

Genome-wide Expression and Location Profiling in *Saccharomyces cerevisiae*:
Experimental and Graphical Analysis

by

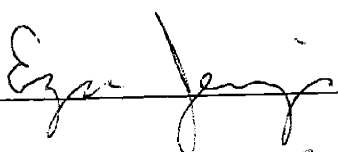
Ezra Jennings
A.B., Molecular Biology
Princeton University, 1993

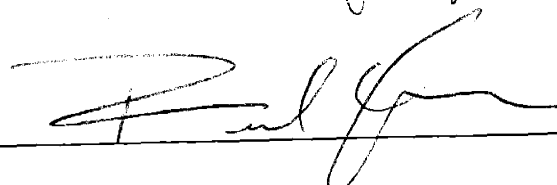
SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

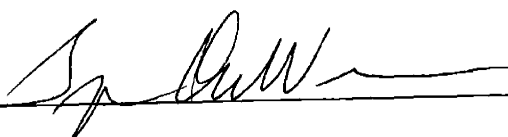
DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
September 2002

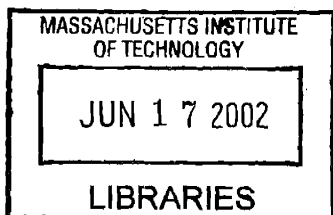
©Ezra Jennings, 2002. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute
publicly copies of this thesis document in whole or in part.

Signature of Author  _____
Department of Biology
June 14, 2002

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

Accepted by  _____
Terry Orr-Weaver
Professor of Biology and
Co-Chairperson, Biology Graduate Committee



ARCHIVES

Dedication

In memory of Raymond

Acknowledgments

I would like to thank all of the people who helped me in one way or another through graduate school. In the Young lab, I would like to thank Frank Holstege for spearheading the microarray effort in the lab; John Barnett, David Chao, Ellen Gadbois, Nancy Hannett, Chris Harbison, Christoph Hengartner, Sang Seok Koh, Tony Lee, Heather Murray, Peter Murray, Vic Myer, Duncan Odom, Dmitry Pokholok, Bing Ren, Joan Richmond, Nicola Rinaldi, François Robert, Ann Schlesinger, Jörg Schreiber, Itamar Simon, Jolyon Terragni, Craig Thompson, Tom Volkert, Chris Wilson, Peter Young, and Julia Zeitlinger for various scientific and non-scientific discussions in the laboratory; and the three members of my room for most of my time in the Young Lab, Jerry Nau, John Wyrick and my baymate, Helen Causton, for scientific discussions and encouragement. Additionally, I would like to thank Charles Tilford, Steve Rozen and Fran Lewitter for discussions and computational assistance.

Outside of the laboratory, I would like to my roommates, Bart and Peter for their encouragement and helpful diversions and to my parents and Alix for their unending support.

Thanks to Prof. Steve Bell, Prof. Eric Lander, Prof. Phil Sharp and Prof. Marc Vidal for their time and willingness to serving on my thesis committee.

I would also like to thank my advisor, Rick Young, for his willingness to let me pursue my interests in computers and to create a non-traditional graduate program.

Genome-wide Expression and Location Profiling in *Saccharomyces cerevisiae*:
Experimental and Graphical Analysis

by

Ezra Jennings

Submitted to the Department of Biology on June 14, 2002
in partial fulfillment of the requirements for the Degree of
Doctor of Philosophy in Biology

Abstract

Genome-wide expression analysis was used to identify genes whose expression depends on the functions of key components of the transcription initiation machinery in yeast. Components of the RNA polymerase II holoenzyme, the general transcription factor TFIIID, and the SAGA chromatin modification complex were found to have roles in expression of distinct sets of genes. The results reveal an unanticipated level of regulation which is superimposed on that due to gene-specific transcription factors, a novel mechanism for co-ordinate regulation of specific sets of genes when cells encounter limiting nutrients, and evidence that the ultimate targets of signal transduction pathways can be identified within the initiation apparatus.

Understanding how DNA-binding proteins control global gene expression and chromosomal maintenance requires knowledge of the chromosomal locations where these proteins function *in vivo*. We developed a microarray method that reveals the genome-wide location of DNA-bound proteins and used this method to monitor binding of gene-specific transcription activators in yeast. A combination of location and expression profiles was used to identify genes whose expression is directly controlled by Gal4 as cells respond to changes in carbon source, and by Thi2 in the absence or presence of thiamin. The results identify pathways that are coordinately regulated by these regulators and reveal novel functions for these regulators. Understanding a transcriptional network such as these will be useful in constructing a cellular regulatory network map.

The use of microarray technology has created new challenges in data analysis for biologists. Visual displays can greatly facilitate the analysis and communication of large quantities of data. We have created a Graphical Display Suite (GDS) that consists of a collection of tools to assist in the visualization of data from genome-wide experiments in *S. cerevisiae*. The GDS is web-accessible, easy to use, and additional components can easily be incorporated into its interface. This suite of tools has proven to be useful in revealing important biological insights.

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

Table of Contents

Title Page	1
Dedication	2
Acknowledgments	3
Abstract	4
Table of Contents	5
Chapter 1: Introduction: Microarrays and Genome-wide Analysis	6
Chapter 2: Dissecting the Regulatory Circuitry of a Eukaryotic Genome	60
Chapter 3: Genome-wide Location Analysis	105
Part I: Genome-wide Location and Function of the DNA Binding Protein Gal4	107
Part II: Thi2 is a Transcriptional Activator of Thiamin Biosynthetic Genes <i>in Vivo</i>	124
Chapter 4: GDS: A Graphical Display Suite for Visualization of Yeast Microarray Data	165
Appendix A: Genome-wide Expression and the World Wide Web	194
Appendix B: Interplay of Positive and Negative Regulators in Transcription Initiation by RNA Polymerase II Holoenzyme	203

Chapter 1

Introduction: Microarrays and Genome-wide Analysis

Introduction

The sequencing of genomes and the ability to immobilize DNA onto solid supports has driven the development of DNA microarrays. Instead of biologists' studying single genes, microarrays have broadened the scope of biology to the analysis of whole genomes. As a result of these massively parallel experiments, huge amounts of data are produced. Unfortunately, conclusions cannot easily be drawn from just looking at the data. Several analytical methods are being applied to microarray data to reveal its underlying structure. As it is difficult for people to visually process large amounts of data, a variety of graphical displays have been very useful in communicating these data. Although many early microarray experiments were generally descriptive, biologists, often with the help of computer scientists and statisticians, are beginning to gain sophisticated insights into several different areas of biology.

In this chapter, I will discuss four major topics, focussing on expression profiling with DNA microarrays. Initially, I will review the different types of microarrays. Then I will provide an overview of data analysis, followed by a discussion of types of visual displays that are used with microarray data. Finally I will highlight some examples of biological insights that have been gained through the use of microarrays.

Microarrays

The origin of microarrays essentially dates back to the development of the Southern blot, where it was discovered that an immobilized DNA could be hybridized to

a labeled probe DNA, thereby identifying the immobilized DNA (Southern, 1975). From the Southern blot eventually came the reverse dot blot, where the probe is immobilized on a solid support and the DNA being tested is labeled and hybridized to the probe (Saiki et al., 1989). The use of glass supports instead of membranes allowed for the spotting of a higher density of probes, making the microarray a high throughput assay. The invention of PCR (Mullis et al., 1986) facilitated the production of cDNA probes and the systematic identification of genomic sequences provided the ability to generate relevant probes for arrays. Refinement in oligonucleotide synthesis made producing thousands of different primers a feasible task for production of cDNA arrays. Advances in *in situ* synthesis with photolithography led to the popularization of oligonucleotide arrays (Fodor et al., 1991). With the convergence of these advancements, microarrays have become an invaluable tool for many biologists who can now perform thousands of experiments in parallel.

There are several different types of microarrays. Most microarrays are DNA-based arrays, but there are also protein arrays, as well as chemical or whole cell arrays. DNA arrays have had two major functions: gene expression profiling and sequencing. Arrays for expression profiling are used with mRNA and report the relative amounts of transcript for every gene in the cell for which there is a probe. Sequencing arrays are hybridized with DNA and read individual bases. DNA can also be hybridized to arrays to determine if regions of DNA are present in a sample. This can be for assaying cellular DNA to determine if a region has been amplified or deleted, or it can be for determining which DNAs have been enriched through a selection such as chromatin immunoprecipitation.

Expression Arrays. DNA microarrays used to study gene expression have been popularized by the commercial availability of arrays from Affymetrix (Lockhart et al., 1996) and the do-it-yourself approach advocated by Patrick Brown and colleagues (DeRisi et al., 1997). These two arrays rely on different technologies for their production and function. Affymetrix arrays use oligonucleotides synthesized *in situ* to measure mRNA levels (Lockhart et al., 1996). The oligonucleotides are built up directly on the solid support of the array using photolithography (Fodor et al., 1991). A mask is created such that when light is shined upon the mask covering the array, the light that passes through the holes in the mask is able to activate the chemical groups present on the array. The next nucleotide is added and by this process unique oligonucleotides are constructed at each position on the array, which is called a feature. As there are four different bases in DNA and Affymetrix usually uses 20 to 25-base long oligonucleotides, 80 to 100 different masks are required to create a single array (Lockhart and Winzeler, 2000). The photolithographic masks are expensive to construct, but once they are made they can be used to create large numbers of arrays relatively cheaply.

Probes for more than 15,000 genes can be placed on each array. A probe set monitors the mRNA level from a single gene and actually consists of up to 40 different oligonucleotides species. Half of these oligonucleotides are a perfect match to the sequence of interest. The other half of the oligonucleotides are identical to the first half except for a single mismatch in the central position of the oligonucleotide. The expression level for each gene is calculated using a formula that effectively subtracts mismatch signal from perfect match signal. The principle behind this is that any non-

specific hybridization to the perfect match oligonucleotides should also hybridize to the mismatch oligonucleotides and thereby not contribute to the final measure of that particular mRNA in the sample.

There are advantages and disadvantages to the Affymetrix system compared to other DNA array platforms. One advantage is the use oligonucleotides as probes. Oligonucleotides allow for higher specificity than cDNAs, and allow for discrimination among closely related genes or isoforms of a single gene. Affymetrix chips are also encased in a cassette, which allows for thorough mixing of the sample (target) with the probes on the array. This cassette is also a disadvantage in that Affymetrix chips can only be scanned on Affymetrix scanners. Additionally, these scanners read at only one wavelength, and consequently only a single sample can be applied to a single chip. Two-color scanners allow for two samples to be hybridized to a single chip and therefore allow direct comparison on a single chip.

Other researchers are devising methods of producing oligonucleotide arrays. A group from the University of Wisconsin has developed a maskless photolithographic technique that relies on micromirrors (Singh-Gasson et al., 1999). Rosetta Inpharmatics has used an ink-jet printing technology to synthesize 60-mer oligonucleotides *in situ* on glass slides (Blanchard et al., 1996; Hughes et al., 2001). Ink jet synthesis avoids costly mask production and requires only one printing cycle for each nucleotide added to the ends of the growing oligonucleotides, as opposed to four cycles for photolithographic methods. Qiagen sells through its Operon subsidiary sets of 70mer oligonucleotides that probe all known genes in a variety of organisms, including yeast and humans. These oligonucleotides have been printed on glass slides using commercially available spotting

robots and provide a simpler alternative to *in situ* synthesis (T.L. Volkert and J. Terragni, pers. comm.).

Spotting cDNAs is the other popular technology used for making DNA arrays. Although several vendors are selling high density spotted cDNA arrays, many laboratories are printing their own arrays. In theory all that is required are glass slides coated with a positively charged molecule such as poly-L-lysine or amino-silane, the desired cDNAs to use as probes, and a robot to perform the printing. In practice getting a functional array can be tricky. Isolation of cDNAs from an organism such as *S. cerevisiae* that has relatively few introns can be performed by using PCR with total genomic DNA or total cDNA. However, in many cases performing PCR from whole cDNA can be difficult due to the complexity of the cDNA, and therefore researchers generally order preidentified clones of cDNAs from vendors or collections. Unfortunately, there is significant error in the annotation of clones from many sources, so ideally each clone should be partially sequenced to verify its identity.

As most commercially available scanners can read glass slide arrays, two differentially labeled samples can be applied to a single array, allowing for direct comparison of the samples without having to account for differences between arrays. Improved dyes may eventually allow for the hybridization of a third control sample to facilitate comparison of samples across multiple arrays. DNA with an unknown sequence can also be spotted onto slides; only those spots of interest are later sequenced (Hayward et al., 2000). A problem with cDNA arrays is the lack of specificity of the probes for closely related genes. In general, large (500-1000 bp) cDNAs are unable to

discriminate between closely related genes or multiple isoforms of a single gene, a problem that can generally be solved with the use of oligonucleotide probes.

An alternative method of monitoring gene expression levels is SAGE (serial analysis of gene expression), which does not use an array-based approach at all. SAGE involves the concatenation and cloning of 9-10 base pair tags from cDNAs into vectors, which are subsequently sequenced (Velculescu et al., 1995). In theory, nine base pairs is sufficient to produce a unique relationship between tag and gene for most genomes, but it has been shown that there can be ambiguity in the assignment of a SAGE tag to a gene in studies with human cells (Lee et al., 2002). For ambiguous tags, identification of the region between the tag and the polyadenylation site allows for a more distinguishable identifier of a transcript (Chen et al., 2002). Unfortunately, SAGE is very labor intensive and costly due to extensive sequencing.

Sequencing Arrays. Another application of a DNA microarray is to determine the sequence of a sample of known origin, or more specifically, to determine genotypic differences among a group of samples. An early example of a sequencing array used oligonucleotides with poly-dT tails crosslinked to nylon membranes to determine genotypes at two loci (Saiki et al., 1989). Southern proposed *in situ* synthesis on a glass slide to create the complete set of oligonucleotides of a particular length (e.g. 8 nucleotides; Southern et al., 1992). Small stretches of sequence could be determined by assembling the results of the hybridization. More recently, Affymetrix designed high density oligonucleotide arrays to sequence a portion of human mitochondrial DNA and of HIV (Chee et al., 1996; Kozal et al., 1996). For each base to be read there are four

probes that are identical except for the nucleotide at the central reading position. In a sample that is homozygous at a particular nucleotide, one of the four probes will detect both alleles. In a heterozygous sample, each allele will be detected individually. This technique has also been used to determine single nucleotide polymorphisms (SNPs) (Hacia et al., 1999; Wang et al., 1998). Although this technique is very powerful for surveying a large number of potential genetic differences among samples, it is not useful for detecting insertions or deletions.

An alternative to hybridization-based arrays for sequencing are ones based on primer extension and mass spectrometry, developed by Sequenom (Little et al., 1997). In this system (reviewed in Jurinke et al., 2001), the polymorphic region is amplified by PCR with one of the primers having a biotin group at its 5' end. The PCR product is immobilized on solid support via the biotin and denatured, leaving only one strand of DNA attached to the surface. The PROBE (primer oligo base extension) reaction is then run where the addition of oligonucleotide primers designed to be complementary to the region of DNA adjacent the polymorphic region are added to the array (Braun et al., 1997). Primer extension is carried out in the presence of three deoxynucleotides and one dideoxynucleotide. The dideoxynucleotide is chosen such that the result will be different length extensions of the primer depending on the polymorphic sequence. The extended primers' masses are determined by MALDI-TOF MS (matrix-assisted laser desorption/ionization time-of-flight mass spectrometry). This technique eliminates possible ambiguities that may arise from the analysis of heterozygous samples with hybridization-based methods. Additionally, it functions with insertions and deletions and

can be used to analyze small tandem repeats. However, many PCR reactions must still be performed for each polymorphic region to be sequenced.

Arrays for Genome Location Profiling. In addition to the ability to monitor the expression levels of genes, it is now possible to determine where a particular protein is binding to DNA. The assay is based on the chromatin immunoprecipitation, a procedure that uses formaldehyde to crosslink molecules to one another, followed by an immunoprecipitation with an antibody directed against the protein of interest (Braunstein et al., 1993; Orlando and Paro, 1993). The crosslinks are reversed and specific regions of DNA that were bound to the protein, as well as unenriched DNA from whole cell extracts, are independently amplified by PCR and detected by gel electrophoresis. For genome-wide location analysis, the enriched and unenriched DNAs are labeled with different fluorophores and hybridized to an array with the appropriate DNAs spotted on it. Therefore all regions of the genome can be queried simultaneously instead of interrogating individual regions by PCR.

As many proteins interact with DNA in the promoter regions of genes, glass slides with the intergenic regions printed on them have proven to be a useful tool for this sort of genome-wide location analysis. By using the intergenic arrays in combination with standard cDNA (open reading frame) arrays, virtually the entire yeast genome can be queried. These arrays have been used successfully with yeast to determine where three transcriptional regulators, Gal4, Ste12, and Rap1, bind to DNA (Lieb et al., 2001; Ren et al., 2000), to determine which genes are bound by known regulators of the cell cycle (Iyer et al., 2001; Simon et al., 2001), and to identify candidate autonomously replicating

sequences that are recognized by the ORC and MCM proteins (Wyrick et al., 2001). This technique has also been applied to human cells to determine to which promoters the cell cycle transcription factor E2F4 binds, although only a fraction of potential human gene promoters were on the arrays (Ren et al., 2002; Weinmann et al., 2002).

Protein Arrays. Although DNA arrays have become accessible and popular in recent years, protein arrays are still under development (Kodadek, 2001; Templin et al., 2002; Wilson and Nock, 2002). Protein arrays can be divided into two functional groups: protein identity arrays and protein function arrays (Kodadek, 2001). Each group has its non-array predecessor. The idea of parallel analysis of protein levels in cells has been around for years and carried out by using two-dimensional gel electrophoresis to separate the proteins (reviewed in Fey and Larsen, 2001). In general, radiolabeled cell lysates are used for visual identification of spots and individual spots on the gel can be identified by mass spectrometry. In theory, if 2-D gels could be run with high reproducibility, once a spot had been identified for a given organism, its coordinates on the gel would be known and could be used to identify it again on subsequent runs of a 2-D gel. In practice, this has not proved very practical and alternatives for measuring the proteome are being developed. The yeast two-hybrid method (Fields and Song, 1989) has been applied on a genome-wide level for yeast (Ito et al., 2001; Ito et al., 2000; Uetz et al., 2000) and hepatitis C virus (Flajolet et al., 2000), and on a pilot scale for the mouse (Suzuki et al., 2001), in an attempt to define functions for many proteins by determining which proteins interact with each other. Unfortunately, highly parallel two-hybrid screens are very labor

intensive, do not work well with transcription factors or membrane-associated proteins, and can be prone to high rates of false positives (reviewed in Serebriiskii et al., 2001).

Many of the protein arrays that have been constructed at this point are protein function arrays, in that a set of proteins of interest are deposited on a solid support and a particular assay is performed. MacBeath and Schreiber (2000) spotted three different purified proteins in quadruplicate and used fluorescently labeled proteins known to interact with the spotted proteins as probes. Although this study did not describe a functionally useful array, it demonstrated that protein arrays could be constructed using available glass slides and microarray robots. This technology was extended to almost the entire yeast proteome, where over 5,800 yeast open reading frames were expressed and purified as GST-His₆ fusions and spotted onto nickel-coated glass slides (Zhu et al., 2001). In this study they used biotinylated calmodulin as a probe and detected known interactions as well as uncharacterized ones. They also used biotinylated liposomes as probes to detect protein-lipid interactions.

There are currently very few examples of protein identification arrays, as these are much more difficult to construct. Either a reagent (e.g. an antibody) that uniquely identifies a protein must be identified and spotted onto a chip, or mass spectrometry can be used. Brown and colleagues (Haab et al., 2001) used 115 antibody-antigen pairs to create either an antigen array or an antibody array for antibody or antigen identification, respectively. This array was constructed using a standard microarray printer and a glass slide and demonstrates the potential for a protein detecting array using antibodies. Another group deposited over 50 antibodies against various CD cell surface antigens onto an array (Belov et al., 2001). They applied whole cells to the array to generate a profile

of the immune system of the cell donor. Several companies are now beginning to offer first generation prefabricated antibody arrays, making this technology accessible to a wider audience.

Ciphergen has developed a product they call the ProteinChip Array, which can be used for measuring protein levels from a sample (reviewed in Weinberger et al., 2002). The assay is composed of two parts: an initial sample fractionation followed by protein detection. Fractionation is carried out by passing the sample over arrays with varying chromatographic properties that are then washed to remove weakly bound proteins. A matrix is then applied to the array and allowed to crystallize so that the proteins become embedded in the array. The arrays are then subjected to surface enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOFMS) to determine the quantity and masses of the individual proteins on the array. Identification of interesting peaks from the SELDI readout can be accomplished by further purification of the sample followed by digestion with a protease and SELDI again. Masses of the proteolytic fragments can then be matched to a database for protein identification. This approach is intriguing as it obviates the need for the generation of protein-specific identification reagents. It remains to be seen whether the complexity of the system will foster widespread usage.

There are several technological hurdles involved with protein arrays. Evaporation and proper folding are typically problems with proteins on glass slides. In one study, proteins were printed in 40% glycerol in an attempt to circumvent this problem (MacBeath and Schreiber, 2000). To reduce the potential for evaporation with standard glass slides, Synder's group used a disposable silicone elastomer to create nanowells (Zhu

et al., 2000). They purified GST fusions of 119 kinases from yeast and deposited them into nanowells for detection of phosphorylation of a variety of substrates. This method also has the added benefits of being able to perform assays in solution and of avoiding cross contamination. A different approach to keeping proteins hydrated is to immobilize them in gel pads that are placed on a glass slide (Arenkov et al., 2000). However, the preparation of gel pads is currently difficult and therefore may not be the optimal method for addressing this problem.

Tissue microarrays. The tissue microarray (TMA) is another example of converting a cumbersome single sample assay and making it miniature and highly parallel. Up to 1,000 different tissue sections can be arrayed on a single glass slide and about 300 identical slides can be produced from a typical tissue specimen (Kononen et al., 1998). TMAs can be used for assays used with larger tissue samples such as fluorescence *in situ* hybridization or immunohistochemistry. Currently, most uses of TMAs have focused on analysis of tumor samples (Mousses et al., 2002).

There are a number of advantages to using TMAs. One is being able to assay control samples in parallel with the samples of interest. Also, comparison among various tissue samples is easier due to the fact that each sample is treated identically in the assay. Assays should be highly reproducible if tissue sections originate from the same sample (Kallioniemi et al., 2001). Consumption of assay reagents is also reduced relative to the number of different tissue samples that are assayed. However, TMAs are more suited to research rather than individual clinical diagnosis, as analysis of such small sections of a

tissue may not provide complete information from a heterogeneous tissue sample (Kallioniemi et al., 2001).

Cell microarrays. With the cell microarray, it is now possible to transfect thousands of different DNA samples into cultured cells in parallel. DNA samples are spotted onto a glass slide in an aqueous gelatin solution and a lipid transfection reagent is added to the spots. The slide is then placed in a dish and covered with culture media containing adherent cells. The resulting slide has groups of cells growing on it that have taken up the various spotted DNAs (Ziauddin and Sabatini, 2001).

Cell microarrays could be used as a means of determining drug targets. In initial studies, the drug FK506 was labeled and added to cells expressing GFP or FKBP12. Only cells expressing FKBP12, a known target for FK506, showed retention of the label (Ziauddin and Sabatini, 2001). Another use for cells microarrays could be to observe the effects of proteins on cellular physiology. cDNAs cloned into expression vectors would be individually spotted onto the array. For cells that exhibited the desired physiological change, the DNA corresponding to that group of cells could be sequenced and identified (Ziauddin and Sabatini, 2001). Similarly, plasmids expressing small interfering RNA duplexes could be spotted to determine the effects of repression of a particular gene (Brummelkamp et al., 2002; Mousses et al., 2002; Yu et al., 2002).

There are a number of benefits to using cell microarrays. For expression cloning, it is not necessary to isolate the cDNA of interest from the cells, as the cDNA is already present in the plasmid that was spotted. Transient phenotypes can also be detected as the arrays use live cells growing on them. Additionally, it may be possible to represent the

entire human genome on several slides, allowing for use of minimal quantities of drugs or staining reagents and the automation of morphological analysis with scanners that are already equipped to read glass microscope slides (Ziauddin and Sabatini, 2001).

Unfortunately, many of the uses of cell microarrays require full-length genes cloned into expression vectors. However, these may be soon available as the Harvard Institute for Proteomics completes its full length expression (FLEX) project (Brizuela et al., 2001).

Small molecule microarrays. The creation of small molecule microarrays allows for the simultaneous screening of many compounds with a protein or other biomolecule. The basic approach involves using a robot to spot small molecules to a derivatized glass slide (MacBeath et al., 1999). The small molecules share a common functional group that allows for covalent attachment to the slide. This approach was taken to isolate a small molecule, uretupamine, that binds to the yeast protein, Ure2p (Kuruvilla et al., 2002). The molecule appears to be specific for Ure2p as it had no effect on a strain harboring a URE2 deletion, as determined by genome-wide expression profiling (Kuruvilla et al., 2002). Small molecule arrays, if they are able to be produced cheaply and stored for a reasonable period of time, may provide the means for easily isolating compounds in a high throughput manner that bind specifically to individual proteins.

Carbohydrate microarrays. This technology is still in its infancy but a new study reports that dextrans can be immobilized on unmodified nitrocellulose coated slides and detected with antibodies (Wang et al., 2002). Further studies will need to be performed to determine whether this approach will work for the wide variety of potentially

interesting carbohydrates. As different carbohydrate structures are associated with different microbes, these arrays may be useful in the future for determining whether an individual has been exposed to a particular pathogen.

Data Analysis

The discussion of data analysis here will focus primarily on genome-wide expression analysis, although some of the principles can be applied to genome-wide location analysis and other microarray-based techniques. The first step in data analysis after hybridizing the sample to the array is data acquisition. A scanning laser produces an image of the array so that the amount of fluorescently labeled material on the array can be quantified. A variety of software packages are capable of finding features or spots on the array with minimal human intervention and associating spots with probe names as provided by the manufacturer of the array. Intensity values may be adjusted to account for background signal.

The researcher now has values for each gene on the array. If two different fluorophores were used on the same array, or if two different arrays are to be compared to each other, the probe intensities must be scaled or normalized to each other. A popular method for normalization is using the median or trimmed mean intensity of each array as a scaling factor. (The term 'array' may be used here to refer to each scan of an array using a particular wavelength.) The assumption made with this method is that the overall distribution of transcript levels does not change significantly among the arrays. An alternative method is to use a subset of the probes on the array to derive a scaling factor.

These may be housekeeping genes that are believed not to change expression among the arrays. Alternatively, they may be genes that are not actually present in the organism being studied but equal amounts of transcripts representing these genes are spiked in prior to hybridization. Naturally, there must be probes on the array for these spiked in control genes.

Most researchers are interested in determining which genes have altered their expression significantly in different conditions or among different samples. Taking the ratio of the expression values describes in a single value the change of a gene's expression relative to a reference. It is often useful to take the logarithm of this ratio so that arithmetic and statistics with these ratios become more straightforward (Kalocsai and Shams, 2001). Unfortunately some methods of analysis can produce negative values for some genes. As ratios of negative numbers are meaningless in this context and it is not possible to take the logarithm of a negative number, negatively valued data must be thresholded to a positive number. The choice of this positive number will vary by the platform and is often determined empirically.

Many studies report genes that have changed by some threshold. A two-fold change in a gene is often used, somewhat arbitrarily, as the amount of change required that is worth reporting. An alternative to using ratios is to standardize the expression level of a gene across many experiments to have a mean of zero and a standard deviation of one. Genes that change in expression by some number of standard deviations may then be reported as having significantly changed. There is a trend away from reporting significance based on fold-change alone, but also to incorporate absolute feature intensity into a statistic that delivers a confidence value for a significant change in expression.

Two groups have developed such error models that use feature intensities and consider both additive and multiplicative noise (Hughes et al., 2000; Ideker et al., 2000). In general, genes that have a lower intensity on the array require a higher ratio of change between two experiments to be called significant. Additionally, by reporting a confidence value for each gene, a p-value threshold does not have to be chosen but rather can be incorporated into models that make use of the data. Once these initial steps of analysis have been performed to determine values for individual genes, it may be of interest to look for trends or structure in the data. A selection of some of the more popular techniques in data analysis is discussed here.

Clustering techniques have been around for years, but have been popularized for use with microarray data by Michael Eisen's Cluster program (Eisen et al., 1998). The basic principle is to reveal structure within the data by creating subgroups of genes that are expressed similarly. The Cluster program performs hierarchical clustering, where each cluster is a component of a larger cluster thereby producing a tree of clusters. There is also non-hierarchical clustering, in which the number of clusters is fixed and there is no tree-like structure to the clusters.

Within the general approach of clustering, there are two methods of creating clusters. Clusters can be created by iteratively combining genes into groups in what is referred to as agglomerative clustering. Alternatively, there is divisive clustering which involves repeatedly dividing the set of genes into subgroups. Divisive clustering was used in a study of colon tissues to separate samples that were normal and cancerous

(Alon et al., 1999). However, agglomerative clustering is more popular in the microarray community, most likely due to Eisen's program.

Usually gene expression ratios are used for clustering. If the amplitude of the changes in gene expression are less important than the overall shape of curve of expression for a given gene, the expression ratios can be transformed such that their mean is zero and their standard deviation is set to one. Alternatively, for each gene, the minimal and maximal changes in expression can be set to negative and positive one, respectively, with the other values being scaled appropriately.

Hierarchical clustering. The first step in hierarchical clustering requires determining the distance between every pair of genes. Distance can be measured by a number of metrics, among which Euclidean distance and the Pearson correlation coefficient are popular. The two genes that are closest together are joined to form a cluster. The matrix of distances only has to be updated now with the distance between the new cluster and all other genes, as all other distances are unchanged. There are several methods for deciding how clusters are linked, or what point in the cluster is used as a reference for determining its distance from other genes or clusters (Anderberg, 1973). A common choice is to use average-linkage clustering where the average value of the cluster, or centroid, is used in calculating its distance to another cluster (Quackenbush, 2001). There is also single-linkage and complete-linkage clustering, which use the minimum and maximum distances between members of two clusters, respectively. The process of fusing clusters continues until there is a single cluster. The structure of the component clusters is preserved and it is left to the investigator to decide

which level of branching to use to report individual clusters. Clustering is not limited to one dimension; both the genes and the samples can be clustered so that samples that have the most similar profiles will be grouped together (Alon et al., 1999).

A potential problem with hierarchical clustering is that once an error is made in constructing a cluster, it cannot be undone and may contribute to the formation of additional clusters. Additionally, the vector that represents large clusters may not reflect the expression patterns of many of the genes in the cluster leading to the fusing of seemingly unrelated clusters (Quackenbush, 2001). Interpretation of the clustering diagram should be done carefully. Although there is no correct choice of linkage or distance metric, each method may produce different results, which vary in their biological relevance. There is also a tendency to interpret the cluster diagram linearly. However, as the diagram is a tree, each cluster can be flipped such that there is no change to the tree's structure, but the linear order of the leaves is altered. A program is now available that optimizes the orientations of the leaves to reveal the most structure in the data (Bar-Joseph et al., 2001).

k-means clustering. An alternative to hierarchical clustering is non-hierarchical clustering, where the number of clusters is fixed and determined in advance. A popular non-hierarchical method is *k-means* clustering, where the data are partitioned into *k* clusters. The optimal number of clusters to choose may be determined through trial and error or through some knowledge of the structure of the data which may have been obtained through hierarchical clustering or principal component analysis (see below). The starting centroids, or seeds, of the clusters need to be determined. There are several

methods of choosing seeds that include randomly selecting k genes from the set of all genes to be clustered, creating random seeds, or devising a method for choosing centroids that spans the diversity in the dataset (Anderberg, 1973). Genes are then assigned to the cluster whose seed they are closest to. Closeness is defined by a distance metric such as those used for hierarchical clustering. Instead of taking the seed approach, all genes can initially be assigned to clusters based on additional knowledge from another method.

With the genes partitioned into starting clusters, each gene is then reexamined to determine which cluster's centroid it is closest to. If it is closer to another cluster than the one to which it is currently assigned, then it is moved, and the centroids of the clusters that lost and gained a member are updated. This process of moving genes continues for a predetermined number of iterations or until some degree of stabilization is reached.

Self-organizing maps (SOM). A self-organizing map, like k -means clustering, is a divisive clustering approach (Kohonen, 1995). The user of an SOM program, such as Genecluster, specifies a two-dimensional geometry of the clusters (e.g. 5 clusters wide by 6 clusters high) (Tamayo et al., 1999). Before genes are actually placed into clusters, the SOM "learns" reference vectors for each cluster. This process involves repeatedly sampling genes from the data and finding which reference vector for a cluster the sampled gene is closest to. The reference vector of the closest cluster is adjusted toward the sampled gene according to a function that is specified by the implementation of the SOM. The reference vectors of the clusters that are topographically nearby are also adjusted, but to a smaller extent. During this process, the distance that the vectors are moved decreases as the number of iterations increases. After a fixed number of iterations

of adjusting the reference vectors, the genes are assigned to the clusters they are closest to. The resulting SOM is an organized collection of clusters with neighboring clusters being similar to each other.

Principal component analysis (PCA). The data from multiple arrays, which each monitor the expression of many genes, has as many dimensions as arrays. For most collections of experiments, not all dimensions of the data may make a significant contribution toward its overall structure. Therefore, the effective dimensionality of the data is actually lower. PCA, also known as singular value decomposition, can be a useful technique for reducing the dimensionality of gene expression data to a point where patterns are more easily detectable (Alter et al., 2000; Holter et al., 2000). Through matrix factorization, it extracts the vectors that represent the most variance in the data. It is often useful to plot the first three principal components in three dimensions for visualization. Although it may not be clear how to divide the data into groups from visualization, PCA is often a useful technique when combined with non-hierarchical clustering approaches to guide how many clusters may most effectively separate the data.

Support vector machines (SVM). Support vector machines are a popular choice of a supervised learning method for microarray data analysis (Brown et al., 2000).

“Supervised learning” means that information not included in the data is applied when classifying the data. SVMs use a training set to learn how to classify samples or genes into two groups. Examples of both groups must be provided in the training set. Once the SVM has learned how to classify from the training set, it can be applied to the data set to

divide the data into two classes. It learns how to classify by transforming the data into a higher dimensional feature space. The SVM uses a kernel function that functions as a distance metric in this feature space to find the hyperplane that divides the samples into two groups (Quackenbush, 2001). A soft margin can be used which allows for some small number of samples to be incorrectly classified. Initial choices of the kernel function and soft margin may not provide the best results. Therefore it may be necessary to use more complex kernel functions until the proper classification is attained (Brown et al., 2000).

Display Methods

One of the many difficulties researchers face with using microarrays is how to communicate their results. In general, a table of gene names and numbers is cumbersome and may not be able to highlight the most interesting features of the data. Biologists are experienced with producing a figure depicting their data that contains only a few pieces of information. Typically, this is an autoradiogram or a simple graph. Readers are accustomed to seeing these figures and can readily understand them. However, with microarrays biologists now need to be able to relate the results of a single experiment that may have many hundreds or thousands of data points. False color images of microarray scans can provide the reader with an impression of what the microarray looks like but not more than a few spots can be identified before the image becomes too confusing. It is the challenge to the presenters of the data to devise graphical representations of their

microarray results that can be easily digested by the audience. Some popular visual displays are grouped into three categories and discussed below.

Biological-related representations. As the usual audience for biological microarray experiments are biologists themselves, a visual display that is familiar to them can communicate the salient points effectively. A variety of approaches have been developed and tried since microarray data has begun to litter the journals. Brown and colleagues demonstrated how a simple coloring of a well-known metabolic chart could highlight their conclusions effectively (De Risi et al., 1997). They were exploring the effects of the switch from fermentation to respiration upon gene expression in yeast. As genes involved in glycolysis and the citric acid cycle were changing their expression levels, the use of a metabolic chart depicting these processes provided an intuitive representation of the data.

If the results of an experiment yield genes that are involved in related cellular processes, they may be best represented by depicting a simplified version of a cell with only the relevant genes included. By doing this, the viewer is provided with the context for where these gene products are functioning, and how they are relevant to the experiment (e.g. Ogawa et al., 2000). If the genes of interest from a genome-wide experiment are physically near each other in the genome, drawing chromosomes and coloring the relevant genes can be effective. This type of display was used to communicate the effects of the loss of histones on gene expression (Wyrick et al., 1999). It is immediately obvious from seeing this display that genes at the telomeres are derepressed.

Abstractions. A very popular method of displaying genome-wide expression data is the output of Michael Eisen's TreeView program (Eisen et al., 1998). In this display, each gene under a particular condition is represented as a horizontal bar and is colored according to its change in expression – red typically indicates increased expression and green indicates decreased expression. This format is very convenient for displaying time course data as many horizontal bars can be drawn end to end, thereby creating a large rectangle of horizontal bars with localized patches of red and green. Others have used variations on this theme where either the colors are changed (e.g. to red and blue, which is more friendly to red-green color blind viewers) or the horizontal bars have been replaced with squares creating a grid-like display (e.g. Pomeroy et al., 2002).

The accumulation and analysis of genome-wide data is allowing researchers to describe biological networks or connections. Ball and stick models intuitively represent networks. Colored circles, or mixes of other shapes to communicate additional information, often denote genes or proteins. Connections between genes or proteins are represented by sticks or arrows to indicate regulation. The length of the connector may reflect the strength of an interaction or the confidence in it. Similarly its color can indicate the type of connection such as a protein-protein interaction or a protein-promoter interaction. These types of displays have been useful in depicting gene regulatory networks (Hartemink et al., 2001), protein-protein interaction maps (Boulton et al., 2002), or combinations of both (Ideker et al., 2001).

Sometimes it is desirable to view genes as entire sets rather than as a collection of individual genes. For this, Venn diagrams have proven to be very useful (e.g. Bernstein

et al., 2000; Lee et al., 2000). In their simplest form, there are two sets of genes, each represented by a differently colored circle that may be proportional in area to the number of genes in the set. If there are genes that are in both sets, they are denoted by having the circles overlap, often with the overlapping region being proportional to the number of genes in common. Venn diagrams can be used with three circles or ovals, but greater than three sets requires that more complex shapes be used if all possible combinations of the sets are to be represented. The strength of Venn diagrams is their simplicity, which is lost when too many sets are represented.

Graphs. Traditional graphs can be useful in displaying genome-wide data. When the expression changes of only a few genes are to be displayed, bar graphs may be sufficient (e.g. Gray et al., 1998). Pie charts easily represent fractions of genes that are involved in a particular function (e.g. Jelinsky and Samson, 1999). For time course experiments in particular, a standard line graph with a fold change in gene expression on the y-axis and time on the x-axis provides a simple depiction of a gene's expression profile. The profile of many genes may be overlaid on a single graph to emphasize the similar expression pattern of these genes (e.g. Primig et al., 2000). These graphs can be simplified by representing the mean expression profile of a cluster of genes, as is done in a program to create self-organizing maps from gene-expression data (Tamayo et al., 1999). Here, a panel of line graphs represents each cluster, with neighboring graphs being more similar to each other than distant graphs.

Scatter plots can conveniently represent data from genome-wide experiments. In the case of a single experiment, the axes correspond to the signal intensities of the two

conditions in the experiment (e.g. wild type vs. mutant, treated vs. untreated), or for two experiments, the axes correspond to signal intensity ratios (e.g. Hughes et al., 2000). A point is plotted for each gene; points along the diagonal represent those genes that behave similarly between the two conditions, with outliers above and below the diagonal being genes that differ significantly. If the majority of points are along a line that is off the diagonal, it may indicate an error in scaling by a factor that corresponds to the slope of that line.

A three-dimensional scatter plot, where dots are replaced by spheres for easier visibility, can be useful for displaying the results of principal component analysis. Each gene is represented as a sphere floating in three-dimensional space along arbitrary axes. Spheres representing genes that have similar profiles will be positioned together in groups facilitating the identification of clusters of genes. If the dimensions of the original data represent a time course, then these clusters may be genes that are functionally related or coordinately regulated (e.g. Raychaudhuri et al., 2000). If the original data is a collection of clinical samples, the genes may reflect the genotype of a group of samples (e.g. Bittner et al., 2000; Pomeroy et al., 2002).

Another three-dimensional graph that was used to visualize microarray data is a terrain map, which was used to highlight different functional groups of genes in *C. elegans* derived from over 500 microarray experiments (Kim et al., 2001). The x-y dimensions of the map reflect a transformation of correlation coefficients between each pair of genes in their data set and the z-axis denotes the density of genes in a particular region. The map is colored according to the height of the mountains; therefore no

additional information is conveyed by the color, but rather it serves to emphasize the height of individual peaks.

Biological Insights

Microarray technology has progressed such that, in most cases, studies have moved beyond proof of principle experiments and are advancing into discovering new biology. Laboratories have conducted numerous microarray experiments with yeast and mouse samples for many of the same reasons that these systems have been popular for use with other types of experiments: it is relatively easy to perform yeast genetics; yeast can easily be grown to provide adequate amounts of sample for analysis; and the genome is well annotated due to an enormous literature. Mice are close enough to humans evolutionarily to be a relevant system for modeling disease. The ability to perform genetic studies and create transgenic or knockout mice provides the tools for studying molecular mechanisms. As the goal for many researchers is to understand human biology and disease, human cell culture is naturally a very popular system for microarrays.

