual contributions of multiple regulators to expression. Again, perturbations of regulatory networks in the form of targeted deletions of DNA-binding proteins can lead to elucidation of whether regulators bound to common sets of genes interact in antagonistic, additive or synergistic ways.

A last important element to consider in comparisons of binding and expression data is that of time. It is important to remember that expression programs are not discrete events, but continue over the course of minutes, hours, or even days. Even genes that are induced by the same stimulus may differ in their expression response with respect to time. This may be due to variations in binding site affinity, the role of other transcriptional regulators or differences in the ordered recruitment of chromatin regulators. Even methods that are capable of deconvolving co-regulated groups of genes (Bar-Joseph et al., 2003a; Spellman et al., 1998) are of limited value in these comparisons if the time point selected for regulator profiling is inappropriate. The next step in this analysis is a well-sampled time course that measures the changes over time in the binding of a set of regulators (for example upon exposure to peroxide) that can then be analyzed against a similar backdrop of changes in gene expression.

Part III

Finally, we are interested in answering the question, "Can changes in genomic binding profiles help identify mechanisms for the control of transcriptional regulators?" While the types of mechanisms employed in the regulation of these proteins are wide-ranging (Chapter 1), we suggest that the changes observed in the binding of a particular regulator under different conditions can inform investigations into the most likely regulatory mechanisms involved in its control. Such information would be particularly valuable in studies of human transcriptional regulators, whose normal function is required to prevent disease, and which may regulate different sets of genes not only in a condition-specific, but also in a cell type-specific manner. Even now, it appears that examining comparisons of binding profiles may provide new insight (and challenge accepted models of regulatory behavior) for even well studied human transcription regulators (personal communication).

We are currently engaged in testing the predictive power of models of regulatory behavior (Harbison et al., 2004) in yeast. For a subset of these regulators we intend to collect information about the environment-dependent changes in its total abundance, cellular localization and modification state. This information will require fusing data from location analysis with that derived from microscopy, quantitative ELISAs and mass-spectrometry. Beyond the confirmation of current predictions, such information can be used to further refine models of behavioral mechanisms. Ultimately, insights derived from differences in binding behavior of a single regulator profiled from cells grown under different conditions, with different genetic backgrounds or of different cell types

may help to constrain models of regulatory behavior and expedite investigations into the mechanisms by which it operates.

References

Adams, C. C., and Workman, J. L. (1993). Nucleosome displacement in transcription. Cell *72*, 305-308.

Arndt, K., and Fink, G. R. (1986). 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 *83*, 8516-8520.

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D. K., and Jaakkola, T. S. (2003a). Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci U S A *100*, 10146-10151.

Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S., Young, R. A., and Gifford, D. K. (2003b). Computational discovery of gene modules and regulatory networks. Nat Biotechnol *21*, 1337-1342.

Buchman, A. R., and Kornberg, R. D. (1990). A yeast ARS-binding protein activates transcription synergistically in combination with other weak activating factors. Mol Cell Biol *10*, 887-897.

Bulyk, M. L., Gentalen, E., Lockhart, D. J., and Church, G. M. (1999). Quantifying DNA-protein interactions by double-stranded DNA arrays. Nat Biotechnol *17*, 573-577.

Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A *98*, 7158-7163.

Bulyk, M. L., McGuire, A. M., Masuda, N., and Church, G. M. (2004). A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in Escherichia coli. Genome Res *14*, 201-208.

Chiang, D. Y., Moses, A. M., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. Genome Biol *4*, R43.

De Rijcke, M., Seneca, S., Punyammalee, B., Glansdorff, N., and Crabeel, M. (1992). Characterization of the DNA target site for the yeast ARGR regulatory complex, a sequence able to mediate repression or induction by arginine. Mol Cell Biol *12*, 68-81.

Diffley, J. F., and Stillman, B. (1989). Similarity between the transcriptional silencer binding proteins ABF1 and RAP1. Science *246*, 1034-1038.

El Bakkoury, M., Dubois, E., and Messenguy, F. (2000). Recruitment of the yeast MADS-box proteins, ArgRI and Mcm1 by the pleiotropic factor ArgRIII is required

for their stability. Mol Microbiol *35*, 15-31.

Errede, B., and Ammerer, G. (1989). STE12, a protein involved in cell-type-specific transcription and signal transduction in yeast, is part of protein-DNA complexes. Genes Dev *3*, 1349-1361.

GuhaThakurta, D., and Stormo, G. D. (2001). Identifying target sites for cooperatively binding factors. Bioinformatics *17*, 608-621.

Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. Cell *55*, 1137-1145.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J. B., Reynolds, D. B., Yoo, J., *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99-104.

Jarvis, E. E., Clark, K. L., and Sprague, G. F., Jr. (1989). The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the MCM1 gene. Genes Dev *3*, 936-945.

Kumar, R., Reynolds, D. M., Shevchenko, A., Goldstone, S. D., and Dalton, S. (2000). Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. Curr Biol *10*, 896-906.

Kurdistani, S. K., Robyr, D., Tavazoie, S., and Grunstein, M. (2002). Genome-wide binding map of the histone deacetylase Rpd3 in yeast. Nat Genet *31*, 248-254.

Kurdistani, S. K., Tavazoie, S., and Grunstein, M. (2004). Mapping global histone acetylation patterns to gene expression. Cell *117*, 721-733.

Lee, C. K., Shibata, Y., Rao, B., Strahl, B. D., and Lieb, J. D. (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet *36*, 900-905.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet *28*, 327-334.

Mootha, V. K., Handschin, C., Arlow, D., Xie, X., St Pierre, J., Sihag, S., Yang, W., Altshuler, D., Puigserver, P., Patterson, N., *et al.* (2004). Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. Proc Natl Acad Sci U S A *101*, 6570-6575.

Ng, H. H., Robert, F., Young, R. A., and Struhl, K. (2002). Genome-wide location and regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev *16*,

806-819.

Primig, M., Winkler, H., and Ammerer, G. (1991). The DNA binding and oligomerization domain of MCM1 is sufficient for its interaction with other regulatory proteins. Embo J *10*, 4209-4218.

Remenyi, A., Scholer, H. R., and Wilmanns, M. (2004). Combinatorial control of gene expression. Nat Struct Mol Biol *11*, 812-815.

Robert, F., Pokholok, D. K., Hannett, N. M., Rinaldi, N. J., Chandy, M., Rolfe, A., Workman, J. L., Gifford, D. K., and Young, R. A. (2004). Global Position and Recruitment of HATs and HDACs in the Yeast Genome. Mol Cell *16*, 199-209.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273-3297.

Wolberger, C. (1998). Combinatorial transcription factors. Curr Opin Genet Dev *8*, 552-559.

Zeitlinger, J., Simon, I., Harbison, C. T., Hannett, N. M., Volkert, T. L., Fink, G. R., and Young, R. A. (2003). Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell *113*, 395-404.

Appendix A

Supplementary Materials
for "Transcriptional Regulatory Code of a Eukaryotic Genome"
(Chapter 3)

Part I: Supplementary Methods

This paper describes the genomic location of 203 transcriptional regulators, a subset of which are examined under different environmental conditions. We previously reported the genomic binding information for 106 regulators profiled in a single growth condition[1]; we have repeated experiments for 44 of these regulators to improve the quality of the complete dataset (available at http://web.wi.mit.edu/young/regulatory_code). We have also introduced additional data analysis features to reduce noise and improve the results.

Genetic Reagents

The 203 transcriptional regulators were identified by searching the YPD and MIPS databases[2-4] for known and predicted transcription factors and nucleic acid binding proteins. Yeast strains were created for each of the 203 regulators in which a repeated Myc epitope coding sequence was integrated into the endogenous gene encoding the regulator. PCR constructs containing the Myc epitope coding sequence and a selectable marker flanked by regions of homology to either the 5' or 3' end of the targeted gene were transformed into the W303 yeast strain Z1256. Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and Western blotting, respectively.

Growth conditions

Regulators were selected for profiling in a specific environment if they were essential for growth in that environment or if there was other evidence implicating them in regulation of gene expression in that environment.

A brief description of the environmental conditions used follows:

Rich media. Cells were grown in YPD (1% yeast extract/2% peptone/2% glucose) to an OD600 of ~0.8.

Highly hyperoxic. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment with hydrogen peroxide (4 mM final) for 30 minutes.

Moderately hyperoxic. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment with hydrogen peroxide (0.4 mM final) for 20 minutes.

Amino acid starvation. Cells were grown to an OD600 of ~0.6 in synthetic complete medium followed by treatment with the inhibitor of amino acid biosynthesis sulfometuron methyl (0.2 μg/ml final) for two hours.

Nutrient deprived. Cells were grown in YPD to an OD600 of ~0.8 followed by treatment with rapamycin (100 nM final) for 20 minutes.

Filamentation inducing. Cells were grown in YPD containing 1% butanol for either 90 minutes or 14 hours (corresponding to an OD600 of ~0.8).

Mating inducing. Cells were grown in YPD to an OD600 of ~0.8 followed by treatment with the alpha factor pheromone (5 μg/ml) for 30 minutes.

Elevated temperature. Cells were grown in YPD at 30°C to an OD600 of ~0.5 followed by a temperature shift to 37°C for 45 minutes.

Galatose medium. Cells were grown in YEP medium supplemented with galactose (2%) to an OD600 of ~0.8.

Raffinose medium. Cells were grown in YEP medium supplemented with raffinose (2%) to an OD600 of ~0.8.

Acidic medium. Cells were grown in YPD to an OD600 of ~0.5 followed by treatment for 30 minutes with succinic acid (0.05 M final) to reach a pH of 4.0.

Phosphate deprived medium. Cells were grown in synthetic complete medium lacking phosphate to a final OD600 of ~0.8.

Vitamin deprived medium. Cells were grown in synthetic complete medium lacking thiamin to a final OD600 of ~0.8.

Genome-wide Location Analysis

Genome-wide location analysis was performed as previously described[1,5,6]. Bound proteins were formaldehyde-crosslinked to DNA in vivo, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-myc antibody, followed by reversal of the crosslinks to separate DNA from protein[7,8]. Immunoprecipitated DNA and DNA from an unenriched sample were amplified and differentially fluorescently labeled by ligation-mediated PCR. Triplicate samples were hybridized to a microarray consisting of spotted PCR products representing the intergenic regions of the *S.cerevisiae* genome. Detailed protocols are available on the authors' website.

Microarray design

Using the Yeast Intergenic Region Primer set (Research Genetics) we PCR amplified and printed approximately 6000 DNA fragments, representing essentially all of the known intergenic regions in the yeast genome[9]. The average size of the spotted PCR products was 480 bp, and the sizes ranged from 60 bp to 1500 bp.

Raw Data Analysis

The microarrays were scanned using an Axon200B scanner, and the images were analyzed with Genepix 5.0. Columns corresponding to the background subtracted intensities and standard deviation of the background were extracted for further analysis. The intensities for the two channels, representing the immunoprecipitated (test) and

unenriched (control) samples, were normalized by using the median of each channel to calculate a normalization factor, normalizing all datasets to a single median intensity. The log ratio of the intensity in the test channel to the control channel was calculated. To account for biases in the immunoprecipitation reaction, these log ratios were normalized for each spot by subtracting the average log ratio of each spot across all arrays. The intensities in the test channel were then adjusted to yield this normalized ratio. Finally, an error model[10] was used to calculate significance of enrichment on each chip and to combine data for replicates to obtain a final average ratio and significance of enrichment for each intergenic region. Each intergenic region was assigned to the genes it is most likely to regulate, as described on the author's website.

We have included new refinements in our analysis relative to that used in Lee et al.[1]. Notably, we have excluded artefactual spots from analysis, selected more reliable probes for normalization and assigned quality metrics to individual arrays to identify low quality experiments.

Error Estimates

We previously estimated a false positive rate of 6-10% for genome-wide binding data that meets a $P \leq 0.001$ threshold. The present study is focused on DNA regions that are both bound ($P \leq 0.001$) and contain a conserved match to a binding site specificity. Of 47 sites that were used by Lee et al.[1] to determine the error rate and that met our criteria for binding sites, 45 were confirmed by independent gene-specific ChIP experiments. Thus, the frequency of false positives in this dataset is likely to be approximately 4%.

The false negative rate is more difficult to estimate, but it is likely to be approximately 24% in the present genome location dataset. This estimate was derived by determining the number of binding interactions reported in the literature for cell cycle regulators that were not identified in the genome-wide location data at $P \leq 0.001$ and associated with conserved binding sites (12/50). We selected the cell cycle literature for analysis because of the extensive study of this group of regulators and their targets.

Motif Discovery Overview

Binding motifs were identified in a five-step process described in detail below and summarized in Supplementary Figure 2. First, motifs were discovered by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance using uniform metrics and then clustered to yield representative motifs. Conservation-based metrics were used to identify the highest-confidence subset of these motifs. For cases in which multiple significant binding motifs were found for a factor, we used statistical scores or information from the Transfac[11], YPD[12], and SCPD[13] databases to choose a single motif for each regulator. Sequence input files, intermediate motif discovery output, and matrix representations of the finalized motifs are available on the authors' website.

Step 1: Initial Motif Discovery

Motif Discovery Programs have different strengths with respect to finding specificities. To gain as comprehensive an analysis as possible, we applied five different motif-

finding programs to the binding data: AlignACE[14], MEME[15], MDscan[16], the conservation-based method described in Kellis et al.[17] , and a new conservation-based method called CONVERGE (described below). The MEME program was also used to analyze a modified input that incorporated conservation information (see "Probe Sequences").

To make the search more thorough, we ran each of these programs multiple times with different parameters. AlignACE was run using the default settings ten times with different random number seeds, in order to increase the motif space it sampled. The results from the AlignACE runs were grouped together for analysis. MEME was run using the supplied 5th-order Markov background model, the "ZOOPS" motif model, and the "-minsites 20 -dna -revcomp" options. MEME runs were repeated using motif width ranges of 7 to 11 and 12 to 18. To run MDscan, seqeuences were ranked according the *P*-value of binding, and the program was run with the "-s 30 -r 5 -t 10" options. To compensate for the fact that MDscan searches only for motifs of fixed width, the program was run repeatedly, once with each width in the range 8 to 15 bases. The method of Kellis et al. was applied to the data as described[17]. CONVERGE was run twice using motif widths of 8 and 15.

MEME_c
We tested whether we could improve the performance of AlignACE, MEME and MDscan by modifying the input sequences to convey the conservation of each base in the *sensu stricto Saccharomyces* species. Using ClustalW[18] alignments for the *sensu stricto* species[17], we replaced a base in the *Saccharomyces* genome with the letter "N" if it was not conserved in 2/3 or 3/4 of the other genomes. Of the programs we tested, only MEME was able to use the modified sequences.

CONVERGE
We designed CONVERGE to identify motifs that are both over-represented in a set of input sequences and conserved across multiple genomes. CONVERGE input sequences consists of an ungapped DNA sequence corresponding to the primary genome, as well as one or more optional aligned sequences, which may contain gaps. The algorithm is based on the ZOOPS model of MEME and uses a 5th-order Markov background model. However, whereas MEME searches for matches to a motif model across a set of input sequences, CONVERGE searches across the multiple-sequence alignments for each sequence. Specifically, CONVERGE treats the probability of a motif occurring at a site in the alignment as the product of the probabilities of the motif occurring at the same site in each of the aligned sequences. Thus, CONVERGE defines a site as conserved in a flexible manner that depends on the motif being discovered. Full details will be presented elsewhere.

Probe Sequences
Motif discovery programs were applied to the sequences of probes bound with a *P*-value ≤ 0.001. We found that some intergenic regions were highly homologous over their entire length, and consequently skew the results of motif discovery since all subsequences are overrepresented. To remove this bias, we used BLAST[19] to identify

pairs of probes with high sequence similarity over 50% of their lengths. For each pair, the shorter intergenic region was omitted from motif discovery computations. This process removed up to nine regions for some experiments, but less than one on average.

To determine the sequences present on the microarrays, we computed the expected products of the PCR used to construct the arrays. Research Genetics primer sequences were obtained from http://www.resgen.com/products/YeIRP.php3 and the March 2002 revision of the yeast genome was obtained from SGD[20]. Probes that were predicted to amplify more than two different genomic sequences were omitted from the calculations. Twenty five probe sequences neighboring repetitive, non-transcribed features (e.g. telomeric repeats, X elements and Y' elements) were also omitted.

PSSM Representation

Motifs from all programs were converted to a standard position-specific scoring matrix (PSSM) for subsequent analysis. AlignACE and MDscan produce alignments of binding sites, and these were first converted into matrices representing the frequency of each base (A, C, G, T) at each position of the alignments. The method of Kellis et al. represents motifs as text strings containing ambiguity codes, which were also converted to matrices of frequencies. (For example, if a motif contained the letter "S" at a particular position, a value of 0.5 would be assigned to both "C" and "G.") The matrices of base frequencies were converted to probabilities and then were adjusted with 0.001 pseudo-counts in proportion to the 0^{th}-order background probabilities (3.1×10^{-4} pseudocounts for A and T, 1.9×10^{-4} pseudocounts for G and C). Log-likelihood scores were computed by dividing the estimated probabilities by the background probability for each letter and computing the base-2 logarithm. CONVERGE and MEME both provide probability matrices, which were used directly.

