From DNA Sequence to Chromatin Dynamics: Computational Analysis of Transcriptional Regulation

Thesis submitted for the degree of

"Doctor of Philosophy"

by

Tommy Kaplan

Submitted to the Senate of the Hebrew University
May 2008

This work was carried out under the supervision of
Prof. Nir Friedman and Prof. Hanah Margalit

Abstract

All cells of a living organism share the same DNA. Yet, they differ in structure, activities and interactions. These differences arise through a tight regulatory system which activates different genes and pathways to fit the cell's specialization, condition, and requirements. Deciphering the regulatory mechanisms underlying a living cell is one of the fundamental challenges in biology. Such knowledge will allow us to better understand how cells work, how they respond to external stimuli, what goes wrong in diseases like cancer (which often involves disruption of gene regulation), and how it can be fought. In my PhD, I focus on regulation of gene expression from three perspectives. First, I present an innovative algorithm for identifying the target genes of novel transcription factors, based on their protein sequence (Chapter 1). Second, I consider how several transcription factors cooperate to process external stimuli and alter the behavior of the cell (Chapter 2). Finally, I study how the genomic position of nucleosomes and their covalent modifications modulate the accessibility of DNA to transcription factors, thus adding a fascinating dimension to transcriptional regulation (Chapters 3 and 4).

To understand transcriptional regulation, one should first reconstruct the architecture of the cell's regulatory map, thus identifying which genes are regulated by which transcription factors (TFs). As the experimental approaches for mapping protein-DNA interactions are expensive and laborious, they are often accompanied by computational algorithms. These complementary approaches analyze the experimentally verified target genes of a TF, identify short sequence elements in their DNA regulatory regions, and represent them using a probabilistic model (DNA motif). Finally, this motif is used to scan the regulatory DNA regions of additional genes, and identify putative binding sites. Such methods were shown useful, mainly for TFs with enough experimental data to accurately characterize the DNA binding preferences. But what about all the transcription factors with no such extensive data? In my dissertation, I developed a novel structure-based approach applicable also to transcription factors with no prior binding data (Chapter 1). This approach combines sequence data with structural information to identify the residues that directly contact with the DNA, and estimates their nucleotide recognition preferences. Given the sequence of a novel protein from the same structural family, we identify the DNA-binding residues and then use the recognition preferences to construct a probabilistic

model of the DNA sequences it binds. I demonstrated this approach on the C2H2 Zinc Finger protein family, showing high compatibility between the learned DNA-recognition preferences and experimental results. I then predicted the DNA motifs of 29 *Drosophila melanogaster* C2H2 transcription factors, and performed a genome-wide scan for their putative target genes. By analyzing the predicted targets for each TF, along with gene annotation and gene expression data, I showed how the function and activity levels of these proteins can be automatically inferred.

Alternatively, high-throughput experimental assays can be applied to directly map TFs and their target genes on a genomic scale. Together with the experimental group of Erin O'Shea (HHMI/Harvard), I developed a novel analytical approach where the expression level of genes is compared between wild-type and mutant yeast strains (Chapter 2). Our method quantifies the exact contribution each TF has on every target gene in several environmental conditions. We applied our method to gain insights into the mechanistic structure of a prototypical example of transcriptional regulatory networks. We focused on the well-studied HOG signaling network, which controls the response of budding yeast to hyper-osmotic stress. In brief, following external signaling, the MAP kinase Hog1 is imported into the nucleus, where it phosphorylates (and activates) several downstream transcription factors. We reconstructed an accurate and quantitative model of the HOG pathway, and analyzed how it interacts with the general stress (Msn2/4) pathway. This study resulted with a regulatory map, based on the expression level of genes in wild-type and mutant strains. In addition, we reconstructed two complementary regulatory maps. We used chromatin immunoprecipitation assays coupled with high-resolution DNA microarrays, to probe the *in vivo* location of transcription factors along the DNA. To analyze these data, I developed a model-based computational algorithm for identifying the exact position and affinity of genomic binding events. The third regulatory map identifies putative target genes of HOG-related factors by computationally scanning promoter regions for their known DNA motifs. As my analysis shows, all three regulatory maps (based on gene expression, on physical binding, and on motif analysis) coincide in a statistically significant manner - most promoters that contained the DNA motif of a factor were indeed physically bound by it, and their gene expression levels affected by its presence. Yet, we found many additional examples of *latent* binding sites which are not occupied, as well as other sites occupied but *non-functional*. These discrepancies suggest that additional higher-order mechanisms are involved in transcriptional regulation, including the packaging of DNA onto chromatin. Furthermore, it paves the way to identifying and characterizing the role of signal processing in gene regulatory networks through combinatorial regulation of gene expression.

To understand the role of chromatin in transcriptional regulation, I focused on the information stored in the packaging of DNA *per se*. This includes the position of nucleosomes along the DNA, as well as their covalent modifications (*e.g.*, by acetylation and methylation). Both these mechanisms, together with the methylation of DNA, were shown to be involved in the occlusion of DNA sites for transcriptional factors, resulting in various degrees of repression of gene expression.

To directly assay the chromatin state in living cells, I collaborated with the experimental group of Oliver Rando (Harvard/UMass). We used high-resolution tiling arrays to investigate the occurrence of 12 histone modifications on thousands of nucleosomes in the budding yeast (Chapter 3). We found that the 12 histone modifications can be roughly split into two groups of co-occurring (or redundant) modifications. While the first group of modifications was found to be independent of transcription and marks the two nucleosomes surrounding transcription start sites, the other group occurs in gradients through the coding regions of genes, and is strongly associated with their transcription level. Our results oppose the "histone code" hypothesis, and show that histone modifications do not follow a simple and discrete code as previously thought. Nonetheless, my analysis does indicate that the state of chromatin encodes valuable information regarding the relative position and expression levels of underlying genes.

It is also intriguing to reveal what happens when the internal or external state of the cell changes. It was showed that the chromatin alters its state within minutes, as if to match the new transcriptional program needed upon the change. Such alterations can be achieved by recruiting specific chromatin modifying enzymes (such as histone acetylases and histone methylases) to directly modify the chromatin state, or by replacing the old nucleosomes by newer ones altogether. Turnover of nucleosomes was known to exist during replication, when about half the nucleosomes are evicted and transferred to the daughter cell, but does it also happen in a replication-independent manner? To directly examine this, I continued the collaboration with the Rando lab. We designed a pulse experiment using tagged histones, and measured the ratio of old to new nucleosomes along the genome in a time-series. To analyze these data, I developed a mathematical model based on rate equations, and a simple algorithm to estimate the turnover rate at each genomic position (Chapter 4). Surprisingly, we found that nucleosomes are indeed replaced both during and independent of DNA replication, in a wide variety of rates. We showed that this can be partially explained by transcription, as turnover rates correlate with polymerase density over coding regions. Nonetheless, our most important and surprising result is showing that the highest turnover rates are found at promoters and regulatory regions, rather than in the coding regions. We believe these high turnover rates reflect a cellular mechanism to constantly update the chromatin state of regulatory regions, and act as barriers that prevent the spreading of chromatin states between neighboring genes.

Contents

Introduction

All cells of a living organism share the same DNA. Yet, they manifest tremendous variability in their structure, activities and interactions. These differences arise through the differential deployment of the cells' common genetic toolkit, whose basic principles are simple (Figure 1). Specialized proteins (called transcription factors) bind regulatory DNA elements in a sequence-specific manner and, once bound, modulate the expression of neighboring genes (1995; Ptashne and Gann, 2002). As straightforward as this may sound, years after sequencing the first genome, we still know very little about how this regulatory information is actually encoded in the genome.

Deciphering the basic principles of regulation underlying a living cell is a major challenge in biology. Such knowledge would allow us to better understand how cells work, how they respond to external stimuli, what goes wrong in diseases like cancer (which often involves disruption of gene regulation), and how they can be fought. Also, accumulating evidence suggests that much of the phenotypic variability within the human population arises from sequence variations that alter gene expression (Levy *et al.*, 2007). Recognizing and predicting the consequences of this variation have the potential to revolutionize medicine by allowing the personalization of preventative and therapeutic measures (Sadee and Dai, 2005; Hoffman, 2007; Wheeler *et al.*, 2008).

1. Overview

In my PhD I developed and utilized computational, mathematical and statistical methodologies, and analyzed a wide range of biological data to study different aspects of eukaryotic transcriptional regulation.

In Chapter 1, I present an innovative algorithm to identify the binding sites of transcription factors based on their sequence (Kaplan *et al.*, 2005). In this work, I showed how information about the DNA binding residues of a protein, together with learned binding preferences between its amino acids and DNA bases, can be used to predict the binding sites of novel transcription factors of the same structural family. This opens the way to automatically identify putative target genes of a transcription factor, providing valuable information regarding its function and activity.

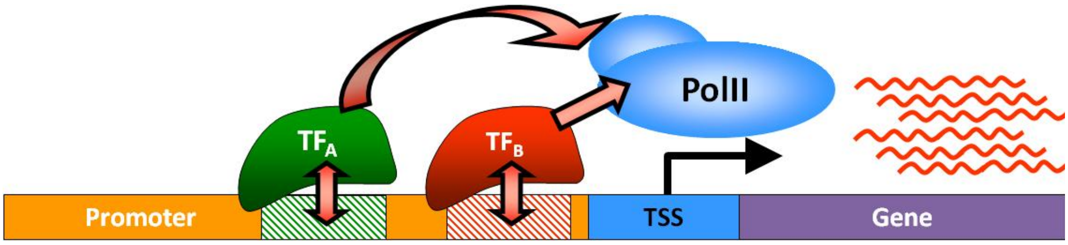


Figure 1. The basics of transcriptional regulation

Regulatory regions of DNA (or promoters; orange) are typically found upstream to the DNA sequence of a gene (purple). Transcription factors can then bind sequence-specific sites within the promoter (shown in green/red), recruit the transcriptional machinery and RNA polymerase II (blue) to the transcription start site (TSS), and bring upon the transcription of multiple mRNA copies of the gene.

In the second part of my research, presented in Chapter 2, I collaborated with the experimental lab of Erin O'Shea (HHMI/Harvard), to understand how external stimuli are processed and integrated by multiple transcription factors. In this study, we also addressed fundamental questions regarding the involvement of chromatin in regulation of gene expression in eukaryotes.

These higher order aspects of transcriptional regulation are further analyzed in the third part of my thesis, presented in Chapters 3 and 4. I have collaborated with the experimental group of Oliver Rando (Harvard/UMass) in two pioneering studies on the role of chromatin in transcriptional regulation and its temporal dynamics (Liu *et al.*, 2005; Dion *et al.*, 2007). For the first time, we were able to characterize the covalent modification patters of thousands of nucleosomes at a single-nucleosome resolution (Liu *et al.*, 2005). This unique exciting data allowed us to address the first principles of the “epigenetic code”, thought to control the state of the chromatin (Strahl and Allis, 2000). Our results were further supported by additional studies with similar results in yeast (Pokholok *et al.*, 2005) and higher organisms (Bernstein *et al.*, 2005; Barski *et al.*, 2007; Bernstein *et al.*, 2007; Mikkelsen *et al.*, 2007). Finally, I continued my collaboration with the Rando lab, now focusing on the dynamics of nucleosome exchange (Dion *et al.*, 2007). We designed a pulse experiment to study the locus-specific incorporation rates of tagged nucleosomes in replication-coupled and replication-independent manners. I developed a mathematical model based on rate equations in non-homogenous Poisson processes, and analyzed the time series experimental data. Our results revealed that nucleosomes are replaced in both replication-coupled and replication-independent manners. Furthermore, we showed that nucleosome exchange occurs in higher rates than previously thought, and suggested possible functionalities for this striking phenomenon. Our results were further supported by additional studies in yeast (Jamai *et al.*, 2007; Rufiange *et al.*, 2007) and fruit fly (Mito *et al.*, 2007).

2. Understanding eukaryotic transcriptional regulation

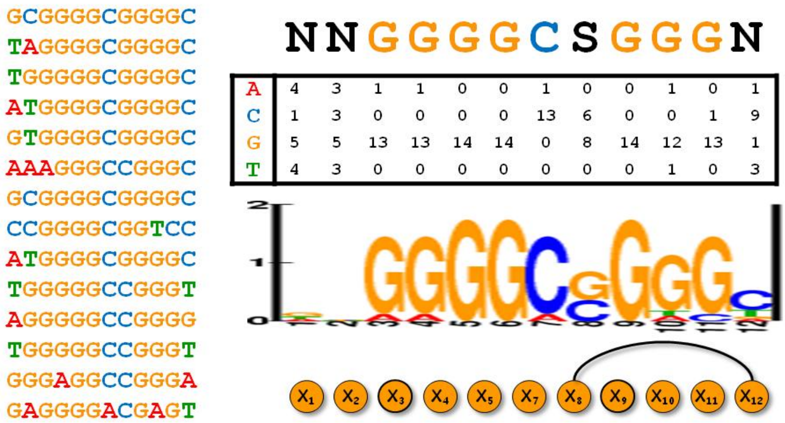
The basic principles of transcriptional regulation in eukaryotic cells are quite simple. The DNA sequence of genes is preceded by a regulatory region of DNA (often called promoter) to which specialized proteins called transcription factors (TFs) bind. Each transcription factor typically binds a relatively short sequence-specific DNA site (~6-20bp). Once bound, these factors can recruit the general transcriptional machinery and bring upon the transcription of multiple mRNA copies of the gene (Figure 1). Alternatively, transcription factors can act as transcriptional repressors, by binding to the promoter region and inhibiting the expression of the gene.

To shed light on the tight transcriptional control governing the expression levels of genes in an eukaryotic cell, one must first reconstruct a transcriptional blueprint (or a regulatory map), specifying which genes are being regulated by which transcription factors. A common practice in identifying protein-DNA interactions is through experimental assays, including biochemical and molecular assays and genetic manipulations of regulatory DNA regions (Latchman, 1995). As reliable and accurate as such direct experimental approaches usually are, they involve laborious experimental work, and are limited to low throughput.

2.1 Modeling and identifying regulatory elements

The recent availability of complete genomic sequences motivated attempts to accompany these experimental methods by complementary computational studies, which will identify protein-DNA interactions through *in silico* analyses. The idea was to build a DNA motif (or a binding site model) for each transcription factor, specifying its DNA-binding preference using a probabilistic model (Stormo, 2000). Then, these descriptive models can be used to scan regulatory regions and identify additional genes targeted by the same factor.

Toward this goal, experimentally verified transcription factor binding sites were extracted from databases such as TRANSFAC (Wingender *et al.*, 2001) or JASPAR



GCGGGGCGGGGC
TAGGGGCGGGGC
TGGGGGCGGGGC
ATGGGGCGGGGC
GTGGGGCGGGGC
AAAGGGCCGGGC
GCGGGGCGGGGC
CCGGGGCGGTCC
ATGGGGCGGGGC
TGGGGGCCGGGT
AGGGGGCCGGGG
TGGGGGCCGGGT
GGGAGGCCGGGA
GAGGGGACGAGT

NNGGGGCSGGGN

A	4	3	1	1	0	0	1	0	0	1	0	1
C	1	3	0	0	0	0	13	6	0	0	1	9
G	5	5	13	13	14	14	0	8	14	12	13	1
T	4	3	0	0	0	0	0	0	0	1	0	3

Figure 2. Modeling binding sites

The aligned set of binding sites (left) can be described using a variety of computational models, from a consensus sequence (top right), through a positional count matrix which specifies the abundance of each nucleotide at each position (also visualized as a sequence logo, below), to a dependency model which captures the probabilistic dependencies between positions in the binding sites (bottom).

(Sandelin *et al.*, 2004). These sites were then aligned and described by either a consensus motif or by descriptive probabilistic models which specify the relative abundance of each nucleotide at each position (Figure 2) (Stormo, 2000). Once the binding sites of a transcription factor are described by such a model, it can be further used to scan regulatory regions of other genes, thus identifying additional binding sites and putative target genes of the same factor (Quandt *et al.*, 1995; Bailey and Gribskov, 1998; Barash *et al.*, 2005).

Such models (often referred to as position-specific scoring matrix, or PSSM), inherently assume that positions within the binding site are independent. While this may be true in some cases, experimental results show that for several structural families of transcription factors, inner-dependencies of positions within binding sites exist (Benos *et al.*, 2002; Bulyk *et al.*, 2002). As we and other showed, the probabilistic representations of DNA motifs can be extended to capture such inner dependencies. For most transcription factors, these enhanced models allow for higher accuracy in modeling their binding site, resulting with a better *in silico* reconstruction of transcriptional regulatory maps (Barash *et al.*, 2003; King and Roth, 2003; Zhou and Wong, 2004; Ben-Gal *et al.*, 2005; Sharon and Segal, 2007).

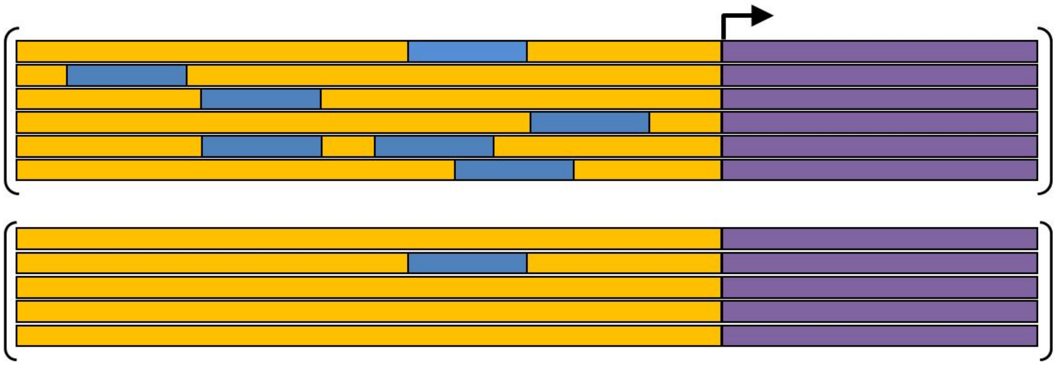


Figure 3. Over-representative motifs in the regulatory sequences of co-regulated genes

High-throughput gene-expression or chromatin immunoprecipitation (ChIP) studies allow the identification of co-expressed and co-regulated genes, respectively (top genes). Statistical and computational algorithms were then developed to identify short motifs (blue) enriched among the promoters of those groups, in comparison to a control set of genes (bottom).

2.2 Using high-throughput experimental data

In the previous sections, I described how regulatory maps of protein-DNA interactions can be reconstructed using low-throughput experimental assays and complementary *in silico* approaches. Although useful, this strategy was limited by the rate in which experimentally verified binding sites could be collected, to characterize the DNA motif for each TF.

About a decade ago, technological advances revolutionized biology by allowing high-throughput assays for sequencing (e.g., the budding yeast *Saccharomyces cerevisiae* genome; (Clayton *et al.*, 1997)), for simultaneously measuring the expression level of thousand of genes using *DNA microarrays* (DeRisi *et al.*, 1997; Holstege *et al.*, 1998; Spellman *et al.*, 1998; Gasch *et al.*, 2000; Hughes *et al.*, 2000) or their *in vivo* binding by regulatory proteins using genomic *chromatin immunoprecipitation* (ChIP) studies (Ren *et al.*, 2000; Iyer *et al.*, 2001; Simon *et al.*, 2001). Suddenly, it was possible to identify which genes alter their expression levels following deletions of transcription factors, and which genes are physically bound by a transcription factor.

To distinguish between direct and indirect target genes, and to overcome experimental noise, the promoter regions of these genes were computationally scanned to identify occurrences of the regulator's DNA motif. This was done by statistical methods which recognized over-represented motifs in the regulatory regions of putative co-regulated sets of genes, in comparison to a control set of genes (Figure 3) (Bailey and

Elkan, 1994; Hertz and Stormo, 1999; Barash *et al.*, 2001; Liu *et al.*, 2001; Barash *et al.*, 2003; Osada *et al.*, 2004).

Such methods were applied to reconstruct the transcriptional regulatory map of the cell in a wide range of organisms, including *E. coli*, yeast, and higher eukaryotes such as worm, fly, and human (Hughes *et al.*, 2000; Lee *et al.*, 2002; Harbison *et al.*, 2004; MacIsaac *et al.*, 2006).

2.3 *Ab initio* prediction of target genes using structural knowledge

As valuable as such high-throughput experiments are in directly mapping protein-DNA interactions, their applications are still limited due to the extensive labor, costly reagents and expensive microarrays required. It was therefore suggested to apply complementary studies which rely on structural knowledge regarding transcription factors, and their DNA binding preferences. Each factor tends to bind sequence-specific sites, according to its structural family and the specific residues through which protein-DNA contacts are accomplished. In general, every amino acid can bind different nucleotides, based on its physic-chemical characteristics. For example, analysis of solved protein-DNA complexes revealed a strong tendency for Arginine and Lysine to interact with Guanine, whereas Glutamic acid tends to interact with Cytosine (Mandel-Gutfreund *et al.*, 1995; Kono and Sarai, 1999). This concept was suggested as a gateway for predicting the DNA motifs of novel proteins, based on their sequence (Kono and Sarai, 1999; Mandel-Gutfreund *et al.*, 2001). In more details, the protein sequence of a query protein can be *threaded,* using the solved structure of another transcription factor used as a *structural template.* This offers a fast and easy way for identifying which residues in the query protein interact with which positions along the DNA. Finally, the physicochemical properties of the amino acids located at these DNA-binding positions are used to predict its DNA motif (Kono and Sarai, 1999; Luscombe *et al.*, 2001; Mandel-Gutfreund *et al.*, 2001; Benos *et al.*, 2002; Benos *et al.*, 2002; Endres *et al.*, 2004; Havranek *et al.*, 2004). Additional studies focused on specific structural families, and showed that the same amino acid may have different binding preferences depending on its positional context (Choo and Klug, 1994; Choo and Klug, 1994; Kono and Sarai, 1999). We found it only natural to further develop such structure-based approaches to predict the DNA motif of transcription factors based on their sequence, while allowing for context-specific amino acid-nucleotide interactions (Chapter 1). Unfortunately, such specific binding preferences involve many parameters, and could not have been accurately estimated from the limited number of solved protein-DNA complexes. Instead, I applied machine learning algorithms to automatically predict the structural alignment of transcription factors and their natural DNA sites, as taken from public databases (Wingender *et al.*, 2001; Sandelin *et al.*, 2004).

3. Experimental studies of transcriptional regulation - Lessons from the HOG pathway

In the previous section I focused on computational and experimental strategies for identifying the target genes of each transcription factor, thus reconstructing transcriptional regulatory maps. I now wish to examine the activity of transcription factors in a broader context, focusing on their role as combinatorial processing units. According to this view, signaling pathways propagate intra- and extra-cellular information regarding the state of the cell and its environment into the eukaryotic nucleus. Then, this information is processed by a network of transcription factors that regulate the expression of genes accordingly (Barrett and Palsson, 2006; Davidson, 2006). For several pathways, detailed circuit diagrams were constructed, showing how signals influence the activity and expression levels of transcription factors, and how these changes are translated to changes in the mRNA expression levels of their target genes (Ben-Tabou de-Leon and Davidson, 2007). For example, regulatory diagrams were constructed for the flagella gene network in *E. coli* (Kalir and Alon, 2004) or for small portions of developmental pathways in higher organisms (Davidson *et al.*, 2002; Levine and Davidson, 2005; Stathopoulos and Levine, 2005).

3.1 Cellular processing of external signals by the HOG pathway

To address this goal, and study how a transcriptional network of several transcription factors processes external signaling and regulate the expression levels of hundreds of genes, we decided to focus on one prototypical pathway as a model system. One of the most studied transcriptional networks is the HOG pathway in budding yeast, controlling the cellular response to hyper-osmotic stress (Figure 4). When yeast cells are exposed to high levels of extra-cellular osmolyte (e.g., salt), they undergo a rapid transcriptional reprogramming, involving hundreds of genes. This facilitates the stimulation of various cellular actions, including glycolysis to enhance the production of glycerol as a compatible solute, diminishing cellular translation levels and cell

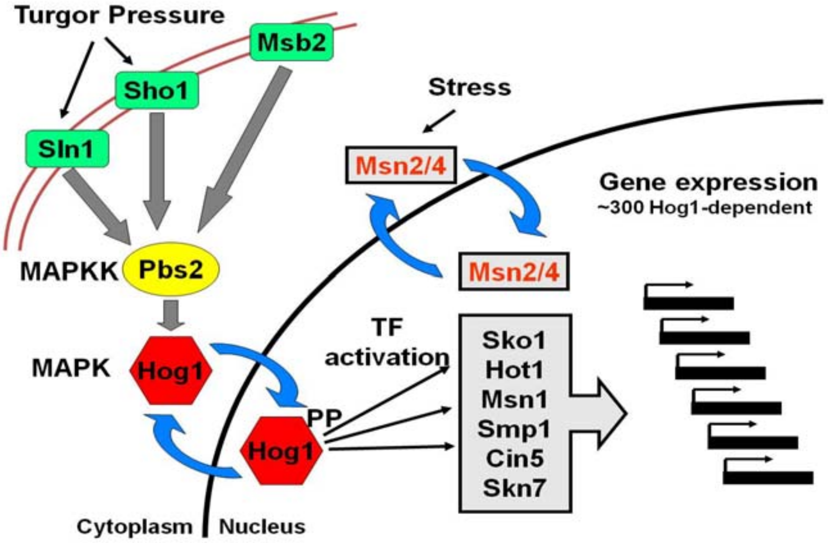


Figure 4. The yeast HOG signaling pathway

Upon induction of extra-cellular turgor pressure, the osmosensors Sho1, Sln1 and Msb2 initiate a signaling cascade which results in phosphorylating the MAP-kinase-kinase Pbs2, which in turn phosphorylates Hog1. The phosphorylated Hog1 is transported into the nucleus, where in activates several TFs.

cycle arrest, repairing the cell wall damage caused by cell shrinkage, as well as a general response to stress (Hohmann *et al.*, 2007). This dramatic switch was shown to be orchestrated by the mitogen-activated protein kinase (MAPK) Hog1, with partial involvement of the paralogous general stress factors Msn2 and Msn4 (Posas *et al.*, 2000; Rep *et al.*, 2000; Yale and Bohnert, 2001; O'Rourke and Herskowitz, 2004). Following activation, Hog1 is imported into the nucleus, where it phosphorylates several downstream transcription factors, which in turn modulate the expression of their specific target genes (Proft and Serrano, 1999; Rep *et al.*, 1999; Rep *et al.*, 2000; Alepuz *et al.*, 2001; Rep *et al.*, 2001; de Nadal *et al.*, 2002; Proft and Struhl, 2002; Alepuz *et al.*, 2003; Proft *et al.*, 2005). Despite intense efforts, our understanding of the transcriptional mechanisms through which the HOG pathway acts, is still rather limited. Gene expression assays in mutant strains suggested that the target genes of these Hog1-regulared factors, as well as the Msn2/4, significantly overlap (Rep *et al.*, 1999; Rep *et al.*, 2000; Alepuz *et al.*, 2001). The extent of this combinatorial control of gene expression, and the mechanisms through which it is achieved, are still unknown.

3.2 Strategies for analyzing high-throughput ChIP-chip and gene expression data

To address these questions, and to dissect the hyperosmotic-induced transcriptional response of the cell into the specific components controlled by each TF, I collaborated with the lab of Erin O'Shea (HHMI/Harvard). We decided to build a more complete model of the HOG pathway by comparing the expression levels of genes in a wide range of genetic and environmental conditions using DNA microarrays (DeRisi *et al.*, 1997; Holstege *et al.*, 1998; Spellman *et al.*, 1998; Gasch *et al.*, 2000; Hughes *et al.*, 2000). We designed a web of multiple partially-overlapping experiments, and measured the expression of yeast genes prior to and following hyper-osmotic stress at a range of mutant strains, including the single and double deletions of the HOG pathway key players (Hog1, Msn2/4, Sko1, and Hot1). To analyze these data and estimate the contribution of each HOG-related TF to the expression level of each gene, I developed a statistical-based regression algorithm (see Methods).

To further validate these expression-based data, which might also include some indirect effects, we decided to directly map the *in vivo* binding of HOG-regulated proteins using chromatin immunoprecipitation coupled with hybridization to DNA microarrays (ChIP-chip; Figure 5). The original ChIP-chip assays (Ren *et al.*, 2000; Iyer *et al.*, 2001; Simon *et al.*, 2001; Lee *et al.*, 2002; Harbison *et al.*, 2004) relied on promoter arrays, and were therefore limited in their resolution to ~1Kb. These arrays allowed to identify which promoters are bound by the TF and which are not, but did not facilitate the exact identification of binding locations. The recent development of high-resolution tiling arrays, allowed for more accurate genome-wide ChIP-chip assays (Cawley *et al.*, 2004; Kim *et al.*, 2005; Pokholok *et al.*, 2005; Qi *et al.*, 2006; Kim *et al.*, 2007; Li *et al.*, 2008) or analysis tools (Buck *et al.*, 2005; Gibbons *et al.*, 2005; Li *et al.*, 2005; Qi *et al.*, 2006). We therefore directly tested the *in vivo* protein-DNA interactions using high-resolution genome wide ChIP-chip assays. To identify the location and affinity of binding events, I developed a computational model-based algorithm for the analysis of high-resolution ChIP-chip data (see Methods). For compatibility, our gene expression and ChIP-chip measurements were done for the same strains and conditions.

Both these assays aimed to quantitatively identify where HOG-related factors bind, and what are their transcriptional effects. To complement this view, we also analyzed

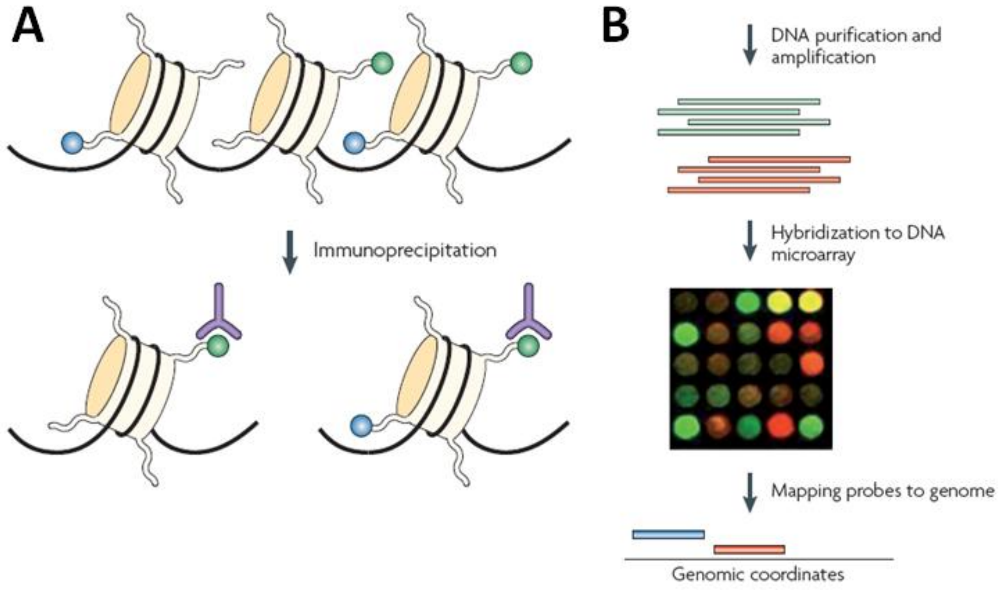

A
Immunoprecipitation
B
DNA purification and amplification
Hybridization to DNA microarray
Mapping probes to genome
Genomic coordinates

Figure 5. Chromatin immunoprecipitation (ChIP) combined with DNA microarrays (ChIP-chip) or high-throughput sequencing (ChIP-Seq)

(A) Modified nucleosomes are immunoprecipitated using modification-specific antibodies (shown in green and purple). **(B)** DNA is amplified, color-labeled and hybridized to a DNA microarray. **(C)** Alternatively to (B), high-throughput sequencing methods (e.g. Illumina's Solexa), purify the DNA, ligate adapters, bind the DNA to a flow cell, amplify, and sequence DNA ends. (Adapted from Zhao *et al*., 2008)

the genomic sequence of yeast, identifying the positions these TF *could* bind. Toward this end, we developed and applied several computational algorithms to characterize and identify HOG-related binding sites (Bailey and Elkan, 1994; Hughes *et al.*, 2000; Liu *et al.*, 2001; Harbison *et al.*, 2004; Barash *et al.*, 2005; Gordon *et al.*, 2005; Habib *et al.*, 2008). All these results are reported in Chapter 2.

3.3 Discrepancies between latent, silent and functional binding sites - Higher order aspects of transcriptional regulation

When the number of genes regulated by a single transcription factor is estimated based on gene expression data, on sequence analysis, or on *in vivo* binding, we come across a very puzzling phenomenon. On the one hand, experimental studies that rely on gene expression data usually estimate the number of targets between 50 and 500 (Holstege *et al.*, 1998; Hughes *et al.*, 2000). On the other hand, computational sequence-based algorithms that scan the regulatory regions of genes for occurrences of known DNA regulatory motifs, typically find thousands of putative target genes for

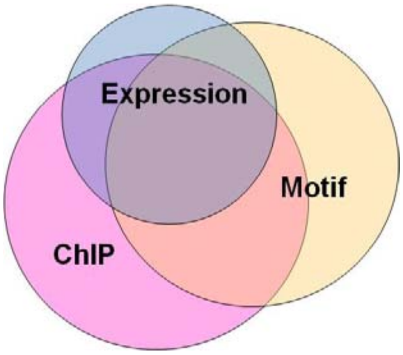


Figure 6. Discrepancies between sequence, ChIP and expression
Putative target genes of a transcription factor, according to gene expression analyses (Expression), in vivo binding (ChIP) and motif analysis (Motif)

each TF (Kaplan *et al.*, 2005; MacIsaac and Fraenkel, 2006; MacIsaac *et al.*, 2006; Yang *et al.*, 2006). When directly measuring the *in vivo* protein-DNA interactions using ChIP-chip, the typical number of bound promoters usually varies between 100 to 1000 (Chapter 2) (Lee *et al.*, 2002; Cawley *et al.*, 2004; Harbison *et al.*, 2004; Wei *et al.*, 2006; Zeitlinger *et al.*, 2007; Li *et al.*, 2008). Yet, the overlap between all these sets of putative targets is only partial (Yang *et al.*, 2006), as illustrated in Figure 6. It is therefore of great interest to understand why DNA regions that contain *bona fide* binding sites are not bound (or latent), and why other binding sites are bound with no observed change in expression level of nearby genes (hence non-functional). Usually these discrepancies are too substantial to be simply justified as experimental noise, or to be explained due to technical aspects (such as the computational representation of DNA motifs or the thresholds used for identifying target genes). Instead, there are a growing number of evidences that link such cases to higher-order mechanisms of transcription regulation. In the next sections, I will address one such mechanism, by reviewing the roles of DNA packaging in transcriptional regulation.

4. The role of chromatin in transcriptional regulation

4.1 Chromatin: DNA packaging, histones & nucleosomes

In eukaryotic cells the DNA is wrapped around nucleosomes, globular complexes of histone proteins, to form the tightly packed chromatin (Figure 7) (Luger *et al.*, 1997). This packaging plays a crucial role in fitting the long chromosomes into the small nuclei and in protecting the DNA from physical damage. Chromatin also plays a functional role in transcriptional regulation, by modulating the affinity of DNA to the transcriptional machinery. On one hand, the occlusion of binding sites for transcription factors was shown to result in transcriptional repression (Venter *et al.*, 1994; Bernstein *et al.*, 2004; Buck and Lieb, 2006). Alternatively, nucleosomes can activate transcription through displacement in spatial juxtaposition of transcription factor binding sites (Stunkel *et al.*, 1997; Workman, 2006; Lam *et al.*, 2008), as if to create a transcriptional funnel (Nemeth and Langst, 2004; Kolesov *et al.*, 2007; Narlikar *et al.*, 2007). This plasticity in nucleosome positioning was shown to be related to ATP-dependent chromatin remodeling enzymes, such as the SWI/SNF and ISWI complexes (Figure 8) (Wu and Winston, 1997; Workman and Kingston, 1998; Vignali *et al.*, 2000; Lusser and Kadonaga, 2003; Langst and Becker, 2004; Cairns, 2005; Whitehouse *et al.*, 2007).

In addition to the exact positioning of nucleosomes, transcription is controlled by covalent modification of the histones. Several residues in the highly-conserved histone proteins are subject to multiple types of covalent modification, including methylation, acetylation, phosphorylation, ubiquitylation, Sumoylation, and ADP-ribosylation (Strahl and Allis, 2000; Turner, 2000; Berger, 2002; Schreiber and Bernstein, 2002; Turner, 2002; Kurdistani and Grunstein, 2003; Kurdistani *et al.*, 2004; Bannister *et al.*, 2005; Bernstein *et al.*, 2005; Liu *et al.*, 2005; Pokholok *et al.*, 2005; Vakoc *et al.*, 2005; Millar and Grunstein, 2006; Nathan *et al.*, 2006; Kouzarides, 2007).

To better understand transcriptional regulation, one must therefore consider the epigenetic context of the DNA sequence, as reflected by the information stored in its packaging.

4.2 Characterization of nucleosome positions

Once the positions of enough nucleosomes were established, several studies identified conserved sequence features that characterize nucleosomal DNA relatively to linker DNA, and suggested sequence-based algorithms for identifying nucleosome positions along genomic regions (Ioshikhes *et al.*, 1996; Ioshikhes *et al.*, 2006; Segal *et al.*, 2006; Peckham *et al.*, 2007; Yuan and Liu, 2008). Although such *in silico* methods are tempting, their accuracy is arguable, with only a small improvement over random positioning (Segal, 2008; Valouev *et al.*, 2008; Yuan and Liu, 2008). Moreover, recent genome-wide studies showed that the position of nucleosomes strongly

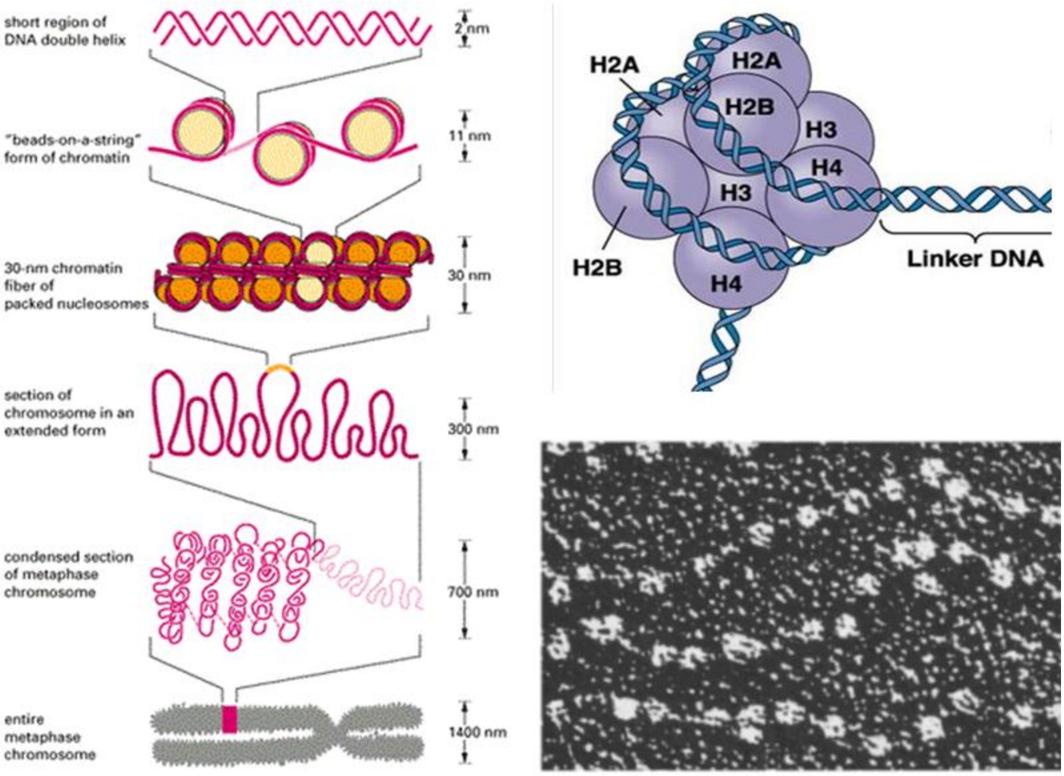

short region of DNA double helix
2 nm
"beads-on-a-string" form of chromatin
11 nm
30-nm chromatin fiber of packed nucleosomes
30 nm
section of chromosome in an extended form
300 nm
condensed section of metaphase chromosome
700 nm
entire metaphase chromosome
1400 nm
H2A
H2A
H2B
H3
H4
H3
H2B
H4
Linker DNA

Figure 7. DNA packaging

Various packaging degrees of the DNA (left). The double-stranded DNA is wrapped around nucleosomes - globular complexes of histone proteins (top right). This initial packaging (often referred to as "beads on a string", bottom right) can be further packed in additional layers, to produce the condensed chromosomes.

depends on chromatin remodeling enzymes (Whitehouse *et al.*, 2007) and dynamically changes to reflect reprogramming of the transcriptional program in changing environmental conditions (Schones *et al.*, 2008; Shivaswamy *et al.*, 2008).

High-throughput methods for determining the positioning of nucleosomes usually involve the breaking of genomic DNA (e.g., by sonication) followed by nucleosomal chromatin immunoprecipitation (ChIP) coupled with dense tiling DNA microarrays (Pokholok *et al.*, 2005; Schones and Zhao, 2008). For more accurate results, nucleosomal DNA can be better purified by linker DNA digestion with micrococcal nuclease (MNase) (Raisner *et al.*, 2005; Yuan *et al.*, 2005; Lee *et al.*, 2007; Whitehouse *et al.*, 2007). Alternatively, massive sequencing techniques such as Roche's 454 or Illumina's Solexa 1G sequencers, were harnessed to replace the use of tiling arrays (ChIP-Seq) (Albert *et al.*, 2007; Barski *et al.*, 2007; Mikkelsen *et al.*, 2007; Mavrich *et al.*, 2008; Shivaswamy and Iyer, 2008). Nucleosomal positions, and

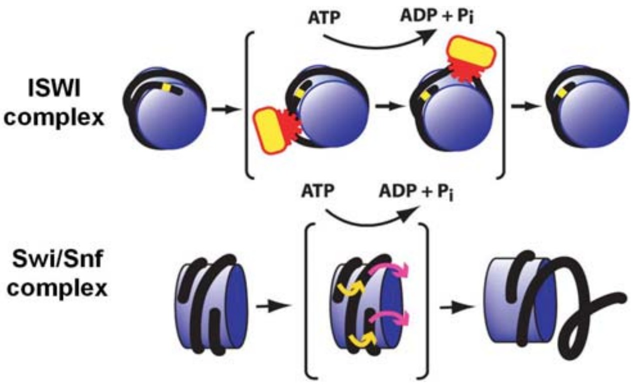


Figure 8. ATP-dependent chromatin remodeling

The ISWI sub-family complex (top) is an ATP-dependent DNA-translocating enzyme that disrupts the nucleosome-DNA contact to allow reestablishment at a different position relative to the DNA. In contrary, the Swi/Snf complex (bottom) rearranges one superhelical turn of nucleosomal DNA, causing the other superhelical turn of DNA to be displaced from the nucleosome (Lusser & Kadonaga, 2003)

in some cases, nucleosomal occupancies, can then be inferred by analyzing these data (Yuan *et al.*, 2005; Ozsolak *et al.*, 2007; Schones *et al.*, 2008; Yassour *et al.*, 2008).

4.3 Mapping histone modifications

To date, over 60 different histone residues are known to be covalently modified, and the actual number of modifications taking place in chromatin further exceeds this. Methylation of lysines or arginines may take several forms (e.g., mono-, di-, or tri-methylation of lysines), whereas some positions can be both acetylated and methylated (Figure 9) (Berger, 2002; Turner, 2002; Peterson and Laniel, 2004; Kouzarides, 2007).

These modifications were shown to involve a wide range of enzymes, including histone acetyltransferases and deacetylases (Sterner and Berger, 2000), methyltransferases and demethylases (Zhang and Reinberg, 2001) and more. Although the "division of labor" between histone modifiers and their target residues is not yet well characterized, the complex regulation of histone modifications and their relation to transcription factors (both sequence-specific and general) is now beginning to be revealed (Strahl and Allis, 2000; Guo *et al.*, 2006; Kouzarides, 2007; Pham *et al.*, 2007; Steinfeld *et al.*, 2007). Unique patterns of chromatin modifications were shown to be connected to various processes along the DNA. DNA repair damage involves the H2A variant H2AX, the phosphorylation of H2AS129 and H4S1, the

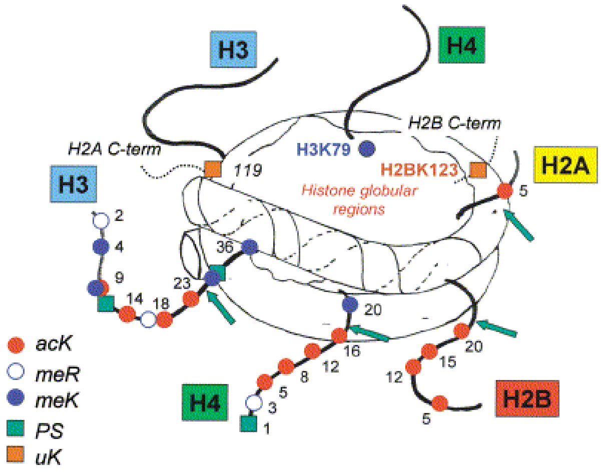


Figure 9. Histone modifications

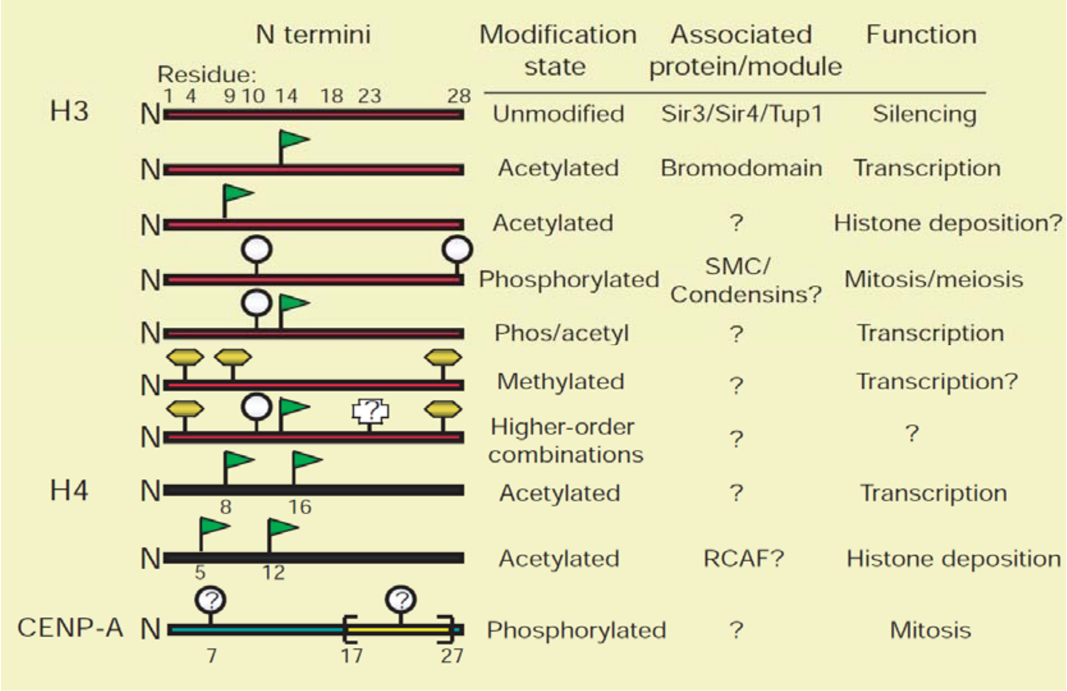
Various residues along the histone tail domains are subject to covalent modifications, including acetylated lysine, methylated arginine, methylated lysine, phosphorylated serine, and ubiquitinated lysine (Turner, 2002).

methylation of H4K20, and the acetylation of H3K56 (Kouzarides, 2007). During replication, the acetylation of specific residues at the N-terminus of histones H3 (K4) and H4 (K5, 8, and 12) is related to the activation of origins of replications (Kouzarides, 2007). Moreover, studies in animal cells showed how different chromatin states are associated with different chromosome condensation and replication times (Zhang *et al.*, 2002; Hashimshony *et al.*, 2003; Lin *et al.*, 2003; Fischle *et al.*, 2005; Wu *et al.*, 2005), mostly through the methylation of H3K9 and of the DNA.

Yet, the most studied aspect of the chromatin state is in transcription control, where it modulates the availability of DNA and the expression of underlying genes through several mechanisms (Shilatifard, 2006; Workman, 2006; Li *et al.*, 2007). For example, acetylation of lysines neutralizes their positive charge, thus weakening their affinity to the negatively charged DNA (Hong *et al.*, 1993) and to neighboring nucleosomes (Luger *et al.*, 1997). In addition, acetylated lysines can be bound by bromodomain transcription factors (Dhalluin *et al.*, 1999; Peterson and Laniel, 2004). While methylation of lysines does not affect their charge, it is recognized by the chromodomain transcription factors (Bannister *et al.*, 2001; Lachner *et al.*, 2001). Moreover, histone modifications offer an unprecedented platform for epigenetic memory. Few examples include tri-methylation of H3K4 by Set1 histone methylase which marks active promoters even after the transcriptional inactivation (Ng *et al.*, 2003), or the well ordered activation of the HO promoter in the yeast cell-cycle, by a combination of chromatin remodeling enzymes (Cosma *et al.*, 1999).