Nevertheless, microarrays have been created for expression analysis of most other model systems including *Escherichia coli* (Selinger et al., 2000), *Bacillus subtilis* (Ye et al., 2000), *Arabidopsis thaliana* (Ruan et al., 1998), *Caenorhabditis elegans* (Reinke et al., 2000), *Drosophila melanogaster* (White et al., 1999), *Xenopus laevis* (Altmann et al., 2001), zebrafish (Herwig et al., 2001), and rats (Chabas et al., 2001). Additionally, numerous pathogens such as *Candida albicans* (Murad et al., 2001), *Caulobacter crescentus* (Laub et al., 2000), *Helicobacter pylori* (Ang et al., 2001), *Hemophilus*

influenzae (de Saizieu et al., 1998), *Mycobacterium tuberculosis* (Wilson et al., 1999), *Plasmodium falciparum* (Hayward et al., 2000), *Pyrococcus furiosus* (Schut et al., 2001), *Streptococcus pneumoniae* (de Saizieu et al., 1998), *Streptomyces coelicolor* (Huang et al., 2001a), cytomegalovirus (Bresnahan and Shenk, 2000) and human herpes virus 8 (Paulose-Murphy et al., 2001) have been profiled, providing annotation of these genomes and exploration of pathogenic mechanisms (Cummings and Relman, 2000). Some examples of the uses for microarrays in biology and the types of biological discoveries that have been made are discussed below.

Cancer classification. Traditional methods of classifying cancer cell samples have focused on cell morphology, histochemical and cytogenetic analysis, and characterization with antibodies. The use of these assays relies on human experience to accurately classify a sample as a particular form of cancer. Additionally, different cancers may have nearly indistinguishable appearances with current assays, but may be distinct diseases at the molecular level. As different cancers respond better to certain treatments, properly classifying a tumor is crucial.

Using microarrays for expression analysis can reveal the molecular signature of a tumor. These signatures can then be used for assigning a sample to a particular molecular class of cancer. By collecting many signatures and correlating these samples with clinical outcome, physicians should eventually be able to prescribe treatments with the added knowledge of a molecular diagnosis. A study led by Golub analyzed acute leukemia samples and created a predictor to classify samples into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) (Golub et al., 1999). Additional

analysis was able to distinguish between ALL that was from a T-cell lineage versus a B-cell lineage. Follow-up studies revealed molecularly discernable subtypes within both T-cell and B-cell lineage ALL (Armstrong et al., 2002; Ferrando et al., 2002; Yeoh et al., 2002). Other classification studies have discovered melanoma (Bittner et al., 2000), B-cell lymphoma (Alizadeh et al., 2000; Shipp et al., 2002), and breast cancer subtypes (Perou et al., 2000; Sorlie et al., 2001; van't Veer et al., 2002), and have distinguished among types of brain tumors (Pomeroy et al., 2002) and normal and cancerous prostate tissue (Dhanasekaran et al., 2001; Singh et al., 2002; Welsh et al., 2001).

Diagnostics. Microarrays show promise for being diagnostic tools both in cancer and in infection. For cancer, the ability to analyze an entire genome instead of relying on a few select markers greatly enhances the ability to make an accurate assessment of the type of cancer the patient is presenting. In addition to the classification studies that were described above, two groups have published expression profile surveys of a wide variety of cancers (Ramaswamy et al., 2001; Su et al., 2001). These types of studies may provide the foundation for a database of cancer profiles that would be available to all for comparison and may guide the production of microarrays constructed specifically for cancer diagnostics.

In the case of infection, diagnosis may be difficult due to an inability to culture the pathogen. Microarrays that can identify foreign nucleic acid from a human sample would circumvent this problem. One approach being taken to address this is to make arrays that probe for the 16S rRNA and antibiotic resistance genes (Ye et al., 2001). As ribosomal RNA is very abundant, it may be possible to minimize amplification

procedures. In addition, regions of the 16S rRNA are highly polymorphic among species and therefore should provide a unique fingerprint for most species (Woese, 1987). Initial studies with *Bacillus* and *Mycobacterium* show promise for this approach (Liu et al., 2001; Troesch et al., 1999). With the realistic threat of biological warfare in the 21st century, it is critical to develop new assays to detect a variety of pathogens.

An alternative approach is to perform expression analysis on human cells that are infected by an unknown pathogen. A potential advantage of this method is that transient or previously undetectable exposures to a pathogen could be detected (Cummings and Relman, 2000). Several studies have been performed examining the host response to individual pathogens by expression analysis (references in Diehn and Relman, 2001). By aggregating these expression signatures into a common resource, it may be possible to create an index of profiles that can be used for diagnosing an infection. It has also been demonstrated that the expression profiles of human cells infected with bacteria exhibit a common gene expression signature (Boldrick et al., 2002; Huang et al., 2001b; Nau et al., 2002). It would be useful to be able to determine whether a patient's symptoms were due to a bacterial infection or were caused by something else.

Drug target discovery. Microarray data has the potential to be a tool to discover the target of an existing drug. Ideally, the drug inactivates the function of a single protein. The molecular signature of the effects of the drug is determined by expression profiling. That profile is then compared to profiles of mutants that have a single gene inactivated by genetic methods. Presumably, if the drug knocks out the function of a single protein, then the expression profiles of mutant cells and drug-treated cells are very

similar. This methodology was proposed and successfully tested on known drugs with known targets (Marton et al., 1998). The same group profiled yeast exposed to the drug dyclonine and found that the drug-treated expression profile compared well to the expression profile of a strain carrying a deletion of the *erg2* gene (Hughes et al., 2000). Additional experiments revealed that *erg2* deletion mutants and dyclonine-treated cells exhibited buildup of the same chemical intermediates and that overexpression of *ERG2* resulted in increased resistance to dyclonine (Hughes et al., 2000). This application of microarrays may prove useful to pharmaceutical companies, but requires large databases of existing expression profiles to actually be a viable tool in the discovery of drug targets.

Gene networks. For systems with few components such as bacteriophage lambda, assembling a genetic regulatory circuit appears to be a manageable problem (McAdams and Shapiro, 1995). However, building a network for even a single-celled eukaryote will be tremendously more difficult. Nevertheless, with the sequencing of genomes and availability of microarray technology, the initial steps are ready to be attempted. The highly parallel nature of microarrays makes them ideal for studying gene networks. For each microarray experiment, large quantities of quantitative data are produced, thereby providing foundations for models. Perturbations to the system can be applied and monitored with subsequent experiments.

Two groups have used the galactose metabolism system to test strategies for network assembly. Ideker and colleagues performed genome-wide expression analysis with yeast strains that harbored a deletion in any one of nine components of the galactose system and were grown in the presence or absence of galactose (Ideker et al., 2001).

Protein quantitation data from yeast growing with or without galactose and protein-protein interaction data from the literature were integrated with the genome-wide expression data to create a protein-interaction network. From their assembled network they proposed refinements to the existing model of galactose metabolism regulation and tested some of their hypotheses. Gifford and coworkers took advantage of over 50 existing genome-wide expression data sets from yeast to explore models of galactose regulation (Hartemink et al., 2001). They used Bayesian networks to differentiate among simple regulatory models and were able to recapitulate the currently accepted model. These types of approaches should both prove useful in analyzing gene networks as more sophisticated hypotheses are developed.

The cell cycle is a well-studied circular network that has been subjected to genome-wide expression analysis (Cho et al., 1998; Spellman et al., 1998). A recent study used genome-wide location profiling to find direct targets of nine transcriptional activators (Simon et al., 2001). They were able to identify genes that are likely regulated by each of the activators in different stages of the cell cycle. Importantly, for each activator, they found that it was regulated by another transcriptional activator earlier in the cell cycle. Additionally, they identified other regulators, such as cyclins, as transcriptional targets of the activators. As the number of genome-wide location profiles of transcriptional activators under various environmental conditions grows, the pieces of a genome control map may fall into place.

Developmental programs. Several groups have used microarrays to gain a better understanding of development programs in multicellular organisms. Microarray studies

are well suited to this task as transcriptional cascades effect many developmental events (Davidson et al., 2002). The strategy used currently for most expression profiling studies involves extracting RNA from whole organisms at various developmental stages. An obvious caveat with this procedure is that a mixture of cell types is being sampled in every experiment. Therefore, information is lost about the cell types in which observable changes in gene expression are occurring and gene expression changes in a single cell type that are drowned out by the presence of other cells.

Two major microarray-based studies of *Drosophila* have explored gene expression changes during metamorphosis and mesoderm development (Furlong et al., 2001; White et al., 1999). Development in *C. elegans* has also been examined by profiling oocytes, six stages of development from eggs to adults, and aged worms (Hill et al., 2000). In each case, genes were identified that exhibit developmental stage-specific expression and coregulation at all time points tested. Additional work, perhaps using genome-wide location analysis, will be required to determine what are the direct effectors of these developmental programs.

Gene regulation. Performing clustering analysis with microarray expression data can reveal genes that are coexpressed. Genes that are coexpressed are often coregulated. Although there is gene regulation at the level of mRNA stability, it is believed that much of gene regulation occurs at the transcriptional level. The cis-acting promoter sequences upstream of transcriptional start sites are recognized by transcription factors that serve to regulate gene expression. Several groups have developed computational tools to identify

these cis-acting sequences from sequence data (Brazma et al., 1998; Roth et al., 1998; van Helden et al., 1998; Wolfsberg et al., 1999).

As yeast has a very well-annotated genome and many genome-wide expression datasets are available, it is often used to demonstrate the power of these techniques. A study from the Church lab clustered yeast cell cycle data and identified sequence motifs that were enriched in each cluster (Tavazoie et al., 1999). Many of the motifs agreed with published observations and new motifs were identified that appear to be biologically significant. Another study by the same group used multiple genome-wide expression datasets to examine the contributions of pairs of sequence motifs toward coordinate gene regulation (Pilpel et al., 2001). They found new synergistic motif combinations and observed regulatory connections between biological processes.

As with many types of *in silico* studies, wet experiments are necessary for validation. In the studies mentioned here, genome-wide expression data was the principal source of input data. These types of studies should become more fruitful as more data sources (e.g. genome-wide location analysis, protein-protein interactions) are fused to create a higher quality input dataset.

Gene function. A hint of a gene's function can often be inferred from sequence homology. However, a more detailed description requires experimentation. In a study to identify molecular determinants of metastasis, an *in vivo* selection was performed to isolate highly metastatic melanoma cells (Clark et al., 2000). Expression profiling with these cells revealed several genes that correlated with the metastatic phenotype, which included the RhoC gene. Metastasis was enhanced by overexpression of RhoC and

inhibited by a RhoC dominant-negative mutant, prescribing a function to RhoC in tumor invasion.

Alternatively, multiple profiling experiments can be performed under different conditions. The function of the gene of interest is suggested by the genes with which it is coexpressed. In a study of over 300 genome-wide expression profiles of yeast, a cluster of genes was found that was induced in ergosterol-related experiments (Hughes et al., 2000). In addition to known ergosterol biosynthetic genes, there were genes of unknown function in that cluster. Further experimentation with one of these genes, *ERG28*, demonstrated that *erg28* mutants produce less ergosterol than wild-type cells and accumulate apparent ergosterol synthesis intermediates. Additionally, expression profiles of *erg28* mutants clustered together with other ergosterol mutants, which is consistent with a role for Erg28 in ergosterol biosynthesis.

Genome analysis. Although most microarray studies have focused on measuring cellular levels of RNAs, microarrays have also been used to measure DNA content. As with an array for expression analysis, all DNA loci can be probed at once. For example, several studies have determined the extent of genome amplification and deletion in several cancer cell lines using either cDNA- or BAC-based array comparative genomic hybridization (CGH) (Pollack et al., 1999; Snijders et al., 2001).

A number of groups have analyzed genome content across several microbial species. As the live vaccine for tuberculosis, BCG, is an attenuated form of *M. bovis*, one group determined the genomic differences among various BCG strains, *M. bovis* and *M. tuberculosis* (Behr et al., 1999). Another group printed microarrays reflecting the two

sequenced strains of *H. pylori* and examined genomic differences among 15 strains to better understand differences in virulence and to define a minimal set of genes required (Salama et al., 2000). Similar studies have been performed with *S. aureus* and *V. cholerae* (Dziejman et al., 2002; Fitzgerald et al., 2001).

Conclusion

There are still many challenges facing researchers who wish to use microarrays. The microarray is a developing technology that reaches a relatively narrow audience. More companies are beginning to sell DNA microarrays but prices are still high. Many academic laboratories are printing their own arrays, but even more lack the resources to do so, or do not have access to a core facility to provide the service for them.

The availability of useful protein identity arrays has been widely anticipated but has yet to materialize. Reagent based detection systems (e.g. antibodies that recognize every protein) may prove too costly and laborious to develop for an entire proteome and may not result in the specificity required. Alternative splicing and post-translational modifications could easily double the number of unique protein species, making the task of protein detection that much more challenging. This approach may be feasible for measuring a select group of proteins, however. Quantitative mass spectrometry-based methods hold promise for overcoming these problems, but may require equipment and expertise that is not readily available.

Although the human genome has been almost completely sequenced (Lander et al., 2001; Venter et al., 2001), there is still a debate regarding the identity of all of the

genes. In fact, by tiling probes across human chromosomes 21 and 22, as many as ten-fold more transcripts may be present than initially expected (Kapranov et al., 2002). As more genomes are sequenced, comparative genomics should help identify coding regions to assist current gene prediction methods. With a better knowledge of human open reading frames, better probes for expression analysis will be able to be designed. As human promoters are generally poorly defined, comparative genomics will help to identify conserved sequence elements (Wasserman et al., 2000). Knowing the boundaries of promoters will improve efforts to select relevant probes for location analysis in humans.

There is promise for the use of microarrays in diagnostics. However, the technical expertise required to perform microarray experiments will have to be minimized if it is to reach the greatest number of patients. Approaches using peripheral blood to perform expression profiling on blood cells or using microarrays to detect various pathogens in the blood stream are feasible, but require amplification of the sample RNA. Small biopsies generally suffer from the same problem. Several groups have devised systems for amplification of RNA to attempt to overcome this obstacle (Baugh et al., 2001; Luo et al., 1999; Pabon et al., 2001; Wang et al., 2000). As the readout from a microarray experiment is the sum of effects from an entire population of cells, eventually it would be ideal to be able to perform genome-wide expression analysis on single cells.

For every microarray experiment that is performed, thousands of pieces of data are produced. Often the investigator is only interested in a subset of the data, or choosing to analyze the data to ask a certain question. By making microarray data easily

accessible, others can analyze and query available datasets in new ways. Several groups have established data repositories for microarray data (Gershon, 2002; Sherlock et al., 2001). Additionally, a central location for data would make it easier for investigators to search for datasets of interest. A challenge associated with creating databases for microarray data is what information should be stored. The Microarray Gene Expression Data (MGED) group has proposed a minimum set of information, called MIAME, that must be provided about every microarray experiment that is published so that any researcher who uses the data has enough information to do their own analysis (Brazma et al., 2001). MGED and others are hoping that journals will demand that papers publishing microarray experiments provide data in MIAME format to ensure that others are able to use the data for their own analyses (Knight, 2002).

With all of this data being produced, larger questions previously unanswerable are beginning to be addressed. Microarray-based expression profiling reports the message levels of thousands of genes at once. Location analysis reveals DNA-protein interactions across the genome. Combine these data with numerous genome sequences and studies of protein-protein interactions through two-hybrid analysis or mass spectrometry of purified protein complexes, and teams of biologists and computer scientists will be able to start modeling the molecular state of the cell. These *in silico* models should guide investigators to the next experiment to perform, further accelerating the pace of discovery.

In this work, I will describe several biological studies using microarrays as well as the creation of graphical tools for use with microarray data. In chapter 2, I will present a study using genome-wide expression profiling to analyze components of the

transcriptional machinery in yeast. This was one of the first studies to use microarrays to profile transcription factor mutants and was published in 1998 (Holstege et al., 1998). Chapter 3 is divided into two parts. The first part reports the invention of the genome-wide location profiling technique and its use with a well-studied transcriptional activator, Gal4. This work was published as a part of Ren et al. (2000). The second part describes a study where genome-wide expression and location profiling were used to produce a model for a transcriptional regulatory network of thiamin biosynthetic gene regulation. This study provides one piece of the cellular transcriptional regulatory circuitry and may be combined with other studies to build a complete regulatory map. The last chapter describes a set of computer programs for creating graphical displays that communicate the results of genome-wide expression and location experiments.

References

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* *403*, 503-511.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci U S A* *96*, 6745-6750.
- Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* *97*, 10101-10106.
- Altmann, C. R., Bell, E., Sczyrba, A., Pun, J., Bekiranov, S., Gaasterland, T., and Brivanlou, A. H. (2001). Microarray-based analysis of early development in *Xenopus laevis*. *Dev Biol* *236*, 64-75.
- Anderberg, M. R. (1973). *Cluster Analysis for Applications* (New York and London, Academic Press).
- Ang, S., Lee, C. Z., Peck, K., Sindici, M., Matrubutham, U., Gleeson, M. A., and Wang, J. T. (2001). Acid-induced gene expression in *Helicobacter pylori*: study in genomic scale by microarray. *Infect Immun* *69*, 1679-1686.
- Arenkov, P., Kukhtin, A., Gemmell, A., Voloshchuk, S., Chupeeva, V., and Mirzabekov, A. (2000). Protein microchips: use for immunoassay and enzymatic reactions. *Anal Biochem* *278*, 123-131.
- Armstrong, S. A., Staunton, J. E., Silverman, L. B., Pieters, R., den Boer, M. L., Minden, M. D., Sallan, S. E., Lander, E. S., Golub, T. R., and Korsmeyer, S. J. (2002). MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* *30*, 41-47.
- Bar-Joseph, Z., Gifford, D. K., and Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* *17*, S22-29.
- Baugh, L. R., Hill, A. A., Brown, E. L., and Hunter, C. P. (2001). Quantitative analysis of mRNA amplification by in vitro transcription. *Nucleic Acids Res* *29*, E29.
- Behr, M. A., Wilson, M. A., Gill, W. P., Salamon, H., Schoolnik, G. K., Rane, S., and Small, P. M. (1999). Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* *284*, 1520-1523.

- Belov, L., de la Vega, O., dos Remedios, C. G., Mulligan, S. P., and Christopherson, R. I. (2001). Immunophenotyping of leukemias using a cluster of differentiation antibody microarray. *Cancer Res* 61, 4483-4489.
- Bernstein, B. E., Tong, J. K., and Schreiber, S. L. (2000). Genomewide studies of histone deacetylase function in yeast. *Proc Natl Acad Sci U S A* 97, 13708-13713.
- Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A., *et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* 406, 536-540.
- Blanchard, A. P., Kaiser, R. J., and Hood, L. E. (1996). High-density oligonucleotide arrays. *Biosens Bioelectron* 11, 687-690.
- Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O., and Relman, D. A. (2002). Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proc Natl Acad Sci U S A* 99, 972-977.
- Boulton, S. J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D. E., and Vidal, M. (2002). Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* 295, 127-131.
- Braun, A., Little, D. P., and Koster, H. (1997). Detecting CFTR gene mutations by using primer oligo base extension and mass spectrometry. *Clin Chem* 43, 1151-1158.
- 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.
- Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., *et al.* (2001). Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet* 29, 365-371.
- Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* 8, 1202-1215.
- Bresnahan, W. A., and Shenk, T. (2000). A subset of viral transcripts packaged within human cytomegalovirus particles. *Science* 288, 2373-2376.
- Brizuela, L., Braun, P., and LaBaer, J. (2001). FLEXGene repository: from sequenced genomes to gene repositories for high-throughput functional biology and proteomics. *Mol Biochem Parasitol* 118, 155-165.
- Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr., and Haussler, D. (2000). Knowledge-based analysis of microarray gene

expression data by using support vector machines. *Proc Natl Acad Sci U S A* 97, 262-267.

Brummelkamp, T. R., Bernards, R., and Agami, R. (2002). A System for Stable Expression of Short Interfering RNAs in Mammalian Cells. *Science* 296, 550-553.

Chabas, D., Baranzini, S. E., Mitchell, D., Bernard, C. C., Rittling, S. R., Denhardt, D. T., Sobel, R. A., Lock, C., Karpuj, M., Pedotti, R., *et al.* (2001). The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* 294, 1731-1735.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. *Science* 274, 610-614.

Chen, J., Lee, S., Zhou, G., and Wang, S. M. (2002). High-throughput GLGI procedure for converting a large number of serial analysis of gene expression tag sequences into 3' complementary DNAs. *Genes Chromosomes Cancer* 33, 252-261.

Cho, R. J., Campbell, M. J., Winzler, 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.

Clark, E. A., Golub, T. R., Lander, E. S., and Hynes, R. O. (2000). Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* 406, 532-535.

Cummings, C. A., and Relman, D. A. (2000). Using DNA microarrays to study host-microbe interactions. *Emerg Infect Dis* 6, 513-525.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., *et al.* (2002). A genomic regulatory network for development. *Science* 295, 1669-1678.

De Risi, 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.

de Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W., and Mous, J. (1998). Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nat Biotechnol* 16, 45-48.

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.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., and Chinnaiyan, A. M. (2001). Delineation of prognostic biomarkers in prostate cancer. *Nature* 412, 822-826.

Diehn, M., and Relman, D. A. (2001). Comparing functional genomic datasets: lessons from DNA microarray analyses of host-pathogen interactions. *Curr Opin Microbiol* 4, 95-101.

Dziejman, M., Balon, E., Boyd, D., Fraser, C. M., Heidelberg, J. F., and Mekalanos, J. J. (2002). Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc Natl Acad Sci U S A* 99, 1556-1561.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868.

Ferrando, A. A., Neubergh, D. S., Staunton, J., Loh, M. L., Huard, C., Raimondi, S. C., Behm, F. G., Pui, C.-H., Downing, J. R., Gilliland, D. G., *et al.* (2002). Gene expression signatures define novel oncogenic pathways in T cell acute lymphoblastic leukemia. *Cancer Cell* 1, 75-87.

Fey, S. J., and Larsen, P. M. (2001). 2D or not 2D. Two-dimensional gel electrophoresis. *Curr Opin Chem Biol* 5, 26-33.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-246.

Fitzgerald, J. R., Sturdevant, D. E., Mackie, S. M., Gill, S. R., and Musser, J. M. (2001). Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc Natl Acad Sci U S A* 98, 8821-8826.

Flajolet, M., Rotondo, G., Daviet, L., Bergametti, F., Inchauspe, G., Tiollais, P., Transy, C., and Legrain, P. (2000). A genomic approach of the hepatitis C virus generates a protein interaction map. *Gene* 242, 369-379.

Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T., and Solas, D. (1991). Light-directed, spatially addressable parallel chemical synthesis. *Science* 251, 767-773.

Furlong, E. E., Andersen, E. C., Null, B., White, K. P., and Scott, M. P. (2001). Patterns of gene expression during *Drosophila* mesoderm development. *Science* 293, 1629-1633.

Gershon, D. (2002). Microarray technology: an array of opportunities. *Nature* 416, 885-891.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

Gray, N. S., Wodicka, L., Thunnissen, A. M., Norman, T. C., Kwon, S., Espinoza, F. H., Morgan, D. O., Barnes, G., LeClerc, S., Meijer, L., *et al.* (1998). Exploiting chemical

libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533-538.

Haab, B. B., Dunham, M. J., and Brown, P. O. (2001). Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol* 2.

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., *et al.* (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* 22, 164-167.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac Symp Biocomput*, 422-433.

Hayward, R. E., Derisi, J. L., Alfadhli, S., Kaslow, D. C., Brown, P. O., and Rathod, P. K. (2000). Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol Microbiol* 35, 6-14.

Herwig, R., Aanstad, P., Clark, M., and Lehrach, H. (2001). Statistical evaluation of differential expression on cDNA nylon arrays with replicated experiments. *Nucleic Acids Res* 29, E117.

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. (2000). Genomic analysis of gene expression in *C. elegans*. *Science* 290, 809-812.

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.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., and Fedoroff, N. V. (2000). Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci U S A* 97, 8409-8414.

Huang, J., Lih, C. J., Pan, K. H., and Cohen, S. N. (2001a). Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. *Genes Dev* 15, 3183-3192.

Huang, Q., Liu, D., Majewski, P., Schulte, L. C., Korn, J. M., Young, R. A., Lander, E. S., and Hacohen, N. (2001b). The plasticity of dendritic cell responses to pathogens and their components. *Science* 294, 870-875.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., *et al.* (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat Biotechnol* 19, 342-347.

- 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.
- Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929-934.
- Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. E. (2000). Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol* 7, 805-817.
- Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98, 4569-4574.
- Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc Natl Acad Sci U S A* 97, 1143-1147.
- 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.
- Jurinke, C., van den Boom, D., Cantor, C. R., and Koster, H. (2001). Automated genotyping using the DNA MassArray technology. *Methods Mol Biol* 170, 103-116.
- Kallioniemi, O. P., Wagner, U., Kononen, J., and Sauter, G. (2001). Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum Mol Genet* 10, 657-662.
- Kalocsai, P., and Shams, S. (2001). Use of Bioinformatics in Arrays. In *DNA Arrays: Methods and Protocols*, J. B. Rampal, ed. (Totowa, N.J., Humana Press, Inc.), pp. 223-236.
- Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P., and Gingeras, T. R. (2002). Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296, 916-919.
- Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., and Davidson, G. S. (2001). A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087-2092.

- Knight, J. (2002). Minimum standards set out for gene-expression data. *Nature* 415, 946.
- Kodadek, T. (2001). Protein microarrays: prospects and problems. *Chem Biol* 8, 105-115.
- Kohonen, T. (1995). *Self-organizing Maps*, second edn (Berlin, Heidelberg, New York, Springer-Verlag).
- Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat Med* 4, 844-847.
- Kozal, M. J., Shah, N., Shen, N., Yang, R., Fucini, R., Merigan, T. C., Richman, D. D., Morris, D., Hubbell, E., Chee, M., and Gingeras, T. R. (1996). Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nat Med* 2, 753-759.
- Kuruvilla, F. G., Shamji, A. F., Sternson, S. M., Hergenrother, P. J., and Schreiber, S. L. (2002). Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature* 416, 653-657.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.
- Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M., and Shapiro, L. (2000). Global analysis of the genetic network controlling a bacterial cell cycle. *Science* 290, 2144-2148.
- Lee, S., Clark, T., Chen, J., Zhou, G., Scott, L. R., Rowley, J. D., and Wang, S. M. (2002). Correct identification of genes from serial analysis of gene expression tag sequences. *Genomics* 79, 598-602.
- Lee, T. I., Causton, H. C., Holstege, F. C., Shen, W. C., Hannett, N., Jennings, E. G., Winston, F., Green, M. R., and Young, R. A. (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature* 405, 701-704.
- 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.
- Little, D. P., Braun, A., O'Donnell, M. J., and Koster, H. (1997). Mass spectrometry from miniaturized arrays for full comparative DNA analysis. *Nat Med* 3, 1413-1416.
- Liu, W. T., Mirzabekov, A. D., and Stahl, D. A. (2001). Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ Microbiol* 3, 619-629.

- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* *14*, 1675-1680.
- Lockhart, D. J., and Winzeler, E. A. (2000). Genomics, gene expression and DNA arrays. *Nature* *405*, 827-836.
- Luo, L., Salunga, R. C., Guo, H., Bittner, A., Joy, K. C., Galindo, J. E., Xiao, H., Rogers, K. E., Wan, J. S., Jackson, M. R., and Erlander, M. G. (1999). Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nat Med* *5*, 117-122.
- MacBeath, G., Koehler, A. N., and Schreiber, S. L. (1999). Printing small molecules as microarrays and detecting protein-ligand interactions en masse. *J Am Chem Soc* *121*, 7967-7968.
- MacBeath, G., and Schreiber, S. L. (2000). Printing proteins as microarrays for high-throughput function determination. *Science* *289*, 1760-1763.
- Marton, M. J., DeRisi, J. L., Bennett, H. A., Iyer, V. R., Meyer, M. R., Roberts, C. J., Stoughton, R., Burchard, J., Slade, D., Dai, H., *et al.* (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med* *4*, 1293-1301.
- McAdams, H. H., and Shapiro, L. (1995). Circuit simulation of genetic networks. *Science* *269*, 650-656.
- Mousses, S., Kallioniemi, A., Kauraniemi, P., Elkahloun, A., and Kallioniemi, O. P. (2002). Clinical and functional target validation using tissue and cell microarrays. *Curr Opin Chem Biol* *6*, 97-101.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harb Symp Quant Biol* *51*, 263-273.
- Murad, A. M., d'Enfert, C., Gaillardin, C., Tournu, H., Tekaia, F., Talibi, D., Marechal, D., Marchais, V., Cottin, J., and Brown, A. J. (2001). Transcript profiling in *Candida albicans* reveals new cellular functions for the transcriptional repressors CaTup1, CaMig1 and CaNrg1. *Mol Microbiol* *42*, 981-993.
- Nau, G. J., Richmond, J. F., Schlesinger, A., Jennings, E. G., Lander, E. S., and Young, R. A. (2002). Human macrophage activation programs induced by bacterial pathogens. *Proc Natl Acad Sci U S A* *99*, 1503-1508.
- Ogawa, N., DeRisi, J., and Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol Biol Cell* *11*, 4309-4321.

- Orlando, V., and Paro, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. *Cell* 75, 1187-1198.
- Pabon, C., Modrusan, Z., Ruvolo, M. V., Coleman, I. M., Daniel, S., Yue, H., and Arnold, L. J., Jr. (2001). Optimized T7 amplification system for microarray analysis. *Biotechniques* 31, 874-879.
- Paulose-Murphy, M., Ha, N. K., Xiang, C., Chen, Y., Gillim, L., Yarchoan, R., Meltzer, P., Bittner, M., Trent, J., and Zeichner, S. (2001). Transcription program of human herpesvirus 8 (kaposi's sarcoma- associated herpesvirus). *J Virol* 75, 4843-4853.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., *et al.* (2000). Molecular portraits of human breast tumours. *Nature* 406, 747-752.
- Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet* 29, 153-159.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23, 41-46.
- Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., *et al.* (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436-442.
- Primig, M., Williams, R. M., Winzeler, E. A., Tevzadze, G. G., Conway, A. R., Hwang, S. Y., Davis, R. W., and Esposito, R. E. (2000). The core meiotic transcriptome in budding yeasts. *Nat Genet* 26, 415-423.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet* 2, 418-427.
- Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., *et al.* (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A* 98, 15149-15154.
- Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac Symp Biocomput*, 455-466.
- Reinke, V., Smith, H. E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S. J., Davis, E. B., Scherer, S., Ward, S., and Kim, S. K. (2000). A global profile of germline gene expression in *C. elegans*. *Mol Cell* 6, 605-616.

- Ren, B., Cam, H., Takahashi, Y., Volkert, T., Terragni, J., Young, R. A., and Dynlacht, B. D. (2002). E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* 16, 245-256.
- 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.
- Ruan, Y., Gilmore, J., and Conner, T. (1998). Towards Arabidopsis genome analysis: monitoring expression profiles of 1400 genes using cDNA microarrays. *Plant J* 15, 821-833.
- Saiki, R. K., Walsh, P. S., Levenson, C. H., and Erlich, H. A. (1989). Genetic analysis of amplified DNA with immobilized sequence-specific oligonucleotide probes. *Proc Natl Acad Sci U S A* 86, 6230-6234.
- Salama, N., Guillemin, K., McDaniel, T. K., Sherlock, G., Tompkins, L., and Falkow, S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc Natl Acad Sci U S A* 97, 14668-14673.
- Schut, G. J., Zhou, J., and Adams, M. W. (2001). DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: evidence for a new type of sulfur-reducing enzyme complex. *J Bacteriol* 183, 7027-7036.
- Selinger, D. W., Cheung, K. J., Mei, R., Johansson, E. M., Richmond, C. S., Blattner, F. R., Lockhart, D. J., and Church, G. M. (2000). RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat Biotechnol* 18, 1262-1268.
- Serebriiskii, I. G., Khazak, V., and Golemis, E. A. (2001). Redefinition of the yeast two-hybrid system in dialogue with changing priorities in biological research. *Biotechniques* 30, 634-636, 638, 640 passim.
- Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C., Dwight, S. S., Kaloper, M., Weng, S., Jin, H., Ball, C. A., *et al.* (2001). The Stanford Microarray Database. *Nucleic Acids Res* 29, 152-155.
- Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., *et al.* (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med* 8, 68-74.
- 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.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., *et al.* (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* *1*, 203-209.

Singh-Gasson, S., Green, R. D., Yue, Y., Nelson, C., Blattner, F., Sussman, M. R., and Cerrina, F. (1999). Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol* *17*, 974-978.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., Kimura, K., *et al.* (2001). Assembly of microarrays for genome-wide measurement of DNA copy number. *Nat Genet* *29*, 263-264.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., *et al.* (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* *98*, 10869-10874.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* *98*, 503-517.

Southern, E. M., Maskos, U., and Elder, J. K. (1992). Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* *13*, 1008-1017.

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.

Su, A. I., Welsh, J. B., Sapinoso, L. M., Kern, S. G., Dimitrov, P., Lapp, H., Schultz, P. G., Powell, S. M., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. (2001). Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* *61*, 7388-7393.

Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., Oda, H., Endo, T., Kondo, S., Bono, H., Okazaki, Y., and Hayashizaki, Y. (2001). Protein-protein interaction panel using mouse full-length cDNAs. *Genome Res* *11*, 1758-1765.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* *96*, 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet* *22*, 281-285.

- Templin, M. F., Stoll, D., Schrenk, M., Traub, P. C., Vohringer, C. F., and Joos, T. O. (2002). Protein microarray technology. *Trends Biotechnol* 20, 160-166.
- Troesch, A., Nguyen, H., Miyada, C. G., Desvarenne, S., Gingeras, T. R., Kaplan, P. M., Cros, P., and Mabilat, C. (1999). Mycobacterium species identification and rifampin resistance testing with high-density DNA probe arrays. *J Clin Microbiol* 37, 49-55.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623-627.
- van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281, 827-842.
- van't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530-536.
- Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270, 484-487.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. *Science* 291, 1304-1351.
- Wang, D., Liu, S., Trummer, B. J., Deng, C., and Wang, A. (2002). Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat Biotechnol* 20, 275-281.
- Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., *et al.* (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077-1082.
- Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T., and Marincola, F. M. (2000). High-fidelity mRNA amplification for gene profiling. *Nat Biotechnol* 18, 457-459.
- Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000). Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 26, 225-228.
- Weinberger, S. R., Dalmasso, E. A., and Fung, E. T. (2002). Current achievements using ProteinChip Array technology. *Curr Opin Chem Biol* 6, 86-91.
- Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H., and Farnham, P. J. (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev* 16, 235-244.

- Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F., Jr., and Hampton, G. M. (2001). Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. *Cancer Res* 61, 5974-5978.
- White, K. P., Rifkin, S. A., Hurban, P., and Hogness, D. S. (1999). Microarray analysis of *Drosophila* development during metamorphosis. *Science* 286, 2179-2184.
- Wilson, D. S., and Nock, S. (2002). Functional protein microarrays. *Curr Opin Chem Biol* 6, 81-85.
- Wilson, M., DeRisi, J., Kristensen, H. H., Imboden, P., Rane, S., Brown, P. O., and Schoolnik, G. K. (1999). Exploring drug-induced alterations in gene expression in *Mycobacterium tuberculosis* by microarray hybridization. *Proc Natl Acad Sci U S A* 96, 12833-12838.
- Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* 51, 221-271.
- Wolfsberg, T. G., Gabrielian, A. E., Campbell, M. J., Cho, R. J., Spouge, J. L., and Landsman, D. (1999). Candidate regulatory sequence elements for cell cycle-dependent transcription in *Saccharomyces cerevisiae*. *Genome Res* 9, 775-792.
- 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.
- 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.
- Ye, R. W., Tao, W., Bedzyk, L., Young, T., Chen, M., and Li, L. (2000). Global gene expression profiles of *Bacillus subtilis* grown under anaerobic conditions. *J Bacteriol* 182, 4458-4465.
- Ye, R. W., Wang, T., Bedzyk, L., and Croker, K. M. (2001). Applications of DNA microarrays in microbial systems. *J Microbiol Methods* 47, 257-272.
- Yeoh, E.-J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., *et al.* (2002). Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* 1, 133-143.
- Yu, J. Y., DeRuiter, S. L., and Turner, D. L. (2002). RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc Natl Acad Sci U S A* 99, 6047-6052.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., *et al.* (2001). Global analysis of protein activities using proteome chips. *Science* 293, 2101-2105.

Zhu, H., Klemic, J. F., Chang, S., Bertone, P., Casamayor, A., Klemic, K. G., Smith, D., Gerstein, M., Reed, M. A., and Snyder, M. (2000). Analysis of yeast protein kinases using protein chips. *Nat Genet* 26, 283-289.

Ziauddin, J., and Sabatini, D. M. (2001). Microarrays of cells expressing defined cDNAs. *Nature* 411, 107-110.

Chapter 2

Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Published as: 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, 717-28.

My Contributions to This Project

This project was begun by Frank Holstege, and involved collaborations with Eric Lander and Todd Golub at the Whitehead/MIT Center for Genome Research. Frank used Affymetrix oligonucleotide arrays to study the role components of the RNA polymerase II holoenzyme and TFIID play in regulating global yeast gene expression. I joined him early in the project to develop a system of data analysis and presentation, as there was no precedent for how to accomplish this. John Wyrick joined the project soon afterwards and worked with Frank to profile components of the transcription apparatus. Tony Lee and Christoph Hengartner provided supporting molecular biology and genetic experiments. This work is described in this chapter.

Summary

Genome-wide expression analysis was used to identify genes whose expression depends on the functions of key components of the transcription initiation machinery in yeast. Components of the RNA polymerase II holoenzyme, the general transcription factor TFIID, and the SAGA chromatin modification complex were found to have roles in expression of distinct sets of genes. The results reveal an unanticipated level of regulation which is superimposed on that due to gene-specific transcription factors, a novel mechanism for co-ordinate regulation of specific sets of genes when cells encounter limiting nutrients, and evidence that the ultimate targets of signal transduction pathways can be identified within the initiation apparatus.

Introduction

Much of biological regulation occurs at the level of transcription initiation. Genes contain promoter sequences which are bound by transcriptional activators and repressors (Struhl, 1995; Ptashne and Gann, 1997). Activators recruit the transcription initiation machinery, which for protein-coding genes consists of RNA polymerase II and at least 50 additional components (Orphanides et al., 1996; Roeder, 1996; Greenblatt, 1997; Hampsey, 1998; Myer and Young, 1998). The transcription initiation machinery includes factors which bind to DNA, cyclin-dependent kinases which regulate polymerase activity, and acetylases and other enzymes which modify chromatin (Burley and Roeder, 1996; Kingston et al., 1996; Roth and Allis, 1996; Steger and Workman, 1996; Tsukiyama and Wu, 1997; Hengartner et al., 1998; Struhl, 1998)

Our understanding of eukaryotic gene expression remains limited in several ways. The complete set of transcriptional regulators has yet to be identified. How these regulators interact with and regulate components of the transcriptional machinery is not yet clear. The functions of just a fraction of the components of the transcriptional machinery are understood, and then only with respect to a small set of genes. Cells must adjust genome expression to accommodate changes in their environment and in their programs of growth control and development, but precisely how coordinate remodeling of genome expression is accomplished for signal transduction pathways or for the cell cycle clock has yet to be learned.

Genome-wide expression monitoring has recently become feasible with the description of complete genome sequences and through the development of cDNA and high-density oligonucleotide array technology (Lockhart et al, 1996; Chee et al., 1996; DeRisi et

al., 1997; Lashkari et al., 1997; Wodicka et al., 1997). Expression profiling has been used to examine differences in gene expression when yeast are grown in various media (Wodicka et al., 1997) and has revealed how yeast genome expression is remodeled during the metabolic shift from fermentation to respiration (DeRisi et al., 1997) and during the cell cycle (Cho et al., 1998). Expression profiling is also being used to improve our understanding of various aspects of human biology and disease (DeRisi et al., 1996; Schena et al., 1996), and to facilitate drug development (Gray et al., 1998). The data generated with genome-wide expression monitoring technology describes the level of each mRNA species in a population, but this data alone does not always produce significant new biological insights. Our knowledge of genome-wide transcriptional regulation is incomplete, making it difficult to understand how such genome-wide expression signatures transpire.

We are exploring the ability of genome-wide expression analysis to provide insights into the transcriptional regulatory circuitry of eukaryotic cells. Such study should provide the foundation and context for interpreting mechanistic studies in control of gene expression. One approach to this problem is to identify the set of genes regulated by each promoter-binding transcription factor. However, if the transcription initiation apparatus itself plays an important role in regulation of gene expression, then it is important to determine the extent to which each gene in the genome depends on the function of key components of the transcription machinery for its expression. We describe here the mRNA population of yeast cells, the requirement for key components of the transcriptional machinery in expression of this population, the observation that certain components of the general apparatus are themselves regulated when cells encounter limiting nutrients, evidence that the ultimate targets of signal transduction pathways can be identified within the initiation apparatus, and additional insights into genome-wide regulatory circuitry. It is well recognized that in

eukaryotes transcriptional control is in large part due to the combinatorial action of promoter-specific activators at enhancer and promoters. Our results reveal that the general transcription factors add an additional level of combinatorial control of eukaryotic gene expression.

Results

Features of the study

The study described here was designed to assess the requirement for key components of the RNA polymerase II transcriptional machinery. This was accomplished by using high density oligonucleotide arrays (HDAs) (Wodicka et al., 1997) to determine the effects of mutations in these components genome-wide. Detailed information and databases supporting all aspects of this study can be found on the World Wide Web at <http://web.wi.mit.edu/young/expression/>.