Step 2: Motif Scoring and Significance Testing

We tested the significance of each motif by comparing how often it was found in the bound and unbound probes. To encapsulate different approaches to measuring motif over-representation, we employed three different metrics: Enrichment, ROC AUC, and for motifs discovered by the method described in Kellis et al., the "CC4" score. The enrichment score is a direct measure of the occurrence of a motif among bound probes compared to all possible gene targets, but does not distinguish between the number of motifs occurrences within each intergenic region. The ROC AUC metric is more sensitive to cases in which the number of motif occurrences is a distinguishing factor. Finally, the CC4 metric provides a way to account for the importance of the conservation of the motif among bound probes. These scores were compared to significance thresholds obtained from calculations on randomized selections of intergenic regions as described below in "Significance Thresholds"

Enrichment score

To obtain the enrichment score, the hypergeometric distribution was used to compare the frequency of the motif in the bound probes to that which would be expected if the intergenic regions were selected at random from the genome. A sequence was considered to contain a motif if it contained at least one or more sites scoring at least 70% of the maximum possible score of the matrix.

A *P*-value for the enrichment was computed according to the formula:

$$p = \sum_{i=b}^{\min(B,g)} \frac{\binom{B}{i}\binom{G-B}{g-i}}{\binom{G}{g}} \qquad (5)$$

where B is the number of bound intergenic regions and G is the total number of intergenic regions represented on the microarray (or the genome). The quantities b and g represent the number of intergenic regions of B and G matching the motif. The quantity $-log_{10}(p)$ is referred to as the enrichment score.

*ROC AUC (*Receiver Operating Characteristic Area Under Curve)
The ROC AUC refers to the area under a receiver operating characteristic curve which is assembled by ranking the sets of bound and unbound probes according to the number of motif matches they contain, and plotting the fractional rankings against each other. We used the method and code described by Clarke and Granek[21].

Conservation CC4
Motifs discovered using the method of Kellis et al.[17] were judged according to the CC4 metric, in which the occurrence of a conserved motif among the bound probes is compared to the expected ratio observed among all 3-gap-3 motifs in among the same set of bound probes. The binomial probability of the observed ratio was computed, and is reported in terms of the equivalent z-score.

Significance Thresholds
We observed that motif discovery programs produce motifs with high over-representation metrics (such as "Enrichment" and "ROC AUC") even when applied to random selections of intergenic regions. To identify the true motifs, we converted the scores from each metric into the empirical probability that a motif with a similar score could be found by the same program in randomly selected sequences. We accepted only those motifs with a *P*-value ≤ 0.001. We selected this stringent threshold to minimize false positives, and because we observed empirically that it identified the correct motifs for many regulators with known specificity. To estimate these thresholds, we ran each program 50 times on randomly selected sequences on sets of 10, 20, 30, 40, 50, 60, 70, 80, 100, 120, 140, and 160 probes.

The observed scores from these random runs were parameterized by a normal distribution. The critical values equivalent to a *P*-value of 0.001 are provided in Supplementary Table 8 for each program and each metric. If the empirical distribution was not normal (by the Shapiro-Wilk test), the corresponding metric was not used to evaluate motifs generated by the relevant program for regulators with a similar number of bound probes.

For a particular experiment, we employed the threshold derived from the randomization set that had the size closest to the number of bound probe sequences. For example, suppose a motif found by performing ten runs of AlignACE on 32 intergenic

sequences had an enrichment score of 25. The relevant score distribution has been obtained by performing ten runs of AlignACE on each of 50 randomly selected sets of 30 intergenic sequences. The resulting distribution of enrichment scores has a mean of 14.1 and standard deviation of 2.1, and the enrichment that corresponds to significance of $P \leq 0.001$ is thus 20.43. Since the score of the candidate motif is higher, it is considered significant.

Step 3: Motif Clustering and Averaging

K-medoids Clustering

The set of significant motifs for each experiment was then clustered via k-medoids clustering[22] using the distance metric described below. The k-medoids algorithm was performed 500 times to find a clustering with a minimal sum of inter-cluster distances. To find the optimal number of clusters, this process was first performed with 10 clusters, and then repeated with incrementally fewer clusters until all average distances between members of a cluster and medoids of other clusters were sufficiently large (greater or equal to 0.18).

Inter-Motif Distance

We constructed a distance metric to aid in the comparison of motifs. The distance D between two aligned motifs "a" and "b" is defined as,

$$D(a,b) = \frac{1}{w}\sum_{i=1}^{w}\frac{1}{\sqrt{2}}\sum_{L\in\{ACGT\}}(a_{i,L} - b_{i,L})^2 \qquad (1)$$

where w is the motif width, and $a_{i,L}$ and $b_{i,L}$ are the estimated probabilities of observing base L at position i of motifs a and b, respectively. The normalizations by w and $\sqrt{2}$ facilitate the interpretation as a fractional distance. For example, a distance of 0.20 indicates that the two motifs differ by about 20%.

In practice, the optimal alignment of motifs is not known. We therefore use the minimum distance between motifs among all alignments in which the motifs overlap by at least seven bases, or when the motifs are shorter, by 2 bases fewer than the shortest motif length. Alignments to the reverse complements of the motifs are included.

Motif Averaging

A single motif representing each cluster was computed by averaging the probabilities at each matrix position of the aligned motifs comprising the cluster. Low-information positions on the flanks of the averaged motifs were removed.

Step 4: Conservation Testing for Averaged Motifs

We tested the conservation of averaged motifs, and focused subsequent analysis on the motifs that met two conservation criteria: First, we required that the frequency of conserved instances of the motif compared to all instances of the motif be at least as high within bound intergenic regions as among all intergenic regions. Second, we required that discovered motifs have at least three conserved instances that are bound.

We considered a sequence a match to a motif if it had a score of at least 60% of the motif maximum. We defined a "conserved instance" to mean that the aligned sequence of at least two other *sensu stricto* species also matched the motif. In cases where fewer than two aligned sequences were available, a site was treated as "not conserved."

Step 5: Assignment a Single Motif to Each Regulator

Often, the motif discovery process produced several significant, distinct averaged motifs (3 on average.). These motifs could represent the desired binding specificity of the protein, or they might arise from the specificity of binding partners or have other biological significance. To identify those motifs representing the binding specificity of the profiled transcription factor, we compared the specificities to binding data in the Transfac[11], YPD[12], and SCPD[13] databases, when available, using the same inter-motif distance metric used for clustering (see above.) There were 21 regulators for which no such data were available. In these cases we chose the motif with the best enrichment score.

Specificity data from these databases is sometimes available in the forms of raw sequences, ambiguity codes, and matrices. For regulators without matrices, we assembled a single consensus sequence to represent the body of experimentally determined specificity information and converted it to a PSSM as described above. Since there is no way to independently assess the quality of the motifs assembled from the databases, we used a permissive threshold to detect similarity between the discovered motifs and the database motifs. Motifs scoring below 0.24 were accepted as matches, while motifs with scores less than 0.35 were examined manually. The scores for the motifs that were used in the Regulatory Code Map are provided in Supplementary Table 2.

Motifs Derived from the Literature

We used a motif derived from the databases for the remaining regulators for which either: (1) Too few intergenic regions (<10) were bound for effective motif discovery, (2) discovered motifs similar to the literature were eliminated by the conservation in Step 4, or (3) none of the discovered motifs matched the literature in Step 5. These motifs were only included if they had at least one conserved instance that was bound. The resulting compendium of 102 motifs (Supplementary Table 3) was used in all subsequent analysis.

Regulatory Code Map

Binding motifs for 102 regulators (Supplementary Table 3) were fused with location analysis data and conservation data to produce a map of active binding sites in intergenic regions. The entire map is available at http://web.wi.mit.edu/fraenkel/regulatory_map/. The map was constructed by finding all conserved occurrences of each motif within intergenic regions bound by the corresponding factor.

We used a binding P-value threshold of $P \leq 0.001$ and the definition of conservation as described in the "Conservation Test" section above. Variants of the map constructed with different binding and conservation thresholds are also available online.

Distributions of distances from the start codon (ATG) of open reading frames to binding sites in the adjacent upstream region were derived from the above data. These were compared to a distribution calculated on ten thousand "randomized" genomes in which the binding sites in each intergenic region were redistributed randomly and independently between the adjacent genes. The region from –100 to –500 (grey area in Figure 2c) contains many more binding sites than expected.

Promoter Classification

Promoters were classified based on the aggregate binding data from all experiments. A promoter was defined as having multiple regulator architecture if more than one regulator bound in the aggregate data, regardless of the number of regulators that bound in any particular condition. Similarly, a promoter was assigned to the single regulator architecture if it was bound by exactly one regulator in the aggregate data.

Regulators that had a tendency to use the repetitive motif architecture were identified by chi-square analysis. For each regulator, we calculated the number of promoters containing a single site and the number containing multiple sites. These values were then compared to the expected values based on the average for all factors.

Co-occurring regulatory motifs were determined based on P values representing the probability, based on the hypergeometric distribution, of finding the observed number of intergenic regions (or more) bound by both regulators under the null hypothesis that binding for the two regulators is independent.

Regulator Behaviour Classification

The binding of each regulator was compared in pair-wise fashion for every environmental condition in which that regulator was studied. Only regions bound at $P \leq 0.001$ and containing conserved matches to the corresponding motif were included in this analysis. Some regulators fall into multiple categories depending on exactly which conditions are compared.

For the "condition invariant" category the ratio of the overlap of bound probes for a regulator was greater than 0.66, and the ratio of the number of bound probes was between 0.66 and 1.5.

For the "condition enabled" category the regulator bound to no probes in one environment.

For the "condition expanded" category the ratio of the overlap of bound probes for a regulator was greater than 0.66, and the ratio of the number of bound probes was less than 0.66 or greater than 1.5.

For the "condition altered" category the regulator bound at least one probe in both environments and the ratio of the overlap of bound probes was less than 0.66.

Experimental Confirmation of Predicted Specificity

We compared the discovered motifs to those in the literature using an automated method, and selected the regulator for which the discrepancy was the greatest, Cin5 (Supplementary Table 2). The discovered motif, TTAcrTAA, contains a one base insertion compared to the previously reported site[23], TTACTAA. The previously known site is poorly enriched in the probes bound by Cin5 ($P \leq 0.02$), while the discovered motif is very strongly enriched ($P \leq 10^{-38.4}$).

We used a gel-shift assay to test whether the specificity for Cin5 that we inferred from our in vivo data also represented the in vitro properties for this regulator (Supplementary Figure 3). The DNA-binding domain of Cin5 was cloned into a derivative of the pET-32 vector (Novagen) fused to thioredoxin and a poly-histidine peptide, expressed in *E. coli*, and purified by affinity chromatography. Protein was incubated with a Cy5-labeled oligonucleotide containing the sequence gcgacaTTACCTAAgggc and challenged with unlabeled competitor containing either the same sequence or the previously published binding site (gcgacaTTACTAAagggc[23]). The reactions were analyzed on 10% acrylamide gels run in 0.5x TBE. Similar results were obtained for a probe containing the core sequence of TTACGTAA.

Bibliography

1. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804. (2002).
2. Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* **25**, 28-30 (1997).
3. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73 (1999).
4. Costanzo, M. C. et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75-9 (2001).
5. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9. (2000).
6. Simon, I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697-708. (2001).
7. Aparicio, O. M. *Current Protocols in Molecular Biology* (ed. al., F. M. A. e.) (John Wiley and Sons, New York, 1999).
8. Orlando, V. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* **25**, 99-104 (2000).
9. Tessier, D. et al. *A DNA Microarrays Fabrication Strategy for Research Laboratories*. (eds. Rehm, H. & Reed, G.) (Wiley-VCH, Weinheim, Germany, 2002).
10. Hughes, T. R. et al. Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-26 (2000).
11. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).
12. Csank, C. et al. Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol* **350**, 347-73 (2002).
13. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-11 (1999).
14. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45 (1998).
15. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-9 (1995).
16. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA

binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **20**, 835-9 (2002).
17. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
18. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
20. Dwight, S. S. et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* **30**, 69-72 (2002).
21. Clarke, N. D. & Granek, J. A. Rank order metrics for quantifying the association of sequence features with gene regulation. *Bioinformatics* **19**, 212-8 (2003).
22. Hastie, T., Tibshirani, R. & Friedman, J. *The elements of Statistical Learning; Data mining, inference and prediction* (Springer-Verlag, New York, 2001).
23. Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol Cell Biol* **17**, 6982-93 (1997).

Part II: Supplementary Tables

Supplementary Table 1. List of regulators and environmental conditions examined*

A1	Dat1	Hap3	Met18	Pho4[11]	Sig1[1]	Swi4	YDR266C
Abf1	Dig1[5,6]	Hap4[2,3]	Met28[3]	Pip2	Sip3	Swi5	YDR520C
Abt1	Dot6	Hap5[3]	Met31[3]	Ppr1	Sip4[3]	Swi6	YER051W
Aca1	Ecm22	Hir1	Met32[3]	Put3[2,3]	Skn7[1,2,7]	Tbs1	YER130C
Ace2	Eds1	Hir2	Met4[3]	Rap1[3]	Sko1	Tec1[5,6]	YER184C
Adr1[3,7]	Fap7	Hir3	Mga1[1]	Rco1	Smk1	Thi2[12]	YFL044C
Aft2[1,2]	Fhl1[1,3,4]	Hms1	Mig1[8]	Rcs1[1,2,3]	Smp1	Tos8	YFL052W
Arg80[3]	Fkh1	Hms2	Mig2[1]	Rdr1	Snf1	Tye7	YGR067C
Arg81[3]	Fkh2[1,2]	Hog1	Mig3	Rds1[1]	Snt2	Uga3[3,4]	Yhp1
Aro80[3]	Fzf1	Hsf1[1,2,7]	Mot3[1,2,3]	Reb1[1,2]	Sok2[5]	Ume6[1]	YJL206C[1,2]
Arr1[1]	Gal3	Ifh1	Msn1	Rfx1	Spt10	Upc2	YKL222C
Ash1[5]	Gal4[8,9]	Ime1[1]	Msn2[1,2,4,7,10]	Rgm1	Spt2	Usv1	YKR064W
Ask10	Gal80	Ime4[1]	Msn4[1,2,4,10]	Rgt1[8]	Spt23	War1	YLR278C
Azf1	Gat1[3,4,7]	Ino2	Mss11[5]	Rim101[1,2]	Srd1	Wtm1	YML081W
Bas1[3]	Gat3	Ino4	Mth1[8]	Rlm1[5]	Stb1	Wtm2	YNR063W
Bye1	Gcn4[3,4]	Ixr1	Ndd1	Rlr1	Stb2	Xbp1[2,7]	Yox1
Cad1[1,3]	Gcr1	Kre33	Ndt80	Rme1	Stb4	Yap1[1,2,7]	YPR022C
Cbf1[3]	Gcr2[3]	Kss1[5,6]	Nnf2	Rox1[1,2]	Stb5	Yap3[1]	YPR196W
Cha4[3]	Gln3[3,4]	Leu3[3]	Nrg1[1,2]	Rph1[1,2,3]	Stb6	Yap5[1]	Yrr1
Cin5[1,2]	Gts1	Mac1[1]	Oaf1	Rpi1	Ste12[5,6]	Yap6[1,2]	Zap1
Crz1	Gzf3[1,4]	Mal13	Opi1	Rpn4[1,2]	Stp1[3]	Yap7[1,2]	Zms1
Cst6	Haa1	Mal33[1,2]	Pdc2	Rtg1[3,4]	Stp2	YBL054W	
Cup9	Hac1	Mbf1	Pdr1[2]	Rtg3[1,2,3,4]	Stp4	YBR239C	
Dal80[4]	Hal9	Mbp1[1,2]	Pdr3	Rts2	Sum1	YBR267W	
Dal81[3,4]	Hap1	Mcm1[5,6]	Phd1[5]	Sfl1	Sut1	YDR026C	
Dal82[3,4]	Hap2[4]	Mds3	Pho2[1,2,3,11]	Sfp1[1,2,3]	Sut2	YDR049W	

[1] Highly hyperoxic
[2] Mildly hyperoxic
[3] Amino acid starved
[4] Nutrient deprived
[5] Filamentation
[6] Mating
[7] Heat
[8] Galactose
[9] Raffinose
[10] Acidic
[11] Phosphate deprived
[12] Vitamin deprived

*All regulators were profiled in rich medium. A subset of these were profiled in at least one other environmental condition, as indicated. A complete description of the conditions can be found at the authors' website.

Supplementary Table 2. Similarity of discovered specificities to literature

Regulator	Distance[1]	Discovered	Literature
Abf1	0.143	rTCAytnnnnAcg	rTCAyTnnnnACGw
Ace2	0.18	tGCTGGT	GCTGGT
Aft2	0.15	rCACCC	ATCTTCAAAAGTGCACCCATTTGCAGGTGC
Azf1	0.203	YwTTkcKkTyyckgykky	TTTTTCTT
Bas1	0.045	TGACTC	TGACTC
Cad1	0.089	mTTAsTmAkC	TTACTAA
Cbf1	0.105	tCACGTG	rTCACrTGA
Cin5	0.324	TTAcrTAA	TTACTAA
Fkh1	0.123	gtAAAcAA	GGTAAACAA
Fkh2	0.212	GTAAACA	GGTAAACAA
Gal4	0.11	CGGnnnnnnnnnnncCg	CGGnnnnnnnnnnCCG
Gat1	0.004	aGATAAG	GATAA
Gcn4	0.123	TGAsTCa	ArTGACTCw
Gln3	0.148	GATAAGa	GATAAGATAAG
Hap1	0.191	GGnnaTAnCGs	CGGnnnTAnCGG
Hap4	0.146	gnCcAAtcA	YCNNCCAATNANM
Hsf1	0.198	TTCynnnnnnTTC	TTCTAGAAnnTTCT
Ino2	0.236	CAcaTGc	ATTTCACATC
Ino4	0.163	CATGTGaa	CATGTGAAAT
Leu3	0.131	cCGgtacCGG	yGCCGGTACCGGyk
Mbp1	0.073	ACGCGt	ACGCGT
Mcm1	0.181	CCnrAtnngg	wTTCCyAAwnnGGTAA
Msn2	0.308	mAGGGGsgg	mAGGGG
Nrg1	0.042	GGaCCCT	CCCT
Pdr1	0.301	ccGCCgRAwr	CCGCGG
Pho4	0.096	CACGTGs	cacgtkng
Rap1	0.181	cayCCrtrCa	wrmACCCATACAyy
Rcs1	0.184	ggGTGcant	AAmTGGGTGCAkT
Reb1	0.055	TTACCCG	TTACCCGG
Rpn4	0.049	GGTGGCAAA	GGTGGCAAA
Sip4	0.184	CGGnynAATGGrr	yCGGAyrrAwGG
Skn7	0.228	GnCnnGsCs	ATTTGGCyGGsCC
Stb5	0.058	CGGnstTAta	CGG
Ste12	0.087	tgAAAC	ATGAAAC
Sum1	0.221	gyGwCAswaaw	AGyGwCACAAAAk
Sut1	0.295	gcsGsgnnsG	CGCG
Swi4	0.122	CgCsAAA	CnCGAAA
Swi6	0.214	CGCgaaa	CnCGAAA
Tec1	0.064	CATTCyy	CATTCy
Tye7	0.193	tCACGTGa	CAnnTG
Ume6	0.16	taGCCGCCsa	wGCCGCCGw
Yap1	0.124	TTaGTmAGc	TTAsTmA
Yap7	0.15	mTkAsTmA	TTACTAA
Zap1	0.085	ACCCTmAAGGTyrT	ACCCTAAAGGT

[1]Distance from known specificity was computed using the distance metric described in Supplementary Methods.