It was therefore hypothesized that different modifications (or combinations of such) are associated with distinct transcriptional contexts (Figure 10) (Strahl and Allis, 2000). This hypothesis, known as the "histone code hypothesis", was characterized in several theoretical studies (Turner, 2000; Berger, 2002; Turner, 2002). Technological advances allowed experimental testing of this hypothesis in a high-throughput manner, using immunoprecipitation of modified nucleosomes. The first genomic studies used promoter-based DNA microarrays, with relatively long probes (~1Kb). This resulted with low resolution data, typically the average modification patterns over ~5-6 neighboring nucleosomes (Bernstein *et al.*, 2002; Robyr *et al.*, 2002; Robyr and Grunstein, 2003; Kurdistani *et al.*, 2004; Schubeler *et al.*, 2004). Later studies allowed the characterization of histones modifications in higher resolution. For example, Pokholok *et al.,* (2005) used tiling microarrays of the entire yeast genome, allowing a resolution of ~4 nucleosomes. In the study presented in Chapter 3 (Liu *et al.*, 2005), we used a combination of accurate linker DNA digestion with much denser arrays to achieve single-nucleosome characterization of nucleosome modification. Our analysis showed that the packaging of DNA encodes a fair amount of information regarding the transcription level of underlying genes and the position of nucleosomes with regard to the transcription start site. Yet, we found no evidence of a discrete 'histone code', as proposed before (Liu *et al.*, 2005). Instead, the acetylation and methylation patterns of nucleosomes were found to be relatively smooth and redundant, changing gradually along the promoters and coding regions of genes, with several differences between active and silenced genes (Liu *et al.*, 2005; Millar *et al.*, 2006; Yuan *et al.*, 2006; Li *et al.*, 2007; Rando, 2007; Steinfeld *et al.*, 2007).

These findings were further supported by recent studies in mammalian models, which used chromatin immunoprecipitation followed by dense tiling arrays (Bernstein *et al.*, 2005; Koch *et al.*, 2007) or by massive sequencing (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007). Yet, the exact characterization of chromatin modifications and their roles in transcriptional regulation are still under intense scrutiny and dispute.



N termini	Modification state	Associated protein/module	Function
H3	Unmodified	Sir3/Sir4/Tup1	Silencing
	Acetylated	Bromodomain	Transcription
	Acetylated	?	Histone deposition?
	Phosphorylated	SMC/ Condensins?	Mitosis/meiosis
	Phos/acetyl	?	Transcription
	Methylated	?	Transcription?
	Higher-order combinations	?	?
H4	Acetylated	?	Transcription
	Acetylated	RCAF?	Histone deposition
CENP-A	Phosphorylated	?	Mitosis

Figure 10. The histone code hypothesis

Histone modifications at selected residues were hypothesized to be related to biological events in a combinatorial code. (Strahl & Allis, 2000)

4.4 Dynamics in chromatin

Chromatin packaging of DNA and its relation to almost every process that takes place along the DNA raises questions regarding the temporal dynamics of chromatin. On one hand, the tight connection to transcriptional activity suggests high plasticity of chromatin, both in terms of nucleosome positioning and the covalent modification of histones. Indeed, histone acetylation was shown to be rapidly reversible within minutes (Vogelauer *et al.*, 2000; Waterborg, 2001) and to reflect transcriptional reprogramming of the cell (Berger, 2002; Krebs, 2007; Schones *et al.*, 2008; Shivaswamy *et al.*, 2008). On the other hand, histone modifications are involved in forming stable epigenetic marks which persist over multiple cycles of replication (Kouzarides, 2007). Such mechanisms involve additional aspects of chromatin, such as the methylation of DNA, or the formation of higher-order heterochromatic structures (Rea *et al.*, 2000; Noma *et al.*, 2001; Peters *et al.*, 2002; Rice *et al.*, 2003; Grewal and Rice, 2004; Lande-Diner and Cedar, 2005).

To address the mechanisms involved in such a wide range of time scales, the dynamics of nucleosome exchange in living cells should be directly probed.

Photobleaching studies of fluorescent histones showed that the core histone H3-H4 tetramers are relatively stable, whereas the H2A-H2B dimers are replaced during transcription. These studies also characterized the dynamics of additional histone variants such as H3.3 or H2A.Z (Kimura and Cook, 2001; Kimura, 2005). Yet, such microscopical studies examine the behavior of the chromatin in bulk, and were soon followed by locus-specific assays that suggested that nucleosomes are evicted upon activation of their underlying genes and are reassembled in trans upon repression (Boeger *et al.*, 2003; Reinke and Horz, 2003; Bernstein *et al.*, 2004; Kristjuhan and Svejstrup, 2004; Lee *et al.*, 2004; Schwabish and Struhl, 2004; Shivaswamy and Iyer, 2008).

Those assays showed that specific histones are sometimes exchanged, and that nucleosomal occupancy is related to passages of the transcriptional machinery. Yet, the question of nucleosome turnover remained. In replication, about half of the nucleosomes are passed to the daughter cell, and are replaced by novel newly synthesized nucleosomes. It was not clear whether nucleosomes are also exchanged in a replication-independent manner. Recent studies used chromatin immuno-precipitation of epitope-tagged histone variants, to show that indeed there is replication-independent turnover of nucleosomes in eukaryotes, ranging from the budding yeast *Saccharomyces cerevisiae* to the fruit fly *Drosophila melanogaster* (See Chapter 4) (Linger and Tyler, 2006; Dion *et al.*, 2007; Jamai *et al.*, 2007; Mito *et al.*, 2007; Rufiange *et al.*, 2007). These studies also highlighted the functional context of nucleosome turnover, and suggested its involvement in several processes, including DNA damage checkpoints following replication, transcriptional control through binding site occlusion, transcriptional insulation via chromatin barriers, and a mechanism for resetting outdated histone modification patterns.

Goals of Research

In my research I aimed to learn about the various mechanisms that control the expression of genes in eukaryotes. Toward this end, I developed and utilized computational tools to integrate various types of genome-scale experimental data.

My first goal was to reconstruct regulatory transcriptional networks from heterogeneous experimental and genomic data. This required that the direct interactions between transcription factors and their target genes were identified. Toward this end, and to accompany experimental high-throughput data of protein-DNA interactions, I developed a computational method based on sequence and structural information. I aimed to use solved protein-DNA complexes as structural templates for novel proteins, thus identifying their DNA-binding positions. I then intended to learn the DNA binding preferences of each residue, and predict the DNA binding preferences of each TF. Moreover, to achieve higher accuracy I planned to learn four different sets of binding preferences, depending on the residue's position within the DNA binding domain. Unfortunately, such a detailed model could not have been accurately learned using structural information, due to the limited number of solved protein-DNA complexes. I therefore sought alternative sources of data, and eventually employed annotated sets of transcription factors and their cognate DNA sites (Wingender *et al.*, 2001; Sandelin *et al.*, 2004). To pinpoint the exact binding position of each pair, I developed a machine learning strategy and aligned the protein-DNA pairs. This allowed me to estimate context-specific amino acid-nucleotide binding preferences, to predict the DNA binding preferences of novel TFs, and to identify putative target genes of these TFs (Chapter 1).

I then focused on another aspect of gene regulation, the processing of stimuli by a combinatorial network of transcription factors. I aimed to understand how external information is being processed and integrated by the cell, and how decisions are being made and executed to reprogram the transcriptional response to external stimuli. To answer these questions, we applied genetic manipulations and measured the expression of genes in the absence of key regulators. Unfortunately, such experiments are somewhat inconclusive, since they cannot distinguish if a gene is being regulated by the transcription factor solely, or through the combinatorial interactions with additional regulators. To overcome this, we decided to dissect the transcriptional response of each gene into more specific components (including the expression level of the gene due to TF_A, due to TF_B, or due to their cooperative effect per se). In this project, I collaborated with the laboratory of Prof. Erin O'Shea (HHMI/Harvard), who measured the gene expression data in a range of additional mutant strains. My goal was to analyze these data and estimate the transcriptional components of each gene. Toward this end, I developed a statistical-based algorithm. Finally, to support these results from a mechanistic view, we wished to verify that genes supposedly regulated by a certain TF (according to our expression-based algorithm) also contain its DNA motif, and are bound by it *in vivo*. Toward this end, I developed a model-based algorithm to identify and characterize binding events using *chromatin immunoprecipitation* data we collected and measured using dense tiling arrays (Methods & Chapter 2).

My final goal was to shift from the naïve conception of DNA as a linear sequence, and focus on the packaging of DNA into *chromatin* and its relation to transcription. Specifically, I wished to accompany the recent mapping of nucleosome positioning in yeast cells by some experimental characterization of their epigenetic "state". Toward this end, I collaborated with the laboratory of Oliver Rando (Harvard/UMass), who directly measured the covalent modification pattern of thousands of nucleosomes. Given these unique data, my goal was to find the basic principles that govern the modification patterns of nucleosomes. First, I intended to identify how many different "types" of nucleosome exists. By applying computational algorithms to automatically cluster the modification data, I was hoping to find a clear partition of the nucleosomes to several groups based on their modification patterns. Second, I wished to identify combinatorial interactions between specific modifications. I then aimed to link the modification pattern of a nucleosome to its genomic location and the transcriptional state of its underlying gene. Does a promoter nucleosome look different than a nucleosome over the coding region? Are the nucleosomes over repressed genes marked differently than those over active genes? Using statistical algorithms, I showed that indeed this was the case. These results led us to additional questions. If the state of nucleosomes is tightly linked to the transcriptional activity levels of underlying genes, what happens when the cell undergoes a transcriptional reprogramming? There should be some mechanisms to allow dynamic changes in the state of chromatin. Toward this end, we designed a series of measurements to test whether nucleosomes are exchanged, both during and independent of DNA replication. I developed an analytical model to analyze these measurements and estimate the turnover rates of nucleosomes, in minutes (Methods, Chapter 4).

Methodology

Mathematical modeling of nucleosomes turnover

To analyze the data presented in Chapter 4, I developed a mathematical model which interprets the time-series measurements of the Flag- and Myc-tagged histones, based on the estimation of turnover rate parameters (in 1/minutes) for each nucleosome (Figure 11).

The model consists of several components:

- *M(t)* – the amount of Myc-H3 molecules in the free histone pool at time *t*.
- *F(t)* – the amount of Flag-H3 molecules in the free histone pool at time *t*.
- $P_l\,(t)$ – the probability that a specific nucleosome at locus *l* at time *t* contains Flag-H3.
- $R_l\,(t)$ – the predicted Flag to Myc log-ratio measured at locus *l* at time *t*.

In the model, the amount of Myc-H3 and Flag-H3 molecules is determined by the production rate and degradation rate of each type of protein. Thus,

$$\frac{d}{dt}M(t) = \alpha_M - \beta_M M(t)$$
$$\frac{d}{dt}F(t) = \alpha_F(t) - \beta_F F(t)$$

where α_M, β_M, $\alpha_F(t)$, and β_F, are the production and degradation rates of Myc-H3 and Flag-H3, respectively. We assume that the production and degradation rate of the Myc-H3 production are constant. Thus, its levels reach steady state equilibrium

$$M(t) = \frac{\alpha_M}{\beta_M}.$$

Similarly, we assume a fixed degradation rate for Flag- H3. Its production rate, however, is assumed to be zero up a particular time point t_0, where the response to galactose has been completed, from which it is produced at some fixed rate α_F.

$$\alpha_F(t) = \begin{cases} 0 & t < t_0 \\ \alpha_F & t \geq t_0 \end{cases}$$

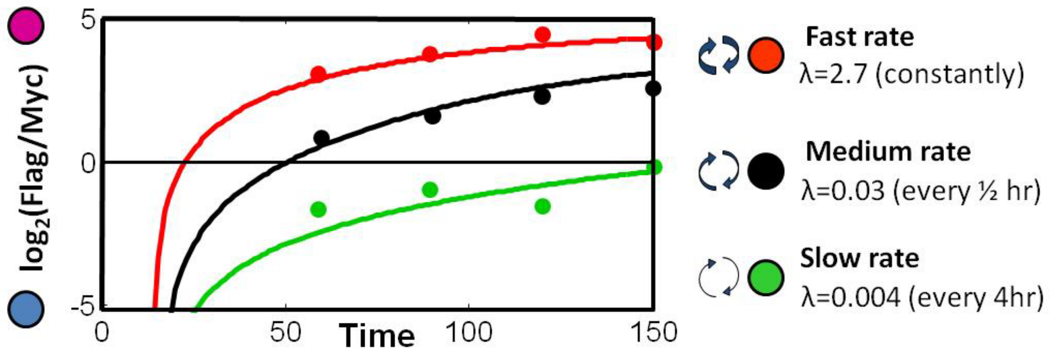


Figure 11. Mathematical model of replication-independent nucleosomes turnover

Shown are the Flag/Myc ratios (in log$_2$) for three genomic loci, over a time-series of measurements. These data were fitted by our model, resulting with the turnover rates of: a fast nucleosome (red, constantly replaced), a medium one (black, replaced every half hour on average), and a slow one (green, replaced every four hours, on average).

Solving the dynamical system under this assumption results in the following solution:

$$F(t) = \begin{cases} 0 & t < t_0 \\ \frac{\alpha_F}{\beta_F}\left(1 - e^{-\beta_F (t-t_0)}\right) & t \geq t_0 \end{cases}$$

We assume that the H3 protein recruited to the DNA at time t is Flag-H3 with probability $\frac{F(t)}{F(t)+M(t)}$, which is the relative proportion of Flag- H3 in the free histone pool at time t, according to the model.

We model turnover events at a particular locus as a Poisson processes. The rate of the process determines the frequency of turnover events. We assume that each turnover event samples a random H3 protein from the free pool. We also assume that the turnover rate does not depend on the particular H3 variant present at the location. We model the turnover rate at a specific genomic location l by the rate parameter λ_l. Thus, the distribution of durations between turnover events is $p(\Delta) = \frac{1}{\lambda_l} e^{-\lambda_l \Delta}$. These assumptions imply that the change in $P_l(t)$ (the probability that a H3 protein at locus l is Flag-H3) depends on the rate of new turnover events at that locus and on the probability of sampling Flag-H3 at that time.

$$\frac{d}{dt}P_l(t) = \lambda_l\left(\frac{F(t)}{F(t)+M(t)} - P_l(t)\right)$$

The initial condition for this equation is $P_l(t) = 0$ since there are no Flag-H3 proteins at the initialization of the experiment. For a given choice of λ_l and trajectories *F(t)* and *M(t)* we solve this equation numerically using ODE45 function in MATLAB 7.0 (rel 14).

Assuming that this probability describes how each nucleosome at location *l* behaves in all the cells, the predicted Flag to Myc log-ratio at this location is then $\log_2 \frac{P_l(t)}{1-P_l(t)}$. The microarray protocol implies that an equal amount of Flag- and Myc-tagged sequences will be hybridized. This implies that when there is a disproportional amount of one of the tags in the cell, the ratios will be normalized by the ratio of two tags in the system. Thus, the expected normalized log-ratio is

$$R_l(t) = \log_2 \frac{P_l(t)}{1-P_l(t)} - N(t)$$

where *N(t)* is a time-dependent normalizing factor.

To recap, given the parameters t_0, α_M, β_M, $\alpha_F(t)$, and β_F, we construct a dynamic model of the Flag to Myc ratio in the free histone pool at each time point. Given these we can fit a rate parameter λ_l for each location and a global normalizing factor *N(t)* for each measured time point. We jointly fit these parameters by minimizing the root mean squared error (RMSD) between the measured log-ratios and the predicted normalized log-ratios. To fit the global parameters (t_0, α_M, β_M, $\alpha_F(t)$, and β_F), we apply the fitting procedures for different values on a predetermined range and choose the one resulting in the smallest RMSD fit.

To apply this model to results from printed tiling arrays, we first transformed the measured ratios at each probe to nucleosome level measurements (as described above). To apply the model to Agilent arrays, we assume each probe represents a separate nucleosome and thus treated each probe as a locus.

Chapter 1 - Paper

Ab initio prediction of transcription factor targets using structural knowledge

Tommy Kaplan, Nir Friedman, Hanah Margalit

PLoS Computational Biology (2005) **1**(1):e1

Ab Initio Prediction of Transcription Factor Targets Using Structural Knowledge

Tommy Kaplan[1,2], **Nir Friedman**[1*], **Hanah Margalit**[2*]

1 School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel, **2** Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University, Jerusalem, Israel

Current approaches for identification and detection of transcription factor binding sites rely on an extensive set of known target genes. Here we describe a novel structure-based approach applicable to transcription factors with no prior binding data. Our approach combines sequence data and structural information to infer context-specific amino acid–nucleotide recognition preferences. These are used to predict binding sites for novel transcription factors from the same structural family. We demonstrate our approach on the Cys_2His_2 Zinc Finger protein family, and show that the learned DNA-recognition preferences are compatible with experimental results. We use these preferences to perform a genome-wide scan for direct targets of *Drosophila melanogaster* Cys_2His_2 transcription factors. By analyzing the predicted targets along with gene annotation and expression data we infer the function and activity of these proteins.

Citation: Kaplan T, Friedman N, Margalit H (2005) Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comp Biol 1(1): e1.

Introduction

Specific binding of transcription factors to *cis*-regulatory elements is a crucial component of transcriptional regulation. Previous studies have used both experimental and computational approaches to determine the relationships between transcription factors and their targets. In particular, probabilistic models were employed to characterize the binding preferences of transcription factors, and to identify their putative sites in genomic sequences [1,2]. This approach is useful when binding data are available, but cannot be applied to proteins without extensive experimental binding studies. This difficulty is particularly emphasized in view of the genome projects, where new proteins are classified as DNA-binding according to their sequence, yet there is no information about the genes they regulate.

To address the challenge of profiling the binding sites of novel proteins, we propose a family-wise approach that builds on structural information and on the known binding sites of other proteins from the same family. We use solved protein–DNA complexes [3] to determine the exact architecture of interactions between nucleotides and amino acids at the DNA-binding domain. Although sharing the same structure, different proteins from a structural family have different binding specificities because of the presence of different residues at the DNA-binding positions. To predict their binding site motif, we need to identify the residues at these positions and understand their DNA-binding preferences.

In previous studies, we used the empirical frequencies of amino acid–nucleotide interactions [4,5] in solved complexes (from various protein families) to build a set of "DNA-recognition preferences." This approach assumed similar DNA-binding preferences of the amino acids for all structural domains and at all binding positions. However, there are clear experimental indications that this assumption is not always valid: a particular amino acid may have different binding preferences depending on its positional context [6–8]. To estimate these context-specific DNA-recognition preferences, we need to determine the appropriate context of each residue, which may depend on its relative position and orientation with respect to the nucleotide. Then, we need to collect statistics about the DNA-binding preferences in this context. This can be achieved from an ensemble of solved protein–DNA complexes from the same family. Unfortunately, sufficient data of this type are currently unavailable.

To overcome this obstacle, we propose to estimate context-specific DNA-recognition preferences from available sequence data using statistical estimation procedures. The input of our method is a set of pairs of transcription factors and their target DNA sequences [2]. We then identify the residues and nucleotides that participate in protein–DNA interaction, and collect statistics about the DNA-binding preferences of residues under different contexts of the binding domain. These are then used to discover the binding site of other transcription factors from the same family, for which no targets are known.

We apply our approach to the Cys_2His_2 Zinc Finger DNA-binding family. This family is the largest known DNA-binding family in multicellular organisms [9] and has been studied extensively [10]. Members of this family bind DNA targets according to a stringent binding model [11,12], which maps the exact interactions between specific residues in the DNA-binding domain with nucleotides at the DNA site (Figure 1). We use many Zinc Finger proteins together with their native DNA targets (extracted from the TRANSFAC database [2]), and apply an iterative expectation maximization (EM)

Received January 10, 2005; Accepted February 11, 2005; Published June 24, 2005
DOI: 10.1371/journal.pcbi.0010001

Abbreviations: EM, expectation maximization; GO, Gene Ontology

Editor: Andrej Sali, University of California at San Francisco, United States of America

*To whom correspondence should be addressed. E-mail: nir@cs.huji.ac.il (NF), hanah@md.huji.ac.il (HM)

Synopsis

Cells respond to dynamic changes in their environment by invoking various cellular processes, coordinated by a complex regulatory program. A main component of this program is the regulation of transcription, which is mainly accomplished by transcription factors that bind the DNA in the vicinity of genes. To better understand transcriptional regulation, advanced computational approaches are needed for linking between transcription factors and their targets. The authors describe a novel approach by which the binding site of a given transcription factor can be characterized without previous experimental binding data. This approach involves learning a set of context-specific amino acid–nucleotide recognition preferences that, when combined with the sequence and structure of the protein, can predict its specific binding preferences. Applying this approach to the Cys_2His_2 Zinc Finger protein family demonstrated its genome-wide potential by automatically predicting the direct targets of 29 regulators in the genome of the fruit fly *Drosophila melanogaster.* At present, with the availability of many genome sequences, there are numerous proteins annotated as transcription factors based on their sequence alone. This approach offers a promising direction for revealing the targets of these factors and for understanding their roles in the cellular network.

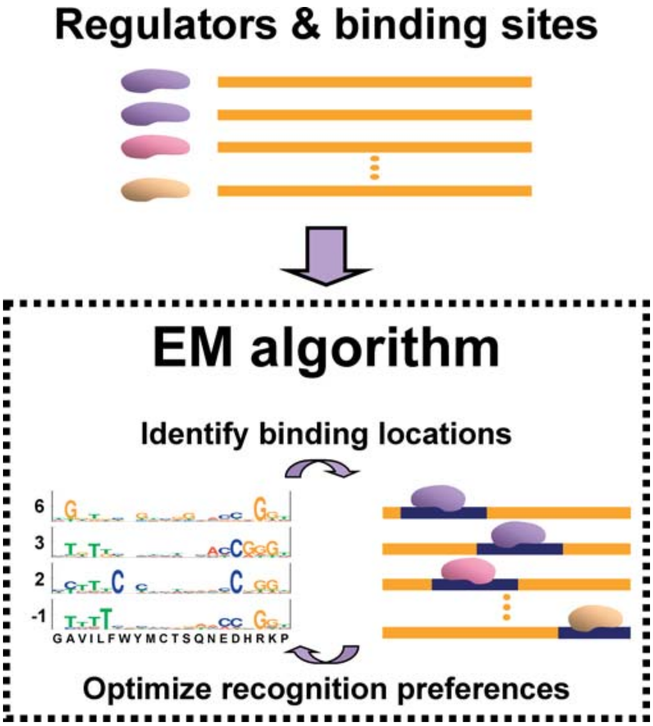

Regulators & binding sites
EM algorithm
Identify binding locations
6
3
2
-1
GAVILFWYMCTSQNEDHRKP
Optimize recognition preferences

Figure 2. Estimating DNA-Recognition Preferences

The DNA-recognition preferences are estimated from unaligned pairs of transcription factors and their DNA targets [2] (above). The EM algorithm [13] is used to simultaneously assess the exact binding positions of each protein–DNA pair (bottom right), and to estimate four sets of position-specific DNA-recognition preferences (bottom left).
DOI: 10.1371/journal.pcbi.0010001.g002

algorithm [13] to estimate position-specific DNA-recognition preferences (Figure 2). These are used in turn for predicting the DNA binding site motifs of novel proteins in the family (Figure 3), and for performing a genome-wide scan for putative targets.

Results

In Silico Reconstruction of DNA-Recognition Preferences

In order to estimate the context-specific DNA-recognition preferences of the Cys_2His_2 Zinc Finger DNA-binding family we used the canonical binding model learned from the solved protein–DNA complex of Egr-1 [11,12]. According to this model, the binding specificity of each Zinc Finger domain is determined by residues at four key positions (see Figure 1). We aimed to learn a different set of DNA-recognition preferences for each of the four key positions. These sets should express the probability of every amino acid to interact with each nucleotide. Since the number of solved protein–DNA complexes is insufficient to estimate these preferences directly, we resorted to sequence data of proteins and their DNA targets. We extracted 455 protein–DNA pairs from the TRANSFAC 7.3 database [2] (see Materials and Methods). Unfortunately, the exact binding locations of these DNA targets are not pinpointed, and thus we employed statistical tools to infer them (see Figure 2; Materials and Methods). We then used the protein–DNA binding model to identify the interacting residues and nucleotides, and collect statistics on their binding preferences (see Materials and Methods). Based on these we estimated four sets of DNA-recognition preferences (Figure 4; Tables S1 and S2), showing both context-independent preferences (such as the preference of lysine for guanine) and context-dependent ones (e.g., the preference of aspartic acid for cytosine). Table S3 shows the 10%–90% confidence intervals of the estimated probabilities.

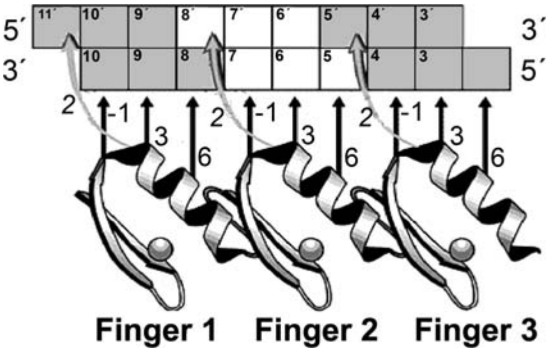


Figure 1. The Canonical Cys_2His_2 Zinc Finger DNA Binding Model

Residues at positions 6, 3, 2, and −1 (relative to the beginning of the α-helix) at each finger interact with adjacent nucleotides in the DNA molecule (interactions shown with arrows). (Figure adapted from a figure by Prof. Aaron Klug, with permission.)
DOI: 10.1371/journal.pcbi.0010001.g001

Learned Recognition Preferences Are Consistent with Experimental Results

We evaluated the four reconstructed sets of DNA-recognition preferences by comparing them with experimental data. First, we compared the derived preferences with qualitative preferences based on phage-display experiments [10] and found the two to be consistent (data not shown). Second, we predicted binding site models for Egr-1 variants for which experimental binding data were available [14], using their sequences and our estimated preferences. These models were used to score the binding of Egr-1 variants to a set of DNA targets that were tested in the experimental

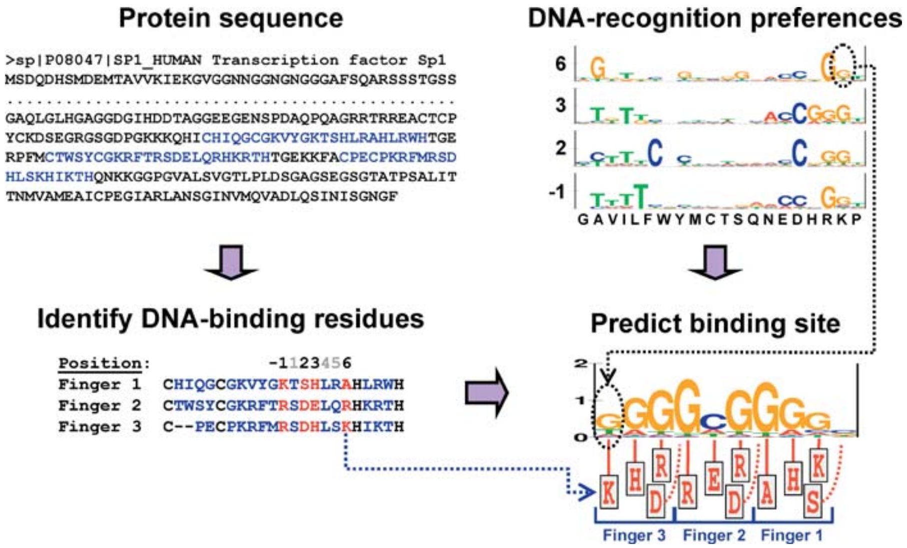

Protein sequence
>sp|P08047|SP1_HUMAN Transcription factor Sp1
MSDQDHSMDEMTAVVKIEKGVGGNNGGNGNGGGAFSQARSSSTGSS
GAQLGLHGAGGDGIHDDTAGGEEGENSPDAQPQAGRRTRREACTCP
YCKDSEGRGSGDPGKKKQHICHIQGCGKVYGKTSHLRAHLRWHTGE
RPFMCTWSYCGKRFTRSDELQRHKRTHTGEKKFACPECPKRFMRSD
HLSKHIKTHQNKKGGPGVALSVGTLPLDSGAGSEGSGTATPSALIT
TNMVAMEAICPEGIARLANSGINVMQVADLQSINISGNGF
DNA-recognition preferences
6
3
2
-1
G A V I L F W Y M C T S Q N E D H R K P
Identify DNA-binding residues
Position: -1123456
Finger 1 CHIQGCGKVYGKTSHLRAHLRWH
Finger 2 CTWSYCGKRFTRSDELQRHKRTH
Finger 3 C--PECPKRFMRSDHLSKHIKTH
Predict binding site
2
1
0
K H D R R E D R A H S K
Finger 3
Finger 2
Finger 1

Figure 3. Predicting the DNA Binding Site Motifs of Novel Transcription Factors

The protein's DNA-binding domains are identified using the Cys_2His_2 conserved pattern (top left). The residues at the key positions (6, 3, 2 and −1) of each finger (marked in red in the bottom left panel) are then assigned onto the canonical binding model (bottom right), and the sets of position-specific DNA-recognition preferences (top right panel) are used to construct a probabilistic model of the DNA binding site. For example, the lysine at the sixth position of the third finger faces the first position of the binding site (dotted blue arrow). We predict the nucleotide probabilities at this position using the appropriate recognition preferences (dotted black arrow).

DOI: 10.1371/journal.pcbi.0010001.g003

study. We found that our predictions were highly correlated with the experimentally measured binding affinities [14] (Table S4).

Next, we evaluated the ability of the estimated recognition preferences to identify binding sites within genomic sequences. We compiled a dataset of binding sites of ten Cys_2His_2 transcription factors. These involved 43 experimentally verified binding sites within natural genomic promoter sequences with a total length of 14,534 bp (Table S5). Using the recognition preferences, we predicted the binding site models of the ten transcription factors and used them to scan the respective promoter regions for putative binding sites (Figure 5A and 5B; see Materials and Methods). To prevent bias by known sites in our training data, we applied a "leave protein out" cross-validation analysis, and predicted the DNA binding model of a protein using DNA-recognition preferences that were learned from a reduced dataset, from which all its binding sites were removed. Our method marked 30 locations as putative binding sites, out of which 21 matched experimental knowledge (sensitivity of 49% and specificity of 70%, $p < 10^{-48}$; see Table S6).

Benos et al. [15] proposed a method (SAMIE) to estimate Cys_2His_2 Zinc Finger position-specific binding preferences from in vitro SELEX binding experiments. We compared the predictions of the known binding sites within promoter regions provided by our position-specific recognition preferences to those of Benos et al. [15] and of Mandel-Gutfreund et al. [5] (Figure 5C; Table S7). These results suggest that predictions based on our recognition preferences out-perform the predictions based on the other methods.

To further evaluate our predictions, we used the binding locations of Sp1 along human Chromosomes 21 and 22, as mapped by genome-wide chromatin immunoprecipitation [16]. We compiled two datasets of 1-kb-long sequences: one dataset included sequences that exhibited highly significant binding, and the other dataset included sequences that showed no binding at all (to be used as a control; see Materials and Methods). We used the DNA-recognition preferences to predict a binding site model for Sp1, and scanned the genomic sequences with it. We identified Sp1 binding sites in 45% of the experimentally bound sequences, and in only 5% of the control sequences (Figure 5D).

Ab Initio Genome-Wide Prediction of Transcription Factor Binding Sites

In the past few years many genomes were solved, yielding sequences of thousands of putative transcription factors. However, only little is currently known about the binding specificities of these factors and about their target genes. To address this problem, we applied our predictive scheme to the *Drosophila melanogaster* genome in a fully automated manner. We first scanned the sequences of 16,201 putative gene products and identified 29 canonical Cys_2His_2 Zinc Finger transcription factors with three or four fingers (see Materials and Methods). We then used their sequences and the estimated DNA-recognition preferences to compile a binding site model for each transcription factor, as in Figure 3 (see Figure S1 and Table S8 for detailed models). Finally, we used these binding site models to scan the upstream promoter regions of 15,665 *D. melanogaster* genes. Multiple putative direct targets were predicted for each Zinc Finger, as detailed at http://compbio.cs.huji.ac.il/Zinc. The number of putative direct target genes for each transcription factor and the overlap between targets of different factors are shown in Figures S2 and S3. Interestingly, several Zinc Fingers have similar residues at the DNA-binding positions, and are therefore predicted to bind similar sites and to have mutual predicted targets (see

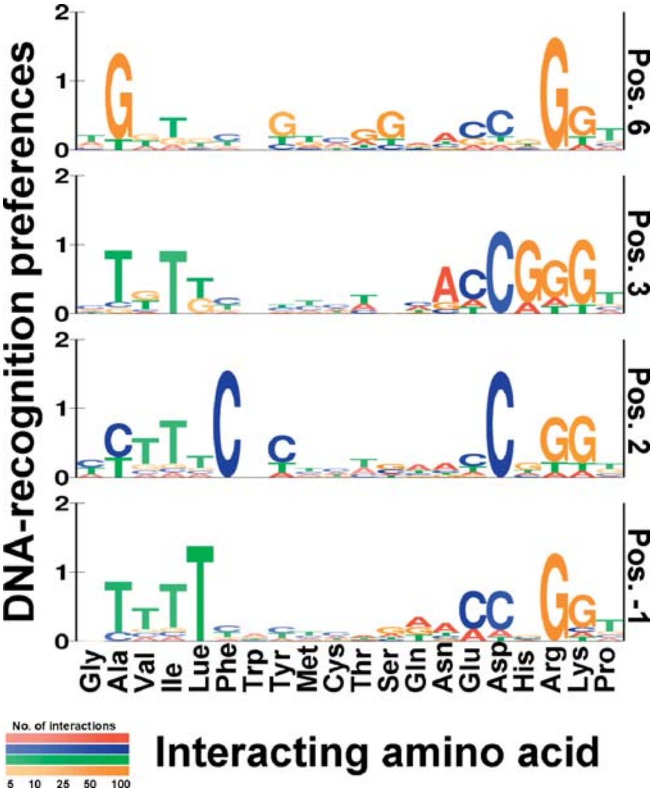


Figure 4. Four Sets of Position-Specific DNA-Recognition Preferences in Zinc Fingers

The estimated sets of DNA-recognition preferences for the DNA-binding residues at positions 6, 3, 2, and −1 of the Zinc Finger domain are displayed as sequence logos. At each position, the associated distribution of nucleotides is displayed for each amino acid. The total height of letters represents the information content (in bits) of the position, and the relative height of each letter represents its probability. Color intensity indicates the level of confidence for a given amino acid at a certain position (where paler colors indicate lower confidence due to low occurrences of the amino acid at the specific position in the training data) (see Tables S1 and S2 for full data). Some of the DNA binding preferences are general, regardless of the residue's position within the zinc finger (e.g., lysine's tendency to bind guanine), while others are position-dependent (e.g., the tendency of phenylalanine to bind cytosine only when in position 2).

DOI: 10.1371/journal.pcbi.0010001.g004

Figures S1 and S3). In *D. melanogaster,* this phenomenon has been reported for at least some transcription factors (e.g., Sp1 and Btd) [17].

To infer the function of the 29 transcription factors, we employed the functional annotations of their predicted target genes (based on the Gene Ontology [GO] terms [18]). The target sets of most transcription factors (21 out of 29) were found to be significantly enriched with at least one GO term (Figure 6A). For some of the transcription factors, the enriched GO terms match prior biological knowledge. For example, the putative targets of Glass were found to be enriched with terms related to photoreceptor cell development, consistent with previous studies that linked the Glass transcription factor with eye photoreceptor development [19]. Similarly, the putative targets of Btd and Sp1 were enriched with developmental terms, such as neurogenesis, development, and organogenesis. Indeed these regulators are known to play essential roles in mechanosensory development [17]. Furthermore, our analysis suggests possible functions for unknown proteins, as well as new annotations for some of the already known regulators (see Figure S4 for complete results).

We further evaluated the function and activity of the 29 transcription factors based on the mRNA expression profiles of their target genes (Figure 6B). We used expression data from early embryogenesis [20], as well as data from the entire life cycle of *D. melanogaster* [21]. In each experiment and for each transcription factor, we tested whether its putative targets showed similarity in their expression patterns and differed from the rest of the genes (see Materials and Methods). Such coherent expression supports the suggested relationship between the factor and the genes it is predicted to regulate. Out of the 29 transcription factors we examined, 21 showed such significant associations in at least one embryogenesis experiment, suggesting their active roles throughout early developmental stages (Figure 6B). These transcription factors include many known developmental regulators that are active during embryonic development (e.g., Btd, Sp1, Glass, Odd-skipped, and Stripe) [18,22], as well as other proteins, whose function is currently unknown. Similar results were obtained using the full life cycle gene expression data [21], mapping putative time points at which each regulator is predicted to be active (Figure 6B).

Note that the expression profiles are based on whole embryos, and therefore ignore spatially differential expression patterns. Thus, the correct function of some tissue-specific Zinc Finger proteins may be obscured in these data. Additional insight may be gained by focusing on expression data in homogeneous regions. Specifically, Butler et al. [23] compared gene expression in two homogeneous parts of the *Drosophila* imaginal wing disc—the body wall and the hinge-wing pouch. In our analysis we used the ratios between the expression levels in the two regions, and examined putative targets for enrichment in one of the regions. We then inferred the regulatory role of a transcription factor (activator or repressor) using its own expression pattern. For example, the putative targets of Stripe show higher expression levels in the body wall than the rest of the genes (enrichment p-value ≤ 0.0002). Stripe itself is enriched more than 9-fold in the body wall, relative to the wing-hinge region. This suggests that Stripe functions mainly in the body-wall region, where it activates its target genes. Indeed, this is consistent with the known role of Stripe as an activator of epidermal muscle attachment genes [24]. Using the same reasoning, we inferred the regulatory roles of four additional *D. melanogaster* transcription factors within the imaginal wing disc, three of which were previously uncharacterized (Table 1).

Discussion

In this paper we propose a general framework for predicting the DNA binding site sequence of novel transcription factors from known families. Our framework combines structural information about a specific DNA-binding domain with examples of binding sites for proteins in the family. We apply a statistical estimation algorithm to the canonical Cys_2His_2 Zinc Finger DNA-binding family, and derive a set of DNA-recognition preferences for each residue at each interacting position in the Zinc Finger DNA-binding domain.

We apply these preferences and predict the binding site models of novel proteins from the same family. Finally, we use the predicted models in genome-wide scans and identify the proteins' putative direct target genes.

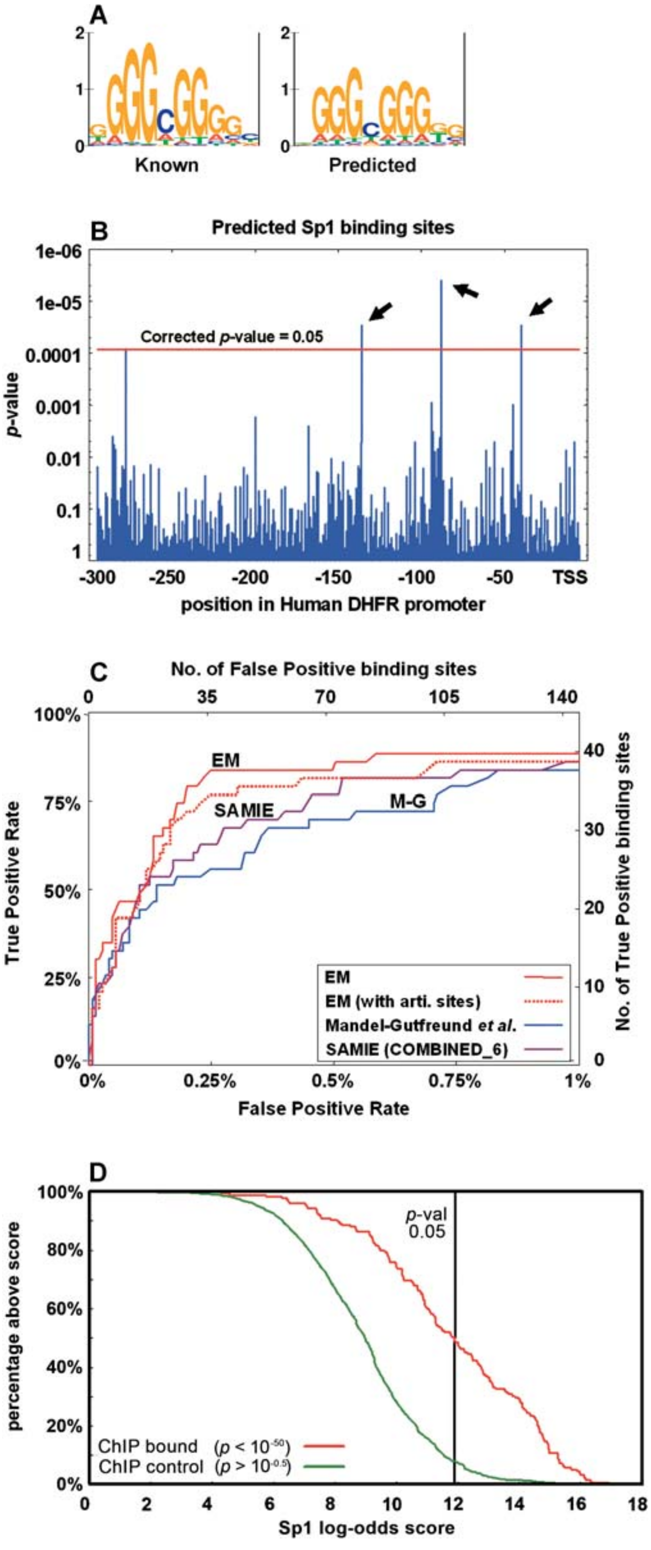


Figure 5. Validation of DNA-Recognition Preferences

(A) The predicted binding site model of human Sp1 protein is compared to its known site (matrix V$SP1_Q6 from TRANSFAC [2], based on 108 aligned binding sites). To prevent bias by known Sp1 sites in our training data, the set of DNA-recognition preferences was estimated from the TRANSFAC data after removing all Sp1 sites.

(B) Scanning the 300-bp-long promoter of human dihydrofolate reductase (DHFR) by the predicted Sp1 binding model. The *p*-value of each potential binding site is shown (*y*-axis). Four positions achieved a significant *p*-value (higher than the horizontal red line), out of which three are known Sp1 binding sites [41] (arrows).

(C) A summary of in silico binding experiments for 21 pairs of Zinc Finger transcription factors and their target promoters. Shown is the tradeoff between false positive rate (*x*-axis) and true positive rate (*y*-axis) as the significance threshold for putative binding sites is changed. For every threshold point, our set of recognition preferences (EM) achieves better accuracy than the preferences of Mandel-Gutfreund et al. [5] (M-G) and Benos et al. [15] (SAMIE). Interestingly, when the DNA-recognition preferences were estimated from training data expanded to include TRANSFAC's artificial sequences, inferior results were obtained (dotted red line).

(D) Cumulative distribution of Sp1 scores among the sequences of targets/non-targets of unbiased chromatin immunoprecipitation scans of human Chromosomes 21 and 22 [16]. The predicted Sp1 motif appears in 45% of the experimentally bound sequences but in only 5% of the control sequences.

DOI: 10.1371/journal.pcbi.0010001.g005

Structure-based approaches for prediction of transcription factor binding sites have recently gained much interest [5,8,15,25–29]. Most of the current structural approaches define a binding model based on solved protein–DNA complexes, and attempt to identify DNA subsequences that best fit the amino acids that are determined as interacting with the DNA. Previous studies [4,8] used ensembles of solved protein–DNA complexes (from all DNA-binding domains) to extract general parameters for amino acid–base recognition. Some studies used only the counts of amino acid–nucleotide pairs to derive these parameters [4], whereas others also considered the spatial arrangements [8]. However, for fine grained definition of such potentials, a much larger set of solved protein–DNA complexes is needed than is currently available. An alternative approach to estimate DNA-recognition preferences is to extract them separately for each DNA-binding domain. However, here again, the data of solved complexes are insufficient to allow such derivation.

In a recent study, Benos et al. [15] assigned position-specific DNA-recognition preferences for the Cys$_2$His$_2$ Zinc Finger family. The model they used is similar to ours, with two significant differences. First, they relied on data from in vitro selection assays, such as SELEX and phage display, to train their recognition preferences. Second, their assays screened artificial sequences, both artificial proteins and artificial DNA targets. In contrast, we rely on previously published information of natural binding sites. Our approach does not require specialized experiments, and more importantly, it captures the specificity of natural proteins to DNA sequences. As we showed, our preferences are consistent with independent experimental results [6,7,10] and are superior to such preferences derived by the other computational approaches [5,15]. In addition, previous studies showed that there are discrepancies between SELEX-derived motifs and those derived from natural binding sites [30,31]. Indeed, our method yielded inferior predictions when information on artificial binding sequences was included in our training data. Figure 4C shows that our set of recognition preferences is superior to previous models in identifying genomic binding sites. When comparing the predictions by the various recognition preferences to measured affinities of DNA artificial sequences [14], we report results similar to those of Benos et al. (see Table S4).

Analysis of the Estimated DNA-Recognition Preferences

Analysis of the estimated recognition preferences suggests that the protein–DNA recognition code is not deterministic, but rather spans a range of preferences. Moreover, our

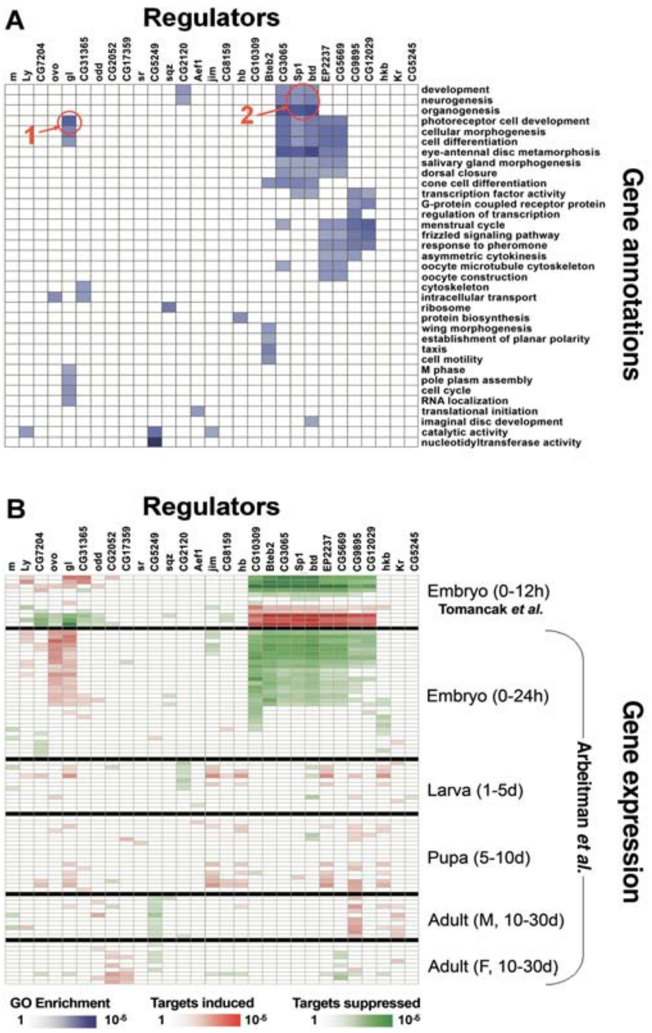


Figure 6. Inferring the Function and Activity of Zinc Finger Transcription Factors in *D. melanogaster*

(A) Similar gene annotation enrichment among the putative target sets of 29 transcription factors in *D. melanogaster*. Blue cells correspond to significant overabundance of a GO term (row) among the predicted targets of a protein (column), using a hyper-geometric test. The binding sites of most factors show enrichment in at least one GO term. For some of the regulators, the enriched GO terms match prior biological knowledge. For example, the putative targets of Glass (gl) were found to be enriched with terms related to photoreceptor cell development (red circle 1). Similarly, the putative targets of Buttonhead (btd) and Sp1 were enriched with developmental terms such as neurogenesis, development, and organogenesis (red circle 2). Closely related GO annotations are not shown; see Figure S4 for full results.

(B) Deducing the activity of the 29 transcription factors using gene expression patterns. Expression data from early (0–12 h) embryogenesis [20] and data from the entire *Drosophila* life cycle [21] are used to test whether the putative direct targets of a regulator are expressed differently than the rest of the genes in a given experiment. Red cells correspond to significant enrichment of overexpressed targets using a Kolmogorov-Smirnov test, while green cells correspond to enrichment of underexpressed targets. For most of the regulators the analysis resulted in at least one significant embryogenesis experiment, suggesting an active role in early developmental stages (above). Similar results were obtained using the full life cycle gene expression data (below).