The yeast transcriptome

Knowledge of the levels of all detectable mRNA species in yeast is useful for evaluating the degree to which these levels depend on any one component of the transcription apparatus. To obtain this information and to assess the reproducibility of the HDA technology, RNA was harvested from two independent wild type cultures and compared using two sets of HDAs on two separate days. The HDAs used here can score mRNA levels for up to 6181 genes. Of the 5460 genes whose mRNA levels were accurately determined and compared in both experiments, 99% of the mRNAs differed no more than 1.7 fold, and only 35 transcripts (0.65%) showed more than a two-fold change. In order to prevent these minimal variations from influencing the results, all experiments were performed in duplicate. The levels determined for the 5460 transcripts in wild type yeast cells and additional information derived from this experiment can be found under “Yeast mRNA population” on the Web site. The SAGE method has previously been used to determine values for 4465 transcripts, the result of which has been termed the yeast transcriptome (Velculescu et al.,

1997). The sensitivity of the HDA technology permitted a determination of the levels of many additional gene products, and revealed that transcripts from 80% of expressed yeast genes exist at steady state levels of 0.1 to 2 molecules/cell.

Dependence of genome expression on key components of transcriptional machinery

At any one promoter, the transcriptional machinery might include the RNA polymerase II core enzyme, the general transcription factors (GTFs), the core Srb/mediator complex, the Srb10 CDK complex, the Swi/Snf complex, and the SAGA complex, among others (Figure 1). One or more subunits of each of these components has been investigated for its role in genome-wide gene expression through the use of mutations which affect either the function or the physical presence of the subunit (Table 1). Loss-of-function mutations in various components of the transcription apparatus were constructed or obtained from various investigators (see Study Design on the web site for details). Two types of mutations have proven to be useful in this study. For essential components of the apparatus, temperature-sensitive (ts) mutations are valuable because they allow the investigator to examine effects on gene expression at any point after inactivating the factor. For nonessential components, we have used either point mutations which knock out the catalytic function of known enzymatic activities, or complete deletion mutations. In each experiment, a mutant cell and its isogenic wild-type counterpart are grown to mid-log phase, the two populations are harvested, RNA is prepared, and hybridization to HDAs is carried out, all in duplicate.

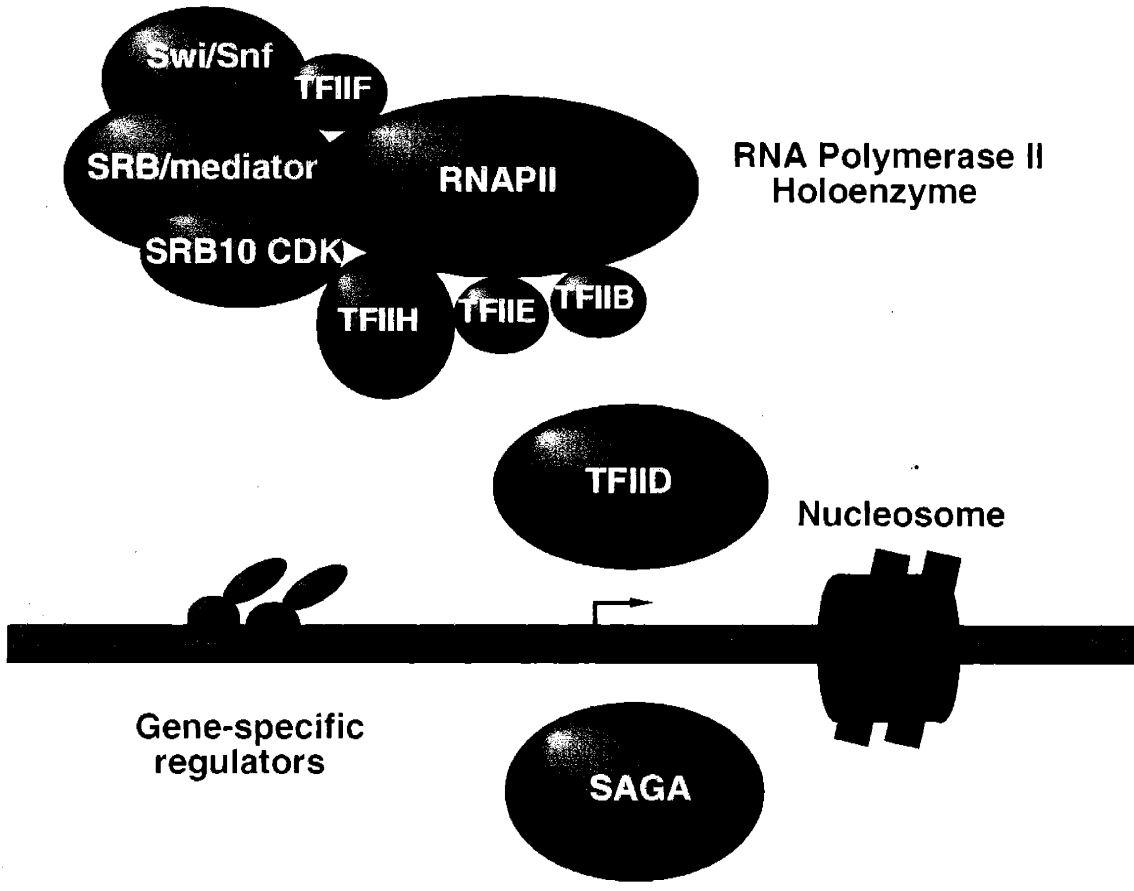


Figure 1. Model of RNA polymerase II transcription initiation machinery.

The machinery depicted here encompasses over 85 polypeptides in 10 (sub) complexes: core RNA polymerase II (RNAPII) consists of 12 subunits, TFIID 9 subunits, TFIIE 2 subunits, TFIIIF 3 subunits, TFIID 14 subunits, core SRB/mediator more than 16 subunits, Swi/Snf complex 11 subunits, Srb10 kinase complex 4 subunits and SAGA 13 subunits (see web site for more details). As detailed in Table 1, representative subunits of these complexes were chosen for analysis of genome-wide transcription dependence.

Table 1. Transcriptional Machinery

Complex and Subunit	Features	Fraction of genes dependent on subunit function
RNA Polymerase II Rpb1	Largest subunit, mRNA catalysis, contains CTD	100%
Srb/mediator (core)		
Srb4	Target of Gal4 activator	93%*
Srb5	Unknown function	16%
Med6	Role in activation of some genes	10%
Srb CDK complex		
Srb10	CTD kinase, negative regulator	3%
Swi/Snf		
Swi2	ATP-dependent chromatin remodelling	6%
General Transcription Factors		
TFIID (TAF _{II} 145)	Large TBP-associated factor, histone acetylase	16%
(TAF _{II} 17)	Component of both TFIID and SAGA	67%
TFIIE (Tfa1)	Promoter opening	54%
TFIIH (Kin28)	CTD kinase	87%*
SAGA		
Gcn5	Histone acetylase	5%
TAF _{II} 17	Component of both TFIID and SAGA	67%

*Srb4 and Kin28 results were essentially identical to Rpb1, but because of the stringency applied by the fit algorithm, a minimal estimate is produced.

Dependence on core RNA polymerase II

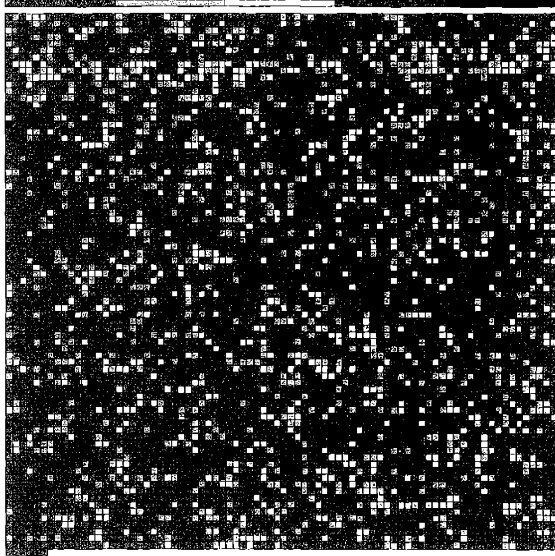
To determine the genome-wide dependence of gene expression on core RNA polymerase II, RNA was isolated from an *rpb1-1* ts cell and its wild-type counterpart 45 minutes after a shift to the nonpermissive temperature and was hybridized to HDAs. Because *rpb1-1* cells shut down transcription of protein-coding genes immediately after a temperature shift, these cells have been used by us and other investigators to determine the half-life of various yeast mRNAs (Nonet et al., 1987; Herrick et al., 1990). The 45 minute time point was used for the analysis of all ts mutants in this study because it is sufficiently long to detect a significant (i.e. a two-fold or more) loss of mRNA levels for 94% of detectable gene products without any loss of rRNA (Nonet et al., 1987). In addition, the 45 minute time point is short enough to minimize the potentially complicating effects of cell cycle arrest and cell death.

The results of genome wide expression analysis of the *rpb1-1* mutant as compared to an isogenic wild type strain are shown in a grid format in Figure 2A. The grid shows the change in mRNA level for each gene, beginning with the left-most gene on chromosome I and proceeding in a linear fashion, left to right, through chromosome I, then II, then III, etc., until the last gene on the right arm of chromosome XVI is reached at the lower right hand corner. 5735 genes were scored in this analysis. The vast majority of mRNAs are reduced more than two-fold in the mutant cells relative to wild type cells, and this reduction provides an apparent half-life for each of the mRNA species (see Yeast mRNA Population on the web site). The value determined with this approach is an approximation, but is useful for comparative purposes. Comparison of this data with that obtained for another ts factor

A. RPB1

Fold Change

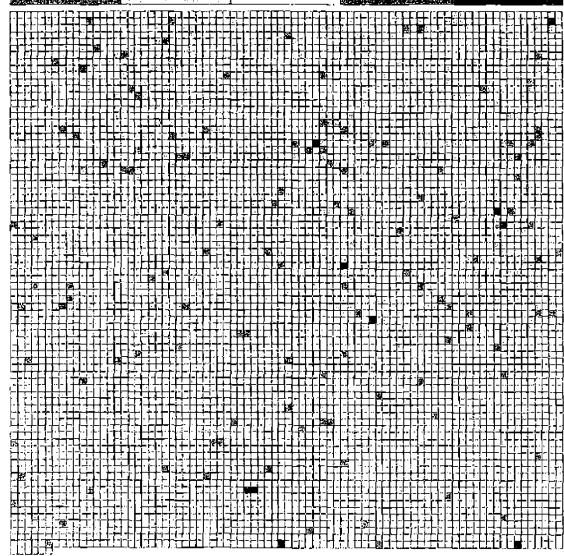
-4 -2 +2 +4



B. MED6

Fold Change

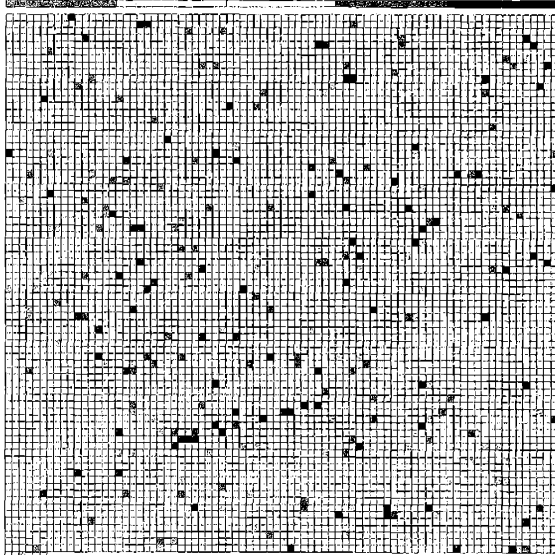
-4 -2 +2 +4



C. SRB10

Fold Change

-4 -2 +2 +4



D. SWI2

Fold Change

-4 -2 +2 +4

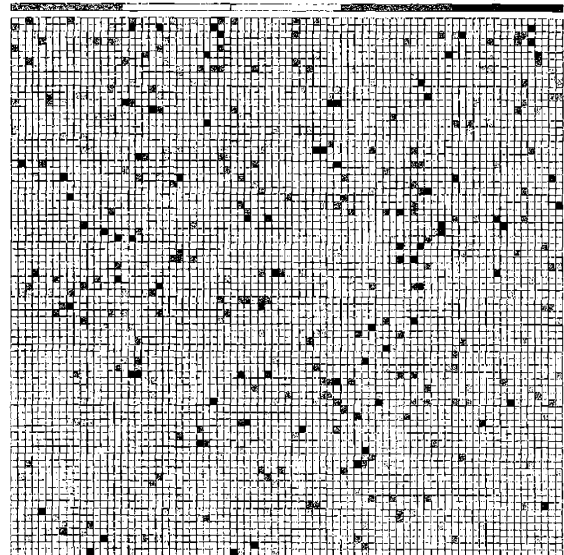


Figure 2. Genome-wide expression data for selected components of the RNA polymerase II holoenzyme.

Data reflecting the change in mRNA levels when a mutant is compared to its isogenic wild type counterpart is presented in a grid format. In the grid, the upper left grid square represents the left-most gene on chromosome I, and the squares to its right represent adjacent genes, proceeding in a linear fashion through chromosome I, then II, then III, etc., until the last gene on the right arm of chromosome XVI is reached at the bottom of the grid. The results are shown for (A) Rpb1, (B) Med6, (C) Srb10 and (D) Swi2.

identifies the set of genes whose expression is equivalently dependent on RNA polymerase II and the factor of interest.

There is a set of genes whose mRNAs are not significantly reduced in the mutant cells. These consist of genes that have stable messages as well as genes whose mRNA levels are slightly elevated in the mutant cells relative to wild-type. In this latter group are many known heat shock or stress response genes (e.g. SSA4, SSA3, HSP26, HSP30, HSP42 and SSL2), plus additional ORFs of unknown but perhaps related function. Similar results were obtained using ts mutants in other general transcription factors.

Dependence on Srb/mediator core subunits

The Srb/mediator complex is tightly associated with RNA polymerase II in a complex which has been termed the holoenzyme (Koleske and Young, 1994; Kim et al., 1994). Srb4 is an essential component of the Srb/mediator complex (Thompson et al., 1995; Kim et al., 1994; Hengartner et al., 1995). A ts mutant in Srb4 (*srb4-138*) was previously used to obtain evidence that several protein coding genes require the function of Srb4, and are thus likely to have the holoenzyme form of RNA polymerase II recruited to their promoters (Thompson and Young, 1995). Genome-wide expression analysis provides a more rigorous test of the model that expression of all protein-coding genes is dependent on Srb4. The experiment was carried out with the same protocol used with the Rpb1 ts mutant. Of the 5361 genes whose mRNA expression levels could be compared (i.e. those that had a greater than two-fold decrease in the experiment with Rpb1 ts and were scored in the Srb4 ts experiment), 93% showed a decrease that closely fit the decrease observed in the Rpb1 ts experiment. Of the mRNAs that did not closely fit the Rpb1 ts decay, only 2 could be found that reproducibly showed large differences in their decay in the two experiments performed. Furthermore, the

set of genes whose mRNAs are not significantly reduced in the Rpb1 ts mutant exhibit the same behavior in the Srb4 ts experiment. The results indicate that genome-wide expression is as dependent on Srb4 as it is on core RNA polymerase II (see Genome-Wide Expression Data on the web site for details). Because Srb4 is associated tightly and exclusively with the RNA polymerase II holoenzyme (Koleske and Young, 1994; Kim et al., 1994; Myers et al., 1998), we can infer that the Srb4-containing RNA polymerase II holoenzyme is generally required for transcription.

Med6 is another essential component of the Srb/Mediator complex and appears to be physically associated with Srb4 (Li et al., 1995; Myers et al., 1998; Lee et al., 1998). A Med6 ts mutant has been generated and used to demonstrate that Med6 is necessary for full induction of *GAL*, *SUC2*, *MFA1*, and *PYK1* genes but is not required for expression of several others (Lee et al., 1997). The genome-wide dependence of gene expression on Med6 was determined with this Med6 ts strain as described above for Rpb1. The results indicate that the expression of 10% of yeast genes are as dependent on Med6 as they are on Rpb1 (Figure 2B; see the web site for detailed information).

The reduction in mRNA levels observed in ts mutants soon after a temperature shift (i.e., 45 minutes) is likely a consequence of primary effects due to factor inactivation because of the time required to produce most secondary effects involves a substantial reduction in both a transcript and its translation product. Nonetheless, the results obtained in this type of experiment must be regarded as the sum of primary and secondary effects. To identify the set of genes whose change in expression is most likely a direct consequence of the loss of function of the ts factor, we compare data from ts inactivation of RNA polymerase II with that obtained by ts inactivation of any other factor. Comparison of the two data sets reveals the transcripts with equivalent decay kinetics in *rpb1-1* and the other ts mutant (see

Technology, Protocols and Data Analysis on the Web site for details). For those genes affected by ts disruption of Med6 where such a comparison could be made, the mRNAs of 506 genes decreased with similar kinetics in the Med6 and Rpb1 experiments. Thus, the expression of 10% of yeast genes are as dependent on Med6 as they are on Rpb1. These 506 genes are most likely to have a direct requirement for Med6 function. The genes whose transcript levels do not fit the Rpb1 kinetics could have a direct, but partial, requirement for Med6 function, or the effects observed at these genes are a secondary consequence of some other gene's altered mRNA levels. The 506 genes we have identified which require Med6 function to the same extent as Rpb1 function are those at which promoter-associated transcriptional regulators are most likely to function through interactions with Med6.

Srb5 is a component of the Srb/mediator complex whose function is also not known (Thompson et al., 1993; Kim et al., 1994; Koleske and Young, 1994; Hengartner et al, 1995; Myers et al., 1998). To determine the genome-wide dependence of gene expression on Srb5, a strain lacking an *SRB5* gene and its wild type counterpart were compared (see the web site for detailed information). The results indicate that 16% of all genes require Srb5 function for their expression. With the *SRB5* deletion strain and other constitutive mutants analyzed here, it is not possible to distinguish between results which are a direct consequence of the loss of Srb5 function and those which are due to a secondary effect such as the loss of another transcriptional regulator. Nonetheless, these results provide important information in that they reveal the complete set of genes which are directly or indirectly affected by loss of Srb5 function. It was striking that expression of many genes central to the pheromone response pathway are dramatically affected by the loss of Srb5, as discussed below.

Dependence on Srb10 CDK complex

Srb10 is cyclin dependent kinase which is part of a holoenzyme subcomplex containing Srb8, -9, -10 and -11 proteins (Liao et al., 1995; Hengartner et al., 1995). Srb10 and its associated proteins have been proposed to form a negative regulatory complex that functions through phosphorylation of the RNA polymerase II CTD (Hengartner et al., 1998). To determine how gene expression depends on Srb10, RNA was isolated from an Srb10 point mutant which lacks catalytic activity, and the expression profile was compared to that of its wild type counterpart. The results are shown in a grid format in Figure 2C. Of the 5626 genes that were scored, 173 gene products showed 2-fold or greater increases in mRNA levels in the mutant relative to the wild type. This indicates that Srb10 is normally a negative regulator of these 173 genes (approximately 3% of the genome). As discussed below, it is notable that nearly half of these genes are derepressed during nutrient deprivation.

Dependence on Swi/Snf

Swi2 ATPase activity plays an essential role in the ability of the Swi/Snf complex to remodel chromatin (Laurent et al., 1993; Cote et al., 1994; Khavari et al., 1993). This activity is thought to facilitate activator and transcription apparatus binding to promoter regions for a small number of genes, thereby overcoming repression by nucleosomes at those promoters (Cote et al, 1994; Imbalzano et al., 1994; Kwon et al., 1994; Burns and Peterson, 1997). Consequently, we anticipated that a small number of genes would be reduced in expression levels in the Swi2/Snf2 mutant. To determine the genome-wide dependence of gene expression on the Swi/Snf complex, RNA was isolated from a Swi2/Snf2 point mutant that lacks ATPase activity and its wild-type counterpart and the two RNA preparations were hybridized to HDAs. The surprising result was that a greater number of genes appear to be negatively regulated by Swi/Snf than are positively regulated (Figure 2D; see the web site for

detailed information). The data show that 203 gene products were elevated 2-fold or more in the mutant relative to the wild type, while just 126 transcripts decreased 2-fold or more (See Genome-Wide Expression Data on the web site). As described below, this result may be explained by recent data indicating that the Swi/Snf complex can catalyze chromatin remodeling in either direction (Schnitzler et al., 1998).

Dependence on general transcription factors

The general transcription factors are necessary to reconstitute promoter-dependent transcription *in vitro* with core RNA polymerase II. These factors include TFIID, TFIIB, TFIIF, TFIIE and TFIIH. Among these factors, TFIIE and TFIIH are of particular interest because numerous reports have suggested that they are in fact not generally required for gene expression (Parvin et al., 1992; Serizawa et al., 1993; Timmers 1994; Holstege, et al., 1995; Sakurai et al., 1997; Kuldell and Buratowski, 1997; Tijerina and Sayre, 1998). Genome-wide expression analysis was carried out on a Kin28 ts cell and its isogenic wild type counterpart using the same experimental protocol used for the Rpb1 ts mutant. Kin28, a CDK subunit of TFIIH, is an RNA polymerase II CTD kinase which is involved in the transition from initiation to elongation (Dahmus, 1996). The results reveal that Kin28 is generally required for expression of protein coding genes (see Genome-Wide Expression Data on the web site). TFIIE is thought to facilitate certain functions of TFIIH. In contrast to the results obtained with Kin28, analysis of genome-wide expression with a Tfa1 ts mutant shows that only 54% of yeast genes require the largest subunit of TFIIE to the same extent as core RNA polymerase II (see Genome-Wide Expression Data on the web site).

The TBP-associated factors (TAF_{II}s) of TFIID are especially interesting because they have been postulated to play important roles in promoter selectivity and gene activation

(Burley and Roeder, 1996; Verrijzer and Tjian, 1996; Lee and Young, 1998). A *ts* mutation in the TFIID subunit TAF_{II}145 (Walker et al., 1997) was used to determine the genome-wide dependence of gene expression on this TAF. Of the 5441 genes that were scored, 1618 gene products were reduced by 2-fold or greater on average in the two comparisons made, 45 minutes after temperature shift. For those genes where a comparison with the Rpb1 experiment could be made, 16% showed a dependence on TAF_{II}145 that was similar to their dependence on Rpb1 (see Genome-Wide Expression on the web site for details).

Interestingly, a large number of genes involved in functions associated with progression through the cell cycle are among the genes most likely to have a direct requirement for TAF_{II}145 function. The TAF_{II}145 *ts* mutant has a cell cycle phenotype: it arrests growth in G1-S after cells are shifted to the nonpermissive temperature. Previous studies showed that several G1-S cyclin genes are expressed at reduced levels in these cells, perhaps accounting for the cell cycle arrest phenotype (Walker et al., 1997). A subset of the genes that have a direct requirement for TAF_{II}145 function and which are involved in functions associated with progression through the cell cycle are listed in Table 2. For example, a significant decrease in mRNA levels was observed for *Ctr9*, which is required for expression

Table 2. Genes That Require Taf145 Function

Gene	Description	Fold Reduction
<u>Cell Cycle</u>		
*DDC1	DNA damage checkpoint protein	10
YER066W	Similar to CDC4 which degrades G1 cyclins	9
SPO1	Possible role in spindle pole body duplication	8
*LTE1	GDP/GTP exchange factor	8
*MKK2	Kinase involved in cell wall integrity	8
*BIM1	Possible role in early spindle pole body assembly	8
*MDM1	Involved in mitochondrial segregation	7
*CTR9	Required for normal expression of G1 cyclins	7
*PAC1	Possible role in spindle pole body orientation	6
*SCP160	Involved in control of chromosome transmission	6
CDC13	Telomere binding protein	6
*TOP3	DNA topoisomerase III	5
*TRX1	Thioredoxin I	5
ARD1	N-acetyltransferase	5
*SCC2	Required for sister chromatid cohesion	5
*CLB2	G2/M cyclin	5
*KIP2	Kinesin related protein	5
*MEC1	Cell cycle checkpoint protein	4
RAD9	DNA repair checkpoint protein	4
*SPC98	Spindle pole body component	4
*BCK1	Kinase involved in cell wall integrity	4
<u>DNA Repair</u>		
*RAD3	Involved in nucleotide excision repair	8
*YHR031C	Possible role in chromosome repair	7
*RAD5	Involved in DNA repair	6
*HSM3	Involved in mismatch repair	6
*RAD50	Involved in recombinational repair	5
*EXO1	Involved in mismatch repair	5
*MSH3	Involved in mismatch repair	5
YER041W	Similar to DNA repair protein, Rad2	5
REV1	Involved in translesion DNA synthesis	4
HDF2	Involved in DNA end-joining repair pathway	4
MSH6	Involved in mismatch repair	4
<u>DNA Synthesis</u>		
*MCM3	Involved in replication initiation, MCM/P1 family	13
RLF2	Chromatin assembly complex, subunit 2	9
*MCM6	Involved in replication initiation, MCM/P1 family	9
REV7	DNA polymerase subunit zeta	7
*MP1	Mitochondrial DNA-directed DNA polymerase	6
*CDC47	Involved in replication initiation, MCM/P1 family	6
*CDC5	Kinase	5
*CDC46	Involved in replication initiation, MCM/P1 family	5
*RFC1	DNA replication protein RFC large subunit	5
*CAC2	Chromatin assembly complex, subunit 1	5

* Gene exhibits equivalent dependence on Taf145 and Rpb1 for normal expression

of G1 cyclins Cln1 and Cln2. In addition, genes which are involved in DNA repair and DNA synthesis are dependent on TAF_{II}145 function. Thus, the G1/S arrest phenotype of TAF_{II}145 mutants may be due to multiple defects in cyclin and chromosome synthesis which occur during this period of the cell cycle.

We next analyzed which genes depend on TAF_{II}17, a histone H3-like TAF which is shared by TFIID and SAGA complexes, for their expression. RNA was isolated from a TAF_{II}17 temperature sensitive cell (TAF17-ts) and its wild type counterpart 45 minutes after a shift to the nonpermissive temperature and was hybridized to HDAs. Of the yeast genes identified in the TAF_{II}17 experiment and appropriate for comparison, 67% are as dependent on TAF_{II}17 function as they are on Rpb1, and are thus most likely to have a direct requirement for TAF_{II}17 function (see Genome-Wide Expression on the web site for details). This indicates that TAF_{II}17 is critical for the expression of a much larger portion of the transcriptome than TAF_{II}145. The presence of TAF_{II}17 in two different complexes may account for this observation.

Dependence on Gcn5 subunit of SAGA

The recent discovery that certain TAFs are components of both the TFIID general transcription factor and the SAGA complex (Grant et al., 1998) makes it particularly interesting to compare the effects of a mutation in a component specific to each complex (TAF_{II}145 in the case of TFIID and Gcn5 in the case of SAGA) with those of a mutation in a component shared by the two complexes (TAF_{II}17). The expression profile of a *GCN5* deletion mutant was compared with its isogenic counterpart (see Genome-Wide Expression

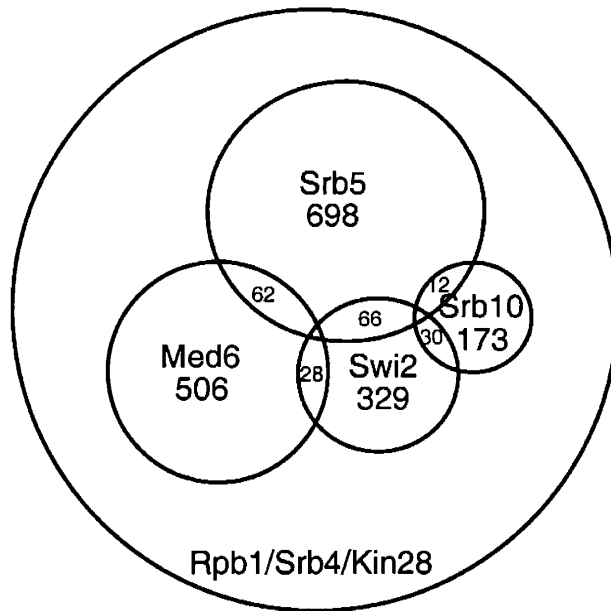
on the Web site for details). Of the 4912 genes which were scored, 185 transcripts were reduced by 2-fold or more and 83 increased by 2-fold or more.

The Gcn5 results indicate that this component of SAGA is necessary for normal expression of no more than 5% of yeast genes. Expression of 16% of protein-coding genes depends on the TAF_{II}145 subunit of TFIID to the same extent they depend on Rpb1. In contrast, the expression of 67% of yeast genes depends on the function of the TAF_{II}17 subunit shared by SAGA and TFIID.

Distinct requirements for components of transcriptional machinery

The analysis carried out thus far indicates that the Rpb1 subunit of core RNA polymerase II, the Srb4 subunit of the Srb/mediator complex, and the Kin28 subunit of the general transcription factor TFIIF are generally required for transcription of protein coding genes. In contrast, expression of only a subset of genes is dependent on Med6, Srb5, Srb10, Swi2, TAF_{II}145, TAF_{II}17, and Gcn5. The sets of genes whose expression requires various RNA polymerase II holoenzyme components are compared in the Venn diagram in Figure 3A. Similarly, the set of genes whose expression requires various TFIID and SAGA components are shown in a Venn diagram in Figure 3B. These diagrams show how distinct sets of genes require the function of distinct components of the transcription machinery. These data suggest that coordinate regulation of large sets of genes could be accomplished by affecting the function of specific components of the transcriptional machinery. If this is the case, then it would be expected that functional relationships exist among some genes within these sets, as has been observed with TAF_{II}145.

A. RNAP II Holoenzyme Components



B. TFIIID and SAGA Components

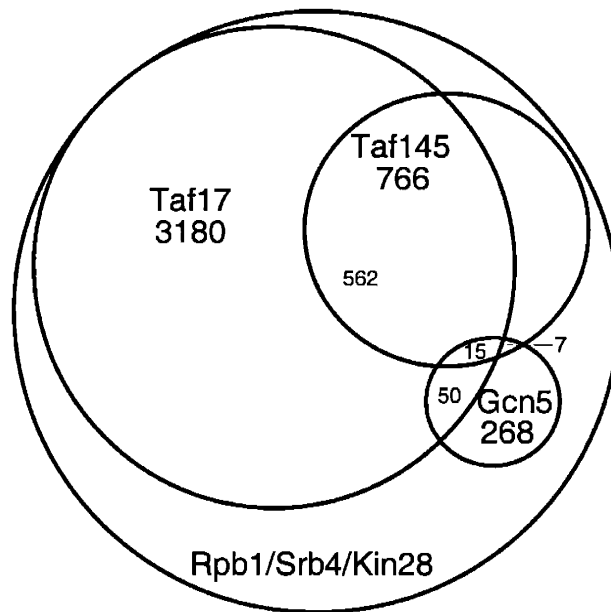


Figure 3. Genome-wide dependence on key components of the transcription machinery

(A) RNA polymerase II holoenzyme components show distinct patterns of genome control. Venn diagram depicting Srb5-, Swi2-, Srb10- and Med6-dependent genes (small circles) in relation to the whole transcriptome (Rpb1-, Srb4- and Kin28-dependent, large circle). The numbers under each subunit name are the sum of genes whose expression depends on that subunit.

(B) Genome control patterns of components of TFIIID and SAGA.

Srb5 has unexpected roles in pheromone response

It was striking that many of the genes whose mRNA levels are most dramatically affected by the loss of *Srb5* fall into the pheromone response pathway. The 15 genes involved in the pheromone response that are expressed at substantially lower levels in the absence of *Srb5* are shown in Figure 4A. Dramatic effects are seen in genes involved in mating factor production and export; the expression of *MFA1* and *MFA2*, the two genes encoding mating pheromone a-factor, are down 28-fold and 11-fold respectively. Additional genes involved in maturation (*STE13*) and export (*STE6*) of mating factor are expressed at substantially lower levels than in the cognate wild type. Furthermore, several components of the signal transduction pathway that responds to mating pheromone are expressed at reduced levels in the *Srb5* mutant. These genes include the receptor for pheromone (*STE2*), subunits of the signaling G-protein (*GPA1*), and the transcription factor which is itself the target of the signaling response and directly regulates subsequent gene expression (*STE12*).

The genome-wide expression profile for the *Srb5* mutant suggests that these cells should exhibit a defect in mating efficiency, a phenotype we had not previously suspected or investigated. Indeed, quantitative mating assays show that the *Srb5* mutant does have a significant defect in mating (Figure 4B). The mating defect was more pronounced than that due to mutations in *Fus3*, a MAP kinase required for cell cycle arrest and cell fusion during mating, but less pronounced than that due to mutations in *STE12*. The defect in mating deficiency exhibited by the *Srb5* mutant may reflect coordinate regulation of the set of pheromone response genes identified through genome-wide expression analysis.

A.

Gene	Description	Fold-reduction
MFA1	Mating pheromone a-factor	28
STE2	Alpha factor receptor	12
MFA2	Mating pheromone a-factor	11
BAR1	Protease that degrades alpha factor	10
SST2	Involved in desensitization to alpha factor	9
FAR1	Inhibitor of CDKs involved in cell-cycle arrest for mating	8
FUS2	Protein required for cell fusion during mating	6
STE6	Membrane transporter; exports a-factor	6
AGA2	a-agglutinin binding subunit	6
AGA1	a-agglutinin anchor subunit	5
STE12	Transcription factor binds to pheromone response element	4
GPA1	GTP-binding subunit of pheromone response pathway	4
STE13	Involved in maturation of alpha factor	4
KAR4	Required for pheromone induction of karyogamy genes	4

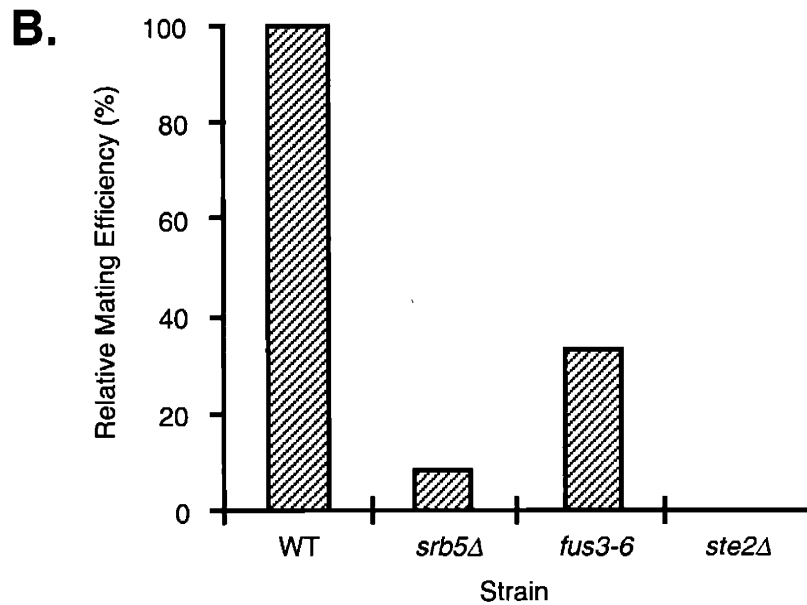


Figure 4. Srb5 is required for expression of pheromone response genes

(A) Pheromone response genes whose expression is reduced in the absence of Srb5.

(B) Cells lacking Srb5 are defective in mating. The mating efficiencies for mutant strains are expressed as a percentage of the mating efficiency of an isogenic wild-type strain. For comparison, strains with mutations in two components of the mating signal transduction pathway (Fus3 and Ste2) are included.

Coordinate regulation of nutrient starvation response genes via Srb10

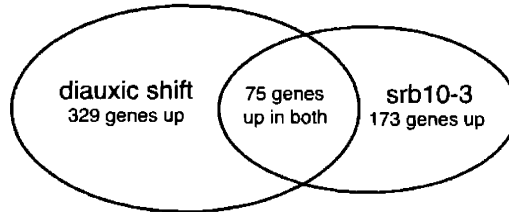
The analysis revealed that Srb10 is a negative regulator of 173 genes. It is notable that nearly half of these genes are derepressed during the nutrient deprivation that occurs during the diauxic shift (DeRisi et al., 1997) (Figure 5). Yeast cells undergo a diauxic shift as nutrients are depleted in culture, and a variety of genes which enable the cell to survive nutrient-limiting conditions are derepressed (Johnston and Carlson, 1992; Yin et al., 1996). These include genes involved in dimorphic morphology (nutrient starved cells alter their morphology to permit foraging for nutrients) and stress responses (starved cells are apparently better able to survive nutrient deprivation when stress proteins are elevated). Srb10 in wild type cells is most likely responsible for repressing this set of genes when cells are in exponential growth on glucose, but no longer performs this function as cells enter the diauxic shift. Coordinate regulation of this set of genes could be accomplished by eliminating the function of Srb10 as cells enter the diauxic shift.

To determine whether Srb10 is physically lost from cells as they enter the diauxic shift, cells containing an epitope-tagged Srb10 protein were grown in YPD media and sampled at various times during the growth curve (Figure 5C). Cell lysates were prepared from each sample and the levels of Srb10 were assayed by Western blot. The data in Figure 5C shows that Srb10 is physically depleted as cells enter the diauxic phase of growth. This result is consistent with evidence that the levels of Srb11, the cyclin partner of Srb10, are reduced when cells are exposed to the limiting nutrient environment in sporulation media (Cooper et al., 1997). It may also explain why a form of yeast holoenzyme purified from commercially available yeast cells lacks the Srb10/Srb11 kinase/cyclin pair (Li et al., 1995; Myers et al., 1998), as these cells are typically grown past mid-log phase. The results thus

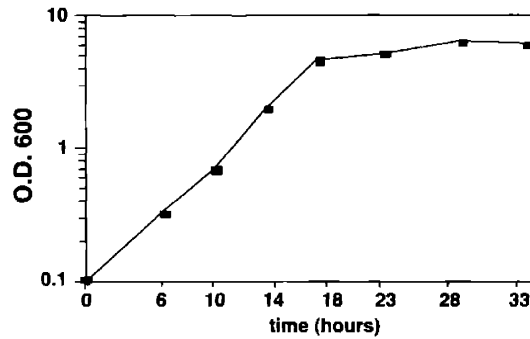
A.

Gene	Description	Fold Up
FLO1	Flocculence cell wall protein	102
SIP18	Induced by osmotic stress	74
YBR116C	Induced by diauxic shift	61
YMR107W	Induced by diauxic shift	32
ALD3	Induced by diauxic shift	28
HSP26	Induced by osmotic stress, diauxic shift	26
GRE1	Induced by osmotic stress, diauxic shift	25
YER150W	Induced by diauxic shift	24
HSP12	Induced by numerous stresses	18
RCK1	Serine/threonine protein kinase	18
FLO11	Flocculence	15
RTA1	Involved in 7-amincholesterol resistance	15
YDR070C	Induced by diauxic shift	13
YBR147W	Induced by diauxic shift	10
CTT1	Induced by osmotic stress, diauxic shift	10
YDL204W	Induced by diauxic shift	10
TKL2	Induced by diauxic shift	10
YGR043C	Induced by diauxic shift	9
YNL194C	Induced by diauxic shift	9
SOL4	Induced by diauxic shift	8
CYC7	Induced by numerous stresses, diauxic shift	8
PUT4	Proline permease, nitrogen induced	8
YKL187C	Induced by diauxic shift	8
NCA3	Life-span determination	8
YML128C	Induced by diauxic shift	8
GPH1	Induced by diauxic shift	8
POT1	Induced by diauxic shift	7

B.



C.



D.

SRB10/SRB10 srb10Δ/srb10Δ



Figure 5. Srb10 CDK represses genes elevated during response to nutrient starvation.

(A) Subset of 173 genes whose expression is derepressed in cells lacking Srb10 kinase activity.

(B) Venn diagram showing the number of genes which are derepressed during the nutrient deprivation which occurs during the diauxic shift and the fraction of these which are derepressed in cells lacking Srb10 kinase activity.

(C) Srb10 protein is depleted from cells as they enter the diauxic shift. The graph shows the growth curve of a yeast strain allowed to grow to stationary phase (33 hours). At specified time points, aliquots from a culture were measured for cell density and equal amounts of cells were harvested for Western blot analysis. Western blots against epitope-tagged Srb10p and a control protein Tub2p (tubulin) show that Srb10 levels decrease substantially as cells enter the diauxic shift (14 to 18 hours).

(D) Cells lacking Srb10 kinase activity exhibit increased pseudohyphal growth. The strains used to assay pseudohyphal growth were derived from L5978, and are congenic to the Σ 1278b genetic background (Liu et al., 1993).

indicate that the nutrient starvation response is mediated, in part, through the physical loss of the Srb10 CDK from the holoenzyme. This novel mechanism provides one example of how coordinate regulation of gene expression can be accomplished through regulation of components of the general initiation machinery.

FLO11, which encodes a cell wall protein that is highly expressed in pseudohyphal cells, is expressed at 15-fold higher levels when Srb10 function is lost (Figure 5A). The dramatic increase in the expression of *FLO11* and other genes whose products are involved in the dimorphic shift led us to determine whether the absence of Srb10 function produces a pseudohyphal phenotype. Both copies of the *SRB10* gene were deleted from a diploid strain that is generally used to assay this phenotype, and colony morphology was examined under the microscope. The results in Figure 5D demonstrate that the loss of Srb10 causes cells to grow preferentially in a pseudohyphal form. This again shows that expression analysis is useful for predicting unexpected phenotypes. More importantly, specific signal transduction pathways control the dimorphic shift (Madhani and Fink, 1998), and these results suggest that one of the ultimate targets of these pathways is the Srb10 kinase.

Discussion

We have characterized the mRNA population of yeast cells and the requirement for key components of the transcriptional machinery in expression of this population using HDA technology. The insights obtained from this analysis include the following. Genome-wide expression is equivalently dependent on Srb4 and Rpb1, suggesting that the Srb4-containing RNA polymerase II holoenzyme is generally recruited to promoters of protein-coding genes. Distinct expression signatures are obtained when a wide variety of components of the transcription apparatus are inactivated, revealing a level of genome regulation which can be superimposed on that due to gene-specific transcription factors. Coordinate regulation of functionally related genes can be effected by regulating a component of the initiation machinery, as exemplified by the regulation of Srb10 and the role of this kinase in the response to nutrient deprivation. The ultimate targets of certain signal transduction pathways can be identified by comparing genome expression signatures from these experiments and those that modify the cellular environment.