Supplementary Table 3. Regulator specificities

Regulator	Discovered specificity[1]	Known specificity[1,2]	Programs[3]
Abf1	rTCAytnnnnAcg	rTCAyTnnnnACGw	A, C, D, K, M, N
Ace2	tGCTGGT	GCTGGT	K
Adr1		GGrGk	
Aft2	rCACCC	...AAAGTGCACCCATT...	A, C, D, M, N
Arr1		TTACTAA	
Ash1		yTGACT	
Azf1	YwTTkcKkTyyckgykky	TTTTTCTT	N
Bas1	TGACTC	TGACTC	A, K, M, N
Cad1	mTTAsTmAkC	TTACTAA	A, C, D, M, N
Cbf1	tCACGTG	rTCACrTGA	A, C, D, K, M, N
Cin5	TTAcrTAA	TTACTAA	A, C, D
Dal80		GATAA	
Dal81		AAAAGCCGCGGGCGGGATT	
Dal82	GATAAG		D, K
Dig1	TgAAAca		A, C, D, K, M, N
Fhl1	TGTayGGrtg		A, C, D, K, M, N
Fkh1	gtAAAcAA	GGTAAACAA	A, C, D, K, M, N
Fkh2	GTAAACA	GGTAAACAA	A, C, D, K, M, N
Gal4	CGGnnnnnnnnnncCg	CGGnnnnnnnnnnCCG	A, K
Gal80		CGGnnnnnnnnnnCCG	
Gat1	aGATAAG	GATAA	K
Gcn4	TGAsTCa	ArTGACTCw	A, C, D, K, M, N
Gcr1		GGCTTCCwC	
Gln3	GATAAGa	GATAAGATAAG	C, D, K
Gzf3		GATAAG	
Hac1		kGmCAGCGTGTC	
Hap1	GGnnaTAnCGs	CGGnnnTAnCGG	C, M
Hap2		CCAAT	
Hap3		CCAAT	
Hap4	gnCcAAtcA	YCNNCCAATNANM	A, C, D, M, N
Hap5		CCAAT	
Hsf1	TTCynnnnnnTTC	TTCTAGAAnnTTCT	A, C, D, K, M, N
Ime1	AAkGAAAnkwA		A
Ino2	CAcaTGc	ATTTCACATC	C, D, M, N
Ino4	CATGTGaa	CATGTGAAAT	A, C, D, K, M, N
Leu3	cCGgtacCGG	yGCCGGTACCGGyk	A, D, K, M
Mac1		GAGCAAA	
Mbp1	ACGCGt	ACGCGT	A, C, D, K, M, N
Mcm1	CCnrAtnngg	wTTCCyAAwnnGGTAA	A, C, D, M, N
Met31		AAACTGTGG	
Met32		AAACTGTGG	
Met4	RMmAwsTGKSgyGsc		C
Mot3		yAGGyA	
Msn2	mAGGGGsgg	mAGGGG	M

Regulator	Discovered specificity[1]	Known specificity[1,2]	Programs[3]
Msn4		mAGGGG	
Ndd1	CCnrAwnnGG		A, D
Nrg1	GGaCCCT	CCCT	A, C, D, M, N
Opi1		TCGAAyC	
Pdr1	ccGCCgRAwr	CCGCGG	M
Pdr3		TCCGCGGA	
Phd1	scnGCngg		A, D, N
Pho2	SGTGCGsygyG		N
Pho4	CACGTGs	cacgtkng	D, K, N
Put3		CGGnnnnnnnnnnCCG	
Rap1	cayCCrtrCa	wrmACCCATACAyy	A, C, D, M, N
Rcs1	ggGTGcant	AAmTGGGTGCAkT	C, D, M, N
Rds1	kCGGCCGa		D, N
Reb1	TTACCCG	TTACCCGG	A, C, D, K, M, N
Rfx1	TTgccATggCAAC		D
Rgt1		CGGAnnA	
Rim101		TGCCAAG	
Rlm1		CTAwwwwTAG	
Rlr1	ATTTTCnnCwTt		N
Rox1		ysyATTGTT	
Rph1		CCCCTTAAGG	
Rpn4	GGTGGCAAA	GGTGGCAAA	A, C, D, K, M, N
Rtg3		GGTCAC	
Sfp1	ayCcrtACay		A, C, D, M, N
Sig1	ArGmAwCrAmAA		M
Sip4	CGGnynAATGGrr	yCGGAyrrAwGG	D
Skn7	GnCnnGsCs	ATTTGGCyGGsCC	A, C, D, M, N
Sko1		ACGTCA	
Smp1		ACTACTAwwwwTAG	
Snt2	yGGCGCTAyca		A, C, D, M, N
Sok2	tGCAgnna		A
Spt2	ymtGTmTytAw		M
Spt23	rAAATsaA		C
Stb1	rracGCsAa		C, D, K, M, N
Stb4	TCGgnnCGA		K
Stb5	CGGnstTAta	CGG	D, N
Ste12	tgAAAC	ATGAAAC	A, C, D, K, M, N
Stp1		rCGGCnnnrCGGC	
Sum1	gyGwCAswaaw	AGyGwCACAAAAk	A, C, D, M, N
Sut1	gcsGsgnnsG	CGCG	A, D, M
Swi4	CgCsAAA	CnCGAAA	A, C, D, K, M, N
Swi5		kGCTGr	
Swi6	CGCgaaa	CnCGAAA	A, C, D, M, N
Tec1	CATTCyy	CATTCy	C
Thi2	gmAAcyntwAgA		C, D

Regulator	Discovered specificity[1]	Known specificity[1,2]	Programs[3]
Tye7	tCACGTGa	CAnnTG	A, C, D, M
Uga3		CCGnnnnCGG	
Ume6	taGCCGCCsa	wGCCGCCGw	A, C, D, K, M, N
Xbp1		CTTCGAG	
Yap1	TTaGTmAGc	TTAsTmA	A, C, D, M
Yap3		TTACTAA	
Yap5		TTACTAA	
Yap6		TTACTAA	
Yap7	mTkAsTmA	TTACTAA	A, C, D, M, N
YDR026C	tTACCCGGm		C, D, M, N
Yhp1		TAATTG	
Yox1		YAATA	
Zap1	ACCCTmAAGGTyrT	ACCCTAAAGGT	N

[1]Text representation of the probability matrices. Lowercase letters indicate a weaker preference (less information content at that position of the probability matrix). Ambiguity Codes: S = C or G, W = A or T, R = A or G, Y = C or T, K = G or T, M = A or C, n = A, C, G or T.
[2] Known specificities are taken from the YPD, SCPD, and TRANSFAC databases.
[3] Program Codes: A = AlignACE, C = CONVERGE, D = MDscan, K = Kellis et al., M = MEME, N = MEME_c.

Supplementary Table 4. Overrepresented MIPS categories among single-regulator architecture binding targets

Regulator	*P* value[1]	Enriched MIPS category[2]
Bas1	6.10e-09	nucleotide metabolism*
Fhl1	1.73e-15	ribosome biogenesis
Gal4	2.18e-04	C-compound and carbohydrate metabolism*
Gat1	4.92e-05	nitrogen and sulfur metabolism*
Gat1	2.63e-02	mRNA transcription*
Gat1	4.38e-02	amino acid metabolism
Gcn4	8.72e-12	amino acid metabolism*
Gzf3	2.21e-02	transport mechanism
Hap3	6.03e-03	lipid, fatty-acid and isoprenoid metabolism
Hap3	1.61e-02	allantoin and allantoate transporters
Hap3	2.50e-02	other energy generation activities
Hap4	3.33e-10	respiration
Hap4	1.78e-05	mitochondrial transport
Hap4	1.03e-02	transport mechanism
Hap4	2.12e-02	assembly of protein complexes
Hsf1	6.58e-06	stress response*
Ino4	5.31e-03	lipid, fatty-acid and isoprenoid metabolism*
Mbp1	1.04e-04	DNA processing
Met32	1.13e-04	amino acid metabolism*
Met32	1.21e-03	nitrogen and sulfur metabolism*
Met32	4.64e-02	amino-acid transporters
Mot3	3.89e-02	DNA processing
Msn2	4.40e-02	metabolism of energy reserves (glycogen, trehalose)
Put3	3.45e-02	other transport facilitators
Reb1	2.09e-05	vesicular transport (Golgi network, etc.)
Rfx1	3.57e-02	other protein-synthesis activities
Rox1	3.43e-02	cell death
Rpn4	2.49e-13	proteolytic degradation*
Rtg3	8.50e-03	other transcription activities
Sig1	2.97e-02	cell cycle
Sip4	2.69e-03	glyoxylate cycle
Sip4	1.57e-02	glycolysis and gluconeogenesis
Stb4	4.02e-02	allantoin and allantoate transporters
Stb5	2.42e-02	electron transport and membrane-associated energy conservation
Ste12	5.56e-03	cell differentiation*
Sut1	5.37e-03	glyoxylate cycle
Swi6	7.96e-03	nitrogen and sulfur metabolism
Thi2	1.15e-02	mRNA transcription*
Thi2	2.45e-02	metabolism of vitamins, cofactors, and prosthetic groups

[1] *P* values represent the probability, based on the hypergeometric distribution, of finding the observed number of genes (or more) with the specified MIPS Level 2 category under the null hypothesis that the genes were selected at random. The values have been corrected for testing multiple categories using Bonferroni correction.

[2] An asterisk (*) indicates that the category is also associated with the regulator itself.

Supplementary Table 5. Regulators with a preference for repetitive motifs

Regulator	*P* value[1]	Non-repetitive	Repetitive
Dig1	1.43e-08	O: 25 E: 45	O: 38 E: 17
Mbp1	2.99e-08	O: 34 E: 56	O: 44 E: 21
Swi6	7.36e-06	O: 34 E: 50	O: 37 E: 20
Sok2	1.34e-05	O: 13 E: 24	O: 21 E: 9
Bas1	2.84e-04	O: 6 E: 12	O: 12 E: 5
Ste12	5.57e-04	O: 48 E: 62	O: 39 E: 24
Swi4	7.29e-04	O: 27 E: 38	O: 26 E: 14
Phd1	7.89e-03	O: 15 E: 21	O: 15 E: 8
Aft2	9.73e-03	O: 22 E: 29	O: 19 E: 11
Swi5	1.05e-02	O: 11 E: 16	O: 12 E: 6
Sfp1	3.03e-02	O: 7 E: 10	O: 8 E: 4
Ino2	4.77e-02	O: 11 E: 15	O: 10 E: 5

[1]*P* values represent the one-tailed probability, based on the chi-square distribution, of finding the observed number of non-repetitive and repetitive motif architecture promoters under the null hypothesis that the distribution for each regulator is the same as the average distribution for all regulators. O = observed number of occurrences; E = expected number of occurrences.

Supplementary Table 6. Co-occurring regulator pairs[1]

Ace2, Fkh2	Dig1, Swi4	Mbp1, Stb1	Rlm1, Sko1
Ace2, Swi5	Dig1, Swi6	Mbp1, Swi4	Rox1, Sut1
Aft2, Rcs1	Dig1, Tec1	Mbp1, Swi6	Sip4, Stp1
Arr1, Yap3	Fhl1, Rap1	Mcm1, Ndd1	Skn7, Sok2
Azf1, Gzf3	Fhl1, Sfp1	Mcm1, Ste12	Skn7, Sut1
Bas1, Met4	Fkh1, Fkh2	Mcm1, Swi4	Skn7, Swi6
Cad1, Yap1	Fkh2, Mcm1	Mcm1, Swi6	Skn7, Xbp1
Cad1, Yap7	Fkh2, Ndd1	Mcm1, Tec1	Sko1, Sok2
Cbf1, Met31	Fkh2, Swi6	Met31, Met32	Sok2, Sut1
Cbf1, Met32	Gat1, Spt23	Met31, Met4	Sok2, Swi6
Cbf1, Met4	Gcn4, Gln3	Met32, Met4	Spt23, Yox1
Cbf1, Pho4	Gcn4, Leu3	Mot3, Rox1	Stb1, Swi4
Cbf1, Tye7	Gcr1, Tye7	Mot3, Skn7	Stb1, Swi6
Cin5, Phd1	Gln3, Hap2	Msn2, Msn4	Stb1, Tec1
Cin5, Skn7	Gzf3, Pdr1	Msn4, Nrg1	Ste12, Swi4
Cin5, Sok2	Hap2, Hap3	Nrg1, Rlm1	Ste12, Swi6
Cin5, Sut1	Hap2, Hap4	Nrg1, Skn7	Ste12, Tec1
Cin5, Xbp1	Hap2, Hap5	Phd1, Rox1	Swi4, Swi6
Cin5, Yap6	Hap3, Hap5	Phd1, Skn7	Swi4, Tec1
Dal82, Gat1	Hap4, Hap5	Phd1, Sok2	Swi6, Tec1
Dal82, Gln3	Hsf1, Msn4	Phd1, Sut1	Yap1, Yap7
Dal82, Hap2	Ino2, Ino4	Phd1, Swi6	Yap6, Yap7
Dig1, Mcm1	Ino4, Sko1	Rap1, Sfp1	
Dig1, Ste12	Mac1, Rcs1	Rim101, Yox1	

[1]Shown are co-occurring regulator pairs ($P \leq 0.005$). P values represent the probability, based on the hypergeometric distribution, of finding the observed number of intergenic regions (or more) bound by both regulators under the null hypothesis that binding for the two regulators is independent.

Supplementary Table 7. Behaviour classifications of regulators[1]

Condition invariant[2]	Condition enabled[3]	Condition expanded[4]	Condition altered[5]
Fhl1	Adr1	Bas1	Adr1
Gal4	Arr1	Cad1	Aft2
Gcn4	Ash1	Cbf1	Cad1
Hsf1	Dal81	Cin5	Cin5
Leu3	Fhl1	Dal82	Dal80
Put3	Gat1	Fkh2	Dal82
Ste12	Hap4	Gal4	Dig1
Ume1	Hsf1	Gcn4	Fkh2
Yap7	Mot3	Gln3	Gat1
	Msn2	Hap2	Gln3
	Pdr1	Mac1	Gzf3
	Phd1	Mbp1	Hap4
	Pho2	Mcm1	Hap5
	Put3	Met31	Mbp1
	Rap1	Met32	Mot3
	Rgt1	Met4	Msn2
	Rim101	Nrg1	Msn4
	Rlm1	Rcs1	Phd1
	Rph1	Rds1	Pho4
	Rpn4	Reb1	Reb1
	Rtg3	Rox1	Rox1
	Sfp1	Rpn4	Rtg3
	Sig1	Rtg3	Skn7
	Sip4	Skn7	Ste12
	Sok2	Ste12	Tec1
	Stp1		Ume6
	Thi2		Yap1
	Uga3		Yap6
	Xbp1		
	Yap1		
	Yap7		

[1] The binding of each regulator was compared in pairwise fashion for every environmental condition in which that regulator was profiled. Some regulators fall into multiple categories depending on exactly which conditions are compared.

[2] The ratio of the overlap of bound probes for a regulator ($P \leq 0.001$) was greater than 0.66 and the ratio of the number of bound probes was between 0.66 and 1.5.

[3] Regulator bound to no probes in one environment.

[4] The ratio of the overlap of bound probes for a regulator was greater than 0.66 and the ratio of the number of bound probes was less than 0.66 or greater than 1.5.

[5] Regulator bound at least one probe in both environments and the ratio of the overlap of bound probes was less than 0.66.