DOI: 10.1371/journal.pcbi.0010001.g006

analyses show that a residue may have different nucleotide preferences depending on its context. For some amino acids, the qualitative preferences remain the same across various positions, while the quantitative preferences vary (e.g., arginine; see Figure 4). The DNA-binding preferences of other residues change across various positions. For example, histidine at position 3 tends to interact with guanine, while it shows no preference to any nucleotide at all other positions. Another example is the tendency of alanine at position 6 to face guanine. This preference, which was revealed automatically by our analysis, is not consistent with the chemical nature of alanine's side chain nor with general examinations of amino acid–nucleotide interactions [5,8]. We suspect that it is affected by the large number of Sp1 targets in our dataset. This potential interaction was implied before in Sp1 binding sites [32] and may reflect an interaction between the residue at position 2 with the complementary cytosine.

The Protein–DNA Binding Model

In this work, we use a binding model that is based on solved protein–DNA complexes. The model presents a rigid and simplistic representation of the amino acid–base interactions at the Zinc Finger domains. Only some of the Zinc Finger domains (termed "canonical" in this work) use this model for binding, while others maintain more complex interactions. As our results show, by using this model, we manage to recover most of the DNA-binding specificities of amino acids, and use them to predict the binding site models of novel proteins. We believe that this model offers a fair tradeoff between complexity (and number of parameters) and accuracy.

Inter-Position Dependencies in the Binding Site

The Cys_2His_2 binding model inherently assumes that all positions within the binding site are independent of each other. This assumption is used in most computational approaches that model binding sites. Two recent papers [33,34] discuss this issue in the context of the Cys_2His_2 Zinc Finger domain. Their analyses of binding affinity measurements suggest that weak dependencies do exist among some positions of the binding sites of Egr-1. Nonetheless, a reasonable approximation of the binding specificities is obtained even when ignoring these dependencies. In another recent study [35], we evaluated probabilistic models that are capable of capturing inter-position dependencies within binding sites. Our results show that dependencies can be found in the binding sites of many proteins from various DNA-binding domains (especially from the helix-turn-helix and the homeo domains). However, our results also suggest that such models of dependencies do not lead to significant improvements in modeling the binding sites of Zinc Finger proteins. Thus, we believe that the Cys_2His_2 binding model we use here is indeed a reasonable approximation of the actual binding.

Genome-Wide Predictions of Binding Sites and Target Genes

In the current era there is a growing gap between the number of known protein sequences and the number of experimentally verified binding sites. To better understand regulatory mechanisms in newly solved genomes, it is crucial to identify the direct target genes of novel DNA-binding proteins. Our method opens the way for such genome-wide assays. Here we apply it to the Cys_2His_2 Zinc Finger DNA-binding family. By predicting the binding site models of regulatory proteins, one can classify genes into those that contain significant binding sites at their regulatory promoter

Table 1. Analysis of Differential Expression in *D. melanogaster* Imaginal Wing Disc

Transcription Factor			**Targets**		**Transcription Factor Role**	
Name	**Body/Wing ratio**[a]	**Body/Wing Enriched**	**Kolmogorov-Smirnov *p*-Value**[b]	**Body/Wing Enriched**	**Inferred Function**	**Known Function**
Stripe	9.113	Body	1.32×10^{-4}	Body	Activator	Activator
EP2237	2.059	Body	5.48×10^{-4}	Body	Activator	Activator
CG10309	0.321	Wing	2.00×10^{-4}	Body	Repressor	—
CG9895	0.380	Wing	8.45×10^{-3}	Body	Repressor	—
CG14655	0.302	Wing	3.47×10^{-2}	Wing	Activator	—

Butler et al. [23] measured the gene expression levels at two parts of the imaginal wing disc—the body wall and hinge-wing pouch, and computed the ratios between the two. The regulatory functions of transcription factors are analyzed by comparing their ratio with the ratios of their targets. Activators are expected to have the same directional enrichment as their targets, while repressors are expected to have opposite effects. Each group of targets is assigned a *p*-value using a two-tailed Kolmogorov-Smirnov test that compares the ratios in the target group to those of the rest of the genes.

[a] Ratio between the transcription factor's mRNA expression levels at the body wall and the wing-hinge pouch.

[b] *p*-Value of targets' enrichment using a Kolmogorov-Smirnov test.

DOI: 10.1371/journal.pcbi.0010001.t001

regions (hence, putative target genes) and those that do not. As we showed, our approach can scale up to such genome-wide scans and successfully predict the target genes of many novel Zinc Finger proteins in higher eukaryotes. Furthermore, by integrating data from external sources, such as gene expression and GO annotations, it is possible to infer the cellular function and activity of these novel proteins.

Applications to Other DNA-Binding Domains

Theoretically, our approach can be extended to handle other structural families, such as the basic leucine zipper, the homeodomain, and the basic helix-loop-helix, for which enough binding data already exist (1,191, 505, and 201 binding sites per family, respectively). This extension requires that the various proteins in the family show a common DNA binding model, which can be used further for other family members. For such families, our approach should suffice. For other families, where the binding models are more complex and flexible (including other Zinc Finger domains, such as CCCC, CCHC, or even the non-canonical Cys_2His_2), more advanced models and learning techniques will be needed. In spite of these possible difficulties, we believe that structural approaches, such as the one we show here, open promising directions, leading to successful predictions of binding site models and, following that, to accurate identification of the target genes of novel proteins, even on genome-wide scales. Eventually, such approaches will be utilized to reconstruct larger and larger portions of the transcriptional regulatory networks that control the living cell.

Materials and Methods

Sequences of Zinc Finger proteins and their binding sites. We trained a profile hidden Markov model [36] on 31 experimentally determined canonical domains [37], and used it to classify the remaining Cys_2His_2 Zinc Finger domains in TRANSFAC [2] as canonical or non-canonical. From these, we selected proteins with two to four properly spaced canonical fingers. This resulted in 61 canonical Cys_2His_2 Zinc Finger proteins, and 455 protein–binding site pairs. We used these pairs as our training data in subsequent steps. The total number of fingers in this dataset was 1,320, and the total length of all binding sites was 9,761 bp (average length of 21 bp per site).

Identification of DNA-binding residues. The interacting residues in each finger are located at positions 6, 3, 2, and −1 relative to the beginning of the α-helix (see Figure 1). We identify these positions using their relative positioning in the Cys_2His_2 conserved pattern: CX(2–4)CX(11–13)HX(3–5)H. Although, theoretically there can be 20^4 different combinations of amino acids at the interacting positions, we found only 80 different combinations among the 1,320 fingers in our database. Figures S5 and S6 show the abundance of amino acids at the different DNA-binding positions.

The probabilistic model. We describe the binding preferences of a protein using a probabilistic model. For a canonical Egr-1-like Zinc Finger protein, we denote by $A = \{A_{i,p} : i = \{1,\ldots, k\}, p \in \{-1,2,3,6\}\}$ the set of interacting residues in the different four positions of the k fingers (ordered from the N- to the C-terminus). Let $N_1,\ldots, N_L$ be a target DNA sequence. The conditional probability of an interaction with a DNA subsequence, starting from the jth position in the DNA is

$$P(N_{j,\ldots}, N_{j+3k(i-1)}|A) = \prod_{i=1}^{k} P_6(N_{j+3(i-1)}|A_{k+1-i,6}) P_3(N_{j+3(i-1)+1}|A_{k+1-i,3}) P_{-1}(N_{j+3(i-1)+2}|A_{k+1-i,-1}) \qquad (1)$$

where $P_p(N|A)$ is the conditional probability of nucleotide N given amino acid A at position p. These probabilities are the parameters of the model. For each of the four interacting positions there is a matrix of the conditional probabilities of the four nucleotides given all 20 residues. We call these matrices the DNA-recognition preferences.

The model, as described above, does not account for the interactions by the amino acid in position 2 in each finger. According to the canonical binding model (see Figure 1), the amino acid at position 2 interacts with the nucleotide that is complementary to the nucleotide interacting with position 6 of the previous finger. Thus, when we have a base pair interacting with two amino acids, we replace the term $P_6(N_{j+3(i-1)}|A_{k+1-i,6})$ with the term

$$\alpha P_6(N_{j+3(i-1)}|A_{k+1-i,6}) + (1-\alpha)P_2(N_{j+3(i-1)}|A_{k+2-i,2}) \qquad (2)$$

for $i > 1$, where α is a weighting coefficient that depends on the number of examples we have seen while estimating the recognition preferences at each position. Moreover, we add the term $P_2(N_{j+3(i-1)}|A_{k+2-i,2})$, for $i = k + 1$, to capture the last nucleotide, which is in interaction with position 2 of the first finger.

Estimating DNA-recognition preferences. We searched for the DNA-recognition preferences that maximized the likelihood of the DNA sites given the binding proteins. The DNA sequences in our database were reported as containing the binding sites [2], yet the exact binding locations were not pinpointed. Thus, we simultaneously identified the exact binding locations and maximum likelihood recognition preferences using the iterative EM algorithm [13]. Starting with an initial choice of DNA-recognition preferences (possible choices are discussed below), the algorithm proceeds iteratively, by carrying out two steps. In the E-step, the expected posterior probability of binding locations is computed for every protein–DNA pair. This is done using the current sets of preferences. In the M-step, the DNA-recognition preferences are updated to maximize the likelihood of the current binding positions for all protein–DNA pairs based on the distribution of possible binding locations computed in the E-step.

Each iteration of these two steps increases the likelihood of the data until reaching a convergence point [13]. Although the EM algorithm is proved to converge, it does not ensure that the final DNA-recognition preferences are the optimal ones, because of suboptimal local maxima of the likelihood function. This can be overcome by using promising starting points or applying the EM procedure with multiple random starting points (see Figure S7). An additional potential pitfall is over-fitting the recognition preferences of rare residues. To address this problem and ensure that the estimated recognition preferences for rare amino acids are close to uniform distribution (i.e., uninformative), we use a standard method of "pseudo-counts." We do so by adding a constant (0.7 in the results above) to each amino acid–nucleotide count computed at the end of the E-step. This is equivalent to using a Dirichlet prior on the parameters, and then performing a maximum a posteriori estimation rather than maximum likelihood estimation.

We evaluated the robustness and convergence rate of the EM procedure using a 10-fold cross-validation procedure. In each round, we removed a part of the data, trained on the remaining pairs, and tested the likelihood of the held-out protein–DNA pairs. We used this procedure to test various initialization options. Our evaluation shows that the EM algorithm performs best when initialized with the general recognition preferences of Mandel-Gutfreund et al. [5], converging within a few iterations. Similar results were obtained using random initialization points, although the convergence rate was somewhat slower (see Figure S7). Also, in Figure S8 we demonstrate the correlation between the size of the training dataset and the likelihood of test data.

Predicting the binding sites of novel proteins. Given the sequence of a novel Cys_2His_2 Zinc Finger protein, we identified the four key residues at each DNA-binding domain, and utilized the appropriate set of DNA-recognition preferences to construct a probabilistic model of the binding site (see Figure 3).

In silico binding experiments. We used the predicted binding site models to scan genomic sequences for putative binding sites. We scored each possible binding position using the log of the ratio between the probability assigned to it by the model and the background probability (log-odds score). We then estimated the p-value of these scores and applied a Bonferroni correction to account for multiple tests within the same promoter region [38]. Sites with a significant p-value ($\leq$0.05 after Bonferroni correction) were marked as putative binding sites (see Figure 4B).

Comparison with other computational approaches. In a similar manner, we generated probabilistic binding site models for these transcription factors using the recognition preferences of Mandel-Gutfreund et al. [5] and SAMIE [15]. We then scanned the corresponding promoter regions using these models.

Ab initio genome-wide prediction of binding sites. We downloaded genomic sequences of the *D. melanogaster* from FlyBase [22], release 3–1. These include 2-kb regulatory regions upstream from 15,664 genes, and the sequences of 16,201 putative gene products. We scanned the proteins for canonical Zinc Finger domains using the Cys_2His_2 conserved pattern and our profile-HMM model (available at http://compbio.cs.huji.ac.il/Zinc). We found 29 proteins with properly spaced three or four fingers (with distances of 28–31 residues between the beginnings of Zinc Finger domains). We then used the learned sets of DNA-recognition preferences to predict probabilistic binding site models for these putative Zinc Finger transcription factors. Finally, we performed in silico binding experiments by scanning each gene's 2-kb upstream region for two significant binding sites ($p \leq 0.05$ after Bonferroni correction). The matched genes were marked as putative direct targets of the transcription factor.

Enrichment of GO annotations among the target genes. FlyBase GO annotations [18,22] were downloaded from the Gene Ontology Consortium (http://www.geneontology.org) in October 2003. The enrichment p-values were calculated by GeneXPress (http://genexpress.stanford.edu), using a hyper-geometric test that compares the abundance of similarly annotated genes among the putative targets to the rest of the genome. We then applied an FDR correction for multiple hypotheses using a false rate of 0.05 [39], and only significant factors/terms are shown.

Inference of activity/function using gene expression data. We downloaded genome-wide gene expression data from early embryogenesis stages [20] (available from FlyBase; http://www.fruitfly.org/cgi-bin/ex/insitu.pl). The expression level of each gene in each array was transformed to log (base 2) of the ratio of expression to the geometric average of the expression of the gene in all arrays. In addition, we downloaded expression data from along the *Drosophila* life cycle [21] (available from Stanford Microarray Database; http://genome-www5.stanford.edu). These expression data are represented as log (base 2) of expression compared to a reference sample representing all stages of the life cycle.

For each protein and in each experiment, we used a Kolmogorov-Smirnov test to evaluate whether the expression pattern of the putative direct target genes was different from the expression of the rest of the genome. We then corrected the results for multiple hypotheses using an FDR correction [39] (false rate of 0.05). Similarly, we used differential gene expression data from *D. melanogaster* imaginal wing disc [23]. For each gene, we computed the ratio of its expression in the body wall to its expression in the hinge-wing pouch, and performed a two-tailed version of the Kolmogorov-Smirnov test to compare these ratios among the putative targets and the rest of the genome.

Supporting Information

Figure S1. Sequence logos of 29 *Drosophila* Transcription Factors

Found at DOI: 10.1371/journal.pcbi.0010001.sg001 (617 KB PDF).

Figure S2. Number of Predicted Direct Targets

Found at DOI: 10.1371/journal.pcbi.0010001.sg002 (162 KB PDF).

Figure S3. Percentage of Pairwise Coverage between Targets

Found at DOI: 10.1371/journal.pcbi.0010001.sg003 (109 KB PDF).

Figure S4. Results of Complete GO Table

Found at DOI: 10.1371/journal.pcbi.0010001.sg004 (182 KB PDF).

Figure S5. Abundance of DNA-Binding Residues in Training Data

Found at DOI: 10.1371/journal.pcbi.0010001.sg005 (123 KB PDF).

Figure S6. Abundance of Combinations of DNA-Binding Residues in Training Data

Found at DOI: 10.1371/journal.pcbi.0010001.sg006 (123 KB PDF).

Figure S7. Convergence of the EM Algorithm on Held-Out Test Data

Found at DOI: 10.1371/journal.pcbi.0010001.sg007 (106 KB PDF).

Figure S8. Likelihood of Held-Out Test Data Given Different Sizes of the Training Datasets

Found at DOI: 10.1371/journal.pcbi.0010001.sg008 (106 KB PDF).

Table S1. Four Sets of DNA-Recognition Preferences: Probabilities

Found at DOI: 10.1371/journal.pcbi.0010001.st001 (22 KB PDF).

Table S2. Four Sets of Recognition Preferences: Counts

Found at DOI: 10.1371/journal.pcbi.0010001.st002 (20 KB PDF).

Table S3. Confidence Intervals on Four Sets of DNA-Recognition Preferences

Found at DOI: 10.1371/journal.pcbi.0010001.st003 (63 KB PDF).

Table S4. Correlation with Experimentally Measured Binding Affinities

Found at DOI: 10.1371/journal.pcbi.0010001.st004 (514 KB TIF).

Table S5. 21 Protein–DNA Pairs

Found at DOI: 10.1371/journal.pcbi.0010001.st005 (2 MB TIF).

Table S6. Sensitivity and Specificity of Test Set at Different Significance Threshold Values

Found at DOI: 10.1371/journal.pcbi.0010001.st006 (328 KB TIF).

Table S7. Sensitivity and Specificity of Test Set at Different Significance Threshold Values—Other Computational Methods

Found at DOI: 10.1371/journal.pcbi.0010001.st007 (440 KB TIF).

Table S8. Position-Specific Score Matrices of 29 Cys_2His_2 Transcription Factors from *Drosophila melanogaster*

Found at DOI: 10.1371/journal.pcbi.0010001.st008 (55 KB PDF).

Acknowledgments

The authors wish to thank Yael Altuvia, Yoseph Barash, Ernest Fraenkel, Benjamin Gordon, Robert Goldstein, Ruth Hershberg, Dalit May, Lena Nekludova, Aviv Regev, and Eran Segal for helpful discussions. TK is supported by the Yeshaya Horowitz Association through the Center for Complexity Science. NF is supported by the Harry and Abe Sherman Senior Lectureship in Computer Science. This work was supported by grants from the Israeli Ministry of Science and the Israeli Science Foundation. A preliminary version of this manuscript appeared in RECOMB 2005 [40].

Competing interests. The authors have declared that no competing interests exist.

Author contributions. TK, NF, and HM conceived and designed the experiments. TK performed the experiments. TK, NF, and HM analyzed the data and wrote the paper. ■

References

1. Stormo GD (2000) DNA binding sites: Representation and discovery. Bioinformatics 16: 16–23.
2. Wingender E, Chen X, Fricke E, Geffers R, Hehl R, et al. (2001) The TRANSFAC system on gene expression regulation. Nucleic Acids Res 29: 281–283.
3. Luscombe NM, Laskowski RA, Thornton JM (2001) Amino acid–base interactions: A three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res 29: 2860–2874.
4. Mandel-Gutfreund Y, Margalit H (1998) Quantitative parameters for amino acid–base interaction: Implications for prediction of protein–DNA binding sites. Nucleic Acids Res 26: 2306–2312.
5. Mandel-Gutfreund Y, Baron A, Margalit H (2001) A structure-based approach for prediction of protein binding sites in gene upstream regions. Pac Symp Biocomput 2001: 139–150.
6. Choo Y, Klug A (1994) Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage. Proc Natl Acad Sci U S A 91: 11163–11167.
7. Choo Y, Klug A (1994) Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. Proc Natl Acad Sci U S A 91: 11168–11172.
8. Kono H, Sarai A (1999) Structure-based prediction of DNA target sites by regulatory proteins. Proteins 35: 114–131.
9. Tupler R, Perini G, Green MR (2001) Expressing the human genome. Nature 409: 832–833.
10. Wolfe SA, Greisman HA, Ramm EI, Pabo CO (1999) Analysis of zinc fingers optimized via phage display: Evaluating the utility of a recognition code. J Mol Biol 285: 1917–1934.
11. Pavletich NP, Pabo CO (1991) Zinc finger-DNA recognition: Crystal structure of a Zif268–DNA complex at 2.1 A. Science 252: 809–817.
12. Elrod-Erickson M, Benson TE, Pabo CO (1998) High-resolution structures of variant Zif268–DNA complexes: Implications for understanding zinc finger–DNA recognition. Structure 6: 451–464.
13. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39: 1–38.
14. Bulyk ML, Huang X, Choo Y, Church GM (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proc Natl Acad Sci U S A 98: 7158–7163.
15. Benos PV, Lapedes AS, Stormo GD (2002) Probabilistic code for DNA recognition by proteins of the EGR family. J Mol Biol 323: 701–727.
16. Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, et al. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499–509.
17. Schock F, Purnell BA, Wimmer EA, Jackle H (1999) Common and diverged functions of the *Drosophila* gene pair D-Sp1 and buttonhead. Mech Dev 89: 125–132.
18. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–D261.
19. Moses K, Ellis MC, Rubin GM (1989) The glass gene encodes a zinc-finger protein required by *Drosophila* photoreceptor cells. Nature 340: 531–536.
20. Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, et al. (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 3: RESEARCH0088.
21. Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, et al. (2002) Gene expression during the life cycle of *Drosophila melanogaster*. Science 297: 2270–2275.
22. FlyBase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 31: 172–175.
23. Butler MJ, Jacobsen TL, Cain DM, Jarman MG, Hubank M, et al. (2003) Discovery of genes with highly restricted expression patterns in the *Drosophila* wing disc using DNA oligonucleotide microarrays. Development 130: 659–670.
24. Vorbruggen G, Jackle H (1997) Epidermal muscle attachment site-specific target gene expression and interference with myotube guidance in response to ectopic stripe expression in the developing *Drosophila* epidermis. Proc Natl Acad Sci U S A 94: 8606–8611.
25. Suzuki M, Gerstein M, Yagi N (1994) Stereochemical basis of DNA recognition by Zn fingers. Nucleic Acids Res 22: 3397–3405.
26. Steffen NR, Murphy SD, Tolleri L, Hatfield GW, Lathrop RH (2002) DNA sequence and structure: Direct and indirect recognition in protein–DNA binding. Bioinformatics 18: S22–S30.
27. Endres RG, Schulthess TC, Wingree NS (2004) Toward an atomistic model for predicting transcription-factor binding sites. Proteins 57: 262–268.
28. Havranek JJ, Duarte CM, Baker D (2004) A simple physical model for the prediction and design of protein-DNA interactions. J Mol Biol 344: 59–70.
29. Paillard G, Deremble C, Lavery R (2004) Looking into DNA recognition: Zinc finger binding specificity. Nucleic Acids Res 32: 6673–6682.
30. Robison K, McGuire AM, Church GM (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J Mol Biol 284: 241–254.
31. Shultzaberger RK, Schneider TD (1999) Using sequence logos and information analysis of Lrp DNA binding sites to investigate discrepancies between natural selection and SELEX. Nucleic Acids Res 27: 882–887.
32. Berg JM (1992) Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites. Proc Natl Acad Sci U S A 89: 11109–11110.
33. Benos PV, Bulyk ML, Stormo GD (2002) Additivity in protein–DNA interactions: How good an approximation is it? Nucleic Acids Res 30: 4442–4451.
34. Bulyk ML, Johnson PLF, Church GM (2002) Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Res 30: 1255–1261.
35. Barash Y, Elidan G, Friedman N, Kaplan T (2003) Modeling dependencies in protein–DNA binding sites. In: Vingron M, Istrail S, Pevzner P, Waterman M, editors. Proceedings of the Seventh International Conference on Research in Computational Molecular Biology. New York: ACM Press. pp. 28–37.
36. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755–763.
37. Wolfe SA, Nekludova L, Pabo CO (2000) DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29: 183–212.
38. Barash Y, Elidan G, Kaplan T, Friedman N (2005) CIS: compound importance sampling method for protein-DNA binding site p-value estimation. Bioinformatics. 21: 596–600.
39. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B 57: 289–300.
40. Kaplan T, Friedman N, Margalit H (2005) Predicting transcription factor binding sites using structural knowledge. In: Miyano S, Mesirov JP, Kasif S, Istrail S, Pevzner PA, et al., editors. Proceedings of the Ninth International Conference on Research in Computational Molecular Biology: Lecture notes in computer science, Volume 3,500. Berlin: Springer-Verlag. pp. 522–537.
41. Kriwacki RW, Schultz SC, Steitz TA, Caradonna JP (1992) Sequence-specific recognition of DNA by zinc-finger peptides derived from the transcription factor Sp1. Proc Natl Acad Sci U S A 89: 9759–9763.

Chapter 2 – Additional results

Structure and function of a transcriptional network activated by the MAPK Hog1

This chapter summarizes my results from a joint project with the laboratories of Erin O'Shea (HHMI/Harvard) and Aviv Regev (Broad/MIT). In this project, Andrew Capaldi (O'Shea laboratory) was in charge of the experimental side and conducted most of the experiments. I conducted the analytical side leading to the results described here. The project involved frequent interactive discussions between Andrew, myself, and the PIs, leading to a tight cycle between the experiments and their analysis. Additional collaborators include Naomi Habib (Hebrew University) and Ying Liu (Harvard).

ABSTRACT

Cells regulate gene expression using a complex network of transcription factors and promoters. To gain insight into the structure and function of these networks, I developed a computational algorithm to analyze gene expression in both single and multiple mutant strains. This allowed me to build a quantitative model of the Hog1 MAPK-dependent osmotic stress response in budding yeast. My model reveals that Hog1 and the general stress transcription factors (Msn2/4) interact, at both the signaling and promoter level, to process external stimuli, and create a context-dependent response. To support these findings, I developed a model-based algorithm and analyzed high-resolution *in vivo* chromatin immunoprecipitation data. This study lays out a path to identifying and characterizing the role of combinatorial processing in transcriptional regulation.

INTRODUCTION

A full understanding of gene regulation will require the construction of detailed circuit diagrams that describe how signals influence transcription factor (TF) activity and how these TFs cooperate to regulate mRNA levels (Davidson, 2006). However, current experimental approaches used to examine these networks, such as chromatin immunoprecipitation (ChIP) and microarray analysis of strains with a single network component deleted (Harbison *et al.*, 2004; Boyer *et al.*, 2005; Hu *et al.*, 2007), provide only a limited view of their structure and function. For example, when a single mutant analysis is used, an interaction between two network components is inferred if they regulate overlapping gene-sets (e.g. HΔ and MΔ, Figure 1a). However, it is not possible to tell from single-mutant data if two factors act fully cooperatively, independently, or partially cooperatively to regulate gene expression (Potential Mechanisms, Figure 1a). Moreover, the nature of the interaction could vary from one target gene to another. As a result, network models derived from such data are incomplete and likely inaccurate.

To overcome this problem, and distinguish between possible regulatory mechanisms, double mutant (or epistasis) analysis can be applied (Avery and Wasserman, 1992). Here, if two network components H and M act cooperatively to regulate a gene, then the single mutants (HΔ and MΔ) and double mutants (HΔMΔ) will have identical expression defects (Cooperative Mechanism, Figure 1b). By contrast, if H and M act independently, then the expression defect in the double mutant will be the sum of the defects found in the single mutants (Independent Mechanism, Figure 1b). In mechanisms with partial cooperativity, the observed behavior will lie between that found for cooperative and independent mechanisms (Partially Cooperative Mechanism, Figure 1b). This approach has been used previously in conjunction with microarrays to examine regulatory mechanisms and pathway interactions at a coarse-grained or qualitative level (Lee *et al.*, 2000; Roberts *et al.*, 2000; O'Rourke and Herskowitz, 2004; Van Driessche *et al.*, 2005; Hu *et al.*, 2007).

Here I show that double mutant analysis can be used to build a detailed and quantitative model of transcriptional regulation, including the strength and type of each edge in the network and the logic gate at each node (in a given condition). To achieve this goal, I developed a microarray-based strategy that allows us to overcome the significant noise in microarray measurements and accurately quantify the influence and interaction of network factors at individual genes. To do this, my collaborators measured the expression levels of genes in a range of mutant strains, and I calculated the value of what we termed the *expression components* for each gene. In the example of the interacting factors H and M, there are three such expression components (Shown in Figure 1b, "Expression Components" column): the transcriptional activation by the transcription factor H alone (H component), the activation from M alone (M component), and the activation that results from the interaction between H and M (Co component). To determine these values, my algorithm considers the expression in the wild-type, single, and double mutant strains (Figure 2a, arrays C-F). The expression component values for each gene are then regressed using the equations shown in Figure 2a, describing which expression components are measured by each microarray (Figure 2a, equations). Finally, I developed a statistical significance score, which estimates the p-value of each expression component per gene.

To evaluate this strategy, I applied it a well-studied prototypical example of transcriptional regulatory network. We focused on the HOG signaling network, which controls the response of budding yeast to hyper-osmotic stress. In brief, following external signaling, the MAP kinase Hog1 is imported into the nucleus, where it phosphorylates (and activates) several downstream transcription factors.

In osmotic stress, the mitogen activated protein kinase (MAPK) Hog1 and the two paralogous general-stress TFs Msn2 and Msn4 are transported into the nucleus (Roberts *et al.*, 2000) where, together, they activate a transcriptional program involving hundreds of genes (Fig. 1a, Venn diagram, Rep *et al.*, 2000; O'Rourke and Herskowitz, 2004). Studies of strains lacking Hog1 or Msn2/4 have led to a model in which Msn2 and Msn4 function downstream of Hog1 in the osmotic stress response (Rep *et al.*, 2000). However, it is unclear if Hog1 and Msn2/4 act independently, cooperatively, or partially cooperatively to control the expression level of the HOG pathway, and what type of interaction between Hog1 and Msn2/4 occurs in the various target genes.

MATERIAL AND METHODS

Regression of Expression Data using the Mutant Cycle Approach

For each gene, I analyzed the expression data from several mutant strains, and dissected its expression levels to basic components contributed by its regulating transcriptional factors (Hog1, Msn2/4 and additional TFs). As described in Figure 2a, to dissect the interaction between Hog1 and Msn2/4, I compared the gene expression of four strains: wt, *hog1*Δ, *msn2/4*Δ, and *hog1*Δ*msn2/4*Δ using DNA microarrays (measured in triplicates by my collaborators):

- B = wt vs *hog1Δmsn2/4Δ*
- C = wt vs *hog1Δ*
- D = wt vs *msn2/4Δ*
- E = *msn2/4Δ* vs *hog1ΔMsn2/4Δ*
- F = *hog1Δ* vs *hog1Δmsn2/4Δ*

For each gene, I described these measurements (in log scale) as the (noisy) sum of three underlying components: H (the influence of Hog1 alone on expression), M (the influence of Msn2/4 alone on expression), and Co (the effect of the interaction between Hog1 and Msn2/4). This allowed me to rewrite the equations above as:

- B = H+M+Co (as all three components are present in the wt strain and absent from the double-delete strain hog1Δmsn2/4Δ)
- C = H+Co (again, the wt stains contains all three components, whereas hog1Δ lacks H and Co and contains only the M component)
- D = M+Co (symmetric to C; only the H component exists in msn2/4Δ)
- E = H (msn2/4Δ contains the H component alone, while the double-delete strains lacks all three)
- F = M (hog1Δ contains only the M component, while the double-delete strains lacks all three)

This system of equations can be formulated as the following matrix multiplication:

$$\begin{bmatrix} \text{wt vs } \textit{hog1}\,\Delta\textit{msn2/4}\,\Delta \\ \text{wt vs } \textit{hog1}\,\Delta \\ \text{wt vs } \textit{msn2/4}\,\Delta \\ \textit{msn2/4}\,\Delta \text{ vs } \textit{hog1}\,\Delta\textit{msn2/4}\,\Delta \\ \textit{hog1}\,\Delta \text{ vs } \textit{hog1}\,\Delta\textit{msn2/4}\,\Delta \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \begin{bmatrix} \text{H} \\ \text{M} \\ \text{Co} \end{bmatrix}$$

or written as $\boldsymbol{Y = X * \beta + \varepsilon}$, where Y are the measured values, X is the design matrix (specifying which components are present or absent for each array) , β is the actual contribution of the three components to the expression of the gene, and ε is the noise. For every gene, my goal was to find the three values in β which minimizes the errors ε.

To solve this linear model, we applied a multiple linear regression algorithm which minimizes the least squares fit of $\boldsymbol{X*\beta}$, assuming a zero-mean Normal distribution of the errors $\boldsymbol{\varepsilon}$. To solve this equation and regress the values of $\boldsymbol{\beta}$, the equation above $\boldsymbol{X * \beta = Y}$ is multiplied (from the left) by X^T, to get: $\boldsymbol{X^T * X * \beta = X^T * Y}$. In this case, the matrix $\boldsymbol{X^T *X}$ is non-singular, and so we invert $\boldsymbol{X^T *X}$ and use it to multiply the equation (from left), and obtain a unique solution for the vector of regression coefficient $\boldsymbol{\beta = (X^T * X)^{-1} * X^T * Y}$. It is assumed that all the coefficients in $\boldsymbol{\beta}$ have a zero-centered normal distribution, and so we can estimate their variance and covariance values. Specifically, $\boldsymbol{Cov(\beta) = \sigma^2 * (X^T * X)^{-1}}$, where $\boldsymbol{\sigma^2}$ is the variance of the fit. These properties pave the way for testing hypotheses about the estimated values of regression coefficients $\boldsymbol{\beta}$. It should be noted that since $\boldsymbol{Y}$ was actually measured in triplicate, we concatenated the 3 sets of values so that $\boldsymbol{n}=|\boldsymbol{Y}|= 15$. We also replicated the design matrix $\boldsymbol{X}$ to match. This allowed for more accurate regression, by estimating the error in each array separately. Calculations were performed based on the REGRESS function of MATLAB, (version 7.0 R14), and following (DeGroot and Schervish, 2002). The actual expression measurements and resulting $\boldsymbol{\beta}$ components I regressed are shown for few examples in Figure 2b, with the regressed values for all HOG pathway genes in Figure 2c (see Results for more details. Analysis of the regression quality is shown on Figure S1a).

The same approach was applied to dissect the pair-wise interactions between additional two transcription factors, which are controlled by Hog1 (Sko1 and Hot1; See Results), and the general stress regulators Msn2/4Δ (See Figures 3 and S1b). Specifically, I determined the values of three components (SH for the expression component related to Sko1Hot1, M for the Msn2/4 effect, and SHM for the effect of their interaction) by comparing gene expression in the wt strain, *msn2/4Δ*, *sko1Δhot1Δ* and *sko1Δhot1Δmsn2/4Δ,* using the matrix below

$$\begin{bmatrix} \text{wt vs } \textit{sko1}\Delta\textit{hot1}\Delta \\ \text{wt vs } \textit{msn2/4}\,\Delta \\ \textit{msn2/4}\,\Delta \text{ vs } \textit{sko1}\Delta\textit{hot1}\Delta\textit{msn2/4}\,\Delta \\ \textit{sko1}\Delta\textit{hot1}\Delta \text{ vs } \textit{sko1}\Delta\textit{hot1}\Delta\textit{msn2/4}\,\Delta \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \begin{bmatrix} \text{SH} \\ \text{M} \\ \text{SHM} \end{bmatrix}$$

In addition, we extended the Mutant Cycle approach to examine the three-way interaction between Sko1, Hot1 and Msn2/4. For this we used the following experiments:

- wt vs *msn2/4Δ*
- wt vs *sko1Δhot1Δ*
- *sko1Δhot1Δ* vs *sko1Δhot1Δmsn2/4Δ*
- *msn2/4Δ* vs *sko1Δhot1Δmsn2/4Δ*
- wt vs *hot1Δ*
- *hot1Δ* vs *hot1Δmsn2/4Δ*
- *msn2/4Δ* vs *hot1Δmsn2/4Δ*
- wt vs *sko1Δ*
- *sko1Δ* vs *sko1Δmsn2/4Δ*
- *msn2/4Δ* vs *sko1Δmsn2/4Δ*

Here, we decomposed these measurements into the sum of ten components, reflecting the effect of each factor: Sko1, Hot1, Msn24, and each combination of two or three factors: Sko1Hot1, Sko1Msn24, Hot1Msn24, and Sko1Hot1Msn24.

As before, we formulated the measurements as a noisy matrix multiplication:

$$\begin{bmatrix} \text{wt vs } \mathit{msn2/4}\,\Delta \\ \text{wt vs } \mathit{sko1}\Delta\mathit{hot1}\Delta \\ \mathit{sko1}\Delta\mathit{hot1}\Delta \text{ vs } \mathit{sko1}\Delta\mathit{hot1}\Delta\mathit{msn2/4}\,\Delta \\ \mathit{msn2/4}\,\Delta \text{ vs } \mathit{sko1}\Delta\mathit{hot1}\Delta\mathit{msn2/4}\,\Delta \\ \text{wt vs } \mathit{hot1}\Delta \\ \mathit{hot1}\Delta \text{ vs } \mathit{hot1}\Delta\mathit{msn2/4}\,\Delta \\ \mathit{msn2/4}\,\Delta \text{ vs } \mathit{hot1}\Delta\mathit{msn2/4}\,\Delta \\ \text{wt vs } \mathit{sko1}\Delta \\ \mathit{sko1}\Delta \text{ vs } \mathit{sko1}\Delta\mathit{msn2/4}\,\Delta \\ \mathit{msn2/4}\,\Delta \text{ vs } \mathit{sko1}\Delta\mathit{msn2/4}\,\Delta \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} x \begin{bmatrix} \text{Sko1} \\ \text{Hot1} \\ \text{Msn2/4} \\ \text{Sko1Hot1} \\ \text{Sko1Msn2/4} \\ \text{Hot1Msn2/4} \\ \text{Sko1Hot1Msn2/4} \end{bmatrix}$$

and found a β which minimizes the system errors.

Statistical significance. To assign a statistical significance value for each expression components ***β*,** I developed a statistical procedure. According to the null hypothesis $\boldsymbol{H_0}$, which suggests that the gene is not being regulated by none of the TFs (nor by their combinations), each regression coefficient $\boldsymbol{\beta_j}$ should equal zero. In this case, the ratio $\boldsymbol{\beta_j / std(\beta_j)}$ distributes according to a ***t***-distribution with ***(n-p)*** degrees of freedom, where ***n*** is the number of experiments in the cycle (15 in the case of Hog1Msn2/4), and ***p*** is the dimension of ***β*** (3 in the case of Hog1Msn2/4) (DeGroot and Schervish, 2002). As shown above, $\boldsymbol{Cov(\beta) = \sigma^2 * (X^T * X)^{-1}}$ and so $\boldsymbol{std(\beta_j)}$ is the squared root of the ***j***th value along the diagonal of the covariance matrix ***Cov(β).*** By computing the cumulative distribution function of the ***t*** distribution, I estimated the likelihood of $\boldsymbol{\beta_i}$ under $\boldsymbol{H_0}$. This estimation approximates the probability of seeing such a value (or larger) at random, thus serves as a "p-value" for $\boldsymbol{\beta_j}$ (DeGroot and Schervish, 2002). I actually expanded this approach to compute additional critical values. For example, a more interesting hypothesis $\boldsymbol{H_0}$ would allow some minimum regulation by each TF (hence, replace the zero by some activation threshold ***thr*).** We defined ***thr*** to account for a non-marginal expression component (above 1.5-fold), and used a similar approach to test if the contribution of some factor is significantly above the threshold (for ***H*** and ***M***, which are assumed to be positive) or non-marginal (two-tailed version, for ***Co***, which could be either positive or negative). In this case, we assumed that under the null assumption $\boldsymbol{H_0,}$ the mean value of $\boldsymbol{\beta_j}$ is smaller than ***thr***, and so $\boldsymbol{(\beta_j - thr) / std(\beta_j)}$ should has a t-distribution with ***(n-p)*** degrees of freedom.

Accuracy of the Approach

To ensure that the expression components fully and accurately account for the raw microarray data we compared the data predicted from the fitted component values back to the raw data used to calculate these values (Figure S1). These plots demonstrate that the expression components determined in the global fit to the array data accurately and completely describe the expression changes found in the individual mutant strains.

Model-Based Analysis of Genome-wide High-resolution ChIP Data

To further support the expression components which were regressed from gene expression data using the method above, we decided to measure the *in vivo* physical binding of the same transcription factors to DNA (see Results). The experimental procedure included chromatin immunoprecipitation (ChIP) data, coupled with hybridization to dense tiling microarrays. To analyze these data, I developed a computational algorithm, based on a concrete physical model which describes how the signal of each binding event is "smeared" at the ChIP signal (due to the length of the IP'ed DNA fragments). This algorithm allowed me to robustly identify the exact position and strength of protein-DNA binding events.

Estimated Shape of a Binding Site

Due to the typical length of the sheared DNA fragments, a binding event at position ***x*** results in high enrichment of IP'ed DNA at the surrounding probes (Figure S2). This effect decays as the distance between the probe and the binding position increases. In general, the probability of a probe, located ***Δ_x*** bases away from the binding location, to report is proportional to the integration over all fragment lengths (of length ***Δ_x*** or longer), multiplied by the number of possible alignments of the DNA fragments that allow both the binding of the fragment by the target transcription factor and its hybridization to the reporting probe, times the relative abundance of DNA fragments of such length, denoted by ***c(l)***. Thus, the estimate for a peak's shape is given in the following equation:

$$F(\Delta_x) \propto \int_{l=\Delta_x}^{\infty} (l - \Delta_x) c(l) dl$$

The distribution of sheared fragment lengths ***C(l)*** depends on the sonication protocol. My collaborators measured the fragment length distribution created by our protocol using an agarose gel and found a broad distribution of fragment lengths (200-2000 bp) that is well described by a Gamma distribution. This distribution has two parameters that control the mean and standard deviation of fragment length. In subsequent experiments I used these two parameters to define the entire fragment length distribution ***c(l)***.

The Peak Fitting Algorithm

I developed an iterative algorithm to identify all significant binding events that appear in the probes (Figure S2). Briefly, this is done by identifying stretches of enriched probes and attempting to explain (at least part of) their values using the peak model. My algorithm chooses the most probable values for center position and peak height, and computes the statistical significance of this peak. If its p-value falls below 0.01 (see below), and its height exceeds 1.5, it is classified as a binding event, and its expected shape (i.e. predicted enrichments for the probes in ***S***) is subtracted from the actual ratios. This enables us to identify overlapping peaks one at a time (starting from the strongest one), until the remaining data cannot be distinguished from noise. My model-based approach also allows us to naturally integrate data from different replicates, computing the likelihood of the peak based on *all* enrichment values of its probes. I now describe in more details the relevant steps.

Optimization of Peak Parameters

Once a window **S** of consecutive probes with enriched values is found, my algorithm searches for optimal peak parameters to fit the enrichment rations in **S** in a two-step manner. First it enumerates over the peak center point $\boldsymbol{x}$ in a 10 bases resolution, then, for each position, it estimates the optimal height $\boldsymbol{\alpha}$ which minimizes the sum of squared deviations between the log (base 2) of the measured enrichments and ones predicted by peak's shape (Figure S2, the enrichments measured by each probe are shown in red, and the predicted shape of the peak plotted in blue). This is done using Brent's method for one-dimensional minimization. Finally, we report the position $\boldsymbol{x}$ and height $\boldsymbol{\alpha}$ whose fit was the optimal.

Estimating the statistical significance of binding events

The statistical significance of a binding event is estimated by computing an empirical log-likelihood ratio (*LLR) p-value***.** Specifically, the likelihood of the enrichments measured using the set of probes $\boldsymbol{S}$ surrounding the binding event, can be computed using a null model $\boldsymbol{L_0}$. According to this model, there are no binding events, and so all enrichment values originate from noise. To model this, I used a *Normal* distribution whose mean and variance were estimated from the DNA microarray. I then computed the likelihood of the same probes given my model. As before, I used a Normal distribution, only now the distribution mean value was set according to the expected shape of the peak. Finally, each binding event was scored according to the log-likelihood-ratio (LLR) based on my model vs the null model $\boldsymbol{L_{peak}/L_0}$. To assign each score a p-value, I computed 1000 shuffling-based LLR scores in the following way: First I replaced the values of all the measured probes in $\boldsymbol{S}$ with randomly-sampled values from the array. Then I re-estimated the optimal height at this position (now consisting of random enrichment values), and calculate the log-likelihood-ratio score for this shuffled set. The p-value of each binding event is then calculated by comparing the true LLR score to those of the randomly-shuffled sets.

Computing a Bayesian Confidence Interval around a binding event

In addition to estimating the peak's center position and height, I also compute a 99%-confidence interval around the binding position of each peak. This is done by considering the likelihood $\boldsymbol{L_{peak}(x)}$ of the peak's probes $\boldsymbol{S}$, when centered at position $\boldsymbol{x}$. I then use Bayes' rule to compute the posterior probability of the center being at each position $\boldsymbol{x}$, and define the Bayesian Confidence Interval as the region covering 99% of the posterior probabilities.

The entire peak fitting process is sketched in the algorithm below:

Algorithm:

```
1. Estimate shape of a peak $F(\Delta_x)$
2. Initialize enrichment ratio threshold $T$ to 10
3. Set cooling factor $K$ to 0.99
4. Let $B$ be the set of binding events
5. Begin main loop:
    1. For every consecutive set $S$ of probes above threshold $T$
        a) Add flanking probes (up to 2.5Kb away) into set $S$
        b) Find center position $x$ and height of peak $\alpha$ to fit S best.
        c) Calculate likelihood-based p-value of peak
        d) If peak is significant, and its estimated height is above 1.5:
              I. Calculate 99% Bayesian Confidence Interval
             II. Add peak into set of binding events B
            III. Predict values for probes in S using B, and subtract from data
    2. Update the enrichment threshold $T = T$ * cooling factor $K$
    3. Repeat main loop until no new significant binding events are found
```

Definition of Yeast Promoters

Promoter regulatory sequences were defined using sequences and annotations from the UCSC genome browser (sacCer1). For genuine genes (UCSC track *sgdGene*), promoters were defined as the regions upstream to the translation start site, up to 1 Kb or up to the coding regions of upstream genes. As for pseudo and dubious genes (UCSC track *sgdOther*), I defined the regulatory regions as 500-bp upstream to the translation start site, regardless of overlapping coding regions.

Genome-Wide Analysis of Bound Genes

Once the genome-wide ChIP data were analyzed and binding events (or peaks) were identified, the overall IP-based enrichment of each gene was estimated by summing the enrichment values of all binding events occurring over its promoter region.

To distinguish between bound and unbound genes, I set a threshold corresponding to 5% false positive rate over a control group of non-target genes (genes outside of the Hog1 network)

RESULTS

A quantitative model of the Hog1-Msn2/4 Network

To examine the interaction between Hog1 and Msn2/4 in detail, I used the mutant cycle approach (Figure 2a) and determined the value of the three expression components in the system: H, M and Co. Gene expression was examined 20 min after induction of stress (0.4 M KCl) since this is near the peak of the transient response (O'Rourke and Herskowitz, 2004) but is early enough to avoid monitoring secondary effects in the mutant strains (Hog1 and Msn2/4 are transcriptionally inactive in pre-stress conditions). To my surprise, even within the HOG pathway, the influence and interaction of Hog1 and Msn2/4 varied dramatically from gene to gene (Figure 2b). The genes were divided into eight distinct regulatory modes, based on the combination of statistically significant expression components at genes induced in osmotic stress (Figure 2c). From these data it is clear that: (i) Hog1 and Msn2/4 interact, since 190 of the 273 genes in the network have a statistically significant Co component (Groups 1, 2, 5, 7, 8; Figure 2c); and (ii) that both Hog1 and Msn2/4 are activated and enhance the expression of their target genes separately, since significant H or M components are found at 112 (Groups 4-8; Figure 2c) and 64 genes (Groups 2,3, 6-8; Figure 2c), respectively.

It is not possible to translate these interaction data directly into a mechanistic network wiring diagram, specifying which genes are regulated by which transcription factors (or their combinations), and to which extent. This is because the cooperative interaction between Hog1 and Msn2/4 could be established either at the promoter level (Hog1 and Msn2/4 interacting on the DNA itself) or at the signaling level (e.g. the nuclear activity level of Msn2/4 regulated by Hog1, see Figure 1a, bottom part). We surmised that the interaction between Hog1 and Msn2/4 is likely to be established, at least in part, at the signaling level, since Hog1 is a protein kinase and is required for full expression of almost all Msn2/4-dependent genes (190/203; Groups 1, 2, 5-7; Figure 2c). Therefore, to test for activation of Msn2/4 by Hog1, my collaborators monitored the stress-induced import of Msn2/4 into the nucleus in wild-type and *hog1*Δ mutant strains containing GFP-tagged Msn2 or Msn4 and a nuclear marker. In brief, their microscopy analysis showed that Hog1 is imported into the nucleus upon KCl stress and that it contributes to the nuclear localization of Msn2 (two-fold change in nuclear levels of Msn2 in wt and *hog1*Δ strains), although Msn2 is imported into the nucleus by some other pathway.

Given these connections at the signaling level, the data from the Hog1-Msn2/4 mutant cycle (Figure 2c) can be explained by a simple wiring diagram (Figure 3a) in which the Co component is assigned to Hog1-dependent gene activation through Msn2/4 while the H and M components are due to direct activation by Hog1 and Msn2/4, respectively. This Hog1-Msn2/4 network model defines only three classes of genes (Figure 3a): (I) genes regulated by Hog1 alone; (II) genes regulated primarily by Hog1 through Msn2/4 (3 genes by Msn2/4 only); and (III) genes regulated by Hog1 both through Msn2/4 and independently of Msn2/4 (mixed regulation). However, the genes in Classes II (Groups 1-3) and III (Groups 5-8) show distinct behavior in deletion mutants, resulting in several groups in the expression component analysis (Figure 2c). This can be explained if different groups of genes within a given class have different thresholds for gene activation by Msn2/4: high, low or intermediate. For example, genes in Groups 1 (Co) and 5 (H+Co) appear to have a high threshold for activation by Msn2/4 as they are insensitive to the low levels of nuclear Msn2/4 found in the absence of Hog1 (Figure 2c; no M component). In contrast, genes in Groups 3 (M) and 6 (H+M) appear to have a low threshold for activation by Msn2/4 as they are fully activated at the low levels of nuclear Msn2/4 found in the absence of Hog1 (Figure 2c; M but no Co component). Finally, genes in Groups 2 (M+Co) and 7 (H+M+Co) appear to have an intermediate threshold for activation as they are partially activated at the low nuclear level of Msn2/4 (Figure 2c; M and Co component).