Transcriptome

We have estimated the number of mRNA molecules present for all genes in a single wild type haploid cell using HDA data (see Yeast mRNA Population on the web site). This is a more accurate representation of the transcriptome than that previously determined because it is better able to score mRNA species which are expressed at very low levels (5460 genes were scored using HDAs, whereas 4465 genes were scored with SAGE). It is particularly valuable to have information on transcripts from genes expressed at low levels because many of the regulatory components of the cell are expressed at low levels.

Gene-specific regulation via the general transcription machinery

The textbook models describe regulation of eukaryotic gene expression as the recruitment of the general transcription machinery to genes by gene-specific activators. Our results demonstrate that the function of certain key components of the general transcription machinery is required for the expression of distinct sets of genes, as illustrated in Figure 3. It is possible that these components of the transcription machinery are necessary targets for a specific set of gene-specific activators, and the loss of such a component produces a dramatic effect at only those genes under the control of such activators. In this case, the data here provide candidate targets within the initiation machinery for activators at most yeast genes.

The results described here also reveal that a layer of regulation is available to the cell in addition to that provided by gene-specific regulators: the expression of specific sets of genes can be regulated by affecting the availability or function of a specific component of the general machinery. Since various components of the general machinery can be acetylated and phosphorylated (Imhof et al., 1997; Kitajima et al., 1994), it is possible that these modifications serve to regulate these components, and thus the genes which require their functions.

Insights into roles of transcriptional complexes

The components of the transcription apparatus that were the focus of this study were selected because they are among the key subunits of the major multiprotein complexes which have roles in transcription of protein-coding genes. These complexes include the RNA polymerase II core enzyme, the general transcription factors (GTFs), the core Srb/mediator complex, the Srb10 CDK complex, the Swi/Snf complex, and the SAGA complex. We

found that three components were generally required for transcription of protein coding genes (Rpb1, Kin28, Srb4). Two were found to be required for more than half, but not all genes (Tfa1, Taf17). Most components investigated thus far were necessary for transcription of less than a fifth of the genome (Srb5, Med6, Srb10, Swi2, Taf145, Gcn5). In this latter group, the evidence indicates that Srb5, Med6, and Taf145 have predominantly positive roles, Srb10 has an almost exclusively negative role, and Swi2 and Gcn5 can have either a positive or a negative role in gene expression.

General Factors

Because Rpb1 and Srb4 proteins are generally required for expression of protein-coding genes, and they are both associated tightly and exclusively with RNA polymerase II and the mediator complex, respectively (Koleske and Young, 1994; Kim et al., 1994; Myers et al., 1998), we can infer that RNA polymerase II and the core mediator complex are generally required for transcription. Assuming that the function of Kin28 is restricted to TFIIF, the data obtained with the Kin28 mutant demonstrates that TFIIF is a general factor. Since the expression of 54% of yeast genes are as dependent on Tfa1 as they are on Rpb1, we infer that TFIIE is directly involved in expression of at least 54% of protein coding genes, but without knowing the contribution of Tfa2, the other subunit of TFIIE, we cannot eliminate the possibility that TFIIE has roles at additional genes.

SRB/mediator complex

The SRB/mediator core complex is essential for general transcription, as evidenced by the requirement for Srb4, but components such as Srb5 and Med6 have roles at specific subsets of genes. These results are consistent with the proposal that the Srb/mediator

complex is recruited to promoters of most genes together with RNA polymerase II, where it acts in a manner analogous to a signal processor with the capacity to integrate the combinatorial effects of multiple inputs from gene-specific transcriptional activators and repressors (Koleske and Young, 1994; Kim et al., 1994; Koh et al., 1998; Myers et al., 1998; Sun et al., 1998).

Srb10 CDK complex

The function of the Srb10 CDK complex can be defined by the kinase itself, since loss-of-function mutations in any of the four components of this complex produce identical phenotypes (Hengartner et al., 1995; Carlson, 1997). The Srb10 kinase is a negative regulator of a substantial fraction of genes that are repressed when cells grow vegetatively in rich media and are induced as cells experience nutrient deprivation. The genes regulated by Srb10 are involved in the nutrient stress response and in the morphological change that permits foraging for nutrients. Srb10 is physically depleted from cells as they enter the diauxic shift, providing a mechanism for derepression of this set of genes. Srb10 in wild-type cells is thus responsible for repressing this set of genes when cells are in exponential growth on glucose, but no longer performs this function as cells enter the diauxic shift.

Swi/Snf complex

If the function of the Swi/Snf complex is ATP-dependent remodeling of chromatin (Laurent et al., 1993; Cote et al., 1994), then the effects we observe due to the Swi2 ATPase mutation should represent the dependence of genome-wide expression on the entire Swi/Snf complex. The results indicate that a greater number of genes are negatively regulated by Swi/Snf than are positively regulated. This is surprising in view of the model that Swi/Snf-

catalyzed remodeling of chromatin facilitates activator binding. It is possible that chromatin remodeling may facilitate binding of negative factors as well as positive factors. An alternative possibility is suggested by recent data indicating that the Swi/Snf complex can remodel chromatin in both directions: it can convert a repressive nucleosome structure towards a more accessible state and vice versa (Schnitzler et al., 1998). It is thus possible that Swi/Snf helps produce a nucleosome structure conducive to transcription at some promoters, and a structure that is repressive at others.

TFIID and SAGA

The general transcription factor TFIID and the SAGA complex share two features: they both contain a subunit capable of histone acetylation (TAF_{II}145 in the case of TFIID and Gcn5 in the case of SAGA) and they share multiple subunits, among which is the histone H3-like TAF, TAF_{II}17 (Grant et al., 1998). As summarized in Figure 4, the results indicate that Gcn5, TAF_{II}145 and TAF_{II}17 are necessary for expression of 5%, 16%, and 67% of yeast genes, respectively. Two models can account for this data: one posits that TAF_{II}17 functions exclusively within the TFIID and SAGA complexes, and the other that TAF_{II}17 is a component of one or more additional complexes. If TAF_{II}17 functions exclusively within the TFIID and SAGA, then TAF_{II}145 and Gcn5 do not fully represent the functions of the two complexes, since the sum of genes which require TAF_{II}145 and Gcn5 function is much smaller than the number of genes which require TAF_{II}17. In this model, one or both complexes contain subunits which make different contributions to gene expression, as might be expected if different subunits are targets of different transcriptional activators and repressors. The results can also be accommodated in a second model, in which TAF_{II}17 is a

component of one or more complexes in addition to TFIID and SAGA. The results described here lay a useful foundation for the additional experiments necessary to gain a fuller understanding of the roles of TFIID and SAGA subunits in gene expression.

Our data, in conjunction with that of previous studies, reveal several striking similarities between TAF_{II}145 and prokaryotic sigma factors. First, both sigma factors and TAF_{II}145 are components of the general transcription machinery. Second, many sigma factors are required for the expression of a related subset of genes; likewise we have shown that TAF_{II}145 appears to be required for expression of a set of genes involved in chromosomal synthesis and G1/S progression. Finally, both sigma factors and TAF_{II}145 act through core promoter elements by direct DNA contacts.

Development of a genome control map

The results described here support the feasibility of dissecting the regulatory circuitry of the yeast genome by using genome-wide expression analysis on cells with mutations in the transcription apparatus. The set of yeast genes whose expression depends on the functions of key components of the transcriptional machinery has been identified. The genome-wide expression signatures produced by lesions in specific components of the transcription apparatus are quite distinct, making it possible to envision a genome control map. Such a map would identify all the components of the transcriptional machinery which have roles at any particular promoter and the contribution which specific components make to coordinate regulation of genes. The map will facilitate modeling of the molecular mechanisms that regulate gene expression and implicate components of the transcription apparatus in functional interactions with gene-specific regulators.

Experimental Procedures

Detailed information on experimental procedures, genetic reagents, HDA technology and data analysis can be found on the World Wide Web at <http://web.wi.mit.edu/young/expression/> in the section titled Study Design.

Acknowledgments

We are grateful to Gwen Acton, Gerry Fink, Tim Galitski, Nancy Hannett, David Hartemink, Fran Lewitter, Hiten Madhani, Johnny Park, Steve Rozen, Michael Sayre, Kevin Struhl, Pablo Tamayo, Joachim Theilhaber and Elizabeth Winzeler for helpful discussion, comments on the manuscript, suggestions for the Web site and software support. We thank David Allis, Gerard Faye, Young-Joon Kim, Craig Peterson and Michael Sayre for reagents, and are particularly grateful to Vic Myer for providing the tagged Srb10 strain. This work was supported by funds from the NIH, Bristol-Myers Squibb Company, Affymetrix Inc., and Millennium Pharmaceuticals Inc. F.C.P.H. was supported by fellowships from EMBO and the Human Frontier Science Program, E.G.J. is a predoctoral fellow of the Howard Hughes Medical Institute, and J.J.W. is a predoctoral fellow of the National Science Foundation. T.R.G. is a recipient of a Burroughs-Wellcome Fund Career Award in the Biomedical Sciences.

References

- Burley, S. K., and Roeder, R. G. (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem* 65, 769-99.
- Burns, L. G., and Peterson, C. L. (1997). The yeast SWI-SNF complex facilitates binding of a transcriptional activator to nucleosomal sites in vivo. *Mol Cell Biol* 17, 4811-9.
- Carlson, M. (1997). Genetics of transcription regulation in yeast: connections to the RNA polymerase II CTD. *Annu. Rev. Cell Dev. Biol.* 13, 1-23.
- Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S., and Fodor, S. P. (1996). Accessing genetic information with high-density DNA arrays. *Science* 274, 610-4.
- 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. *Molecular Cell* 2, 65-73.
- Cooper, K. F., Mallory, M. J., Smith, J. B., and Strich, R. (1997). Stress and developmental regulation of the yeast C-type cyclin Ume3p (Srb11p/Ssn8p). *EMBO J* 16, 4665-75.
- Cote, J., Quinn, J., Workman, J. L., and Peterson, C. L. (1994). Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science* 265, 53-60.
- Dahmus, M. (1996). Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *J Biol Chem.* 271, 19009-19012
- De Risi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A., and Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genet* 14, 457-60.
- De Risi, 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, 680-6.
- Grant, P. A., Schieltz, D., Pray-Grant, M. G., Steger, D. J., Reese, J. C., Yates, J. R., and Workman, J. L. (1998). A subset of TAF_{II}s are integral components of the SAGA complex required for nucleosome acetylation and transcriptional stimulation. *Cell* 94, 45-53.
- Gray, N.S., Wodicka, L., Thunnissen, A.-M., Norman, T.C., Kwon, S., Espinoza, F.H., Morgan, D.O., Barnes, G., LeClerc, S., Meijer, L., Kim, S.-H., Lockhart, D.J. and Schultz, P. (1998) Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533-538.

- Greenblatt, J. (1997). RNA polymerase II holoenzyme and transcriptional regulation. *Curr Opin Cell Biol* 9, 310-9.
- Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews* 62, 465-503.
- Hengartner, C. J., Myer, V. E., Liao, S.-M., Wilson, C. J., Koh, S. S., and Young, R. A. (1998). Temporal regulation of RNA polymerase II by Srb10 and Kin 28 cyclin-dependent kinases. *Molecular Cell* 2, 43-53.
- Hengartner, C. J., Thompson, C. M., Zhang, J., Chao, D. M., Liao, S. M., Koleske, A. M., Okamura, S., and Young, R. A. (1995). Association of an activator with an RNA polymerase II holoenzyme. *Genes and Development* 9, 897-910.
- Herrick, D., Parker, R., and Jakobson, A. (1990). Identification and comparison of stable and unstable mRNAs in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 10, 2269-2284.
- Holstege, F. C., Tantin, D., Carey, M., van der Vliet, P. C., and Timmers, H. Th. M. (1995). The requirement for the basal transcription factor IIE is determined by the helical stability of promoter DNA. *EMBO J* 14, 810-9.
- Imbalzano, A. N., Kwon, H., Green, M. R., and Kingston, R. E. (1994). Facilitated binding of TATA-binding protein to nucleosomal DNA. *Nature* 370, 481-5.
- Imhof, A., Yang, X. J., Ogryzko, V. V., Nakatani, Y., Wolffe, A. P., and Ge, H. (1997). Acetylation of general transcription factors by histone acetyltransferases. *Curr Biol* 7, 689-92.
- Johnston, M., and Carlson, M. (1992). In *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle and J. R. Broach, eds. (Cold Spring Harbor: Cold Spring Harbor Laboratory Press), pp. 193.
- Khavari, P. A., Peterson, C. L., Tamkun, J. W., Mendel, D. B., and Crabtree, G. R. (1993). BRG1 contains a conserved domain of the SWI2/SNF2 family necessary for normal mitotic growth and transcription. *Nature* 366, 170-4.
- Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H., and Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* 77, 599-608.
- Kingston, R. E., Bunker, C. A., and Imbalzano, A. N. (1996). Repression and activation by multiprotein complexes that alter chromatin structure. *Genes and Development* 10, 905-920.
- Kitajima, S., Chibazakura, T., Yonaha, M., and Yasukochi, Y. (1994). Regulation of the human general transcription initiation factor TFIIF by phosphorylation. *J Biol Chem* 269, 29970-7.

- Koh, S. S., Ansari, A. Z., Ptashne, M., and Young, R. A. (1998). An activator target in the RNA polymerase II holoenzyme. *Molecular Cell* *1*, 895-904.
- Koleske, A. J., and Young, R. A. (1994). An RNA polymerase II holoenzyme responsive to activators. *Nature* *368*, 466-9.
- Kuchin, S., and Carlson, M. (1998). Functional relationships of Srb10-Srb11 kinase, carboxy-terminal domain kinase CTDK-I, and transcriptional corepressor Ssn6-Tup1. *Mol Cell Biol* *18*, 1163-71.
- Kuldell, N. H., and Buratowski, S. (1997). Genetic analysis of the large subunit of yeast transcription factor IIE reveals two regions with distinct functions. *Mol Cell Biol* *17*, 5288-98.
- Kwon, H., Imbalzano, A. N., Khavari, P. A., Kingston, R. E., and Green, M. R. (1994). Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex. *Nature* *370*, 477-81.
- Lashkari, D. A., De Risi, 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. *Proc Natl Acad Sci U S A* *94*, 13057-62.
- Laurent, B. C., Treich, I., and Carlson, M. (1993). The yeast SNF2/SWI2 protein has DNA-stimulated ATPase activity required for transcriptional activation. *Genes Dev* *7*, 583-91.
- Lee, T. I., Wyrick, J. J., Koh, S. S., Jennings, E. G., Gadbois, E. L., and Young, R. A. (1998). Interplay of positive and negative regulators in transcription initiation by RNA polymerase II holoenzyme. *Mol. Cell. Biol.* *18*, 4455-4462.
- Lee, T. I. and Young, R.A. (1998) Regulation of gene expression by TBP-associated proteins. *Genes & Dev.* *12*, 1398-1408.
- Lee Y.C., Min S., Gim B.S., and Kim Y.J. (1997). A transcriptional mediator protein that is required for activation of many RNA polymerase II promoters and is conserved from yeast to humans. *Mol Cell Biol* *17*, 4622-32
- Li, Y., Bjorklund, S., Jiang, Y. W., Kim, Y. J., Lane, W. S., Stillman, D. J., and Kornberg, R. D. (1995). Yeast global transcriptional regulators Sin4 and Rgr1 are components of mediator complex/RNA polymerase II holoenzyme. *Proc Natl Acad Sci U S A* *92*, 10864-8.
- Liao, S. M., Zhang, J., Jeffery, D. A., Koleske, A. J., Thompson, C. M., Chao, D. M., Viljoen, M., van Vuuren, H. J., and Young, R. A. (1995). A kinase-cyclin pair in the RNA polymerase II holoenzyme. *Nature* *374*, 193-6.
- Liu, H., Styles, C.A., and Fink, G.R. (1993) Elements of the yeast pheromone response pathway required for filamentous growth of diploids. *Science* *262*, 1741-1744.

- Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* *14*, 1675-1680.
- Madhani, H.D., and Fink, G.R. (1998) The control of filamentous differentiation and virulence in fungi. *Trends Cell Biology* *8*, 348-353.
- Myer, V. and Young, R.A. (1998) RNA polymerase II holoenzymes and subcomplexes. *J. Biol. Chem.*, in press.
- Myers, L. C., Gustafsson, C. M., Bushnell, D. A., Lui, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R. D. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes Dev.* *12*, 45-54.
- Nonet, M., Scafe, C., Sexton, J., Young, R.A. (1987) Eukaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. *Mol Cell Biol* *7*, 1602-11
- Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev* *10*, 2657-83.
- Parvin, J. D., Timmers, H. T., and Sharp, P. A. (1992). Promoter specificity of basal transcription factors. *Cell* *68*, 1135-44.
- Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. *Nature* *386*, 569-77.
- Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* *21*, 327-35.
- Roth, S. Y., and Allis, C. D. (1996). Histone acetylation and chromatin assembly: a single escort, multiple dances? *Cell* *87*, 5-8.
- Sakurai, H., Ohishi, T., and Fukasawa, T. (1997). Promoter structure-dependent functioning of the general transcription factor III_E in *Saccharomyces cerevisiae*. *J Biol Chem* *272*, 15936-42.
- Schena, M. (1996). Genome analysis with gene expression microarrays. *Bioessays* *18*, 427-31.
- Schnitzler, G., Sif, S., and Kingston, R. E. (1998). Human SWI/SNF interconverts a nucleosome between its base state and a stable remodeled state. *Cell* *94*, 17-27.
- Serizawa, H., Conaway, J. W., and Conaway, R. C. (1993). Phosphorylation of C-terminal domain of RNA polymerase II is not required in basal transcription. *Nature* *363*, 371-4.

- Steger, D. J., and Workman, J. L. (1996). Remodeling chromatin structures for transcription: what happens to the histones? *Bioessays* 18, 875-84.
- Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. *Genes Dev* 12, 599-606.
- Struhl, K. (1995). Yeast transcriptional regulatory mechanisms. *Annu Rev Genet* 29, 651-74.
- Sun, X., Zhang, Y., Cho, H., Rickert, P., Lees, E., Lane, W., and Reinberg, D. (1998). NAT, a human complex containing Srb polypeptides that function as a negative regulator of activated transcription. *Molecular Cell* 2, 1-11.
- 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, 1361-75.
- Thompson, C. M., and Young, R. A. (1995). General requirement for RNA polymerase II holoenzymes in vivo. *Proc Natl Acad Sci U S A* 92, 4587-90.
- Tijerina, P., and Sayre, M. H. (1998). A debilitating mutation in transcription factor IIE with differential effects on gene expression in yeast. *J Biol Chem* 273, 1107-13.
- Timmers, H. Th. M. (1994). Transcription initiation by RNA polymerase II does not require hydrolysis of the beta-gamma phosphoanhydride bond of ATP. *EMBO J* 13, 391-9
- Tsukiyama, T., and Wu, C. (1997). Chromatin remodeling and transcription. *Curr Opin Genet Dev* 7, 182-91.
- Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D., Jr., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell* 88, 243-51.
- Verrijzer, C. P., and Tjian, R. (1996). TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem Sci* 21, 338-42.
- Walker, S. S., Shen, W. C., Reese, J. C., Apone, L. M., and Green, M. R. (1997). Yeast TAF(II)145 required for transcription of G1/S cyclin genes and regulated by the cellular growth state. *Cell* 90, 607-14.
- 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.
- Yin, Z., Smith, R. J., and Brown, A. J. (1996). Multiple signalling pathways trigger the exquisite sensitivity of yeast gluconeogenic mRNAs to glucose. *Mol Microbiol* 20, 751-64.

Chapter 3

Genome-wide Location Analysis

My Contributions to This Project

For the Gal4 part of this project, I performed the genome-wide expression data and analysis of the galactose/glucose experiment. I also created a web site to communicate the data and provide access to additional results not provided in the manuscript. For the study of Thi2, I conducted all of the experiments and performed all of the data analysis. The Thi2-myc strain was constructed by Nancy Hannett, the intergenic arrays used for location analysis were printed by Itamar Simon, Christopher Harbison, Jolyon Terragni and Thomas Volkert, and the the Operon oligonucleotide arrays were printed by Jolyon Terragni and Thomas Volkert.

Chapter 3

Part I:

Genome-wide Location and Function of the DNA Binding Protein Gal4

Published as a part of: Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000). Genome-wide location and function of DNA binding proteins. *Science* 290, 2306-9.

Summary

Understanding how DNA-binding proteins control global gene expression and chromosomal maintenance requires knowledge of the chromosomal locations where these proteins function *in vivo*. We developed a microarray method that reveals the genome-wide location of DNA-bound proteins and used this method to monitor binding of gene-specific transcription activators in yeast. A combination of location and expression profiles was used to identify genes whose expression is directly controlled by Gal4 as cells respond to changes in carbon source. The results identify pathways that are coordinately regulated by Gal4 and reveal novel functions for this regulator. Genome-wide location analysis will facilitate investigation of gene regulatory networks, gene function, and genome maintenance.

Introduction

Many proteins bind to specific sites in the genome to regulate genome expression and maintenance. Transcriptional activators, for example, bind to specific promoter sequences and recruit chromatin modifying complexes and the transcription apparatus to initiate RNA synthesis (Lee and Young, 2000; Malik and Roeder, 2000; Ptashne and Gann, 1997). The reprogramming of gene expression that occurs as cells move through the cell cycle, or when cells sense changes in their environment, is effected in part by changes in the DNA-binding status of transcriptional activators. Distinct DNA-binding proteins are also associated with origins of DNA replication, centromeres, telomeres, and other sites, where they regulate chromosome replication, condensation, cohesion, and other aspects of genome maintenance (Dutta and Bell, 1997; Kelly and Brown, 2000). Our understanding of these proteins and their functions is limited by our knowledge of their binding sites in the genome.

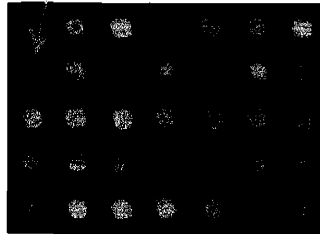
Results and Discussion

The genome-wide location analysis method we have developed allows protein-DNA interactions to be monitored across the entire yeast genome. The method combines a modified Chromatin Immunoprecipitation (ChIP) procedure, which has been previously used to study protein-DNA interactions at a small number of specific DNA sites (Orlando, 2000), with DNA microarray analysis. Briefly, cells were fixed with formaldehyde, harvested, disrupted by sonication, and DNA fragments crosslinked to a protein of interest were enriched by immunoprecipitation with a specific antibody. After reversal of the crosslinks, the enriched DNA was amplified and labeled with a fluorescent dye (Cy5) using ligation-mediated PCR (LM-PCR). A sample of DNA that was not enriched by immunoprecipitation was subjected to LM-PCR in the presence of a different fluorophore (Cy3), and both IP-enriched and unenriched pools of labeled-DNA were hybridized to a single DNA microarray containing all yeast intergenic sequences (Figure 1). A single-array error model (Roberts et al., 2000) was adopted to handle noise associated with low intensity spots and to permit a confidence estimate for binding (p-value). When independent samples of 1ng of genomic DNA were amplified with the LM-PCR method, signals for greater than 99.8% of genes were essentially identical within the error range (p-value < 10^{-3}). The IP-enriched/unenriched ratio of fluorescence intensity obtained from three independent experiments was used with a weighted average analysis method to calculate the relative binding of the protein of interest to each sequence represented on the array.

To investigate the accuracy of the genome-wide location analysis method, we used it to identify sites bound by the transcriptional activator Gal4 in the yeast genome. Gal4

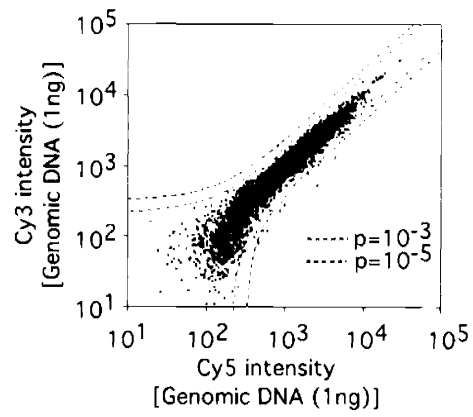
A.

Binding site



■ IP-enriched DNA
■ unenriched DNA
■ merged

B.



C.

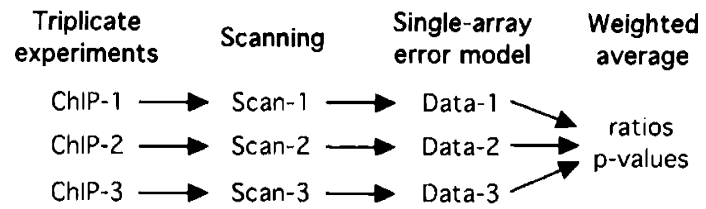


Figure 1. The Genome-wide location profiling method.

A. Close-up of a scanned image of a microarray containing DNA fragments representing 6361 intergenic regions of the yeast genome. The arrow points to a spot where the red intensity is over-represented, identifying a region bound in vivo by the protein under investigation.

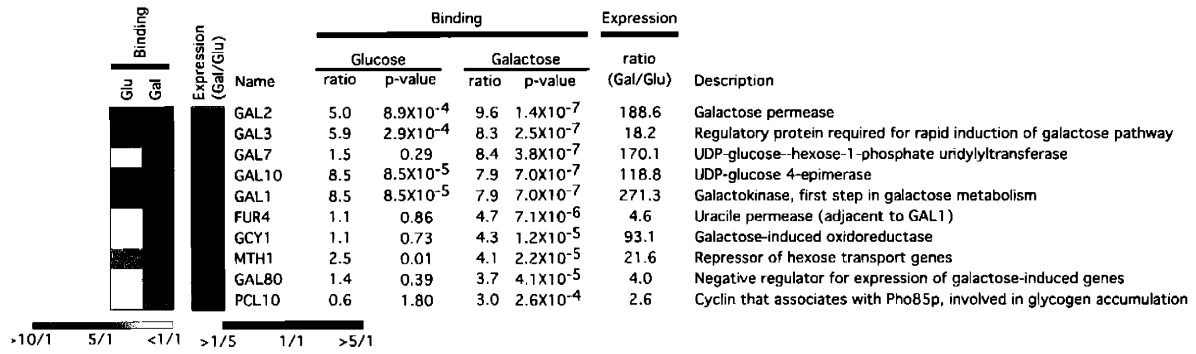
B. Analysis of Cy3- and Cy5-labeled DNA amplified from 1 ng of yeast genomic DNA using a single array error model (Roberts et al., 2000). The error model cutoffs for p-values equal to 10^{-3} and 10^{-5} are displayed.

C. Experimental design. For each factor, three independent experiments were performed and each of the three samples were analyzed individually using a single-array error model. The average binding ratio and associated p-value from the triplicate experiments were calculated using a weighted average analysis method.

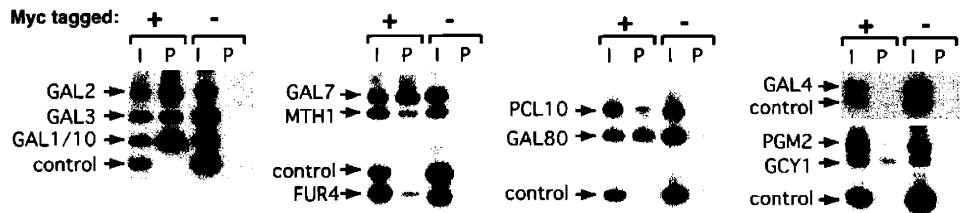
activates genes necessary for galactose metabolism and is among the best characterized transcriptional activators (Johnston and Carlson, 1992; Ptashne and Gann, 1997). Ten genes were found to be bound by Gal4 (p-value ≤ 0.001) and induced in galactose using our analysis criteria (Figure 2A). These included seven genes previously reported to be regulated by Gal4 (*GAL1*, *GAL2*, *GAL3*, *GAL7*, *GAL10*, *GAL80*, and *GCY1*). The *MTH1*, *PCL10*, and *FUR4* genes were also bound by Gal4 and activated in galactose. Each of these results was confirmed by conventional ChIP analysis (Figure 2B), and *MTH1*, *PCL10*, and *FUR4* activation in galactose was found to be dependent on Gal4 (Figure 2C). Both microarray and conventional ChIP showed that Gal4 binds to *GAL1*, *GAL2*, *GAL3* and *GAL10* promoters under glucose and galactose conditions, but the binding was generally weaker in glucose. The consensus Gal4 binding sequence that occurs in the promoters of these genes (CGGN₁₁CCG) can also be found at many sites through the yeast genome where Gal4 binding is not detected; therefore, sequence alone is not sufficient to account for the specificity of Gal4 binding *in vivo*. Previous studies of Gal4-DNA binding have suggested that additional factors such as chromatin structure contribute to specificity *in vivo* (Chasman et al., 1990; Marmorstein et al., 1992).

The identification of *MTH1*, *PCL10* and *FUR4* as Gal4-regulated genes reveals new functions for Gal4 and explains how regulation of several different metabolic pathways can be coordinated (Figure 2D). *MTH1* encodes a transcriptional repressor of certain *HXT* genes involved in hexose transport (Schmidt et al., 1999). Our results suggest that the cell responds to galactose by increasing the concentration of its galactose transporter at the expense of other transporters. In other words, while Gal4 activates expression of the galactose transporter gene *GAL2*, Gal4 induction of the *MTH1* repressor gene leads to reduced levels of glucose transporter expression. The Pcl10 cyclin associates with

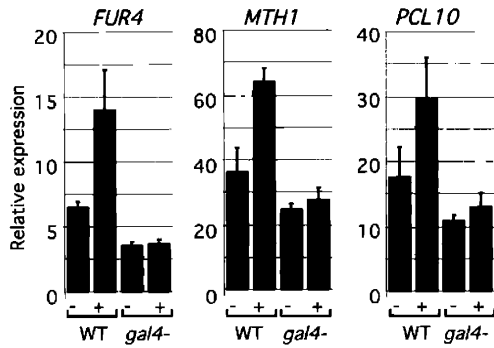
A.



B.



C.



D.

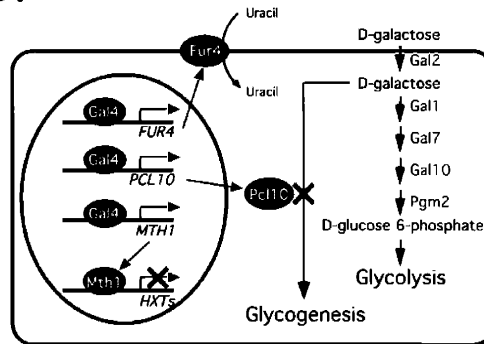


Figure 2. Genome-wide location of Gal4 protein.

A. Genes whose promoter regions were bound by myc-tagged Gal4 (p-value < 0.001) and whose expression levels were induced at least 2-fold by galactose are listed. The weight-averaged ratios and p-values are shown for Gal4 binding in galactose and glucose.

Binding ratios are also displayed using a blue/white color scheme and galactose/glucose expression ratios are displayed using a red/green color scheme.

B. Confirmation of microarray data for each gene in panel A using conventional chromatin IP procedure. Strains with (+) or without (-) a myc-tagged Gal4 protein were grown in galactose. Amplification of the unenriched DNA (I) and IP-enriched DNA (P) is shown. *ARN1* (control) was used as a negative control.

C. Galactose-induced expression of *FUR4*, *MTH1* and *PCL10* is Gal4-dependent.

Samples from wild type and gal4- strains were taken before and after addition of galactose. The expression of *FUR4*, *MTH1*, and *PCL10* was monitored by quantitative RT-PCR and quantified by phosphorimaging.

D. Model summarizing the role of Gal4 in galactose-dependent cellular regulation. The products of genes newly identified as directly regulated by Gal4 are shown as green circles; those previously identified are shown in blue.

Pho85p and appears to repress the formation of glycogen (Huang et al., 1998). Thus, the observation that *PCL10* is Gal4-activated suggests that reduced glycogenesis occurs to maximize the energy obtained from galactose metabolism. *FUR4* encodes a uracil permease (Jund et al., 1988) and its induction by Gal4 may reflect a need to increase intracellular pools of pyrimidines to permit efficient UDP addition to galactose catalyzed by Gal7.

We have shown that a combination of genome-wide location and expression analysis can identify the global set of genes whose expression is controlled directly by transcriptional activators *in vivo*. The application of location analysis to two yeast transcriptional activators revealed how multiple functional pathways are coordinately controlled *in vivo* during the response to specific changes in the extracellular environment. All of the known targets for Gal4 were confirmed, and new functional modules were discovered that are regulated directly by this factor.

Expression analysis with DNA microarrays allows investigators to identify changes in mRNA levels in living cells, but the inability to distinguish direct from indirect effects limits the interpretation of the data in terms of the genes that are controlled by specific regulatory factors. Genome-wide location analysis provides information on the binding sites at which proteins reside through the genome under various conditions *in vivo*, and will prove to be a powerful tool for further discovery of global regulatory networks.

Experimental Procedures

Genome-wide expression analysis

S. cerevisiae FY98 (MATa *ura3-52 leu2Δ1*) (Madison and Winston, 1997) was grown in galactose-containing medium (YEPGalactose) and half of the culture was switched to glucose-containing medium (YPD), where they were grown for 6 hours before harvesting (the final OD₆₀₀ was 0.5–0.7). Total RNA was isolated by using a hot-phenol method (Ausubel et al., 1994). Poly(A)+ RNA was purified from total RNA with Oligotex oligo(dT) selection step (Qiagen, Chatsworth, CA). Poly(A)+ RNA was converted into double-stranded cDNA by using a modified oligo(dT) primer with a T7 RNA polymerase promoter sequence at the 5' end. *In vitro* transcription of the double-stranded cDNA was performed with T7 RNA polymerase (T7 Megascript kit, Ambion, Austin, TX) and included biotin-labeled CTP and UTP (biotin-11-CTP, biotin-16-UTP, Enzo Diagnostics) to generate cRNA. The cRNA, along with a set of 5 poly-A tagged controls, was hybridized to a set of four oligonucleotide arrays (GeneChip Ye6100 arrays, Affymetrix, Santa Clara, CA). After washing, the arrays were stained with streptavidin-phycoerythrin (Molecular Probes, Eugene, OR) and read with a confocal scanner (Hewlett Packard). Intensities were captured using GeneChip software (Affymetrix) and a single raw expression level for each gene was determined. Individual mRNA levels were scored if the computer algorithm used for analysis (Wodicka et al., 1997) returned a "present" call in both the two galactose and the two glucose expression profiles for that gene or if the expression levels of that gene changed in the same direction and were greater than background levels in both wild type and mutant comparisons. A decrease was called if an mRNA dropped more than two-fold in both comparisons.

Epitope Tagging of Yeast Strains

The Gal4 activator was tagged with a 18 copies of the myc epitope by inserting the epitope coding sequence into the normal chromosomal loci of these genes. Vectors developed by (Cosma et al., 1999) were used for amplifying a fragment that contains the repeated myc tag coding sequence flanked by 50 bp from both sides of the stop codon of the gene. The PCR products were transformed into the W303 strain Z1256 (MATa, ade2-1, trp1-1, can1-100, leu2-3,112, his3-11,15, ura3) to generate the GAL4::18myc::TRP1 tagged strain. Clones were selected for growth on TRP- plates, the insertion of the tagged sequence was confirmed by PCR, and expression of the epitope-tagged protein was confirmed by western blotting using an anti-Myc antibody (9E11).

Chromatin Immunoprecipitation and Genome-wide Location Analysis

Briefly, yeast strains containing a myc-tagged version of the protein of interest were grown to mid log phase (OD 0.6–1.0), fixed with 1% formaldehyde for 20 min at room temperature and overnight at 4°C, harvested, and disrupted by mixing with glass beads on a Vibrax-VXR at 4°C for 2 hours. DNA was sheared to an average length of 400 bp by sonication. The DNA fragments crosslinked to the protein were enriched by immunoprecipitation with anti-myc specific monoclonal antibody (9E11). After reversal of the crosslinks, proteins were digested and the enriched DNA was purified over a QiaQuick (Qiagen) PCR purification column. For gene-specific analysis, PCR with the appropriate primers was performed with the enriched DNA or with a sample that was not enriched by immunoprecipitation. For genome-wide location analysis, the enriched DNA was amplified and labeled with a fluorescent dye (Cy5) with the use of a ligation-mediated polymerase chain reaction (LM-PCR). A sample of DNA that was not enriched by immunoprecipitation

was subjected to LM-PCR in the presence of a different fluorophore (Cy3), and both immunoprecipitation (IP)-enriched and -unenriched pools of labeled DNA were hybridized to a single DNA microarray containing all yeast intergenic sequences.

Images of Cy3 and Cy5 fluorescence intensities were generated by scanning the arrays using a GSI Lumonics Scanner. The Cy3 and Cy5 images were analyzed using ArrayVision software, which defined the grid of spots and quantified the average intensity of each spot and the surrounding background intensity. The background intensity was subtracted from the spot intensity to give the final calculated spot intensity. The intensity of the two channels was normalized according to the median. For each spot, the ratio of corrected Cy5/Cy3 intensity was computed. Each experiment was carried out in triplicate, and a single-array error model was used to handle noise, to average repeated experiments with appropriate weights, and to rank binding sites by p-value (Roberts et al., 2000). Additional details regarding this analysis can be found at <http://web.wi.mit.edu/young/location/>.

The intergenic regions present on the array were assigned to the gene or genes found transcriptionally downstream. Where a single intergenic region contains promoters for two divergently transcribed genes, the intergenic region was assigned to the gene or genes induced in the presence of galactose, as defined by our genome-wide expression analysis. Promoter regions detected with a p-value < 0.001 were included for further analysis.

Microarray design and production

The 6361 intergenic regions were amplified using the Yeast Intergenic Region Primers (Research Genetics) primer set. 50 μ L PCR reactions were performed in 96-well plates with each primer pair with the following conditions: 0.25 μ M of each primer, 20 ng of yeast genomic DNA, 250 μ M of each dNTP, 2 mM MgCl₂, 1X PCR buffer (Perkin Elmer), and

0.875 units of Taq DNA polymerase (Perkin Elmer). PCR amplification was performed in MJ Research Thermocyclers beginning with 2 minute denaturation at 95°C, followed by 36 cycles of 30 seconds at 92°C, 45 seconds at 52°C, and 2 minutes at 72°C, with a final extension cycle of 7 minutes at 72°C. 1 μ L of each PCR reaction mix was then reamplified in a 100 μ L PCR reaction using universal primers (Life Technologies) with the same reagent concentrations and the following thermocycling conditions: 3 minutes at 94°C, followed by 25 cycles of 30 seconds at 94°C, 30 seconds at 60°C, and 1 minute at 72°C, with a final extension cycle of 7 minutes at 72°C. Each PCR product was verified by gel electrophoresis. The PCR products were then precipitated with 2-propanol, washed with 70% ethanol, dried overnight, and resuspended in 20 μ L of 3XSSC. The resuspended DNA was transferred to 384 well plates and printed on GAPS-coated slides (Corning) using a Cartesian robot (Cartesian Technologies). The printed slides were rehydrated, snap-dried, and UV crosslinked in UV Stratalinker (Stratagene) set at 60 mJoules. The slides were then stored under vacuum for at least 2 days prior to hybridization.

Acknowledgments

We are grateful to Pia Cosma, Kim Nasmyth and Vic Myer for reagents, protocols and helpful discussions. Supported by funds from National Institutes of Health, Helen Hay Whitney foundation, National Cancer Institute of Canada, National Science Foundation, Howard Hughes Medical Institute, European Molecular Biology Organization, and the Human Frontier Science Program.

References

- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K. (1994). *Current Protocols in Molecular Biology*, Current Protocols).
- Chasman, D. I., Lue, N. F., Buchman, A. R., LaPointe, J. W., Lorch, Y., and Kornberg, R. D. (1990). A yeast protein that influences the chromatin structure of UASG and functions as a powerful auxiliary gene activator. *Genes Dev* *4*, 503-14.
- Cosma, M. P., Tanaka, T., and Nasmyth, K. (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell* *97*, 299-311.
- Dutta, A., and Bell, S. P. (1997). Initiation of DNA replication in eukaryotic cells. *Annu Rev Cell Dev Biol* *13*, 293-332.
- Huang, D., Moffat, J., Wilson, W. A., Moore, L., Cheng, C., Roach, P. J., and Andrews, B. (1998). Cyclin partners determine Pho85 protein kinase substrate specificity in vitro and in vivo: control of glycogen biosynthesis by Pcl8 and Pcl10. *Mol Cell Biol* *18*, 3289-99.
- Johnston, M., and Carlson, M. (1992). Regulation of Carbon and Phosphate Utilization. In *The Molecular and Cellular Biology of the Yeast Saccharomyces*, E. W. Jones, J. R. Pringle and J. R. Broach, eds. (Cold Spring Harbor, N-Y: Cold Spring Harbor Laboratory Press), pp. 193-281.
- Jund, R., Weber, E., and Chevallier, M. R. (1988). Primary structure of the uracil transport protein of *Saccharomyces cerevisiae*. *Eur J Biochem* *171*, 417-24.
- Kelly, T. J., and Brown, G. W. (2000). Regulation of Chromosome Replication. *Annu Rev Biochem* *69*, 829-880.
- Lee, T. I., and Young, R. A. (2000). Transcription of eukaryotic protein-coding genes. *Ann. Rev. Genetics* *34*.
- Madison, J., and Winston, F. (1997). Evidence that Spt3 functionally interacts with Mot1, TFIIA, and TATA-binding protein to confer promoter-specific transcriptional control in *Saccharomyces cerevisiae*. *Mol Cell Biol* *17*, p287-295.
- Malik, S., and Roeder, R. G. (2000). Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. *Trends Biochem Sci* *25*, 277-83.
- Marmorstein, R., Carey, M., Ptashne, M., and Harrison, S. C. (1992). DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* *356*, 408-14.

Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked- chromatin immunoprecipitation. *Trends Biochem Sci* 25, 99-104.

Ptashne, M., and Gann, A. (1997). Transcriptional activation by recruitment. *Nature* 386, 569-77.

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., Tyers, M., Boone, C., and Friend, S. H. (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287, 873-80.

Schmidt, M. C., McCartney, R. R., Zhang, X., Tillman, T. S., Solimeo, H., Wolf, S., Almonte, C., and Watkins, S. C. (1999). Std1 and Mth1 proteins interact with the glucose sensors to control glucose-regulated gene expression in *Saccharomyces cerevisiae*. *Mol Cell Biol* 19, 4561-71.

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

Chapter 3

Part II:

Thi2 is a Transcriptional Activator of Thiamin Biosynthetic Genes *in Vivo*

Summary

Thiamin is a cofactor for several critical metabolic enzymes. In *S. cerevisiae*, thiamin can be synthesized *de novo* or taken up from the surroundings when it is available. A combination of genome-wide expression and location analysis with yeast grown under conditions of low and high thiamin revealed that Thi2 binds to the promoters of thiamin biosynthetic genes only in the absence of thiamin. Additional experiments confirmed that Thi2 is itself induced by low thiamin and demonstrated that it is a transcriptional activator *in vivo*. Our results demonstrate how a change in environmental conditions can cause an organism to reprogram gene expression by inducing the expression of a transcriptional activator. Understanding a transcriptional network such as this will be useful in constructing a cellular regulatory network map.

Introduction

With the availability of complete genome sequences, biologists can begin to systematically attempt to decipher the regulatory circuitry of the genome. As *S. cerevisiae* has a relatively well annotated genome and is genetically tractable, it is an ideal system for using genome-wide approaches to elucidate the genome control network. Several groups have undertaken enormous efforts using a variety of techniques such as the two-hybrid system (Ito et al., 2001; Newman et al., 2000; Uetz et al., 2000), immunoprecipitation followed by mass spectrometry (Gavin et al., 2002; Ho et al., 2002), parallel deletion analysis (Tong et al., 2001; Winzeler et al., 2000), and genome-wide expression analysis of deletion mutants (Hughes et al., 2000), to begin to catalog biological connections. However, it is difficult to deconvolute what are the precise causes for the effects observed. Alternatively, one can cause minor changes to the physiology of an organism in a way such that a small number of genes are changing their expression. This approach may allow for the understanding of portions of the regulatory circuitry. The fusion of several different minicircuits will lead to the foundations of a genome-wide control map.

An alteration of the growth environment by the removal of a single nutrient would be expected to cause changes in gene expression that are predominantly in response to the absence of the nutrient. An example of such a nutrient is the vitamin thiamin (Vitamin B1). Thiamin pyrophosphate, which is synthesized from thiamin, is required for the function of several key enzymes that are involved in metabolism. Two enzymes that are in amino acid synthesis pathways, acetolactate synthase, which performs the first step in

branched chain amino acid synthesis, and transketolase, which is required for aromatic amino acid synthesis, require thiamin as a cofactor (Poulsen and Stougaard, 1989; Schaaff-Gerstenschlager et al., 1993; Sundstrom et al., 1993). Pyruvate, the product of glycolysis, is converted to ethanol by pyruvate decarboxylase during fermentation and to acetyl-CoA by pyruvate dehydrogenase for respiration. Both these enzymes require thiamin for their activity (Hohmann, 1991; Hohmann and Cederberg, 1990; Steensma et al., 1990). Alpha-ketoglutarate dehydrogenase, an enzyme similar to pyruvate dehydrogenase, also uses thiamin (Repetto and Tzagoloff, 1989). Additionally, there are two other proteins of unknown function in the yeast sequence database that exhibit sequence similarity to pyruvate decarboxylase and consequently are predicted to require thiamin as a cofactor (Hohmann and Meacock, 1998).

In humans, thiamin cannot be synthesized and therefore must be supplied in the diet. Beriberi, a disease that affects the cardiovascular, muscular, gastrointestinal and nervous systems, results from the thiamin deficiency (Begley, 1996). It is primarily seen in Southeast Asian populations, but in developed countries is most often seen in alcoholics (Krishna et al., 1999). Interestingly, HIV-infected individuals have also been found to be at increased risk for thiamin deficiency (Muri et al., 1999).

Exogenous thiamin can be used by yeast or it can be synthesized by the organism in its absence (Iwashima et al., 1973). Several genes that code for enzymes that are involved in thiamin biosynthesis or transport have been identified in *S. cerevisiae* (Enjo et al., 1997; Llorente et al., 1999; Nosaka et al., 1993; Nosaka et al., 1994; Praekelt et al., 1994; Schweingruber et al., 1986; Singleton, 1997). These genes have been found to be repressed in the presence of exogenous thiamin, and induced in its absence. Additional

unidentified proteins are involved in synthesis of early precursors of thiamin. As the mechanisms of synthesis of these early precursors appear to be different in prokaryotes from eukaryotes, they have been hard to identify (Estramareix and David, 1996).

Three proteins, Pdc2, Thi3, and Thi2, have been identified genetically as potential regulators of thiamin biosynthetic genes (Hohmann and Meacock, 1998; Nishimura et al., 1992a; Nishimura et al., 1992b). Pdc2 has been described as a regulator of pyruvate decarboxylase and has an asparagine-rich box that may be involved in transcriptional activation (Hohmann, 1993; Hohmann and Meacock, 1998). Deletion of *PDC2* is lethal, indicating that if it does play a role in thiamin biosynthetic gene regulation, it has other vital cellular functions as well (Winzeler et al., 1999). Under conditions of low thiamin, deletion of *THI2* or *THI3* prohibits the activation of several thiamin biosynthetic genes (Enjo et al., 1997; Nosaka et al., 1993; Nosaka et al., 1994). Thi3 is similar to pyruvate decarboxylase and appears to exhibit a decarboxylase activity involved in the production of isoamyl alcohol (Dickinson et al., 1997). Thi2 has a Zn₂-Cys₆ domain that is characteristic of several fungal transcription factors such as Gal4, and therefore seemed to be the most likely candidate to be a transcriptional regulator of thiamin biosynthetic genes.

It is unknown how thiamin biosynthetic genes are activated and why they are not expressed in the presence of thiamin. Therefore, we have performed genome-wide expression analysis to reveal changes in gene expression due to the absence of thiamin. We have identified Thi2 as a transcriptional regulator that is itself induced in the absence of thiamin. Using genome-wide location analysis, we found that it binds to promoters of

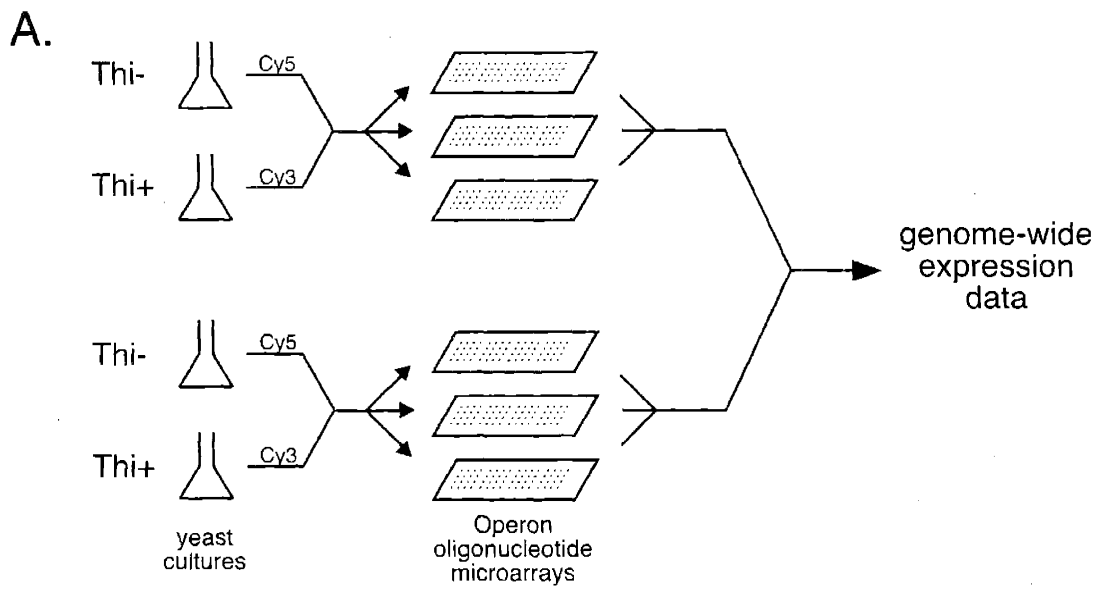
thiamin biosynthetic genes, many of which have a conserved element in their promoter.
The methods used here may be a useful model for analyzing activator function *in vivo*.

Results

Genome-wide Expression Analysis in the Absence of Thiamin

To understand how yeast cells reprogram gene expression in response to the absence of thiamin, we performed genome-wide expression analysis. Duplicate cultures of cells were grown in synthetic media with or without thiamin, RNA was isolated and labeled cDNA was prepared in the presence of fluorescently labeled nucleotides. Labeled samples were hybridized in triplicate to oligonucleotide arrays and scanned (Figure 1). An error model was used to determine which genes significantly increased or decreased in expression (Roberts et al., 2000). As summarized in Table 1, 252 genes increased in expression and 45 decreased in expression when yeast are grown in the absence of thiamin. As expected, many genes involved in thiamin biosynthesis and transport are induced (Enjo et al., 1997; Llorente et al., 1999; Nosaka et al., 1993; Nosaka et al., 1994; Praekelt et al., 1994; Schweingruber et al., 1986; Singleton, 1997). Additionally, our data agree with the findings that *PET18*, a gene whose protein is homologous to the C-terminal domains of Thi20, Thi21 and Thi22, and also *YLR004C*, which codes for a protein similar to Dal5, an allantoate permease, increase expression in the absence of thiamin (Llorente and Dujon, 2000; Llorente et al., 1999). However, we do not see increased expression of *PDC5* as previously reported, although this work relied on a lacZ reporter (Muller et al., 1999).

Our expression data also extend our knowledge of thiamin-regulated genes. Among these genes is Thi2, which codes for a potential regulator of thiamin biosynthesis (Nishimura et al., 1992a). Also included in these genes are two gene families,



B.

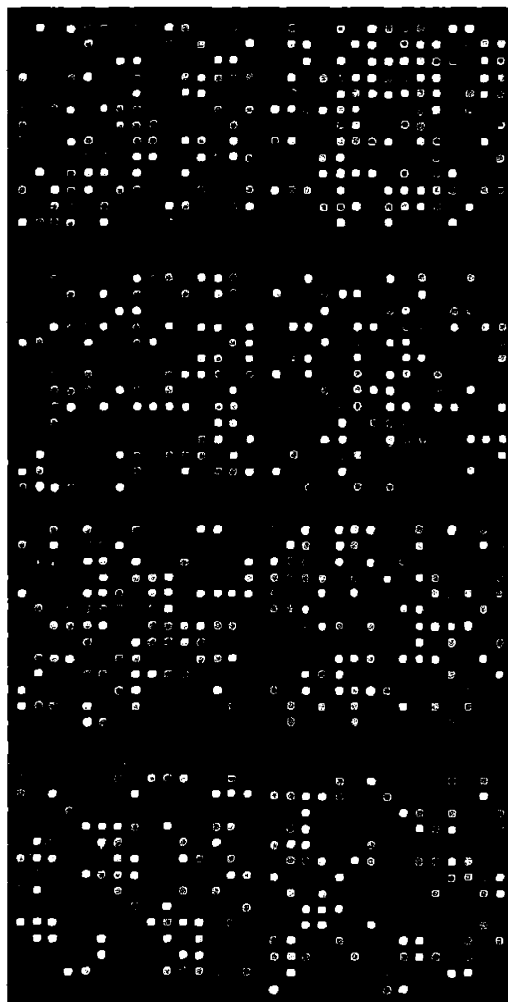


Figure 1. Genome-wide expression analysis.

A. Schematic of the experimental design.

B. A false color image of an Operon oligonucleotide array.

Table 1. Genome-wide expression analysis

Increased Expression ^a	Decreased Expression ^a	Functional Category ^b
125	0	Ribosomal Proteins
19	0	Cofactors Metabolism
14	0	RNA Pol I Transcription/Ribosome Biogenesis
10	8	Carbon Metabolism/Energy
9	2	Transcription/Splicing/Chromatin
8	3	Cell Growth/Division and DNA Synthesis
7	1	Protein Synthesis
6	1	Nucleotide Metabolism
5	0	Phosphate Metabolism
5	1	Transport
4	1	Protein Destination
4	0	Transposon and Viral-related Proteins
1	5	Amino-acid Metabolism
1	0	Ionic Homeostasis
0	4	Cell Rescue/Defense
34	19	Unknown Function/Unclassified
252	45	Total

^aGenes were considered to have changed in expression if they scored with a p-value < 0.005.

^bFunctional categories were based on the functional classification catalogue from the Comprehensive Yeast Genome Database at MIPS (Mewes et al. 2002).

SNO2/SNO3 and *SNZ2/SNZ3*, that are thought to be involved in pyridoxine biosynthesis (Ehrenshaft et al., 1999; Osmani et al., 1999). The other notable class of genes increasing in expression is the genes encoding nearly all of the ribosomal proteins and some genes involved in RNA polymerase I transcription and ribosomal biogenesis. Of the handful of genes that decrease in expression in the absence of thiamin, no functional grouping of genes predominated.

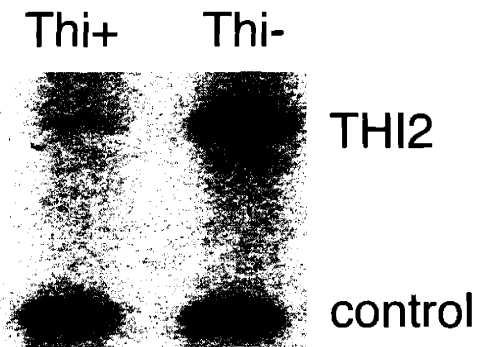
Induction of THI2

The observation from genome-wide expression analysis that *THI2* message was induced in the absence of thiamin led us to examine this more carefully. Using RT-PCR, we confirmed the induction of *THI2* mRNA in the absence of thiamin (Figure 2A). Additionally, Thi2 protein levels are correspondingly higher in the absence of thiamin. A western blot was performed with whole cell extracts of a yeast strain harboring a chromsomally integrated copy of *THI2* with 18 copies of the myc epitope fused to its C-terminus and probed with an anti-myc antibody (Figure 2B). This epitope-tagged Thi2 protein is functional as cells grown in the absence of thiamin are viable, unlike a mutant that is deleted for Thi2 (data not shown). Our observations indicate that both *THI2* mRNA and Thi2 protein levels are barely detectable when cells are supplied with exogenous thiamin, but are highly induced when thiamin is limiting.

Thi2 Activates Transcription in Vivo

Mutations in Thi2 exhibit defective expression of thiamin biosynthetic genes (Nishimura et al., 1992a; Nosaka et al., 1993; Nosaka et al., 1994). Thi2 also contains a

A.



B.

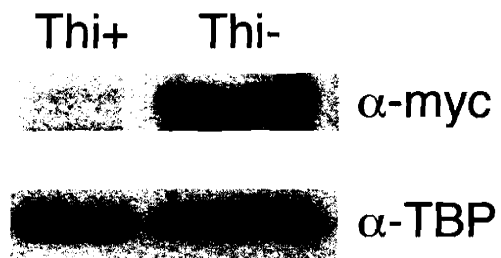


Figure 2. Thi2 induction in the absence of thiamin.

A. RT-PCR with primers for *THI2* and control mRNAs from yeast grown in the absence (Thi-) or presence (Thi+) of thiamin.

B. Western blot of whole cell extracts from a Thi2-myc strain grown in YPD (Thi+), or synthetic media lacking thiamin (Thi-). The top panel was probed with anti-myc antibody and the bottom panel was probed with an antibody against TBP, to demonstrate equal loadings of protein.

Zn₂-Cys₆ domain, present in several fungal transcriptional activators, however some proteins with this domain, such as Leu3, Arg81, and Ume6 can function as transcriptional repressors (Todd and Andrianopoulos, 1997). Therefore, we wished to determine whether Thi2 can function as a transcriptional activator *in vivo*. The *THI2* gene was fused to a LexA DNA binding domain and introduced into a strain harboring a lacZ reporter gene with 8 LexA operator sites upstream of it. Beta-galactosidase activity was measured from cells growing in mid-log phase in synthetic complete media lacking the appropriate nutrients to maintain the reporter and effector plasmids. As seen in Figure 3, beta-galactosidase activity was comparable to that induced by LexA-Gal4, and much higher than background levels of LexA alone or LexA-Gal4 in the absence of the reporter. These results indicate that Thi2 can function as a transcriptional activator *in vivo*.

Genome-wide location analysis of Thi2

The induction of Thi2 synthesis and its ability to activate transcription when tethered to a promoter suggest that Thi2 functions as a transcriptional activator. We wished to determine which promoters are bound by Thi2 *in vivo* to better characterize the role of Thi2 in the cell. Therefore, we performed genome-wide location analysis with an anti-myc antibody and extracts of our myc-tagged Thi2 strain that was grown in either rich media (YPD) or in synthetic media lacking thiamin (Thi-). Genome-wide location analysis consists of conventional chromatin immunoprecipitation (ChIP) followed by amplification of the enriched DNA by ligation-mediated PCR in the presence of fluorescently labeled nucleotides (Iyer et al., 2001; Ren et al., 2000). DNA from whole

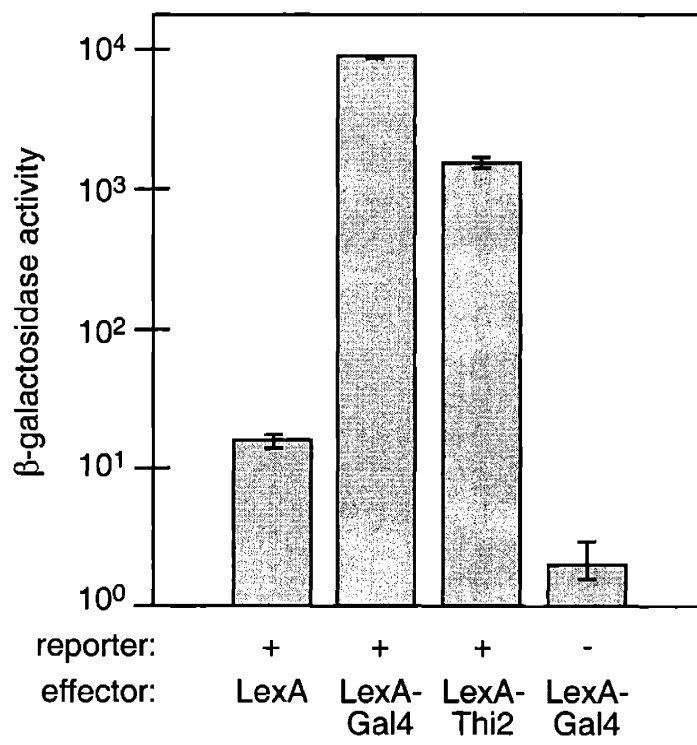


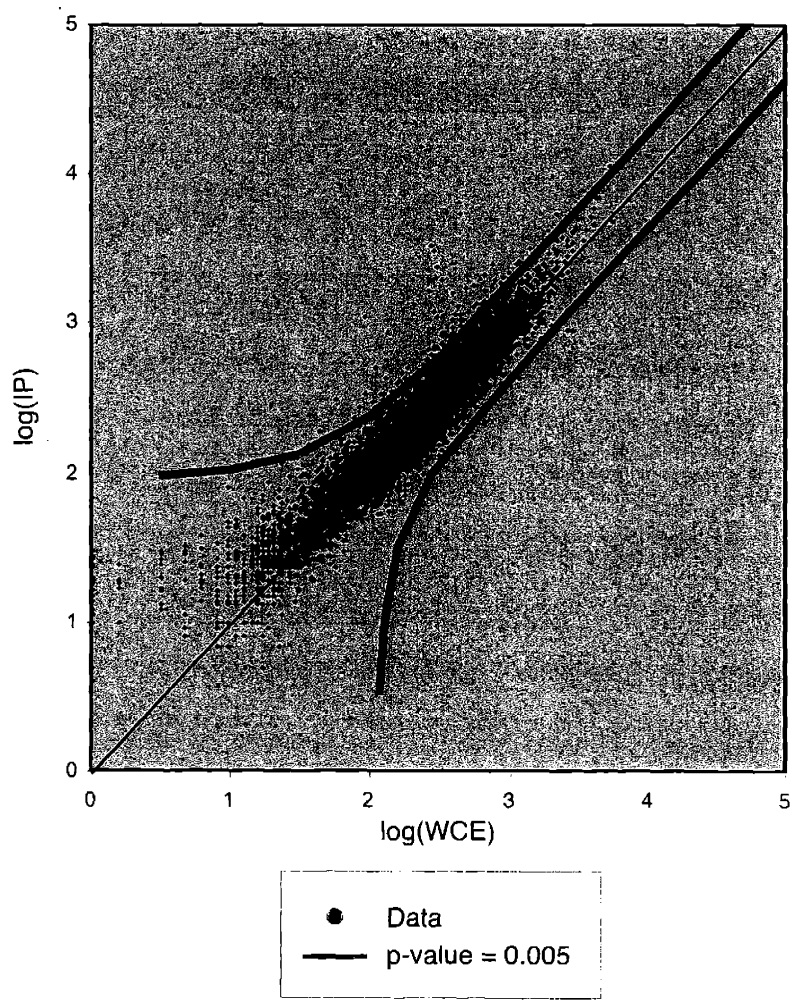
Figure 3. Thi2 activates transcription *in vivo*.

Beta-galactosidase activity, measured in Miller units (Miller, 1972), from yeast strains with 8 LexA operator sites upstream of a lacZ reporter and either LexA, LexA-Gal4, or LexA-Thi2. An additional control of LexA-Gal4 with no reporter was also included.

cell extracts is also amplified and labeled but with nucleotides containing a different fluorophore. The two pools of DNA are then hybridized to DNA arrays with every intergenic region in *S. cerevisiae* printed on them (Ren et al., 2000; Simon et al., 2001). We used a previously developed error model (Ren et al., 2000; Roberts et al., 2000), with slight modifications (see Experimental Procedures), to define promoters that are bound by Thi2 with high probability ($p < 0.005$). A representative scatter plot from a genome-wide location experiment from cells grown in Thi- media is shown in Figure 4A. One of the promoters that was bound in the genome-wide experiment, *THI80*, was chosen for verification with conventional ChIP. *THI80* promoter DNA was more highly enriched than a region of control DNA by the immunoprecipitation (Figure 4B).

Genome-wide location profiling, using a p-value < 0.005 , indicates that Thi2 binds to 23 promoters in Thi+ (YPD) media and 94 promoters in Thi- media. As ~30 genes would be expected to be bound by chance (0.5% of ~6000 total yeast genes) and Thi2 is barely detectable in Thi+ media (Figure 2), the promoters bound by Thi2 in Thi+ media likely represent noise. To enrich for the most biologically relevant targets, we focussed on those promoters that are bound by Thi2 in the absence of thiamin and their corresponding genes that show increased expression when thiamin is limiting. Many of the genes involved in thiamin synthesis have their promoters bound by Thi2, indicating that Thi2 is very likely a direct regulator of these genes (Figure 5). *PHO3*, which codes for an acid phosphatase that was previously described as being repressed by thiamin, and a homolog, *PHO5*, are also targets. Extracellular thiamin may be available as thiamin monophosphate or pyrophosphate but is imported into the cell in its phosphate-free form. Therefore if thiamin is limiting, it would benefit the cell to increase extracellular

A.



B.

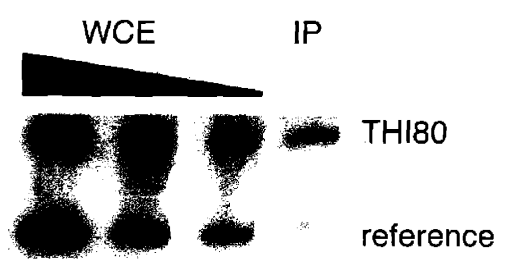


Figure 4. Genome-wide location analysis.

A. Representative results from a genome-wide location experiment. Each spot on the array is represented as a point on the graph. A thick line indicates the $p=0.005$ boundary.

B. Conventional chromatin immunoprecipitation with primers for *THI80* and a reference DNA.

YPD	Thi-	Binding	Expression (Thi-/Thi+)	Name	Binding				Expression		Description
					YPD		Thiamine-		ratio (Thi-/Thi+)	ratio	
					ratio	p-value	ratio	p-value			
				YDR438W	1.3	0.095	14	1.5x10 ⁻¹⁰	6.7	Protein of unknown function	
				FCY22	1.4	0.066	13	2.1x10 ⁻¹⁰	2.5	Purine/cytosine permease with similarity to Fcy2p	
				THI21	1.3	0.107	12	3.8x10 ⁻¹⁰	5.5	Hydroxymethylpyrimidine phosphate kinase domain fused to Pet18p-like domain	
				PHO3	0.8	0.720	6.6	1.0x10 ⁻⁸	36	Acid phosphatase, constitutive, thiamine-binding protein	
				THI80	0.9	0.619	5.7	9.2x10 ⁻⁸	3.8	Thiamine pyrophosphokinase	
				THI20	1.0	0.549	5.5	2.8x10 ⁻⁸	14	Hydroxymethylpyrimidine phosphate kinase domain fused to Pet18p-like domain	
				YLR004C	1.0	0.560	4.6	1.5x10 ⁻⁷	45	Similarity to Dal5p and other members of the allantoin permease family	
				THI6	0.8	0.826	4.2	2.9x10 ⁻⁷	7.3	Thiamine-phosphate pyrophosphorylase / hydroxyethylthiazole kinase	
				THI22	1.2	0.238	4.2	2.4x10 ⁻⁷	20	Hydroxymethylpyrimidine phosphate kinase domain fused to Pet18p-like domain	
				SNZ3	1.0	0.509	3.6	1.2x10 ⁻⁶	34	Putative pyridoxine (vitamin B6) biosynthetic enzyme	
				SNO3	1.0	0.509	3.6	1.2x10 ⁻⁶	20	Putative pyridoxine (vitamin B6) biosynthetic enzyme	
				SNZ2	1.0	0.490	3.2	4.0x10 ⁻⁶	49	Putative pyridoxine (vitamin B6) biosynthetic enzyme	
				SNO2	1.0	0.490	3.2	4.0x10 ⁻⁶	19	Putative pyridoxine (vitamin B6) biosynthetic enzyme	
				MAK32	1.8	0.005	2.7	4.8x10 ⁻⁵	2.6	Required for structural stability of L-A dsRNA-containing particles	
				PHO5	1.3	0.206	2.5	1.8x10 ⁻³	42	Acid phosphatase, repressible	
				YGR283C	1.0	0.495	2.4	1.8x10 ⁻⁴	2.7	Protein of unknown function	
				THI7	1.1	0.331	2.2	1.8x10 ⁻³	21	Thiamine transport protein	
				RPL25	0.9	0.611	2.2	5.7x10 ⁻⁴	6.8	Ribosomal protein L25	
				YDR541C	1.6	0.023	2.1	7.2x10 ⁻⁴	8	Weak similarity to dihydroflavonol-4-reductases	
				THI4	0.8	0.909	1.9	1.8x10 ⁻³	170	Thiamine-repressed protein essential for growth in the absence of thiamine	

<1/1 5/1 >10/1 <1/5 1/1 >5/1

Figure 5. Genome-wide location analysis and genome-wide expression analysis reveal biologically relevant targets of Thi2.

The genes whose promoters are bound by Thi2 and whose expression increases in the absence of thiamin are listed. Ratio is the ratio of fluorescent intensities of IP-enriched DNA over whole cell extract DNA. p-value is the probability, derived from the error model, that Thi2 is not bound the promoter indicated.

phosphatase levels to be able to import as much thiamin as possible. Included in this list are genes that have limited annotation (*YDR438W*, *YLR004C*, *YGR283C*, *YDR541C*) as well as genes that would not appear to be involved in thiamin biosynthesis (*FCY22*, *MAK32*). None of these genes was found to be essential for growth in the absence of thiamin (data not shown; Llorente and Dujon, 2000).

Thi2 Consensus Binding Site

Having determined *in vivo* targets for Thi2, it was possible to examine the promoters that are bound by Thi2 and search for a consensus sequence that is enriched in these promoters relative to the set of all promoters. Several programs have been written to facilitate this task, among which is AlignACE (Roth et al., 1998). We used AlignACE with the promoter sequences of the genes listed in Figure 5 to explore whether a binding site for Thi2 was apparent. The sequence, GGNAACYNWWAGA, found by AlignACE, was present with no more than one mismatch and in either orientation in 17 of the 20 promoters that were bound by Thi2 and whose gene was induced in the absence of thiamin (Table 2). This consensus sequence is included in a region of the *PHO3* promoter that enabled it to respond to thiamin and also caused a mobility shift in a gel retardation assay with whole cell extracts grown under conditions of limiting thiamin (Nosaka et al., 1992). Examination of the 73 genes whose promoters are bound by Thi2 but not induced in the absence of thiamin reveals that 18 genes have Thi2 consensus binding sites.

Table 2. A consensus binding site for Thi2

Gene ^a	Site ^b
FCY22	AGTAACTCTTAGA
PHO3	GGAAACTCAAAGA
MAK32	GGAAACAAAAAGA
RPL25	AGAAACTTAAAGA
RPL25	GGAAACCGTTAGA
SNO2/SNZ2	GGAAACTGTAAGA
SNO3/SNZ3	GGAAACTGTAAGA
THI20	GGAAACCCTTAGA
THI21	GGAAACCCTTAGA
THI22	GGAAACCTTTAGA
THI4	GGTAACTGATAGA
THI6	GGCAATCGTAAGA
THI7	GGCAACCTCTAGA
THI80	GGCAACTATTAGA
THI80	AGAAATTTGTAGA
YDR438W	AGTAACTATTAGA
YLR004C	GGAAACTCAAAGA
consensus	GG-AACY-WWAGA

^aGenes whose promoters were bound by Thi2 in thiamine-media and were induced in the absence of thiamine

^bPotential Thi2 binding sites in the promoter of the indicated gene

Discussion

Using genome-wide expression analysis, we have determined how yeast reprogram gene expression in the absence of thiamin. We have also shown that Thi2, a protein implicated in thiamin biosynthetic gene regulation is itself induced in the absence of thiamin, acts as a transcriptional activator *in vivo*, and binds to the promoters of many of the genes involved in thiamin biosynthesis. A putative binding site for Thi2 appears in many of the promoters that appear to be *in vivo* targets of Thi2.

As expected, the absence of thiamin causes all of the genes implicated in thiamin biosynthesis to be induced. In addition to thiamin biosynthetic genes, two gene families, *SNO2/3* and *SNZ2/3*, that have been implicated in pyridoxine biosynthesis are also induced. As pyridoxine is a precursor molecule used in the biosynthesis of thiamin (Figure 6), it would be appropriate for the cell to increase pyridoxine synthesis in the absence of thiamin. This observation is consistent with the observation that synthesis of pyridoxine in *Saccharomyces uvarum* is inhibited by the presence of thiamin in the growth media (Minami et al., 1982).

It is puzzling why the cell would increase expression of nearly every subunit of the ribosome in the absence of thiamin. Examination of many other genome-wide expression datasets indicates that cells may decrease expression of ribosomal subunits under stress conditions and increase them when cells adjust to the stress or return to more favorable conditions (Causton et al., 2001; Gasch et al., 2000). It would not seem that the absence of thiamin would be a favorable condition for the cell, although this may reflect our lack of understanding of the biology behind nutrient uptake and biosynthesis.

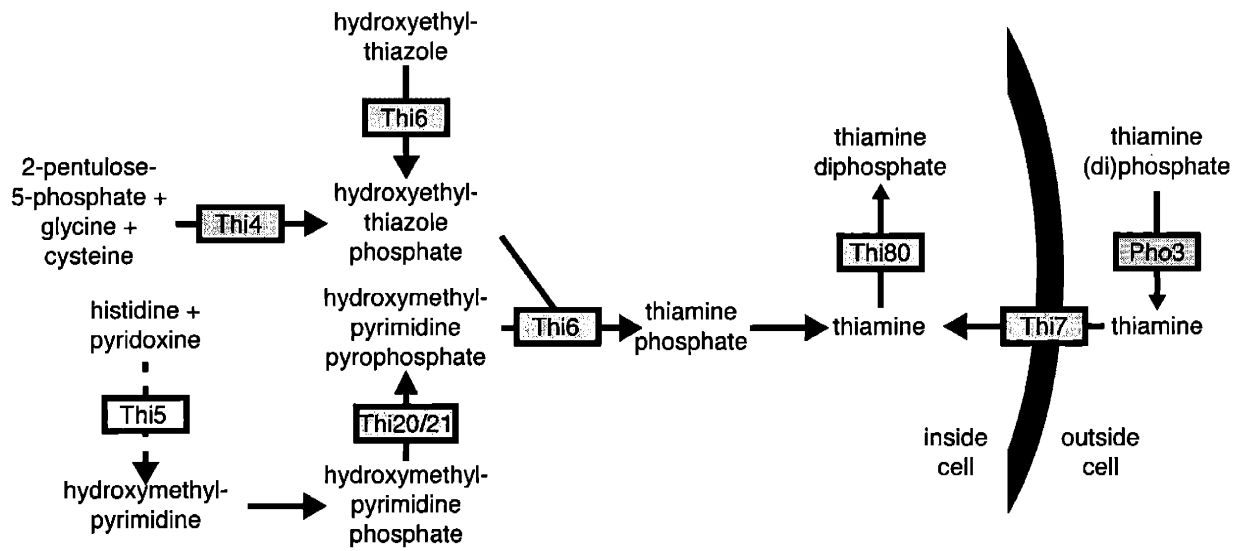


Figure 6. Thiamin biosynthesis is regulated by Thi2.

A metabolic chart depicting thiamin biosynthesis (Hohmann and Meacock, 1998). All genes are induced in the absence of thiamin. Genes that are colored are bound by Thi2 in the absence of thiamin.

Whether the increased expression of ribosomal genes is carried out by Thi2, is not clear. One of the genes encoding ribosomal subunits appear to be directly regulated by Thi2, although there are 16 additional ribosomal protein genes that have a putative Thi2 binding site (with one mismatch to the consensus allowed) in their promoter. It is possible that the cell is able to sense an imbalance in the availability of ribosomal subunits and adjusts the levels of other subunits accordingly.

One of the other genes that is induced in the absence of thiamin is *ILV5*, which catalyzes the second step in isoleucine and valine biosynthesis. The first step in this biosynthetic pathway is carried out by *Ilv2*, acetolactate synthase, which uses thiamin pyrophosphate as a cofactor. It is likely that the amount of thiamin available to the cells in these experiments is greater than the amount that they produce in the absence of exogenous thiamin (Byrne and Meacock, 2001). Therefore, *Ilv2* activity may be limited by the amount of available thiamin, and consequently the cell may want to increase the amount of *Ilv5* available to continue to drive the reactions of branched-chain amino acid synthesis forward.

Among the previously implicated regulators of thiamin biosynthetic genes, only *THI2* itself is induced – neither *THI3* nor *PDC2* are induced, although they may be activated post-transcriptionally. From our experiments here, it appears that Thi2 is the regulator of thiamin biosynthetic genes at the transcriptional level. Thi3 has a thiamin binding domain (Dickinson et al., 1997), and Pdc2 regulates the activity of pyruvate decarboxylases (Hohmann, 1993), which have thiamin binding domains (Green, 1989). Therefore, Thi3 and Pdc2 may be involved in thiamin concentration sensing and transmitting the signal for activation of *THI2* transcription.

There appear to be several different mechanisms of activating transcriptional regulators. An activator such as Gal4 is always present, but its activity is repressed by a protein-protein interaction with Gal80 (Johnston, 1987). Phosphorylation may cause the regulator to be activated as is the case with Ste12 (Song et al., 1991). Msn2 resides in the cytosol but is translocated to the nucleus when its activity is required (Gorner et al., 1998). *GCN4* is transcribed continuously, but its translation is inhibited when it is not needed (Hinnebusch, 1996). From our work here, we find that Thi2 falls into the class of transcriptional regulators that includes Ime1 and are themselves transcriptionally induced when their function is required (Kassir et al., 1988). It is interesting that the cell has evolved different mechanisms of regulating transcriptional regulators. It is possible that these mechanisms reflect the urgency of the transcriptional response required to respond to a signal.

The results from genome-wide location analysis of Thi2 and the genome-wide expression analysis of cells growing in the absence or presence of thiamin indicate that Thi2 is a direct regulator of the genes in the thiamin biosynthetic pathway (Figure 6). Interestingly, we do not see binding of Thi2 to the promoters of *THI5*, and its homologs *THI11*, *THI12* and *THI13*. The Operon oligonucleotide arrays we used for genome-wide expression analysis can not distinguish among these four genes as they are identical, but we see on average 85 fold more signal from the thiamin- channel than the thiamin+ channel for the spots that probe for *THI5*, *THI12* and *THI13*. It appears that there was no DNA spotted on the arrays for the *THI11* probe. These genes may not be directly regulated by Thi2 and instead induced by another transcriptional regulator. However, *THI5* and *THI12* are immediately downstream of *SNZ3* and *SNZ2*, respectively, which are

induced by the absence of thiamin and have promoter regions that are bound by Thi2. Therefore it is tempting to speculate that Thi2 is still activating the transcription of *THI5* and *THI12*, but from a distance. *THI11* and *THI13* may either not be expressed in the absence of thiamin or may be activated by another regulator.

In addition to the genes known to be involved in thiamin biosynthesis, there are other genes that have promoters that are bound by Thi2 and are induced in the absence of thiamin. Two of these genes, *FCY22* and *YLR004C*, encode permeases that may be used for importing molecules that are precursors to thiamin (Paulsen et al., 1998). However, neither of these genes is essential for viability in the absence of thiamin indicating that if they do import thiamin precursors, there are other permeases that can import the necessary molecules in their absence.

By defining a set of high confidence *in vivo* targets for Thi2, we could use AlignACE to search for a sequence motif that occurred in most of the promoters bound by Thi2. The sequence we report, GGNAACYNWWAGA, is shorter than the 20 base pair region in the *PHO3* promoter that was found to mediate repression by thiamin (Nosaka et al., 1992), and larger than the 7 base pair site that (Hohmann and Meacock, 1998) report to find in the promoters of several thiamin biosynthetic genes. Orthologs were found in *Saccharomyces paradoxus* for 13 of the genes that are induced in the absence of thiamin and whose promoters are bound by Thi2. The proposed Thi2 binding site is completely conserved in 11 of these orthologous promoters, consistent with the notion that this site is a relevant regulatory sequence (M. Kamvysselis, pers. comm.).

A consensus site is not present in three promoters that are potentially regulated by Thi2. However, of these promoters, two of the promoters had a consensus site with two

mismatches. Why we were not able to find this site in the remaining target for Thi2 is unclear. 473 genes have a Thi2 consensus site with no more than one mismatched base in their upstream intergenic region. As it is unlikely that most of these promoters are actually recognized by Thi2, chromatin structure or some other mechanism must be preventing Thi2 from binding to these sites.

Inspection of the consensus sequence we present does not reveal an obvious symmetry. Many other proteins with Zn₂-Cys₆ domains that have been studied bind to symmetrical sites and dimerize through a coiled coils (Todd and Andrianopoulos, 1997). Sequence analysis of Thi2 with MultiCoil does not reveal a significant match to a coiled coil (J. Newman pers. comm.; Wolf et al., 1997). This observation and the structure of the consensus binding site imply that Thi2 recognizes DNA as a monomer.

A remaining question from this study is how is Thi2 induced in the absence of thiamin. From a broad study using genome-wide location analysis with over 100 transcriptional regulators, no regulators were found to bind to the *THI2* promoter with high confidence (Lee et al., 2002). It is possible that Thi2 expression increases in the absence of thiamin due to an activator increasing its transcription or due to the inactivation of a repressor that keeps *THI2* from being transcribed in the presence of thiamin. A genetic screen could be devised to uncover candidate regulators for either scenario: mutants could be recovered that have increased *THI2* expression in the presence of thiamin, or that lack *THI2* expression in the absence of thiamin. Some work has begun in this area, although mutants were isolated but not identified that had derepressed gene expression of thiamin biosynthetic genes, not regulators of thiamin biosynthesis (Burrows et al., 2000). However, as more genome-wide expression and location experiments are

performed and their datasets become available, it is perhaps more likely that a clue to the identity of a regulator of Thi2 will arise from this approach.

Experimental Procedures

Microarray production

Operon oligonucleotide arrays were constructed using DNA from the Operon Yeast Genome Oligo Set, which contains 6307 70mer oligonucleotides. Approximately 1 nl of DNA was printed from 1 mg/ml stock solutions in 3X SSC, 1.5 M betaine on GAPS-II aminosilane coated slides (Corning) using a Cartesian PixSys robot. Printed slides were baked for 2 hr at 80°C, UV crosslinked in a UV Stratalinker (Stratagene) at 300 mJoules, and stored under vacuum desiccation until use. Intergenic arrays were constructed as described in Part I of this chapter.