Supplementary Table 8. Motif score significance cutoffs ($P \leq 0.001$)

Enrichment Score[1]

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME_c
10	12.70	20.32	11.78	13.54	n/a
20	11.96	21.14	12.95	12.89	9.81
30	11.43	20.43	13.30	12.57	n/a
40	11.34	20.62	14.04	11.64	7.53
50	10.74	19.94	12.23	12.81	7.43
60	10.50	19.71	10.95	12.37	n/a
70	10.34	18.30	13.25	11.34	n/a
80	10.20	19.40	12.84	11.93	n/a
100	9.36	20.31	11.56	10.58	2.91
120	n/a	18.59	13.14	10.94	n/a
140	8.14	18.52	11.26	10.87	n/a
160	n/a	20.04	11.38	9.77	n/a

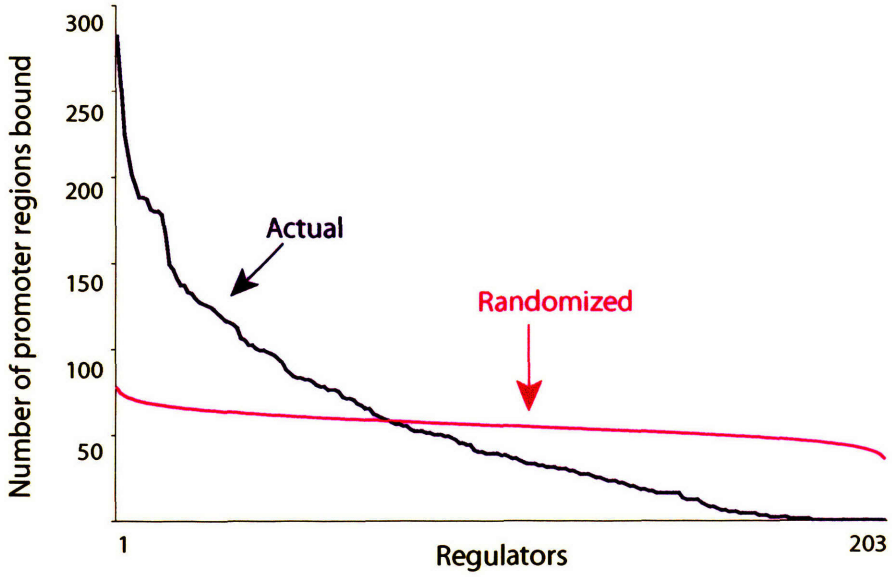
ROC a.u.c.[1]

Number of sequences	Converge	AlignACE	MDscan	MEME	MEME_c
10	n/a	n/a	n/a	n/a	n/a
20	0.812	0.842	0.857	0.925	n/a
30	0.758	0.773	0.793	0.831	0.785
40	0.720	0.713	0.758	0.764	0.737
50	0.687	0.674	0.719	0.737	0.711
60	0.670	0.662	0.688	0.706	0.654
70	0.663	0.641	0.686	0.684	0.664
80	0.643	0.626	0.670	0.675	0.648
100	0.634	0.615	0.664	0.633	0.606
120	0.624	0.604	0.629	0.624	0.602
140	0.608	n/a	0.634	n/a	0.590
160	0.594	0.580	0.613	0.593	0.588

[1]Motif score significance $P \leq 0.001$ thresholds for "Enrichment" and "ROC a.u.c." specificity metrics obtained from calculations on randomized selections of intergenic regions as described in Methods. Entries containing "n/a" denote that the empirical distribution was not normal. The threshold for the CC4 metric (4.95) is not dependent on the number of sequences.

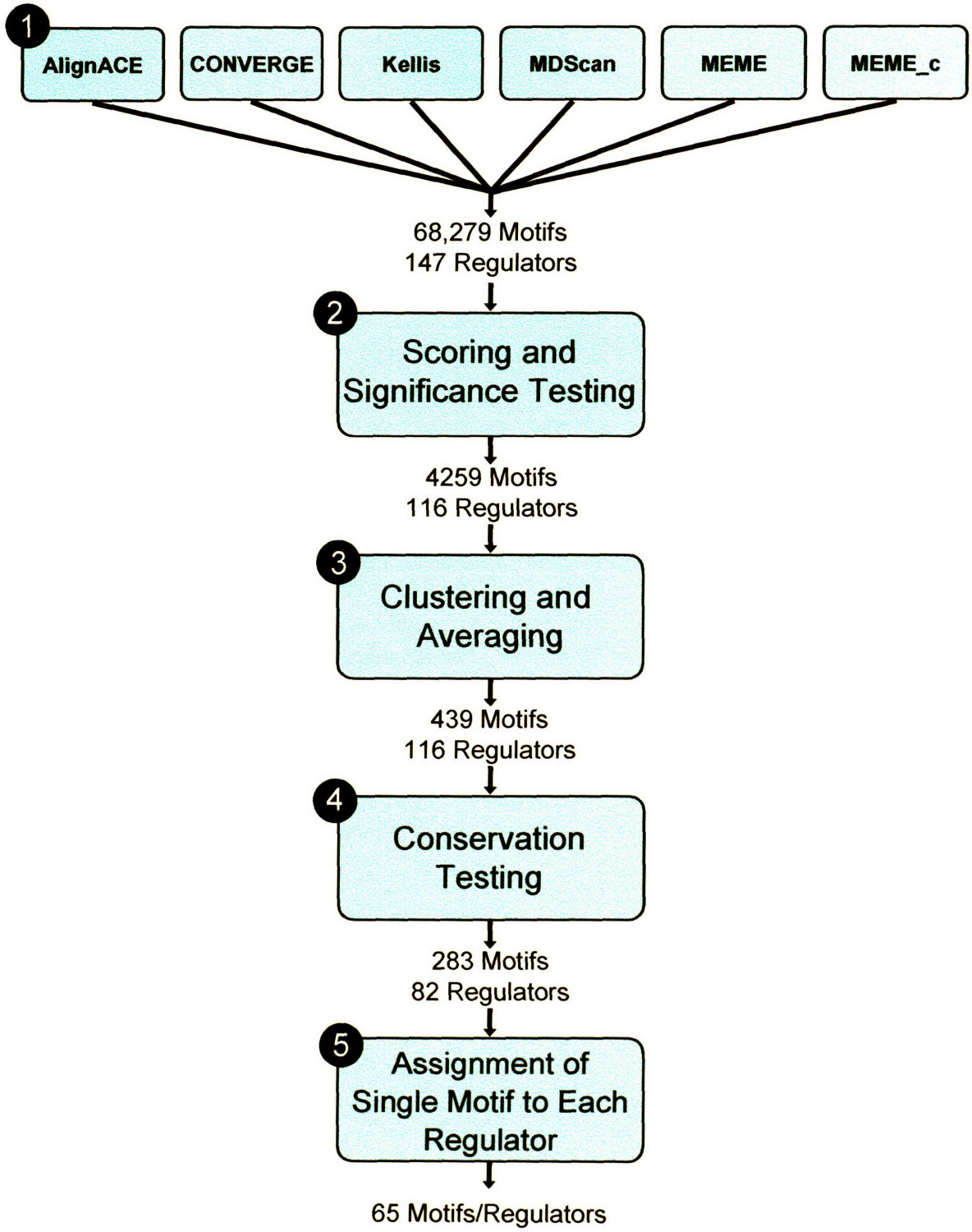
Part III: Supplementary Figures

Supplementary Figure 1



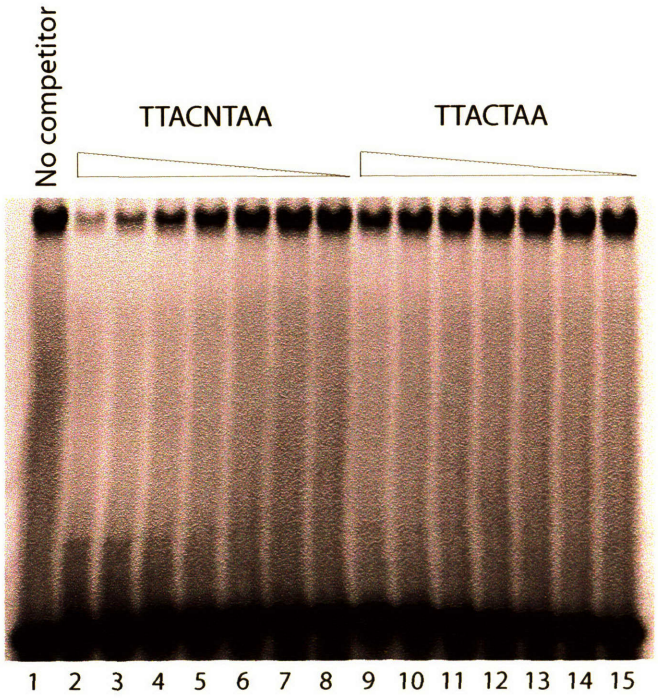
Distribution of the number of promoter regions bound per regulator (blue). For regulators profiled under multiple conditions, the union of promoter regions bound under all conditions is reported. An average of randomized distributions for the same set of *P* values randomly assigned among regulators and promoter regions is shown in pink.

Supplementary Figure 2



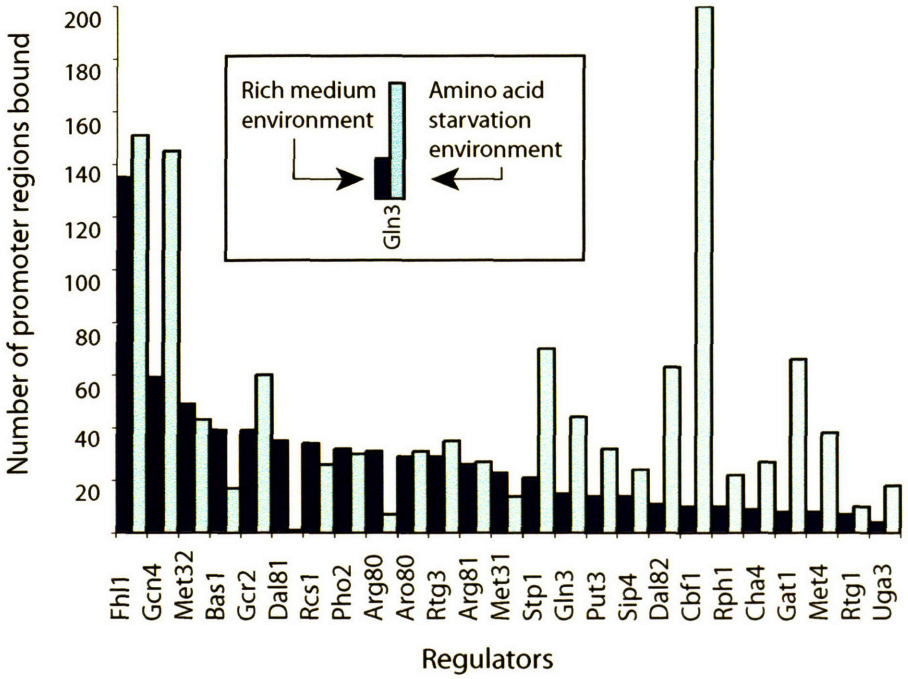
Overview of motif discovery and assignment. Motifs were identified by applying a suite of motif discovery programs to the intergenic sequences identified by the binding data. The resulting specificity predictions were filtered for significance and then clustered to yield representative motifs. Conservation-based metrics were used to identify the highest-confidence subset of these motifs. For cases in which multiple significant binding motifs were found for a factor, we used statistical scores or information from specificity databases to choose a single motif for each regulator. A complete description of the method can be found in Supplementary Methods.

Supplementary Figure 3



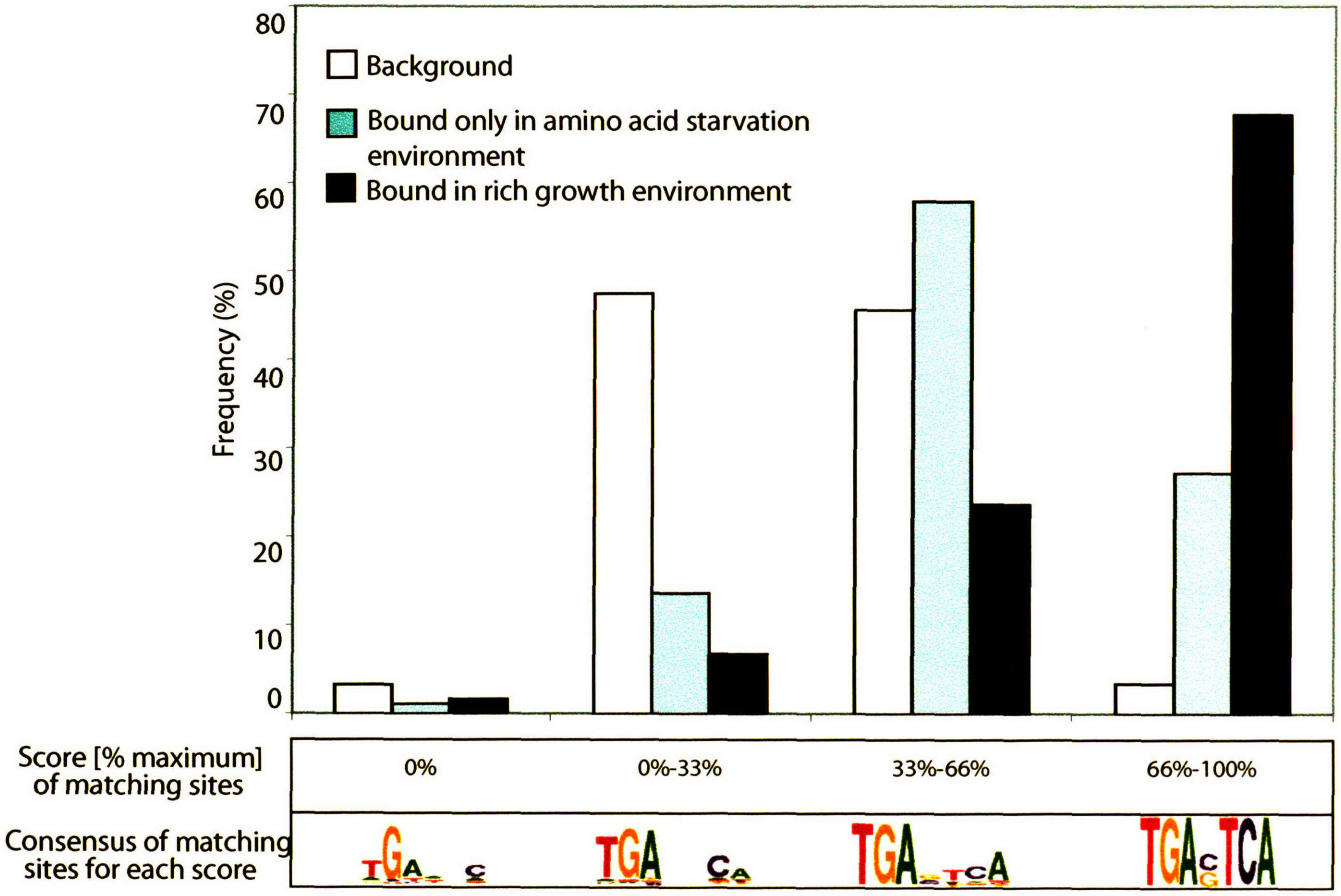
Comparison of Cin5 binding to two sequences. Recombinant Cin5 was purified from bacteria and incubated with a Cy5-labeled oligonucleotide containing the sequence (gcgacaTTACCTAAgggc) and challenged with one of two unlabeled competitors: the same sequence (lanes 2-8) or the previoulsy published binding site (gcgacaTTACTAAagggc; lanes 9-15). The concentration of each competitor was varied in 3-fold steps. The probe based on our discovered motif was approximately 27-fold better in competing away the shifted band compared to the probe based on the previously published specificity. Similar results were obtained for a probe containing a core sequence of TTACGTAA.

Supplementary Figure 4



Pairwise comparison of the number of promoter regions bound under two different conditions for 25 regulators (based solely on genome-wide location data with $P \leq 0.001$). Dark blue bars represent the number of promoter regions bound under growth in rich medium; light blue bars represent the number of promoter regions bound under growth in amino acid starvation medium.

Supplementary Figure 5



Quality of Gcn4 binding sites among intergenic regions bound under different conditions. Each intergenic region was scored based on the quality of the best matching subsequence to the Gcn4 binding specificity (TGASTCA). In rich media conditions 68% of the intergenic regions contain high-quality matches to the Gcn4 specificity. Under starvation conditions the levels of Gcn4 protein rise, and the set of bound intergenic regions expands. Of the newly bound regions, only 27% contain high-quality matches. By contrast, only 3% of all intergenic regions contain matches of this quality.

Appendix B

Remodeling of Yeast Genome Expression in Response to Environmental Changes

Published as: Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. *Mol Biol Cell.* **2**:323-37.

My contributions to this project

Upon joining the lab, I began experiments investigating the transcriptional response to changes in osmolarity. I performed genome-wide expression analysis on yeast grown in medium containing elevated concentrations of either salt or of sorbitol. These data were subsumed into a larger study on environmental responses that was authored by Helen Causton and Bing Ren of our lab.

Summary

We used genome-wide expression analysis to explore how gene expression in *Saccharomyces cerevisiae* is remodeled in response to various changes in extracellular environment, including changes in temperature, oxidation, nutrients, pH and osmolarity. The results demonstrate that over half of the genome is involved in various responses to environmental change and identify the global set of genes induced and repressed by each condition. These data implicate a substantial number of previously uncharacterized genes in these responses, and reveal a signature common to environmental responses that involves approximately 10% of yeast genes. The results of expression analysis with *MSN2/MSN4* mutants support the model that the Msn2/4 activators induce the common response to environmental change. These results provide a global description of the transcriptional response to environmental change and extend our understanding of the role of activators in effecting this response.

Introduction

The ability to respond rapidly to fluctuations in temperature, nutrients and other environmental changes is important for competitive fitness and cell survival. Understanding the response of cells to environmental changes is of interest because it can provide clues to the molecular apparatuses that enable cells to adapt to new environments and the molecular mechanisms that have evolved to regulate the remodeling of gene expression that occurs in new environments.

