TEL AVIV UNIVERSITY
GEORGE S. WISE FACULTY OF LIFE SCIENCES
GRADUATE SCHOOL

Field

Computational Biology

Subject

The Influence of Chromatin Modifiers on Transcription

Submitted by

Israel Steinfeld

This work was carried out in partial fulfillment of the requirements for a MSc. degree in the Department of Molecular Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University.

Advisors

Prof. Martin Kupiec Prof. Ron Shamir

October 2007

ABSTRACT

Transcription regulation is fundamental in many biological processes in all living organisms. A particularly extensively studied area in transcription regulation is that of genes, where the major transcriptional program is governed by transcription factors. These factors have affinity to specific sequences in the DNA, upstream of the transcription start site of the genes they regulate. Yet, not all phenotypes can be explained by regulation in the DNA level. For example, different cell types, having the same DNA content, carry out different transcriptional programs. Hence, it is clear that other factors participate in the complexity and diversity of transcription regulation.

Another major factor, found in recent years to play an important role in transcription regulation, is chromatin structure. A condensed chromatin structure can prevent the access of external factors, such as transcription factors, thereby preventing execution of sequence-based transcription programs. Many factors that influence chromatin structure have been identified, but the transcriptional programs in which they participate are still poorly understood. In various cases chromatin modifiers participate in transcriptional control together with DNA bound transcription factors. Novel high-throughput experimental methods allow the genome-wide identification of binding sites for transcription factors, as well as the quantification of gene expression under various environmental and genetic conditions.

In this thesis we study the contribution of chromatin structure to transcription. To do so we have developed a new statistical model methodology that uses and the vast amount of available data to dissect the intricate relationships of chromatin modifiers and transcription factors. Using our methodology we were able to measure and characterize the dependency of transcription factors on specific chromatin modifiers in carrying out their transcriptional programs.

Our methodology was applied to one of the most widely used and basic eukaryote model organism, *Saccharomyces cerevisiae*. We collected a diverse compendium of gene expression profiles, comprising 170 experiments of strains defective for chromatin modifiers, taken from 26 different studies. Our method succeeds in identifying known intricate genetic interactions between chromatin modifiers and transcription factors and uncovers many novel genetic interactions. Our analysis gives the first comprehensive picture of the contribution of chromatin structure to transcription in a eukaryote.

Acknowledgments

Three people followed me through the work presented in this thesis, both my advisors and Amos Tanay. From each of my advisors I have learned an amazing world of research. Both having the same goal of unveiling the mysteries of life, each armed with different tools and point of view. I thank my advisors for the patience, the support, the rich guidance and foremost for giving me the chance to research the wonderful world of biology through their eyes. It was a privilege.

Special thanks and world of gratitude goes to Amos Tanay for sharing with me a bit of his exceptional way of thinking, his guidance, for his productive and practical remarks, and basically teaching me almost everything I know in the field of computational biology.

Last but not least, I thank the group of researchers, members of the Kupiec and Shamir laboratories. Although I didn't have the chance to work with every one of them, each enriched my world with helpful discussions and encouragement. Thank you all for making this journey an interesting one.

Contents

1 INTRODUCTION 6

1.1 BIOLOGICAL PROCESSES ARE REGULATED AT THE TRANSCRIPTIONAL LEVEL 6
1.2 HIGH THROUGHPUT TECHNOLOGIES 7
1.2.1 DNA microarrays measure gene expression at genomic scale 8
1.2.2 TF binding via chromatin immuno-precipitation and microarrays 8
1.3 CHROMATIN STRUCTURE AND CHROMATIN MODIFICATION FACTORS 9
1.4 THE INFLUENCE OF CHROMATIN MODIFIERS ON TRANSCRIPTION 12

2 RESULTS 13

2.1 THE CM COMPENDIUM 13
2.2 THE MODEL 17
2.3 THE STATISTICAL TEST 17
2.4 UME6 REGULATION 18
2.4.1 Expanding the analysis of Ume6 regulation 22
2.5 SYSTEMATIC EXPLORATION OF ALL CM-TF INTERACTIONS 24
2.6 CM-TF INTERACTION RESULTS 26
2.6.1 Gcn4 as a repressor of amino acid biosynthetic genes 26
2.6.2 Regulation of Yap6 through repression by Tup1 30
2.6.3 TBP -dependent Transcription Factors 33

3 DISCUSSION 37

3.1 EXPANDING THE CM COMPENDIUM 37
3.2 IMPROVING THE STATISTICAL MODEL 38
3.3 WORKING WITH DIFFERENT ORGANISMS 38

4 METHODS 40

4.1 K-S ANALYSIS 40
4.2 YEAST GENOME 40
4.3 DATA PREPARATION 40
4.4 ALTERED GENE GROUPS AND THEIR OVERLAP TEST 41
4.5 ANNOTATION ENRICHMENT 41
4.6 HIERARCHICAL CLUSTERING 41