Incorporation of Sko1 and Hot1 into the Network Model

To explain how Hog1 activates genes independently of Msn2/4 (112 genes with an H component, Groups 4-8 Figure 2c), we focused on the transcription factors activated by Hog1 following salt induction. Hog1 was shown to phosphorylate, activate and/or bind to the TFs, Msn1, Smp1, Sko1, Hot1 and Cin5, but few target genes have been identified for these factors (Proft and Serrano, 1999; Rep *et al.*, 1999; de Nadal *et al.*, 2003; Nevitt *et al.*, 2004). It was previously suggested that Sko1 acts as a *transcriptional repressor* in the absence of stress, switches to act as a *transcriptional activator* following osmotic stress (Proft and Serrano, 1999; Proft and Struhl, 2002). My collaborators measured the expression levels of strains lacking one or more of these factors, and found that only Sko1 and Hot1 play a significant role in the osmotic stress response.

To incorporate these factors into the network model, we applied the mutant cycle approach and dissected the influence of, and interaction between, Sko1/Hot1 (together) and Msn2/4 (Figure 3b, red cycle). Interestingly, my analysis found only few genes with positive interaction (AND-like) between these factors (Figure 3c), opposed to a much larger number of genes with negative interaction (OR-like). Most of these genes are weakly affected by Sko1/ Hot1 in the presence of Msn2/4, but induce transcription up to 100-fold in the absence of Msn2/4 (Figure 3c, bottom). My analysis also indicates that additional 10 genes are activated by Sko1/Hot1 and Msn2/4 independently (i.e. with no significant cooperativity). I found a striking correlation (R=0.90, Figure 3d) between the original H component (as determined by the Hog1-Msn2/4, Figure 2a) and its representation by the sum of the Sko1/Hot1 component (Figure 3b, red cycle) plus the extent of transcriptional repression by Sko1 prior to induction. This analysis suggests that Msn2/4-independent gene induction by Hog1 occurs almost entirely through Sko1 and Hot1. My collaborators further addressed this point directly by measuring the transcriptional effect of deleting Hog1 in the absence of Sko1, Hot1 and Msn2/4.

Detailed transcriptional dissection of Hog1 transcriptional network

To further examine the influence that Sko1, Hot1 and Msn2/4 have on gene expression individually, and quantify the interaction between Sko1-Msn2/4, Hot1-Msn2/4, Sko1-Hot1 and Sko1-Hot1-Msn2/4, I extended the mutant cycle algorithm to look at three-way interactions (black and red cycles and black components, Figure 3b). This allowed me to fully dissect transcriptional regulatory interactions at gene promoters and accurately measure the influence of Hot1 and Sko1 separately, even where they act redundantly with Msn2/4. These expression components were used to expand our initial Hog1-Msn2/4 network model (shown in Figure 3a) into a detailed model of the Hog1 transcriptional network in salt-induced osmotic stress (Figure 3e).

Analysis of genome-wide *in vivo* binding of Sko1 and Hot1

To validate our network model, and gain insight into its structure, my collaborators used genome-wide chromatin immunoprecipitation assays (Ren *et al.*, 2000; Iyer *et al.*, 2001), followed by hybridization to dense tiling array of the yeast genome.

Specifically, we measured the *in vivo* binding locations of Sko1 and Hot1, both prior to and following KCl-induced hyper-osmotic stress. To analyze these data, I developed a model-based algorithm to identify the exact location of binding sites, and their relative enrichment. As opposed to algorithms presented in parallel studies (Buck *et al.*, 2005; Gibbons *et al.*, 2005; Kim *et al.*, 2005; Li *et al.*, 2005; Qi *et al.*, 2006), my model is based on the typical peak-shaped signal of each binding event, caused by the variation in the length and binding position among the immunoprecipitated DNA fragments (see Methods). This analysis dramatically reduced the number of false positive calls due to noisy probes, since each binding event is actually characterized using 8-12 of its neighboring probes (see Methods, Figures 4a and S2a). Moreover, due to the iterative nature of my algorithm, it facilitates the identification of several binding events in vicinity of each other (Figure S2b-c).

Analysis of the ChIP-chip data revealed that most Sko1 and Hot1 binding events occur inside promoter regions (80% of the 100 peaks with an enrichment ratio >5 in KCl). I then identified all promoter binding events and estimated their statistical significance. I used the conservative assumption that all binding events in the promoter regions of ~6000 genes outside of the Hog1 network (273 genes in Figure 2c) are spurious, and can be treated as a background reference (see Methods).

This analysis revealed an excellent agreement with the Sko1/Hot1 target genes identified through gene expression analysis (Groups I and II of Figure 3a): 65-80% of the genes repressed by Sko1 (27 total), activated by Sko1 (52 total), or activated by Hot1 (15 total) were found to be significantly bound by the factor in the appropriate condition (Figure 4b). I also found higher than expected binding of Sko1 and Hot1 at other genes within the Hog1 network. In fact, 42 additional Sko1 target genes and 23 additional Hot1 target genes were identified based on ChIP-chip analysis (Figures 4c and d). While some of these bound genes were missed by the expression-based analysis due to barely significant p-values, most of these binding sites are in fact *latent binding sites* (24/42 Sko1 and 17/23 Hot1), with negligible expression components (p-value>0.80).

The ChIP analysis provides additional insight into the dual function of Sko1 in transcriptional regulation. There are three classes of Sko1 binding behavior within the Hog1 transcriptional network: (i) Sko1 binds to the promoter in pre-stress conditions

(YEPD), but is released within 5 min of KCl stress (6 genes in total, including FSH1 and HXT4 shown in details in Figure 4a); (ii) Sko1 is constitutively bound to the promoter (45 genes in total, including YHR033W, HXT1 and HXT5 from Figure 4a); (iii) Sko1 is only recruited to the promoter following stress treatment (37 genes, including YHR087W in Figure 4a). This variable behavior of Sko1 is functionally important due to the dual roles of Sko1 as both a transcriptional repressor (in YEPD) and activator (in KCl). Indeed, these three binding modes are reflected by the resulting expression components of Sko1, with typical activation for genes bound by Sko1 in KCl (Figure 4c, bars, top), as opposed to a combination of repression and activation for genes constantly bound (Figure 4c, bars, bottom). Analysis of the Hot1 bound promoters reveal a slightly simpler picture, mixing constitutive (7 genes) and inducible (28 genes, including YHR087W and HXT1 shown in Figure 4a, right panel).

To summarize, using Sko1 and Hot1, Hog1 can control the expression levels of its target genes in five distinct combinations: constitutive Sko1 binding with/without Hot1; inducible Sko1 binding with/without Hot1; and pre-stress only binding of Sko1. These results highlight the accuracy of our two methods for identifying the target genes of HOG related transcription factors – using the mutant cycle approach and the model-based analysis of high-resolution ChIP-chip data.

Signal Integration in the Hog1 Network

Taken together, these data provide a detailed model of the Hog1 transcriptional network in KCl-induced osmotic stress (Figure 3e). Examination of this network reveals that the external stimuli sent through Hog1 and the general stress (Msn2/4) pathways are integrated at two levels. At the cellular signaling level, Hog1 further activates Msn2/4 by mediating its nuclear import. At the promoter level, activation of Hog1 is transmitted *via* Sko1 and Hot1, which cooperates with Msn2/4 in a variety of distinct modes (Figure 3e).

DISCUSSION

Previous analysis of the Hog1-dependent stress response led to a coarse-grained model of Hog1 function where the MAP-kinase Hog1 regulates gene expression through three independent paths: activation of Msn2/4, activation of Hot1, and de-repression of Sko1, with Sko1/Hot1 acting at only 12 genes (Rep *et al.*, 2000; Hohmann *et al.*, 2007). Using a detailed analysis of gene expression in wt, single and multiple mutant strains, we converted this incomplete and qualitative description into a quantitative and nearly complete network model (Figure 3e). This model shows that the signal from Hog1 is spread out to several transcription factors and then recombined in several distinct structures at the promoters of HOG-related genes (Figure 3e). This network architecture allows stress signals transmitted through Hog1 to enhance the general stress program via Msn2/4, and to fine-tune this reaction using additional transcription factors (Sko1 and Hot1). Overall, our model of the Hog1 network provides insight into the way a signal can create a context-dependent gene expression program using a limited number of transcription factors.

Beyond establishing the structure and function of the Hog1 transcriptional network, these results demonstrate the utility of double mutant analysis, and the overall strategy taken here, for dissecting gene regulatory systems. I have shown that, starting with two or more known/putative network components, it is possible to build a quantitative genome-wide network model and to identify the genes regulated by missing components. By performing a screen for the factors that act on these genes (using computational and experimental means), it is possible to identify the missing components and integrate them into the network model. This approach has immediate application to studying conditionally activated pathways (and drug-pathway interactions) using gene KOs, and can be extended to other systems through the use of RNAi and chemical inhibitors.

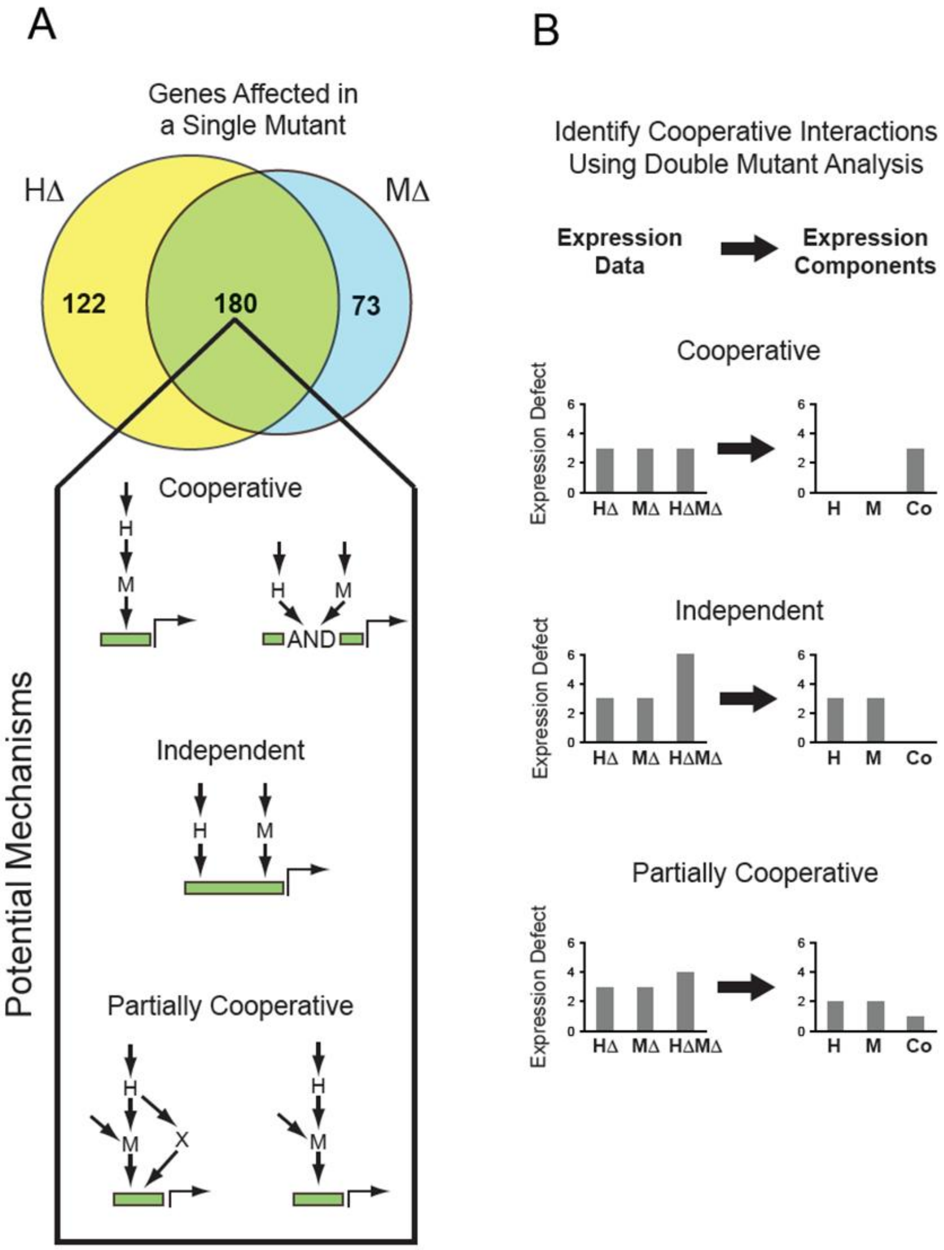


Figure 1. Single and double mutant analysis of gene expression. **(A)** Venn diagram summarizing the overlap in the number of genes with a >2-fold defect in gene expression in the *hog1*Δ (HΔ) and *msn2*Δ *msn4*Δ (MΔ) mutants, following salt induction. Wiring diagrams indicate the possible ways factors H and M can interact to regulate expression of overlapping sets of genes. **(B)** Schematic illustrating the application of the double mutant approach to analyzing transcriptional network structure and function (see text for details).

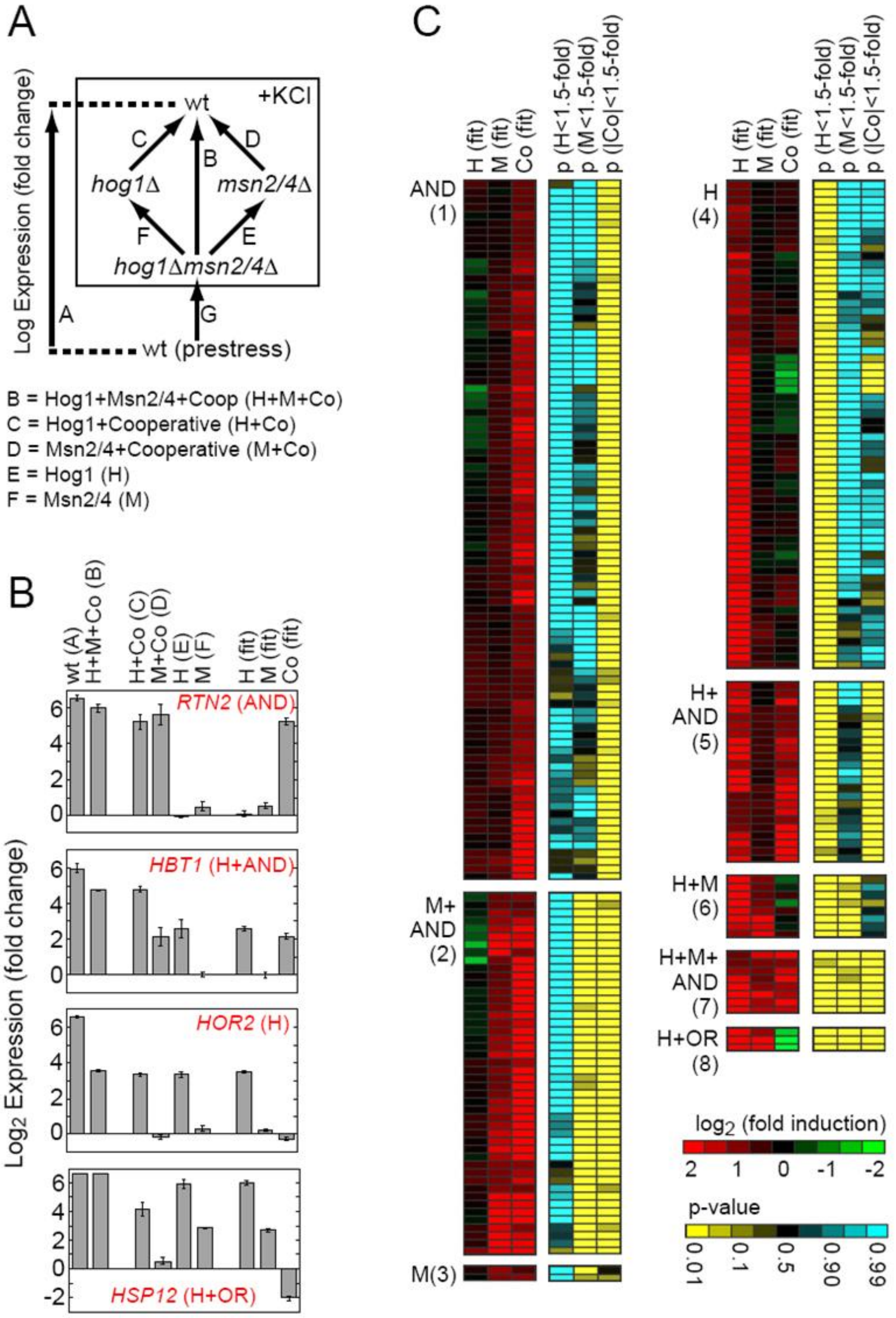


Figure 2. Role of Hog1 and Msn2/4 in osmotic stress-dependent gene induction. **(A)** Schema describing the experiments and equations used to break the influence of Hog1 and Msn2/4 into components. Each arrow represents a single microarray (measured in triplicate) comparing gene expression in two strains. The equations listed below the diagram describe the relationship between the data from each measurement and the underlying expression components. Note here that expression is in Log terms (thus additive) and so an OR-like gate is represented as a negative cooperative component equal to the H or M component (such that H=M=H+M+Co). **(B)** Sample data for four genes showing the errors associated with the microarray measurements and expression component values. **(C)** Heat map showing the regressed value of the expression components (red/green), and their statistical significance (yellow/blue), for all HOG-pathway genes (defined by up-regulation in response to hyper-osmotic stress, either by Hog1 (≥3-fold) or by Msn2/4 (≥2-fold). Each row of six columns shows the data for a single gene. Genes were clustered into groups (1-8) and labeled according to the statistical significance of the various expression components that influence their induction (using a p-value threshold of 0.05; AND = +Co; OR = -Co).

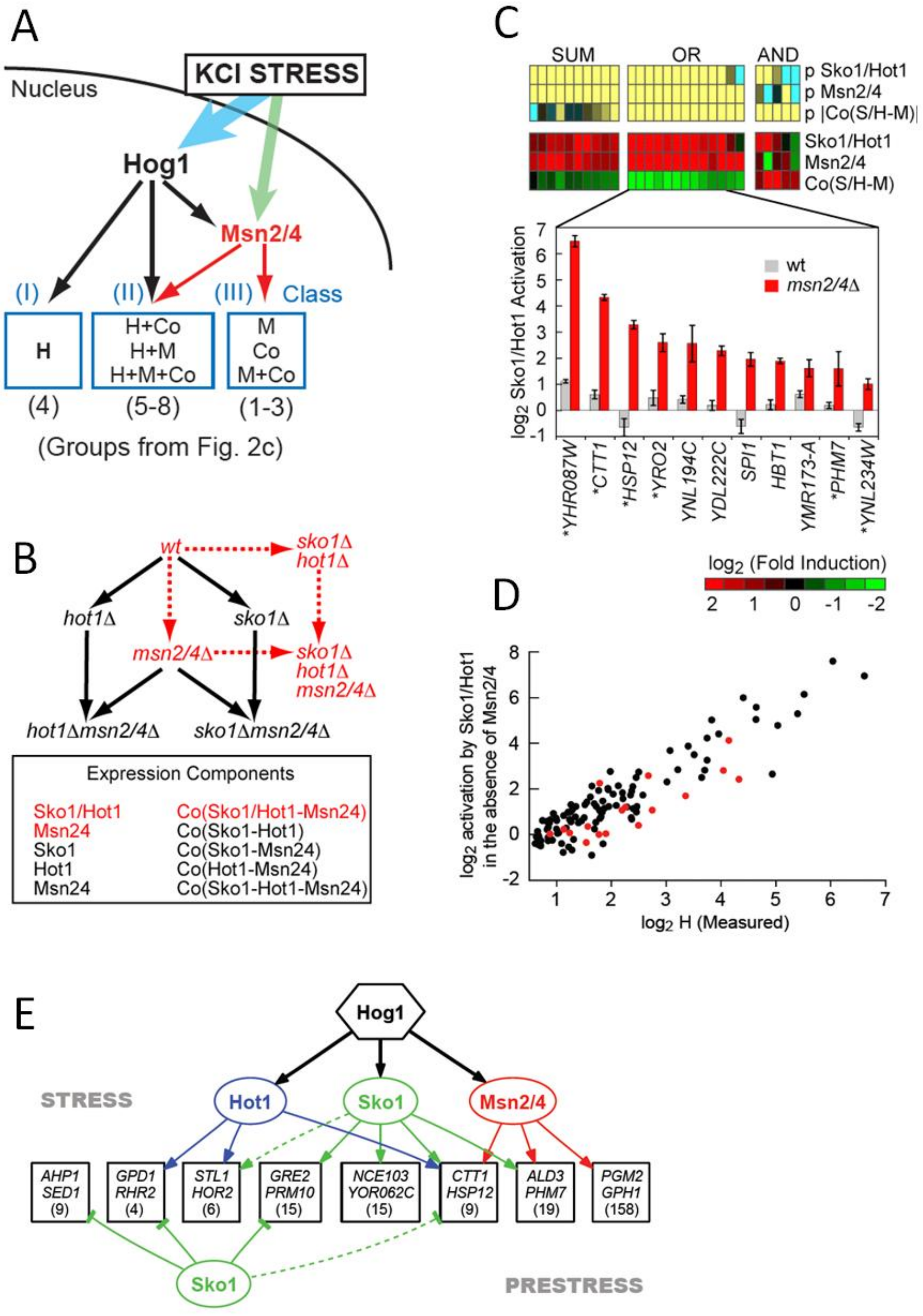


Figure 3. Mechanism of Hog1-dependent gene activation. **(A)** Model of the Hog1 transcriptional network, explaining the expression component data found in Figure 2. **(B)** Schema describing the experiments and equations used to dissect the expression components of Sko1, Hot1 and Msn2/4. **(C)** Interaction between Sko1/Hot1 and Msn2/4. Heat map showing the regressed expression components, and their statistical significance. Only shown are genes with a statistically significant components by Sko1/Hot1 and Msn2/4 (p-value<0.01). The bars show the transcriptional effect of raw data for the sko1Δhot1Δ vs.˜ wt (denoted wt) and sko1Δhot1Δmsn2Δmsn4Δ /msn2Δmsn4Δ (denoted msn2/4Δ) data for 11/13 OR gate genes where Sko1/Hot1 activity is redundant in the wild-type strain. Gene names highlighted by a star are activated by both Sko1 and Hot1 (in some cases redundantly), other genes are just activated by Sko1. **(D)** Correlation between the level of induction measured for Hog1 alone (H component, Figure 2) and its decomposition to regressed Msn2/4-independent expression components, equals the sum of join effect of Sko1/Hot1 in the absence of Msn2/4 (Sko1/Hot1 component, red cycle part Figure 3b) plus the extent of Sko1 repression in YEPD. **(E)** Structure of the transcriptional network activated by the MAPK Hog1. Genes are grouped based on common regulatory mechanisms (denoted by a box with the names of two sample genes) and only shown if two or more genes have the same connections as determined by expression and confirmed by ChIP. Broken lines indicate interactions that that exist for only part of a group. The number in each box refers to the number of genes in a group based on expression data alone. To simplify the figure *latent binding* events are not shown and there is no representation of cooperativity at the promoter level.

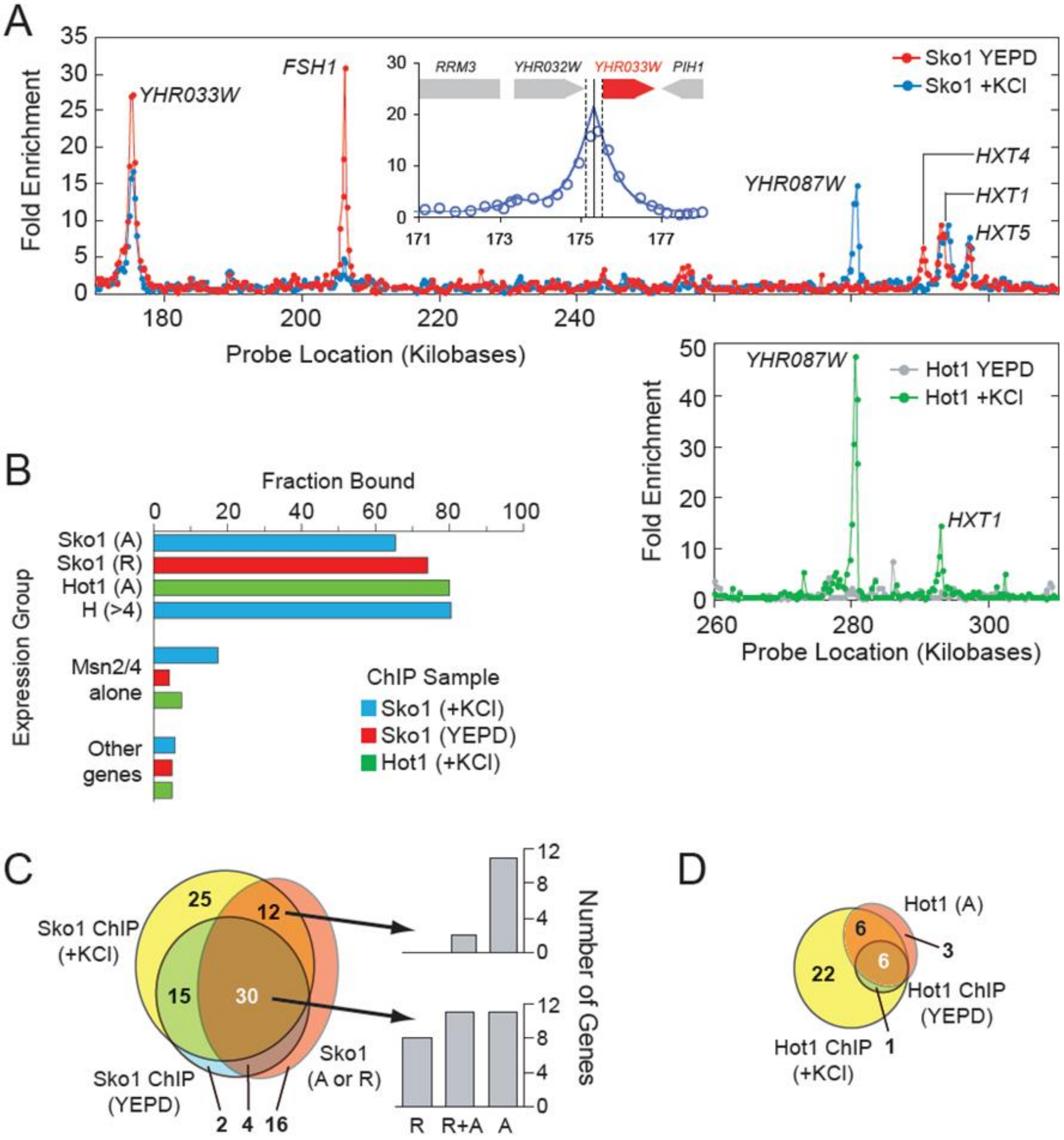


Figure 4. ChIP analysis of Sko1 and Hot1 binding sites **(A)** Sample raw data for Sko1 (upper panel) and Hot1 (right panel) for a region of chromosome 8 (shown ~1% of the genome). Each data point shows the enrichment ratio as measured by one probe on the microarray. The inset shows a typical example of a binding event. The measured enrichment values are shown by circles, while its optimal fit by a model-derived peak shape , used to analyze the data (see Methods) is shown in blue. The vertical solid line shows the optimal binding position, while the dotted lines show the 99% confidence interval. **(B)** Overlap of ChIP and expression data. The target genes for Sko1, Hot1 and Msn2/4 alone ($p<0.05$) were compared to the bound genes identified by the ChIP analysis from the peak fitting ($p<0.05$). **(C)** Venn diagram showing the overlap between ChIP data ($p<0.05$) and expression data ($p<0.058$) for Sko1. The number of binding sites at genes without significant Sko1 induction and/or repression was adjusted for the expected number of false positives. The bar graphs show the number of genes that are repressed (R), repressed and activated (R+A) or just activated, for genes where there is both significant binding and expression data. **(D)** Venn diagram showing the overlap between ChIP data ($p<0.05$) and expression data ($p<0.05$) for Hot1. Again here the number of binding sites at genes without significant Hot1 induction is corrected for the number of false positives expected.

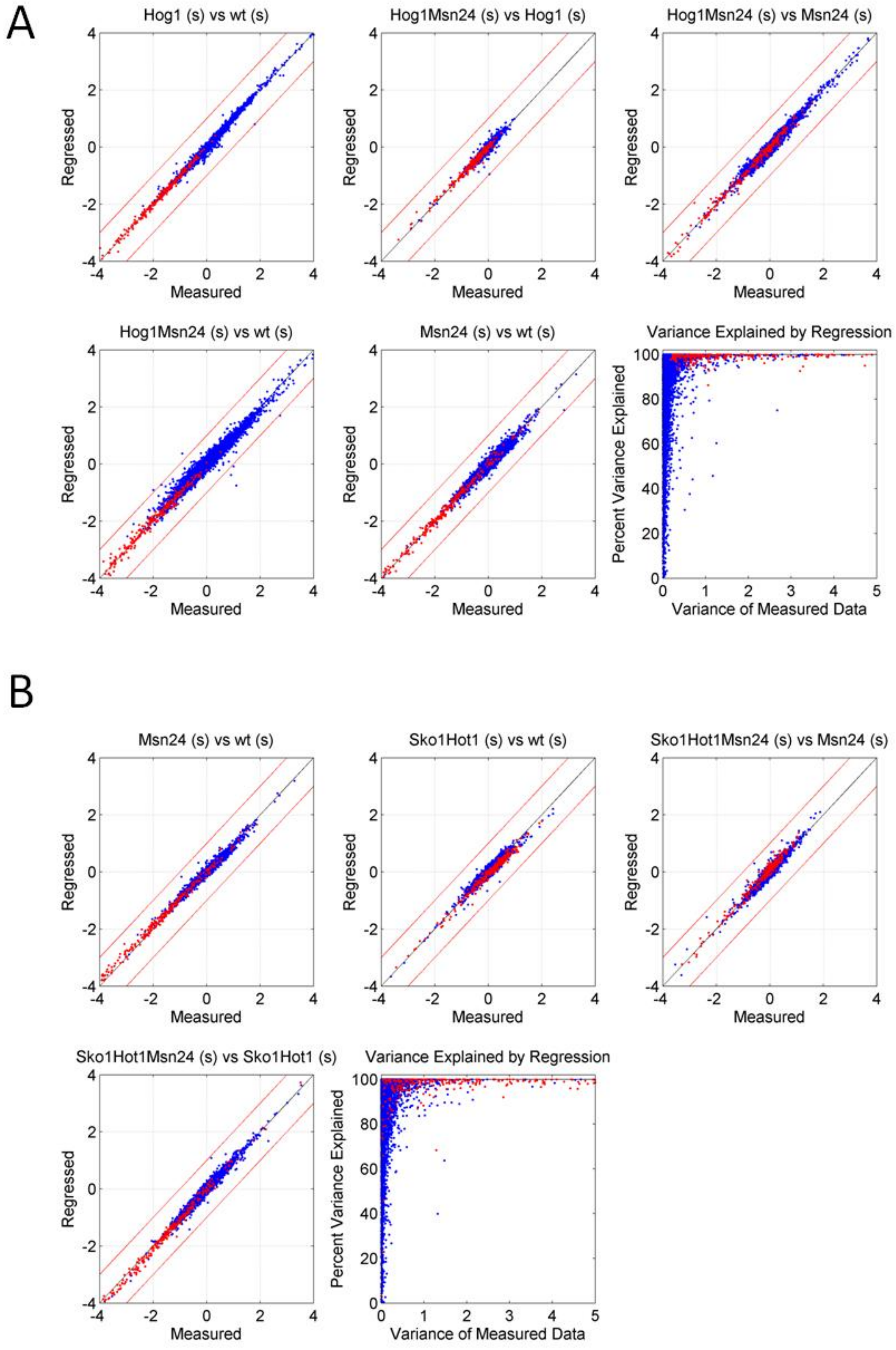


Figure S1. **(A)** Goodness of fit: Hog1 vs Msn2/4 cycle. The data from each microarray experiment B-F in Figure 2a are plotted vs. their reconstruction as combinations of the three regressed values (H, M, and Co). Each point shows the measured (X-axis) vs. regressed (Y-axis) log2 fold-change for a single gene, colored red if included in the Hog1 pathway genes (273 genes in total), or blue if outside of the network. The last panel shows the percent of total variance explained by the fit (color coded as above). For each gene, I computed the variance (V) of expression (over the expression measurements, B-F). The percent of variance explained by the regression is given by 100*(V-R)/V, where R is the variance of the residual data. In general, Hog1 related genes (shown in red) present a high variance, due to significant expression components, and are well fitted. Other genes present low variance (few transcriptional changes). **(B)** Goodness of fit: Sko1/Hot1 vs Msn2/4 cycle. Same as **(A)**, except that the arrays and fitted components are for Figure 3b (red cycle).

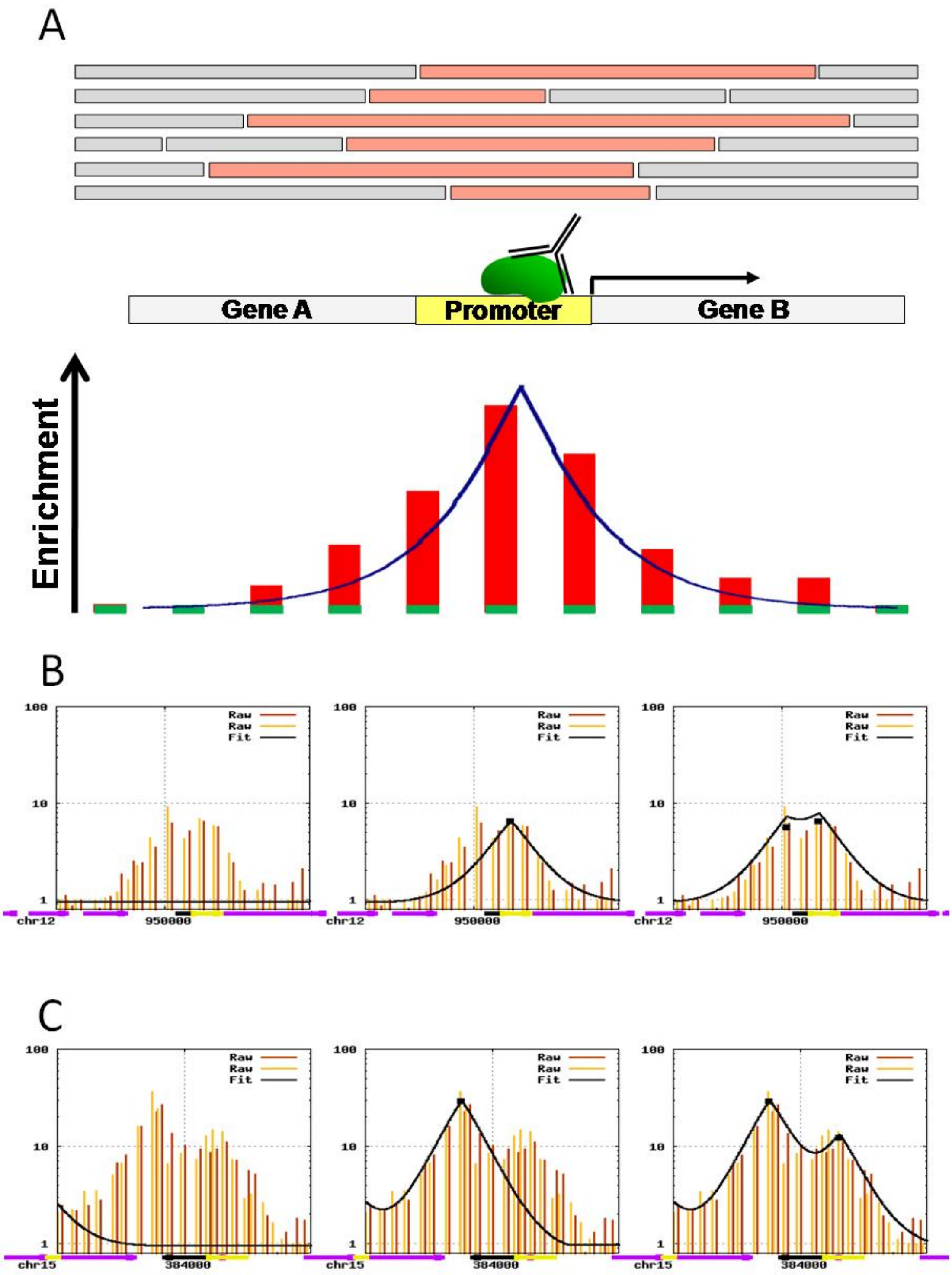


Figure S2. Model-based analysis of genome-wide high-resolution ChIP data. **(A)** In the chromatin immunoprecipitation procedure, TFs (green) are crosslinked to DNA fragments (pink), which are then purified and hybridized to densely tiled genomic microarray (green). The red bars correspond to the relative enrichment of IP'ed DNA, fitted by a model-based estimation of the shape of a binding event (blue). **(B-C)** Iterative steps of the peak fitting algorithm allow the exact identification of several binding events in close proximity. Shown are the fit of two statistically significant binding events by Sko1 in the promoter regions of CIN5 (YOR028C; shown in B) and YLR412C-A (C).

Chapter 3 – Paper

Single-nucleosome mapping of histone modifications in s. cerevisiae

Chih Long Liu*, Tommy Kaplan*, Minkyu Kim, Stephen Buratowski, Stuart L. Schreiber, Nir Friedman

PLoS Biology (2005) **3**(10):e328

* These authors contributed equally to this work.

Single-Nucleosome Mapping of Histone Modifications in *S. cerevisiae*

Chih Long Liu[1,2☯], Tommy Kaplan[3,4☯], Minkyu Kim[5], Stephen Buratowski[5], Stuart L. Schreiber[2], Nir Friedman[3], Oliver J. Rando[1*]

1 Bauer Center for Genomics Research, Harvard University, Cambridge, Massachusetts, United States of America, **2** Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts, United States of America, **3** School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel, **4** Department of Molecular Genetics and Biotechnology, The Hebrew University, Jerusalem, Israel, **5** Department of Biological Chemistry and Molecular Pharmacology, Harvard University, Boston, Massachusetts, United States of America

Covalent modification of histone proteins plays a role in virtually every process on eukaryotic DNA, from transcription to DNA repair. Many different residues can be covalently modified, and it has been suggested that these modifications occur in a great number of independent, meaningful combinations. Published low-resolution microarray studies on the combinatorial complexity of histone modification patterns suffer from confounding effects caused by the averaging of modification levels over multiple nucleosomes. To overcome this problem, we used a high-resolution tiled microarray with single-nucleosome resolution to investigate the occurrence of combinations of 12 histone modifications on thousands of nucleosomes in actively growing *S. cerevisiae*. We found that histone modifications do not occur independently; there are roughly two groups of co-occurring modifications. One group of lysine acetylations shows a sharply defined domain of two hypo-acetylated nucleosomes, adjacent to the transcriptional start site, whose occurrence does not correlate with transcription levels. The other group consists of modifications occurring in gradients through the coding regions of genes in a pattern associated with transcription. We found no evidence for a deterministic code of many discrete states, but instead we saw blended, continuous patterns that distinguish nucleosomes at one location (e.g., promoter nucleosomes) from those at another location (e.g., over the 3′ ends of coding regions). These results are consistent with the idea of a simple, redundant histone code, in which multiple modifications share the same role.

Citation: Liu CL, Kaplan T, Kim M, Buratowski S, Schreiber SL, et al. (2005) Single-nucleosome mapping of histone modifications in *S. cerevisiae.* PLoS Biol 3(10): e328.

Introduction

Nucleosomes play many roles in transcriptional regulation, ranging from repression through occlusion of binding sites for transcription factors [1], to activation through spatial juxtaposition of transcription factor-binding sites [2]. There are two main ways in which cells modulate nucleosomal influences on gene expression. One way is through chromatin remodelling, using the energy of adenosine triphosphate hydrolysis to modulate nucleosomal structure, often resulting in changed nucleosomal location [3]. Alternatively, covalent histone modifications have many effects on transcription. Histone proteins have highly conserved tails, which are subject to multiple types of covalent modification, including acetylation, methylation, phosphorylation, ubiquitination, sumoylation, and adenosine-diphosphate ribosylation [4–9].

Histone acetylation has been the subject of decades of research, whereas histone methylation has come under intense scrutiny more recently. Lysine acetylation neutralizes lysine's positive charge, and can influence gene expression in at least two ways. Firstly, charge neutralization can affect contacts between the positively charged histone tail and negatively charged neighbouring molecules, such as adjacent linker DNA [10], or acidic patches on histones in nucleosomes [11]. Alternatively, acetyl-lysine is bound by the bromodomain, a protein domain found in many transcriptional regulators; thus, acetylation might affect recruitment of protein complexes [12]. Histone acetylation is rapidly reversible, and acetyl groups turn over rapidly in vivo, with half-lives on the order of minutes [13], allowing for rapid gene expression changes in response to signals [14]. Acetylation of histone lysines has been associated with both transcriptional activation and transcriptional repression [15–17]. The outcome of acetylation depends on which lysine is acetylated and the location of the modified nucleosome. A recent genome-scale study of histone acetylation in yeast revealed a complicated relationship between histone modification and transcriptional output [18].

Histone methylation has been best characterized by histone 3-lysine 4 (H3K4), wherein methylation is associated with active transcription in multiple organisms, ranging from *Saccharomyces cerevisiae* to mammals. Lysine can be mono-, di-, or tri-methylated, and none of these methylation states will alter lysine's positive charge (under conditions of standard lysine pKa and physiological pH). As a result, it is unlikely that charge–charge interactions are modulated by methylation, which appears instead to affect cellular processes

Received May 10, 2005; Accepted July 16, 2005; Published August 30, 2005
DOI: 10.1371/journal.pbio.0030328

Abbreviations: ARS, autonomously replicating sequence; bp, base pairs; CDS, coding sequences; ChIP, chromatin immunoprecipitation; TSS, transcription start site

Academic Editor: Peter Becker, Adolf Butenandt Institute, Germany

*To whom correspondence should be addressed. E-mail: orando@cgr.harvard.edu

☯These authors contributed equally to this work.

through binding of methyl-lysine–binding proteins. Indeed, methyl-lysine is bound by at least one domain type—the chromodomain [19,20]. In contrast to histone acetylation, histone methylation is long-lived. Although a histone-lysine demethylase (termed LSD1) was recently identified in metazoans. *S. cerevisiae* does not have a homolog of this protein. Even in metazoans, the proposed enzymatic mechanism allows for demethylation of mono- and di-methylated lysine, but not of tri-methylated lysine [21]. Whether or not enzymatic demethylation of tri-methyl-lysine occurs, and whatever other mechanisms allow for replacement of tri-methylated histones (such as histone replacement—[22]), in yeast, H3K4 tri-methylation is associated with active transcription. The histone tri-methylation persists for over an hour after transcription ceases, providing a memory of recent transcription [23].

The discovery of multiple modification types and modified residues suggested that different combinations of histone modifications might lead to distinctive transcriptional outcomes. According to the "histone code" hypothesis, "distinct histone modifications, on one or more tails, act sequentially or in combination to form a 'histone code' that is read by other proteins to bring about distinct downstream events" [6].

This hypothesis has been the subject of much debate, much of it concerning the requirements for histone modifications to form a "code" [4–9]. In this study, we focused on the combinatorial complexity of histone modification patterns. Insights into this complexity require an understanding of which combinations of modifications occur in vivo, and the functional consequences of these combinations. Mutagenesis of histone tails has demonstrated that not all combinations of histone modifications lead to distinct transcriptional states [24]. In addition, genome-wide localization studies of histone modifications in yeast, flies, and mammals have demonstrated that not all possible histone-modification patterns occur in vivo [18,25,26].

A major confounding effect in the interpretation of previous genome-wide studies of histone modifications in vivo is the low resolution of the measurements (~500–1,000 base pairs [bp]) relative to the size of the nucleosome (~146 bp). Thus, the measured ratio for a given spot represents an aggregate that is actually an average of information from several nucleosomes, which complicates analysis. Furthermore, in some studies, acetylation patterns at intergenic and coding regions were measured using different microarrays, precluding a common reference point. Finally, whole genomic DNA has typically been used as the reference DNA in these microarray studies, thereby confounding the measurements of histone modification with underlying variation in nucleosome density [27,28].

To overcome these limitations, we made use of a recently developed, high-density oligonucleotide microarray with ~20-bp resolution. We recently used this microarray to map nucleosome positions across almost half a megabase of the yeast genome [29]. In this study, we use this microarray to measure the levels of 12 different histone modifications in individual nucleosomes. We find that modifications do not occur independently of each other and that a small number of distinct combinations occur in vivo. Different modification patterns are enriched at specific locations in gene or promoter regions, and these patterns are predictive of the transcription level of the underlying gene. Sharp transitions in histone modifications mostly occur near the transcription start site (TSS). Together these results provide a simpler view of histone modification, and suggest that there is little combinatorial information encoded in the histone tails.

Results

High-Resolution Measurement of Histone Modifications Using Tiled Microarrays

Chromatin immunoprecipitation (ChIP) using modification-specific antibodies [30,31] was used to map histone modifications in actively growing yeast cultures. We used a standard ChIP protocol, with one major modification (Figure 1A). In our protocol, formaldehyde-fixed yeast were lysed gently by spheroplasting and osmotic lysis rather than by glass beads, and DNA was digested to mononucleosomes using micrococcal nuclease (rather than sheared to ~500 bp by sonication) (Figure S1). This allowed us to map modifications at nucleosomal resolution. We used antibodies specific to 12 individual modifications, including mono-, di-, and tri-methylation of histone H3K4, as well as acetylation of various lysines on all four histones. Immunoprecipitated DNA was isolated, linearly amplified [32], and labelled with Cy5 fluorescent dye, while mononucleosomal DNA treated under identical conditions was used as the "input" and labelled with Cy3. This choice of input served to control for nucleosomal occupancy differences (to prevent highly modified, low-occupancy nucleosomes from appearing to be poorly modified nucleosomes), as it has been shown that nucleosomes are not always present in every cell in a population [33,34]. Mixtures were hybridized to a tiled microarray covering half a megabase of yeast genomic sequence, including almost all of Chromosome III as well as 230 additional 1-kb promoter regions [29]. This represents approximately 4% of the yeast genome, and includes a total of 356 promoter regions. Finally, to measure active transcription (while avoiding effects of mRNA instability that influence mRNA abundance measurements), we also immunoprecipitated DNA associated with RNA polymerase II (this DNA was sheared by sonication rather than cut with micrococcal nuclease) [35].

A Chromosomal View of Histone Modifications

The resulting data provide a rich view of histone modification over half a megabase of yeast sequence, demonstrating several prominent features (Figure 1B shows a sample stretch). First, histone modifications generally occur in broad domains, and there are few examples of nucleosomes whose modification pattern was significantly different from that of their adjacent nucleosomes. This was not due to limitations in the experimental technique, as we did find multiple examples of punctate nucleosomes that occurred in expected locations (see below). Second, modifications were generally homogeneous for all the probes within a given nucleosome. Third, correlations could be observed between a nucleosome's position relative to coding regions and its modification pattern. For example, most of the open reading frames shown in Figure 1B exhibit a striking pattern of histone H3K4 methylation, with tri-methylation occurring at the 5′ end of the coding region, shifting to di-methylation, and then to mono-methylation. This pattern is clear over

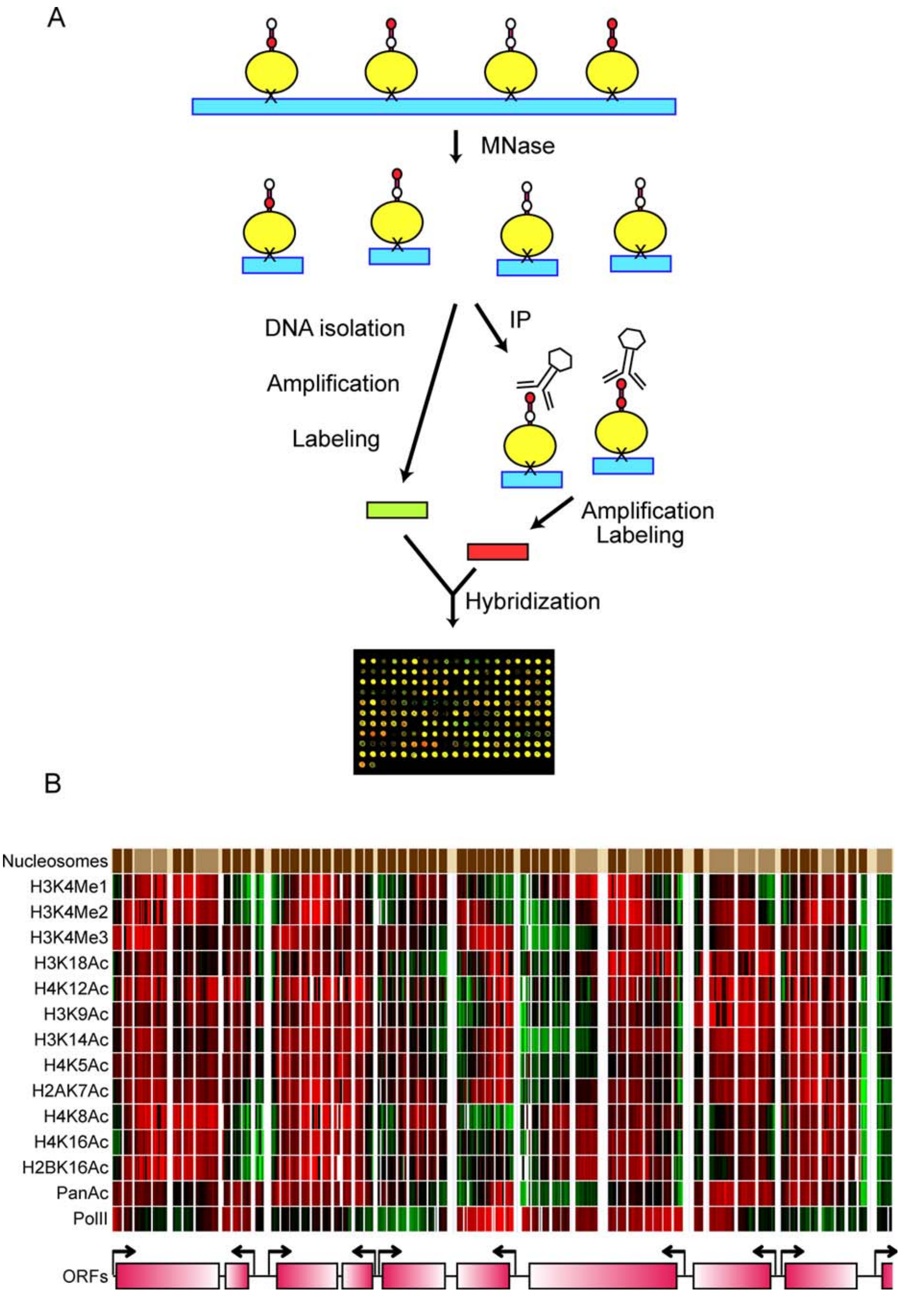


Figure 1. Overview

(A) Nucleosomes are first cross-linked to DNA using formaldehyde. Cross-linked chromatin is digested to mononucleosomes with micrococcal nuclease. Mononucleosomal digests are immunoprecipitated using an antibody specific to a particular histone modification, and immunoprecipitated DNA is isolated and labelled with Cy5. DNA is also isolated from the same nuclease titration step prior to immunoprecipitation, labelled with Cy3, and mixed with Cy5-labeled immunoprecipitated DNA. Labelled DNA is then hybridized to a tiled microarray covering half a megabase of yeast genome.