Significant clues to the mechanisms involved in adaptation to new environments have come from studies of the genes that are expressed in response to specific stresses. For example, cells exposed to elevated temperatures induce transcription of genes encoding heat shock proteins (Craig, 1992). The heat shock proteins are a family of approximately a dozen proteins that are evolutionarily conserved. Studies of heat shock proteins led to the realization that many function as molecular chaperones (Ellis, 1999). Molecular chaperones are critical regulators of protein structure and function, and have roles in almost every cellular process. Some molecular chaperones may even facilitate evolutionary processes (Rutherford and Lindquist, 1998). The importance of molecular chaperones suggests that it will be valuable to identify and further study the complete set of stress-inducible genes. If the number of stress-responsive genes is substantial, their identification could make a significant contribution to functional annotation of an important set of previously uncharacterized genes.

Cells must coordinate adjustments in genome expression to accommodate changes in their environment. Despite our lack of knowledge about the complete set of genes involved in these changes, investigators have identified transcriptional activators and repressors that likely contribute to coordinate remodeling of genome expression. For example, the yeast heat shock transcriptional activator Hsf1 and the canonical sequence it binds have been identified (Kingston et al., 1987; Parker and Topol, 1984; Sorger and Pelham, 1987; Wu, 1985). In the absence of heat shock, Hsf1 is inactive; the molecular chaperone Hsp90 is thought to contribute to this inactivation by binding and sequestering the activator (Ali et al., 1998; Bharadwaj et al., 1999; Duina et al., 1998; Zou et al., 1998). Another set of activators, Msn2 and Msn4, act in concert to induce expression of genes under almost any stress condition. Msn2 and Msn4, normally resident in the cytoplasm, are transported into the nucleus during stress, where they bind to stress response elements (STRE) in promoters (Estruch and Carlson, 1993; Görner et al., 1998; Marchler et al., 1993; Martinez-Pastor et al., 1996; Schmitt and McEntee, 1996). The complete set of genes induced by various environmental changes has not been established, so it is not yet clear that these activators are responsible for coordinate induction of these genes.

Here we describe the temporal expression profiles of yeast cells exposed to seven environmental changes. These transcriptional responses demonstrate that a much larger fraction of the genome is involved in responses to environmental changes than previously appreciated, identify the global set of genes induced and repressed by new conditions, and reveal a signature common to each of the environmental responses. Furthermore,

expression profiles of strains deleted for Msn2/Msn4 reveal the contributions of these activators to coordinate regulation of environmental responses.

Results

We identified environmental conditions that have been frequently selected for study by other investigators. *Saccharomyces cerevisiae* cells in logarithmic phase growth were exposed to various environmental changes and the transcriptional response was monitored using high-density oligonucleotide arrays. These changes involved heat (a shift from 25°C to 37°C), acid (pH 6.0 to 4.0), alkali (pH 6.0 to 7.9), hyperoxia (0.0 mM to 0.4 mM H_2O_2), salt (addition of NaCl to 1.0 M) and osmotic stress (addition of sorbitol to 1.5 M). For each of the conditions cells were grown in YPD and subjected to the new environment when cultures reached OD_{600} 0.5 to 0.8. Labeled 'target' RNA was prepared from cultures harvested immediately before and at various times after the change in environment and hybridized to Affymetrix Genechips, as described previously (Holstege et al., 1998). Additional detailed information and interactive databases supporting this study can be found on the World Wide Web at web.wi.mit.edu/young/environment/. Data on the transcriptional response to nutrient depletion at the diauxic shift were taken from DeRisi *et al.* (DeRisi et al., 1997).

The clustered results shown in Figure 1 reveal several interesting features of the response yeast cells undergo to various environmental changes. A remarkable fraction of the genome is subjected to expression remodeling during these responses. Of the 5594 genes whose expression could be scored in these time courses, expression of 66% (3684) is altered significantly when the data is analyzed as described in Methods. It is clear from much previous work that cells have evolved responses that enhance cell survival and

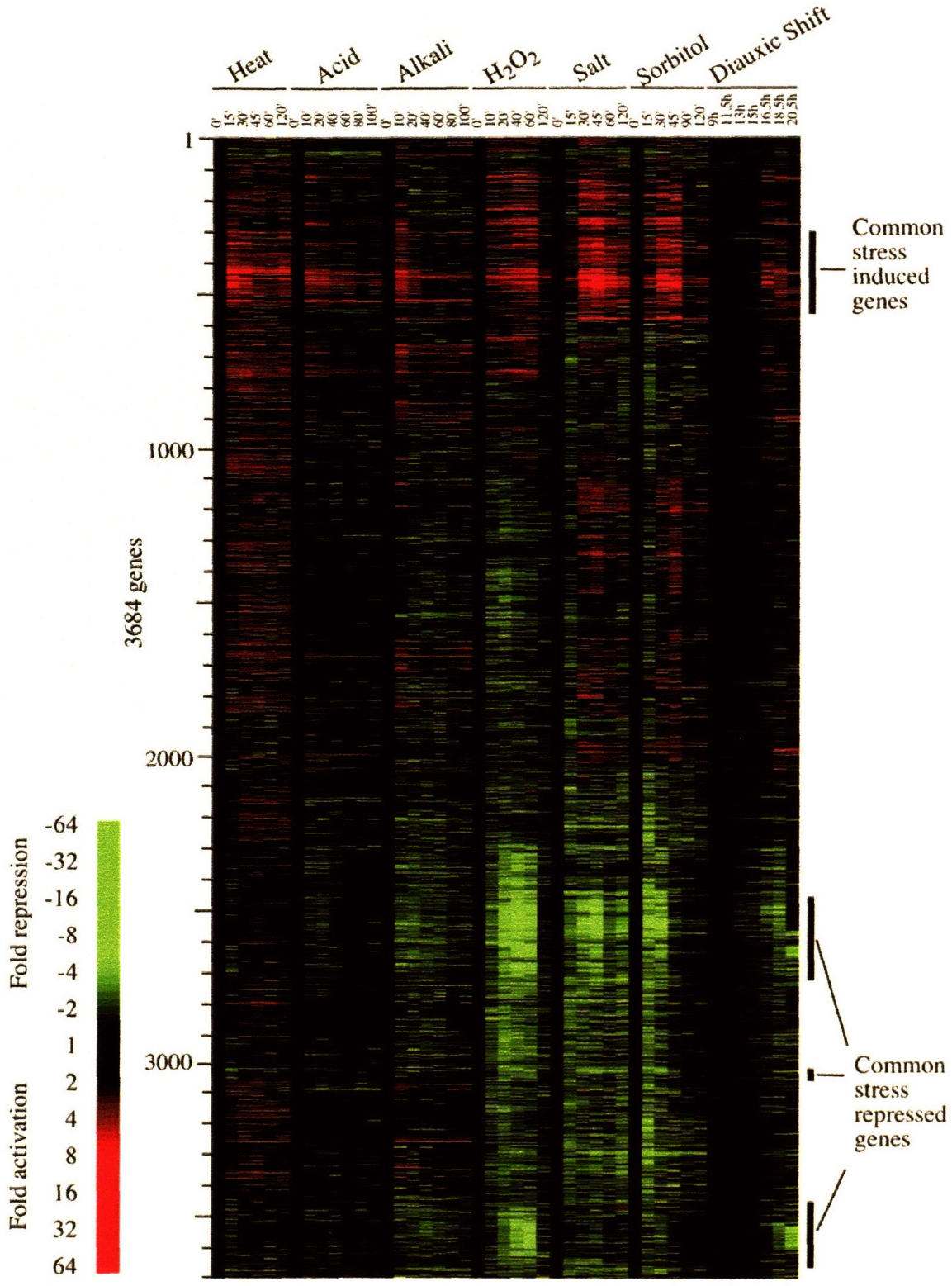

Heat
Acid
Alkali
H_2O_2
Salt
Sorbitol
Diauxic Shift
0' 15' 30' 45' 60' 120'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 120'
0' 15' 30' 45' 60' 120'
0' 15' 30' 45' 90' 120'
9h 11.5h 13h 15h 16.5h 18.5h 20.5h
1
1000
2000
3000
3684 genes
Common stress induced genes
Common stress repressed genes
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64

Figure 1. Time-course expression profiles for cells exposed to changes in environment.

The expression profiles of 3684 genes whose transcript levels changed by three-fold in at least one of the time courses are represented. See Methods for further details of data analysis. Each horizontal strip represents a single gene. The fold change is represented by a color (see color bar). The genes that are induced or repressed in most of the responses to environmental changes are indicated.

fitness in the dynamic environments in which they live, but the extent of the genome involved in these responses in yeast is impressive. The involvement of many genes with unknown cellular roles in these environmental changes implicates these genes in specific responses, and these results should therefore contribute to further functional annotation of the genome. Given the broad scope of expression remodeling that occurs when cells encounter new environments, which is a frequent occurrence outside of the laboratory, the term "stress response" seems inadequate to describe these events. To avoid confusing the large scale effects observed in this study with previously described stress responses, we will tend to refer to the broader effects on gene expression as environmental responses rather than stress responses.

It is also evident from the results shown in Figure 1 that the kinetics of global change, and whether the change is transient or constitutive, varies with the environmental change. For example, cells exposed to a shift in temperature activate expression of heat-responsive genes within 15 minutes of the temperature shift, whereas cells exposed to an increase in salinity take longer to respond. In most responses, there were genes whose expression levels changed transiently and others whose levels remained altered through the entire time course. We infer that the products of the genes that exhibited transient increases are involved in facilitating the transition to the new environment. The genes whose expression levels change to a new level and remain so altered likely encode products that have a continuous role in the cell under the new conditions.

We note that there are a substantial number of genes whose expression patterns are common to most of the environmental changes. We henceforth refer to these as common environmental response (CER) genes.

Common environmental response: genes induced

The response that is common to most of the environmental changes examined here involves 499 genes, of which 216 are induced, and 283 are repressed (the criteria used for analysis are described in Methods). The CER thus involves approximately 10% of the genome; that such a large fraction of the genome is remodeled under a wide range of environmental changes attests to the importance of these genes in cellular adaptation to the environment. The expression data for the induced common environmental response genes are displayed in Figure 2. The genes induced in the common environmental response include those with functions in carbohydrate metabolism, cell stress and the generation of energy.

Many of the genes induced are involved in glycolysis, an increase in which could provide energy needed for the functions of ATP-dependent molecular chaperones and other machinery involved in the response to cellular stress. Genes encoding all the subunits of trehalose synthetase (*TPS1*, *TPS2*, *TPS3*, *TSL1*) were also activated. This might be anticipated from previous studies (Jelinsky and Samson, 1999; Rep et al., 2000), since trehalose is thought to protect cellular components from the detrimental effects of stress by providing energy for the renaturation of cellular structures and possibly by protecting cells and membranes from denaturation.

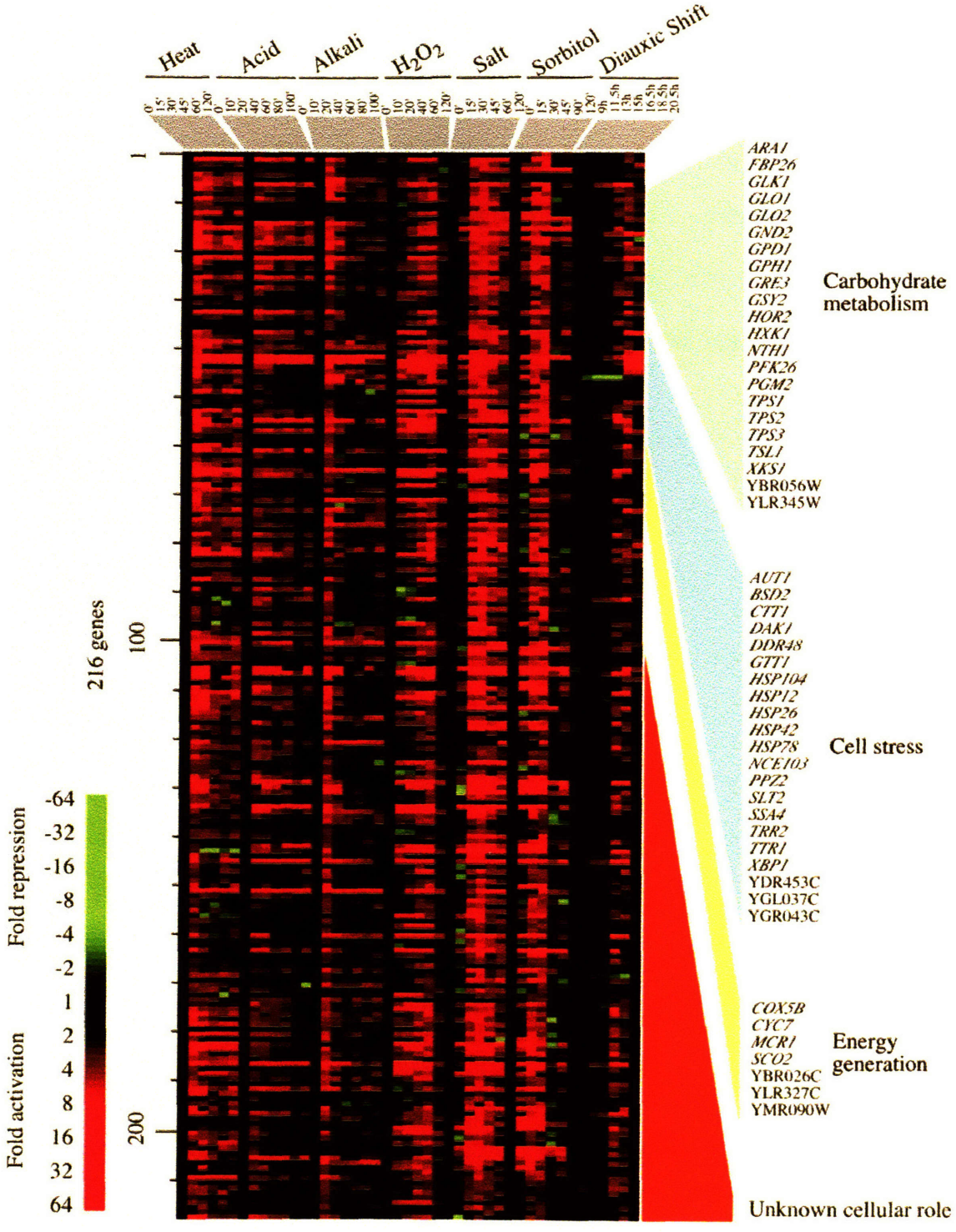

Heat
Acid
Alkali
H_2O_2
Salt
Sorbitol
Diauxic Shift
0' 15' 30' 45' 60' 120'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 120'
0' 15' 30' 45' 60' 120'
0' 15' 30' 45' 90' 120'
9h 11.5h 13h 15h 16.5h 18.5h 20.5h
1
100
200
216 genes
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64
ARA1
FBP26
GLK1
GLO1
GLO2
GND2
GPD1
GPH1
GRE3
GSY2
HOR2
HXK1
NTH1
PFK26
PGM2
TPS1
TPS2
TPS3
TSL1
XKS1
YBR056W
YLR345W
Carbohydrate metabolism
AUT1
BSD2
CTT1
DAK1
DDR48
GTT1
HSP104
HSP12
HSP26
HSP42
HSP78
NCE103
PPZ2
SLT2
SSA4
TRR2
TTR1
XBP1
YDR453C
YGL037C
YGR043C
Cell stress
COX5B
CYC7
MCR1
SCO2
YBR026C
YLR327C
YMR090W
Energy generation
Unknown cellular role

Figure 2. A common environmental response: gene activation.

One of the clusters from the hierarchical tree in Figure 1, containing genes whose expression is induced in most of the environmental responses, is represented. These data were sorted so genes with similar cellular functions are listed together. Details are as described in the legend to Figure 1.

The 'cell stress' genes in the CER include classical heat shock genes (*HSP12*, *HSP26*, *HSP42*, *HSP78*, *HSP104*, *SSA4*, *SSE2*), many of which encode molecular chaperones that facilitate protein folding or maintenance of a particular protein conformation. Genes whose products are involved in protein degradation (*PHB2*, *RPN5*, *UBC5*, *UBC8*, *YPS6*) were also induced, consistent with the notion that damaged or partially denatured proteins need to be degraded in order to prevent the accumulation of protein aggregates.

Other genes induced in the CER include those involved in antioxidant defenses. These function in the degradation of reactive species like hydrogen peroxide that can potentially damage proteins and nucleic acids (*CTT1*). Ion homeostasis genes are also represented in the CER: these are involved in sequestration or metal transport (*BSD2*), or in thioredoxin or glutathione regulation (*TTR1*, YDR435C, YCL035C). These genes are important for maintaining the reducing environment within the cell. Many of the CER genes were thought to be specifically induced in response to changes in the tonicity of the environment, however, our data suggest that many changes in the environment also result in increased membrane permeability and thus induce systems involved in ion transport.

The energy generation genes include those required for respiration, some of which have consensus binding sites for the HAP2,3,4 complex. It has been observed previously that these genes are induced upon nutrient limitation at the diauxic shift (DeRisi et al.,

1997). More than one-half of the induced CER genes do not have known cellular roles, and these can now be annotated as being involved in the CER.

A subset of the CER genes were previously observed to be involved in a variety of stress responses, and have been termed 'general stress response' genes (Moskvina et al., 1998; Treger et al., 1998). These approximately 60 to 190 genes have STRE (Stress Response Element) consensus sequences in their promoter regions; the Msn2 and Msn4 transcription factors bind to these elements and activate transcription of these genes under stress conditions (Boy-Marcotte et al., 1998; Görner et al., 1998; Martinez-Pastor et al., 1996; Schmitt and McEntee, 1996). The relationship between the CER genes and the previously identified general stress genes will be discussed further below.