REFERENCES 42

APPENDIX: SUPPLEMENTS 48

List of Figures and Tables

Figure 1: A model for chromatin modifier-mediated transcription 11
Figure 2A: The CM gene expression compendium 15
Figure 2B: The CM gene expression compendium 16
Table 1: Response of the Ume6 cohort in various CM knockout experiments 20
Figure 3: Distribution of expression values for the Ume6 cohort in various CM knockout experiments *21*
Figure 4: Overlap in altered cohort genes 23
Table 2: K-S scores for the Gcn4 cohorts in RPD3C deleted members 29
Figure 5: Distribution of expression values for the Yap6 cohort in various CM knockout experiments *32*
Table 3: K-S scores in experiments disrupting various TBP interactions 34
Supplementary Table A: The publications from which the gene expression data were obtained 48
Supplementary Table C: All the ORFs excluded from our analysis 50
Supplementary Table D: The full cohort-TATA box occupancy level 51
Supplementary Table E: All cohorts found to be enriched in ribosomal related GO terms using TANGO 53
Supplementary Figure 1: Detailed clustering of the compendium 55
Supplementary Figure 2: Overlap of altered genes of the Yap6 and Skn7 cohorts in three experiments 56

1 INTRODUCTION

Transcription regulation is a basic mechanism for controlling biological processes in all living organisms. In particular, the transcription regulation of the genes, being the main functional entities in the genome, is fundamental in controlling many biological processes. Hence, the characterization of the gene transcriptional programs is central in the ongoing exploration of biological processes.

Modern biology is undergoing an information revolution in the last decade, which is apparent in a shift of thinking and practice. The emergence of novel high-throughput technologies enables the quantification of various biological features in a genomic scale. Although this revolution has great advanced the analysis of the complex regulatory networks, it generated a grave need in tools from other research disciplines, such as computer science, statistics and physics.

In this thesis we rely on the integration of methods taken from the field of mathematics and concepts taken from biology. We describe the statistical tools used, and also give many biological examples and discuss their implication to future research. The utilization of the high-throughput information gives us the ability to understand biological mechanism on the systemic level. Our goal was to portray the influence of chromatin structure on the regulation of gene transcription. The results of this thesis were recently published in *Nature Genetics* [74].

We start with a brief review on the field of transcription regulation, high throughput methods and chromatin. We continue with underlining our methods and introducing the compendium of gene expression profiles we have assembled. We finish with an exploration of the intricate interplay between transcription factors and chromatin dynamics, which is a key part in the transcriptional program.

1.1 Biological processes are regulated at the transcriptional level

The information necessary for carrying out most of the biological processes in any cell is encoded as genes in the DNA. According to the central dogma of biological information flow, biological processes begin at the DNA level, which is transcribed to mRNA and then translated into protein. In the scope of this thesis we consider a gene to be an open reading frame in the genome that encodes for a protein. The proper function of the information flow in both the right time and

place is dependent on various levels of regulation. Both transcription (DNA to RNA) and translation (RNA to protein) are regulated positively or negatively by many cellular factors.

The fate of a given cell at any point in time is determined by its particular program of gene expression. In eukaryotic genomes, gene regulation at the transcription level is governed mainly by proteins that facilitate transcription by binding to gene promoters and recruiting the transcription machinery. Another set of proteins prevents transcription of certain genes by binding to their promoters and preventing the recruitment of the transcription machinery. In this thesis we will refer to both types of regulators (positive and negative) by the general term Transcription Factor (TF).

The current estimate is that there are about 25,000 genes in the human genome. In the yeast *S. cerevisiae,* which has a smaller genome, gene organization is simpler and better understood and the number of predicted genes is approximately 6,000. This high number of genes, even in simple organisms, constitutes a complex system that has to be fine tuned and regulated for specific biological process. The major mechanism of regulation is by usage of TFs. Both the binding and activity of TFs can be regulated in order to control the level of transcription of each gene. As such, each gene has a specific transcriptional program, largely manifested by its TF binding sites, which control the level of its expression at any given time.

1.2 High throughput technologies

With the emergence of new high throughput technologies, the classical way of analyzing a biological phenomenon, with experiments aimed at dissecting the role of one or two proteins in a specific process, is starting to shift to examining the phenomenon at the system level. New technologies and algorithms enable the researcher to perform experiments at a genomic scale, thus allowing to ask much more wide-ranging questions. The rapid collection of information needs to be handled by organization methods and also be analyzed by methods taken from the field of information theory. Not surprisingly methods developed in computer science are utilized in the analysis of specific processes as well as in the dissection of the system as a whole. Here we describe two applications for such novel technologies that have revolutionized the way we explore and examine biological systems.