(B) Example of raw data. Data are shown for all modifications tested, along with PolII data. Red (green) indicates enrichment (depletion), while grey indicates missing data. Data from probes found in linker regions are not shown. Each row represents median data from multiple replicates with one antibody, as indicated (PanAc refers to a nonspecific antibody to acetyl-lysine, which we used to measure bulk acetylation). "Nucleosomes" shows positions of nucleosomes previously described [29], with dark brown for well-positioned nucleosomes, very light brown for linkers, and intermediate brown for delocalized nucleosomes. "ORFs" shows locations of annotated genes. Data shown are for Chromosome III coordinates 58,900 to 72,100.

DOI: 10.1371/journal.pbio.0030328.g001

most expressed open reading frames on Chromosome III, and is consistent with reports that Set1 association with RNA polymerase is responsible for methylation of this lysine [23,36]. Finally, we noticed broad domains of low acetylation occurring over heterochromatic regions on our array—subtelomeric sequences and the silent mating type loci [37] (Figure S2).

Coupling of Modifications to Organization of Transcriptional Units

To analyze the relationship of different modifications to the underlying sequence, we aligned all genes (and their promoters) by their start codon. For example, Figure 2A shows data for histone H4K16 acetylation on aligned genes

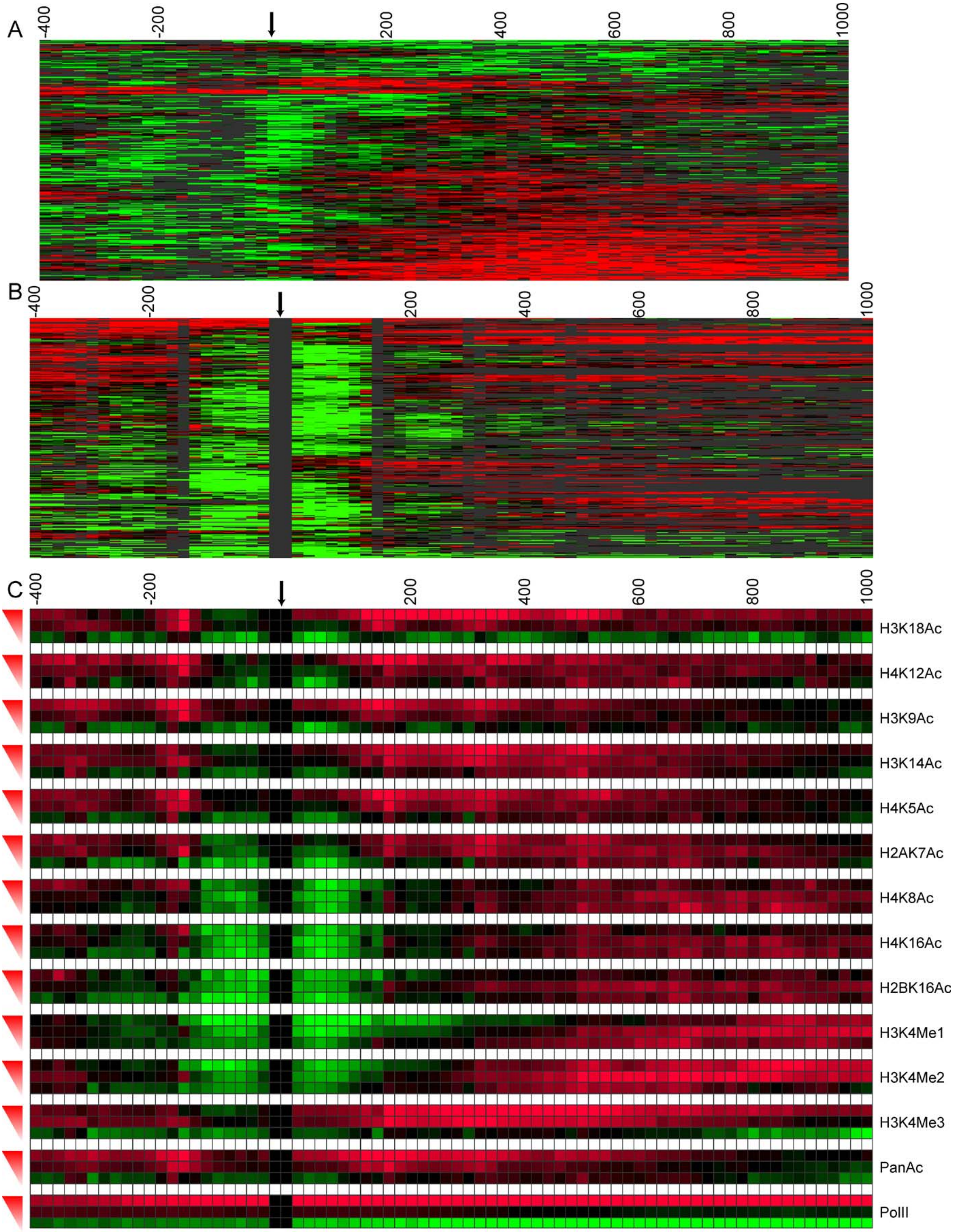

A
-400
-200
200
400
600
800
1000
B
-400
-200
200
400
600
800
1000
C
-400
-200
200
400
600
800
1000
H3K18Ac
H4K12Ac
H3K9Ac
H3K14Ac
H4K5Ac
H2AK7Ac
H4K8Ac
H4K16Ac
H2BK16Ac
H3K4Me1
H3K4Me2
H3K4Me3
PanAc
PolII

Figure 2. Broad Patterns of Histone Modifications

(A) H4K16Ac aligned by ATG. In this representation, the horizontal axis represents location relative to the downstream gene's start codon, and each horizontal line represents one PolII-driven gene. Each cell in the resulting matrix corresponds to the acetylation level at a given microarray probe for one tail position. Red (green) cells mark hyper-acetylated (hypo-acetylated) probes. Non-nucleosomal probes are blackened. We clustered the promoters using a probabilistic agglomerative clustering algorithm (see Materials and Methods). Arrow indicates annotated ATG.

(B) H4K16 aligned by transcriptional start site, as in (A), except that arrow indicates TSS (identified in [29]) and data before and after the TSS are aligned by the first nucleosome in that direction.

(C) Relationship of histone modification patterns to transcription level. Genes were split into three groups based on PolII enrichment, and averaged data for these groups are shown as indicated, aligned as in (B). Transcription level is indicated by red triangles to the left of each set of three rows.

DOI: 10.1371/journal.pbio.0030328.g002

that were clustered to highlight patterns (see Materials and Methods). Clearly notable in this representation is a hypo-acetylated domain adjacent to most start codons. We have recently discovered that TSSs are found in long nucleosome-free regions [29]. By aligning genes by the location of the first nucleosome following the TSS, a clear domain of two hypo-acetylated nucleosomes can be observed at most PolII promoters (Figure 2B). This alignment, therefore, provides a highly informative view of the relationship of histone modifications to the underlying structure of the genome (see Figure S3 for the remaining modifications).

To explore the relationship of these modifications to transcription, we separated genes into "bins" of varying transcriptional activity (see Materials and Methods) and averaged the enrichment data for all aligned genes in each bin (Figures 2C and S4). Several previously identified features of yeast chromatin are apparent. First, histone H3K4 methylation enrichment correlates with transcription levels, and occurs in a 5′ to 3′ gradient (as also seen in Figure 1B) with tri-methyl enrichment at the 5′ end of genes, shifting to di-methyl and then mono-methyl. Histone H3K4 is methylated by Set1, which is associated with elongating RNA polymerase [23,36], and, as noted above, this gradient presumably reflects the kinetics of dissociation of Set1 from the polymerase, convoluted with the ensemble-average location of polymerase. Second, we reproduced previous observations that histone H3K9/K14 acetylation is enriched over the 5′ ends of coding regions [26,38].

Figure 2C also reveals novel locations of particular histone modification patterns. In particular, the two-nucleosome hypo-acetylation domain described above for H4K16 acetylation is surprisingly general, and a nearly identical pattern is also seen for acetylation of H4K8 and of H2B K16 (Figures S3 and 2C). This hypo-acetyl domain does not correlate with transcription levels (as measured by either PolII occupancy or by mRNA abundance [Figures 2C and S4]). Also, the acetylation of these residues at the middle and 3′ ends of coding regions is either uncorrelated (H2BK16) or anticorrelated (H4K8 and K16) with transcription (Figure 2C). We will therefore refer to this group of modifications as the *transcription-independent* modifications, for convenience (and to emphasize the stereotyped promoter-deacetyl domain). A two-nucleosome hypo-acetylation domain is also present at a smaller subset of promoters for the remaining acetylation states, and is generally found preferentially in poorly expressed genes (Figures S3 and 2C). However, the acetylation of these lysines is found at the 5′ end of coding regions, whereas acetylation of the transcription-independent group is largely excluded from 5′ coding regions. We will refer to this 5′-directed group of modifications as the *transcription-dependent* modifications. Acetylation of H2A K7 is an interesting case, as its pattern appears to be a mixture of the two types of patterns described. However, we have recently found that the H2A isoform Htz1 is enriched in a pattern that dramatically parallels the hypo-acetylation domain observed for the transcription-independent modifications (unpublished data), so H2A is expected to be depleted in this region. This, coupled with the 5′-enrichment of acetylation seen for H2A K7, in highly transcribed genes, leads us to include this modification in the transcription-dependent group.

Low Dimensionality of Nucleosome Modification Patterns

The analysis presented above is highly informative, but is based on aggregated data for many promoters, and thus may obscure interesting underlying phenomena. A more informative approach would be to examine the distinct modification patterns at individual nucleosomes. We defined the modification pattern of each nucleosome as the median hybridization value, for each measured antibody, of the probes associated with the nucleosome (usually between six and 15 probes; see Materials and Methods). In addition, we classified nucleosomes according to their positions relative to genome annotations (Figure 3A; see Materials and Methods). We used nine annotation categories that represent nucleosomes in promoter regions, transcribed regions, and other regions (tRNA genes and autonomously replicating sequences (ARSs). These classifications are discussed further below.

Nucleosomes were clustered by modification pattern, using a probabilistic hierarchical agglomerative clustering procedure (see Materials and Methods). As is readily apparent from this clustering (Figure 3B), histone modification patterns span the full possible range of overall modification level, from hypo-acetylated to hyper-acetylated. Nevertheless, a striking aspect of this clustering is the limited range of observed modification patterns. Visual inspection suggests that, as previously noted [18], histone modifications are not independent of each other. Indeed, the matrix of correlations between the 12 modifications shows that there are two groups of strongly correlated acetylations (Figure 3C).

To better understand the effective number of degrees of freedom among the 12 dimensions available, we performed a principal component analysis (see Materials and Methods). Principal component analysis is a technique used to transform a large number of possibly correlated variables to a smaller number of uncorrelated variables, and thereby identify the number of independent dimensions in a dataset. As suggested by the observation above, 81% of the variance in histone modification patterns is captured by the first two principal components (Figure 3D). Moreover, if we examine only the nine acetylations, we can explain 90% of the variance using two components (unpublished data). The first principal component corresponds to overall level of histone modification (Figure S5). The second principal component

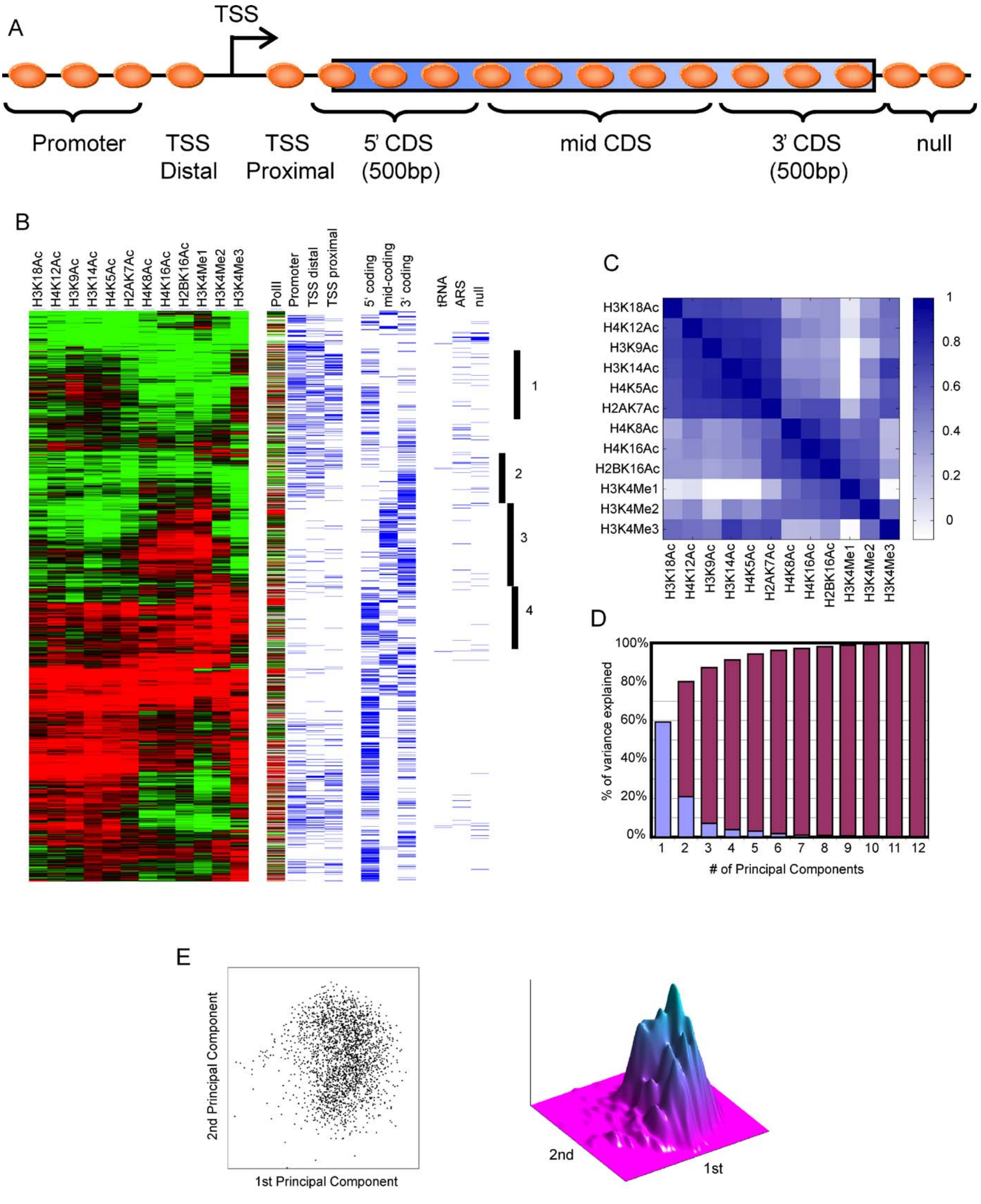


Figure 3. Nucleosome Modification Patterns

(A) Schematic of annotation scheme for nucleosomes based on their position relative to transcribed units. Intergenic nucleosomes were assigned to the following categories: promoter region (anything upstream of a coding region), nucleosome immediately upstream to the TSS ("distal"), and the nucleosome immediately downstream of the TSS ("proximal"). Transcribed regions were separated into 5′, middle, and 3′ CDSs. Finally, to capture features of chromatin not associated with PolII genes, we independently classified nucleosomes associated with ARS sequences, tRNA genes, and Null (any other intergenic region).

(B) Hierarchical clustering of 2,288 nucleosomes. Left panel: each row corresponds to a single nucleosome, and each column to a particular modification. Red (green) denotes hyper-acetylation (hypo-acetylation) in the first nine columns and relative level of methylation in the last three columns. Rows are sorted according to the dendogram built during clustering. PolII shows the PolII occupancy of the gene associated with the nucleosome in question. Right panel: each row corresponds to a nucleosome (matching the left panel), and each column corresponds to an annotation of the nucleosome according to the scheme of (A). A blue cell denotes a positive annotation of the nucleosome with the appropriate column label. Numbers indicate examples of clusters, as follows: (1) nucleosomes enriched for H3K9Ac, H3K14Ac, and H3K4Me3 that are mostly upstream of transcribed regions; (2) strongly hypo-acetylated nucleosomes, mostly at upstream regions or 3′ of coding regions; (3) nucleosomes acetylated at H4K8 and K16, and H2B K16 that are almost exclusively at the middle and 3′-ends of coding regions; and (4) hyper-acetylated and methylated nucleosomes that are mostly found at the 5′-end of coding regions.
(C) The Pearson correlations of the 12 modification levels between different probes show that there are two tightly correlated groups of acetylations at specific residues. The first group consists of H2A K7; H3K9, K14, and K18; and H4K5 and K12. The second group consists of H2B K16; and H4K8 and K16. Mono- and di-methylation of H3K4 are correlated with the second group, while tri-methylation of H3K4 is correlated with the first group.
(D) The percent of variance captured by using different number of components. The *x*-axis denotes the number of components, and the *y*-axis denotes the percent of the variance in the data explained by each components (blue bars) as well as the cumulative percentage explained (red bars).
(E) Representation of all nucleosomes in two-dimensional modification space. In the left panel, each point represents a nucleosome plotted according to the relative level of the first principal component (*x*-axis) and second principal component (*y*-axis) for the modification pattern. The right panel is a three-dimensional plot showing density of points along the plane.
DOI: 10.1371/journal.pbio.0030328.g003

corresponds to the relative levels of the two groups of histone modifications—the transcription-associated modifications that occur in 5′ to 3′ gradients over coding regions, and the group of acetylations characterized by short hypo-acetyl domains surrounding TSS (Figure S5). By projecting each nucleosome to a point in the plane spanned by the first two principal components (Figure 3E), we can visualize the range of observed modifications. There is a large region of allowable modifications that is spanned continuously by different nucleosomes. These results suggest that, at the level of cell populations, there are no discrete states for nucleosome modifications. Instead, nucleosome modification patterns occur continuously over a large range of possible space, though this two-dimensional space is dramatically simplified compared to the 12 dimensions available. In other words, nucleosomes have continuous variation, both in the total level of acetylation, and in the relative ratio of the two groups of modifications, but they do not show much complexity beyond these two axes.

Specific Chromosomal Locations Are Associated with Characteristic Histone Modifications

Notable in Figure 3B is an association of particular modification patterns with specific genomic locations. For example, Cluster 2 consists of hypo-acetylated nucleosomes that are predominantly located within promoter regions and at the 3′ ends of coding regions. We systematically explored these correlations by testing the modification data for statistically significant, location-specific differences in the levels of each modification type (Figure 4A). For example, promoter nucleosomes are globally hypo-acetylated in residues H2A K7 (presumably due to the enrichment of Htz1), H2B K16, and H4K8 and K16 (and, to a lesser extent, H3K18), and are depleted of mono- and di-methylated H3K4. Nucleosomes at 5′ ends of coding regions are enriched for H3K4Me3, as well as H3K18Ac, H4K12Ac, H3K9Ac, H3K14Ac, H4K5Ac, and H2AK7Ac. When we examine the modification patterns of individual nucleosomes in the two-dimensional principal component plot, we can clearly distinguish nucleosomes in promoter regions from those in transcribed regions (Figure 4B). Moreover, of the nucleosomes in transcribed regions, we can distinguish among nucleosomes in the 5′ end, the middle, and the 3′ end of the transcribed region (Figures 4C and S6).

These results show that specific genomic regions are characterized by distinct modification patterns, with little overlap in modification types between the different regions. We conclude that the histone modification patterns are highly informative about the location of nucleosomes along the chromosome, and suggest that, in yeast, nucleosome modification patterns, like nucleosome positioning, exhibit local variation around a basic stereotype that is determined by the chromosomal location.

Variation in Modifications Occurring over Transcribed Regions is Predictive of Transcription Levels

While nucleosomes at different locations are associated with statistically different modification patterns, the correlations are imperfect, as a given nucleosome modification pattern can clearly be found in multiple locations (Figure 4B and 4C). This imperfect association might be due to differences in expression level of the coding regions examined. We therefore separated nucleosome locations (5′ coding, etc.) into bins according to the PolII activity level of the associated transcription unit. Figure 5A shows the modification pattern of each of five nucleosomes (defined by position) for highly PolII-enriched genes, while Figure 5B shows this pattern for PolII-depleted genes. This view emphasizes both the distinction between nucleosomes at various genomic locations (as seen in aggregate in Figure 4) and the transcription-associated variation in the modification pattern at a given location. Figure 5C shows a cartoon of the chromatin structure of an arbitrary yeast gene.

To further explore the relationship between transcription activity and modification pattern at a given location, we tested each location for modifications that were significantly associated with high or low transcription. For example, we consider the nucleosomes near the 5′ ends of those genes with extreme levels of PolII enrichment or depletion (Figure 6A). Consistent with results shown in Figures 2C and 5A and 5B, we see that levels of mono- and tri-methylation of H3K4, as well as the acetylation level of H3K9, H3K14, H2A K7, H4K5, and H4K12 have significant differences between these two classes of 5′ coding region nucleosomes ($p < 0.01$ using *t*-test). We trained a classification method that examines these modifications and predicts whether the nucleosome is part of an expressed coding region or not. We evaluated this classifier using leave-one-out cross-validation (see Materials and Methods) to estimate its accuracy on unseen examples. This evaluation shows that the classifier is correct on 75.4%

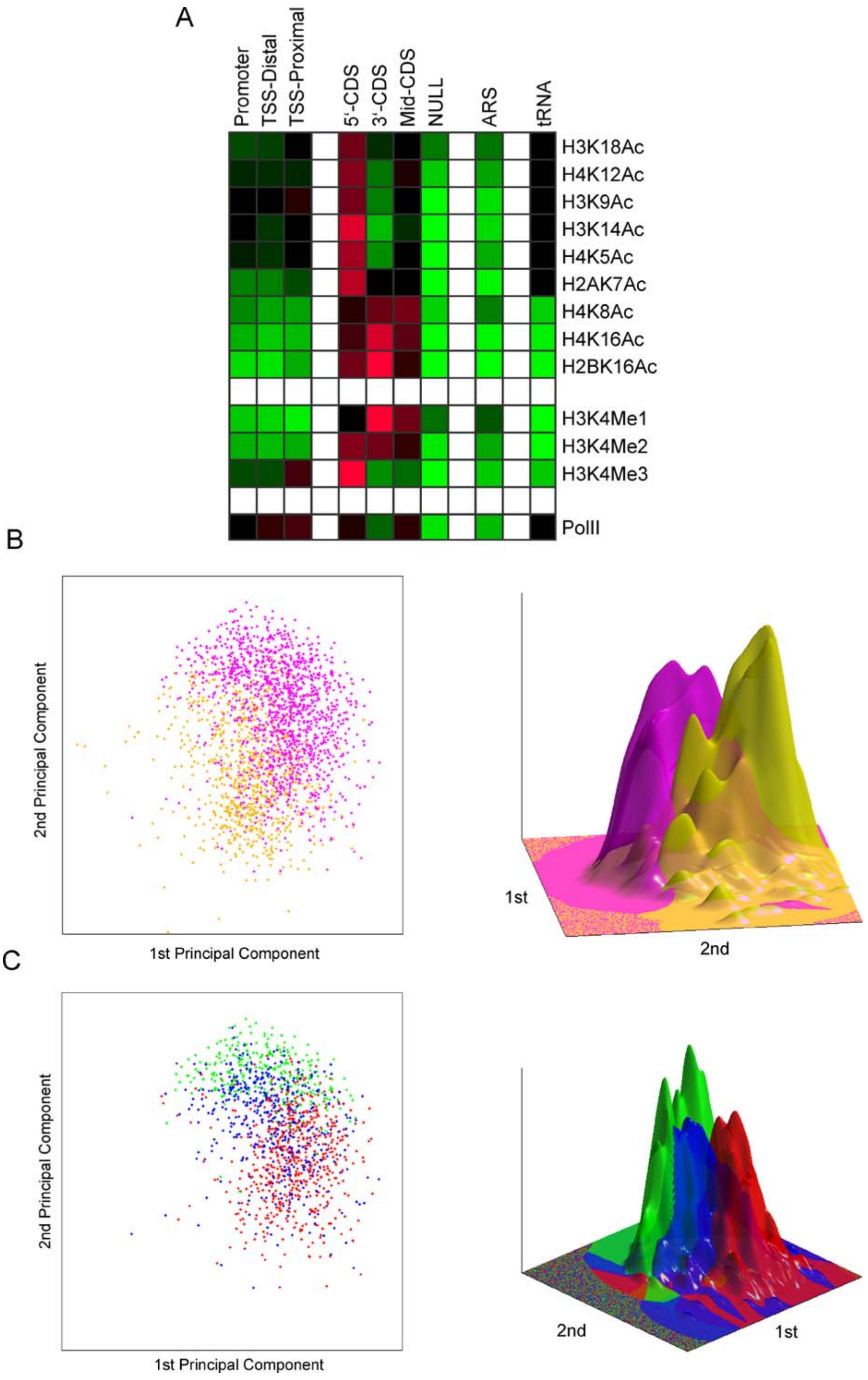


Figure 4. Nucleosome Modifications Relate to Nucleosome Position

(A) Analysis of differential modification for each class of nucleosomes. Rows correspond to specific modifications, and columns correspond to genomic locations. Each cell is coloured by the average modification level of nucleosomes with this annotation. Non-significant (using false discovery rate of 95% on *t*-test *p*-values) cells are blackened.

(B) Promoter nucleosomes (orange) significantly differ from coding region nucleosomes (pink) in their histone modifications pattern. The left panel shows the two types of nucleosomes as points in the plane, where the *x*-axis represents the level of the first principal component, and the *y*-axis represents the second principal component. The right panel shows the density within each class.

(C) Distinction between nucleosomes in transcribed regions. Colours denote 5′-end (red), middle (green), and 3′- end (blue) nucleosomes. Visualization is as described in (B).

DOI: 10.1371/journal.pbio.0030328.g004

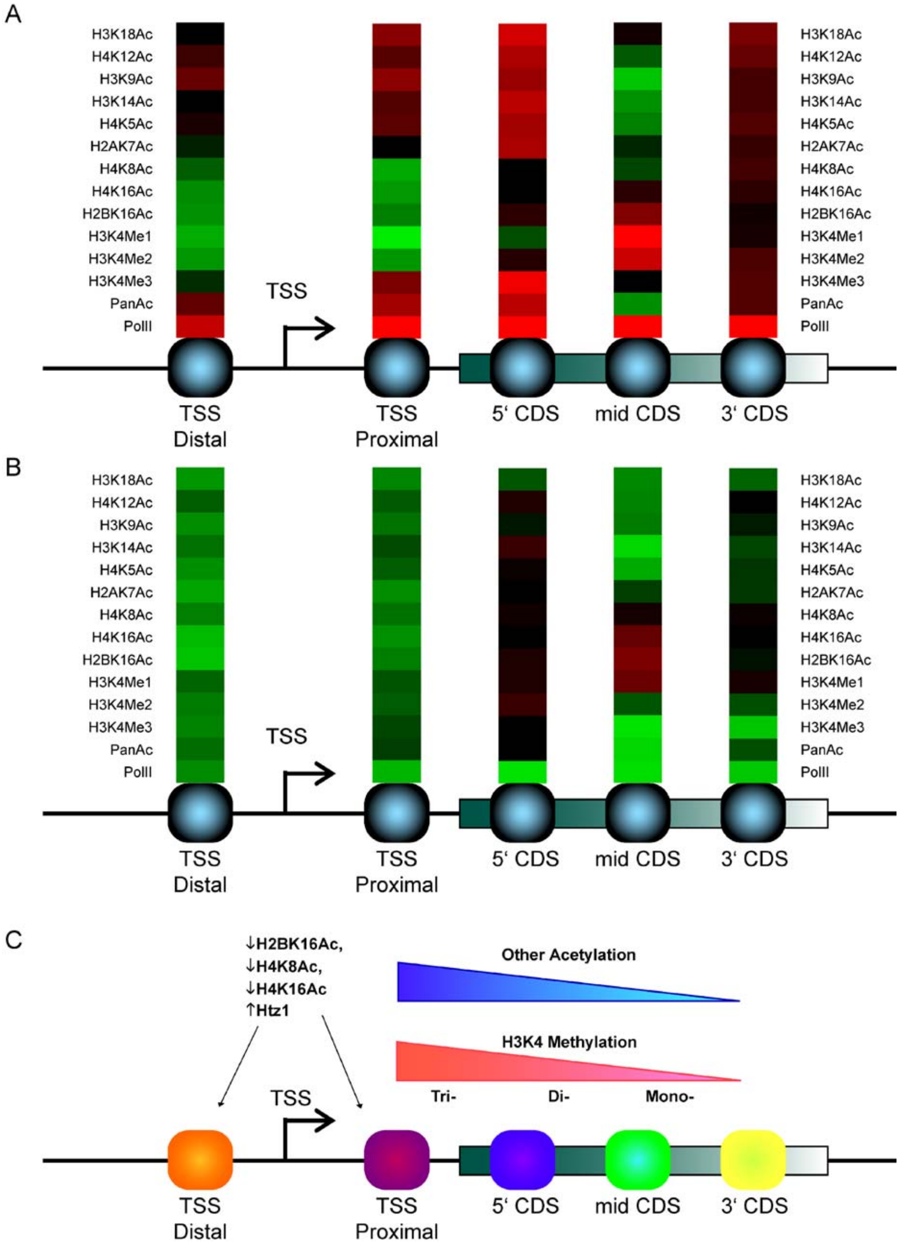


Figure 5. Nucleosome Modifications Partitioned by Location and by Transcription Level

(A) Modification patterns of nucleosomes associated with actively transcribed genes. Genes with high levels of PolII occupancy were grouped, and the modification data for the indicated nucleosome types were averaged.

(B) Modification patterns of nucleosomes associated with poorly transcribed genes, grouped as in (A), except that genes with low levels of PolII were selected.

(C) Schematic view of yeast chromatin architecture. Cartoon view showing chromatin structure of an arbitrary yeast gene. Yeast genes are typically characterized by an upstream nucleosome-free region, which serves as the transcriptional start site [29]. Surrounding this nucleosome-free region are two nucleosomes that exhibit low levels of acetylation at H2BK16, H4K8, and H4K16, and that carry Htz1 in place of the canonical H2A (unpublished data). The remaining acetylations occur in a gradient from 5′ to 3′ over actively transcribed genes. Similarly, actively transcribed genes exhibit a gradient of H3K4 methylation, with trimethylation occurring at the 5′- ends of genes, and di- and mono-methylation occurring over the middle of the coding region. Nucleosomes are coloured to emphasize the different average modification patterns at each indicated location.

DOI: 10.1371/journal.pbio.0030328.g005

of the nucleosomes in the training set (compared to 60.1% when nucleosomes labels are randomly permuted; $p <$ 0.0001). Thus, although expression values are not perfectly encoded by histone modifications, they are clearly reflected in them. We see a similar pattern if we examine nucleosomes in the middle of coding regions (Figure S7). In this case the accuracy is 82.7% (compared to 61.3% by chance; $p <$ 0.0001). Notably, the set of significant modifications in this case is different, and in fact two of the transcription-independent modifications, H4K8 and K16, are both slightly anticorrelated with transcription here.

These results indicate that over coding regions, variation in histone modification patterns is associated with transcription level. For example, the transcription-associated modifications are globally enriched at the 5′ ends of genes, and the level of these modifications is correlated with transcription level. To

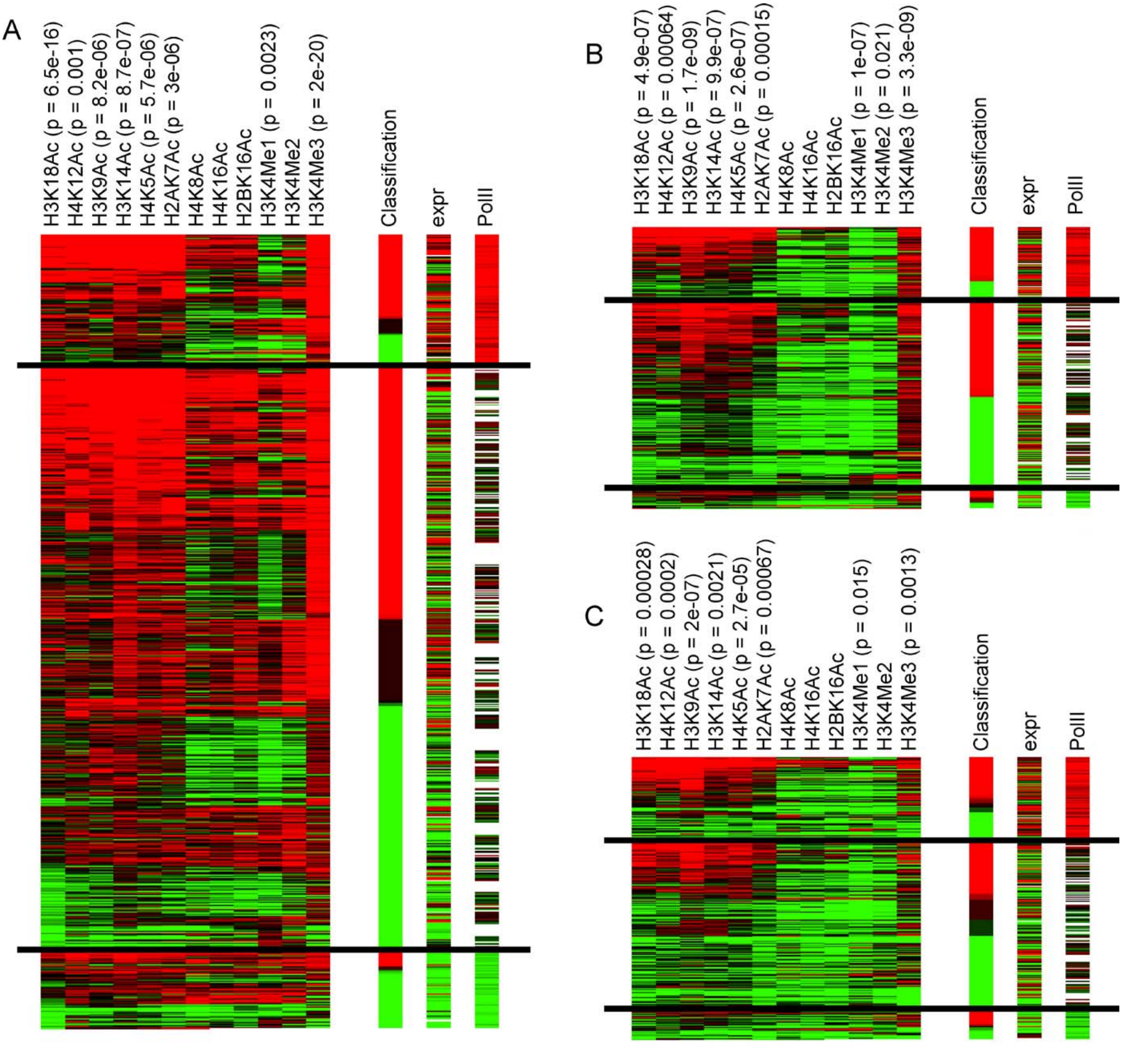


Figure 6. Nucleosome Modifications Relate to Transcription Level

(A) Classification plot of nucleosomes in 5′-coding regions according to PolII occupancy. A classifier was trained to distinguish between nucleosomes with high and low PolII occupancy, and evaluated using leave-one-out cross-validation. Each row corresponds to one nucleosome. Nucleosomes are split into three groups associated with genes corresponding to high, intermediate, and low PolII occupancy level (from top to bottom, respectively). The left 12 columns denote modification patterns of each nucleosome. Modifications with significant differences between high and low nucleosomes are marked with the *p*-value determined by *t*-test. Colours denote relative acetylation/methylation levels. The rightmost three columns correspond to the classifier's prediction of transcription, the expression level (mRNA abundance; see Materials and Methods) and the PolII occupancy of genes. The average accuracy of random classification was 60.71%, with a standard deviation of 4.3%. Accuracy of classifier was 75.38% ($p < 0.0001$).

(B) Classification plot of TSS proximal nucleosomes, labelled as in (A). The average accuracy of random classification was 62.45%, with a standard deviation of 4.75%. Accuracy of classifier was 72.8% ($p = 0.0004$).

(C) Classification plot of TSS distal nucleosomes; as in (A). The average accuracy of random classification was 65.79%, with a standard deviation of 4.22%. Accuracy of classifier was 58.4% ($p = 0.9333$).

DOI: 10.1371/journal.pbio.0030328.g006

explore whether these results hold true for nucleosomes that are not found over transcribed regions, and to thereby test the idea that upstream histone modifications control gene expression, we repeated the classification analysis for nucleosomes surrounding the TSS (Figure 6B and 6C), which are modified in similar ways (Figure 4A) with the exception that the gene-proximal nucleosome is associated with DNA passaged by RNA polymerase, while the gene-distal nucleosome is not. Here, we found that the gene-proximal nucleosome indeed carries information about transcription level—a classification method tested using this nucleosome correctly identified 72.8% of gene expression patterns (as

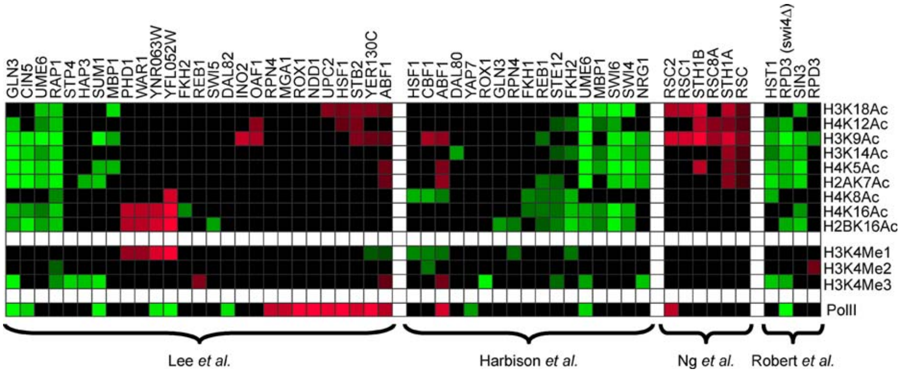


Figure 7. Histone Modifiers

Analysis of differential modification of nucleosomes associated with various transcriptional regulators. Promoter nucleosomes located near binding sites of the indicated factors were tested for enrichment of all modifications relative to the overall promoter modification pattern. Each cell is coloured by the average modification level of nucleosomes with this annotation. Non-significant cells (using false discovery rate of 95% on *t*-test *p*-values) are blackened. Localization data are taken from the indicated studies [39–42].

DOI: 10.1371/journal.pbio.0030328.g007

compared with 62.4% by chance; $p = 0.0004$). In contrast, the gene-distal nucleosome, which is not subjected to the passage of RNA polymerase and associated modifying enzymes, fails to accurately classify transcription levels (58.4%, as compared with 65.7% expected by chance), demonstrating that modification patterns associated with transcribed regions provide a much better predictor of transcription levels than do upstream modification patterns.

Modifications Associated with Transcriptional Regulators

The observed modifications at the two TSS nucleosomes might be either a prerequisite for PolII recruitment or a consequence of this step. Since we measure modification in a single condition, we cannot directly resolve this question. However, we can gain additional insight by examining nucleosomes in promoters reported to be bound by specific chromatin remodelers or by specific transcription factors. Using the results of several recent ChIP studies [39–41], we compiled a set of target promoters for each factor (see Materials and Methods). We then tested for distinct patterns in the promoter nucleosomes. In addition, we analyzed nucleosomes around putative transcription factor binding sites [42] (see Materials and Methods). Our results highlight specific factors that are significantly associated with specific modifications (Figure 7). For instance, we see that promoters of genes bound by the repressor Ume6 are significantly hypo-acetylated at most positions. This finding correlates with previous observations demonstrating recruitment of the HDAC Rpd3 by Ume6 [43,44]. Another interesting example is the significant hyper-acetylation of several positions among the targets of the Rsc remodeling complex. These include H3K9 and, to a lesser extent, H4K12, H3K14, and H4K5. Recently, mutants in the Rsc complex were shown to interact genetically with K14 mutations, a finding supported by binding of the complex to K14-acetylated H3-tail peptides [45].

Modification Boundaries Occur Near Transcriptional Start Sites

The availability of histone modification data at single nucleosome resolution allows analysis of the extent to which modification patterns occur discretely or in broad domains. As noted above and previously reported [44], histones can be deacetylated in a localized manner. However, visual inspection reveals that at locations farther away from the TSS, most histone modifications occur in broad domains. To further investigate this, we searched for sharp boundaries to histone modification domains by identifying pairs of nucleosomes between which a dramatic change occurs (increase or decrease of two standard deviations at one of the tail positions). We found ~100 boundaries for each modification (from 82 to 108). We then examined the locations of these boundaries, finding that most were located adjacent to TSSs. For example, boundaries for modifications associated with transcription, such as H3K4 tri-methyl, occurred across the TSS. This is visualized in Figure 8A, a scatterplot of K4 tri-methylation for adjacent nucleosomes (*x*-axis shows tri-methylation for nucleosome N, *y*-axis shows tri-methylation of N-1). The majority of nucleosomes show high correlation for this modification between adjacent nucleosomes, though there are two small groups of anticorrelated nucleosomes, indicating methylation boundaries. Pairs of nucleosomes that fall to either side of the TSS were plotted separately (grouped according to which strand the gene falls on), showing that most of the K4 tri-methyl boundaries occur at the TSSs, as expected.

We also examined "punctate" nucleosomes—those differing significantly in modification type from the two nucleosomes to either side. We found 44 nucleosomes with a punctate pattern of at least one of the 12 modifications in this study. Examples of punctate nucleosome are shown in Figure 8B and 8C. Most nucleosomes that exhibit this characteristic are found upstream of the TSS. In many cases, this is clearly due to the location of the nucleosome between two TSSs, leading to a single nucleosome exhibiting no transcription-associated modifications, surrounded by nucleosomes with the characteristic transcriptional modifications.

Discussion

Profiling Histone Modification at the Mononucleosome Level

We have mapped, at single-nucleosome resolution, 12 histone modifications in actively dividing cultures of *S. cerevisiae*. This, along with the translational positioning of nucleosomes described previously [29] and location studies

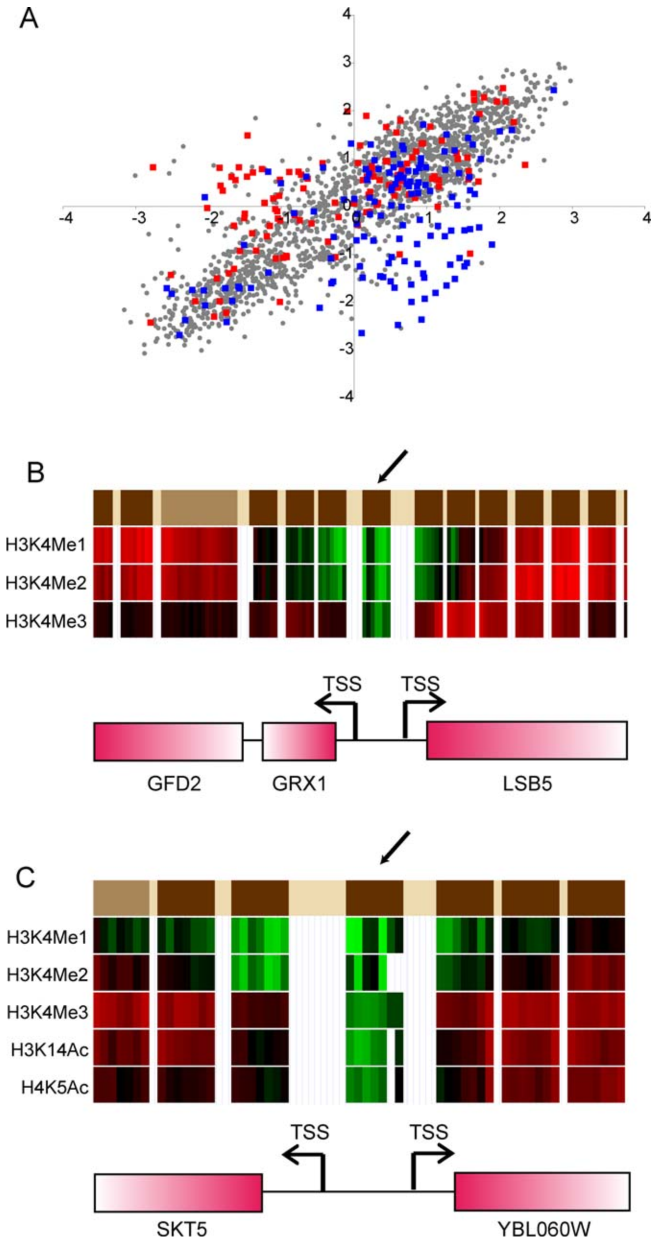


Figure 8. Modification Boundaries

(A) H3K4Me3 boundaries occur across TSSs. The *x*-axis represents the level of H3K4Me3 for a given nucleosome, and the *y*-axis represents the level of this modification for the preceding nucleosome. Pairs of nucleosomes flanking the TSS for a gene on the W strand are plotted as blue squares, and pairs flanking TSSs for genes on the C strand are plotted as red squares. Remaining nucleosome pairs are plotted as grey circles.

(B) Example of a punctate nucleosome. Histone modification plotted as in Figure 1B for a subset of histone modifications. Arrow indicates a nucleosome whose modification pattern differs significantly for H3K4Me3 from nucleosomes to either side. Gene names are as labelled.

(C) Example of a punctate nucleosome, labelled as in (B).

DOI: 10.1371/journal.pbio.0030328.g008

on the H2A isoform Htz1 (unpublished data), provides a draft sequence (see below) of the primary structure of half a megabase of yeast chromatin. We wish to stress the importance of the high resolution of our method for deconvoluting the results of previous studies on histone modification. The use of ~1-kb intergenic and coding probes in standard microarray studies reports on mixtures of multiple nucleosomes. For example, we show that the two nucleosomes immediately adjacent to the TSS are generally deacetylated at H4K16, whereas surrounding nucleosomes are often highly acetylated (Figure 2B). As a result, the acetylation level measured in standard microarray studies will depend on the length of the 5′ untranslated region (which is especially confounding, as this correlates with functional classifications of the encoded genes [46]); the length of the entire intergenic region probed; and the nature of the intergenic region (divergent or parallel genes), as the deacetyl signals from the TSS will be diluted by these additional nucleosomes in a complicated way. Furthermore, the ~300–500-bp standard shear size used in microarray studies results in some sampling of additional nearby nucleosomes outside the borders of the microarray spot. Our methodology eliminates all these confounding variables and also controls for local variation in nucleosome density, thus dramatically simplifying modification mapping.