Common environmental response: genes repressed

The expression data for the repressed set of genes in the CER are displayed in Figure 3. The 283 repressed genes are dominated by genes associated with translation and protein synthesis. Many of these have previously been observed to be repressed in response to specific stresses (DeRisi et al., 1997; Eisen et al., 1998; Jelinsky and Samson, 1999; Kim and Warner, 1983; Lashkari et al., 1997; Rep et al., 2000). The repressed set includes genes for cytoplasmic ribosomal proteins, polymerase I, II and III transcription, tRNA synthetases, proteins required for processing ribosomal RNAs, and a subset of translation initiation factors. The genes repressed in all environmental changes include 106 genes of unknown function.

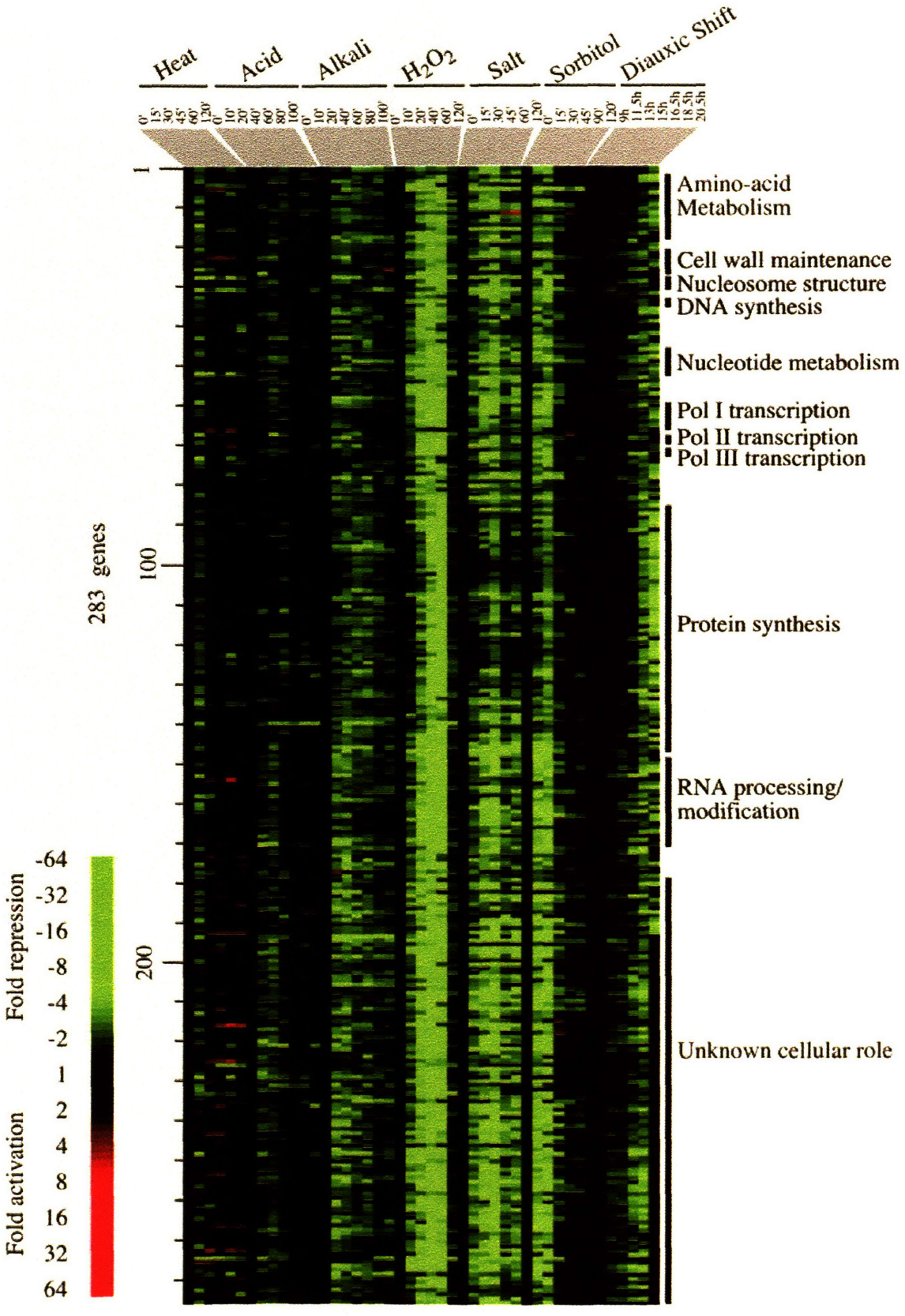

Heat
Acid
Alkali
H_2O_2
Salt
Sorbitol
Diauxic Shift
0' 15' 30' 45' 60' 120'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 120'
0' 15' 30' 45' 60' 120'
0' 15' 30' 45' 90' 120'
9h 11.5h 13h 15h 16.5h 18.5h 20.5h
1
100
200
283 genes
Amino-acid Metabolism
Cell wall maintenance
Nucleosome structure
DNA synthesis
Nucleotide metabolism
Pol I transcription
Pol II transcription
Pol III transcription
Protein synthesis
RNA processing/ modification
Unknown cellular role
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64

Figure 3. A common environmental response: transient repression of the translation apparatus.

Three clusters from the hierarchical tree in Figure 1, containing genes whose expression was repressed in most of the environmental responses, are represented. These clusters were combined and the genes sorted according to their cellular roles. Details are as described in the legend to Figure 1.

The transient but significant reduction in transcripts for the translation apparatus is consistent with the previously noted transient translational arrest that occurs on heat shock, as cellular resources are redirected towards synthesis of stress proteins (Miller et al., 1982; Miller et al., 1979). The loss and then re-establishment of transcript levels for genes encoding the translation apparatus and its regulators is remarkably coordinated for each environmental change, although the dynamics of each response are different, as observed for the induced set of CER genes. For example, genes are significantly repressed for about 60 minutes after treatment with hydrogen peroxide, but are only transiently repressed in the heat and sorbitol time courses (Figure 3).

Differential expression of isozymes

Previous reports have suggested that isozymes and other members of multigene families can be differentially expressed under specific conditions (Rep et al., 2000). To determine whether this is the case for the CER genes under conditions explored in this study, we examined the 405 pairs of homologous genes identified by Wolfe and Shields (Wolfe and Shields, 1997). Of the 316 pairs of genes for which expression data was obtained, 79 pairs of genes contain at least one member of the CER. In 37 of these pairs, one gene of the pair is CER-induced and expression of the other member of the pair is not (Figure 4). Many of these enzymes are involved in glycolysis (*GLK1*, *GLO2*, *GND2*, *HXK1*, *PGM2*) or energy generation (*COX5A*, *CYC7*). The observation that cells differentially express these particular isozymes during changes in environment suggests that one member of these pairs plays a particularly important role in the adaptation to new environments.

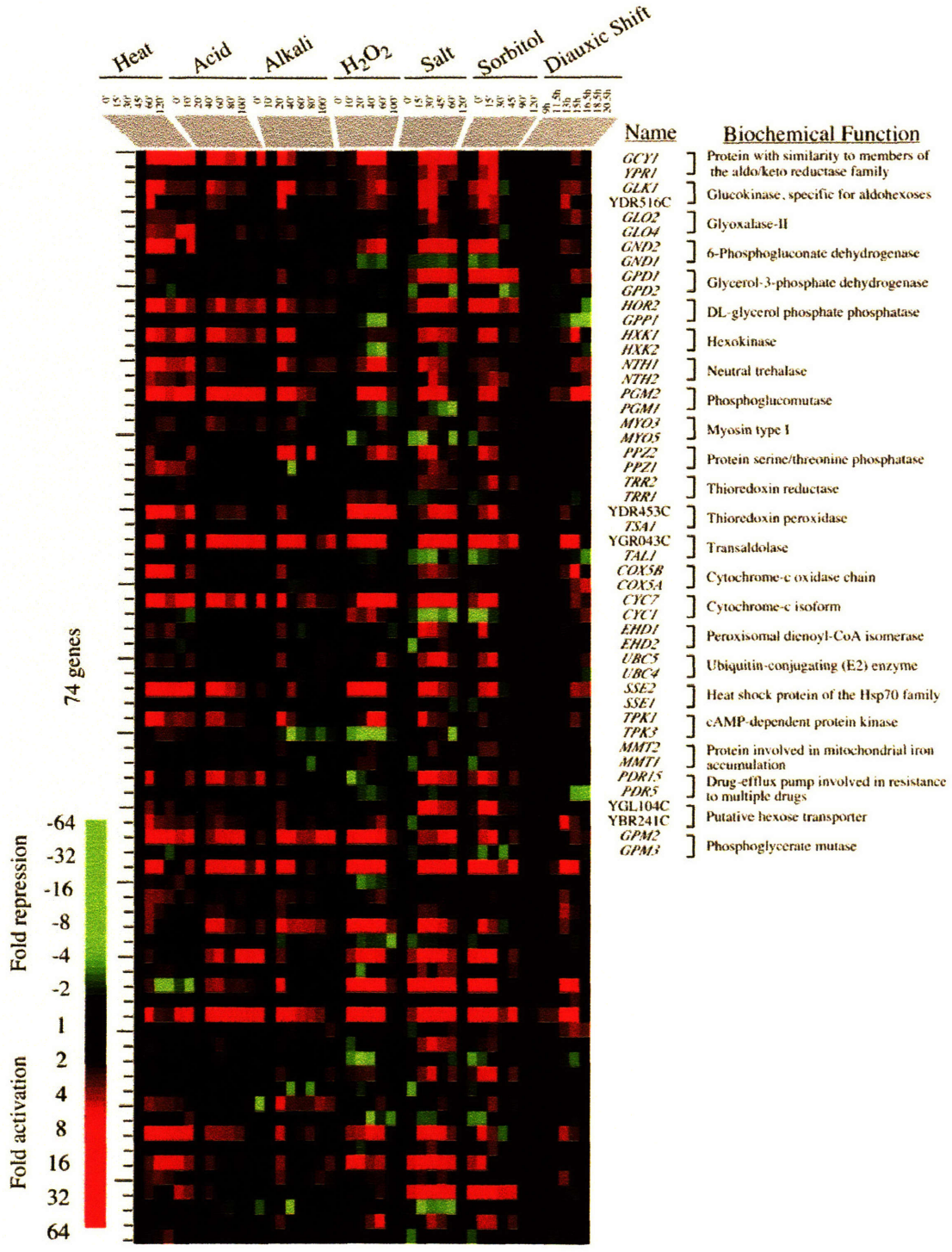

Heat
Acid
Alkali
H_2O_2
Salt
Sorbitol
Diauxic Shift
74 genes
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64
Name
Biochemical Function
GCY1
YPR1
Protein with similarity to members of the aldo/keto reductase family
GLK1
YDR516C
Glucokinase, specific for aldohexoses
GLO2
GLO4
Glyoxalase-II
GND2
GND1
6-Phosphogluconate dehydrogenase
GPD1
GPD2
Glycerol-3-phosphate dehydrogenase
HOR2
GPP1
DL-glycerol phosphate phosphatase
HXK1
HXK2
Hexokinase
NTH1
NTH2
Neutral trehalase
PGM2
PGM1
Phosphoglucomutase
MYO3
MYO5
Myosin type I
PPZ2
PPZ1
Protein serine/threonine phosphatase
TRR2
TRR1
Thioredoxin reductase
YDR453C
TSA1
Thioredoxin peroxidase
YGR043C
TAL1
Transaldolase
COX5B
COX5A
Cytochrome-c oxidase chain
CYC7
CYC1
Cytochrome-c isoform
EHD1
EHD2
Peroxisomal dienoyl-CoA isomerase
UBC5
UBC4
Ubiquitin-conjugating (E2) enzyme
SSE2
SSE1
Heat shock protein of the Hsp70 family
TPK1
TPK3
cAMP-dependent protein kinase
MMT2
MMT1
Protein involved in mitochondrial iron accumulation
PDR15
PDR5
Drug-efflux pump involved in resistance to multiple drugs
YGL104C
YBR241C
Putative hexose transporter
GPM2
GPM3
Phosphoglycerate mutase

Figure 4. Homologous gene pairs are differentially expressed in response to changes in the environment.

There are 37 pairs of genes (74 genes) for which one of the pair is part of the CER and the other is not. Details are described in the text.

Environment-specific responses

Although the CER genes exhibit expression changes in all environmental changes, there are a substantial number of genes whose expression is altered in response to a specific change in environment. The CER accounts for 18 to 38% of the total population of genes whose expression is altered in response to an environmental change. The responses to specific changes in environment, such as salt concentration, involve the remodeling of up to one-third of the genome.

Heat

Genes whose expression is induced and remains elevated after a shift to higher temperature are shown in Figure 5A. Many of these induced genes have functions in protein folding and transport, including *EUG1* and *LHS1*, which are regulated by the unfolded protein response (Chapman et al., 1998; Craven et al., 1996; Travers et al., 2000). The induction of protein folding genes is consistent with the need to contend with the widespread protein denaturation that occurs on heat stress. It has been proposed that retention of proteins in the endoplasmic reticulum and the refolding of heat-denatured glycoproteins is also part of the cellular stress response. Lhs1 is one of the proteins required for the return to secretion competence of heat denatured proteins in the endoplasmic reticulum (Saris and Makarow, 1998). Although the cellular response to heat stress has been extensively studied, 50% of the 854 heat responsive genes do not have defined functions.

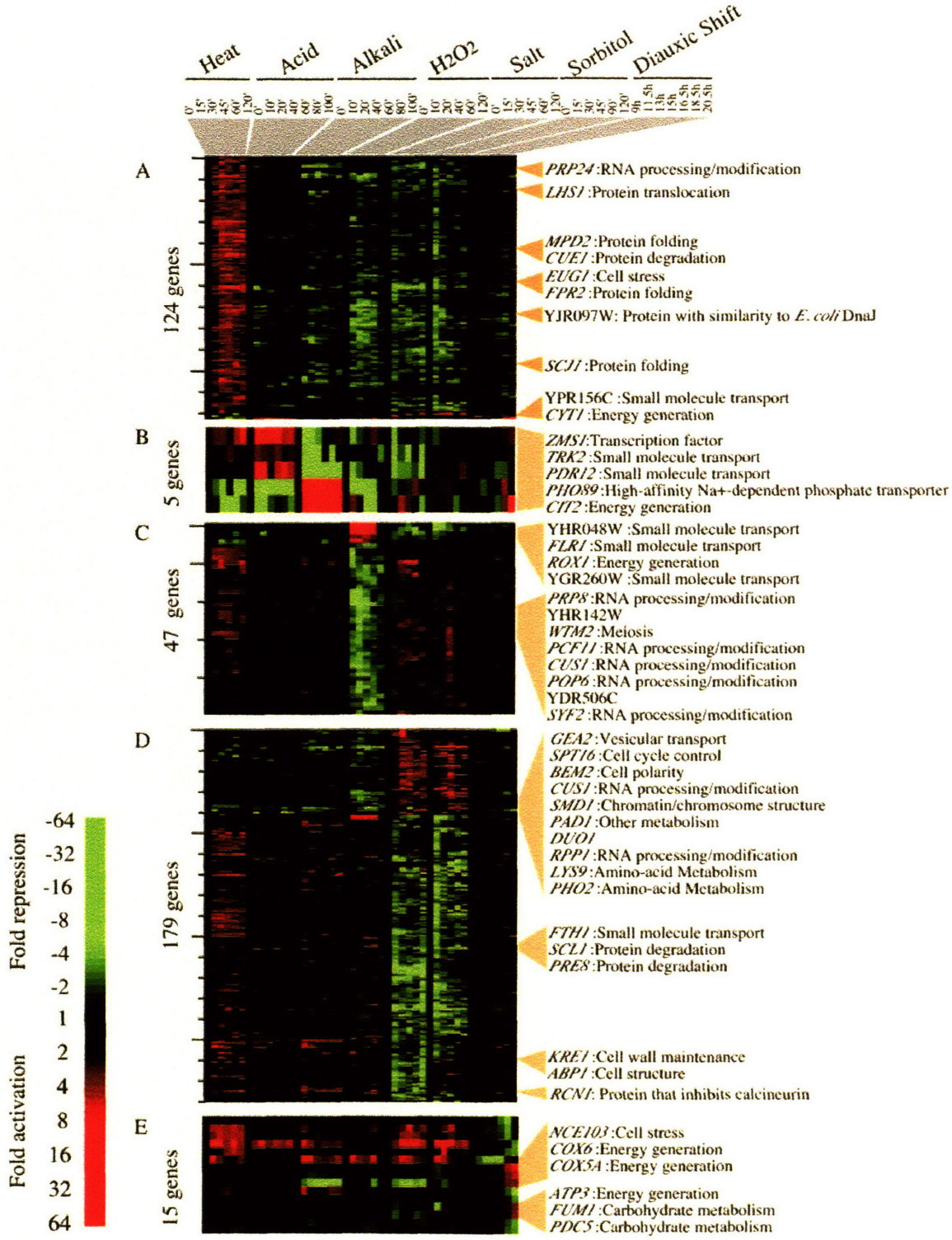

Heat
Acid
Alkali
H2O2
Salt
Sorbitol
Diauxic Shift
0' 15' 30' 45' 60' 120'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 80' 100'
0' 10' 20' 40' 60' 120'
0' 15' 30' 45' 60' 120'
0' 15' 30' 45' 90' 120'
0h 11.5h 13h 15h 16.5h 18.5h 20.5h
A
124 genes
PRP24 :RNA processing/modification
LHS1 :Protein translocation
MPD2 :Protein folding
CUE1 :Protein degradation
EUG1 :Cell stress
FPR2 :Protein folding
YJR097W: Protein with similarity to E. coli DnaJ
SCJ1 :Protein folding
YPR156C :Small molecule transport
CYT1 :Energy generation
B
5 genes
ZMS1:Transcription factor
TRK2 :Small molecule transport
PDR12 :Small molecule transport
PHO89 :High-affinity Na+-dependent phosphate transporter
CIT2 :Energy generation
C
47 genes
YHR048W :Small molecule transport
FLR1 :Small molecule transport
ROX1 :Energy generation
YGR260W :Small molecule transport
PRP8 :RNA processing/modification
YHR142W
WTM2 :Meiosis
PCF11 :RNA processing/modification
CUS1 :RNA processing/modification
POP6 :RNA processing/modification
YDR506C
SYF2 :RNA processing/modification
D
179 genes
GEA2 :Vesicular transport
SPT16 :Cell cycle control
BEM2 :Cell polarity
CUS1 :RNA processing/modification
SMD1 :Chromatin/chromosome structure
PAD1 :Other metabolism
DUO1
RPP1 :RNA processing/modification
LYS9 :Amino-acid Metabolism
PHO2 :Amino-acid Metabolism
FTH1 :Small molecule transport
SCL1 :Protein degradation
PRE8 :Protein degradation
KRE1 :Cell wall maintenance
ABP1 :Cell structure
RCN1: Protein that inhibits calcineurin
E
15 genes
NCE103 :Cell stress
COX6 :Energy generation
COX5A :Energy generation
ATP3 :Energy generation
FUM1 :Carbohydrate metabolism
PDC5 :Carbohydrate metabolism
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64

Figure 5. Environmental response-specific gene expression.