Genome-wide expression analysis

Duplicate cultures of strain W303 strain Z1256 (*MATa*, *ade2-1*, *trp1-1*, *can1-100*, *leu2-3,112*, *his3-11,15*, *ura3*) were grown at 30°C in complete synthetic media lacking thiamin (QBiogene) or the same media supplemented with 400 µg/L thiamin (Sigma) until cells reached OD₆₀₀=0.6-0.8. Cells were harvested by centrifugation, media was decanted and cells were frozen in liquid nitrogen and stored at -80°C. RNA was prepared by the hot phenol extraction method (Ausubel et al., 1994) and mRNA was isolated using Oligotex beads (Qiagen). Labeled cDNA were prepared by combining 2.5 µg of mRNA and 2 µg oligo dT primer in a total reaction volume of 40 µL containing 25 µM of dA,C,GTP, 12.5 µM of dTTP, 400 U Superscript II reverse transcriptase (Invitrogen) and either 12.5 µM Cy3-dUTP or 25 µM of Cy5-dUTP. The reaction was incubated at 23°C for 10 minutes and then at 42°C 2 hours. RNA was digested with 2 U of RNase H and

1.5 U of RNase A at 37°C for 15 minutes. DNA was purified with QIAquick PCR purification columns (Qiagen) and eluted into 50 µL 10 mM Tris pH 8.0. 20 pmol of each dye-labeled sample, which may have been pooled from multiple labeling reactions, was placed on each array. Samples were combined with 5 µg sonicated salmon sperm DNA, 10 µg tRNA, and 20 µg polyA in prehybridization solution (25% formamide, 5X SSC and 0.1% SDS) and hybridized to Operon oligonucleotide arrays that had been in prehybridization solution for 45 minutes at 42°C. For each of the two cultures, three hybridizations were performed at 42°C for 16-20 hr. As a control, two additional hybridizations were performed with two samples of the same mRNA that were labeled with Cy3 and Cy5. Slides were washed in 2X SSC, 0.1% SDS for 5 minutes at 42°C, with additional washes in 0.1X SSC, 0.1% SDS at room temperature for 10 minutes, and then three times in 0.1X SSC at room temperature for one minute before drying under nitrogen. Slides were scanned (Axon) and spot intensities were quantified (GenePix).

Samples from each pair of cultures were hybridized in triplicate, and a single-array error model was used to handle noise, to average repeated experiments with appropriate weights, and to rank binding sites by p-value (Roberts et al., 2000). P-values from both culture replicates were combined using the method of Fisher. Genes that scored with a p-value less than or equal to 0.005 were considered to have changed expression significantly in the experiment.

Beta-galactosidase assay

A LexA-Thi2 expression vector was constructed by using PCR with *S. cerevisiae* genomic DNA to clone the *THI2* coding region and inserting it in frame into the multiple

cloning site of pLexA (Clontech). All LexA fusions and the p8op-lacZ reporter (Clontech) were transformed into strain W303 (Z1256) and grown in synthetic media lacking uracil and histidine. Beta-galactosidase assays were performed in triplicate and as described (Rose and Botstein, 1983).

Epitope tagging, chromatin immunoprecipitation and genome-wide location analysis

Procedures were performed as described in Part I of this chapter with the following exceptions. A Thi2-myc strain (Z1521) was constructed and used for chromatin immunoprecipitation. The tagged strain was grown in either YPD or complete synthetic media lacking thiamin. Intergenic arrays were scanned with an Axon scanner using GenePix software. Binding ratios were mean centered based on over 400 genome-wide location hybridizations (N. Rinaldi and R. Young, unpublished) and were used with the error model to generate probabilities that binding to a particular spot did not occur.

RT-PCR and western blotting

Whole cell extracts of strain Z1256 grown in YPD or complete synthetic media lacking thiamin were prepared and used for RT-PCR in the presence of a radiolabeled nucleotide with the appropriate primers. For western blotting, whole cell extracts from strain Z1521 grown in YPD or complete synthetic media lacking thiamin were prepared and loaded onto a 4-15% tris-glycine polyacrylamide gel. The gel was run at 100V and proteins were transferred to nitrocellulose. The membrane was blocked with 5% milk in TBST overnight at 4°C and probed with the 9E11 antibody that recognizes the myc

epitope, followed by an anti-mouse antibody conjugated to horseradish peroxidase.

Detection was performed with the ECL system (Amersham).

References

- Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., and Struhl, K. (1994). *Current Protocols in Molecular Biology*, Current Protocols).
- Begley, T. P. (1996). The biosynthesis and degradation of thiamin (vitamin B1). *Nat Prod Rep* 13, 177-185.
- Burrows, R. J., Byrne, K. L., and Meacock, P. A. (2000). Isolation and characterization of *Saccharomyces cerevisiae* mutants with derepressed thiamine gene expression. *Yeast* 16, 1497-1508.
- Byrne, K. L., and Meacock, P. A. (2001). Thiamin auxotrophy in yeast through altered cofactor dependence of the enzyme acetohydroxyacid synthase. *Microbiology* 147, 2389-2398.
- 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.
- Dickinson, J. R., Lanterman, M. M., Danner, D. J., Pearson, B. M., Sanz, P., Harrison, S. J., and Hewlins, M. J. (1997). A ¹³C nuclear magnetic resonance investigation of the metabolism of leucine to isoamyl alcohol in *Saccharomyces cerevisiae*. *J Biol Chem* 272, 26871-26878.
- Ehrenshaft, M., Bilski, P., Li, M. Y., Chignell, C. F., and Daub, M. E. (1999). A highly conserved sequence is a novel gene involved in de novo vitamin B6 biosynthesis. *Proc Natl Acad Sci U S A* 96, 9374-9378.
- Enjo, F., Nosaka, K., Ogata, M., Iwashima, A., and Nishimura, H. (1997). Isolation and characterization of a thiamin transport gene, *THI10*, from *Saccharomyces cerevisiae*. *J Biol Chem* 272, 19165-19170.
- Estramareix, B., and David, S. (1996). Biosynthesis of thiamine. *New J Chem* 20, 607-629.
- 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.
- Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., Cruciat, C. M., *et al.* (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141-147.
- 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.

Green, J. B. (1989). Pyruvate decarboxylase is like acetolactate synthase (ILV2) and not like the pyruvate dehydrogenase E1 subunit. *FEBS Lett* 246, 1-5.

Hinnebusch, A. G. (1996). Translational control of *GCN4*: gene-specific regulation by phosphorylation of eIF2. In *Translational Control*, J. W. B. Hershey, M. B. Mathews, and N. Sonenberg, eds. (Plainview, N.Y., Cold Spring Harbor Laboratory Press), pp. 199-244.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S. L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., *et al.* (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180-183.

Hohmann, S. (1991). Characterization of PDC6, a third structural gene for pyruvate decarboxylase in *Saccharomyces cerevisiae*. *J Bacteriol* 173, 7963-7969.

Hohmann, S. (1993). Characterisation of PDC2, a gene necessary for high level expression of pyruvate decarboxylase structural genes in *Saccharomyces cerevisiae*. *Mol Gen Genet* 241, 657-666.

Hohmann, S., and Cederberg, H. (1990). Autoregulation may control the expression of yeast pyruvate decarboxylase structural genes PDC1 and PDC5. *Eur J Biochem* 188, 615-621.

Hohmann, S., and Meacock, P. A. (1998). Thiamin metabolism and thiamin diphosphate-dependent enzymes in the yeast *Saccharomyces cerevisiae*: genetic regulation. *Biochim Biophys Acta* 1385, 201-219.

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.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* 98, 4569-4574.

Iwashima, A., Nishino, H., and Nose, Y. (1973). Carrier-mediated transport of thiamine in baker's yeast. *Biochim Biophys Acta* 330, 222-234.

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.

Johnston, M. (1987). A model fungal gene regulatory mechanism: the GAL genes of *Saccharomyces cerevisiae*. *Microbiol Rev* 51, 458-476.

- Kassir, Y., Granot, D., and Simchen, G. (1988). IME1, a positive regulator gene of meiosis in *S. cerevisiae*. *Cell* 52, 853-862.
- Krishna, S., Taylor, A. M., Supanaranond, W., Pukrittayakamee, S., ter Kuile, F., Tawfiq, K. M., Holloway, P. A., and White, N. J. (1999). Thiamine deficiency and malaria in adults from southeast Asia. *Lancet* 353, 546-549.
- Lee, T.I., Rinaldi, N.J., Robert, F., *et al.* (2002). A transcriptional regulatory network map for *Saccharomyces cerevisiae*. In preparation.
- Llorente, B., and Dujon, B. (2000). Transcriptional regulation of the *Saccharomyces cerevisiae* DAL5 gene family and identification of the high affinity nicotinic acid permease TNA1 (YGR260w). *FEBS Lett* 475, 237-241.
- Llorente, B., Fairhead, C., and Dujon, B. (1999). Genetic redundancy and gene fusion in the genome of the Baker's yeast *Saccharomyces cerevisiae*: functional characterization of a three-member gene family involved in the thiamine biosynthetic pathway. *Mol Microbiol* 32, 1140-1152.
- Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkötter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 30, 31-34.
- Miller, J. H. (1972). *Experiments in Molecular Genetics* (Cold Spring Harbor, New York, Cold Spring Harbor Laboratory).
- Minami, J., Kishi, T., and Kondo, M. (1982). Effects of thiamin on vitamin B6 synthesis in yeasts. *J Gen Microbiol* 128, 2909-2917.
- Muller, E. H., Richards, E. J., Norbeck, J., Byrne, K. L., Karlsson, K. A., Pretorius, G. H., Meacock, P. A., Blomberg, A., and Hohmann, S. (1999). Thiamine repression and pyruvate decarboxylase autoregulation independently control the expression of the *Saccharomyces cerevisiae* PDC5 gene. *FEBS Lett* 449, 245-250.
- Muri, R. M., Von Overbeck, J., Furrer, J., and Ballmer, P. E. (1999). Thiamin deficiency in HIV-positive patients: evaluation by erythrocyte transketolase activity and thiamin pyrophosphate effect. *Clin Nutr* 18, 375-378.
- Newman, J. R., Wolf, E., and Kim, P. S. (2000). A computationally directed screen identifying interacting coiled coils from *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A* 97, 13203-13208.
- Nishimura, H., Kawasaki, Y., Kaneko, Y., Nosaka, K., and Iwashima, A. (1992a). Cloning and characteristics of a positive regulatory gene, THI2 (PHO6), of thiamin biosynthesis in *Saccharomyces cerevisiae*. *FEBS Lett* 297, 155-158.

- Nishimura, H., Kawasaki, Y., Kaneko, Y., Nosaka, K., and Iwashima, A. (1992b). A positive regulatory gene, *THI3*, is required for thiamine metabolism in *Saccharomyces cerevisiae*. *J Bacteriol* 174, 4701-4706.
- Nosaka, K., Kaneko, Y., Nishimura, H., and Iwashima, A. (1993). Isolation and characterization of a thiamin pyrophosphokinase gene, *THI80*, from *Saccharomyces cerevisiae*. *J Biol Chem* 268, 17440-17447.
- Nosaka, K., Nishimura, H., Kawasaki, Y., Tsujihara, T., and Iwashima, A. (1994). Isolation and characterization of the *THI6* gene encoding a bifunctional thiamin-phosphate pyrophosphorylase/hydroxyethylthiazole kinase from *Saccharomyces cerevisiae*. *J Biol Chem* 269, 30510-30516.
- Nosaka, K., Yamanishi, K., Nishimura, H., and Iwashima, A. (1992). Upstream activation element of the *PHO3* gene encoding for thiamine-repressible acid phosphatase in *Saccharomyces cerevisiae*. *FEBS Lett* 305, 244-248.
- Osmani, A. H., May, G. S., and Osmani, S. A. (1999). The extremely conserved *pyroA* gene of *Aspergillus nidulans* is required for pyridoxine synthesis and is required indirectly for resistance to photosensitizers. *J Biol Chem* 274, 23565-23569.
- Paulsen, I. T., Sliwinski, M. K., Nelissen, B., Goffeau, A., and Saier, M. H., Jr. (1998). Unified inventory of established and putative transporters encoded within the complete genome of *Saccharomyces cerevisiae*. *FEBS Lett* 430, 116-125.
- Poulsen, C., and Stougaard, P. (1989). Purification and properties of *Saccharomyces cerevisiae* acetolactate synthase from recombinant *Escherichia coli*. *Eur J Biochem* 185, 433-439.
- Prækel, U. M., Byrne, K. L., and Meacock, P. A. (1994). Regulation of *THI4* (*MOL1*), a thiamine-biosynthetic gene of *Saccharomyces cerevisiae*. *Yeast* 10, 481-490.
- 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.
- Repetto, B., and Tzagoloff, A. (1989). Structure and regulation of *KGD1*, the structural gene for yeast alpha-ketoglutarate dehydrogenase. *Mol Cell Biol* 9, 2695-2705.
- 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.
- Rose, M., and Botstein, D. (1983). Construction and use of gene fusions to *lacZ* (beta-galactosidase) that are expressed in yeast. *Methods Enzymol* 101, 167-180.

- 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.
- Schaaff-Gerstenschlager, I., Mannhaupt, G., Vetter, I., Zimmermann, F. K., and Feldmann, H. (1993). TKL2, a second transketolase gene of *Saccharomyces cerevisiae*. Cloning, sequence and deletion analysis of the gene. *Eur J Biochem* *217*, 487-492.
- Schweingruber, M. E., Fluri, R., Maundrell, K., Schweingruber, A. M., and Dumermuth, E. (1986). Identification and characterization of thiamin repressible acid phosphatase in yeast. *J Biol Chem* *261*, 15877-15882.
- 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.
- Singleton, C. K. (1997). Identification and characterization of the thiamine transporter gene of *Saccharomyces cerevisiae*. *Gene* *199*, 111-121.
- Song, D., Dolan, J. W., Yuan, Y. L., and Fields, S. (1991). Pheromone-dependent phosphorylation of the yeast STE12 protein correlates with transcriptional activation. *Genes Dev* *5*, 741-750.
- Steensma, H. Y., Holterman, L., Dekker, I., van Sluis, C. A., and Wenzel, T. J. (1990). Molecular cloning of the gene for the E1 alpha subunit of the pyruvate dehydrogenase complex from *Saccharomyces cerevisiae*. *Eur J Biochem* *191*, 769-774.
- Sundstrom, M., Lindqvist, Y., Schneider, G., Hellman, U., and Ronne, H. (1993). Yeast TKL1 gene encodes a transketolase that is required for efficient glycolysis and biosynthesis of aromatic amino acids. *J Biol Chem* *268*, 24346-24352.
- Todd, R. B., and Andrianopoulos, A. (1997). Evolution of a fungal regulatory gene family: the Zn(II)₂Cys₆ binuclear cluster DNA binding motif. *Fungal Genet Biol* *21*, 388-405.
- Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., Bader, G. D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C. W., Bussey, H., *et al.* (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* *294*, 2364-2368.
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., *et al.* (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* *403*, 623-627.
- Winzeler, E. A., Liang, H., Shoemaker, D. D., and Davis, R. W. (2000). Functional analysis of the yeast genome by precise deletion and parallel phenotypic characterization. *Novartis Found Symp* *229*, 105-109.

Winzler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., *et al.* (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901-906.

Wolf, E., Kim, P. S., and Berger, B. (1997). MultiCoil: a program for predicting two- and three-stranded coiled coils. *Protein Sci* 6, 1179-1189.

Chapter 4

GDS: A Graphical Display Suite for Visualization of Yeast Microarray Data

My Contributions to this Project

I created the framework for the Graphical Display Suite and wrote the entire user interface for it. I designed and wrote most of the programs with contributions from Jianxin Xie and John Barnett. The genome-wide location profiling experiments were performed by various members of the Young lab and me, and I performed the multi-subunit complex analysis with some assistance from Nicola Rinaldi.

Summary

The use of microarray technology has created new challenges in data analysis for biologists. Visual displays can greatly facilitate the analysis and communication of large quantities of data. We have created a Graphical Display Suite (GDS) that consists of a collection of tools to assist in the visualization of data from genome-wide experiments in *S. cerevisiae*. The GDS is web-accessible, easy to use, and additional components can easily be incorporated into its interface. This suite of tools has proven to be useful in revealing important biological insights.

Introduction

Biologists have traditionally conducted experiments yielding a single or a handful of data points. These results are communicated by showing annotated raw data (e.g. an autoradiogram) or with conventional graphs. The advent of microarray technology has resulted in the production of large quantities of data. Analysis of microarray data can be challenging and cumbersome (Quackenbush, 2001). Although experiments are usually performed to answer a specific question, there may be answers to questions in the data that were not originally posed. Displaying the results of microarray experiments as large tables of numbers is not only not practical due to space constraints in traditional paper publications, but is also not very informative to the reader. Both of these issues can be addressed by the use of graphical displays. A well-designed figure can allow the scientist to quickly identify unexpected trends or structure in the data. This same figure can also communicate a conclusion from many data points much more effectively than the raw numbers themselves.

There are several programs available to facilitate microarray data analysis and visualization. GeneSpring and SpotFire are examples of commercially available software packages. TreeView is a popular program that performs hierarchical clustering and creates a graphical output (Eisen et al., 1998). All of these programs are stand-alone software packages that are dependent on a particular computing platform, thereby limiting their use to the computers on which the programs are installed. Programs such as GeneSpring and SpotFire are powerful, but may be too costly for some researchers. The development of web-based graphical display software that is free and can be run

from any computer connected to the Internet, is advantageous to the growing community of researchers analyzing microarray data.

S. cerevisiae is an excellent system for conducting genome-wide experiments. The yeast genome, compared to the genomes of other organisms, is very well annotated and thereby lends itself to genome-wide experiments for several reasons. Researchers have high confidence in what are described as open reading frames (ORFs) are indeed genes and therefore have high confidence that the probes being used are probing the appropriate ORF. The genome is of low complexity so probe designers do not have to rely on clone sets that may have misidentified clones to develop cDNA probes. The sequence is complete and the hypothetical functions for over half the ORFs in the genome have been described. All published and some unpublished information on each gene and its product is available through curated web sites (Dwight et al., 2002; Mewes et al., 2002). Additionally each candidate ORF has a systematically defined name, which facilitates comparison of data across experiments.

As microarray experiments produce large quantities of data, it can be useful to analyze the data in ways not originally intended when the experiment was designed. Unexpected connections in the data can lead to valuable insights. The connections may be observed by comparing the data with other microarray data sets or functional classifications of genes. Communicating the insights gained from microarray experiments may require creating new types of displays when simple graphs or charts are insufficient. We developed the Graphical Display Suite (GDS) described here to facilitate the analysis and communication of data generated from genome-wide experiments with the yeast, *S. cerevisiae*. A description of each of the tools is presented below.

Implementation

Overview

GDS is a web-accessible application (<http://web.wi.mit.edu/young/gds/>) and displays graphics in the scalable vector graphics (SVG) format. SVG is an XML-based graphics format that describes graphical elements as vectors rather than pixels. The advantage of vector graphics is that when they are scaled they do not suffer from the loss of resolution that pixelated graphics do. Since it is XML-based, it is theoretically readable by humans so that modifications to the graphics can be made in the absence of a drawing program. Text embedded in SVG is displayed as text characters in the specified font and not as a collection of pixels. Consequently SVG text can be copied and pasted elsewhere as well as searched for in browsers and by search engines. SVG also supports scripting events allowing for the graphics to be interactive with the user. Currently, SVG support is not built into web browsers. However, Adobe is one of the proponents of SVG and their SVG viewer is being bundled with their Acrobat Reader which is used to display the ubiquitous Portable Document Format (PDF) files. Increasing numbers of drawing programs support exporting of SVG, which should further popularize this format. GDS was written in Java and uses the Batik class library from the XML project of the Apache group to create and modify SVG files. An SVG browser that is distributed with Batik can display SVG images and also convert them to the popular JPEG format.

The input data for the GDS are not raw data, but rather lists of genes. We have chosen to separate GDS from the initial steps of data analysis and instead have its input be lists of genes for maximal flexibility. Lists of genes would typically be taken from

microarray analysis (e.g. the list of genes that increase in expression by 2-fold or more during heat shock) but are not limited to microarray results. Lists of genes are entered through the web interface or a file containing lists of genes can be uploaded. The format of this file is described in the online help. Genes can be entered as standard ORF names (e.g. YAL001C) or as Saccharomyces Genome Database (SGD) names (Dwight et al., 2002). The user will be notified if invalid gene names were entered and those genes will be removed from the list. Additionally, intergenic region names can be entered. These names consist of an 'i' followed by a standard ORF name, and correspond to the region to the right of the standard ORF name specified (Ren et al., 2000). The user can assign two feature colors and a text color to every gene list. Any shape or other graphical feature in the display that corresponds to a gene found in a particular gene list will be assigned the feature color of that gene list. Similarly, any text associated with the gene in the display will be colored the text color of that gene list. Sixteen standard colors are provided, but the full palette of RGB colors is available. Multiple lists of genes can be entered sequentially with their own feature and text colors. A default or uncolored feature and text color can also be selected. These colors are applied to all features and text elements in the displays before any genes are matched.

Several tools make up the GDS. Each tool creates a different graphical display and colors portions of the display based upon user input. The tools included are Biological Mapper, two styles of Chromosome Displays, Cell Cycle Display, Expression Data Mapper, Transcriptome Bins, Functional Category Circles, Regulator Network Wheel, and Venn Diagram Drawer. Each tool allows the user to relate genome-wide data, in the form of lists of genes, to other types of data such as genome-wide expression

data, genomic structure data, and protein functional categorization data that are built into the program. Some tools rely on genes being present in only one gene list. If genes appear in multiple lists, the user will be prompted to choose a feature color and text color to represent these genes, regardless of how many lists they appear in. These colors will be used in the tools (e.g. chromosome displays, biological mapper) that require genes to appear in only one list.

Biological Mapper

The Biological Mapper tool uses predrawn SVG files that typically contain graphical depictions of metabolic or signal transduction pathways as templates, but can be any drawing that contains graphical representations or names of genes (Figure 1). These files can be created by anyone using a graphical editor that can export SVG, such as Adobe Illustrator, and then uploaded to GDS. The genes that are present in the template are compared with those that are in the gene lists. The genes in the template are colored according to the gene list in which they are found, and the genes not found are colored the default color. This tool provides a convenient means for determining if a group of genes are involved in a variety of biological processes. KEGG also provides this function with an extensive collection of pathways, but the user is limited to a single gene list being displayed on a pathway. Additionally, the pathways depicted are generic and not specific to yeast and therefore may provide potentially confusing diagrams (Kanehisa et al., 2002).

Chromosome Display

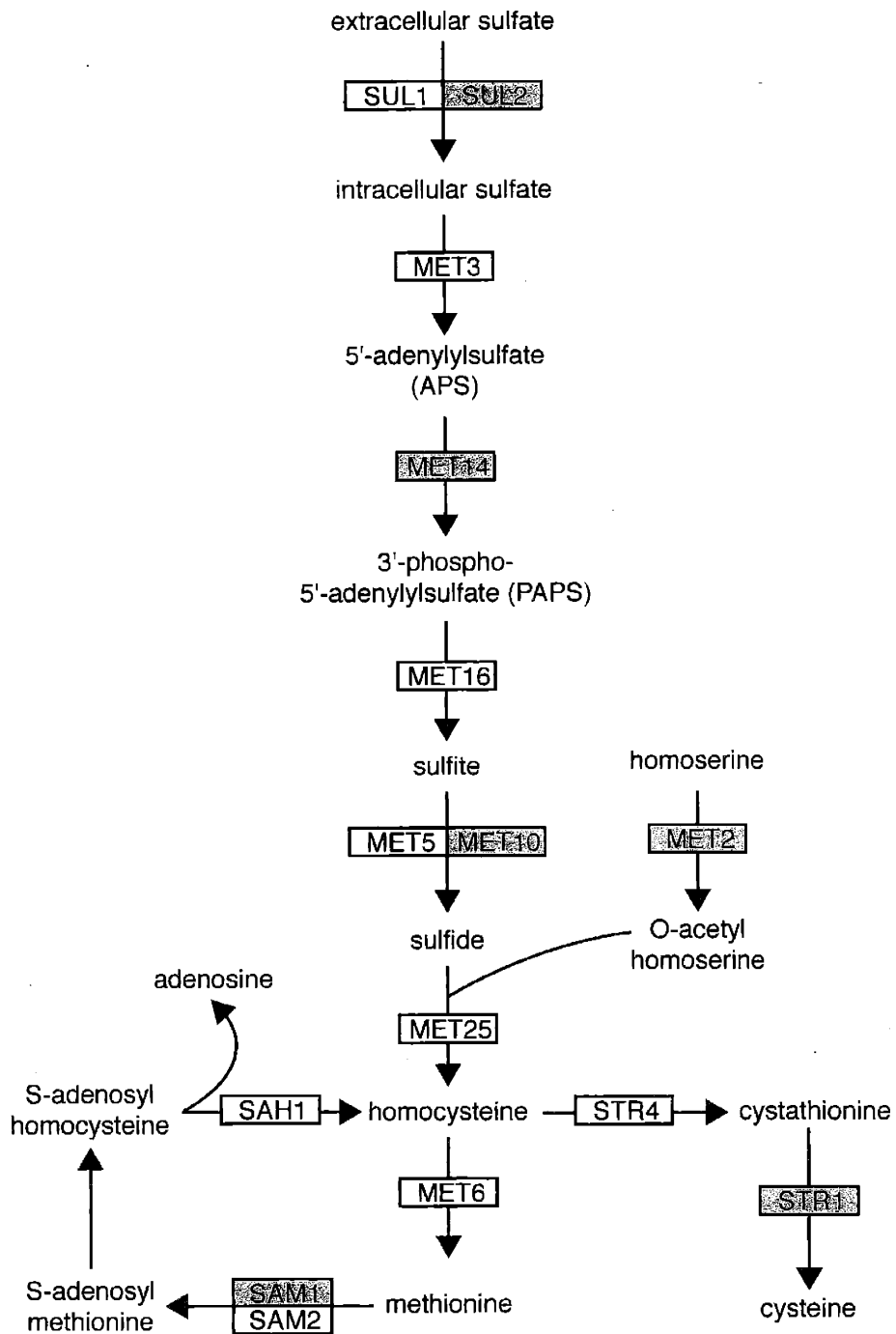


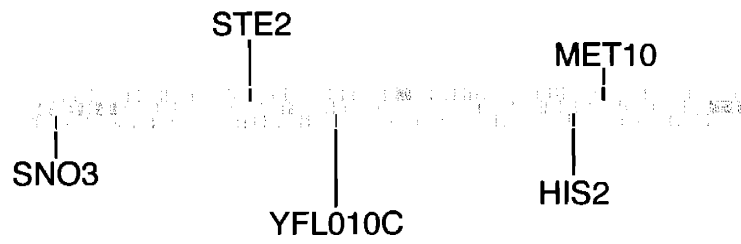
Figure 1. Sample output from the Biological Mapper.

The methionine biosynthesis pathway is shown, which is one of the displays included in the Biological Mapper tool. Genes whose promoters are bound by the Met4 protein, as determined by genome-wide location analysis (Lee et al., 2002), are shaded.

There are two versions of the chromosome display, one in which the genome is drawn to scale and one in which each genomic feature is of a uniform width. The to-scale version displays each gene or intergenic region as a thin rectangle, where the length of the rectangle is proportional to the size of the feature in the genome and its horizontal position in the display reflects its position in the genome (Figure 2A). There are three rows of rectangles: Watson strand ORFs are on the top and Crick strand ORFs are on the bottom with intergenic regions drawn in the middle. Genes that are found in any of the gene lists entered by the user have their corresponding rectangle colored according to the feature color in the gene list. The names of the genes that are colored are optionally displayed above or below the rectangles with a line connecting the text label to the rectangle. Individual text labels can be hidden by clicking on the corresponding rectangle or they can be moved by clicking and dragging the text label. Clicking and releasing on the text label will spawn a new browser window with a SGD report for that gene (Dwight et al., 2002). Each chromosome is displayed individually as multiple large images overload many browsers.

The other version of the chromosome display tool displays every gene or intergenic region as a hollow rectangle of uniform size in the order it appears along the chromosome without respect to which strand the coding information is on (Figure 2B). Yellow circles denote centromeres. If a gene appears in a gene list, the rectangles are filled according to the feature color of the appropriate gene list, otherwise they are filled according to the default color. The names of the genes associated with each rectangle can be displayed by moving the mouse over the rectangle. Clicking on the rectangle will open a new browser window with the SGD report for that gene. The chromosomes are

A.



B.



Figure 2. Sample output from the Chromosome Displays.

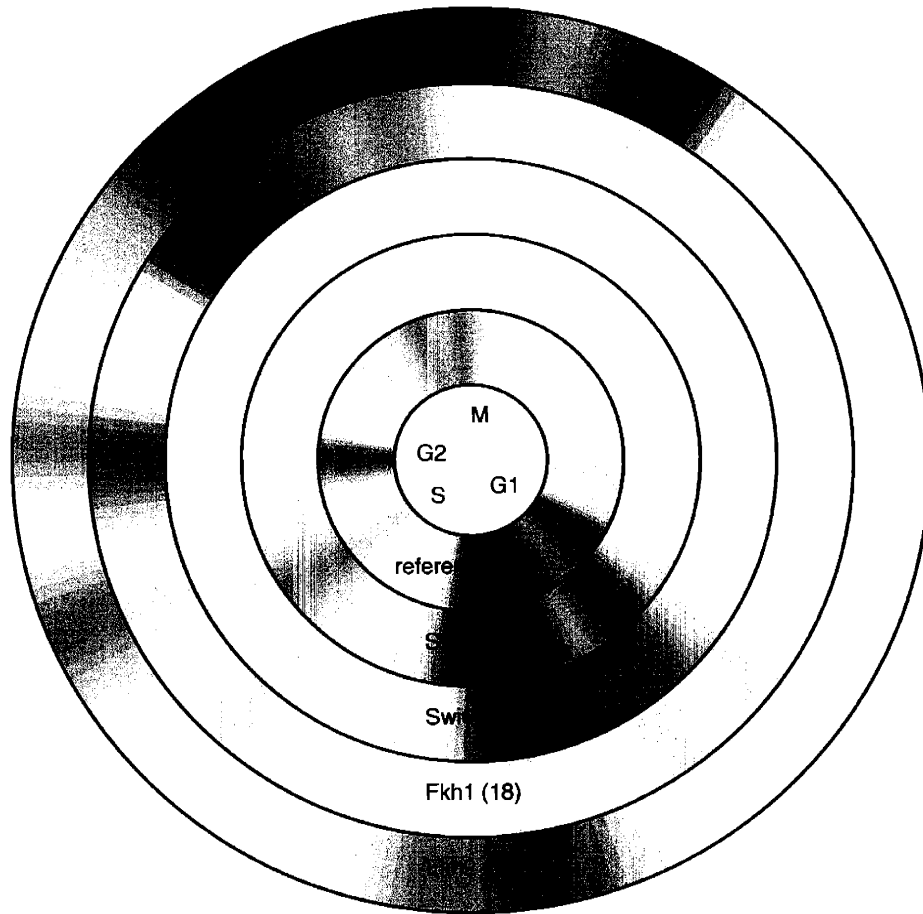
Chromosome VI is shown with five genes highlighted with **A.** features that are proportional to their length in the genome or **B.** features that are all the same width, regardless of the size of the gene.

initially drawn such that the two arms of the chromosome are split into two parts at the centromere. By clicking on the centromere, the orientation of each arm of the chromosome can be flipped. Clicking again on the centromere causes the two arms to be joined together at the chromosome. The use of this not-to-scale version may be useful when information about the intergenic regions is not needed and the lengths of genomic features are not important.

Cell Cycle Display and Expression Data Mapper

These tools compare genes entered by the user to predefined sets of genes from genome-wide expression data. For the Cell Cycle Display, genes in each gene list are compared to the set of cell cycle-regulated genes as determined by Spellman et al. (1998). The gene lists are drawn as concentric circles with the first gene list being in the center and the last gene list being on the outside (Figure 3A). Reference circles can be drawn before or after any of the circles representing user-defined gene lists. The reference band is equivalent to a user's entering the entire collection of cell cycle-regulated genes as a gene list. The cell cycle is subdivided into a number of bins defined by the user, and genes from each gene list that are cell cycle-regulated are placed in a bin according to the phase of the cell cycle in which its expression peaks. The bin that contains the most genes is colored with the highest intensity of the feature color for that gene list and the color intensity of all other bins are scaled relative to the most populated bin. For visual effect, a blurring is created around each bin that has genes in it. The user can control the extent of blurring by adjusting the bandwidth parameter. The Cell Cycle Display was

A.



B.

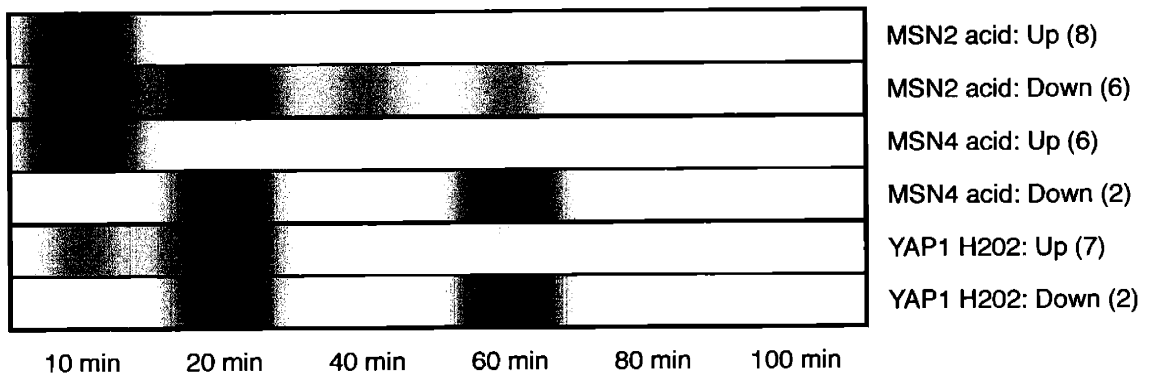


Figure 3. Sample output from the Cell Cycle Display and the Expression Data Mapper.

A. Genes whose promoters are bound by the Swi4, Swi6, Fkh1 and Fkh2 proteins, as determined by genome-wide location analysis, were used with the Cell Cycle Display (Simon et al., 2001). The reference band comes from all cell cycle regulated genes as defined by Spellman et al. (1998).

B. Promoters that are bound, according to genome-wide location analysis (Lee et al., 2002), by Msn2 or Msn4 when cells are grown under acidic conditions, or by Yap1 when cells are grown in the presence of hydrogen peroxide, were used with the acid expression data set in the Expression Data Mapper (Causton et al., 2001).

used in Simon et al. (2001) to illustrate where in the cell cycle nine different transcriptional activators function.

The Expression Data Mapper takes the concept of the Cell Cycle Display, but applies it to linear time course expression data (Figure 3B). As a result, the output is drawn as stacked rectangles instead of concentric circles, with the leftmost portion of each rectangle corresponding to the beginning of the time course and the rightmost portion being the end of the time course. The time courses available include the effect of heat shock, osmotic shock, and treatments with acid, alkali, hydrogen peroxide, high salt and rapamycin (Causton et al., 2001; Hardwick et al., 1999). This display provides a simple method for highlighting an interaction between the user's data and genome-wide expression data.

Functional Category Circles

It is not always anticipated in a genome-wide expression or location experiment which genes' expression will be affected under the experimental condition tested. Being able to easily group sets of genes into functional categories can provide valuable insights into understanding what cellular processes are being affected by the condition or mutant being examined. The MIPS database has created a functional catalog where they have placed each yeast gene into a category based on the gene's prospective function (Mewes et al., 2002). The catalog is designed in a hierarchical fashion such that there are several broad categories that all genes fall into, with several tiers of subcategories below where the genes are more precisely classified.

The Functional Category Circles tool displays the first two tiers of the MIPS functional catalog as circles (Figure 4). The larger circles depict the first tier of the hierarchy. Smaller circles within the larger circles are subcategories of the category specified by the larger circle. The area of the smaller circles is proportional to the number of genes in that subcategory. After the user selects one of the entered lists of genes, a hypergeometric test is performed to determine whether the number of genes that fall into the particular category is greater than would be expected by chance for a list of genes of the size entered (Tavazoie et al., 1999). A p-value is returned from the hypergeometric distribution and the circle is colored red if the p-value is below a threshold set by the user. Darker intensities of red correspond to each order of magnitude that the p-value is below the threshold.

Transcriptome Bins

The Transcriptome Bins tool provides a means to determine if a particular list of genes contains genes that are predominantly transcribed at a particular rate (Figure 5). This tool is only able to use one or two lists of genes at a time. The output is a bar graph with the x-axis corresponding to groups of genes that are all transcribed within a range of transcriptional frequencies. The number of total genes in each group is not the same but the ranges included in each group are intended to create general classes of genes. The height of each bar on the graph reflects the percentage of genes found in a particular group. The data used for assigning transcriptional frequencies to genes comes from Holstege et al. (1998).

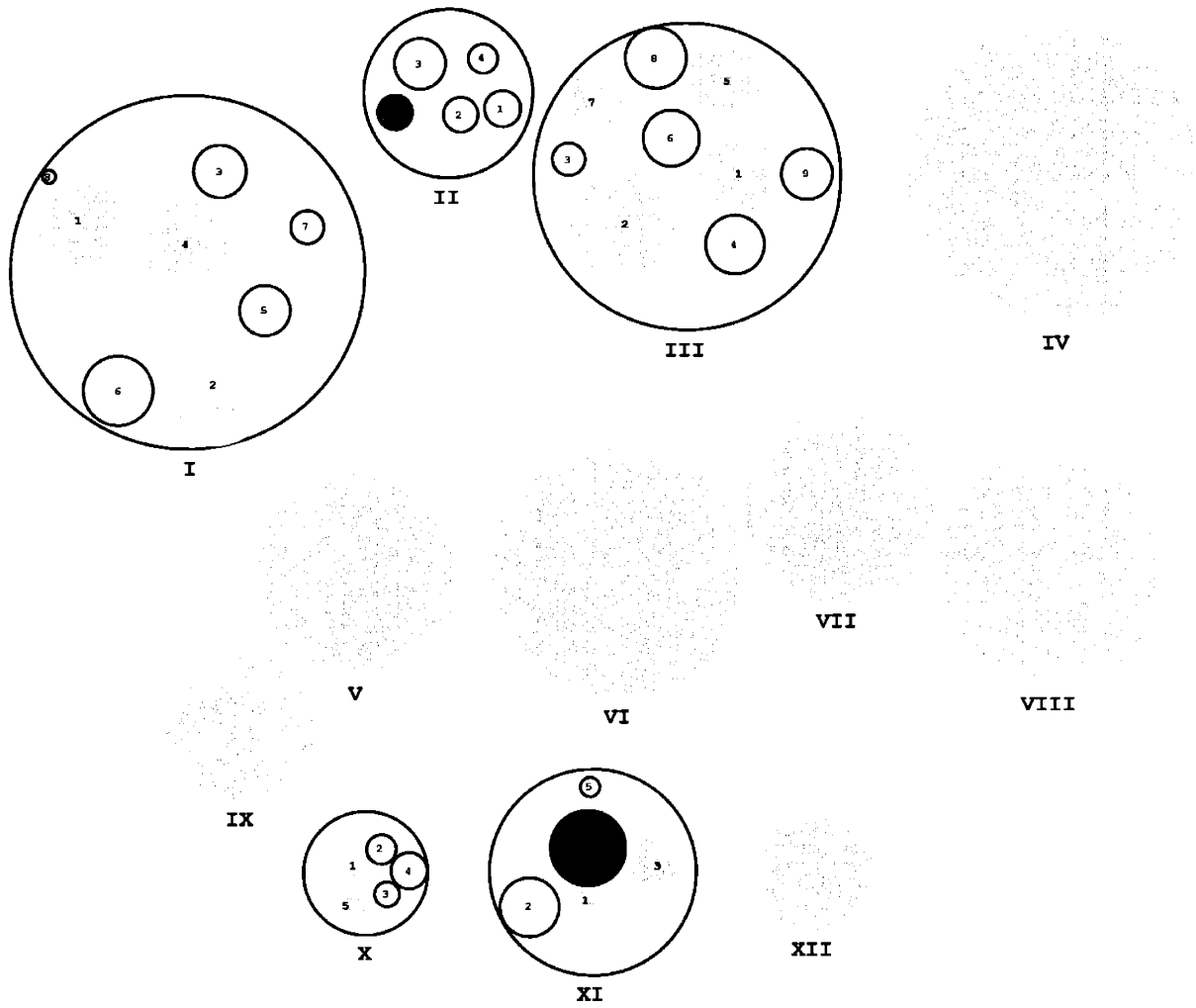


Figure 4. Sample output from the Functional Category Circles.

Promoters that are bound by Msn2 when cells are grown under acidic conditions as determined by genome-wide location analysis (Lee et al., 2002) were used with the Functional Category Circles tool.