We note, however, that our study is subject to the same issues with antibody specificity that remain a crucial limitation of ChIP studies—the epitope accuracy of any ChIP study is determined by the specificity of the antibodies used. We used the state-of-the-art in antibodies (see Materials and Methods), but improvements in antibody specificity may improve the fidelity of these experiments. In addition, ensemble measurements such as those presented here necessarily provide population averages, and we cannot rule out the possibility that small subpopulations of cells in different phases of the cell cycle, or in different epigenetic states, might be characterized by modification patterns that are obscured in the population average. Finally, this study does not provide a complete sequence of chromatin's primary structure in our tiled region. A complete view of the primary structure requires the addition of all additional modifications, including core domain modifications, and, ideally, the conformations of the nucleosomes studied.

Histone Tail Modifications Occur in Two Groups that Vary Quantitatively

This mapping has allowed us to investigate combinatorial questions raised by the framing of histone modifications as a "code." Most importantly, we have shown that many histone modifications are highly correlated with one another, resulting in few discrete histone modification patterns. However, we cannot say whether these modifications occur in the same nucleosome or whether the correlations are due to a mixture of partially modified nucleosomes at a given location. Some modified residues may be correlated because histone-modifying enzymes are not strongly residue-specific [8,47], whereas other correlations may be due to histone-modifying enzymes that are either recruited to chromatin by association with other types of modification, or preferentially act on tails carrying another modification [48–50]. Still other modifications may be correlated because the relevant modifying enzymes may be targeted by association with similar complexes, such as RNA polymerase [23,51]. These correlations suggest a high level of redundancy in yeast histone modification, implying that the code is extremely simple, carrying only a tiny fraction of the maximum possible amount of information. Indeed, as principal component analysis shows, we can compress the 12-dimensional space of possible modification patterns onto two main axes, with only a minor loss of accuracy.

This raises the important question of why so many different modifications occur in the cell, yet such a small subset of combinations is used. We suggest only a few possible answers. First, the loss of a positive charge that occurs with lysine acetylation should reduce the free energy of interaction with a negative charge by approximately 1–3 kcal/mol. Thus, loss of multiple positive charges could lead to much greater free energy changes in an interaction, and to a much more pronounced change in interactions than would be caused by a single acetylation. Furthermore, we note that at any given nucleosome location the quantitative level of acetylation varies, allowing for the possibility of "rheostat"-like control of transcription levels. This is consistent with recent mutagenesis studies showing that transcriptional response to H4K→R mutations is largely continuous and analogue, rather than discrete and digital [24]. Second, it is possible that multiple modifications occur together in order to cause several distinct required events to occur, whether they be co-occurring structural changes in the nucleosome or the 30-nm fibre, or recruitment of protein complexes that function together. This has been observed at the human interferon-β promoter, wherein activation of the promoter causes Gcn5-dependent acetylation of H3K9/14 and H4K8, whose acetylation recruits TFIID and hSWI/SNF, respectively [52]. If these protein complexes tend to function together, then the recruiting modifications will be correlated. Third, if modifications that occur together at steady-state do not occur simultaneously, but rather in a temporal cascade [6], this enables the possibility of complex signal filtering behaviour. For example, if one histone acetylase were to acetylate a single lysine, and that acetyl-lysine were to recruit a distinct histone acetylase that acetylated another lysine, then a requirement for both acetylations for transcription to occur would produce a low-pass filter. This filter would reject transient spikes in signalling pathways and allow transcriptional outcomes only in response to sustained signalling. A careful examination of the temporal response of histone modifications to signalling will help determine if this might occur for the correlated modifications. Finally, if one modification recruits enzymes that modify the remaining residues, then having multiple modifications allows for switch-like behaviour [53,54].

Stereotyped Promoter Architecture

One of the two groups of histone modifications exhibits a striking, stereotyped pattern in promoter regions. Nucleosomes immediately adjacent to the TSS are hypo-acetylated at H2BK16, H4K8, and H4K16. This hypo-acetylation does not correlate with transcription levels, and the inability of the histone modification pattern at the gene-distal TSS-adjacent nucleosome to accurately reflect transcriptional activity of the associated gene (Figure 6C) does not support the idea that upstream modifications are causal for transcription.

In separate work, we have identified this di-nucleosomal domain that flanks the TSS as highly enriched for the H2A isoform Htz1 (demonstrating that these nucleosomes do not appear deacetylated due to some artifactual difficulty with immunoprecipitation). Also, this enrichment is independent of transcription (unpublished data). In other words, the majority of promoter nucleosome-free regions in yeast are surrounded on either side by nucleosomes with hypo-acetylated H2BK16, hypo-acetylated H4K8 and K16, and Htz1 in place of H2A. These results raise two questions: how does this domain arise, and what is its functional role in transcription?

Previous reports have shown that Rpd3 deacetylates one to three nucleosomes when recruited to promoters [44], consistent with the width of this deacetylation domain. However, the generality of the pattern observed here suggests that multiple distinct deacetylases function in this localized manner, because Rpd3 is present at only a subset of the promoters analyzed [31,43]. Alternatively, it is possible that these nucleosomes turn over rapidly (due to the presence of some assembly of chromatin-remodelling activities at promoters), and that the histone isoform and modification pattern exhibited reflects the composition of free histones in the nucleoplasm. In either case, the function of this domain remains elusive at present.

Relationship of Histone Modifications to Transcription

We have described a group of histone modifications that co-occur, and that are preferentially found at the 5′ ends of actively transcribed genes. This relationship between histone modification patterns, location relative to coding regions, and transcript abundance, would be expected if histone modification played a largely passive, rather than instructive, role in transcription, with nucleosomes being modified by various enzymes associated with RNA polymerase. This is clearly the case, for example, for PolII-associated Set1, which is responsible for the correlation between H3K4 tri-methylation over the 5′ end of coding regions and corresponding transcription levels. A similar type of mechanism appears to hold for the Set2-mediated tri-methylation of H3K36, which occurs over transcribed genes [55]. However, mutant studies have shown abundant transcriptional defects associated with mutations in histone-modifying enzymes [56,57]. These studies cannot determine whether histone modification is instructive or permissive for transcription—in other words, whether histone modifications initiate a chain of events that result in transcription, or whether that gene is associated with a non-permissive chromatin structure that must be antagonized using the modification in question. We suggest that the transcription-associated modifications play a permissive role in gene expression, and that the transcriptional defects in histone-modification mutants result from a partial inability of RNA polymerase to transit unmodified nucleosomes [58,59], or to a failure to recruit factors required for efficient transcription [60]. However, we do not rule out the possibility that histone modifications play both roles, with an initial mark that is causal for a transcription pattern subsequently "erased" by modifications occurring with the resultant transcription.

The Histone Code

Taken together, these results do not support a model for the histone code in which a vast set of widely varying modification combinations play complicated instructive roles in transcriptional regulation. Instead, these results further extend genome-wide studies in *Drosophila,* which show that histone modifications occur in few independent combinations [25], and suggest that these patterns are often the result, rather than the cause, of transcription. These results therefore emphasize a role for modifications of the histone tails as facilitators of transcription. It will be of great interest in future studies to assay the dynamic nature of histone modifications during changes in transcription, and the establishment of histone modification patterns during DNA replication.

Materials and Methods

Yeast culture. An aliquot of 450 ml of BY4741 *bar1*Δ cells was grown to an A_{600} OD of 0.9 in 2-L flasks shaking at 200 rpm in a 28 °C water bath. Formaldehyde (37%) was added to a 1% final concentration, and the cells were incubated for 15 min at 25 °C, shaking, at 90 rpm. Then, 2.5 M glycine was added to a final concentration of 125 mM, to quench the formaldehyde. The cells were inverted and let to stand at 25 °C for 5 min. The cells were spun down at 3,000 × g for 5 min at 4 °C and washed twice, each time with an equal volume of ice-cold sterile water.

Micrococcal nuclease digestion. The cell pellets were resuspended in 39 ml Buffer Z (1 M sorbitol, 50 mM Tris-Cl [pH 7.4]), 28 μl of β-ME (14.3 M, final concentration 10 mM) was added, and cells were vortexed to resuspend. Then, 1 ml of zymolyase solution (10 mg/ml in Buffer Z; Seikagaku America, Falmouth, Massachusetts, United States) was added, and the cells were incubated at 28 °C, shaking at 200 rpm, in 50-ml conical tubes, to digest cell walls. Spheroplasts were then spun at 3,000 × g, 10 min, at 4 °C. Spheroplast pellets were resuspended and split into aliquots of 600 μl of NP-S buffer (0.5 mM spermidine, 1 mM β-ME, 0.075% NP-40, 50 mM NaCl, 10 mM Tris [pH 7.4], 5 mM $MgCl_2$, 1 mM $CaCl_2$) per 90-ml cell culture equivalent. Forty units of micrococcal nuclease (Worthington Biochemical, Lakewood, New Jersey, United States) were added, and the spheroplasts were incubated at 37 °C for 20 min—this was determined in initial titrations to yield > 80% mononucleosomal DNA (see Figure S1), but to repeat these results an independent titration should be carried out as a preliminary study. The digestion was halted by shifting the reactions to 4 °C and adding 0.5 M EDTA to a final concentration of 10 mM.

ChIP. All steps were done at 4 °C unless otherwise indicated. For each aliquot, Buffer L (50 mM Hepes-KOH [pH 7.5], 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate) components were added from concentrated stocks (10–20×) for a total volume of 800 μl per aliquot. Each aliquot was incubated with 80–100 μl of 50% Sepharose Protein A Fast-Flow bead slurry (Sigma, St. Louis, Missouri, United States) equilibrated in Buffer L for 1 h on a tube rotisserie rotator. The beads were pelleted with a 1-min spin at 3,000 × g, and approximately 2.5%–5% of the supernatant was set aside as ChIP input material. With the remainder, antibodies were added to each aliquot (20% of a 450-ml cell culture) in the following volumes: 25 μl anti-H3K4Me1 Ab (affinity purified; Abcam, Cambridge, Massachusetts, United States), 6 μl anti-H3K4Me2 Ab (affinity purified; Abcam), 6 μl anti-H3K4Me3 Ab (affinity purified; Abcam), 4 μl anti-H4K16Ac Ab (whole antiserum; Abcam), 9 μl anti-H4K5Ac Ab (whole antiserum; Abcam), 3 μl anti-H3K14Ac Ab (whole antiserum; Upstate Cell Signaling Solutions, Charlottesville, Virginia, United States), 3 μl anti-H2AK7Ac Ab (whole antiserum; Upstate), 2 μl, anti-H4K8Ac Ab (whole antiserum; Abcam), 15 μl, anti-H4K12Ac Ab (whole antiserum; Abcam), 25 μl anti-Ac Ab (whole antiserum; Abcam), 16 μl anti-H3K9Ac Ab (affinity purified; Abcam), 25 μl anti-H2BK16Ac (L) (whole antiserum; Abcam), and 3 μl anti-H3K18Ac Ab (whole antiserum; gift of M. Grunstein). We also used 3 μl of a distinct antibody to H4K16Ac (whole antiserum; gift of M. Grunstein) to assess specificity of different sources of antibody. Replicates using this antibody were as correlated with each other as they were with replicates using the Abcam antibody.

These were incubated, rotating, overnight (~16 h), after which the sample was transferred to a tube containing 80–100 μl of 50% Protein A bead slurry. The sample was incubated with the beads for 1 h for the immunoprecipitation, after which the beads were pelleted by a 1-min spin at 3,000 × g. After removal of the supernatant, the beads were washed with a series of buffers in the following manner: 1 ml of the buffer would be added, and the sample rotated on the tube rotisserie for 5 min, after which the beads would be pelleted in a 30-s spin at 3,000 × g and the supernatant removed. The washes were performed twice for each buffer in the following order: Buffer L, Buffer W1 (Buffer L with 500 mM NaCl), Buffer W2 (10 mM Tris-HCl [pH 8.0], 250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1mM EDTA), and 1× TE (10 mM Tris, 1 mM EDTA [pH 8.0]). After the last wash, 125 μl of elution buffer (TE [pH 8.0] with 1% SDS, 150 mM NaCl, and 5 mM dithiothreitol) was added to each sample, and the beads were incubated at 65 °C for 10 min, with frequent mixing. The beads were spun for 2 min at 10,000 × g, and the supernatant was removed and retained. The elution process was repeated once for a total volume of 250 μl of eluate. For the ChIP input material set aside, elution buffer was added for a total volume of 250 μl. After overlaying the samples with mineral oil, the samples were incubated overnight at 65 °C to reverse cross-links.

Antibody specificity. A significant concern with ChIP studies is the epitope specificity of the antibodies used. High correlations between different modifications could arise if two antibodies cross-reacted. We note four reasons that this is unlikely to be a major problem for this study. First, if antibodies did indeed cross-react, then the resulting profiles should look like some weighted average (depending on relative affinities of the two antibodies) of the two "pure" profiles. If there were a third modification pattern (besides what we term the *transcription-dependent* and *transcription-independent* patterns), then the two antibodies in question would be expected to show a third *mixed* pattern, distinct from the two patterns described, and this was not observed. On the other hand, if only two true patterns do exist but there is cross-reactivity for antibodies, the mixed profile is expected to show a 5′ gradient of acetylation, along with two deacetyl nucleosomes adjacent to the TSS. This pattern was seen for H2AK7, but, as we note, this is likely due to the replacement of H2A with Htz1 at the TSS-adjacent nucleosomes. Furthermore, this pattern was not seen for the H3K14 antibody, which recognizes lysine in the context of a similar site to that of H2AK7 (GGKA). So we do not believe that these antibodies are cross-reacting.

Second, we repeated experiments for one of the epitopes in this study (H4K16) with two distinct antibodies, and the results were indistinguishable. One of these antibodies, from the Grunstein lab, was previously tested for cross-reactivity by attempting ChIP from strains carrying the H4K16R mutation [37].

Third, there are two pairs of antibodies for which cross-reaction is most likely to be a concern: H4K5 and K12 (both lysines occur in the context of GKGG), and H2AK7 and H3K14 (both occur in the context of GGKA). However, within each pair, the two antibodies are more highly correlated with other antibodies in their group than with the other antibody with a similar recognition site (see Figure 3C). If these antibodies had cross-reacted, then their profiles should be the most highly correlated. In addition, technical literature from Upstate shows that both the H2AK7 and H3K14 acetylation antibodies fail to immunoprecipitate DNA from yeast strains carrying the appropriately mutated recognition site.

Finally, it is worth noting that even if a pair or two of antibodies cross-reacted, the point that histone modifications occur at reduced dimensionality would still hold. Instead of 12 dimensions reducing to two dimensions, we would say, for example, that 10 dimensions reduce to two. This is not, to our thinking, a significant change in the central message of this study. In addition, it would not challenge the other main points of the manuscript, that the two TSS-adjacent nucleosomes exhibit a stereotyped modification pattern and that most of the histone modification that correlates with transcription levels occurs over coding regions.

Protein degradation and DNA purification. After cooling the samples down to room temperature, each sample was incubated with an equal volume of proteinase K solution (1× TE with 0.4 mg/ml glycogen, and 1 mg/ml proteinase K) at 37 °C for 2 h. Each sample was then extracted twice with an equal volume of phenol and once with an equal volume of 25:1 chloroform:isoamyl alcohol. Phase-lock gel tubes were used to separate the phases (light gel for phenol, heavy gel for chloroform:isoamyl alcohol). Afterwards, 0.1 volume 3.0 M sodium acetate [pH 5.3] and 2.5 volumes of 100% ice-cold ethanol were added, and the DNA was allowed to precipitate overnight at –20 °C. The DNA was pelleted by centrifugation at 14,000 × g for 15 min at 4 °C, washed once with cold 70% ethanol, and spun at 14,000 × g for 5 min at 4 °C. After removing the supernatant, the pellets were allowed to dry and then were resuspended in 20 μl 10 mM Tris-Cl, 1 mM EDTA [pH 8.0], and 0.5 μg of RNase A was added. The samples were incubated at 37 °C for 1 h, and then treated with 7.5 units of calf intestinal alkaline phosphatase in a 30-μl volume supplemented with NEB Buffer 3 (10× concentration of 100 mM NaCl, 50 mM Tris-HCl [pH 7.9], 10 mM $MgCl_2$, 1 mM dithiothreitol). The samples were then incubated for a further 1 h at 37 °C and then cleaned up with the Qiagen MinElute Reaction Cleanup Kit (Qiagen, Valencia, California, United States), following manufacturer's directions, except with an elution volume of 20 μl.

Linear amplification of DNA. The samples were amplified, with a starting amount of 125 ng for ChIP input materials and up to 75 ng for ChIP samples, using the DNA linear amplification method described in BMC Genomics 4:19 [32].

Microarray hybridization. RNA produced from the linear amplification (3 μg) was used to label probe via the amino-allyl method as described at http://www.microarrays.org. Labelled probes were hybridized onto a yeast tiled oligonucleotide microarray [29] at 65 °C for 16 h, and washed as described at http://www.microarrays.org. The arrays were scanned at 5-μm resolution with an Axon Laboratories (Sunnyvale, California, United States) GenePix 4000B scanner running GenePix 5.1.

Image analysis and data processing. Array features were filtered using the autoflagging feature of GenePix 5.1 with the following criteria defining features to be discarded: [Flags] = [Bad], or [Flags] = [Absent], or [Flags] = [Not Found], or LCase([ID]) = "empty", or LCase([ID]) = "blank", or ([SNR 635] < 3 and [SNR 532] < 3), or [F Pixels] < 100, or ([F Pixels] < 150 and [Circularity] < 75).

The remaining features for each array were then block-normalized by calculating the average net signal intensity for each channel in a given block, and then taking the product of this average and the net signal intensity for each filtered array feature in the block. Afterwards, all block-normalized array features were normalized using a global average net signal intensity as the normalization factor.

Each histone tail modification epitope was chromatin-immunoprecipitated in three to six biological replicates, with additional technical replicates of the microarray hybridizations. Outlying replicates were removed (with a minimum remainder of three replicates), and the median was calculated and used for subsequent data analysis.

Normalization of modification and PolII data. Each assay was repeated three to six times, and median values per probe were calculated. Measurements for each antibody were first log (base 2) transformed and then normalized (to mean of zero and variance of one).

Data availability. Data can be viewed at http://compbio.cs.huji.ac.il/Nucs. Data are downloadable at http://www.cgr.harvard.edu/chromatin, and have been deposited in GEO.

Clustering of aligned genes. The genes were clustered using PCluster, a probabilistic hierarchical clustering algorithm [61]. Probes at locations relative to gene reference point, either beginning of coding sequence (CDS) (Figure 2A) or TSS (Figure 2B), are used as attributes of the gene. Linker probes (based on the nucleosome locations of [29]) were discarded and treated as missing values.

Splitting genes into transcriptional groups. Each gene was assigned a transcription activity value based on the average enrichment of PolII along CDS probes. Genes with less than five CDS probes were removed to reduce noise. We then used thresholds of 0.75 and –0.75 to classify genes as highly, mid-, and untranscribed. This resulted in 75 highly transcribed genes, 192 intermediate genes, and 57 poorly transcribed genes. We also repeated the analysis presented in Figure 2C using mRNA abundance rather than PolII occupancy to bin genes (Figure S4), and the results were qualitatively indistinguishable.

Averaging probes into nucleosomal-based data. A total of 24,947 probes were assigned to 2,288 nucleosomes using a four-probe minimum size cutoff [29]. We used the hand-called set of nucleosome positions (these were generated by inspection and adjustment of the automated hidden Markov model calls; these positions are provided in the dataset associated with [29]), as that set covered a slightly greater fraction of the genome. Results are qualitatively unchanged when only HMM calls are used (unpublished data). For each antibody, the nucleosomal values were set by the median levels of relevant probes.

Genomic classification of nucleosomes. Nucleosomes were annotated based on their relative position to nearby genes. Nucleosomes in the first (or last) 500 bp of annotated genes were annotated as 5′ CDS (or 3′ CDS) nucleosome. Other CDS nucleosomes were annotated as mid-CDS. The two TSS adjacent nucleosomes were annotated as TSS distal (5′) and proximal (3′) nucleosomes. Nucleosomes upstream (up to 1 kb or closer to non-dubious CDSs) were annotated as promoter nucleosomes. Nucleosomes around tRNA genes (200 bp from each side) or ARS elements (200 bp from each side) were annotated as tRNA or ARS nucleosomes. Other nucleosomes were annotated as null. In certain cases, we allowed more than one annotation per nucleosome; for instance, a nucleosome between two divergent genes can be annotated as TSS-proximal for one gene, and a promoter nucleosome for another one.

Single nucleosome clustering. Nucleosomes were clustered using PCluster [61], treating each nucleosome as a vector of 12 values.

Principal component analysis. Principal component analysis was applied to the nucleosomal modification data of 2,288 nucleosomes versus 12 modifications using MATLAB 6.5 (rel 13) procedure "princomp." Density visualization was done using Parzen windows density estimator with Gaussian kernels (with standard deviation of 0.3) .

Genomic enrichment of modifications. We compared the modifications of nucleosomes affiliated with each genomic location (promoter, TSS distal, etc.) to all other nucleosomes, using a standard two-tail *t*-test. To correct for multiple hypotheses, we used a 5% false discovery rate procedure [62]. The average change was then calculated for < modification, genomic location > pairs with significant *p*-values.

Transcription-specific modifications. To identify specific modifications at genomic locations with significant correlations to expression levels of nearby genes, we trained a classification method to predict whether a nucleosome was associated with genes enriched or depleted for PolII. To prevent biased results, we applied a leave-one-out cross-validation procedure in which the tested nucleosome was removed from the training set, and a classifier was trained on the rest of the nucleosomes and used to predict the held-out nucleosome label. We used a Naive Bayes classifier [63] using the implementation described [64]. We then classified the held-out nucleosome, based on the probability of its modification pattern under each of the classes. We computed the overall accuracy of classification and a *p*-value by repeating the same leave-one-out procedure with randomly reshuffled nucleosome labels.

Functional classification of nucleosomes. We used recent genomic studies [39–41] and compiled a set of target promoters for each factor. We then tested the promoter and TSS-distal and TSS-proximal nucleosomes of these genes for enrichment of specific modifications. In addition, we created a subset of the target nucleosomes of Harbison et al., by restricting the nucleosomes to those up to 100 bp away from putative binding sites bound in rich growth conditions [42]. As described earlier, we compared the "bound" nucleosomes to all other promoter/TSS nucleosomes, and used a false discovery rate-corrected two-tail *t*-test.

Supporting Information

Dataset S1. Complete Dataset

Individual worksheets contain data for all individual replicates before range normalization, for combined median data organized by epitope, and for combined median data after range normalization.

Found at DOI: 10.1371/journal.pbio.0030328.sd001 (48 MB XLS).

Dataset S2. Replicate Reproducibility

Data contain correlations between individual experiments for each antibody.

Found at DOI: 10.1371/journal.pbio.0030328.sd002 (24 KB XLS).

Figure S1. Digestion of Chromatin to Mononucleosomes before Immunoprecipitation

Gels show micrococcal nuclease-digested DNA from multiple independent cultures used for the immunoprecipitations reported here. Molecular markers are as indicated. Blue dots indicate nucleosomal DNA used for immunoprecipitations, while green dots show sonicated DNA from the same culture. Digested DNA used for immunoprecipitation was typically > 80% mononucleosome.

Found at DOI: 10.1371/journal.pbio.0030328.sg001 (674 KB PDF).

Figure S2. Low Levels of Histone Modification over Heterochromatin

Data are plotted as in Figure 1B. Chromosome III coordinates are shown above the modification data. Three panels show data for a portion of (from left to right) TelIIIL, HML, and TelIIIR. Only partial regions of the three are shown, as the remainder was not tiled due to cross-hybridization concerns [29].

Found at DOI: 10.1371/journal.pbio.0030328.sg002 (551 KB PDF).

Figure S3. Broad Patterns of Histone Modifications

Data are aligned by the TSS, and plotted as in Figure 2B for all remaining modifications, as indicated.

Found at DOI: 10.1371/journal.pbio.0030328.sg003 (1.8 MB PDF).

Figure S4. Relationship of Histone Modifications to mRNA Abundance

Genes were grouped into low, medium, and high mRNA abundance classes using data from competitive hybridizations of mRNA versus genomic DNA on cDNA microarrays (CLL and SLS, unpublished data). Low-abundance mRNAs were defined as those with log(2) ratios less than −1, while high-abundance mRNAs were defined as those exhibiting log(2) ratios greater than 1. Histone modification data are averaged and displayed as in Figure 2C, and results are qualitatively indistinguishable from those generated using PolII occupancy to classify genes.

Found at DOI: 10.1371/journal.pbio.0030328.sg004 (676 KB PDF).

Figure S5. Representation of the First Two Principal Components

The first component (left panel) consists of all positive coefficients (plotted on the *y*-axis), and therefore captures the global magnitude of modification (both acetylation and methylation). The second component differentiates between the two groups of correlated modifications (see Figure 3C). Bars indicate different epitopes as indicated.

Found at DOI: 10.1371/journal.pbio.0030328.sg005 (512 KB PDF).

Figure S6. Principal Component Analysis of Nucleosome Modifications

Data plotted as in Figure 4B and 4C, right panels.

Found at DOI: 10.1371/journal.pbio.0030328.sg006 (580 KB PDF).

Figure S7. Nucleosome Modifications Relate to Transcription Level

Classification plot as described in Figure 5, using mid-CDS nucleosomes. The average accuracy of random classification was 61.27%, with a standard deviation of 5.76%. Accuracy of classifier was 82.65% ($p < 0.0001$).

Found at DOI: 10.1371/journal.pbio.0030328.sg007 (397 KB PDF).

Accession Numbers

The Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) accession numbers for the experiments described here are GSM64526–GSM64587, GSM64591, and GSM64592, and are part of series accession number GSE2954.

Acknowledgments

We would like to thank N. Francis, L. Garwin, H. Madhani, T. Maniatis, H. Margalit, A. Murray, A. Regev, B. Suter, and K. Thorn for critical reading of the manuscript and/or helpful discussions. We thank S. Kurdistani and M. Grunstein for their generous gifts of antibodies to H3K18Ac and H4K16Ac. TK is supported by the Yeshaya Horowitz Foundation through the Center for Complexity Science. NF is supported by the Harry and Abe Sherman Senior Lectureship in Computer Science. This research was supported by grants to SB, SLS, NF, and OJR from the National Institute of General Medical Sciences, and to NF from the Israeli Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

Author contributions. CLL, SB, SLS, and OJR conceived and designed the experiments. CLL and MK performed the experiments. CLL, TK, NF, and OJR analyzed the data. NF and OJR wrote the paper. ■

References

1. Venter U, Svaren J, Schmitz J, Schmid A, Horz W (1994) A nucleosome precludes binding of the transcription factor Pho4 in vivo to a critical target site in the PHO5 promoter. Embo J 13: 4848–4855.
2. Stunkel W, Kober I, Seifart KH (1997) A nucleosome positioned in the distal promoter region activates transcription of the human U6 gene. Mol Cell Biol 17: 4397–4405.
3. Langst G, Becker PB (2004) Nucleosome remodeling: One mechanism, many phenomena? Biochim Biophys Acta 1677: 58–63.
4. Turner BM (2000) Histone acetylation and an epigenetic code. Bioessays 22: 836–845.
5. Turner BM (2002) Cellular memory and the histone code. Cell 111: 285–291.
6. Strahl BD, Allis CD (2000) The language of covalent histone modifications. Nature 403: 41–45.
7. Schreiber SL, Bernstein BE (2002) Signaling network model of chromatin. Cell 111: 771–778.
8. Kurdistani SK, Grunstein M (2003) Histone acetylation and deacetylation in yeast. Nat Rev Mol Cell Biol 4: 276–284.
9. Berger SL (2002) Histone modifications in transcriptional regulation. Curr Opin Genet Dev 12: 142–148.
10. Hong L, Schroth GP, Matthews HR, Yau P, Bradbury EM (1993) Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 “tail” to DNA. J Biol Chem 268: 305–314.
11. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251–260.
12. Dhalluin C, Carlson JE, Zeng L, He C, Aggarwal AK, et al. (1999) Structure and ligand of a histone acetyltransferase bromodomain. Nature 399: 491–496.
13. Waterborg JH (2001) Dynamics of histone acetylation in *Saccharomyces cerevisiae*. Biochemistry 40: 2599–2605.
14. Vogelauer M, Wu J, Suka N, Grunstein M (2000) Global histone acetylation and deacetylation in yeast. Nature 408: 495–498.
15. Hebbes TR, Thorne AW, Crane-Robinson C (1988) A direct link between core histone acetylation and transcriptionally active chromatin. Embo J 7: 1395–1402.
16. De Nadal E, Zapater M, Alepuz PM, Sumoy L, Mas G, et al. (2004) The MAPK Hog1 recruits Rpd3 histone deacetylase to activate osmoresponsive genes. Nature 427: 370–374.
17. Wang A, Kurdistani SK, Grunstein M (2002) Requirement of Hos2 histone deacetylase for gene activity in yeast. Science 298: 1412–1414.
18. Kurdistani SK, Tavazoie S, Grunstein M (2004) Mapping global histone acetylation patterns to gene expression. Cell 117: 721–733.
19. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, et al. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature 410: 120–124.
20. Lachner M, O’Carroll D, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410: 116–120.
21. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, et al. (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119: 941–953.
22. Ahmad K, Henikoff S (2002) The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. Mol Cell 9: 1191–1200.
23. Ng HH, Robert F, Young RA, Struhl K (2003) Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Mol Cell 11: 709–719.
24. Dion MF, Altschuler SJ, Wu LF, Rando OJ (2005) Genomic characterization reveals a simple histone H4 acetylation code. Proc Natl Acad Sci U S A: 5501–5506.
25. Schubeler D, MacAlpine DM, Scalzo D, Wirbelauer C, Kooperberg C, et al. (2004) The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. Genes Dev 18: 1263–1271.
26. Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, et al. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169–181.
27. Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL (2004) Global nucleosome occupancy in yeast. Genome Biol 5: R62.
28. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. Nat Genet 36: 900–905.
29. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science: 626–630.
30. Bernstein BE, Humphrey EL, Erlich RL, Schneider R, Bouman P, et al. (2002) Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A 99: 8695–8700.
31. Robyr D, Suka Y, Xenarios I, Kurdistani SK, Wang A, et al. (2002) Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. Cell 109: 437–446.
32. Liu CL, Schreiber SL, Bernstein BE (2003) Development and validation of a T7 based linear amplification for genomic DNA. BMC Genomics 4: 19.
33. Boeger H, Griesenbeck J, Strattan JS, Kornberg RD (2003) Nucleosomes unfold completely at a transcriptionally active promoter. Mol Cell 11: 1587–1598.
34. Schwabish MA, Struhl K (2004) Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. Mol Cell Biol 24: 10111–10117.
35. Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, et al. (2004) The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature 432: 517–522.
36. Krogan NJ, Dover J, Wood A, Schneider J, Heidt J, et al. (2003) The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p: Linking transcriptional elongation to histone methylation. Mol Cell 11: 721–729.
37. Suka N, Suka Y, Carmen AA, Wu J, Grunstein M (2001) Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin. Mol Cell 8: 473–479.
38. Roh TY, Ngau WC, Cui K, Landsman D, Zhao K (2004) High-resolution genome-wide mapping of histone modifications. Nat Biotechnol 22: 1013–1016.
39. Ng HH, Robert F, Young RA, Struhl K (2002) Genome-wide location and

regulated recruitment of the RSC nucleosome-remodeling complex. Genes Dev 16: 806–819.

40. Robert F, Pokholok DK, Hannett NM, Rinaldi NJ, Chandy M, et al. (2004) Global position and recruitment of HATs and HDACs in the yeast genome. Mol Cell 16: 199–209.
41. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298: 799–804.
42. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431: 99–104.
43. Kurdistani SK, Robyr D, Tavazoie S, Grunstein M (2002) Genome-wide binding map of the histone deacetylase Rpd3 in yeast. Nat Genet 31: 248–254.
44. Kadosh D, Struhl K (1998) Targeted recruitment of the Sin3-Rpd3 histone deacetylase complex generates a highly localized domain of repressed chromatin in vivo. Mol Cell Biol 18: 5121–5127.
45. Kasten M, Szerlong H, Erdjument-Bromage H, Tempst P, Werner M, et al. (2004) Tandem bromodomains in the chromatin remodeler RSC recognize acetylated histone H3 Lys14. Embo J 23: 1348–1359.
46. Hurowitz EH, Brown PO (2003) Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. Genome Biol 5: R2.
47. Sterner DE, Berger SL (2000) Acetylation of histones and transcription-related factors. Microbiol Mol Biol Rev 64: 435–459.
48. Pray-Grant MG, Daniel JA, Schieltz D, Yates JR III, Grant PA (2005) Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature 433: 434–438.
49. Lo WS, Trievel RC, Rojas JR, Duggan L, Hsu JY, et al. (2000) Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14. Mol Cell 5: 917–926.
50. Cheung P, Tanner KG, Cheung WL, Sassone-Corsi P, Denu JM, et al. (2000) Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation. Mol Cell 5: 905–915.
51. Wittschieben BO, Otero G, de Bizemont T, Fellows J, Erdjument-Bromage H, et al. (1999) A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme. Mol Cell 4: 123–128.
52. Agalioti T, Chen G, Thanos D (2002) Deciphering the transcriptional histone acetylation code for a human gene. Cell 111: 381–392.
53. Ferrell JE Jr. (1996) Tripping the switch fantastic: How a protein kinase cascade can convert graded inputs into switch-like outputs. Trends Biochem Sci 21: 460–466.
54. Ferrell JE Jr. (1997) How responses get more switch-like as you move down a protein kinase cascade. Trends Biochem Sci 22: 288–289.
55. Krogan NJ, Kim M, Tong A, Golshani A, Cagney G, et al. (2003) Methylation of histone H3 by Set2 in *Saccharomyces cerevisiae* is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol 23: 4207–4218.
56. Bernstein BE, Tong JK, Schreiber SL (2000) Genomewide studies of histone deacetylase function in yeast. Proc Natl Acad Sci U S A 97: 13708–13713.
57. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. Cell 95: 717–728.
58. Protacio RU, Li G, Lowary PT, Widom J (2000) Effects of histone tail domains on the rate of transcriptional elongation through a nucleosome. Mol Cell Biol 20: 8866–8878.
59. Kristjuhan A, Walker J, Suka N, Grunstein M, Roberts D, et al. (2002) Transcriptional inhibition of genes with severe histone h3 hypoacetylation in the coding region. Mol Cell Biol 10: 925–933.
60. Santos-Rosa H, Schneider R, Bernstein BE, Karabetsou N, Morillon A, et al. (2003) Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. Mol Cell 12: 1325–1332.
61. Friedman N (2003) PCluster: Probabilistic agglomerative clustering of gene expression profile. Jerusalem: Hebrew University. 6 p. Available: http://ai.stanford.edu/~erans/module__nets/figures/pcluster.pdf. Accessed 28 July 2005.
62. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Stat Soc B 57: 289–300.
63. Duda RO, Hart PE (1973) Pattern classification and scene analysis. New York: Wiley. 482 p.
64. Ben-Dor A, Friedman N, Yakhini Z (2002) Overabundance Analysis and Class Discovery in Gene Expression Data. Jerusalem: Hebrew University. 26 p. Available: http://www.cs.huji.ac.il/~nirf/Papers/BFY2Full.pdf. Accessed 28 July 2005.

Chapter 4 – Paper

Dynamics of replication-independent histone turnover in budding yeast

Michael F. Dion* Tommy Kaplan*, Minkyu Kim, Stephen Buratowski, Nir Friedman, Oliver J. Rando

Science (2007) **315**, 1405

* These authors contributed equally to this work.

II acceptor adjacent to the proposed donor site, with the 4-OH end available for interaction with both E114 and C1 of the donor sugar. This position of lipid II is comparable to that of subsites +1 and +2 of the substrate in λL (*15*), with the prereaction GT_{51} substrates similar to the postreaction lysozyme products.

We propose that E114 is a Brønsted base and acts to directly abstract a proton from the 4-OH group of the lipid II acceptor. The deprotonated form of E114 may be stabilized by the adjacent R249 residue, strictly conserved as part of motif V. The proton abstraction step probably occurs concomitantly with the electrophilic migration of the donor C1 toward the acceptor 4-OH group (Fig. 4, A and B). In the moenomycin complex, the conserved E171 residue lies closer to the glyceric acid moiety than the phosphate-sugar linkage (the β phosphate in our substrate model), which in combination with pH activity profiles of the *E. coli* PBP1b enzyme (*16*) casts some doubt on whether E171 protonates the sugar-phosphate linkage to assist catalysis. Furthermore, mutants of this residue in *E. coli* PBP1b retain some residual activity, whereas those of our predicted Brønsted base, E114, do not (*9*). If E171 does not act to protonate the substrate, then we propose that it helps to coordinate the pyrophosphate group of the donor, either directly or via a divalent metal cation. The variable pH optima and divalent cation requirements of the GT_{51} family of enzymes (*17–19*) may result from varying local environments of the E171 residue. The S_N2-like reaction occurs between donor and acceptor, causing inversion at the donor C1 anomeric carbon and formation of the β1,4-linked product. The lipid-pyrophosphate leaving group of the donor is then free to diffuse away and be recycled in lipid II synthesis. We propose that translocation of the newly formed product to the donor site is assisted by a higher affinity for the pyrophosphate moiety in the donor site than in the acceptor site, with the conserved positively charged K155, K163, R167, and K168 residues located near the donor pyrophosphate region of the active site (Fig. 4C). This model is again reminiscent of the lysozyme active site, where the +1 and +2 subsites that match the modeled GT_{51} acceptor sugars possess the lowest substrate affinity of all the subsites. These two structures now provide a basis for addressing further questions about the mechanism of this important family of enzymes and for the design of new antibacterials. This work also opens the door for understanding structure and function relationships in other clinically important families of lipid-sugar GTs.

References and Notes

1. C. Goffin, J. M. Ghuysen, *Microbiol. Mol. Biol. Rev.* **62**, 1079 (1998).
2. T. D. Bugg *et al.*, *Biochemistry* **30**, 10408 (1991).
3. C. T. Walsh, *Science* **261**, 308 (1993).
4. R. C. Goldman, D. Gange, *Curr. Med. Chem.* **7**, 801 (2000).
5. P. Butaye, L. A. Devriese, F. Haesebrouck, *Antimicrob. Agents Chemother.* **45**, 1374 (2001).
6. Materials and methods are available as supporting material on *Science* Online.
7. C. Contreras-Martel *et al.*, *J. Mol. Biol.* **355**, 684 (2006).
8. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
9. M. Terrak *et al.*, *Mol. Microbiol.* **34**, 350 (1999).
10. L. Holm, C. Sander, *Nucleic Acids Res.* **27**, 244 (1999).
11. P. Welzel *et al.*, *Tetrahedron* **43**, 585 (1987).
12. J. Halliday, D. McKeveney, C. Muldoon, P. Rajaratnam, W. Meutermans, *Biochem. Pharmacol.* **71**, 957 (2006).
13. P. Welzel, *Chem. Rev.* **105**, 4610 (2005).
14. O. Ritzeler *et al.*, *Tetrahedron* **53**, 1675 (1997).
15. A. K. Leung, H. S. Duewel, J. F. Honek, A. M. Berghuis, *Biochemistry* **40**, 5665 (2001).
16. B. Schwartz *et al.*, *Biochemistry* **41**, 12552 (2002).
17. D. S. Barrett, L. Chen, N. K. Litterman, S. Walker, *Biochemistry* **43**, 12375 (2004).
18. D. Barrett *et al.*, *J. Bacteriol.* **187**, 2215 (2005).
19. M. Terrak, M. Nguyen-Disteche, *J. Bacteriol.* **188**, 2528 (2006).
20. C. Evrard, J. Fastrez, J. P. Declercq, *J. Mol. Biol.* **276**, 151 (1998).
21. We thank S. Withers for helpful mechanistic discussion. We are grateful for beam time and assistance at the Advanced Light Source. The atomic coordinates and structure factors of the apoenzyme and moenomycin complex have been deposited at the Protein Data Bank, with accession numbers 2OLU and 2OLV, respectively. Figures were prepared using PyMol (www.pymol.org). N.C.J.S. is a Howard Hughes Medical Institute (HHMI) international scholar, a Canadian Institute of Health Research (CIHR) scientist, and a Michael Smith Foundation for Health Research (MSFHR) senior scholar. A.L.L. is a MSFHR and CIHR fellow. This work was funded by CIHR and HHMI operating funds and infrastructure funds from the Canada Foundation of Innovation and MSFHR.

Supporting Online Material
www.sciencemag.org/cgi/content/full/315/5817/1402/DC1
Materials and Methods
SOM Text
Fig. S1
Table S1
References

23 October 2006; accepted 24 January 2007
10.1126/science.1136611

Dynamics of Replication-Independent Histone Turnover in Budding Yeast

Michael F. Dion,[1]*† Tommy Kaplan,[2,3]* Minkyu Kim,[4] Stephen Buratowski,[4]
Nir Friedman,[2] Oliver J. Rando[1]†‡

Chromatin plays roles in processes governed by different time scales. To assay the dynamic behavior of chromatin in living cells, we used genomic tiling arrays to measure histone H3 turnover in G1-arrested *Saccharomyces cerevisiae* at single-nucleosome resolution over 4% of the genome, and at lower (~265 base pair) resolution over the entire genome. We find that nucleosomes at promoters are replaced more rapidly than at coding regions and that replacement rates over coding regions correlate with polymerase density. In addition, rapid histone turnover is found at known chromatin boundary elements. These results suggest that rapid histone turnover serves to functionally separate chromatin domains and prevent spread of histone states.

Characterizing the dynamic behavior of nucleosomes is key to understanding the diversity of biological roles of chromatin. Nucleosomes are evicted at many yeast promoters during gene activation (*1–4*) and are reassembled in trans upon repression (*5*). In *Drosophila*, active transcription leads to replacement of histone H3 by the variant isoform H3.3 (*6*, *7*), whereas in budding yeast (whose only H3 is an H3.3 homolog), passage of RNA polymerase II (Pol II) results in eviction of nucleosomes from some (*8*), but not all (*9*), coding regions. In contrast, studies in *Physarum polycephalum* suggest that H3 is not replaced during Pol II transcription (*10*). Furthermore, recent results in yeast suggest that H4 deposition is independent of transcription status (*11*). The disagreement between these studies leads us to map the locus-specific turnover rate of histone H3 at genomic scale so as to address two questions. First, is there evidence for general transcription-dependent H3 turnover? Second, are there additional mechanisms for histone turnover?

To measure turnover rates, we used yeast carrying constitutively expressed Myc-tagged histone H3, as well as an inducible Flag-tagged H3 (*5*) (fig. S1). Flag-H3 was induced in G1-arrested cells for varying amounts of time, chromatin was cross-linked and digested to mononucleosomes (*12*), and Myc- and Flag-tagged histones were immunoprecipitated sepa-

[1]Faculty of Arts and Sciences, Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA. [2]School of Computer Science and Engineering, The Hebrew University, Jerusalem 91904, Israel. [3]Department of Molecular Genetics and Biotechnology, Faculty of Medicine, The Hebrew University, Jerusalem 91120, Israel. [4]Department of Biological Chemistry and Molecular Pharmacology, Harvard University, 240 Longwood Avenue, Boston, MA 02115, USA.

*These authors contributed equally to this work.

†Present address: Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, USA.

‡To whom correspondence should be addressed. E-mail: Oliver.Rando@umassmed.edu

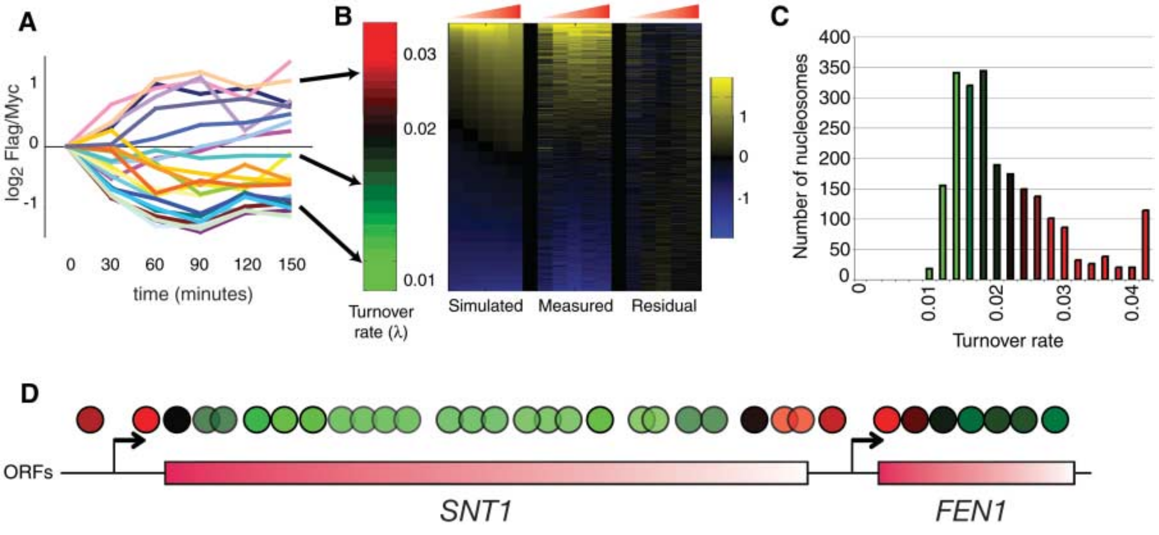


Fig. 1. Time courses of histone turnover in yeast. (**A**) H3 turnover for 23 adjacent nucleosomes in G1-arrested yeast cultures. Flag and Myc were immunoprecipitated at various time points after Flag-H3 induction (*x* axis), and Flag/Myc ratios (*y* axis) were measured by microarray. (**B**) A computational model reduces time course data to a single turnover parameter λ (frequency of histone turnover events, in units of min^{-1}), represented as the leftmost red-to-green color bar. Measured time-course data and data simulated using λ values are represented as blue-yellow heat maps (right). The minor differences (Residual) between measured and simulated data demonstrate that our model captures the majority of histone turnover dynamics during G1 arrest. (**C**) Distribution of turnover rates for nucleosomes in G1-arrested yeast. Binned turnover rates are color coded as in (B). (**D**) Sample genomic stretch, with nucleosomes (A) color coded by turnover rate.

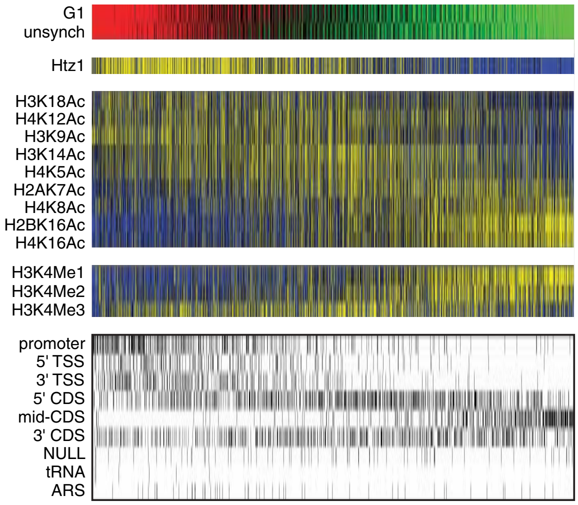


Fig. 2. Relation between histone modifications and H3 turnover, nucleosomes (columns) versus annotations (rows). Nucleosomes are ordered by turnover rate (red-to-green). Modification and Htz1 levels (*12*, *18*) are shown in yellow-to-blue heat maps, where yellow represents enrichment. The bottom panel shows genomic locations (*12*): 5′ and 3′ TSS refer to nucleosomes surrounding the transcriptional start site; promoter indicates other upstream probes. Protein-coding sequences are separated into 5′, middle, and 3′. Other annotations describe autonomously replicating sequences (ARSs), tRNA genes, and Null (any other intergenic region).

rately. Amplified DNAs were competitively hybridized to a 20–base pair (bp) resolution microarray covering 4% of the genome (*13*), yielding Flag/Myc ratios at each time point for each nucleosome on our array (Fig. 1A). We then estimated the turnover rate (number of H3 replacement events per unit of time) of each nucleosome using a simple analytical model that fits the experimental data with a small number of parameters (*14*, *15*) (Fig. 1B).

To test the validity of our results, we repeated the experiment in unsynchronized yeast (fig. S2), observing well-correlated but consistently faster turnover rates, as expected given global H3 deposition during genomic replication (Fig. 2 and fig. S3). We analyzed turnover rates in G1-arrested cells across the entire yeast genome using commercial microarrays with ~265-bp resolution (*16*, *17*) (fig. S4) and obtained a high correlation between rates from the two distinct measurement platforms (fig. S5). We also measured whole-genome histone occupancy (*1*, *3*) (*13*), finding that H3 replacement rates were weakly anticorrelated with H3 occupancy (fig. S6).