1.2.1 DNA microarrays measure gene expression at genomic scale

First introduced in 1995 [60], microarray technology is a powerful tool enabling, in one experiment, to have a quantification of sequence information at a genomic scale. Microarrays are used for measuring both RNA (e.g transcripts levels), and DNA (e.g. DNA copy number variations and SNP detection). The basic technology used in microarrays consists of DNA fragments attached covalently to a solid surface. The DNA fragments are grouped according to sequence identity and act as probes. By measuring the hybridization level of the tested sample to each fragment set, we can quantify the amount of the features measured. The ability to generate a collection of probes that represent the gene ensemble of an entire genome, allows the researcher to measure each of the genes transcript level in a particular environment and time.

Two main array technologies are currently used: spotted arrays and oligonucleotide microarrays. In spotted arrays the probes are usually cDNAs or oligonucleotides that are spotted on the array and correspond to mRNA. The experiment is a comparison between two samples, each labeled with different fluorophores (e.g. cyanine 3 and cyanine 5), resulting in the ratio of expression for each gene. In oligonucleotide arrays the probes are usually of a fixed length of about 25 nucleotides and are synthesized to match the gene in question. The design of proper controls enables the absolute quantification of RNA levels, but to compare two strains, two arrays need to be used.

The common use of expression profiling is to pinpoint particular genes that participate in a biological process and indeed the microarray technology has become a common tool in many research laboratories, exploring any basic mechanism, from heat shock response in *E. coli* [61] to cell cycle progression in *S. cerevisiae* [62] and cancer prognosis in humans [63].

As gene expression profiles are accumulated in the repositories, we can start to organize the transcriptional programs imposed on the genes. This will be a first step in uncovering the key players that participate in the regulation of biological processes.

1.2.2 TF binding via chromatin immuno-precipitation and microarrays

Recent studies have shown that DNA microarrays can be used in combination with immuno-precipitation, to associate genomic sequences to particular cellular factors [1]. The method called ChIP-on-chip or location analysis starts by treating *in-vivo* a cell population of interest with

formaldehyde. The formaldehyde cross-links proteins to the DNA thus keeping the factors involved in DNA-mediated regulation, such as TFs, bound to the DNA at the regulated sites. The next step is fragmentation of the DNA (into fragments of 0.2-1Kb), commonly done by sonication. Using specific antibodies for the factor in study, the DNA fragments attached to the factor are precipitated (hence, immuno-percipitation). At this step the microarray is used to map each precipitated fragment to the genome (total DNA is used as a reference in the microarray analysis).

As described before, TFs bind upstream of their regulated genes, and various studies have used location analysis to map the binding locations of a particular TF in the genome. In this case, the microarray is designed to have probes for the promoters of each gene. The output of such analysis is the gene set that is predicted to be regulated by the TF studied.

The accumulation of experiments carried out for different TFs in different environments have redefined what we know about the network of transcriptional regulation and are fundamental in understanding the transcriptional programs in living cells.

The ChIP-on-chip technology is still improving, both in precipitation procedures, and also in resolution and coverage of the microarrays. These advancements in technology provide data with higher quality and facilitate the research of more complex systems.

1.3 Chromatin structure and chromatin modification factors

In all living cells the DNA is wrapped around proteins called histones, thus forming chromatin. The repeating subunit of the chromatin, the nucleosome, consists of 146 DNA nucleotides wrapped around the histone core, which carries one subunit of each of the four histones: H2A, H2B, H3 and H4. The main function of the chromatin is to pack the DNA efficiently in the cell, but it was also shown to participate in crucial processes such as mitosis, replication, DNA damage, and gene expression [64]. One important aspect of nucleosome regulation is granted by the accessibility of the histone tails to various proteins. Each histone tail, the N-terminus of the histone, has unique characteristics of amino acid composition and length. The function of the histone tails is not clear, but they have been shown to bind other proteins as well as to undergo post translational modifications. These modifications can change the chromatin state and thus participate in the regulation of many chromatin related processes.

The efficiency of a TF in governing transcription depends on various elements. The affinity of the TF for its promoters is one of them; another crucial factor is the chromatin state of the regulated genes. As stated above, chromatin configuration may determine the accessibility of the promoter to external factors and also the performance of the transcription machinery [2,3]. In eukaryotic cells many proteins influence chromatin structure; these are referred to as chromatin modifiers (CMs). Most CMs are believed to work by affecting histones along the chromatin. The presence of such CMs at the vicinity of a transcribed gene could change the efficiency of transcription by enabling the formation of a chromatin structure needed for TF activity [**Figure 1 A**]. For example, some CMs confer a chromatin state that is condensed and compact. This "closed" chromatin structure is less accessible to the transcriptional machinery. If it occurs in the gene's promoter, it will diminish the transcription efficiency of that gene. CMs are also known to act in the opposite direction, causing the chromatin to adopt a less compact configuration, and thus enabling gene expression.