Environmental response-specific genes were selected based on the criteria described in Methods. The genes displayed are a subset of the environmental response-specific genes.

(A) Genes whose expression is uniquely remodeled in response to heat.

(B) Genes whose expression is diametrically regulated in response to acid and alkali.

(C) Genes whose expression is uniquely remodeled in response to hydrogen peroxide.

(D) Genes that respond similarly to salt and sorbitol.

(E) Genes whose expression is specific to the diauxic shift. Some of the response-specific genes are listed on the right.

Acid and Alkali

PDR12 is one of the genes that is regulated in a pH-specific manner (Figure 5B). *PDR12* encodes an ATP-dependent membrane transporter that is highly induced at low pH, but is repressed at high pH. Pdr12 was first identified as a protein induced by sorbic acid and may function to export carboxylate anions out of cells (Holyoak et al., 1999; Piper et al., 1998). The transcription profile of *PDR12* was used as a template to identify genes whose expression is regulated in a reciprocal manner in acid and in alkali. We were interested in identifying those genes whose expression is either reset and maintained at a higher level, or reset and maintained at a lower level during the environmental change. We identified four genes by these criteria (Figure 5B). Two of these genes, *ZMS1* and *TRK2*, are activated by acid treatment, and repressed by alkali treatment. They encode a zinc-finger family transcription factor and a potassium transporter (Ko et al., 1990), respectively. Another two, *CIT2* and *PHO89*, are repressed by acid and activated by alkali. They encode a peroxisomal citrate synthase and sodium-phosphate symporter. The Trk2 potassium transporter is responsible for a K+ current at low pH, and its activity is low at neutral or high pH (Bihler et al., 1999). By contrast, the Pho89 sodium-phosphate symporter catalyzes sodium-dependent phosphate uptake, and its activity is high at alkaline pH (Martinez and Persson, 1998). Thus, for Trk2 and Pho89, both differential transcriptional regulation and protein activity contribute to adaptation of yeast to changes in the pH environment.

Hydrogen Peroxide

The response to hyperoxia involves about a third of the genome and differs from the other stress responses in that the maximal effects on gene expression occur slightly later than in the other stress time courses (Figure 1). Despite this, much of the transcriptome has returned to prestress levels by two hours after the addition of hydrogen peroxide. As expected, *ROX1*, which encodes a transcriptional repressor of hypoxic genes, was among the hyperoxia specific genes (Figure 5C) (Deckert et al., 1995). ROX1 was induced and then repressed, consistent with its autoregulatory activity (Deckert et al., 1995).

High Salt and High Osmolarity

The high salt response is likely to be a composite response to both the osmotic and ionic changes caused by the addition of sodium chloride, whereas, the sorbitol response is expected to be specific for osmotic changes. However, the sets of genes involved in the responses to high salt and to sorbitol are remarkably similar to each other (Figure 5D), suggesting that most of the changes in gene expression are in response to the change in osmotic conditions. There are some genes whose expression is reset in response to a change in the ion concentration. For example, the levels of *RCN1* (YKL159C) drop 8 fold in salt and only 1.4 fold in sorbitol over the time course. Rcn1 is a known inhibitor of calcineurin, which functions to stimulate the transcription of the gene encoding the primary sodium transporter, *ENA1*. It is notable that many of the genes known to be induced in high saline conditions, including *ENA1*, are induced by multiple stresses, suggesting that ion homeostasis is a critical response for most environmental changes.

Diauxic Shift

There are only a few genes whose expression changes at least three-fold in response to nutrient limitation at the diauxic shift, but whose expression changes less than two fold in response to other environmental changes. These include genes required for respiration (e.g. *COX5A* and *COX6)*, indicative of a shift from fermentative to respiratory growth, as observed previously (DeRisi et al., 1997).

Requirement for the Msn2/Msn4 activators

Because the transcriptional activators Msn2 and Msn4 are thought to be involved in induction of genes common to many stress responses, we tested whether Msn2/Msn4 are required for activation of the induced set of CER genes. We selected the acid response for this experiment because the CER contributes to a large percentage of the acid response. Figure 6 compares the response of wild type and *msn2msn4* deleted strains to acid. Of the 193 genes induced in the CER whose expression could be measured in the *msn2msn4* strain, 147 were induced more than two fold in acid. Msn2/Msn4 appear to be required for the induced expression of 136 (93%) of these genes. It is also notable that these activators appear to have a function in overcoming transcriptional repression, as most acid-induced genes are repressed upon treatment with acid in the *msn2msn4* strain.

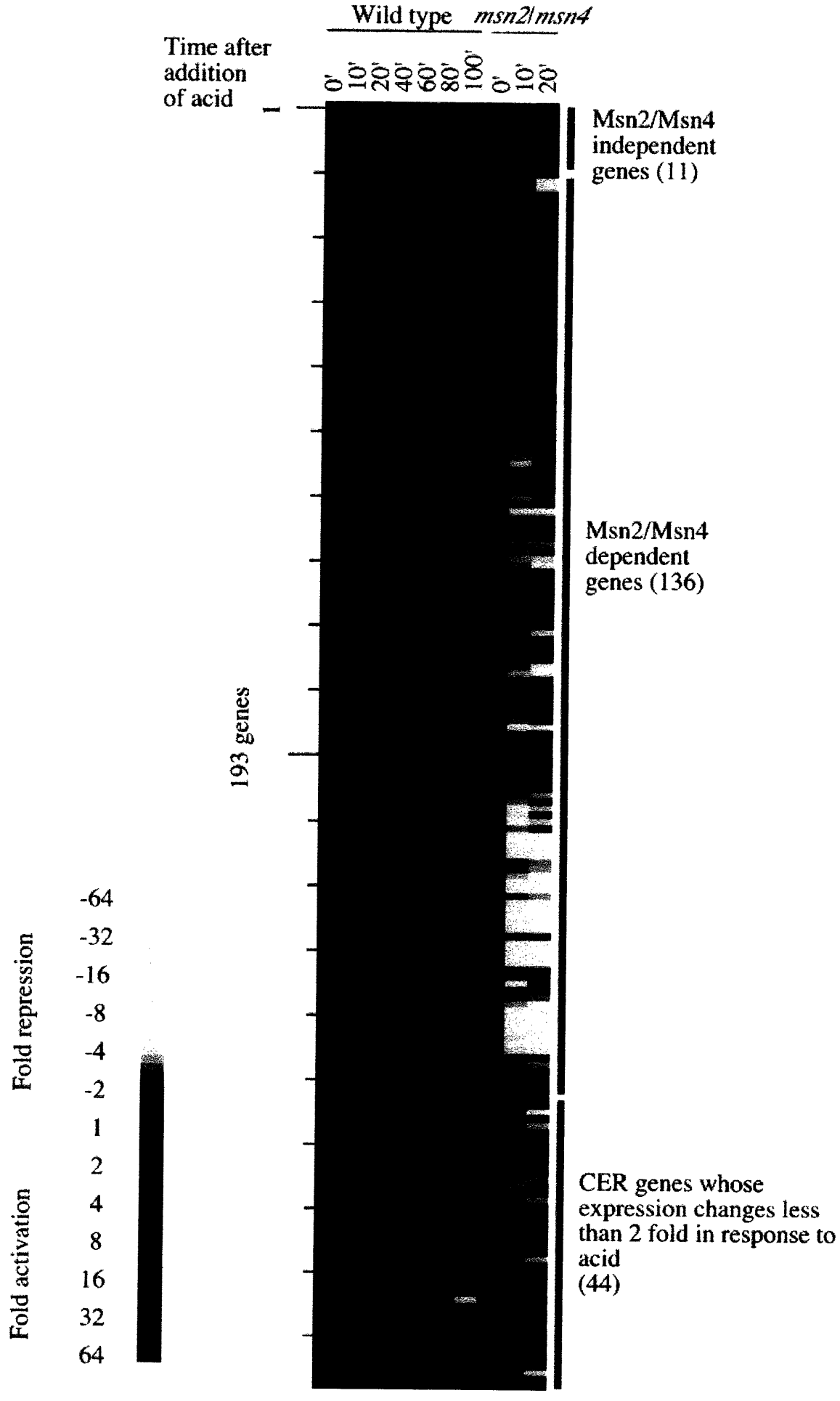

Wild type
msn2/msn4
Time after addition of acid
0'
10'
20'
40'
60'
80'
100'
0'
10'
20'
1
193 genes
Msn2/Msn4 independent genes (11)
Msn2/Msn4 dependent genes (136)
CER genes whose expression changes less than 2 fold in response to acid (44)
Fold repression
Fold activation
-64
-32
-16
-8
-4
-2
1
2
4
8
16
32
64

Figure 6. Induction of most acid-induced genes depends on Msn2/Msn4.

The genes whose expression changed by at least three-fold in response to acid are shown for a wild type strain and a strain deleted for *MSN2* and *MSN4*. Msn2/4 dependent genes are shown with a bar. Details are as described in the legend to Figure 1.

Discussion

Genome-wide expression analysis was used to explore how gene expression is remodeled in response to changes in extracellular environment, including changes in temperature, oxidation, nutrients, pH and osmolarity. We found that approximately two-thirds of the genome is involved in the response to environmental changes. The inclusion of a large fraction of genes in environmental responses reveals the importance of expression remodeling in adapting to environmental changes, and implicates a substantial set of genes with previously uncharacterized cellular roles in these responses. The results of this study identify a common set of genes that are induced and repressed in most of these responses, and these we call common environment response (CER) genes. We find that Msn2/Msn4 are involved in the activation of nearly all of the genes that are induced in the CER.

Approximately 66% of yeast genes are involved in responding to the changes in environment that we have examined here when using stringent analysis criteria. Only genes whose expression changed more than 3 fold in at least one treatment and whose rescaled fluorescent intensities differed from that of the untreated controls by more than 100 units were selected. In addition, those genes that could not be reliably detected in more than 30% of the time points were discarded. Given the large number of genes involved in these responses, it is striking that only ~1000 (17%) yeast genes are thought to be essential for viability under standard laboratory conditions (Winzeler et al., 1999). The large difference in these numbers emphasizes the fact that life has evolved under

PAGES (S) MISSING FROM ORIGINAL

PAGE 213 IS MISSING

conditions in which the environment is continually changing, in contrast to those in the laboratory.

CER

Our results indicate that approximately ~10% of yeast genes are induced or repressed in common when yeast cells respond to a wide variety of environmental changes. It seems likely that this common environmental response is due to a common effect on cells when exposed to almost any change that affects permeability of the cell wall or integrity of protein structure. The permeability of the cell wall increases in response to heat, spheroplasting and ethanol, and the consequences of increased permeability may contribute to a common response (Adams and Gross, 1991; Carratù et al., 1996). Diverse changes in environmental conditions may generally induce molecular chaperones because of their effects on the structural integrity of proteins and this, in turn, may require substantial increases in energy production since many molecular chaperones employ the energy of ATP hydrolysis (Lindquist, 1992).

The identification of a CER helps to explain the phenomena of tolerance and cross-protection, in which pretreatment of cells with a mild environmental change provides protection against a more severe change of either the same or a different nature (Lewis et al., 1995; Park et al., 1997). Not all changes in conditions provide cross-protection, for example, ethanol treatment does not result in increased tolerance to heat, although the converse is true (Piper, 1995). This suggests that some of the environmental change-specific genes play important roles in the response to individual changes in

with our observation that nearly 40% of the genes affected by nutrient limitation at the diauxic shift are also CER genes. These pathways appear to play roles in regulation of both the CER induced and repressed genes, via a variety of mechanisms (Beck and Hall, 1999; Boy-Marcotte et al., 1998; Görner et al., 1998; Klein and Struhl, 1994; Neuman-Silberg et al., 1995; Smith et al., 1998).

The regulation of some of the CER repressed genes has also been reported to involve the protein kinase C pathway (Nierras and Warner, 1999). Our data show that many of the genes which function to restore membrane integrity and to deal with the consequences of membrane damage are part of the CER, consistent with the involvement of the PKC pathway, which is thought to respond to changes in membrane integrity.

The set of CER repressed genes include a large number of genes that encode components of the translation apparatus, including cytoplasmic ribosomal protein genes. The 137 cytoplasmic ribosomal protein genes are coordinately regulated and their transcripts have relatively short half-lives (Li et al., 1999). Transcription of ribosomal protein genes can account for up to half of the RNA polymerase II-mediated transcription initiation events in the cell (Warner, 1999). A transient reduction in the synthesis of ribosomal protein mRNAs would permit energy and other resources to be diverted towards the synthesis and use of molecular chaperones, along with other mechanisms involved in surviving a change in the environment (Jelinsky and Samson, 1999).

The results described here implicate a substantial set of genes with previously

uncharacterized cellular roles in the response to environmental change. These include 118 of the 216 CER induced genes and 106 of then 283 CER repressed genes. These genes can now be defined as environmental change responsive genes.

Exercising the genome

The time course expression profiles described here involve changes in a substantial fraction of the genes in the yeast genome. Studies of the yeast cell cycle revealed that 500 to 800 genes change expression levels significantly during cell cycle progression (Cho et al., 1998; Spellman et al., 1998). Environmental change also "exercises" a large fraction of the genome, in many cases a larger portion than seen with the cell cycle. The availability of genome-wide expression data involving significant portions of the genome should prove to be valuable for efforts to map the regulatory circuits of yeast cells and should serve as a useful foundation for future efforts to increase our understanding of the molecular mechanisms involved in adaptation to environmental change. The large number of genes involved and the temporal changes that occur in these environmental responses should also provide a rich source of information for computational modeling of regulatory networks.

Acknowledgments

Strain Z985 (1097) was the kind gift of Helmut Ruis. We were alerted to the possibility that isozymes are differentially expressed during environmental changes in public talks by members of the laboratory of Dr. Patrick Brown. This work was supported by funding from the Bristol-Meyers Squibb Company, Affymetrix Inc., and Millenium Pharmaceuticals Inc. and grants from the National Institutes of Health (R.A.Y.), the Helen Hay Whitney Foundation (B.R.) and Howard Hughes Medical Institute (E.G.J.).

Methods

Strains

ATCC-201388 *Mata his3*D*1 leu2*D*0 met15*D*0 ura3*D*0*

Z985 (1097) *Mata ade2-1 can1-100* GAL^{+} *his3-11,15 leu2-3,112* psi^{+} *trp1-1 ura3 msn2*D::*HIS3 msn4*D::*TRP1*

wt-P82a *Mata ade2-1 can1-100 leu2-3,112 trp1-1 ura3-1 hsc82*D::*LEU2 hsp82*D::*LEU2 his3::HSP82*

Growth Conditions

Heat Shock

An overnight culture of strain wt-P82a was grown in YPD (1% yeast extract/2% peptone/2% glucose) and used to inoculate 2 (4L) flasks containing 1500 ml of YPD. These were grown to O.D.$_{600}$= 0.5 at 25°C. The temperature shift to 37°C was carried out by the addition of an equal volume of YPD prewarmed to 49°C. Cells were harvested by centrifugation at times 15', 30', 45', 60' and 120' after the temperature shift and from duplicate cultures immediately before the temperature shift (time 0').

The strain wr-P82a was used in this study because it is isogenic to many strains that are used by this and other laboratories to investigate the heat shock responses. Although this strain harbors a null mutation in the HSC82 gene, it is functionally wild type, because it constitutively expresses the HSP82 gene. The HSP82 and HSC82 gene products are ~97% identical and are functionally equivalent (Borkovich et al., 1989).

Acid

An overnight culture of strain ATCC-201388 was grown in YPD and used to inoculate 2 (2L) flasks containing 200 ml of YPD. These were grown to $O.D._{600}$= 0.5 at 30°C. 20ml of succinnic acid (0.5 M, titrated to pH 3.0 with Tris base) was added to a final concentration of 0.05 M. This brought the pH to 4.0. The media remained at pH 4.0 throughout the experiment. Cells were harvested by centrifugation at times 10', 20', 40', 60', 80' and 100' after the addition of acid and from duplicate cultures immediately before the addition of acid (time 0'). The pH of each sample was measured using a pH meter from Orion Research.

Z985 (1097) was grown under the same conditions and cells were harvested at 0', 10' and 20' after the addition of acid.

Alkali

The experiment was carried out as described for acid, except that 20ml Tris-HCl (1M, pH8.25) was added to a final concentration of 0.1M. This brought the pH to 7.9. The media remained at pH 7.9 throughout the experiment.