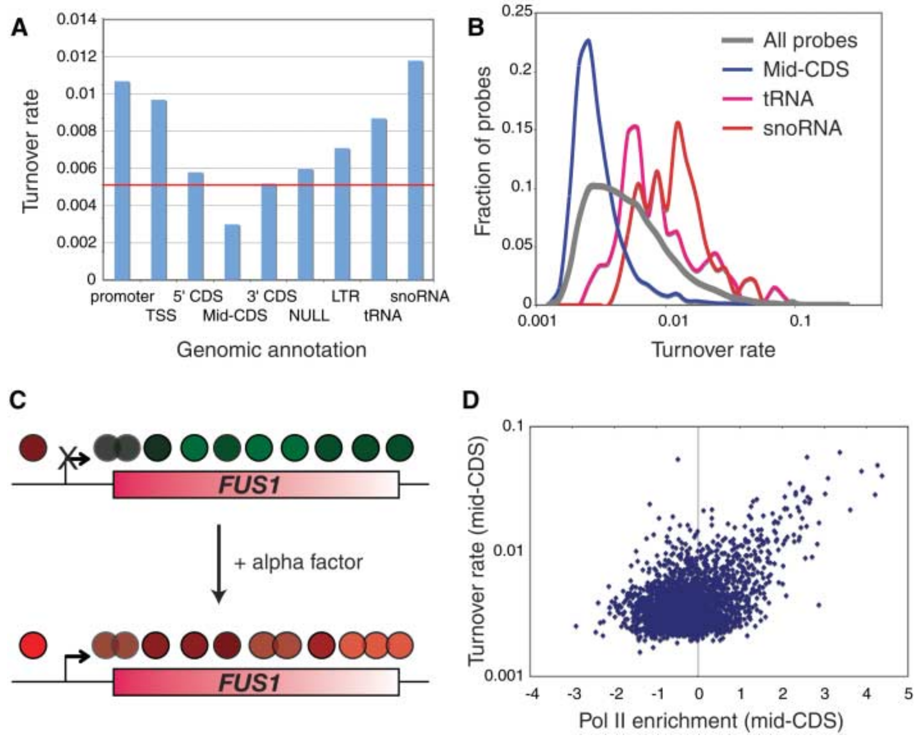
These results are consistent with those expected of H3 replacement from a free pool of H3 and demonstrate that we can recover semiquantitative turnover rates from time-course experiments. The time required for production of Flag-H3 (30 to 45 min) limits our ability to measure the rates of the hottest nucleosomes, which accumulate Flag-H3 before any protein can be detected by Western blot. We therefore caution against literal interpretation of turnover rates, because parameter choices (e.g., Flag-H3 degradation rate) affect absolute turnover rates; however, over a wide range of parameters, the ratio between estimated rates is robust. The resulting rate estimates span one to two orders of magnitude (depending on measurement platform) between "cold" nucleosomes that rarely turn over and hot ones whose replacement rate is faster than the time granularity of our experiment (Fig. 1C and Fig. 3B).

We compared high-resolution turnover rates to previously measured features of these nucleosomes (*12*, *17*, *18*) (Fig. 2 and fig. S7). Nucleosomes over protein-coding regions were coldest, whereas promoter nucleosomes were generally hot. Correspondingly, hot nucleosomes were depleted of the histone modifications that are "stereotypically" depleted surrounding the transcription start site (TSS) (*12*) and were conversely enriched for the histone H2A variant Htz1 (*16*, *18*).

These results are notable for two reasons. First, they suggest that replacement of TSS-adjacent nucleosomes with an appropriately modified nucleoplasmic pool could be partially responsible for promoter patterns of histone modification. Second, erasure of histone modifications due to rapid turnover would result in a steady-state picture of stereotyped promoter chromatin that does not capture transient states, potentially hiding any number of informative histone modification events.

Analysis of median replacement rates for various genomic loci confirmed that the most rapid turnover occurs over promoters, tRNA, and small nucleolar RNA genes (Fig. 3, A and B, and fig. S8). Most unexpected, given the dynamic H3.3 replacement over *Drosophila* genes (*7*, *19*), was the slow H3 turnover over protein-coding genes. Indeed, the coldest probes, mid–coding region probes, cover 28% of the genome yet account for only 10% of turnover. Despite the slower H3 turnover in coding regions, relative variation of turnover rates among coding regions might correlate with polymerase activity. For example, histone turnover over the alpha factor–inducible gene *FUS1* is more rapid in alpha factor–arrested cells than in unsynchronized cells (Fig. 3C and fig. S9). We therefore measured Pol II enrichment across the entire yeast genome, finding that polymerase enrichment over genes exhibited good correlation ($r^2 = 0.54$, $P < 6 \times 10^{-17}$)

Fig. 3. Slow histone replacement over protein-coding genes. (**A**) Median turnover rates for genomic annotations (from whole-genome data). (**B**) Probe-level distributions of transcribed regions compared with the entire data set. *X* axis (logarithmic scale) shows turnover rate. *Y* axis shows fraction of probes within each rate bin. (**C**) *FUS1* coding region and associated nucleosomes, color coded according to turnover rates from high-resolution microarray experiments on unsynchronized yeast cultures (top), and G1-arrested cultures (bottom). (**D**) Scatter plot of coding region histone turnover (whole-genome data) versus $\log_2$ of Pol II enrichment.



with histone replacement rates (Fig. 3D). This is consistent with RNA polymerase passage evicting nucleosomes in some cases, although many highly transcribed genes (*RPL37B*, for example) exhibit low turnover rates.

Although polymerase passage and the resulting histone eviction represent a plausible first step for coding region histone turnover, they are unlikely to account for the bulk of histone replacement (Fig. 3A). Promoters of hot coding regions tend to be hot, but the converse is not true: Most hot promoters were adjacent to cold coding regions (e.g., Fig. 1D). Moreover, replacement rates at promoters were, unlike those at coding regions, poorly correlated with polymerase abundance, either at the promoter or over the coding region (fig. S10), making it unlikely that promoter turnover is solely a result of polymerase activity.

To systematically characterize promoter histone turnover, we tested the hottest subset of promoters for enrichment of published experimental and computational annotations (table S6). The hottest promoters include those carrying binding sites for a subset of transcription factors (such as Rap1, Reb1, Gcn4, and Adr1), those upstream of genes regulated by chromatin-modulating complexes (e.g., Ssn6/Tup1, Mediator, SAGA, Swi/Snf, and Sir), and those upstream of genes associated with nuclear pore components (e.g., Cse1, Mlp1, Nup116, and Nup2). Clustering hot promoters based on enriched annotations yielded independent clusters (Fig. 4A and fig. S11), such as a group of hot promoters associated with nuclear pore components (*20*). These separate clusters suggest that the many enrichments identified potentially

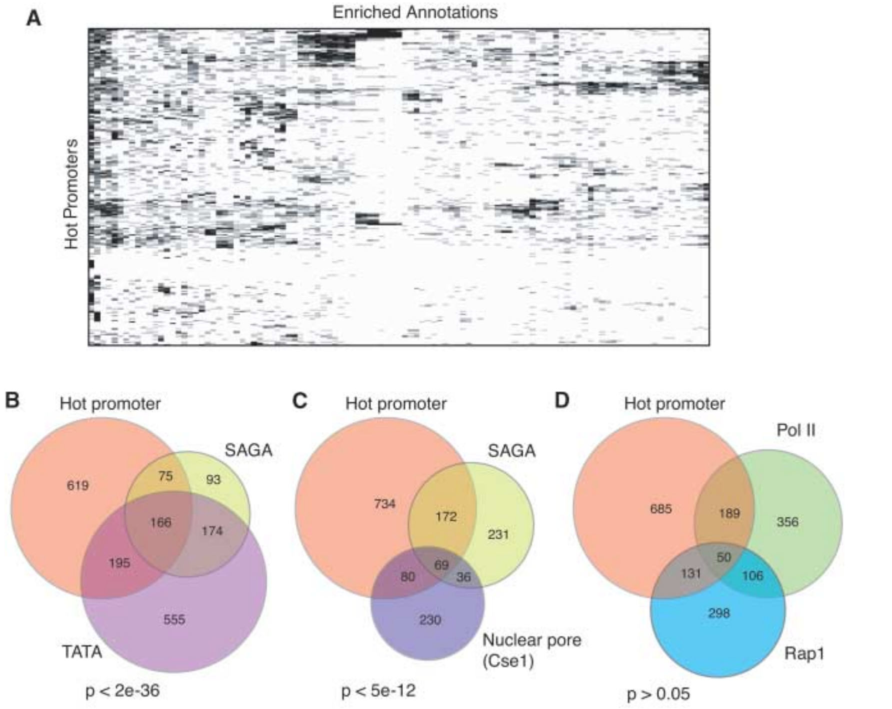


Fig. 4. Rapid turnover at promoters is associated with multiple partially overlapping features. (**A**) Hot promoters were tested for significantly enriched ($p < 10^{-7}$) annotations. Cluster diagram shows hot promoters as rows, annotations (table S6 and fig. S11) as columns. Black bars indicate positive annotations for a given promoter. (**B** to **D**) Overlap between hot promoters and pairs of enriched annotations. *P* value shows significance of overlap between pairs of annotations, given the extent of their overlap with hot promoters (hypergeometric distribution). SAGA-dominated genes are enriched for TATA-containing promoters (B) and are moderately correlated with Cse1-bound genes (C), whereas promoters with Rap1 sites are not enriched upstream of genes exhibiting high Pol II levels in our experiment (D).

reflect multiple, partially overlapping mechanisms for rapid promoter turnover (Fig. 4, B to D). Some enrichments suggest clear hypotheses about the mechanism for rapid turnover (e.g., rapid histone replacement at Swi/snf-regulated promoters may well be a consequence of Swi/snf action), whereas other enrichments are less illuminating (e.g., what causes rapid replacement at nuclear pores?).

Many features of hot nucleosomes (including Htz1, tRNA genes, nuclear pore association, and Rap1 and Reb1 sites) are associated with boundaries that block heterochromatin spreading in yeast (*21*–*24*). How do boundaries block lateral spreading (*25*) of chromatin states? Suggested mechanisms include long gaps between nucleosomes, or recruited acetylases that compete with spreading deacetylation (*26*, *27*). The rapid H3 replacement at boundary-associated regions suggests an alternative hypothesis: that constant replacement of nucleosomes serves to erase a laterally spreading chromatin domain before it spreads any further (fig. S12). To investigate the role of Htz1 (whose role in boundary function is poorly understood) in histone replacement, we measured Flag-H3 incorporation in *htz1*Δ mutants, finding globally slowed H3 incorporation but few locus-specific effects (*14*). Further experiments will be required to untangle this relationship and to evaluate the role of rapid turnover at chromatin boundaries.

We have measured H3 replacement rates throughout the yeast genome, finding that nucleosomes over coding regions are replaced at high transcription rates, although most turnover occurs over promoters and small RNA genes. What function is served by histone replacement at promoters? Rapid turnover could transiently expose occluded transcription factor binding sites or it could ensure, by erasure of promoter chromatin marks, that transcriptional reinitiation occurs only in the continuing presence of an activating stimulus. Whatever the function, one important implication is that steady-state localization studies of histone marks could be confounded by dilution with histones carrying the average modification levels of the free histone pool, making dynamic or genetic studies key to deciphering any instructive roles of histone marks in transcriptional control. Finally, rapid turnover occurs at chromatin boundaries [see also (*28*)]. We propose that erasure of histone marks (or associated proteins) by rapid turnover delimits the spread of chromatin states. We further speculate that the widespread histone turnover at promoters throughout the compact yeast genome could serve, in a sense, to "expand" the genome by preventing chromatin states of adjacent genes from affecting each other.

References and Notes

1. B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, S. L. Schreiber, *Genome Biol.* **5**, R62 (2004).
2. H. Boeger, J. Griesenbeck, J. S. Strattan, R. D. Kornberg, *Mol. Cell* **11**, 1587 (2003).
3. C. K. Lee, Y. Shibata, B. Rao, B. D. Strahl, J. D. Lieb, *Nat. Genet.* **36**, 900 (2004).
4. H. Reinke, W. Horz, *Mol. Cell* **11**, 1599 (2003).
5. U. J. Schermer, P. Korber, W. Horz, *Mol. Cell* **19**, 279 (2005).
6. K. Ahmad, S. Henikoff, *Mol. Cell* **9**, 1191 (2002).
7. Y. Mito, J. G. Henikoff, S. Henikoff, *Nat. Genet.* **37**, 1090 (2005).
8. M. A. Schwabish, K. Struhl, *Mol. Cell. Biol.* **24**, 10111 (2004).
9. A. Kristjuhan, J. Q. Svejstrup, *EMBO J.* **23**, 4243 (2004).
10. C. Thiriet, J. J. Hayes, *Genes Dev.* **19**, 677 (2005).
11. J. Linger, J. K. Tyler, *Eukaryot. Cell* **5**, 1780 (2006).
12. C. L. Liu *et al.*, *PLoS Biol.* **3**, e328 (2005).
13. G. C. Yuan *et al.*, *Science* **309**, 626 (2005).
14. Materials and methods are available as supporting material on *Science* Online.
15. Genomic turnover rates can be viewed at the University of California, Santa Cruz, Genome Browser on *S. cerevisiae*; http://compbio.cs.huji.ac.il/Turnover
16. B. Guillemette *et al.*, *PLoS Biol.* **3**, e384 (2005).
17. D. K. Pokholok *et al.*, *Cell* **122**, 517 (2005).
18. R. M. Raisner *et al.*, *Cell* **123**, 233 (2005).
19. K. Ahmad, S. Henikoff, *Proc. Natl. Acad. Sci. U.S.A.* **99**, (Suppl. 4), 16477 (2002).
20. J. M. Casolari *et al.*, *Cell* **117**, 427 (2004).
21. D. Donze, C. R. Adams, J. Rine, R. T. Kamakaka, *Genes Dev.* **13**, 698 (1999).
22. K. Ishii, G. Arib, C. Lin, G. Van Houwe, U. K. Laemmli, *Cell* **109**, 551 (2002).
23. M. D. Meneghini, M. Wu, H. D. Madhani, *Cell* **112**, 725 (2003).
24. Q. Yu *et al.*, *Nucleic Acids Res.* **31**, 1224 (2003).
25. L. N. Rusche, A. L. Kirchmaier, J. Rine, *Annu. Rev. Biochem.* **72**, 481 (2003).
26. X. Bi, J. R. Broach, *Curr. Opin. Genet. Dev.* **11**, 199 (2001).
27. Y. H. Chiu, Q. Yu, J. J. Sandmeier, X. Bi, *Genetics* **165**, 115 (2003).
28. Y. Mito *et al.*, *Science* **315**, 1408 (2007).

29. We thank K. Ahmad, N. Francis, A. Gasch, N. Habib, A. Jaimovich, R. Kupferman, H. Margalit, and I. Wapinski for critical reading of the manuscript. We thank P. Korber for the generous gift of the USY6 strain. O.J.R. is supported in part by a Career Award in Biomedical Sciences from the Burroughs Wellcome Fund. This research was supported by grants to O.J.R., S.B., and N.F. from the National Institute of General Medical Sciences, NIH; to O.J.R. from the Human Frontiers Science Program; and to N.F. from the Israeli Science Foundation. O.J.R. designed the experiments, and M.F.D. carried them out. S.B. designed, and M.K. carried out, Pol II chromatin immunoprecipitation. T.K., N.F., and O.J.R. analyzed the data. O.J.R. and N.F. wrote the paper.

Supporting Online Material

www.sciencemag.org/cgi/content/full/315/5817/1405/DC1
Materials and Methods
Figs. S1 to S18
Tables S1 to S6

18 August 2006; accepted 6 February 2007
10.1126/science.1134053

Histone Replacement Marks the Boundaries of cis-Regulatory Domains

Yoshiko Mito,[1,2] Jorja G. Henikoff,[1] Steven Henikoff[1,3]*

Cellular memory is maintained at homeotic genes by cis-regulatory elements whose mechanism of action is unknown. We have examined chromatin at *Drosophila* homeotic gene clusters by measuring, at high resolution, levels of histone replacement and nucleosome occupancy. Homeotic gene clusters display conspicuous peaks of histone replacement at boundaries of cis-regulatory domains superimposed over broad regions of low replacement. Peaks of histone replacement closely correspond to nuclease-hypersensitive sites, binding sites for Polycomb and trithorax group proteins, and sites of nucleosome depletion. Our results suggest the existence of a continuous process that disrupts nucleosomes and maintains accessibility of cis-regulatory elements.

Chromatin can be differentiated by the replication-independent replacement of one histone variant with another (*1*). For example, histone H3.3 is deposited throughout the cell cycle, replacing H3 that is deposited during replication (*2*–*4*). Unlike replication-coupled assembly of H3, which occurs in gaps between old nucleosomes on daughter helices, the insertion of H3.3 is preceded by disruption of preexisting histones during transcription and other active processes (*3*, *5*). We have previously shown that H3.3 replacement profiles resemble those for RNA polymerase II (*2*), which suggests that gradual replacement of H3.3 occurs in the wake of transiting polymerase to repair disrupted chromatin (*1*). Here, we ask whether histone replacement and nucleosome occupancy are also distinctive at cis-regulatory elements.

Log-phase *Drosophila melanogaster* S2 cells were induced to produce biotin-tagged H3.3 for two or three cell cycles (*2*). DNA was extracted from streptavidin pull-down assay and input material, labeled with Cy3 and Cy5 dyes, and cohybridized to microarrays. To provide a standard, we profiled biotin-tagged H3-containing chromatin in parallel. Analysis of H3.3/H3 levels over the entire *3*R chromosome arm revealed that the ~350-kb bithorax complex (BX-C) region displays the lowest H3.3/H3 ratio of any region of comparable size on *3*R, and the Antennapedia

[1] Basic Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98109, USA. [2] Molecular and Cellular Biology Program, University of Washington, Seattle, WA 98195, USA. [3] Howard Hughes Medical Institute, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA.

*To whom correspondence should be addressed. E-mail: steveh@fhcrc.org

Discussion and Conclusions

During my PhD studies, I aimed to gain further understanding of the basic principles of transcriptional regulation in eukaryotes. The regulation of gene expression is crucial for the proper cell functions and is achieved by a complex network of regulators acting at various levels. Throughout my studies I explored the various mechanisms involved in regulation of gene expression. I began my journey from the near atomic level of protein-DNA interactions (Chapter 1). I continued by developing sophisticated mathematical models for transcription factor binding sites (Barash *et al.*, 2003) and their genomic applications (Barash *et al.*, 2005) (not included in this dissertation). I then developed computational algorithms to explore the combinatorial interactions between transcription factors (Chapter 2). Finally, I focused on the packaging of DNA and its temporal dynamics (Chapters 3 and 4).

This discussion is divided into three parts: The first part concerns Chapter 1, and reviews various aspects of our manuscript from 2005, as well as additional updates in the field. The second part deals with strategies to analyze more complex regulatory systems, in terms of combinatorial interactions between transcription factors, as presented in Chapter 2. Finally, I discuss the role of higher order aspects of transcriptional regulation, as reflected by the studies presented in Chapters 3 and 4.

Ab initio prediction of transcription factor targets using structural knowledge

The ever growing availability of genomic sequences, together with the limited rate in which proteins can be experimentally characterized, has created a huge number of novel proteins, known by their sequence only. This regards also transcription factors, which are identified as such by their sequence only, but for which there is no information about the binding sites or target genes. To bridge this gap, I developed a novel computational algorithm, which uses structural information in order to predict the DNA motifs (and putative binding sites) of transcription factors from their sequence. The concept of predicting transcription factor binding sites based on structural knowledge was presented several years prior to the publication of our work (Kono and Sarai, 1999; Mandel-Gutfreund *et al.*, 2001; Benos *et al.*, 2002; Endres *et al.*, 2004; Havranek *et al.*, 2004). It relies on two well studied concepts in structural biology. First is the use of the 3D structure of one protein as a *structural template* for other proteins from the same family. This dogma has been employed for many years in various structure prediction algorithms (Lemer *et al.*, 1995; Rost and Sander, 1996; Rost *et al.*, 1997; Karplus *et al.*, 1999; Skolnick and Fetrow, 2000). Second, the energetic preference of protein-DNA interactions are general, and arise from the physicochemical properties of amino acids and nucleotides (von Hippel, 1994; Mandel-Gutfreund and Margalit, 1998; Luscombe *et al.*, 2001; Mandel-Gutfreund *et al.*, 2001). When put together, these concepts allow to use known protein-DNA structures to analyze the sequence of novel proteins, identify which residues will interact with the DNA, and then predict their binding preferences. Additional studies focused on specific structural families, and showed that a particular amino acid may have different binding preferences depending on its positional context (Choo and Klug, 1994; Choo and Klug, 1994; Kono and Sarai, 1999).

My study contributed to this line of studies by two means. First, the previous studies relied on solved protein-DNA complexes to estimate the amino acid-nucleotide binding preferences. Those structures provide details at the atomic level, but their number is limited. For example, when we started our study, the largest structural compilation of protein-DNA interactions, by Mandel-Gutfreund and Margalit (1998), was based upon 53 solved complexes, and surveyed 218 nucleotide-amino acid interactions in total (major groove interactions only). To overcome this limitation, I developed a computational algorithm to expand our input data and also process sequence pairs of transcription factors and their cognate natural binding sites. Such data are much easier to obtain. For example, by extracting the sequences of experimentally verified binding sites from the TRANSFAC database (Wingender *et al.*, 2001), I estimated the recognition preferences based on 5367 interactions, from 455 protein-DNA pairs. This allowed me to estimate the binding preferences using 25 times more data. Second, this wealth of data (due to our sequence-based approach) allowed me to estimate more specialized, context-specific binding preferences. While previous studies estimated the amino acid-nucleotide preferences based on protein-DNA interactions from a variety of structural families, I focused on a single family (the C2H2 zinc finger domain) and learned a different set of binding preferences for every key position along the protein's DNA-binding domain. As I showed, while some binding preferences are general (e.g. lysine's tendency to bind guanine), others were position-specific (e.g. the tendency of phenylalanine at position 2 of the zinc finger DNA-binding domain to bind cytosine) (Kaplan *et al.*, 2005). A similar approach was simultaneously taken by Benos *et al.,* (2001), using artificial SELEX data. As I showed using held-out test data (cross validation), both the use of natural sequence data, and the estimation of position-specific binding preferences, dramatically contributed to the accuracy of our predictions (shown in Figure 5C of Chapter 1, or in Supp Tables 6,7 of Kaplan *et al.*, (2005)). I then applied this algorithm to predict the DNA motifs of 29 C2H2 zinc finger proteins in *Drosophila melanogaster*, identified the putative target genes of each, and analyzed gene expression data and genomic annotations to estimate the function and activity levels of each factor (Chapter 1).

Since the publication of our manuscript, this concept was further developed by additional studies. Some focused on developing better models to approximate the physical energy in protein-DNA interactions (Morozov *et al.*, 2005; Moroni *et al.*, 2007; Siggers and Honig, 2007), while others focused on the computational parts of the models (Cho *et al.*, 2008). The type of data used in training the model is also of great importance. Benos *et al.,* (2001) used artificial SELEX data to characterize protein-DNA interactions. While this allowed the exploitation of additional protein-DNA interactions, my results suggest that synthetic data (including SELEX) lead to biased results when applied to genomic studies (Chapter 1). Technological advances from the Bulyk laboratory at Harvard promise to provide high-throughput identification of quasi-natural binding sites using protein binding microarrays (PBMs). These arrays span a huge range of short dsDNA sequence, to which epitope-tagged TFs bind preferentially. Such a technology opens an opportunity for large-scale characterization of the binding preferences of a purified TF in a single day. In addition, recent works concentrated on expanding the approach I described to additional structural families, including the basic leucine zipper (bZip, Grigoryan and Keating, 2006), the helix-turn-helix (Moroni *et al.*, 2007), and homeodomain (Liu and Bader, 2007).

This study provided the first automated characterization of novel transcription factors, using their sequence-based predicted DNA motifs to identify target genes on a genome-wide scale. As such, it contributed greatly to the scientific community. As every model, it includes some caveats and limitations, which should be addressed in future extensions. The main caveat of my algorithm relates to the rigid architecture of protein-DNA interactions it relies on. In the structural domain we analyzed, the canonical binding model (Elrod-Erickson *et al.*, 1998) which implies the same architecture of amino acid-nucleotide contacts between various C2H2 zinc finger protein and their DNA binding sites, usually holds. I showed how other C2H2 proteins (e.g. GLI), where our DNA motif predictions might not hold due to a rather different binding architecture (Pavletich and Pabo, 1993), can be identified using a probabilistic computational classifier. To further extend our algorithm to handle additional, more flexible, DNA binding domains, including Helix-Turn-Helix (HLH) and basic leucine zipper (bZIP) domains, we must allow a larger degree of variability in the specific protein-DNA contacts. Energy-based approaches were proven useful in offering this flexibility for protein-protein docking problems (Baker and Sali, 2001; Gray *et al.*, 2003; Schueler-Furman *et al.*, 2005), with promising applications to protein-DNA interactions (Endres *et al.*, 2004; Havranek *et al.*, 2004; Endres and Wingreen, 2006; Siggers and Honig, 2007).

Another possible issue is posed by our limited knowledge on modeling transcription factors binding sites. Most approaches that describe the sequences bound by some TF use a probabilistic model (DNA motif) and inherently assume that positions within the binding sites are independent of each other. This independence (or additivity) assumption was shown to hold for C2H2 zinc finger domains (Benos *et al.*, 2002; Bulyk *et al.*, 2002), although it might not be the case for other structural families. In a previous study, we have shown these inner-dependencies to exist in the binding sites of many transcription factors, from various structural families (Barash *et al.*, 2003). I believe that such dependencies should be incorporated into future sequence-based methods for predicting transcription factor binding sites. For example, using the complex mathematical models developed by me and others (Agarwal and Bafna, 1998; Barash *et al.*, 2003; King and Roth, 2003; Zhou and Wong, 2004; Ben-Gal *et al.*, 2005; Sharon and Segal, 2007).

Complex regulatory systems

Eukaryotic cells regulate gene expression using a complex network of signaling pathways, transcription factors and promoters. Studies, such as the one I presented in Chapter 1, facilitate the reconstruction of the transcriptional regulatory map of direct protein-DNA interactions, but are of limited help in understanding how transcription factors interact to control the expression of genes. For example, the binding of two transcription factors to the same promoter may result in a transcriptional outcome that equals the sum of their individual effects (hence, act independently). Alternatively, they might interact to further activate the gene (e.g. by stabilizing each other) or to achieve a weaker transcriptional outcome than expected.

In Chapter 2, we developed and applied a novel strategy which combined genetic and computational tools to gain insights into the structure and function of the well-studied transcriptional network that controls the cellular response to osmotic stress. This was done by integrating gene expression data from various mutant strains, to quantitatively estimate the contribution of single and pairs of TFs to the expression of target genes. We found that *transcriptional calculus*, where the expression level of a gene is presented as the sum of TF-specific components, is surprisingly accurate. The transcriptional response to osmotic stress of ~90% of the HOG genes we analyzed could have been neatly dissected into the sum of contributions by specific transcription factors (Sko1, Hot1 and Msn2/4), with only few genes presenting a statistically significant residual (Chapter 2). We also found that combinatorial regulation of gene expression, for example by binary *AND* or *OR* gates, occurs for most genes (Chapter 2). In addition to the expression-based analysis, we analyzed the regulatory system from two supplementary perspectives. First, we scrutinized the DNA regulatory sequence of genes, characterizing the DNA motif of transcription factors and identifying their putative target genes. Second, we pinpointed the *in vivo* binding positions of these TFs along the genome, using a novel model-based algorithm to analyze a series of chromatin immunoprecipitation experiments, followed by hybridization to a densely tiled DNA microarray. These complementary studies allowed us, for the first time, to gain insights into the mechanisms of a transcriptional regulation - starting from the DNA sequence level (as reflected in the computational analysis of binding sites), through the transcriptional mechanism level (reflected by our measurements of *in vivo* physical binding), to their final outcome (reflected by changes in the expression levels of target genes). We found that the majority (~60%) of regulatory regions that contain the recognition sequence of a TF, were also physically bound by it in a statistically significant manner. From those bound genes, ~75% presented a significant transcriptional response 20 minutes after the induction of hyperosmotic shock. These high numbers imply a rather comprehensive understanding of the transcriptional mechanisms involved in the majority of HOG pathway genes. On the other hand, these numbers depict significant discrepancies between the target genes identified using sequence, physical binding and expression. These differences emphasize the overall complexity of eukaryotic transcriptional control, and suggest the role of higher order regulatory mechanisms, such as chromatin.

Previous analyses of the osmotic stress response in yeast led to a coarse-grained transcriptional model of the HOG pathway, where the Hog1 kinase regulates gene expression through the activation of the general stress regulator Msn2/4 as well as additional transcription factors (Rep *et al.*, 2000; Hohmann *et al.*, 2007). While microarrays were previously used to portray the whole-genome transcriptional response of yeast cells to hyper-osmotic stress (Roberts *et al.*, 2000; O'Rourke and Herskowitz, 2004), we extended this strategy by measuring gene expression data from a series of yeast strains, including the wild type strain, mutant strains lacking HOG-related TFs (hog1Δ, msn2/4Δ, sko1Δ, hot1Δ), and strains lacking combinations of TFs (e.g. hog1Δmsn2/4Δ). This allowed us to computationally dissect the transcriptional response of genes into modular components corresponding to the contribution of single or pairs of) transcription factors to the expression levels of genes. Our study can be also viewed from an epistatic perspective (Cordell, 2002), where the interaction between two genes is estimated by comparing some phenotype between the wt strain, the two knockout strains, and the double-knockout strain (Avery and Wasserman, 1992; Van Driessche *et al.*, 2005; Collins *et al.*, 2006). We extend this classic view by simultaneously measuring the expression levels of multiple genes, then treating each one as a different phenotype. As we have shown, the epistatic interaction between the two master regulators of the HOG pathway (Hog1 and Msn2/4) depends on the gene used as phenotype.

The genetic/computational approach presented in Chapter 2 yielded insights as to the propagation of external stimuli to the nucleus (through Hog1 and Msn2/4), and their translation to transcriptional instructions. Such approach cannot be applied to any transcriptional network, as it requires a fair amount of prior knowledge. This includes the factors participating in this network whose transcriptional components should be measured using deletion strains. Moreover, our algorithm, like previous epistatic studies, is capable of identifying the cooperative response of transcription factors, but fails to determine where such interactions occur. For example, genes with AND-like gates between Hog1 and Msn2/4 can be interpreted in two distinct manners. One explanation reflects a combinatorial interaction at the promoter level, where both factors are required prior to transcriptional activation. Alternatively, such AND-like interactions can be explained at the signaling level, where one factor (Hog1) is related to nuclear localization of the second (Msn2/4), which in turn activates the gene. In both cases, the presence of a single factor would not suffice to activate the gene. To resolve this, we also identified the *in vivo* physical binding locations of Hog1-regulated transcription factors, and found that these two explanations are not mutually exclusive. While for some genes, the higher nuclear concentration of Msn2/4 is predominant, others are controlled by combinatorial regulation at the promoter level. This transcriptional network is a fine example of achieving a smooth range of transcriptional behaviors, using only 4 main regulators (Hog1, Msn2/4, Sko1 and Hot1). Further investigation is needed to understand how exactly this combinatorial regulation is encoded in the promoter regions of HOG genes.

In this study, I developed a computational model-based algorithm to analyze high-resolution ChIP data. Such algorithm were previously proposed in few studies (Buck *et al.*, 2005; Gibbons *et al.*, 2005; Kim *et al.*, 2005; Li *et al.*, 2005; Qi *et al.*, 2006). Most these studies naively identified ChIP-enriched genomic regions (Buck *et al.*, 2005; Gibbons *et al.*, 2005; Kim *et al.*, 2005; Li *et al.*, 2005), whereas our method is based on a concrete computational model to estimates the expected shape of enrichments around a binding event (Chapter 2). This model-based approach allowed us to integrate data from neighboring probes and identify the position of physical binding events and their affinity (or height) with significant robustness to measurement noise. The recent development of next-generation sequencing platforms, such as Illumina's Solexa sequencers (Barski *et al.*, 2007; Johnson *et al.*, 2007; Robertson *et al.*, 2007), or the Roche/454 Genome Sequencers (Margulies *et al.*, 2005) paved the way to ChIP-sequencing assays. Here, following chromatin immunoprecipitation, the purified DNA fragments are being sequenced rather than hybridized to a microarray as in ChIP-chip assays. The ChIP-sequencing platform offers an accurate assay to identify *in vivo* binding events at a competitive expense (at least for high eukaryotes, where a single ChIP-chip assay requires several micoarrays due to the size of the genome). Published Solexa-based ChIP-sequencing data (Johnson *et al.*, 2007; Mikkelsen *et al.*, 2007; Robertson *et al.*, 2007) bears a great similarity to ChIP-chip results, suggesting that my model-based algorithm tools can be easily extended to handle such data.

The discrepancies we identified between the predicted target genes of transcription factors based on sequence motifs, on *in vivo* ChIP analysis, or on expression arrays, shed light onto the mechanistic principles of transcriptional regulation, and lead to fascinating future directions. Obviously, some discrepancies might arise from threshold definitions. As previous studies suggested, there are no clear definitions of activation thresholds. While the few genes constantly bound by transcription factors will have a strong transcriptional effect, thousands of additional genomic loci are also bound at low-affinities by transcription factors, sometimes resulting with transcriptional activation (Tanay, 2006; Li *et al.*, 2008). For consistency reasons, we eventually set the thresholds for all three types of annotations to allow exactly 5% of false positive calls (where we expression-based transcriptional groups are used as "truth"). Additional reasons for the observed discrepancies might arise from experimental noise, or from our limited abilities in modeling binding sites. More interesting are the biological mechanisms capable of explaining such discrepancies, such as the role of chromatin in transcriptional regulation. Prior studies demonstrated how nucleosomes can serve to modify the occupancy of the DNA at the transcription start site (Cosma *et al.*, 1999) or at transcription factor binding sites (Narlikar *et al.*, 2007). A fascinating extension of our work relates the ability of nucleosome positioning data to explain the experimental inconsistencies. For example, high nucleosomal occupancy which occludes unbound motifs (i.e. bona fide binding sites with no supporting ChIP data).

Higher order aspects of transcriptional regulation

The studies I presented in Chapters 3 & 4 (Liu *et al.*, 2005; Dion *et al.*, 2007) were pioneering in characterizing the chromatin landscape of a eukaryotic genome at a single-nucleosome resolution.

In the first paper, we used published data regarding the position of nucleosome across the budding yeast genome (Yuan *et al.*, 2005), and mapped their acetylation and methylation patterns. My computational analysis showed that some phenomena are limited to specific nucleosomes (such as the punctual deacetylation of the two nucleosomes surrounding transcription start site sites), whereas others modifications tend to appear in gradients across the coding regions of transcribed genes (e.g., the methylation patterns of histone H3 lysine 4 (H3K4)). By analyzing the modification data together with the genomic positions of nucleosomes, I characterized the stereotypical modification patterns of various nucleosomes types (e.g., nucleosomes over promoters of highly-transcribed genes). I also showed that most of the epigenetic information stored in the chromatin can be summarized into two simple overlaid patterns, and that no discrete combinatorial code is found. Although histone modifications were measured before, our study was the first to characterize the epigenetic landscape at a single nucleosome resolution. Such resolution is crucial in order to relate the specific modification patterns to genomic loci, or to test the similarity between adjacent nucleosomes. For example, this high-resolution allowed us to observe a unique pattern of modification on the first nucleosome upstream to the transcription start site, or to identify specific nucleosomes at the boundaries of two genes. A parallel study by the Young laboratory at the Whitehead Institute (Pokholok *et al.*, 2005) used genome-wide arrays of lower resolution to characterize a rather different set of histone modifications. Nonetheless, their study also observed some of our main phenomena (modifications correlated with transcription and the decaying patterns of acetylation and methylation patterns across coding regions). Our observations were further confirmed in higher eukaryotes, including plant, fly, mouse and human using chromatin immunoprecipitation followed by hybridization (Schubeler *et al.*, 2004; Bernstein *et al.*, 2005; Kim *et al.*, 2005; Roh *et al.*, 2005; Roh *et al.*, 2006; Bernstein *et al.*, 2007; Koch *et al.*, 2007; Zhang, 2008) or deep sequencing (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007).

Previous hypotheses suggested that the histone modification can be interpreted by some form of a “histone code” (Strahl and Allis, 2000). I have employed computational and statistical analysis tools to identify probabilistic dependencies between different histone positions (in case of a combinatorial code), as well as correlations between the modifications to external data, such as gene expression levels, localization of transcription and chromatin factors involved in regulating these genes, the modifications of adjacent nucleosomes, etc. Yet, I could not identify a clear, discrete code to interpret the histone modifications (Liu *et al.*, 2005), nor was it found by the other studies mentioned. We should take extra caution before dismissing the histone code hypothesis. We should keep in mind that the current studies mapped only the modifications for which reliable antibodies existed, and do not cover the full arsenal of modifications. A discrete code may still exist over the unmapped modifications. In addition, our data were collected in a single-nucleosome resolution from a population of yeast cells. Theoretically it is possible that our measurements average over the values of specific cells, concealing possible differences. This caveat could be overcome by future technological advances that will measure the modifications at a single molecule level.

In the second paper (presented in Chapter 4), we measured the replacement rates of histone H3 across the yeast genome. We were intrigued by the molecular mechanisms that allow for a wide range of chromatin plasticity. On the one hand, cells maintain the chromatin state over generations (e.g. in silenced genomic loci). On the other hand, rapid changes in the histone modification patterns (as part of transcriptional reprogramming) occur only minutes after the induction of a novel stimulus. Such changes can be achieved by recruiting chromatin modifying enzymes to directly "fix" the old nucleosomes. Alternatively, it is possible to replace the old nucleosomes altogether by newer ones. While the first approach was previously addressed (by both ChIP-chip and gene expression studies (Robyr *et al.*, 2002; Robert *et al.*, 2004; Pokholok *et al.*, 2005)), genome-scale applications of nucleosome turnover were somewhat overlooked. We therefore used a yeast mutant strain bearing a galactose-induced tagged-histones (Schermer *et al.*, 2005), to estimate their genome-wide locus-specific integration, based on a time-series of ChIP-chip measurements. To analyze these data, I developed a mathematical model, and translate the array-based time-series genome-wide measurements into actual rates (in minutes). We found turnover rates over coding regions to correlate with the transcription levels of underlying genes. We also found that nucleosomes over regulatory regions are replaced in much higher rates than those over coding regions. These two phenomena were further supported by parallel studies in budding yeast (Jamai *et al.*, 2007; Rufiange *et al.*, 2007) and fruit fly (Mito *et al.*, 2007). The interpretation of the first phenomenon is rather straightforward. High transcription levels of genes involve multiple passages of the transcriptional machinery and polymerase II. This high activity over the DNA, was previously shown to cause nucleosomal instability (Kristjuhan and Svejstrup, 2004; Schwabish and Struhl, 2004), which may partially explain the high turnover rates over coding regions of active genes. Less clear, however, is the relation between high turnover rates over regulatory regions and transcriptional control. We suggested three possible explanations (which are not necessarily mutually exclusive) to this striking phenomenon. First, the high turnover rate over regulatory regions may serve as a regulatory mechanism per se, by exposing occluded transcription factor binding sites to their regulators. Alternatively, given the unique modification pattern of promoter nucleosomes and its relation to transcription, high turnover rates may serve to constantly reset the transcriptional program ensuring that transcriptional re-initiation only occurs during activating stimuli. Such tight regulation is often obtained by additional means, e.g., by a well orchestrated series of events prior to transcription initiation, which is followed by a rapid disassembly. Well studied examples include the promoter of human interferon β (Agresti and Bianchi, 2003) and the cell-cycle regulation of HO gene in yeast (Cosma *et al.*, 1999). A third explanation views dynamic nucleosomes as chromatin barriers, suggesting they prevent the spreading of chromatin states across the regulatory regions of adjacent genes (Noma *et al.*, 2001). Our two studies relate the chromatin landscape around genes to their expression levels, and further emphasize the importance of chromatin in transcriptional regulation, together with the actual DNA sequence.

Some caveats might affect our estimations of turnover rates. First, our estimations rely on a transgenic system where histone H3 is over-expressed. Theoretically, this over-abundance might cause a bias toward higher turnover rates. Replacement of H3, however, causes full nucleosomal disassembly, suggesting that this bias is limited due to stoichiometric constraints (hence, normal activation levels of the other histones). More worrying are experimental biases due to non-linear response of DNA microarrays. To control for this effect, we repeated our measurements using both a printed array (Liu *et al.*, 2005; Yuan *et al.*, 2005) and a commercial one (Guillemette *et al.*, 2005; Pokholok *et al.*, 2005). We observed excellent correlation between the turnover rates estimated by the two platforms (Dion *et al.*, 2007), although our measurements indicated a “quenching” effect (limited dynamic range) in the printed arrays. These possible artifacts indicate that the turnover rates I calculated should not be literally interpreted as the expected number of turnover events per nucleosome per minute. Our *qualitative* results, however, are valid, and supported by parallel studies that used other tagged histones or experimental platforms to measure nucleosome exchange (Linger and Tyler, 2006; Jamai *et al.*, 2007).

So far I showed how the modification patterns of nucleosomes (Chapter 3) are well connected to their turnover rates (Chapter 4), as well as to the position of nearby genes and their expression levels. This opens the way for several exciting future directions, related to transcription, nucleosome exchange, and histone modifications. The first question involves the dynamics of histone modification and nucleosome positioning following transcriptional reprogramming. To this end, I analyzed published gene expression data, measuring the transcriptional response of yeast cells to various types of stress (Gasch *et al.*, 2000), and identified the conditions which cause the strongest transcriptional change. Based on this, together with the Rando laboratory, we carefully selected four modification sites and designed a time-series of measurements. This experiment will describe the temporal change in histone modifications following induction of stress. These experiments are about to be completed, promising fascinating insights into the dynamics of DNA packaging and their relation to transcriptional regulation.

Second, to probe the *causal connection* between histone modification and transcription, we designed a parallel set of experiments, where the endogenous RNA polymerase II was replaced by a temperature-sensitive variant. This allows to "shut-off" the activity of the polymerase by shifting the cells to high temperature. Comparison of the histone modifications in this strain to a wild-type strain (in the same environmental conditions), will reveal if and to which extent does transcriptional activity control histone modification patterns.

Third, I am interested in studying the crosstalk between histone modifications and nucleosomes turnover. The recently discovered acetylation of Lysine 56 at histone H3 (H3K56) encompasses a promising opportunity for such questions. Acetylation of H3K56 was recently shown to mark newly incorporated nucleosomes during DNA replication (Hyland *et al.*, 2005; Masumoto *et al.*, 2005; Xu *et al.*, 2005; Ozdemir *et al.*, 2006; Rufiange *et al.*, 2007). Following our discoveries about replication-independent nucleosome exchange, we decided to pursue this direction and test whether acetylation of H3K56 also relates to *replication-independent* incorporation of nucleosomes. I aim to analyze the acetylation of H3K56 in view of the exact timing of *replication-coupled* exchange of nucleosomes (based on external estimations, Raghuraman *et al.*, 2001; Yabuki *et al.*, 2002), as well as the *replication-independent* exchange rates (based on our estimations, Dion *et al.*, 2007). We have already measured the H3K56 acetylation levels in yeast cells, both in unsynchronized mid-log yeast, and in yeast advancing synchronously through the cell cycle. My preliminary results suggest that about half of the H3K56 acetylation and deacetylation events can be explained by nucleosome exchange. This suggests that the locus-specific recruitment of histone modifying enzymes may play a role in regulating the other half.

Concluding remarks

Understanding gene regulation is crucial for understanding how cells function and how they adjust to changes in the external environment or internal state. This highly complex process involves a stratified network of regulatory layers, including the sequence of DNA and its packaging, as well as higher order mechanisms, such as the organization of chromosomes in the nucleus.

In my thesis, I presented several studies in which I developed and applied sophisticated computationally tools to provide insights into various levels of transcriptional regulation. These studies contributed to the characterization and identification of regulatory sites, to our understanding of combinatorial transcriptional regulation, and to the fine-grained description of the chromatin state, and as a whole present an expedition toward understanding of gene regulation at the systems level.

References

Agarwal, P and Bafna, V (1998). Detecting non-adjoining correlations with signals in DNA. ***Proc. of the 2nd Conf. Research in Computational molecular biology***: 2-8.

Agresti, A and Bianchi, ME (2003). HMGB proteins and gene expression. ***Curr Opin Genet Dev,*** 13(2): 170-8.

Albert, I, Mavrich, TN, Tomsho, LP*, et al.* (2007). Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. ***Nature,*** 446(7135): 572-6.

Alepuz, PM, de Nadal, E, Zapater, M*, et al.* (2003). Osmostress-induced transcription by Hot1 depends on a Hog1-mediated recruitment of the RNA Pol II. ***EMBO J,*** 22(10): 2433-42.

Alepuz, PM, Jovanovic, A, Reiser, V*, et al.* (2001). Stress-induced map kinase Hog1 is part of transcription activation complexes. ***Mol Cell,*** 7(4): 767-77.

Avery, L and Wasserman, S (1992). Ordering gene function: the interpretation of epistasis in regulatory hierarchies. ***Trends Genet,*** 8(9): 312-6.

Bailey, TL and Elkan, C (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. ***Proc Int Conf Intell Syst Mol Biol,*** 2: 28-36.

Bailey, TL and Gribskov, M (1998). Combining evidence using p-values: application to sequence homology searches. ***Bioinformatics,*** 14(1): 48-54.

Baker, D and Sali, A (2001). Protein structure prediction and structural genomics. ***Science,*** 294(5540): 93-6.

Bannister, AJ, Schneider, R, Myers, FA*, et al.* (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. ***J Biol Chem,*** 280(18): 17732-6.

Bannister, AJ, Zegerman, P, Partridge, JF*, et al.* (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. ***Nature,*** 410(6824): 120-4.

Barash, Y, Bejerano, G and Friedman, N (2001). A Simple Hyper-Geometric Approach for Discovering Putative Transcription Factor Binding Sites. ***Algorithms in Bioinformatics: Proc. First International Workshop,*** (2149): 278-293.

Barash, Y, Elidan, G, Friedman, N*, et al.* (2003). Modeling dependencies in Protein-DNA binding sites. ***Proc. of the 7th International Conf. on Research in Computational Molecular Biology***: 28-37.

Barash, Y, Elidan, G, Kaplan, T*, et al.* (2005). CIS: compound importance sampling method for protein-DNA binding site p-value estimation. ***Bioinformatics,*** 21(5): 596-600.

Barrett, CL and Palsson, BO (2006). Iterative reconstruction of transcriptional regulatory networks: An algorithmic approach. ***Plos Computational Biology,*** 2(5): 429-438.

Barski, A, Cuddapah, S, Cui, K*, et al.* (2007). High-resolution profiling of histone methylations in the human genome. ***Cell,*** 129(4): 823-37.

Ben-Gal, I, Shani, A, Gohr, A*, et al.* (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. ***Bioinformatics,*** 21(11): 2657-66.

Benos, PV, Bulyk, ML and Stormo, GD (2002). Additivity in protein-DNA interactions: how good an approximation is it? ***Nucleic Acids Res,*** 30(20): 4442-51.

Benos, PV, Lapedes, AS, Fields, DS*, et al.* (2001). SAMIE: statistical algorithm for modeling interaction energies. ***Pac Symp Biocomput***: 115-26.

Benos, PV, Lapedes, AS and Stormo, GD (2002). Is there a code for protein-DNA recognition? Probab(ilistical)ly. ***Bioessays,*** 24(5): 466-75.

Benos, PV, Lapedes, AS and Stormo, GD (2002). Probabilistic code for DNA recognition by proteins of the EGR family. ***J Mol Biol,*** 323(4): 701-27.

Ben-Tabou de-Leon, S and Davidson, EH (2007). Gene regulation: gene control network in development. ***Annu Rev Biophys Biomol Struct,*** 36: 191.

Berger, SL (2002). Histone modifications in transcriptional regulation. ***Curr Opin Genet Dev,*** 12(2): 142-8.

Bernstein, BE, Humphrey, EL, Erlich, RL*, et al.* (2002). Methylation of histone H3 Lys 4 in coding regions of active genes. ***Proc Natl Acad Sci U S A,*** 99(13): 8695-700.

Bernstein, BE, Kamal, M, Lindblad-Toh, K*, et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. ***Cell,*** 120(2): 169-81.

Bernstein, BE, Liu, CL, Humphrey, EL*, et al.* (2004). Global nucleosome occupancy in yeast. ***Genome Biol,*** 5(9): R62.

Bernstein, BE, Meissner, A and Lander, ES (2007). The mammalian epigenome. ***Cell,*** 128(4): 669-81.

Boeger, H, Griesenbeck, J, Strattan, JS*, et al.* (2003). Nucleosomes unfold completely at a transcriptionally active promoter. ***Mol Cell,*** 11(6): 1587-98.

Boyer, LA, Lee, TI, Cole, MF, *et al.* (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. ***Cell,*** 122(6): 947-56.

Buck, MJ and Lieb, JD (2006). A chromatin-mediated mechanism for specification of conditional transcription factor targets. ***Nat Genet,*** 38(12): 1446-51.

Buck, MJ, Nobel, AB and Lieb, JD (2005). ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. ***Genome Biol,*** 6(11): R97.

Bulyk, ML, Johnson, PL and Church, GM (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. ***Nucleic Acids Res,*** 30(5): 1255-61.

Cairns, BR (2005). Chromatin remodeling complexes: strength in diversity, precision through specialization. ***Curr Opin Genet Dev,*** 15(2): 185-90.