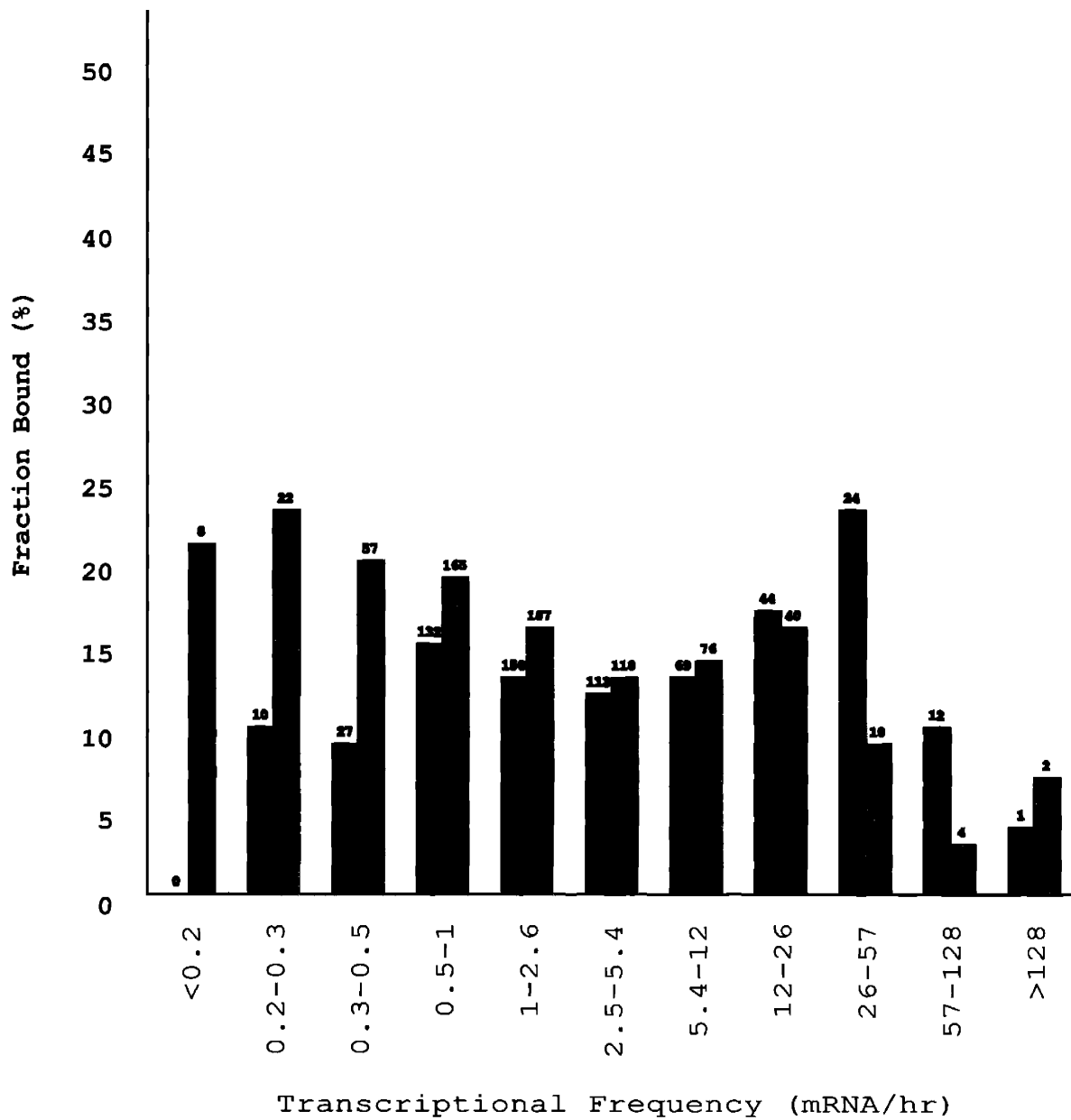


Figure 5. Sample output from the Transcriptome Bins tool.

Genes that are cell cycle regulated (blue; Spellman et al., 1998) or induced by rapamycin (red; Hardwick et al., 1999) were used with Transcriptome Bins tool.

Venn Diagram Drawer

Venn Diagrams are a convenient way to depict the relationship between two sets. With this tool, two gene lists are drawn as circles that are proportional in area to the number of genes that are in each list. The two circles will be drawn to overlap each other if there are genes in common between the two lists. The area of this overlapping region will be proportional to the number of shared genes and will be colored as a combination of colors from the two gene lists. The diameter of the larger circle and which gene lists should be used to draw the circles can be selected by the user.

Venn diagrams have proven to be very useful in communicating results of genome-wide expression analysis (Holstege et al., 1998; Lee et al., 2000). A common question asked of two genome-wide expression data sets is how many genes from the two different experiments are in common. Many genes in common often indicate a biological connection between the two experimental conditions. Conversely, a very small overlap of genes suggests that the experimental conditions are not involved in same biological process.

Regulator Network Wheel

Signal transduction pathways are often depicted as proteins that point to other proteins, implying that one protein transduces a signal through the next. As more is learned about these pathways, it is often the case that many pathways are interconnected and are better termed biological networks. The end points of these networks are usually at the transcriptional level, where a gene is activated or repressed in response to a signal.

Transcriptional networks consist of genes that regulate other genes, which in turn regulate more genes. The data from a study using genome-wide location analysis to identify the targets of 106 transcriptional regulators can be used to determine which regulators are the targets of other regulators (Lee et al., 2002). The Regulator Network Wheel consists of the names of these 106 transcriptional regulators arranged radially in a circle (Figure 6). Arrows drawn across the circle connect regulators with their targets. Each regulator name represents the regulator protein if an arrow originates from it and the promoter that drives expression of the gene coding for the regulator if the arrow points to it. A subset of all connections from the genome-wide location data is drawn based on the genes in the selected gene list with arrows only being drawn from genes that appear in the gene list. Regulators that may self-regulate and bind to their own promoters are displayed in bold face. The regulators are grouped along the circle by general biological function, with each grouping receiving a different color.

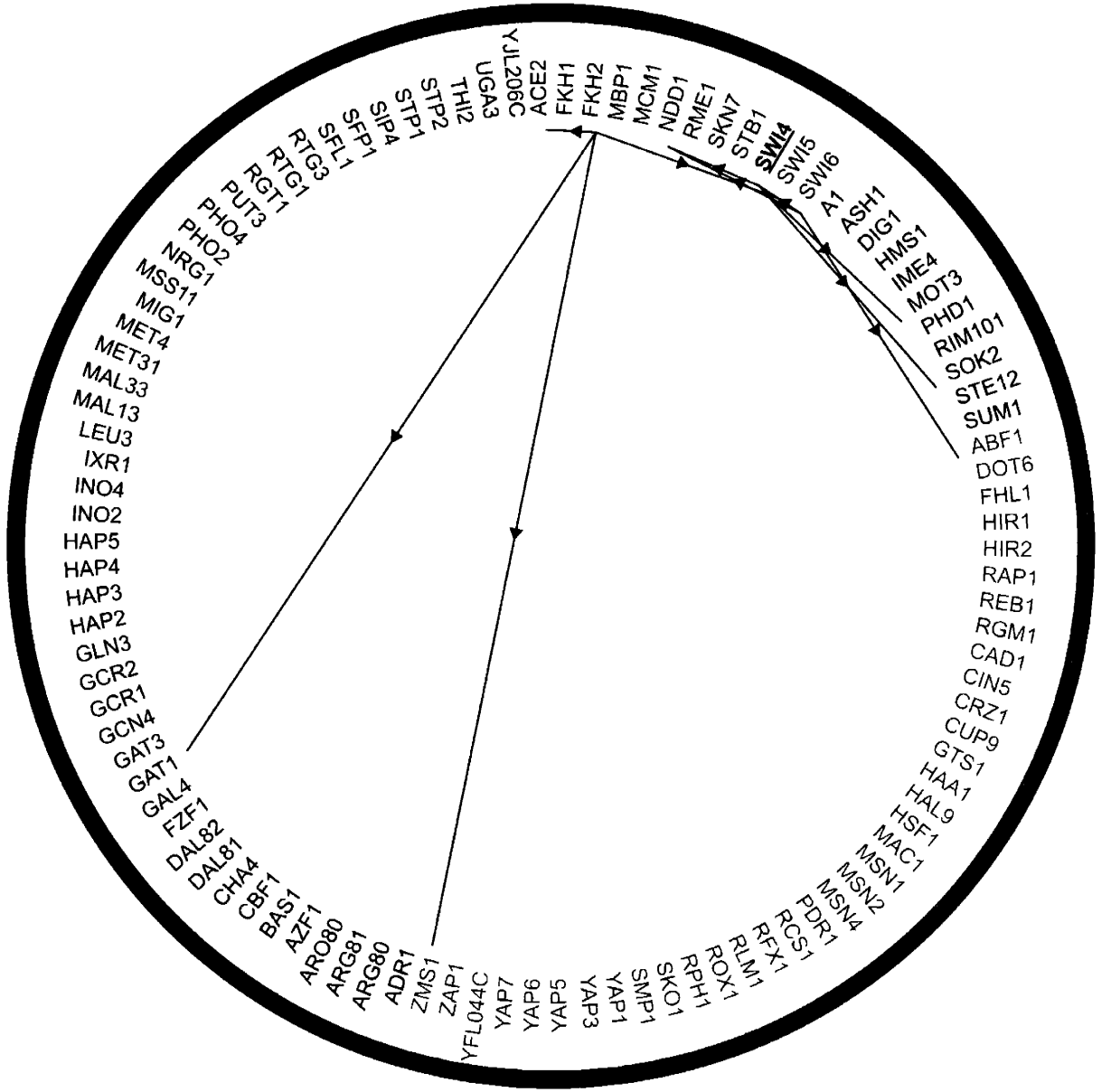


Figure 6. Sample output from the Regulator Network Wheel.

106 transcriptional regulators are displayed in a circle and 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 orange, environmental response in green, and metabolism in blue). Lines with arrows depict binding of a regulator to the gene encoding another regulator. Circles with arrows depict binding of a regulator to the promoter region of its own gene.

Future Directions

Displays from the GDS have been useful in facilitating biological discovery. The chromosomal display was used in a study where genome-wide expression analysis was performed to examine the effects of histone depletion on gene expression (Wyrick et al., 1999). The image revealed that telomere-proximal genes were derepressed upon histone depletion. The GDS is not restricted to being used with genome-wide expression data. GDS has also been used with genome-wide location data to uncover transcriptional regulatory networks from a study where over 100 transcriptional regulators were profiled and should continue to facilitate analysis as this data is explored further (Lee et al., 2002).

GDS allows for a comparison of the results of a microarray experiment or any other experiment that produces a list of genes to many different data sets that represent different facets of biology. The use of visual displays such as those included in GDS enables researchers to communicate their data and to easily look for unexpected biological connections in their data. As more genome-wide data is subjected to this type of analysis, unexpected biological insights will certainly be revealed.

GDS may continue to evolve. As the need for new graphical tools becomes apparent, they can be incorporated into the suite. Additional expression data sets may be added to the Expression Data Mapper and additional pathways may be added to the Biological Mapper. As other genomes become better annotated, these tools could be adapted for the results of experiments from other organisms.

References

- 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.
- Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., *et al.* (2002). Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* 30, 69-72.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868.
- Hardwick, J. S., Kuruvilla, F. G., Tong, J. K., Shamji, A. F., and Schreiber, S. L. (1999). Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc Natl Acad Sci U S A* 96, 14866-14870.
- 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.
- Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. *Nucleic Acids Res* 30, 42-46.
- Lee, T. I., Causton, H. C., Holstege, F. C., Shen, W. C., Hannett, N., Jennings, E. G., Winston, F., Green, M. R., and Young, R. A. (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature* 405, 701-704.
- Lee, T.I., Rinaldi, N.J., Robert, F., *et al.* (2002). A transcriptional regulatory network map for *Saccharomyces cerevisiae*. In preparation.
- Mewes, H. W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., and Weil, B. (2002). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 30, 31-34.
- Quackenbush, J. (2001). Computational analysis of microarray data. *Nat Rev Genet* 2, 418-427.
- 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.

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.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nat Genet* 22, 281-285.

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.

Appendix A

Genome Expression on the World Wide Web

Published as: Jennings, E.G. and Young, R.A. (1999). Genome Expression on the World Wide Web. *Trends Genet* *15*, 202-4.

The study of gene expression traditionally has been pursued through a combination of biochemical, genetic, and molecular biological studies. Genome sequences and new technologies have recently provided new approaches to study gene expression. By using high density DNA microarrays or “DNA chips,” which consist of either oligonucleotides or cDNAs attached to a solid phase, researchers are now able to measure the level of thousands of mRNAs simultaneously. This allows investigators to identify the set of genes influenced by a physiological event or a particular mutation and could ultimately allow biologists to understand the transcriptional program of the cell.

Several groups have performed these kinds of chip experiments with the bakers' yeast, *Saccharomyces cerevisiae*, under a variety of conditions. Yeast currently has several advantages over metazoans for genome-wide expression studies. Since the entire yeast genome has been sequenced, the expression level of every gene can be measured. The small size of the yeast genome, which consists of approximately 6200 genes, means that it takes fewer data points to provide complete information about the organism's transcriptional state. Analysis of genome-wide expression data is more easily performed in the context of the substantial yeast literature. The genetic tractability of yeast permits efficient experimental examination of models that emerge from genome-wide expression data.

Genome-wide expression experiments create enormous quantities of data which must be managed, analyzed and presented in new ways. Since traditional journals are not well suited for the presentation of large amounts of data, researchers have begun to provide access to their data on the World Wide Web. The popularity and relative ease of use of the web makes this forum suitable for the posting and sharing of this data. We discuss some of the sites on the web which have been created to present and analyze genome-wide expression data below.

Web-based Expression Data

Papers featuring genome-wide expression experiments generally have accompanying web sites which are listed in Table 1. There are a few core features that are found in most of these sites. The ability to search a database by gene name allows users to track any gene of interest in a given experiment. Most experiments report expression values as a fold change from some standard experimental condition (e.g. wild type or time zero) relative to a condition of interest (e.g. mutant or time +X). Users can search the database for those genes whose expression has changed a particular amount. When genes are listed, a brief annotation is supplied with the gene name which gives users some understanding of the function of genes with which they are unfamiliar. Finally, most sites allow the user to download their data in tabular form. This is particularly useful for investigators who have devised their own methods to analyze and present data.

Brown and coworkers have used DNA chips to study various physiological processes in yeast by analyzing the state of gene expression over a time course (Chu et al., 1998; DeRisi et al., 1997). The user can query the data based on a minimum or maximum fold change for any number of time points and can therefore effectively retrieve the genes which exhibit a particular pattern of expression. The site accompanying Spellman et al. (1998) provides time course data for the cell cycle but displays it graphically for an easy snapshot of the transcriptional profile of a particular gene.

Other investigators have used genome-wide expression experiments to analyze various components of the transcription apparatus (Holstege et al., 1998; Myers et al., 1999). They use mutants in genes to determine what contribution various transcription factors make toward gene expression. The site supporting Holstege et al. (1998) allows the user to identify genes affected by the loss of a transcription factor and to list them according to functional categories. This provides insight into those transcription factors with roles in the regulation of a particular physiological process.

Transcriptome

Two groups have described the yeast mRNA population in terms of the level of every detectable mRNA species; this population has been called the transcriptome (Table 1). Velculescu et al. (1997) used serial analysis of gene expression (SAGE) to measure the number of copies of a given mRNA per cell. Holstege et al. (1998) used DNA chips. Both data sets are available on the web, searchable by gene name, and provide easy access to additional information about the gene of interest, either through the SGD (Cherry et al., 1998) for SAGE, or through YPD (Hodges et al., 1999) for the DNA chip data.

Analysis

As the study of gene expression through genome-wide analysis is new, the development of methods and tools for analyzing expression data is in its infancy. Eisen et al. (1998) have created a program that allows investigators to perform cluster analysis on expression data. They use color bars to depict changes in gene expression and group genes which change in similar ways so that users can easily see genes whose expression is coordinately regulated. Several groups have written programs to search for over-represented sequence motifs in promoters (Brazma et al., 1998; Roth et al., 1998; van Helden et al., 1998). These programs can be used in conjunction with genome-wide expression data to determine if a particular DNA sequence mediates the regulation of a group of genes.

Future analytical programs will need to address several issues. The output from a given experiment is often a list of genes, many of which may be unfamiliar to the investigator. A program which allows users to take a list of gene products and group them by function, metabolic pathway, or biochemical complex will help uncover how a particular physiological process is regulated. It will be useful to determine and graphically display the intersection of the set of genes affected in one experiment with the set of

genes affected in another; the experiment is to determine whether two cellular processes use similar regulatory mechanisms. Such analysis should help unleash the power of genome-wide expression experiments and reveal more of the transcriptional regulatory circuitry of the cell. Finally, it is particularly interesting to consider the development of a computer program capable of using large amounts of expression data to predict the transcriptional behavior of cells.

Table 1. Genome expression web resources

URL	Description	Reference
Expression		
http://cmgm.stanford.edu/pbrown/explore/	Diauxic shift, TUP1, YAP1	DeRisi et al., 1997
http://cmgm.stanford.edu/pbrown/med2/	MED2 deletion	Myers et al., 1999
http://cmgm.stanford.edu/pbrown/sporulation/	Sporulation	Chu et al., 1998
http://genome-www.stanford.edu/cellcycle/	Cell cycle	Spellman et al., 1998
http://genomics.stanford.edu/yeast/cellcycle.html	Cell cycle	Cho et al., 1998
http://www.hsph.harvard.edu/geneexpression/	Alkating agent	Jelinsky and Samson, 1999
http://mips.gsf.de/proj/yeast/transcription/chrXI_map.html	Chromosome XI	Richard et al., 1997
http://mips.gsf.de/proj/yeast/transcription/mig1_contr.html	MIG1	Klein et al., 1998
http://web.wi.mit.edu/young/expression/	Transcription Apparatus	Holstege et al., 1998
Transcriptome		
http://genome-www.stanford.edu/cgi-bin/SGD/SAGE/querySAGE	SAGE	Velculescu et al., 1997
http://web.wi.mit.edu/young/expression/	RNA Polymerase II mutant	Holstege et al., 1998
Analysis		
http://arep.med.harvard.edu/mrnadata/mrnasoft.html	Promoter Analysis	Roth et al., 1998
http://copan.cifn.unam.mx/Computational_Biology/yeast-tools/	Promoter Analysis	van Helden et al., 1998
http://rana.stanford.edu/clustering/	Clustering	Eisen et al., 1998
http://www.cs.Helsinki.FI/~vilo/Yeast/	Promoter Analysis	Brazma et al., 1998
Supporting Yeast Databases		
http://genome-www.stanford.edu/Saccharomyces/	<i>S. cerevisiae</i> Genome Database (SGD)	Cherry et al., 1998
http://quest7.proteome.com/YPDhome.html	Yeast Protein Database (YPD)	Hodges et al., 1999
http://www.mips.biochem.mpg.de/proj/yeast/	Munich Information Centre for Protein Sequences (MIPS)	Mewes et al., 1999

Acknowledgments

We thank Fran Lewitter for helpful comments. E.G.J. is a predoctoral fellow of the Howard Hughes Medical Institute.

References

- Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. (1998). Predicting gene regulatory elements in silico on a genomic scale. *Genome Res* 8, 1202-1215.
- Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., *et al.* (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res* 26, 73-79.
- 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.
- 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.
- Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868.
- 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.
- 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.
- 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.
- Klein, C. J., Olsson, L., and Nielsen, J. (1998). Glucose control in *Saccharomyces cerevisiae*: the role of Mig1 in metabolic functions. *Microbiology* 144, 13-24.
- Mewes, H. W., Heumann, K., Kaps, A., Mayer, K., Pfeiffer, F., Stocker, S., and Frishman, D. (1999). MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 27, 44-48.
- Myers, L., Gustafsson, C., Hayashibara, K., Brown, P., and Kornberg, R. (1999). Mediator protein mutations that selectively abolish activated transcription. *Proc Natl Acad Sci U S A* 96, p67-72.
- Richard, G. F., Fairhead, C., and Dujon, B. (1997). Complete transcriptional map of yeast chromosome XI in different life conditions. *J Mol Biol* 268, 303-321.

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.

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.

van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* *281*, 827-842.

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D., Jr., Hieter, P., Vogelstein, B., and Kinzler, K. W. (1997). Characterization of the yeast transcriptome. *Cell* *88*, 243-251.

Appendix B

Interplay of Positive and Negative Regulators in Transcription Initiation by RNA Polymerase II Holoenzyme

Published as: Lee, T. I., Wyrick, J. J., Koh, S. S., Jennings, E. G., Gadbois, E. L., and Young, R. A. (1998). Interplay of positive and negative regulators in transcription initiation by RNA polymerase II holoenzyme. *Mol Cell Biol* *18*, 4455-62.

My Contributions to this Project

Ellen Gadbois, a former graduate student in the Young laboratory, performed a genetic selection for suppressors of the temperature sensitive phenotype of *srb4-138*. She had grouped many of the recessive suppressors into complementation groups and identified the genes that were mutated in two of these groups. I continued working with the recessive suppressors by forming more complementation groups from those suppressors she had not grouped. I cloned two additional recessive suppressors, *NOT5* and *CAF1*, both members of the NOT complex. This was consistent with Tony Lee's and Ellen's work that mutations in *NOT1* and *NOT3* could also suppress the defect in *srb4-138*. These results, together with John Wyrick's work with identifying the dominant suppressors of *srb4-138*, and Sang Seok Koh's biochemical analysis of Srb4, Srb6, and Med6 proteins, are described here.

Summary

Activation of protein coding genes involves recruitment of an RNA polymerase II holoenzyme to promoters. Since the *Srb4* subunit of the holoenzyme is essential for expression of most class II genes and is a target of at least one transcriptional activator, we reasoned that suppressors of a temperature sensitive mutation in *Srb4* would identify other factors generally involved in regulation of gene expression. We report here that *MED6* and *SRB6*, both of which encode essential components of the holoenzyme, are among the dominant suppressors, and the products of these genes interact physically with *Srb4*. The recessive suppressors include *NCB1* (*BUR6*), *NCB2*, *NOT1*, *NOT3*, *NOT5* and *CAF1*, which encode subunits of NC2 and the Not complex. NC2 and Nots are general negative regulators which interact with TATA binding protein (TBP). Taken together, these results suggest that transcription initiation involves a dynamic balance between activation mediated by specific components of the holoenzyme and repression by multiple TBP-associated regulators.

Introduction

Expression of mRNA genes in eukaryotes involves the recruitment of RNA polymerase II and other general transcription factors to promoters (Orphanides et al., 1996; Roeder, 1996). Evidence that RNA polymerase II can be found associated with most of the general transcription factors and additional factors essential for initiation *in vivo* suggests that much of the transcription initiation apparatus can be recruited to promoters in a preassembled RNA polymerase II holoenzyme (Chao et al., 1996; Kim et al., 1994; Koleske and Young, 1994; Maldonado et al., 1996; Ossipow et al., 1995; Pan et al., 1997; Shi et al., 1997).

RNA polymerase II holoenzymes consist of RNA polymerase II and a subset of general transcription factors, together with Srb/Mediator proteins. Several lines of evidence indicate that the Srb/Mediator proteins are involved in the response to gene-specific activators. Truncation mutations of the C-terminal domain (CTD) of the largest subunit of RNA polymerase II result in defects in activation (Allison and Ingles, 1989; Gerber et al., 1995; Liao et al., 1991; Scafe et al., 1990), and the Srb proteins were originally identified through genetic interactions with one such truncation mutant (Hengartner et al., 1995; Liao et al., 1995; Thompson et al., 1993). The RNA polymerase II holoenzyme responds to the addition of transcriptional activators *in vitro* while purified polymerase and general factors alone do not (Kim et al., 1994; Koleske and Young, 1994). The Srb/Mediator complex binds to the CTD and can be purified as a separate complex from holoenzyme. This purified Srb/Mediator complex is necessary to reconstitute the ability of a defined transcription system to respond to activators *in vitro* (Hengartner et al., 1995; Kim et al., 1994).

Activators have been shown to bind directly to the Srb/Mediator complex (Hengartner et al., 1995), and genetic and biochemical studies have identified the Srb4 subunit as a target of the well-studied acidic activator Gal4 (Koh et al., 1998).

Temperature sensitive mutations in the essential Srb4 holoenzyme subunit can produce a rapid, general shutdown of mRNA synthesis, demonstrating that Srb4 is required for expression of most protein-coding genes (Thompson and Young, 1995). Because essentially all of the Srb proteins are tightly associated with the holoenzyme in *Saccharomyces cerevisiae* cells, the Srb-containing holoenzyme likely functions in transcription initiation at most class II promoters *in vivo*.

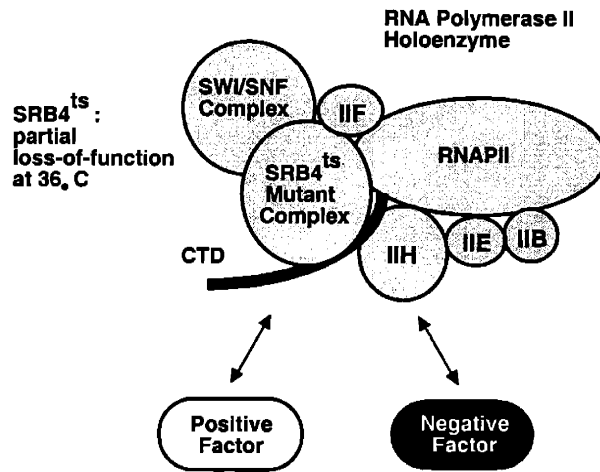
To further investigate the role of Srb4 and the holoenzyme in transcriptional activation, we have isolated and characterized extragenic suppressors of the temperature sensitive phenotype of a *srb4-138* mutant. Srb4 normally has a positive role in transcription initiation, and the Srb4-138 mutation affects the function of the protein at the nonpermissive temperature (Thompson and Young, 1995). Suppressors of Srb4-138 must compensate for the reduced function of the mutant subunit and might therefore include mutations in other positive factors which increased their activity. The suppressors might also include mutations in negative factors which reduced their activity. Indeed, we have identified dominant and recessive suppressors of the ts phenotype of *srb4-138* which occur in positive and negative regulators, respectively. The results described here support a model in which activation mediated by holoenzyme is repressed by general negative regulators associated with TAT box binding protein (TBP).

Results

Fifty-four isolates were obtained in a genetic selection for suppressors of the temperature sensitive phenotype of cells harboring *srb4-138*. Genetic analysis revealed that all 54 extragenic suppressing mutations occurred in eight genes (Figure 1). Eight of the isolates were dominant suppressing mutations that occurred in two genes that encode components of the Srb/Mediator complex. The remaining 46 isolates were recessive suppressing mutations that occurred in six genes whose products are subunits of negative regulators.

MED6* and *SRB6* alleles are dominant suppressors of *srb4-138

Genetic analysis of the 54 genetic suppressors indicated that eight had a dominant suppressing phenotype (Figure 1). Linkage analysis of the eight dominant suppressors revealed that they fall into two groups. Group A consists of seven isolates and group B consists of one isolate. Figure 2A shows the suppressing phenotype of one of the isolates from each group. To identify the gene represented in group A, a genomic DNA library was constructed from one of the dominant suppressing isolates, cells containing the *srb4-138* temperature sensitive mutation were transformed with the library, and recombinant DNA clones that suppressed the temperature sensitive phenotype were isolated. The minimal fragment of genomic DNA with the suppressing phenotype was identified and sequenced, and was found to encode a mutant form of *MED6* (*MED6-101*). The *MED6* dominant mutation is a G to T substitution at nucleotide 454, converting amino acid 152 from

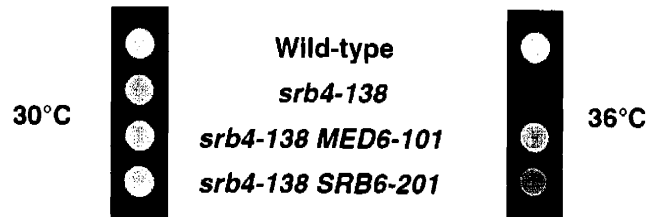


Gene	Dominant Alleles	Recessive Alleles	Deletion Phenotype	Protein Mass (kDa)	Subunit of
MED6	7	0	inviable	32.8	RNA pol II holoenzyme
SRB6	1	0	inviable	13.8	RNA pol II holoenzyme
NCB1	0	5	inviable	15.5	NC2
NCB2	0	1	inviable	16.7	NC2
NOT1	0	18	inviable	240.2	Not complex
NOT3	0	19	viable	94.4	Not complex
NOT5	0	2	viable	65.8	Not complex
CAF1	0	1	viable	49.7	Not complex

Figure 1. Genetic suppressors of the temperature sensitive Srb4-138 mutant RNA polymerase II holoenzyme.

Conditional defects in the essential Srb4 subunit of the holoenzyme might be overcome by compensatory gain of function mutations in a general positive regulator or loss of function mutations in a global negative regulator.

A. Dominant suppressors of *srb4^{ts}* mutant



B. Med6 amino acid sequence

1 MNVTPLDELQWKSPewIQVfGLRTENVLDYFAESPFfdKT
 41 SNNQVIKMQRQFSQLNDPNAAVNMTQNIMTLpdGKNGNLE
 81 EEFAYVDPARRQILFKYPMYMQLEEEELMKLDGTEYVLSSV
 121 REPDFWVIRKQRRTNNSGVGSAKGPEIIPLQDYIIGANI

↓
Y *MED6-101*

161 YQSPTIFKIVQSRLMSTSYHLNSTLESlyDLIEFQPSQGV
 201 HYKVPTDSTTATAATNGNnAGGGSNKSSVRPTGGANMAT
 241 VPSTTNVNMTVNTMGtGGQTI DNgtGRtGNGNMGITTEML
 281 DKLMVTSIRSTPNYI*

C. Srb6 amino acid sequence

1 MSNQALYEkLEqTRtILSVKLAELINMTTIADRNDDDEGS
 41 FAQENSELAVATTsvMMVNNQTMQLIKNVQDLLILTRSIK

↓
H *SRB6-201*

81 EKWLLNqIPVTEHskVTRfDEKQIEELLDNCIETFVAEKT
 121 T*

Figure 2. Dominant mutations in *MED6* or *SRB6* suppress *srb4-138*.

(A) Growth phenotypes of *SRB4* cells and cells containing the *srb4-138* mutation, either alone or with the *MED6-101* or *SRB6-201* mutations. Cells were spotted on YPD medium and incubated at 30°C and 36°C for 2 days.

(B) Sequence of Med6. The suppressing mutation is indicated in boldface type. The *MED6-101* dominant mutation is a G to T substitution at nucleotide 454, converting amino acid 152 from aspartic acid to tyrosine.

(C) Sequence of Srb6. The suppressing mutation is indicated in boldface type. The *SRB6-201* dominant mutation is an A to C transversion at nucleotide 175, converting amino acid 59 from asparagine to histidine.

aspartic acid to tyrosine (Figure 2B). To confirm that the suppressing mutation occurs in the *MED6* gene, a gap-repair method was used with plasmids lacking the *MED6* open reading frame but retaining flanking DNA. Plasmids gap-repaired from the suppressing strain conferred suppression while plasmids repaired from wild-type strains did not, confirming that the suppressing mutation occurs in *MED6*.

It is possible that the dominant suppressing mutations in *MED6* eliminated a requirement for *Srb4* function. To investigate this possibility, the *MED6-101* allele was introduced into a strain with a genomic deletion of *SRB4* covered by a CEN plasmid containing *SRB4* and *URA3*, and the cells were plated on 5-FOA medium to select for cells that lost the plasmid (Figure 3). No cells could be recovered on 5-FOA medium, indicating that some *Srb4* function is essential for cell survival, even in the presence of *MED6-101*. These data suggest that *Srb4-138* protein retains some function, even at temperatures that do not permit cell growth.

To identify the suppressor in group B, a genomic DNA library was prepared from cells containing this dominant mutation and transformed into cells containing the *srb4-138* temperature sensitive mutation. DNA clones that suppressed the temperature sensitive phenotype were recovered and sequenced. The minimal fragment sufficient for the suppressing phenotype was found to contain a mutant allele of *SRB6*. The *SRB6-201* dominant mutation is an A to C transversion at nucleotide 175, converting amino acid 59 from asparagine to histidine. Gap repair analysis confirmed that the suppressing mutation

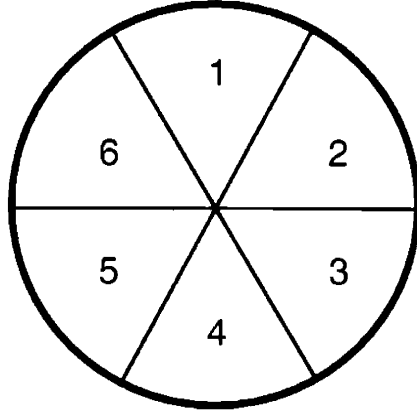
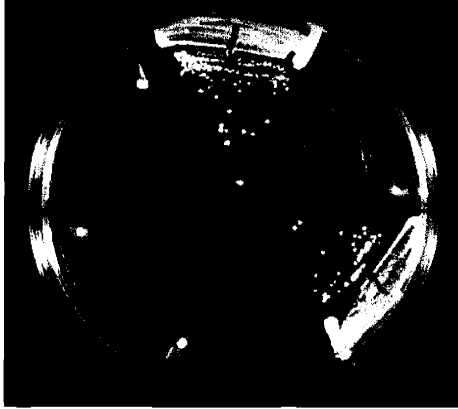


Figure 3. Dominant mutations in *MED6* or *SRB6* do not bypass a requirement for *srb4-138*.

Growth phenotypes of cells containing plasmid-borne copies of *SRB4* or *srb4-138*, either alone or with the *MED6-101* or *SRB6-201* mutations. Cells were streaked on 5-FOA to select against *URA3* versions of *SRB4* or *srb4-138* and incubated at 30°C for 2 days.

Relevant genotypes are described as follows: **1)** *srb4Δ2::HIS3* [pCT127 (*SRB4 LEU2 CEN*)] **2)** *srb4Δ2::HIS3* [pCT15 (*SRB4 URA3 CEN*)] **3)** *srb4Δ2::HIS3* [pCT181 (*srb4-138 LEU2 CEN*)] **4)** *srb4Δ2::HIS3* [pEG39 (*srb4-138 URA3 CEN*)] **5)** *srb4Δ2::HIS3 MED6-101* [pEG39 (*srb4-138 URA3 CEN*)] **6)** *srb4Δ2::HIS3 SRB6-201* [pEG39 (*srb4-138 URA3 CEN*)]

occurs in the *SRB6* gene. As with the *MED6* suppressing alleles, the dominant mutation in *SRB6* was unable to bypass the requirement for some level of *Srb4* function, as cells harboring *SRB6-201* did not restore viability to cells with an *SRB4* deletion (Figure 3).

Med6 and Srb6 associate with Srb4

Srb4, *Med6* and *Srb6* are components of the *Srb*/Mediator complex (Hengartner et al., 1995; Kim et al., 1994; Lee et al., 1997; Myers et al., 1998). *Srb4* and *Srb6* are involved in similar functions *in vivo* (Thompson et al., 1993; Thompson and Young, 1995) and can form a complex *in vitro* (Koh et al., 1998). The observation that dominant mutations in *MED6* can compensate for a partial loss of *Srb4* function might reflect a physical interaction between *Srb4* and *Med6*. We examined pairwise interactions between recombinant *Srb4*, *Med6* and *Srb6* proteins expressed in a baculovirus system (Figure 4). Extracts containing FLAG epitope-tagged *Srb4* or *Med6* were incubated with extracts containing an equimolar amount of untagged *Med6* or *Srb6* protein. The epitope tagged subunit was immunoprecipitated and the pellet was analyzed by Western blotting for the untagged protein. Untagged protein was used in parallel reactions to control for specific immunopurification, and ovalbumin was added to each reaction to control for non-specific aggregation. The results confirmed previous evidence that *Srb4* and *Srb6* can form a complex (Figure 4A) (Koh et al., 1998) and revealed that *Srb4* and *Med6* bind to one another *in vitro* (Figure 4B). There were no detectable interactions between *Med6* and *Srb6* (Figure 4C).

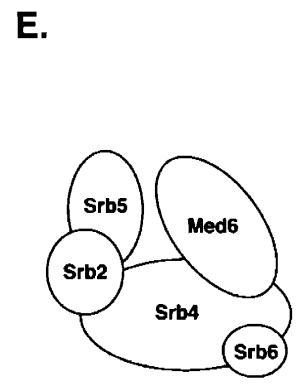
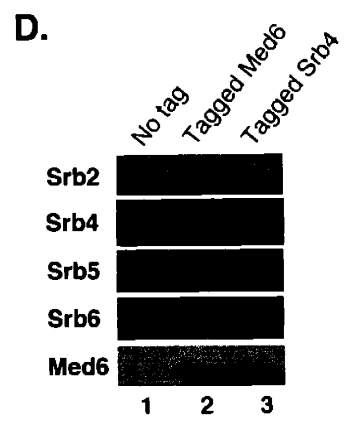
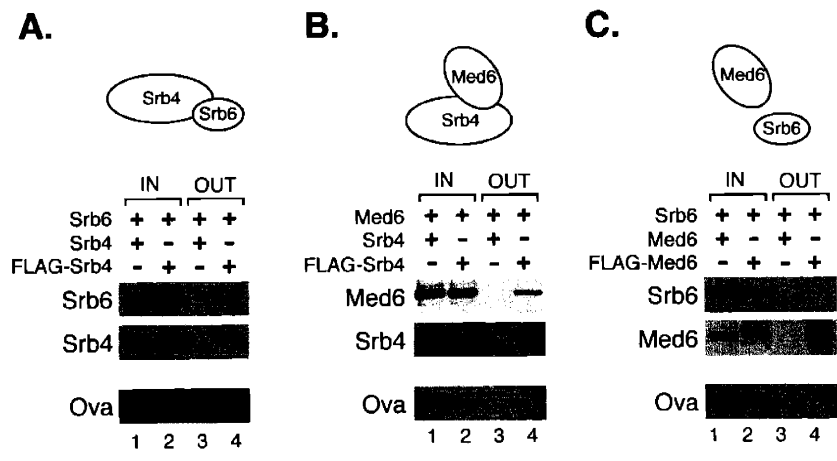


Figure 4. Med6 and Srb6 associate with Srb4

(A-C) Pairwise interactions of Med6 with Srb proteins. An insect cell extract containing one recombinant protein was incubated with an extract containing equimolar amounts of another recombinant protein that lacked (lanes 1 and 3) or contained (lanes 2 and 4) the FLAG epitope tag. Ovalbumin was added to each reaction to serve as a control for specific immunoprecipitation. The epitope-tagged Med6 or Srb4 and bound proteins were immunoprecipitated using anti-FLAG antibody. Fractions (1/10) of the load (IN) and all of the pellets (OUT) were analyzed by Western blotting using specific antibodies. A schematic interpretation of the binary interactions is presented at the top of each panel.

(D) Insect cell extracts containing Med6 and Srb proteins were subjected to coimmunoprecipitation using anti-FLAG antibody. In the control reaction (lane 1, no tag), no tagged recombinant was included. In the other reactions, either Med6 (lane 2, tagged Med6) or Srb4 (lane 3, tagged Srb4) contained FLAG epitope-tag.

(E) A model depicting interactions between Med6 and dominant Srb proteins. This model assumes the stoichiometric association of the five proteins.

Srb2, Srb4, Srb5 and Srb6 form a complex *in vitro* (Koh et al., 1998). Figure 4D shows that Med6 binds to this Srb subcomplex. Extracts containing Srb2, Srb4, Srb5, Srb6 and Med6 proteins were incubated and immunoprecipitated using antibodies against epitope-tagged Srb4 or Med6. In both cases, all five proteins were coimmunoprecipitated (Figure 4D, lanes 2 and 3). The genetic and biochemical data is consistent with the model for an Srb/Mediator subcomplex shown in Figure 4E.

***NCB1* and *NCB2* loss-of-function mutations compensate for *Srb4* defect**

In addition to the eight dominant suppressors identified as alleles of *MED6* and *SRB6*, 46 suppressors were characterized as recessive suppressors of *srb4-138*. We recently reported the identification of one of the recessive suppressors as *NCB1* (*BUR6*), which encodes the large subunit of NC2 (Gadbois et al., 1997). Since NC2 is composed of two subunits, we tested whether any of the other complementation groups involved the *NCB2* gene, which encodes the other subunit of this general negative regulatory factor. An isolate from each complementation group was transformed with a plasmid carrying a wild-type *NCB2* gene, and the transformants were screened for the loss of the recessive suppressing phenotype. One group, consisting of a single isolate, showed this loss of suppression following transformation with the wild-type *NCB2* gene. Subsequent gap repair analysis confirmed that a mutation in *NCB2* was responsible for suppression. Thus, mutations in either subunit of yeast NC2 can cause suppression of *srb4-138* (Figure 5A). The suppressing

A. NC2 suppressors of *srb4^{ts}* mutant



B. NC2 β amino acid sequence

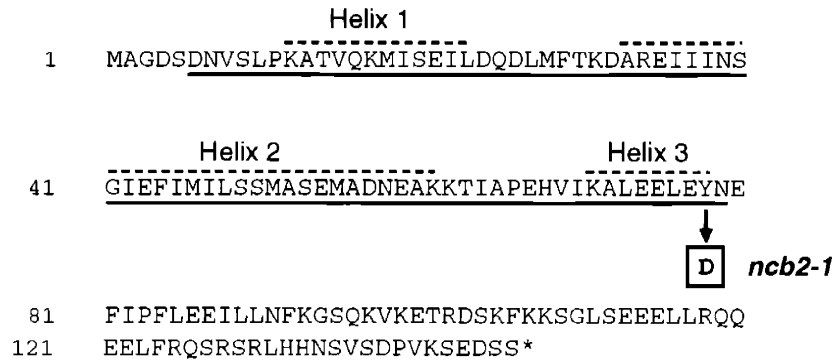


Figure 5. Recessive mutations in either subunit of NC2 suppress *srb4-138*.

(A) Growth phenotypes of *SRB4* cells and cells containing the *srb4-138* mutation, either alone or in conjunction with *ncb1-1* or *ncb2-1* mutations. Cells were spotted on YPD medium and incubated at 30°C and 36°C for 2 days.