Hydrogen Peroxide

An overnight culture of strain ATCC-201388 was grown in YPD and used to inoculate 2 (2L) flasks containing 200 ml of YPD. These were grown to $O.D._{600}$= 0.7 at 30°C. Hydrogen peroxide was added to a final concentration of 0.4 mM. Cells were harvested by centrifugation at times 10', 20', 40', 60' and 100' after the addition of hydrogen peroxide and from duplicate cultures immediately before the addition of

hydrogen peroxide (time 0').

Salt

An overnight culture of strain ATCC-201388 was grown in YPD and used to inoculate 7 (2L) flasks containing 200 ml of YPD. These were grown to $O.D._{600}$= 0.6 - 0.8 at 30°C. Sodium chloride was added to a final concentration of 1.0M. Cells were harvested by centrifugation at times 15', 30', 45', 60', 90' and 120' after the addition of sodium chloride and from duplicate cultures immediately before the addition of sodium chloride (time 0').

Sorbitol

The experiment was carried out as described for salt, except that sorbitol was added to a concentration of 1.5M.

RNA Preparation, Probe Preparation and DNA Chip Hybridization

mRNA isolation, cDNA preparation, biotin-labeled cRNA generation and DNA microarray analysis were performed as described previously (Holstege et al., 1998). Microarray analysis was carried out using Affymetrix 6100 or S98 Yeast Genome chips according to standard protocols.

Data acquisition and Processing

The output files from the scanner were downloaded as text files, and then loaded into a custom-built database (ChipDB) for further analysis.

Clustering Analysis of gene expression profiles in all conditions (Figure 1)

The expression profiles corresponding to individual time points were scaled to a common reference profile based on fluorescent intensities of 5 DNA controls that were added after the preparation of total RNA. The normalized data were downloaded from ChipDB as a text file and the SAS software package used to calculate the fold changes for each gene. A genes fold change value was considered to be reliable, and used for analysis, if the fluorescence intensity value was scored as 'present' in at least one of the time points after the change in conditions, or in both of the zero time points. Genes whose expression changed more than 3 fold in at least one treatment and whose rescaled fluorescent intensities differed from that of the untreated controls by more than 100 units were selected. Genes that could not be reliably detected in more than 30% of the time points for a given time course were discarded. Genes that passed these selection criteria were considered to be the set of genes whose expression is changed in response to environmental change. The log-transformed expression values for the selected genes (a total of 3864 genes) were exported to a text file. A non-parametric correlation matrix (Kendall's Tau similarity metric) was calculated for every pair of genes. The resulting matrix was used to cluster these genes into a hierarchical tree using the average linkage method provided in the Cluster program (Eisen et al., 1998).

Identification of common environmental response genes (Figures 2 and 3)

The common environmental response (CER) genes were identified from the clustering output as genes whose expression was induced or repressed in all conditions.

From this list, the genes that changed at least 2 fold in 5 or more time courses were selected as CER genes.

Identification of specific responses to each environmental change (Figure 5)

The environment specific response genes were identified as genes whose expression was induced, or repressed, at least 3 fold in response to a specific change in environment, but which were not included in the CER.

To identify genes whose expression is specifically reset in response to the shift to 37°C, the genes whose expression was induced, or repressed, by more than 3 fold in response to the temperature shift, were selected. To confer specificity, a second criterion required that genes induced in response to heat were not induced by more than two fold in response to any other environment change and genes repressed in response to heat were not repressed by more than two fold in response to any other environmental change. Pearson correlation coefficients (P.C.C.) were calculated for each of these genes against a synthetic reference pattern. The reference pattern required that gene expression was high at each time point after the temperature shift and low before the temperature shift and in response to other environmental changes (heat 0'=0, heat 15' = 1, heat 30' =1, heat 45' = 1, heat 60' =1, heat 120' = 1, acid 0' = 0, acid 10' = 0, etc. 0 represents an arbitrary low value and 1 represents an arbitrary high value). Genes whose profiles gave a P.C.C. greater than 0.60 were defined as genes whose expression remodeled in response to a change in temperature. This strategy was employed to identify genes reset in response to other environmental changes, except that the reference pattern was modified

accordingly. For example, to find genes whose expression is modified in response to hyperoxia, the reference pattern required a high value after the addition of hydrogen peroxide and a low value for other time courses. For the salt-sorbitol response, the reference pattern demanded a high value after the addition of salt or sorbitol, and low prior to the addition of salt or sorbitol or in response to other environmental changes. For the diauxic shift (DeRisi et al., 1997), the only criterion was that the expression was induced, or repressed, more than 3 fold during diauxic shift, but less than 2 fold in other environmental changes (as described for the selection of heat-specific genes).

To identify genes whose expression levels were reset in a reciprocal manner in response to acid and to alkali, the transcription profile of the *PDR12* gene was used as reference pattern against which P.C.C.s were calculated. The genes selected are those for which the P.C.C. exceeds 0.60 and whose expression level induced, or repressed, by more than 3 fold in acid or alkali, and less than 2 fold in response to any other environmental changes (as described for the selection of heat-specific genes).

Identification of Msn2/Msn4 dependent CER genes (Figure 6)

The genes induced in the CER were examined under acid conditions in wild type yeast and in *msn2msn4* strains. Among the 216 CER acid induced genes, the transcript levels of 193 genes could be reliably measured in the *msn2msn4* mutants. From these genes, 147 genes whose expression is induced at least 2 fold after addition of acid (compared with wild type cells) were used for analysis. The expression values of these genes in wild type and *msn2msn4* strains were clustered using the Cluster program (Eisen

et al., 1998). Genes that are dependent on Msn2/Msn4 for their induction were identified as those with a signature showing a high level of induction in wild type and no induction in the *msn2/msn4* strain.

References

Adams, C. C., and Gross, D. S. (1991). The Yeast Heat Shock Response Is Induced by Conversion of Cells to Spheroplasts and by Potent Transcriptional Inhibitors. J. Bacteriol. *173*, 7429-7435.

Ali, A., Bharadwaj, S., O'Carroll, R., and Ovsenek, N. (1998). HSP90 interacts with and regulates the activity of heat shock factor 1 in *Xenopus* oocytes. Mol. Cell. Biol. *18*, 4949-4960.

Beck, T., and Hall, M. N. (1999). The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. Nature. *402*, 689-692.

Bharadwaj, S., Ali, A., and Ovsenek, N. (1999). Multiple components of the HSP90 chaperone complex function in regulation of the heat shock factor 1 in vivo. Mol. Cell. Biol. *19*, 8033-8041.

Bihler, H., Gaber, R. F., Slayman, C. L., and Bertl, A. (1999). The presumed potassium carrier Trk2p in *Saccharomyces cerevisiae* determines an H+-dependent, K+-independent current. FEBS Lett. *447*, 115-120.

Boy-Marcotte, E., Perrot, M., Bussereau, F., Boucherie, H., and Jacquet, M. (1998). Msn2p and Msn4p Control a Large Number of Genes Induced at the Diauxic Transition Which are Repressed by Cyclic AMP in *Saccharomyces cerevisiae*. J. Bacteriol. *180*, 1044-1052.

Carratù, L., Franceschelli, S., Pardini, C. L., Kobayashi, G. S., Horvath, I., Vigh, L., and Maresca, B. (1996). Membrane lipid perturbation modifies the set point of the temperature of heat shock response in yeast. P.N.A.S. (USA). *93*, 3870-3875.

Chapman, R., Sidrauski, C., and Walter, P. (1998). Intracellular signaling from the endoplasmic reticulum to the nucleus. Ann. Rev. Cell. Dev. Biol. *14*, 459-485.

Cho, R. J., Campbell, M. J., Winzeler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J., and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell. *2*, 65-73.

Craig, E. (1992). The heat shock response of *Saccharomyces cerevisiae*. In The Molecular and Cellular Biology of the Yeast *Saccharomyces cerevisiae*: Gene expression: Cold Spring Harbor Laboratory Press), pp. 501-537.

Craven, R. A., Egerton, M., and Stirling, C. J. (1996). A novel Hsp70 of the yeast ER lumen is required for the efficient translocation of a number of protein precursors. EMBO J. *15*, 2640-2650.

Deckert, J., Perini, R., Balasubramanian, B., and Zitomer, R. S. (1995). Multiple elements and auto-repression regulate Rox1, a repressor of hypoxic genes in *Saccharomyces cerevisiae*. Genetics. *139*, 1149-58.

Deckert, J., Torres, A. M. R., Simon, J. T., and Zitomer, R. S. (1995). Mutational analysis of Rox1, a DNA-binding repressor of hypoxic genes in Saccharomyces cerevisiae. Mol. Cell. Biol. *15*, 6109-6117.

DeRisi, J. L., Iyer, V., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science. *278*, 680-686.

Duina, A. A., Kalton, H. M., and Gaber, R. F. (1998). Requirement for Hsp90 and a CyP-40-type cyclophilin in negative regulation of the heat shock response. J. Biol. Chem. *273*, 18974-18978.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression paptterns. P.N.A.S. (USA). *95*, 14863-14868.

Ellis, R. J. (1999). Molecular chaperones: pathways and networks. Curr. Biol. *9*, R137-R139.

Estruch, F., and Carlson, M. (1993). Two homologous Zinc Finger Genes Identified by Multicopy Suppression in a SNF1 Protein Kinase Mutant of *Saccharomyces cerevisiae*. Mol. Cell. Biol. *13*, 3872-3881.

Görner, W., Durchschlag, E., Martinex-Pastor, M. T., Estruch, F., Ammerer, G., Hamilton, B., Ruis, H., and Schüller, C. (1998). Nuclear localization of the C_2H_2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. Genes and Dev. *12*, 586-597.

Holstege, F. C. P., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. S., Lander, E. S., and Young, R. A. (1998). Dissecting the Regulatory Circuitry of a Eukaryotic Genome. Cell. *95*, 717-728.

Holyoak, C. D., Bracey, D., Piper, P. W., Küchler, K., and Coote, P. J. (1999). The *Saccharomyces cerevisiae* weak-acid-inducible ABC transporter Pdr12 transports fluorescein and preservative anions from the cytosol by an energy-dependent mechanism. J. Bacteriol. *181*, 4644-4652.

Jelinsky, S. A., and Samson, L. D. (1999). Global response of *Saccharomyces cerevisiae* to an alkylating agent. P.N.A.S. (USA). *96*, 1486-1491.

Kim, C. H., and Warner, J. R. (1983). Mild Temperature Shock Alters the Transcription of a Discrete Class of *Saccharomyces cerevisiae* Genes. Mol. Cell. Biol. *3*, 457-465.

Kingston, R. E., Schuetz, T. J., and Larin, Z. (1987). Heat-inducible human factor that binds to a human *hsp70* promoter. Mol. Cell. Biol. *7*, 1530-1534.

Klein, C., and Struhl, K. (1994). Protein kinase A mediates growth-regulated expression of yeast ribosomal protein genes by modulating RAP1 transcriptional activity. Mol. Cell. Biol. *14*, 1920-1928.

Ko, C. H., Buckley, A. M., and Gaber, R. F. (1990). TRK2 is required for low affinity K+ transport in *Saccharomyces cerevisiae*. Genetics. *125*, 305-312.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., Brown, P. O., and Davis, R. W. (1997). Yeast microarrays for genome wide parallel genetic and gene expression analysis. P.N.A.S. (USA). *94*, 13057-13062.

Lewis, J. G., Learmonth, R. P., and Watson, K. (1995). Induction of heat, freezing and salt tolerance by heat and salt shock in *Saccharomyces cerevisiae*. Microbiol. *141*, 687-694.

Li, B., Nierras, C. R., and Warner, J. R. (1999). Transcriptional elements involved in the repression of ribosomal protein synthesis. Mol. Cell. Biol. *19*, 5393-5404.

Lindquist, S. (1992). Heat-shock proteins and stress tolerance in microorganisms. Curr. Op. Genet. Dev. *2*, 748-755.

Marchler, G., Schüller, C., Adams, G., and Ruis, H. (1993). A *Saccharomyces cerevisiae* UAS element controlled by protein kinase A activates transcription in response to a variety of stress conditions. EMBO J. *12*, 1997-2003.

Martinez, P., and Persson, B. L. (1998). Identification, cloning and characterization of a derepressible Na+-coupled phosphate transporter in *Saccharomyces cerevisiae*. Mol. Gen. Genet. *258*, 628-638.

Martinez-Pastor, M. T., Marchler, G., Schüller, C., Marchler-Bauer, A., Ruis, H., and Estruch, F. (1996). The *Saccharomyces cerevisiae* zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress-response element (STRE). EMBO J. *15*, 2227-2235.

Miller, M., Xuong, N.-H., and Geiduschek, E. P. (1982). Quantitative analysis of the heat shock response of *Saccharomyces cerevisiae*. J. Bacteriol. *151*, 311-327.

Miller, M. J., Xuong, N., and Geidushek, E. P. (1979). A response of protein synthesis to temperature shift in the yeast *Saccharomyces cerevisiae*. P.N.A.S. (USA). *76*, 5222-5225.

Moskvina, E., Schüller, C., Maurer, C. T. C., Mager, W. H., and Ruis, H. (1998). A search in the genome of *Saccharomyces cerevisiae* for genes regulated via stress response elements. Yeast. *14*, 1041-1050.

Neuman-Silberg, F. S., Bhattacharya, S., and Broach, J. (1995). Nutrient availability and the *RAS*/cyclic AMP pathway both induce expression of ribosomal protein genes in *Saccharomyces cerevisiae* but by different mechanisms. Mol. Cell. Biol. *15*, 3187-3196.

Nierras, C. R., and Warner, J. R. (1999). Protein kinase C enables the regulatory circuitry that connects membrane synthesis to ribosome synthesis in *Saccharomyces cerevisiae*. J. Biol. Chem. *274*, 13235-13241.

Park, J.-I., Grant, C. M., Attfield, P. V., and Dawes, I. W. (1997). The freeze-thaw stress response of the yeast *Saccharomyces cerevisiae* is growth phase specific and is controlled by nutritional state via the RAS-cyclic AMP signal transduction pathway. Appl. Env. Microbiol. *63*, 3818-3824.

Parker, C. S., and Topol, J. (1984). A *Drosophila* RNA polymerase II transcription factor specific for the heat-shock gene binds to the regulatory site of an *hsp70* gene. Cell. *37*, 273-283.

Piper, P., Mahé, Y., Thompson, S., Pandjaitan, R., Holyoak, C., Egner, R., Mühlbauer, M., Coote, P., and Kuchler, K. (1998). The Pdr12 ABC transporter is required for the development of weak organic acid resistance in yeast. EMBO J. *17*, 4257-4265.

Piper, P. W. (1995). The heat shock and ethanol stress responses of yeast exhibit extensive similarity and functional overlap. FEMS Microbiol. Lett. *134*, 121-127.

Rep, M., Krantz, M., Thevelein, J. M., and Hohmann, S. (2000). The transcriptional response of *Saccharomyces cerevisiae* to osmotic shock. J. Biol. Chem. *275*, 8290-8300.

Rep, M., Reiser, V., Gartner, U., Thevelein, J. M., Hohmann, S., Ammerer, G., and Ruis, H. (1999). Osmotic stress-induced gene expression in *Saccharomyces cerevisiae* requires Msn1p and the novel nuclear factor Hot1p. Mol. Cell. Biol. *19*, 5474-5485.

Rutherford, S. L., and Lindquist, S. (1998). Hsp90 as a capacitor for morphological evolution. Nature. *396*, 336-310.

Saris, N., and Makarow, M. (1998). Transient ER retention as stress response: conformational repair of heat-damaged proteins to secretion-competent structures. J. Cell Sci. *111*, 1575-1582.

Schmitt, A. P., and McEntee, K. (1996). Msn2p, a zinc finger DNA-binding protein, is the transcriptional activator of the multistress response in *Saccharomyces cerevisiae*. P.N.A.S. (USA). *93*, 5777-5782.

Smith, A., Ward, M. P., and Garrett, S. (1998). Yeast PKA represses Msn2p/Msn4p-dependent gene expression to regulate growth, stress response and glycogen accumulation. EMBO J. *13*, 3556-3564.

Sorger, P. K., and Pelham, H. R. B. (1987). Purification and characterization of a heat shock element binding protein from yeast. EMBO J. *6*, 3035-3041.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. Mol. Biol. Cell. *9*, 3273-3297.

Travers, K. J., Patil, C. K., Wodicka, L., Lockhart, D. J., Weissman, J. S., and Walter, P. (2000). Functional and genomic analyses reveal an essential coordination between the unfolded protein response and ER-associated degradation. Cell. *101*, 249-258.

Treger, J. M., Schmitt, A. P., Simon, J. R., and McEntee, K. (1998). Transcriptional factor mutations reveal regulatory complexities of heat shock and newly identified stress genes in *Saccharamyces cerevisiae*. J. Biol. Chem. *273*, 26875-26879.

Warner, J. (1999). The economics of ribosome biosynthesis in yeast. TIBS. *24*, 437-40.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., Bakkoury, M. E., Foury, F., Friend, S., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittman, M., Pai, C., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Veronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmerman, K., Philippsen, P., Johnston, M., and Davis, R. W. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science. *285*, 901-906.

Wolfe, K. H., and Shields, D. C. (1997). Molecular evidence for an ancient duplication of the entire yeast genome. Nature. *387*, 708-713.

Wu, C. (1985). An exonuclease protection assay reveals heat-shock element and TATA-box binding proteins in crude nuclear extracts. Nature. *317*, 84-87.

Zou, J., Guo, Y., Guettouche, T., Smith, D. F., and Voellmy, R. (1998). Repression of Heat Shock Transcription Factor HSF1 Activation by HSP90 (HSP90 Complex) that Forms a Stress-Sensitive Complex with HSF1. Cell. *94*, 471-480.