Cawley, S, Bekiranov, S, Ng, HH, *et al.* (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. ***Cell,*** 116(4): 499-509.

Cho, SY, Chung, M, Park, M, *et al.* (2008). ZIFIBI: Prediction of DNA binding sites for zinc finger proteins. ***Biochemical and Biophysical Research Communications,*** 369(3): 845-848.

Choo, Y and Klug, A (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. ***Proc Natl Acad Sci U S A,*** 91(23): 11168-72.

Choo, Y and Klug, A (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. ***Proc Natl Acad Sci U S A,*** 91(23): 11163-7.

Clayton, RA, White, O, Ketchum, KA, *et al.* (1997). The first genome from the third domain of life. ***Nature,*** 387(6632): 459-62.

Collins, SR, Schuldiner, M, Krogan, NJ, *et al.* (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. ***Genome Biol,*** 7(7): R63.

Cordell, HJ (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. ***Hum Mol Genet,*** 11(20): 2463-8.

Cosma, MP, Tanaka, T and Nasmyth, K (1999). Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. ***Cell,*** 97(3): 299-311.

Davidson, EH (2006). The regulatory genome : gene regulatory networks in development and evolution. Amsterdam ; Boston, Elsevier/Academic Press.

Davidson, EH, Rast, JP, Oliveri, P, *et al.* (2002). A genomic regulatory network for development. ***Science,*** 295(5560): 1669-78.

de Nadal, E, Alepuz, PM and Posas, F (2002). Dealing with osmostress through MAP kinase activation. ***EMBO Rep,*** 3(8): 735-40.

de Nadal, E, Casadome, L and Posas, F (2003). Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. ***Mol Cell Biol,*** 23(1): 229-37.

DeGroot, MH and Schervish, MJ (2002). Probability and statistics. Boston, Addison-Wesley.

DeRisi, JL, Iyer, VR and Brown, PO (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. ***Science,*** 278(5338): 680-6.

Dhalluin, C, Carlson, JE, Zeng, L, *et al.* (1999). Structure and ligand of a histone acetyltransferase bromodomain. ***Nature,*** 399(6735): 491-6.

Dion, MF, Kaplan, T, Kim, M, *et al.* (2007). Dynamics of replication-independent histone turnover in budding yeast. ***Science,*** 315(5817): 1405-1408.

Elrod-Erickson, M, Benson, TE and Pabo, CO (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. ***Structure,*** 6(4): 451-64.

Endres, RG, Schulthess, TC and Wingreen, NS (2004). Toward an atomistic model for predicting transcription-factor binding sites. ***Proteins,*** 57(2): 262-8.

Endres, RG and Wingreen, NS (2006). Weight matrices for protein-DNA binding sites from a single co-crystal structure. ***Phys Rev E Stat Nonlin Soft Matter Phys,*** 73(6 Pt 1): 061921.

Fischle, W, Tseng, BS, Dormann, HL, *et al.* (2005). Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. ***Nature,*** 438(7071): 1116-22.

Gasch, AP, Spellman, PT, Kao, CM, *et al.* (2000). Genomic expression programs in the response of yeast cells to environmental changes. ***Mol Biol Cell,*** 11(12): 4241-57.

Gibbons, FD, Proft, M, Struhl, K, *et al.* (2005). Chipper: discovering transcription-factor targets from chromatin immunoprecipitation microarrays using variance stabilization. ***Genome Biol,*** 6(11): R96.

Gordon, DB, Nekludova, L, McCallum, S, *et al.* (2005). TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. ***Bioinformatics,*** 21(14): 3164-5.

Gray, JJ, Moughon, S, Wang, C, *et al.* (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. ***J Mol Biol,*** 331(1): 281-99.

Grewal, SI and Rice, JC (2004). Regulation of heterochromatin by histone methylation and small RNAs. ***Curr Opin Cell Biol,*** 16(3): 230-8.

Grigoryan, G and Keating, AE (2006). Structure-based prediction of bZIP partnering specificity. ***Journal of Molecular Biology,*** 355(5): 1125-1142.

Guillemette, B, Bataille, AR, Gevry, N, *et al.* (2005). Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. ***PLoS Biol,*** 3(12): e384.

Guo, X, Tatsuoka, K and Liu, R (2006). Histone acetylation and transcriptional regulation in the genome of Saccharomyces cerevisiae. ***Bioinformatics,*** 22(4): 392-9.

Habib, N, Kaplan, T, Margalit, H, *et al.* (2008). A novel Bayesian DNA motif comparison method for clustering and retrieval. ***PLoS Comput Biol,*** 4(2): e1000010.

Harbison, CT, Gordon, DB, Lee, TI, *et al.* (2004). Transcriptional regulatory code of a eukaryotic genome. ***Nature,*** 431(7004): 99-104.

Hashimshony, T, Zhang, J, Keshet, I, *et al.* (2003). The role of DNA methylation in setting up chromatin structure during development. ***Nat Genet,*** 34(2): 187-92.

Havranek, JJ, Duarte, CM and Baker, D (2004). A simple physical model for the prediction and design of protein-DNA interactions. ***J Mol Biol,*** 344(1): 59-70.

Hertz, GZ and Stormo, GD (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. ***Bioinformatics,*** 15(7-8): 563-77.

Hoffman, EP (2007). Skipping toward personalized molecular medicine. ***N Engl J Med,*** 357(26): 2719-22.

Hohmann, S, Krantz, M and Nordlander, B (2007). Yeast osmoregulation. ***Methods Enzymol,*** 428: 29-45.

Holstege, FC, Jennings, EG, Wyrick, JJ, *et al.* (1998). Dissecting the regulatory circuitry of a eukaryotic genome. ***Cell,*** 95(5): 717-28.

Hong, L, Schroth, GP, Matthews, HR, *et al.* (1993). Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA. ***J Biol Chem,*** 268(1): 305-14.

Hu, Z, Killion, PJ and Iyer, VR (2007). Genetic reconstruction of a functional transcriptional regulatory network. ***Nat Genet,*** 39(5): 683-7.

Hughes, TR, Roberts, CJ, Dai, H, *et al.* (2000). Widespread aneuploidy revealed by DNA microarray expression profiling. ***Nat Genet,*** 25(3): 333-7.

Hyland, EM, Cosgrove, MS, Molina, H, *et al.* (2005). Insights into the role of histone H3 and histone H4 core modifiable residues in Saccharomyces cerevisiae. ***Mol Cell Biol,*** 25(22): 10060-70.

Ioshikhes, I, Bolshoy, A, Derenshteyn, K, *et al.* (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. ***J Mol Biol,*** 262(2): 129-39.

Ioshikhes, IP, Albert, I, Zanton, SJ, *et al.* (2006). Nucleosome positions predicted through comparative genomics. ***Nat Genet,*** 38(10): 1210-5.

Iyer, VR, Horak, CE, Scafe, CS, *et al.* (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. ***Nature,*** 409(6819): 533-8.

Jamai, A, Imoberdorf, RM and Strubin, M (2007). Continuous histone H2B and transcription-dependent histone H3 exchange in yeast cells outside of replication. ***Mol Cell,*** 25(3): 345-55.

Johnson, DS, Mortazavi, A, Myers, RM, *et al.* (2007). Genome-wide mapping of in vivo protein-DNA interactions. ***Science,*** 316(5830): 1497-502.

Kalir, S and Alon, U (2004). Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. ***Cell,*** 117(6): 713-20.

Kaplan, T, Friedman, N and Margalit, H (2005). Ab initio prediction of transcription factor targets using structural knowledge. ***PLoS Comput Biol,*** 1(1): e1.

Karplus, K, Barrett, C, Cline, M, *et al.* (1999). Predicting protein structure using only sequence information. ***Proteins,*** Suppl 3: 121-5.

Kim, TH, Abdullaev, ZK, Smith, AD, *et al.* (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. ***Cell,*** 128(6): 1231-45.

Kim, TH, Barrera, LO, Zheng, M, *et al.* (2005). A high-resolution map of active promoters in the human genome. ***Nature,*** 436(7052): 876-80.

Kimura, H (2005). Histone dynamics in living cells revealed by photobleaching. ***DNA Repair (Amst),*** 4(8): 939-50.

Kimura, H and Cook, PR (2001). Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B. ***J Cell Biol,*** 153(7): 1341-53.

King, OD and Roth, FP (2003). A non-parametric model for transcription factor binding sites. ***Nucleic Acids Res,*** 31(19): e116.

Koch, CM, Andrews, RM, Flicek, P*, et al.* (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. ***Genome Research,*** 17(6): 691-707.

Kolesov, G, Wunderlich, Z, Laikova, ON*, et al.* (2007). How gene order is influenced by the biophysics of transcription regulation. ***Proc Natl Acad Sci U S A,*** 104(35): 13948-53.

Kono, H and Sarai, A (1999). Structure-based prediction of DNA target sites by regulatory proteins. ***Proteins,*** 35(1): 114-31.

Kouzarides, T (2007). Chromatin modifications and their function. ***Cell,*** 128(4): 693-705.

Krebs, JE (2007). Moving marks: Dynamic histone modifications in yeast. ***Molecular Biosystems,*** 3(9): 590-597.

Kristjuhan, A and Svejstrup, JQ (2004). Evidence for distinct mechanisms facilitating transcript elongation through chromatin in vivo. ***EMBO J,*** 23(21): 4243-52.

Kurdistani, SK and Grunstein, M (2003). Histone acetylation and deacetylation in yeast. ***Nat Rev Mol Cell Biol,*** 4(4): 276-84.

Kurdistani, SK, Tavazoie, S and Grunstein, M (2004). Mapping global histone acetylation patterns to gene expression. ***Cell,*** 117(6): 721-33.

Lachner, M, O'Carroll, D, Rea, S*, et al.* (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. ***Nature,*** 410(6824): 116-20.

Lam, FH, Steger, DJ and O'Shea, EK (2008). Chromatin decouples promoter threshold from dynamic range. ***Nature,*** 453(7192): 246-50.

Lande-Diner, L and Cedar, H (2005). Silence of the genes--mechanisms of long-term repression. ***Nat Rev Genet,*** 6(8): 648-54.

Langst, G and Becker, PB (2004). Nucleosome remodeling: one mechanism, many phenomena? ***Biochim Biophys Acta,*** 1677(1-3): 58-63.

Latchman, DS (1995). Gene regulation : a eukaryotic perspective. London ; New York, Chapman & Hall.

Lee, CK, Shibata, Y, Rao, B*, et al.* (2004). Evidence for nucleosome depletion at active regulatory regions genome-wide. ***Nat Genet,*** 36(8): 900-5.

Lee, TI, Causton, HC, Holstege, FC*, et al.* (2000). Redundant roles for the TFIID and SAGA complexes in global transcription. ***Nature,*** 405(6787): 701-4.

Lee, TI, Rinaldi, NJ, Robert, F*, et al.* (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. ***Science,*** 298(5594): 799-804.

Lee, W, Tillo, D, Bray, N*, et al.* (2007). A high- resolution atlas of nucleosome occupancy in yeast. ***Nature Genetics,*** 39(10): 1235-1244.

Lemer, CM, Rooman, MJ and Wodak, SJ (1995). Protein structure prediction by threading methods: evaluation of current techniques. ***Proteins,*** 23(3): 337-55.

Levine, M and Davidson, EH (2005). Gene regulatory networks for development. ***Proc Natl Acad Sci U S A,*** 102(14): 4936-42.

Levy, S, Sutton, G, Ng, PC*, et al.* (2007). The diploid genome sequence of an individual human. ***PLoS Biol,*** 5(10): e254.

Li, B, Carey, M and Workman, JL (2007). The role of chromatin during transcription. ***Cell,*** 128(4): 707-19.

Li, W, Meyer, CA and Liu, XS (2005). A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. ***Bioinformatics,*** 21 Suppl 1: i274-82.

Li, XY, MacArthur, S, Bourgon, R*, et al.* (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. ***PLoS Biol,*** 6(2): e27.

Lin, CM, Fu, H, Martinovsky, M*, et al.* (2003). Dynamic alterations of replication timing in mammalian cells. ***Curr Biol,*** 13(12): 1019-28.

Linger, J and Tyler, JK (2006). Global replication-independent histone H4 exchange in budding yeast. ***Eukaryot Cell,*** 5(10): 1780-7.

Liu, CL, Kaplan, T, Kim, M*, et al.* (2005). Single-nucleosome mapping of histone modifications in S. cerevisiae. ***PLoS Biol,*** 3(10): e328.

Liu, LA and Bader, JS (2007). Ab initio prediction of transcription factor binding sites. ***Pac Symp Biocomput***: 484-95.

Liu, X, Brutlag, DL and Liu, JS (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. ***Pac Symp Biocomput***: 127-38.

Luger, K, Mader, AW, Richmond, RK*, et al.* (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. ***Nature,*** 389(6648): 251-60.

Luscombe, NM, Laskowski, RA and Thornton, JM (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. ***Nucleic Acids Res,*** 29(13): 2860-74.

Lusser, A and Kadonaga, JT (2003). Chromatin remodeling by ATP-dependent molecular machines. ***Bioessays,*** 25(12): 1192-200.

MacIsaac, KD and Fraenkel, E (2006). Practical strategies for discovering regulatory DNA sequence motifs. ***PLoS Comput Biol,*** 2(4): e36.

MacIsaac, KD, Wang, T, Gordon, DB*, et al.* (2006). An improved map of conserved regulatory sites for Saccharomyces cerevisiae. ***BMC Bioinformatics,*** 7: 113.

Mandel-Gutfreund, Y, Baron, A and Margalit, H (2001). A structure-based approach for prediction of protein binding sites in gene upstream regions. ***Pac Symp Biocomput***: 139-50.

Mandel-Gutfreund, Y and Margalit, H (1998). Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. ***Nucleic Acids Res,*** 26(10): 2306-12.

Mandel-Gutfreund, Y, Schueler, O and Margalit, H (1995). Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. ***J Mol Biol,*** 253(2): 370-82.

Margulies, M, Egholm, M, Altman, WE*, et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. ***Nature,*** 437(7057): 376-80.

Masumoto, H, Hawke, D, Kobayashi, R*, et al.* (2005). A role for cell-cycle-regulated histone H3 lysine 56 acetylation in the DNA damage response. ***Nature,*** 436(7048): 294-8.

Mavrich, TN, Jiang, C, Ioshikhes, IP*, et al.* (2008). Nucleosome organization in the Drosophila genome. ***Nature***.

Mikkelsen, TS, Ku, M, Jaffe, DB*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. ***Nature,*** 448(7153): 553-60.

Millar, CB and Grunstein, M (2006). Genome-wide patterns of histone modifications in yeast. ***Nature Reviews Molecular Cell Biology,*** 7(9): 657-666.

Millar, CB, Xu, F, Zhang, KL*, et al.* (2006). Acetylation of H2AZ Lys 14 is associated with genome-wide gene activity in yeast. ***Genes & Development,*** 20(6): 711-722.

Mito, Y, Henikoff, JG and Henikoff, S (2007). Histone replacement marks the boundaries of cis-regulatory domains. ***Science,*** 315(5817): 1408-1411.

Moroni, E, Caselle, M and Fogolari, F (2007). Identification of DNA-binding protein target sequences by physical effective energy functions: free energy analysis of lambda repressor-DNA complexes. ***Bmc Structural Biology,*** 7.

Morozov, AV, Havranek, JJ, Baker, D*, et al.* (2005). Protein-DNA binding specificity predictions with structural models. ***Nucleic Acids Research,*** 33(18): 5781-5798.

Narlikar, L, Gordan, R and Hartemink, AJ (2007). A nucleosome-guided map of transcription factor binding sites in yeast. ***PLoS Comput Biol,*** 3(11): e215.

Nathan, D, Ingvarsdottir, K, Sterner, DE*, et al.* (2006). Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. ***Genes Dev,*** 20(8): 966-76.

Nemeth, A and Langst, G (2004). Chromatin higher order structure: opening up chromatin for transcription. ***Brief Funct Genomic Proteomic,*** 2(4): 334-43.

Nevitt, T, Pereira, J, Azevedo, D*, et al.* (2004). Expression of YAP4 in Saccharomyces cerevisiae under osmotic stress. ***Biochem J,*** 379(Pt 2): 367-74.

Ng, HH, Robert, F, Young, RA*, et al.* (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. ***Mol Cell,*** 11(3): 709-19.

Noma, K, Allis, CD and Grewal, SI (2001). Transitions in distinct histone H3 methylation patterns at the heterochromatin domain boundaries. ***Science,*** 293(5532): 1150-5.

O'Rourke, SM and Herskowitz, I (2004). Unique and redundant roles for HOG MAPK pathway components as revealed by whole-genome expression analysis. ***Mol Biol Cell,*** 15(2): 532-42.

Osada, R, Zaslavsky, E and Singh, M (2004). Comparative analysis of methods for representing and searching for transcription factor binding sites. ***Bioinformatics,*** 20(18): 3516-25.

Ozdemir, A, Masumoto, H, Fitzjohn, P*, et al.* (2006). Histone H3 lysine 56 acetylation: a new twist in the chromosome cycle. ***Cell Cycle,*** 5(22): 2602-8.

Ozsolak, F, Song, JS, Liu, XS*, et al.* (2007). High-throughput mapping of the chromatin structure of human promoters. ***Nat Biotechnol,*** 25(2): 244-8.

Pavletich, NP and Pabo, CO (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. ***Science,*** 261(5129): 1701-7.

Peckham, HE, Thurman, RE, Fu, Y*, et al.* (2007). Nucleosome positioning signals in genomic DNA. ***Genome Res,*** 17(8): 1170-7.

Peters, AH, Mermoud, JE, O'Carroll, D*, et al.* (2002). Histone H3 lysine 9 methylation is an epigenetic imprint of facultative heterochromatin. ***Nat Genet,*** 30(1): 77-80.

Peterson, CL and Laniel, MA (2004). Histones and histone modifications. ***Curr Biol,*** 14(14): R546-51.

Pham, H, Ferrari, R, Cokus, SJ*, et al.* (2007). Modeling the regulatory network of histone acetylation in Saccharomyces cerevisiae. ***Mol Syst Biol,*** 3: 153.

Pokholok, DK, Harbison, CT, Levine, S*, et al.* (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. ***Cell,*** 122(4): 517-27.

Posas, F, Chambers, JR, Heyman, JA*, et al.* (2000). The transcriptional response of yeast to saline stress. ***J Biol Chem,*** 275(23): 17249-55.

Proft, M, Gibbons, FD, Copeland, M*, et al.* (2005). Genomewide identification of Sko1 target promoters reveals a regulatory network that operates in response to osmotic stress in Saccharomyces cerevisiae. ***Eukaryot Cell,*** 4(8): 1343-52.

Proft, M and Serrano, R (1999). Repressors and upstream repressing sequences of the stress-regulated ENA1 gene in Saccharomyces cerevisiae: bZIP protein Sko1p confers HOG-dependent osmotic regulation. ***Mol Cell Biol,*** 19(1): 537-46.

Proft, M and Struhl, K (2002). Hog1 kinase converts the Sko1-Cyc8-Tup1 repressor complex into an activator that recruits SAGA and SWI/SNF in response to osmotic stress. ***Mol Cell,*** 9(6): 1307-17.

Ptashne, M and Gann, A (2002). Genes & signals. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory Press.

Qi, Y, Rolfe, A, MacIsaac, KD*, et al.* (2006). High-resolution computational models of genome binding events. ***Nat Biotechnol,*** 24(8): 963-70.

Quandt, K, Frech, K, Karas, H*, et al.* (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. ***Nucleic Acids Res,*** 23(23): 4878-84.

Raghuraman, MK, Winzeler, EA, Collingwood, D*, et al.* (2001). Replication dynamics of the yeast genome. ***Science,*** 294(5540): 115-21.

Raisner, RM, Hartley, PD, Meneghini, MD*, et al.* (2005). Histone variant H2A.Z marks the 5 ' ends of both active and inactive genes in euchromatin. ***Cell,*** 123(2): 233-248.

Rando, OJ (2007). Global patterns of histone modifications. ***Current Opinion in Genetics & Development,*** 17(2): 94-99.

Rea, S, Eisenhaber, F, O'Carroll, D*, et al.* (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. ***Nature,*** 406(6796): 593-9.

Reinke, H and Horz, W (2003). Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. ***Mol Cell,*** 11(6): 1599-607.

Ren, B, Robert, F, Wyrick, JJ*, et al.* (2000). Genome-wide location and function of DNA binding proteins. ***Science,*** 290(5500): 2306-9.

Rep, M, Krantz, M, Thevelein, JM*, et al.* (2000). The transcriptional response of Saccharomyces cerevisiae to osmotic shock. Hot1p and Msn2p/Msn4p are required for the induction of subsets of high osmolarity glycerol pathway-dependent genes. ***J Biol Chem,*** 275(12): 8290-300.

Rep, M, Proft, M, Remize, F*, et al.* (2001). The Saccharomyces cerevisiae Sko1p transcription factor mediates HOG pathway-dependent osmotic regulation of a set of genes encoding enzymes implicated in protection from oxidative damage. ***Mol Microbiol,*** 40(5): 1067-83.

Rep, M, Reiser, V, Gartner, U*, et al.* (1999). Osmotic stress-induced gene expression in Saccharomyces cerevisiae requires Msn1p and the novel nuclear factor Hot1p. ***Mol Cell Biol,*** 19(8): 5474-85.

Rice, JC, Briggs, SD, Ueberheide, B*, et al.* (2003). Histone methyltransferases direct different degrees of methylation to define distinct chromatin domains. ***Mol Cell,*** 12(6): 1591-8.

Robert, F, Pokholok, DK, Hannett, NM*, et al.* (2004). Global position and recruitment of HATs and HDACs in the yeast genome. ***Mol Cell,*** 16(2): 199-209.

Roberts, CJ, Nelson, B, Marton, MJ*, et al.* (2000). Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. ***Science,*** 287(5454): 873-80.

Robertson, G, Hirst, M, Bainbridge, M, *et al.* (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. ***Nat Methods,*** 4(8): 651-7.

Robyr, D and Grunstein, M (2003). Genomewide histone acetylation microarrays. ***Methods,*** 31(1): 83-9.

Robyr, D, Suka, Y, Xenarios, I, *et al.* (2002). Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. ***Cell,*** 109(4): 437-46.

Roh, TY, Cuddapah, S, Cui, K, *et al.* (2006). The genomic landscape of histone modifications in human T cells. ***Proc Natl Acad Sci U S A,*** 103(43): 15782-7.

Roh, TY, Cuddapah, S and Zhao, K (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. ***Genes Dev,*** 19(5): 542-52.

Rost, B and Sander, C (1996). Bridging the protein sequence-structure gap by structure predictions. ***Annu Rev Biophys Biomol Struct,*** 25: 113-36.

Rost, B, Schneider, R and Sander, C (1997). Protein fold recognition by prediction-based threading. ***J Mol Biol,*** 270(3): 471-80.

Rufiange, A, Jacques, PE, Bhat, W, *et al.* (2007). Genome-wide replication-independent histone H3 exchange occurs predominantly at promoters and implicates H3K56 acetylation and Asf1. ***Molecular Cell,*** 27(3): 393-405.

Sadee, W and Dai, Z (2005). Pharmacogenetics/genomics and personalized medicine. ***Hum Mol Genet,*** 14 Spec No. 2: R207-14.

Sandelin, A, Alkema, W, Engstrom, P, *et al.* (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. ***Nucleic Acids Res,*** 32(Database issue): D91-4.

Schermer, UJ, Korber, P and Horz, W (2005). Histones are incorporated in trans during reassembly of the yeast PHO5 promoter. ***Mol Cell,*** 19(2): 279-85.

Schones, DE, Cui, K, Cuddapah, S, *et al.* (2008). Dynamic regulation of nucleosome positioning in the human genome. ***Cell,*** 132(5): 887-98.

Schones, DE and Zhao, K (2008). Genome-wide approaches to studying chromatin modifications. ***Nature Reviews Genetics,*** 9(3): 179-191.

Schreiber, SL and Bernstein, BE (2002). Signaling network model of chromatin. ***Cell,*** 111(6): 771-8.

Schubeler, D, MacAlpine, DM, Scalzo, D, *et al.* (2004). The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. ***Genes Dev,*** 18(11): 1263-71.

Schueler-Furman, O, Wang, C, Bradley, P, *et al.* (2005). Progress in modeling of protein structures and interactions. ***Science,*** 310(5748): 638-42.

Schwabish, MA and Struhl, K (2004). Evidence for eviction and rapid deposition of histones upon transcriptional elongation by RNA polymerase II. ***Mol Cell Biol,*** 24(23): 10111-7.

Segal, E, Fondufe-Mittendorf, Y, Chen, L, *et al.* (2006). A genomic code for nucleosome positioning. ***Nature,*** 442(7104): 772-8.

Segal, MR (2008). Re-cracking the nucleosome positioning code. ***Stat Appl Genet Mol Biol,*** 7: Article14.

Sharon, E and Segal, E (2007). A Feature-Based Approach to Modeling Protein-DNA Interactions. ***Proc. of the 11th International Conf. on Research in Computational Molecular Biology***: 77-91.

Shilatifard, A (2006). Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. ***Annu Rev Biochem,*** 75: 243-69.

Shivaswamy, S, Bhinge, A, Zhao, Y, *et al.* (2008). Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. ***PLoS Biol,*** 6(3): e65.

Shivaswamy, S and Iyer, VR (2008). Stress-dependent dynamics of global chromatin remodeling in yeast: Dual role for SWI/SNF in the heat shock stress response. ***Molecular and Cellular Biology,*** 28(7): 2221-2234.

Siggers, TW and Honig, B (2007). Structure-based prediction of C2H2 zinc-finger binding specificity: sensitivity to docking geometry. ***Nucleic Acids Research,*** 35(4): 1085-1097.

Simon, I, Barnett, J, Hannett, N, *et al.* (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. ***Cell,*** 106(6): 697-708.

Skolnick, J and Fetrow, JS (2000). From genes to protein structure and function: novel applications of computational approaches in the genomic era. ***Trends Biotechnol,*** 18(1): 34-9.

Spellman, PT, Sherlock, G, Zhang, MQ, *et al.* (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. ***Mol Biol Cell,*** 9(12): 3273-97.

Stathopoulos, A and Levine, M (2005). Genomic regulatory networks and animal development. ***Dev Cell,*** 9(4): 449-62.

Steinfeld, I, Shamir, R and Kupiec, M (2007). A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. ***Nat Genet,*** 39(3): 303-9.

Sterner, DE and Berger, SL (2000). Acetylation of histones and transcription-related factors. ***Microbiol Mol Biol Rev,*** 64(2): 435-59.

Stormo, GD (2000). DNA binding sites: representation and discovery. ***Bioinformatics,*** 16(1): 16-23.

Strahl, BD and Allis, CD (2000). The language of covalent histone modifications. ***Nature,*** 403(6765): 41-5.

Stunkel, W, Kober, I and Seifart, KH (1997). A nucleosome positioned in the distal promoter region activates transcription of the human U6 gene. ***Mol Cell Biol,*** 17(8): 4397-405.

Tanay, A (2006). Extensive low-affinity transcriptional interactions in the yeast genome. ***Genome Res,*** 16(8): 962-72.

Turner, BM (2000). Histone acetylation and an epigenetic code. ***Bioessays,*** 22(9): 836-45.

Turner, BM (2002). Cellular memory and the histone code. ***Cell,*** 111(3): 285-91.

Vakoc, CR, Mandat, SA, Olenchock, BA*, et al.* (2005). Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. ***Mol Cell,*** 19(3): 381-91.

Valouev, A, Ichikawa, J, Tonthat, T*, et al.* (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. ***Genome Res***.

Van Driessche, N, Demsar, J, Booth, EO*, et al.* (2005). Epistasis analysis with global transcriptional phenotypes. ***Nat Genet,*** 37(5): 471-7.

Venter, U, Svaren, J, Schmitz, J*, et al.* (1994). A nucleosome precludes binding of the transcription factor Pho4 in vivo to a critical target site in the PHO5 promoter. ***EMBO J,*** 13(20): 4848-55.

Vignali, M, Hassan, AH, Neely, KE*, et al.* (2000). ATP-dependent chromatin-remodeling complexes. ***Mol Cell Biol,*** 20(6): 1899-910.

Vogelauer, M, Wu, J, Suka, N*, et al.* (2000). Global histone acetylation and deacetylation in yeast. ***Nature,*** 408(6811): 495-8.

von Hippel, PH (1994). Protein-DNA recognition: new perspectives and underlying themes. ***Science,*** 263(5148): 769-70.

Waterborg, JH (2001). Dynamics of histone acetylation in Saccharomyces cerevisiae. ***Biochemistry,*** 40(8): 2599-605.

Wei, CL, Wu, Q, Vega, VB*, et al.* (2006). A global map of p53 transcription-factor binding sites in the human genome. ***Cell,*** 124(1): 207-19.

Wheeler, DA, Srinivasan, M, Egholm, M*, et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. ***Nature,*** 452(7189): 872-6.

Whitehouse, I, Rando, OJ, Delrow, J*, et al.* (2007). Chromatin remodelling at promoters suppresses antisense transcription. ***Nature,*** 450(7172): 1031-U3.

Wingender, E, Chen, X, Fricke, E*, et al.* (2001). The TRANSFAC system on gene expression regulation. ***Nucleic Acids Res,*** 29(1): 281-3.

Workman, JL (2006). Nucleosome displacement in transcription. ***Genes & Development,*** 20(15): 2009-2017.

Workman, JL and Kingston, RE (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. ***Annu Rev Biochem,*** 67: 545-79.

Wu, L and Winston, F (1997). Evidence that Snf-Swi controls chromatin structure over both the TATA and UAS regions of the SUC2 promoter in Saccharomyces cerevisiae. ***Nucleic Acids Res,*** 25(21): 4230-4.

Wu, R, Terry, AV, Singh, PB*, et al.* (2005). Differential subnuclear localization and replication timing of histone H3 lysine 9 methylation states. ***Mol Biol Cell,*** 16(6): 2872-81.

Xu, F, Zhang, K and Grunstein, M (2005). Acetylation in histone H3 globular domain regulates gene expression in yeast. ***Cell,*** 121(3): 375-85.

Yabuki, N, Terashima, H and Kitada, K (2002). Mapping of early firing origins on a replication profile of budding yeast. ***Genes Cells,*** 7(8): 781-9.

Yale, J and Bohnert, HJ (2001). Transcript expression in Saccharomyces cerevisiae at high salinity. ***J Biol Chem,*** 276(19): 15996-6007.

Yang, A, Zhu, Z, Kapranov, P*, et al.* (2006). Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. ***Mol Cell,*** 24(4): 593-602.

Yassour, M, Kaplan, T, Jaimovich, A*, et al.* (2008). Nucleosome positioning from tiling microarray data. ***Bioinformatics,*** to appear.

Yuan, GC and Liu, JS (2008). Genomic sequence is highly predictive of local nucleosome depletion. ***PLoS Comput Biol,*** 4(1): e13.

Yuan, GC, Liu, YJ, Dion, MF*, et al.* (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. ***Science,*** 309(5734): 626-30.

Yuan, GC, Ma, P, Zhong, W*, et al.* (2006). Statistical assessment of the global regulatory role of histone acetylation in Saccharomyces cerevisiae. ***Genome Biol,*** 7(8): R70.

Zeitlinger, J, Zinzen, RP, Stark, A*, et al.* (2007). Whole-genome ChIP-chip analysis of Dorsal, Twist, and Snail suggests integration of diverse patterning processes in the Drosophila embryo. ***Genes Dev,*** 21(4): 385-90.

Zhang, J, Xu, F, Hashimshony, T*, et al.* (2002). Establishment of transcriptional competence in early and late S phase. ***Nature,*** 420(6912): 198-202.

Zhang, X (2008). The epigenetic landscape of plants. ***Science,*** 320(5875): 489-92.

Zhang, Y and Reinberg, D (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. ***Genes Dev,*** 15(18): 2343-60.

Zhou, Q and Wong, WH (2004). CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. ***Proc Natl Acad Sci U S A,*** 101(33): 12114-9.

תוכן העניינים

כאשר הדנ"א משוכפל ומחצית הנוקלאוזומים מורשים לתא הבת ומוחלפים בחדשים. אולם, האם תופעה זו של תחלופת נוקלאוזומים מתרחשת גם בנפרד ממחזור התא? כדי לענות על שאלה זו באופן ישיר, המשכתי בשיתוף הפעולה עם מעבדתו של אוליבר רנדו. תכננו ניסוי המבוסס על שמרים מוטנטים בהם אחד החלבונים המרכיבים את הנוקלאוזום (ההיסטון H3) שוכפל, סומן, והוצב תחת בקרה חיצונית. מנגנון זה איפשר לגדל תאי שמר, לעצור את מחזור התא באופן חיצוני, להתחיל את ייצור ההיסטון המוטנטי (המסומן), ולבסוף להעריך בסדרה של מדידות בטכנולוגיית ChIP-chip את כמות הנוקלאוזום המסומן בכל עמדה לאורך הגנום (פרק 4). כדי לנתח נתונים אלו, ולתרגמם לקצבי התחלופה הממוצעים עבור כל נוקלאוזום, פיתחתי מודל מתמטי המבוסס על משוואות קצב, ואלגוריתם חישובי לשערוך הקצבים. להפתעתי, הקצבים שהתקבלו פרשו טווח רחב של התנהגויות. החל בנוקלאוזומים שלא הוחלפו כלל (למעט תחלופה עקב חלוקת התא) ועד לנוקלאוזומים שהתחלפו בקצב גבוה מרזולוציית הניסוי. ניתוח סטטיסטי שערכתי קישר את קצבי התחלופה למיקום הנוקלאוזום לאורך הגן, ולרמת ביטוי הגן. באופן כללי, נוקלאוזומים לאורך גנים פעילים התחלפו בקצב גבוה יותר מכאלה לאורך גנים מושתקים, בהתאמה עם ממצאים קודמים שקישרו בין פעילות על הדנ"א (למשל שעתוק) ואי-יציבות נוקלאוזומלית. בנוסף לכך, מצאתי כי נוקלאוזומים הממוקמים באזורי הבקרה נטו לקצבי תחלופה גבוהים פי 4 (בממוצע) מקצבי התחלופה של נוקלאוזומים הנמצאים לאורך רצף הגן המקודד לחלבון. תופעה זו מפתיעה, היות ודווקא באזורים המקודדים, היכן שהדנ"א משמש כתבנית לשעתוק ועובר טלטלות פיזיות, ניתן היה לצפות לקצבים גבוהים יותר. אני מציע מספר הסברים אפשרים לתופעה שמצאנו, ביניהם: (1) אינטראקציות חלבון-דנ"א המתרחשות באזורי הבקרה גורמות לאי-יציבות נוקלאוזומלית; (2) תבנית המודיפיקציות המאפיינת נוקלאוזומים באזורי בקרה מעודדת תחלופה גבוהה באמצעות מנגנון לא ידוע; (3) קצב התחלופה הגבוה באזורי בקרה משמש כמנגנון מובנה לשינויים מהירים באריזת הדנ"א וע"י כך משמש כפלטפורמה לרענון בקרת השעתוק; ו- (4) קצב התחלופה הגבוה באזורי בקרה משמש כמחסום פיזי להתפשטות מצב הכרומטין לאורך הדנ"א וכך מאפשר שונות ברמות הביטוי בין גנים סמוכים. מחקרים נוספים ידרשו כדי לאפיין טוב יותר את התופעה הזו ואת הקשר הסיבתי שלה לבקרת שעתוק.

לסיכום, בעבודה זו פיתחתי שיטות חישוביות, מתמטיות וסטטיסטיות, וניתחתי נתונים גנומיים ממגוון סוגים כדי ללמוד אודות ההיבטים השונים של בקרת שעתוק בתאים אאוקריוטים. כפי שהראתי, בקרה זו מתרחשת במגוון רמות, החל בקישור בין גורמי שעתוק לבין הדנ"א ובאינטראקציות בינם לבין עצמם, וכלה באריזת הדנ"א על-גבי נוקלאוזומים, באופן סימונם ע"י מודיפיקציות קוולנטיות, ובשינויים הדינמיים המעורבים באריזה זו לאורך חיי התא. ככלל, מחקריי תורמים לידיעותינו אודות בקרת השעתוק ברמה המנגנונית והמערכתית, ולהבנה עמוקה יותר של התהליכיים התאיים והמחלות האנושיות המערבות שיבושים בהם.

מתפקדים (non functional binding sites), אשר גם בהיותם קשורים לא השפיעו על ביטוי הגן. פערים אלה בין רצף הדנ"א, קישור החלבון, והשפעתו על ביטוי גנים, תומכים בקיום מנגנונים נוספים המעורבים בבקרת שעתוק ביצורים אאוקריוטים. כך למשל אריזת הדנ"א על-גבי קומפלקסים חלבוניים הקרויים נוקלאוזומים, אשר משפיעה על נגישותו לחלבוני בקרה.

כדי להבין את הקשר בין בקרת שעתוק לבין אריזת הדנ"א על גבי הכרומטין, התמקדתי במידע האפיגנטי המקודד באריזת הדנ"א. מידע זה כולל הן את מיקום הנוקלאוזומים לאורך הדנ"א, והן את סימונם ע"י מודיפיקציות קוולנטיות (הצמדת שיירים ביוכימיים כגון מתיל או אצטיל לחומצות אמינו ספציפיות על-גבי הנוקלאוזום). שני המנגנונים הנ"ל, ביחד עם מתילציה של הדנ"א, קושרו לדחיסה מרחבית של הדנ"א, הסתרתו מפני פקטורי שעתוק ועצירת ביטוי גנים. מטרתי לפיכך היתה לאפיין את מצב הכרומטין בתאים חיים. לשם כך שיתפתי פעולה עם מעבדתו של אוליבר רנדו מאוניברסיטת הרווארד (Oliver Rando), אשר מדד בטכנולוגיית ChIP-chip את תבנית האצטילציה והמתילציה של נוקלאוזומים לאורך הדנ"א בסקלה גנומית וברזולוציה של נוקלאוזום בודד (פרק 3). הנתונים שנאספו מאפשרים לבדוק לראשונה את נכונות היפותזת הקוד האפיגנטי (Histone code hypothesis). מודל תאורטי זה גורס שניתן לפרש את מצב הכרומטין בעזרת קוד פשוט, כך שלכל מודיפיקציה יש משמעות ברורה. כדי לנתח את הנתונים, קרי רמות האצטילציה והמתילציה שנמדדו ב-12 עמדות שונות על גבי אלפי נוקלאוזומים לאורך הגנום, הפעלתי מגוון כלים אלגוריתמיים וסטטיסטיים. התרכזתי בהצלבת רמות האצטילציה/מתילציה בין העמדות השונות, וכן בקשר בין תבנית המודיפיקציות של כל נוקלאוזום לרצף דנ"א המלופף סביבו. האנליזה שערכתי מטילה ספק בנכונות ההיפותזה, היות ולא מצאתי קוד שכזה (לפחות לא בין 12 העמדות שחקרנו). יחד עם זאת, ניתוח הנתונים הראה כי 12 העמדות שחקרנו מתחלקות באופן גס לשתי קבוצות מקבילות. בקבוצה הראשונה מסומנים באופן ייחודי שני הנוקלאוזומים הגובלים באתרי תחילת שעתוק, בעוד ששינויים בעמדות מהקבוצה השניה נטו להופיע באופן הדרגתי לאורך גנים. בנוסף, מצאתי קשר הדוק בין סימון הנוקלאוזומים לבין רמת הביטוי של הגנים הארוזים עליהם. לפיכך, למרות שאין ממצאים התומכים בקיום "קוד", הרי שניתוח סטטיסטי מראה כי מהתבוננות בתבנית המודיפיקציות של נוקלאוזומים ניתן להסיק אודות מיקומם היחסי לאורך הגן ורמת הביטוי שלו.

כאשר מצב התא או סביבתו משתנים, תכנית הבקרה השעתוקית מתעדכנת בתוך דקות כדי לספק את צרכי התא. יחד עם שינוי זה ברמות ביטוי הגנים, מתרחש שינוי מקביל במצב הכרומטין. שינוי זה מושג ע"י אנזימים ספציפיים אשר מזיזים נוקלאוזומים, או לחלופין מעדכנים את מצב האצטילציה והמתילציה שלהם ע"י הוספת ופירוק אותם שיירים (אצטיל ומתיל). יחד עם זאת, קיימת אפשרות לדרך נוספת לעדכון המצב האפיגנטי - ע"י החלפת הנוקלאוזומים המסומנים בנוקלאוזומים חדשים. תופעה דומה מתרחשת במהלך מחזור התא,

חלבון-דנ"א. בהנתן רצף חלבון חדש מאותה משפחה מבנית, ניתן להשתמש במודל הקישור הפתור כדי לזהות את חומצות האמינו אשר קושרות את הדנ"א ולנבא את העדפות הקישור (כלומר אילו רצפי דנ"א חלבון זה יקשור). הדגמתי את יכולות השיטה בעזרת חלבונים ממשפחת ה- C2H2 zinc finger בכך שניתחתי את גנום זבוב הפירות (*Drosophila melanogaster*) וניבאתי מוטיבי קישור לכ-29 פקטורי שעתוק. לאחר מכן סרקתי את רצפי הבקרה המוכרים בגנום הזבוב, ואפיינתי מאות גני מטרה לכל פקטור שעתוק. ע"י שילוב מידע זה עם נתונים נוספים, כגון רמות ביטוי הגנים בשלבים שונים של התפתחות הזבוב, או חלוקתם לקבוצות פונקציונליות שונות, הגעתי לתובנות חדשות בנוגע לאופי ורמות הפעילות של כל פקטור שעתוק.

התפתחויות טכנולוגיות שהושגו בשנים האחרונות מאפשרות למדוד את הקשר בין פקטורי שעתוק לגנים בהיקף גנומי וברזולוציה גבוהה. בשיתוף עם המעבדה של ארין או'שה מאוניברסיטת הרווארד (Erin O'Shea), התמקדתי בפיתוח שיטה להבנת המבנה הפנימי והמנגנונים הקומבינטוריים של רשת בקרת שעתוק (פרק 2). האסטרטגיה שלנו מבוססת על שימוש מובנה בזנים מוטנטים (בהם חסר פקטור שעתוק אחד או יותר) בכדי למדוד ולכמת את השפעת פקטורי השעתוק על ביטוי גנים. השתמשנו בשיטה זו בכדי לחקור את רשת בקרת השעתוק אשר מתווכת את תגובת שמר האופים (*Saccharomyces sereviciae*) ללחץ אוסמוטי גבוה (HOG pathway). בתגובה ללחץ אוסמוטי (למשל בתנאי מליחות קיצוניים), החלבון Hog1 מוכנס לגרעין התא, שם הוא מזרחן ומפעיל מספר פקטורי שעתוק משניים. בעזרת האלגוריתם שפיתחתי, ניתחתי את הנתונים הגנומיים שנאספו, ובניתי מודל כמותי מדוייק של בקרת השעתוק ברשת זו. במקביל לרשת תפקודית זו המבוססת על רמות ביטוי של גנים, בנינו שתי רשתות מקבילות המבוססות על נתונים מסוג שונה. תחילה, נעשה שימוש בטכנולוגיית ChIP-chip בכדי לזהות אתרי דנ"א הנקשרים בתא החי באופן פיזי ע"י פקטורי שעתוק. בשיטה זו רצפי דנ"א הקשורים לחלבון מושקעים בעזרת נוגדנים, מזוקקים, ומזוהים בעזרת מערכי דנ"א (DNA microarrays) אשר מכסים בצפיפות את גנום השמר. כדי לנתח את הנתונים הביולוגיים האלה, פיתחתי אלגוריתם חישובי אשר מזהה את המיקום המדוייק וקובע את עוצמת הקישור היחסית של כל אתר קישור בגנום. על ידי כך שירטטתי את החיווט הפיזי של רשת בקרת השעתוק. לבסוף, בנינו גרסא שלישית של אותה רשת בקרה בהתבסס על רצף הדנ"א בלבד, ע"י איפיון רצפי ההכרה של פקטורי השעתוק (מוטיבי דנ"א). כפי שהראיתי, שלוש רשתות הבקרה השונות (המבוססות על תפקוד גנים, על קישור פיזי לדנ"א, ועל מוטיבים קצרים ברצף הדנ"א), חופפות זו את זו באופן ניכר – רוב הגנים שהכילו את מוטיב הדנ"א של פקטור שעתוק אכן נמצאו קשורים אליו פיזית, ויתרה מזו, פקטור השעתוק נמצא מעורב בבקרה על רמות הביטוי של הגן. יחד עם זאת, מצאתי מקרים רבים של אתרי קישור חבויים (latent binding sites), אשר נותרו פנויים למרות התאמתם הרבה למוטיב הקישור של פקטור השעתוק. בנוסף, מצאתי מספר רב של אתרי קישור לא

תקציר

כל התאים בייצור חי מכילים מידע גנטי זהה אשר מקודד ברצף הדנ"א. אולם, תאים מסוגים שונים נבדלים זה מזה במבנה ובפעילות. הבדלים אלה מושגים בעזרת בקרות שונות, כאשר הבולטת בהן היא בקרת השעתוק, אשר אחראית על הפעלת גנים ומסלולים שונים בהתאם לסוג התא, למצבו ולצרכיו. אחד האתגרים החשובים ביותר בביולוגיה הוא הבנת מנגנוני בקרת השעתוק בתא חי. ידע זה יאפשר לנו להבין טוב יותר כיצד תאים עובדים, כיצד הם מגיבים לאותות חיצוניים, מה משתבש במחלות (לדוגמא בסרטן, אשר מערב שיבושים בבקרת גנים), ואילו צעדים אנו יכולים לנקוט כנגד שיבושים אלה. ברמה הבסיסית ביותר, בקרת שעתוק מושגת ע"י פקטורי שעתוק - חלבונים הנקשרים לדנ"א בסמיכות לגנים ומעודדים (או מונעים) את שעתוקם. במהלך עבודה זו, אני מתרכז בשלושה היבטים מרכזיים של בקרת שעתוק. ראשית, אני מציג גישה חדשנית לזיהוי גני המטרה של פקטורי שעתוק חדשים, בהתבסס על רצף החלבון (פרק 1). שנית, אני בוחן את המנגנונים בהם מספר פקטורי שעתוק פועלים יחדיו כדי לעבד אותות חיצוניים ולתרגמם לשינויים בבקרת השעתוק התאית (פרק 2). לבסוף, אני מתרכז במימד מרתק של בקרת השעתוק, הקשר בין אריזת דנ"א על-גבי נוקלאוזומים, ובפרט מיקומם לאורך מולקולת הדנ"א והאופן בו הם מסומנים קוולנטית, לבין נגישות הדנ"א לקישור פקטורי שעתוק ורמות הביטוי של הגנים המקודדים בדנ"א (פרקים 3 ו- 4).

כדי להבין כיצד מושגת בקרה כה הדוקה של ביטוי גנים, עלינו לתאר במדויק את רשת השעתוק התאית, ובכך לאפיין אילו גנים מבוקרים ע"י אילו פקטורי שעתוק. היות ומיפוי ישיר של אינטראקציות חלבון-דנ"א בכלים ניסויים הוא לרוב יקר ומייגע, פותחו כתחליף אלגוריתמים חישוביים לזיהוי אתרי קישור של פקטורי שעתוק. באופן טיפוסי גישות חישוביות אלו כוללת: (1) ניתוח תוצאות ניסויים בהיקף גנומי וזיהוי חלק מגני המטרה של פקטור השעתוק; (2) ניתוח רצף הדנ"א באזורי הבקרה של גנים אלה וזיהוי אתרים קצרים אליהם הוא נקשר; (3) ייצוג אתרים אלה בעזרת מודל הסתברותי (מוטיב דנ"א); ו- (4) שימוש במוטיב זה לשם סריקת אזורי בקרה של גנים נוספים וזיהוי אתרי קישור חדשים. אסטרטגיה חישובית זו הוכחה כיעילה, בעיקר עבור פקטורי שעתוק אשר אתרי הקישור שלהם כבר אופיינו במספר ניסויים רב. אולם מה בנוגע לחלבונים אשר טרם נבחנו? למשל, כאלה אשר נמצאו בעזרת פרוייקטי הגנום המתפרסמים חדשות לבקרים? במהלך הדוקטורט שלי פיתחתי גישה חישובית לאיפיון מוטיב הדנ"א של חלבון, גם בהעדר מידע ניסויי אודותיו (פרק 1). בהסתמך על מידע מקומפלקסים מבניים של חלבון ומולקולת הדנ"א, האלגוריתם שפיתחתי מנתח את רצף החלבון ומזהה את העמדות אשר נקשרות באופן ישיר למולקולת הדנ"א. נתונים אלו משמשים לאיפיון הקישור בין החלבון לדנ"א (ע"י קביעת חומצות האמינו והבסיסים ביניהם מתקיים קשר), ומתורגמים להעדפות הסטטיסטיות של אינטראקציות

עבודה זו נעשתה בהדרכתם של
פרופ' ניר פרידמן ופרופ' חנה מרגלית

ניתוח חישובי של בקרת שעתוק:

מרצף הדנ"א ועד דינמיקה של כרומטין

חיבור לשם קבלת תואר דוקטור לפילוסופיה

מאת

תומר קפלן

הוגש לסינט האוניברסיטה העברית, בירושלים
מאי 2008