(B) Sequence of NC2 β . The suppressing mutation is indicated in boldface type. The *ncb2-1* recessive mutation is a T to G substitution at nucleotide 232, converting amino acid 78 from tyrosine to aspartic acid. The histone fold motif is underlined and the helices of the motif are indicated by dashed lines.

allele (*ncb2-1*) was sequenced and found to affect the histone fold motif which is important for the stable interaction of the two NC2 subunits (Arents and Moudrianakis, 1995; Baxevanis and Landsman, 1997; Goppelt and Meisterernst, 1996; Mermelstein et al., 1996). The suppressing mutation is a T to G substitution at nucleotide 232, converting a highly conserved tyrosine to aspartic acid within the histone fold motif (Figure 5B). This defect is similar to that identified for the suppressing mutation in the other subunit of NC2; the *ncb1-1* mutation truncates the histone fold motif of this protein (Gadbois et al., 1997). Like *NCB1*, the *NCB2* gene is essential (Gadbois et al., 1997; Kim et al., 1997), and so the missense mutation must cause a partial functional defect in the small NC2 subunit.

***NOT1, NOT3, NOT5* and *CAF1* loss-of-function mutations compensate for *Srb4* defect**

Since two of the recessive complementation groups define genes encoding a known negative regulator of transcription, we expected that additional negative regulators might be represented among the other complementation groups. Previous genetic and biochemical studies indicated that *MOT1* (Auble et al., 1994; Davis et al., 1992; Wade and Jaehning, 1996), the *NOT* genes (Collart and Struhl, 1993; Collart and Struhl, 1994; Oberholzer and Collart, 1998), and histones (Han and Grunstein, 1988; Lenfant et al., 1996; Recht et al., 1996; Wan et al., 1995 and reviewed in Grunstein, 1990; Wolffe, 1997) all negatively regulate transcription. Consequently, representative isolates of the unidentified complementation groups were transformed with wild-type *MOT1, NOT1, NOT2, NOT3, NOT4, NOT5, HTA1-HTB1*, or *HHT1-HHF1* and tested for viability at the restrictive temperature. Based on the loss of suppression of the *srb4-138* phenotype when transformed

with a copy of the wild-type gene, three complementation groups were found to represent recessive suppressing alleles of *NOT1*, *NOT3* and *NOT5* (Figure 6). The identities of these suppressors were confirmed by linkage analysis or gap repair analysis. In addition, a disruption of the *NOT3* gene also suppressed *srb4-138*, indicating that a loss-of-function mutation in *NOT3* could alleviate the holoenzyme defect (data not shown). The suppressing alleles of these genes were sequenced and the recessive mutations identified. The *not1-10* suppressing allele is a G to A substitution at nucleotide 5828 converting amino acid 1943 from glycine to aspartic acid. The *not 3-10* suppressing allele is a 19 bp duplication of nucleotides 1620 to 1638 that results in a frameshift and truncation of the protein. The *not 5-10* suppressing allele is a C to G mutation at nucleotide 1443 converting amino acid 481 from phenylalanine to leucine. As described for the previously identified *ncb1-1* suppressor (Gadbois et al., 1997), suppressing alleles of *NOT* genes are able to rescue global transcriptional defects in polyA mRNA expression caused by *srb4-138* at the restrictive temperature (data not shown).

The Not proteins have recently been shown to associate with the Ccr4/Caf1 regulatory complex (Liu et al., 1998). Since three of the recessive complementation groups define *NOT* genes, we examined whether *CCR4* or *CAF1* was also represented among the recessive suppressors of *srb4-138*. The sole isolate of the last unidentified complementation group was transformed with wild-type *CCR4* or *CAF1* and tested for viability at the restrictive temperature. Based on the loss of suppression of the *srb4-138* phenotype when

Not complex suppressors of *srb4^{ts}* mutant



Figure 6. Recessive mutations in *NOT1*, *NOT3*, *NOT5* or *CAF1* suppress *srb4-138*.

Growth phenotypes of *SRB4* cells and cells containing the *srb4-138* mutation, either alone or in conjunction with *not1-10*, *not3-10*, *not5-10* or *caf1-10* mutations. Cells were spotted on YPD medium and incubated at 30°C and 36°C for 2 days.

transformed with a copy of the wild-type gene, this complementation group represented a recessive suppressing allele of *CAF1* (Figure 6). The identity of this suppressor was confirmed by gap-repair analysis. The suppressing allele of this gene, *caf1-10* was sequenced and the recessive mutation identified as an A to G substitution at nucleotide 739 that converts amino acid 247 from asparagine to aspartic acid. A disruption of *CAF1* does not suppress the temperature sensitive phenotype of *srb4-138* (data not shown).

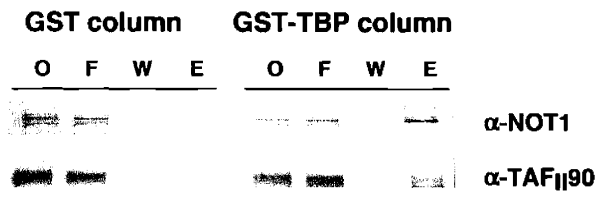
If the Not proteins are general negative regulators, then loss of function mutations in Not proteins might be expected to suppress defects due to at least some temperature sensitive RNA polymerase II mutants. Indeed, in an independent selection for suppressors of temperature sensitive mutations in the second largest subunit of RNA polymerase II (*RPB2*), we have identified recessive suppressing mutations in *NOT1* and *NOT2* (T. Lee, unpublished data).

NC2 and Not proteins associate with TBP

Med6 and Srb6 have previously been identified as components of the mediator subcomplex of RNA polymerase II holoenzyme (Koleske and Young, 1994; Lee et al., 1997). Quantitative Western analysis revealed that NC2 α , NC2 β , Not1 and Not3 are not components of the holoenzyme (data not shown). Evidence that yeast NC2 binds TBP and represses transcription (Gadbois et al., 1997; Goppelt and Meisterernst, 1996; Goppelt et al., 1996; Kim et al., 1997) led us to investigate whether Not proteins also bind to TBP. Whole cell extract was prepared and passed over a GST-TBP column, and various fractions were analyzed by Western analysis with anti-Not1 antibody. Under conditions where TAF_{II}s are

also retained, Not1 bound to the GST-TBP column (Figure 7A). To confirm that Not1 interacts with TBP, yeast TBP was epitope tagged at its N-terminus with the FLAG epitope, immunoprecipitated from yeast cell extracts, and associated proteins were identified by western blot analysis (Figure 7B). TAF_{II}s, NC2, Mot1 and Spt3, each of which has previously been shown to bind TBP (Auble et al., 1994; Eisenmann et al., 1992; Gadbois et al., 1997; Goppelt et al., 1996; Poon et al., 1994; Poon and Weil, 1993; Reese et al., 1994), were used as positive controls. Not1, TAF_{II}s, NC2, Mot1 and Spt3 were all found to coimmunopurify with TBP. In contrast, RNA polymerase II was not associated with the immunopurified TBP preparation.

A.



B.

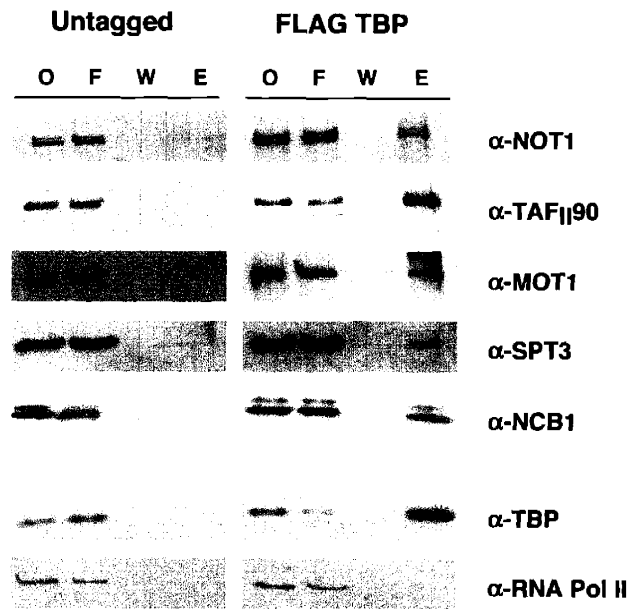


Figure 7. TBP-associated proteins.

(A) Western blot analysis of eluate from TBP affinity column eluates. Crude cell extract was passed over GST or GST-TBP columns. Bound proteins were eluted with 2M KCl and probed with antibodies against Not1 and Taf Π 90. Lanes indicated are onput (O), flowthrough (F), wash (W) and eluate (E)

(B) Western blotting of immunoprecipitations from crude fractions of whole cell extract derived from cells with (FLAG-TBP) or without (untagged) FLAG epitope tagged TBP reveals that Not proteins interact, directly or indirectly with TBP. Under these conditions, TBP also interacts with other proteins previously described as TBP-interacting factors, including Taf Π 90, NC2, Spt3, and Mot1, but not RNA polymerase II. Lanes indicated are onput (O), flowthrough (F), wash (W) and eluate (E).

Discussion

Genetic selections can provide substantial new insights into the function of complex biological systems. Genetic and biochemical characterization of suppressors of RNA polymerase II mutations previously led us and others to the holoenzyme model. The isolation and characterization of eight genes found in a selection for suppressors of the *srb4-138* allele provides additional insights into the holoenzyme components which are involved in transcription activation and the set of TBP regulators which appear to be general negative regulators of class II genes.

Functional interactions among holoenzyme subunits implicated in activation

In principle, dominant suppressors of an *srb4* temperature-sensitive mutant could reveal compensatory mutations in positive factors which are involved in class II gene expression. In fact, dominant mutations compensating for the *srb4* mutation occurred in two genes whose products are also Srb/Mediator subunits, *MED6* and *SRB6*. These two proteins are essential components of the holoenzyme and contribute to the response to activators *in vivo* and *in vitro* (Hengartner et al., 1995; Lee et al., 1997; Thompson et al., 1993).

The functional interactions suggested by the genetic analysis are supported by physical interactions seen with recombinant Srb4, Srb6 and Med6. Srb4 interacts with both Med6 and Srb6 *in vitro*. Taken together, the genetic and biochemical results further refine our model for Srb subunit interactions within the holoenzyme (Koh et al., 1998) and extend it to incorporate Med6 (Figure 4E).

General negative transcription factors associated with TBP

Recessive mutations that suppress the *srb4-138* defect occurred in the genes encoding both subunits of the negative regulator NC2. Genetic and biochemical evidence indicates that NC2 is a general negative regulator of transcription which is essential for yeast cell viability (Gadbois et al., 1997; Goppelt and Meisterernst, 1996; Goppelt et al., 1996; Inostroza et al., 1992; Kim et al., 1996; Kim et al., 1997; Meisterernst and Roeder, 1991; Mermelstein et al., 1996; Prelich, 1997; White et al., 1994; Yeung et al., 1997; Yeung et al., 1994). The protein represses transcription by binding to promoter-bound TBP and preventing the association of TFIIA and TFIIB during formation of the preinitiation apparatus (Goppelt et al., 1996; Kim et al., 1996; Mermelstein et al., 1996).

Recessive suppressing mutations also occurred in genes encoding subunits of the Not complex. Previous experiments indicated that the Not protein complex can act as a negative regulator at several genes (Collart and Struhl, 1993; Collart and Struhl, 1994). A recent report suggests that some components of this complex may have a positive role at certain genes, but it is not yet clear whether this role is direct (Liu et al., 1998). The evidence presented here indicates that Not proteins, like NC2, have a general negative regulatory function. Loss of function mutations in *NOT1* can suppress defects in both *SRB4* and RNA polymerase II subunit (*RPB2*) mutations. Similarly, loss of function mutations in NC2 suppress defects of both *SRB4* and *SRB6* mutations. The observations that mutations in NC2 or Not proteins can suppress defects in Srb and Rpb subunits of the RNA polymerase II holoenzyme indicates that NC2 and Not proteins contribute to a general level of

transcriptional repression that must be overcome during transcription initiation by the holoenzyme.

While the mechanism of repression by NC2 involves TBP binding, the mechanism of repression by Not proteins has not been clear. We have found that Not1 associates with TBP, consistent with genetic evidence that Not mutations can relieve defects due to specific TBP mutations (Collart, 1996). Thus, Not proteins may be one of several factors, including NC2, that contribute to gene regulation by regulating TBP activity (Lee and Young, 1998).

The balance between activation and repression

Genetic analysis of suppressors of the *srb4-138* mutation has revealed functional links between holoenzyme subunits involved in activation and two general negative regulators that associate with TBP. It is notable that recessive mutations in both *NCB* and *NOT* genes have previously been observed to compensate for defects in activation. The *NOT* genes were identified in a screen for suppressors of a defect in the GCN4 transcriptional activator (Collart and Struhl, 1993; Collart and Struhl, 1994). A mutation in *NCB1* (*BUR6*) can compensate for the loss of the Upstream Activating Sequence (UAS) in the *SUC2* gene (Prelich, 1997; Prelich and Winston, 1993). These results are consistent with the model that activators generally recruit holoenzymes in a manner that is dependent on Srb4, Srb6 and Med6 function, and that NC2 and Not complexes generally inhibit transcription activation by this pathway.

Experimental Procedures

Yeast manipulations

Yeast strains and plasmids are listed in Table 1. Details of strain and plasmid constructions are available upon request. Yeast media was prepared as described (Thompson et al., 1993). Yeast transformations were done using a lithium acetate procedure (Schiestl and Gietz, 1989). Plasmid shuffle techniques were performed as described (Boeke et al., 1987) using 5-fluoro-orotic acid (5-FOA) as a selective agent against *URA3* plasmids. Plasmids were recovered from yeast as described (Hoffman and Winston, 1987).

DNA methods

DNA manipulations were performed as described (Sambrook et al., 1989). PCR amplifications were performed with Vent DNA polymerase (New England Biolabs) or Taq DNA polymerase (Perkin Elmer) as described by the manufacturer.

Selection and analysis of *srb4-138* suppressors.

Two ml YPD cultures of the yeast strain Z628 were grown overnight at 30°C, plated at a density of 3×10^6 cells/plate and placed at 36°C. Suppressors arose at a frequency of approximately one in 2×10^6 cells. One colony was picked from each plate, further colony purified, and subsequently retested for ability to grow at 36°C.

To exclude intragenic revertants, the *srb4-138 LEU2* plasmids were recovered from strains harboring suppressing mutations and transformed into Z811. Cells were streaked on

5-FOA to select against the *URA3* version of *srb4-138* and assayed for growth at 36°C on YPD, and those that grew were considered to have a suppressing mutation linked to the original plasmid-borne copy of *srb4-138*.

Dominant and recessive growth phenotypes were determined by mating the suppressors in the Z628 background to Z811, and assaying growth at 36°C on YPD. Diploids able to grow at 36°C contained a dominant suppressor. Diploids unable to grow at 36°C contained a recessive suppressor. To facilitate linkage analysis, the mating type of approximately half of the dominant suppressors and half of the recessive suppressors was switched by inducing expression of a plasmid-born *HO* gene under the control of a galactose inducible promoter.

Random spore analysis of the dominantly suppressing mutations was used to determine if two independent isolates were likely to contain mutations in the same gene. Haploids, each containing the *srb4-138* mutation and an independently isolated suppressing mutation, were mated to each other to form diploids. These diploids were sporulated on plates and a small quantity of spores scraped off and shaken overnight at 30°C in 0.5 ml of 30 mM β-mercaptoethanol and 100 ng/ml Zymolase 100 T (ICN). 0.5 ml of 1.5% NP-40 and 0.4 g glass beads were added and the mixture incubated on ice for 15 min. The suspension was then vortexed 3 min, incubated on ice 5 min, vortexed 2 min, and the glass beads allowed to settle for 10 min at room temperature. The supernatant was removed, spun 2 min, the pellet washed once in water, then resuspended in water and a portion plated onto YPD. Approximately fifty of the haploid offsprings were assayed for their ability to grow at 36°C.

If all haploids were able to grow at 36°C, then the two suppressing isolates were assumed to contain mutations in the same gene.

Dominantly suppressing mutations were assayed for the ability to bypass the requirement for *Srb4*. Strains harboring dominant suppressors and carrying a *LEU2* plasmid with *srb4-138* were transformed with a *URA3* version of *srb4-138*. Transformants were grown in synthetic complete $\text{Ura}^- \text{Leu}^+$ to permit loss of the *LEU2* plasmid. The resultant strains were streaked on 5-FOA to select against the *URA3* containing plasmid. Cells harboring dominant mutations could not survive on 5-FOA indicating that there was still a requirement for *srb4-138* even in the context of *MED6-101* or *SRB6-201*.

Genetic complementation of the recessive alleles involved mating haploids, each containing the *srb4-138* mutation and an independently isolated suppressing mutation, to form diploids and assessing the ability of these diploids to grow at 36°C. Diploids able to grow at 36°C were assumed to contain suppressing mutations in the same gene. Genomic clones of each complementation group were used to confirm the identity of each member of the complementation group and to identify additional members.

Cloning of dominant suppressors of *srb4-138*

Genomic DNA clones containing *MED6-101* and *SRB6-201* were isolated by taking advantage of their ability to dominantly suppress the *srb4-138* temperature sensitive phenotype. Genomic DNA was isolated from strains containing the dominant suppressing alleles of *MED6* and *SRB6* (Z848 and Z847, respectively). Libraries were constructed in a

yeast centromeric plasmid containing the *URA3* gene as a selectable marker (Thompson et al., 1993). These libraries were transformed into yeast cells containing *srb4-138* and genomic clones were isolated from Ura⁺ transformants able to grow at 36°C. When necessary, the mutant genes were further subcloned.

Complementation analysis

Complementation groups containing mutant alleles of *NCB2*, *NOT1*, *NOT3*, *NOT5* and *CAF1* were identified by transforming Z828 with a pCT3 plasmid containing wild-type *NCB2* (pRY7212); Z829 with a YCP50 plasmid containing wild-type *NOT1* (gift of M. Collart); Z830 with a pRS316 plasmid containing wild-type *NOT3* (gift of M. Collart); Z864 with a pRS316 plasmid containing wild-type *NOT5* and Z862 with a pRS316 plasmid containing wild-type *CAF1* (pRY7288). The resulting strains no longer grew at the nonpermissive temperatures, indicating that the suppression phenotype was reversed by the wild-type *NCB2*, *NOT* and *CAF1* genes. Confirmation that these represented the suppressor-containing genes was obtained through linkage analysis (*NOT1* and *NOT3*) and gap repair (*NCB2*, *NOT5* and *CAF1*).

Genetic linkage analysis

The identities of *not1* and *not3* alleles as suppressors of *srb4-138* were confirmed by genetic linkage analysis. The *URA3* gene was integrated next to the *NOT1* gene in Z836 using *SacI*-digested pES183 (gift of E. Shuster). The resulting strain, Z837 was mated to Z829. The resulting diploid strain was sporulated and 20 tetrads were dissected. Analysis of

the resulting spores showed that temperature-sensitive phenotype always cosegregated with the Ura⁺ phenotype, indicating that the suppressing allele was tightly linked to the *NOT1* gene. For *NOT3*, the *URA3* gene was integrated next to the *NOT3* gene in Z836 using *EagI*-digested pRS306 with *NOT3* (gift of M. Collart). The resulting strain, Z838 was mated to Z830. The resulting diploid strain was sporulated and 20 tetrads were dissected. Analysis of the resulting spores showed that temperature-sensitive phenotype always cosegregated with the Ura⁺ phenotype, indicating that the suppressing allele was tightly linked to the *NOT3* gene.

Sequence analysis

Suppressors of the temperature sensitive phenotype of *srb4-138* were recovered by a plasmid gap repair technique (Orr-Weaver et al., 1983). Gap-repaired plasmids carrying suppressing alleles of *MED6*, *SRB6*, *NCB2*, *NOT5* and *CAF1* were sequenced (Research Genetics). Suppressing alleles of *NOT1* and *NOT3* were obtained by PCR of genomic DNA from strains Z829 and Z830, respectively. PCR products were directly sequenced by Research Genetics.

Expression of recombinant Med6

The *MED6* ORF was cloned into baculoviral transfer vectors by PCR amplification of the gene using the plasmid pET-MED6 (Lee et al., 1997) and oligonucleotides 5'-GGAAGATCTATGAACGTGACACCGTTGGAT-3' and 5'-TGCTCTAGATCATATGTAGTTTGGGGTGGGA-3'. Recombinant baculoviruses were

generated and used to infect Sf21 insect cells. Insect cell extracts were prepared as described (Koh et al., 1997).

Immunoprecipitation of Srb4, Med6 and Srb6

Coimmunoprecipitation experiments were performed to test interactions of Med6 with various Srb proteins. An insect cell extract containing FLAG epitope-tagged Med6 or Srb4 was incubated with an extract containing an equimolar amount of untagged, recombinant Srb4, Srb6 or Med6 for 3 hours on ice. Controls included the use of ovalbumin and the use of Med6 or Srb4 lacking FLAG epitope in the respective reactions. The anti-FLAG M2 antibody coupled agarose beads (Eastman Kodak), equilibrated in the buffer MTB (Hengartner et al., 1995), were added to the reactions and incubated for 3 hours at 4°C with constant agitation. Beads were precipitated and washed extensively with MTB. Proteins in the pellet were eluted by boiling in sample buffer and analyzed by Western blot. For the experiment shown in Figure 4D, insect cell extracts containing those five recombinant proteins were prepared by coinfecting the cells with the recombinant baculoviruses at a multiplicity of infection (m.o.i.) of 5-10. Coimmunoprecipitations were performed as above using anti-FLAG M2 antibody.

Antibody Reagents

A portion of Not1 (amino acids 1266 to 1442) was purified as a fusion to glutathione-S-transferase (GST) from *E. coli* DH5 α according to previously published methods (Smith and Johnson, 1988). The purified fusion protein was injected into rabbits to raise polyclonal

antisera. Anti-polII Western blots were performed with the mouse monoclonal 8WG16 antibody. All other Western blots were performed with rabbit polyclonal antisera. Anti-Spt3 antibody was the kind gift of J. Madison and F. Winston. Anti-TAF_{II}90 antibody was the kind gift of J. Reese and M. Green.

GST-TBP affinity chromatography

TATA box binding protein (TBP) affinity chromatography was performed as described (Reese et al., 1994) with the following modifications. To make whole cell extract, yeast strain BJ926 was grown to an OD of 3 in YPD at 30°C, harvested after washing in 150 mM Tris acetate pH 7.9, 50 mM potassium acetate and stored at -80°C. 130 g of thawed cell pellet was resuspended in 68 ml 3X lysis buffer (450 mM Tris acetate pH 7.9, 30% glycerol, 15 mM EDTA, 15 mM EGTA, 30 mM sodium fluoride, 1.8 mM sodium vanadate, 30 μM antipain-HCl, 15 mM benzamidine, 3 μg/ml aprotinin, 3 μg/ml leupeptin, 3 μg/ml pepstatin, 0.25 mM PMSF, 15 μM chymostatin). Cells were disrupted by bead beating for 20 cycles of 30 seconds of beating followed by 30 seconds of cooling in a stainless steel bead beater filled with 200 ml 0.4 - 0.6 micron glass beads washed in 1X lysis buffer. After beating, DTT and Na₂S₂O₅ were added to 0.5 and 0.1 mM respectively. The crude extract was centrifuged for 20 minutes at 10,000 rpm in a Sorval GSA rotor. 1/9th volume of 3M (NH₄)₂SO₄ (pH 7.9) was added slowly and the mixture was stirred gently for 20 minutes and degassed. 1/100th volume 10% polymin-P was added dropwise and the extract was stirred gently for 20 minutes and degassed. The extract was centrifuged for 90 minutes at 42,000 rpm in a Ti45 rotor

(Beckman) and the supernatant (180 ml, 36 mg/ml) was frozen and stored at -80°C . Prior to use, extract was thawed and dialyzed against buffer T(100) until the conductivity was equivalent to buffer T(150). Buffer T is 20 mM HEPES-KOH (pH 7.6), 10 mM magnesium acetate, 5 mM EGTA, 5 mM DTT, 20% glycerol, 0.5 $\mu\text{g/ml}$ each of leupeptin, pepstatin A, aprotinin, antipain-HCl, chymostatin, bestatin, 2 mM benzamidine-HCl, 0.5 mM PMSF and potassium acetate added to the concentration (in mM) indicated in parentheses.

GST-yTBP and GST columns were prepared as described (Reese et al., 1994). Yeast whole cell extract (~ 100 mg) was diluted in buffer T(150) to a total volume of 30 ml. 15 ml of dilute extract was incubated with 1.0 ml of GST-yTBP or GST agarose at 4°C with rotation. Resin was collected by gentle centrifugation (2,000 rpm, 1 minute) and washed with 10 column volumes of buffer T(150). Bound proteins were eluted with 2M potassium chloride. Peak fractions were pooled and dialyzed into buffer T(100).

Construction of FLAG-tagged TBP yeast strain

Plasmid RY7269 was constructed by PCR amplification with two sets of primers. The first set of primers generated a 1 kb fragment that incorporated the FLAG epitope behind the initial ATG of the open reading frame. This fragment was digested at the 5' end with *XhoI* and at the 3' end with *PspI406I* (an endogenous site at nucleotide 14 of the TBP open reading frame). The second set of primers generated a 2 kb fragment including the TBP open reading frame and approximately 1 kb of 3' downstream sequence. This fragment was digested at the 5' end with *PspI406I* and at the 3' end with *XmaI*. The two digested fragments were then ligated into the *LEU2* vector pRS315 digested with *XhoI* and *XmaI*. The

resulting construct was transformed into yeast strain BYΔ2 (Cormack et al., 1991) which has a genomic deletion of TBP covered by a wild-type copy of TBP on a *URA3* plasmid.

Selection against the *URA3* plasmid using 5-FOA generated strain Z850 and confirmed that the tagged version of TBP was fully functional and able to complement the TBP deletion.

Immunoprecipitation of FLAG-TBP

A crude fraction of yeast extract was prepared from yeast strains Z849 and Z850. Briefly, whole cell extract was prepared as described (Koleske et al., 1996). 50 mg of whole cell extract was diluted in buffer A(150) and passed over a 2 ml Bio-Rex 70 column equilibrated in buffer A(150). Buffer A is 20 mM HEPES-KOH, pH 7.6, 1 mM EDTA, 20% glycerol, 1 mM DTT, 0.5 mM PMSF, 1 mM benzamidine, and protease inhibitors as described above. The number in parentheses indicates the concentration of potassium acetate in mM. The columns were washed with 20 column volumes of buffer BH(150) and followed by elution with 10 column volumes of buffer BH(300) and buffer BH(600). Buffer BH is 20 mM HEPES-KOH, pH 7.6, 1 mM EDTA, 10% glycerol, 0.5 mM PMSF, 1 mM benzamidine, and protease inhibitors as described above. The number in parentheses indicates the concentration of potassium acetate in mM. Peak fractions of the BH(600) eluate were pooled and used for immunoprecipitations.

Approximately 100 μg of the BH(600) fraction was diluted with 4 volumes of buffer BH(0). Samples (approximately 1.5 ml each) were first cleared by incubation with 20 μl of anti-FLAG M1 affinity gel and then immunoprecipitated with 20 μl of anti-FLAG M2 affinity gel for 2 hours at 4°C with rotation. Beads were collected by centrifugation at 8,000

x g and washed five times with BH buffer supplemented with various concentrations of potassium acetate. Bound proteins were eluted by boiling for 1 minute in SDS-PAGE sample buffer without DTT. After centrifugation, additional sample buffer with DTT was added to the supernatant. Typically, 1/100 of the load and flowthrough and 1/5 of the eluate was loaded for Western blot analysis.

Table 1. Yeast Strains

<u>Strain</u>	<u>Genotype</u>
Z22	<i>Mat a ura3-52 his3Δ200 leu2-3,112</i>
Z579	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 [pCT127 (SRB4 LEU2 CEN)]</i>
Z628	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z804	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 ncb1-1 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z811	<i>Mat α ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 [RY7215 (srb4-138 URA3 CEN)]</i>
Z828	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 ncb2-1 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z829	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 not1-10 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z830	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 not3-10 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z836	<i>Mat α ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 [RY2882 (srb4-138 LEU2 CEN)]</i>
Z837	<i>Mat α ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 not1/URA3 [RY2882 (srb4-138 LEU2 CEN)]</i>
Z838	<i>Mat α ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 not3/URA3 [RY2882 (srb4-138 LEU2 CEN)]</i>
Z847	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 SRB6-201 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z848	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 MED6-101 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z849	<i>Mat a ura3-52 leu2-PET56 spt15Δ2 [YCp86 (SPT15 URA3 CEN)]</i>
Z850	<i>Mat a ura3-52 leu2-PET56 spt15Δ2 [RY7269 (SPT15 5' FLAG tag LEU2 CEN)]</i>
Z862	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 caf1-10 [pCT181 (srb4-138 LEU2 CEN)]</i>
Z864	<i>Mat a ura3-52 his3Δ200 leu2-3,112 srb4Δ2::HIS3 not5-10 [pCT181 (srb4-138 LEU2 CEN)]</i>

Acknowledgments

We thank P. Sharp, M. Green, V. Myer and H. Madhani for advice and discussions. We thank F. Holstege, M. Collart, M. Green, Y.-J. Kim, J. Madison, J. Reese, K. Struhl, C. Wilson, and F. Winston for kind gifts of extracts, strains, plasmids and antibodies. We thank A. S. Lee for technical assistance. J.J.W. is a predoctoral fellow of the National Science Foundation. E. J. is a predoctoral fellow of the Howard Hughes Medical Institute. This work was supported by National Institutes of Health grants to R.A.Y.

References

- Allison, L. A., and Ingles, C. J. (1989). Mutations in RNA polymerase II enhance or suppress mutations in GAL4. *Proc Natl Acad Sci U S A* 86, 2794-8.
- Arents, G., and Moudrianakis, E. (1995). The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proc Natl Acad Sci U S A* 92, p11170-4.
- Auble, D. T., Hansen, K. E., Mueller, C. G., Lane, W. S., Thorner, J., and Hahn, S. (1994). Mot1, a global repressor of RNA polymerase II transcription, inhibits TBP binding to DNA by an ATP-dependent mechanism. *Genes Dev* 8, 1920-34.
- Baxevanis, A. D., and Landsman, D. (1997). Histone and histone fold sequences and structures: a database. *Nucleic Acids Res* 25, 272-3.
- Boeke, J. D., Trueheart, J., Natsoulis, G., and Fink, G. R. (1987). 5-Fluoroorotic acid as a selective agent in yeast molecular genetics. *Methods Enzymol* 154, 164-75.
- Chao, D. M., Gadbois, E. L., Murray, P. J., Anderson, S. F., Sonu, M. S., Parvin, J. D., and Young, R. A. (1996). A mammalian SRB protein associated with an RNA polymerase II holoenzyme. *Nature* 380, 82-5.
- Collart, M. A. (1996). The NOT, SPT3, and MOT1 genes functionally interact to regulate transcription at core promoters. *Mol Cell Biol* 16, 6668-76.
- Collart, M. A., and Struhl, K. (1993). CDC39, an essential nuclear protein that negatively regulates transcription and differentially affects the constitutive and inducible HIS3 promoters. *Embo J* 12, 177-86.
- Collart, M. A., and Struhl, K. (1994). NOT1(CDC39), NOT2(CDC36), NOT3, and NOT4 encode a global-negative regulator of transcription that differentially affects TATA-element utilization. *Genes Dev* 8, 525-37.
- Cormack, B. P., Strubin, M., Ponticelli, A. S., and Struhl, K. (1991). Functional differences between yeast and human TFIID are localized to the highly conserved region. *Cell* 65, 341-8.
- Davis, J. L., Kunisawa, R., and Thorner, J. (1992). A presumptive helicase (MOT1 gene product) affects gene expression and is required for viability in the yeast *Saccharomyces cerevisiae*. *Mol Cell Biol* 12, 1879-92.

- Eisenmann, D. M., Arndt, K. M., Ricupero, S. L., Rooney, J. W., and Winston, F. (1992). SPT3 interacts with TFIID to allow normal transcription in *Saccharomyces cerevisiae*. *Genes Dev* 6, 1319-31.
- Gadbois, E. L., Chao, D. M., Reese, J. C., Green, M. R., and Young, R. A. (1997). Functional antagonism between RNA polymerase II holoenzyme and global negative regulator NC2 in vivo. *Proc Natl Acad Sci U S A* 94, 3145-50.
- Gerber, H. P., Hagmann, M., Seipel, K., Georgiev, O., West, M. A., Litingtung, Y., Schaffner, W., and Corden, J. L. (1995). RNA polymerase II C-terminal domain required for enhancer-driven transcription. *Nature* 374, 660-2.
- Goppelt, A., and Meisterernst, M. (1996). Characterization of the basal inhibitor of class II transcription NC2 from *Saccharomyces cerevisiae*. *Nucleic Acids Res* 24, 4450-5.
- Goppelt, A., Stelzer, G., Lottspeich, F., and Meisterernst, M. (1996). A mechanism for repression of class II gene transcription through specific binding of NC2 to TBP-promoter complexes via heterodimeric histone fold domains. *Embo J* 15, 3105-16.
- Grunstein, M. (1990). Histone function in transcription. *Annu Rev Cell Biol* 6, 643-78.
- Han, M., and Grunstein, M. (1988). Nucleosome loss activates yeast downstream promoters in vivo. *Cell* 55, 1137-45.
- 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.
- Hoffman, C. S., and Winston, F. (1987). A ten-minute DNA preparation from yeast efficiently releases autonomous plasmids for transformation of *Escherichia coli*. *Gene* 57, 267-72.
- Inostroza, J. A., Mermelstein, F. H., Ha, I., Lane, W. S., and Reinberg, D. (1992). Dr1, a TATA-binding protein-associated phosphoprotein and inhibitor of class II gene transcription. *Cell* 70, 477-89.
- Kim, J., Parvin, J. D., Shykind, B. M., and Sharp, P. A. (1996). A negative cofactor containing Dr1/p19 modulates transcription with TFIIA in a promoter-specific fashion. *J Biol Chem* 271, 18405-12.
- Kim, S., Na, J. G., Hampsey, M., and Reinberg, D. (1997). The Dr1/DRAP1 heterodimer is a global repressor of transcription in vivo. *Proc Natl Acad Sci U S A* 94, 820-5.

- Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H., and Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* 77, 599-608.
- Koh, S., Ansari, A., Ptashne, M., and Young, R. (1998). An activator target in the RNA polymerase II holoenzyme. *Mol Cell* 1, p895-904.
- Koh, S. S., Hengartner, C. J., and Young, R. A. (1997). Baculoviral transfer vectors for expression of FLAG fusion proteins in insect cells. *Biotechniques* 23, 622-4, 626-7.
- Koleske, A. J., Chao, D. M., and Young, R. A. (1996). Purification of yeast RNA polymerase II holoenzymes. *Methods Enzymol* 273, 176-84.
- Koleske, A. J., and Young, R. A. (1994). An RNA polymerase II holoenzyme responsive to activators. *Nature* 368, 466-9.
- Lee, T. I., and Young, R. A. (1998). Regulation of gene expression by TBP-associated proteins. *Genes Dev* 12, 1398-408.
- Lee, Y. C., Min, S., Gim, B. S., and Kim, Y. J. (1997). A transcriptional mediator protein that is required for activation of many RNA polymerase II promoters and is conserved from yeast to humans. *Mol Cell Biol* 17, 4622-32.
- Lenfant, F., Mann, R. K., Thomsen, B., Ling, X., and Grunstein, M. (1996). All four core histone N-termini contain sequences required for the repression of basal transcription in yeast. *Embo J* 15, 3974-85.
- Liao, S. M., Taylor, I. C., Kingston, R. E., and Young, R. A. (1991). RNA polymerase II carboxy-terminal domain contributes to the response to multiple acidic activators in vitro. *Genes Dev* 5, 2431-40.
- Liao, S. M., Zhang, J., Jeffery, D. A., Koleske, A. J., Thompson, C. M., Chao, D. M., Viljoen, M., van Vuuren, H. J., and Young, R. A. (1995). A kinase-cyclin pair in the RNA polymerase II holoenzyme. *Nature* 374, 193-6.
- Liu, H., Badarinarayana, V., Audino, D., Rappsilber, J., Mann, M., and Denis, C. (1998). The NOT proteins are part of the CCR4 transcriptional complex and affect gene expression both positively and negatively. *EMBO J* 17, p1096-106.
- Maldonado, E., Shiekhhattar, R., Sheldon, M., Cho, H., Drapkin, R., Rickert, P., Lees, E., Anderson, C. W., Linn, S., and Reinberg, D. (1996). A human RNA polymerase II complex associated with SRB and DNA-repair proteins. *Nature* 381, 86-9.

- Meisterernst, M., and Roeder, R. G. (1991). Family of proteins that interact with TFIID and regulate promoter activity. *Cell* *67*, 557-67.
- Mermelstein, F., Yeung, K., Cao, J., Inostroza, J. A., Erdjument-Bromage, H., Egelson, K., Landsman, D., Levitt, P., Tempst, P., and Reinberg, D. (1996). Requirement of a corepressor for Dr1-mediated repression of transcription. *Genes Dev* *10*, 1033-48.
- Myers, L. C., Gustafsson, C. M., Bushnell, D. A., Lui, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R. D. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. *Genes and Development* *12*, 45-54.
- Oberholzer, U., and Collart, M. A. (1998). Characterization of NOT5 that encodes a new component of the Not protein complex. *Gene* *207*, 61-69.
- Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes Dev* *10*, 2657-83.
- Orr-Weaver, T. L., Szostak, J. W., and Rothstein, R. J. (1983). Genetic applications of yeast transformation with linear and gapped plasmids. *Methods Enzymol* *101*, 228-45.
- Ossipow, V., Tassan, J. P., Nigg, E. A., and Schibler, U. (1995). A mammalian RNA polymerase II holoenzyme containing all components required for promoter-specific transcription initiation. *Cell* *83*, 137-46.
- Pan, G., Aso, T., and Greenblatt, J. (1997). Interaction of elongation factors TFIIIS and elongin A with a human RNA polymerase II holoenzyme capable of promoter-specific initiation and responsive to transcriptional activators. *J Biol Chem* *272*, 24563-71.
- Poon, D., Campbell, A. M., Bai, Y., and Weil, P. A. (1994). Yeast Taf170 is encoded by MOT1 and exists in a TATA box-binding protein (TBP)-TBP-associated factor complex distinct from transcription factor IID. *J Biol Chem* *269*, 23135-40.
- Poon, D., and Weil, P. A. (1993). Immunopurification of Yeast TATA-binding Protein and Associated Factors. *J. Biol. Chem.* *268*, 15325-15328.
- Prelich, G. (1997). *Saccharomyces cerevisiae* BUR6 encodes a DRAP1/NC2alpha homolog that has both positive and negative roles in transcription in vivo. *Mol Cell Biol* *17*, 2057-65.
- Prelich, G., and Winston, F. (1993). Mutations that suppress the deletion of an upstream activating sequence in yeast: involvement of a protein kinase and histone H3 in repressing transcription in vivo. *Genetics* *135*, 665-76.

- Recht, J., Dunn, B., Raff, A., and Osley, M. A. (1996). Functional analysis of histones H2A and H2B in transcriptional repression in *Saccharomyces cerevisiae*. *Mol Cell Biol* *16*, 2545-53.
- Reese, J. C., Apone, L., Walker, S. S., Griffin, L. A., and Green, M. R. (1994). Yeast TAFIIS in a multisubunit complex required for activated transcription. *Nature* *371*, 523-7.
- Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends Biochem Sci* *21*, 327-35.
- Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor: Cold Spring Harbor Laboratory Press).
- Scafe, C., Chao, D., Lopes, J., Hirsch, J. P., Henry, S., and Young, R. A. (1990). RNA polymerase II C-terminal repeat influences response to transcriptional enhancer signals. *Nature* *347*, 491-4.
- Schiestl, R. H., and Gietz, R. D. (1989). High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr Genet* *16*, 339-46.
- Shi, X., Chang, M., Wolf, A. J., Chang, C. H., Frazer-Abel, A. A., Wade, P. A., Burton, Z. F., and Jaehning, J. A. (1997). Cdc73p and Paf1p are found in a novel RNA polymerase II-containing complex distinct from the Srbp-containing holoenzyme. *Mol Cell Biol* *17*, 1160-9.
- Smith, D. B., and Johnson, K. S. (1988). Single-step purification of polypeptides expressed in *Escherichia coli* as fusions with glutathione S-transferase. *Gene* *67*, 31-40.
- 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*, 1361-75.
- Thompson, C. M., and Young, R. A. (1995). General requirement for RNA polymerase II holoenzymes in vivo. *Proc Natl Acad Sci U S A* *92*, 4587-90.
- Wade, P. A., and Jaehning, J. A. (1996). Transcriptional corepression in vitro: a Mot1p-associated form of TATA-binding protein is required for repression by Leu3p. *Mol Cell Biol* *16*, 1641-8.
- Wan, J. S., Mann, R. K., and Grunstein, M. (1995). Yeast histone H3 and H4 N termini function through different GAL1 regulatory elements to repress and activate transcription. *Proc Natl Acad Sci U S A* *92*, 5664-8.

White, R. J., Khoo, B. C., Inostroza, J. A., Reinberg, D., and Jackson, S. P. (1994). Differential regulation of RNA polymerases I, II, and III by the TBP-binding repressor Dr1. *Science* 266, 448-50.

Wolffe, A. P. (1997). Histones, nucleosomes and the roles of chromatin structure in transcriptional control. *Biochem Soc Trans* 25, 354-8.

Yeung, K., Kim, S., and Reinberg, D. (1997). Functional dissection of a human Dr1-DRAP1 repressor complex. *Mol Cell Biol* 17, 36-45.

Yeung, K. C., Inostroza, J. A., Mermelstein, F. H., Kannabiran, C., and Reinberg, D. (1994). Structure-function analysis of the TBP-binding protein Dr1 reveals a mechanism for repression of class II gene transcription. *Genes Dev* 8, 2097-109.