CMs are usually divided into two main groups according to their biochemical activity: factors that utilize ATP, and factors that are ATP-independent. Among the ATP-independent CMs a widely explored group comprises the histone acetyltransferases (HATs) and the histone deacetylases (HDACs) [2]. The addition of acetyl groups to specific lysine residues on the N-terminal histone tails by the HATs is believed to create a less condensed chromatin structure. Previous work showed that hyperacetylated regions are in general highly transcribed while hypoacetylated regions are silent [4]. Other ATP-independent CMs are the histone methyltransferases. The methylation of N-terminal histone tail has been linked to transcription activation and repression in many organisms [5,6]. Additional histone modifications such as phosphorylation and ubiquitylation are known, and the mechanisms by which these modifications affect transcription constitute one of the most active areas of current research.

An additional group of CMs are the ATP-dependent chromatin remodelers [3]. These highly conserved modifiers usually act as multi-protein complexes that contain an ATPase subunit. The mechanism by which they act is still unclear. Some remodelers are able to destabilize the nucleosomes, allowing the binding of factors to the DNA; others can shift the position of nucleosomes along the chromatin, affecting chromatin structure [2].

A B

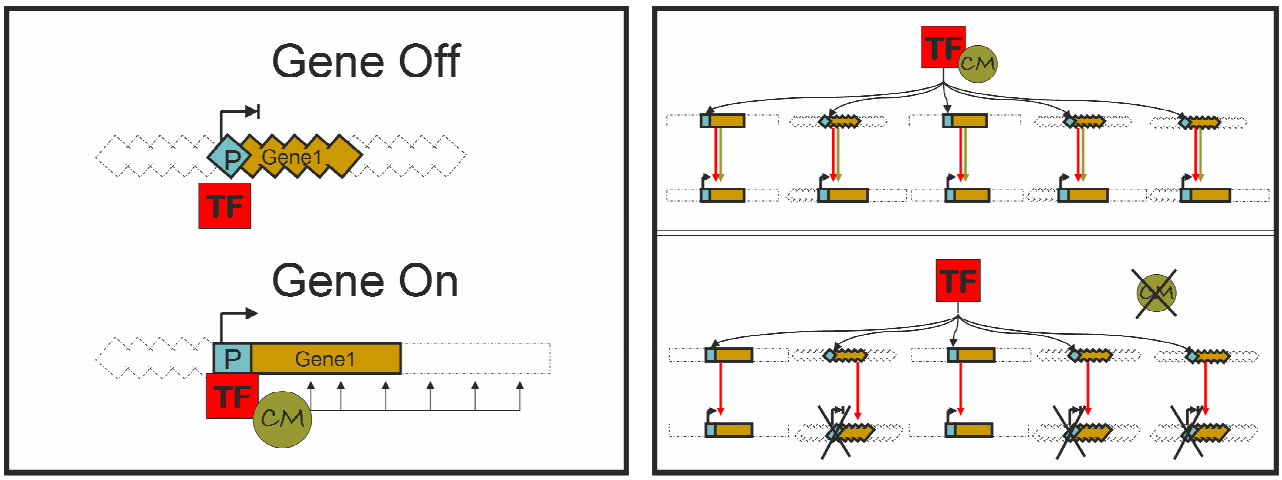


Figure 1: A model for chromatin modifier-mediated transcription. *A) In a "closed" chromatin structure (upper diagram) the transcriptional machinery is less accessible to the gene's promoter; in that situation the efficiency of transcription diminishes. The activity of a CM relaxes the chromatin into an "open" structure (lower diagram), which promotes transcription by facilitating accessibility. In some cases CMs are known to act in the opposite direction, causing the chromatin to adopt a more compact configuration, and thus preventing gene expression [3,4,6,34]. B) Upon activation of a TF, each of the TF target genes is induced according to various parameters, among others its chromatin structure. The interaction between TF and CM enables the activation of those genes located in regions with "closed" chromatin (upper diagram). In strains mutated for CM genes, the absence of the CM will lead to changes in the expression levels of those genes that depend on the CM for transcription (lower diagram).*

1.4 The influence of chromatin modifiers on transcription

As both CMs and TFs carry out related functions with respect to transcription, cooperation is expected between them in transcriptional programs. Such programs can be generated combinatorially, where each factor has its own group (cohort) of regulated genes, and cohorts of different factors intersect in various environments generating a specific program. In this thesis we explore the physical cooperation of CMs and TFs. One simple model of cooperation predicts that some TFs might require the recruitment of a CM to facilitate their activity [**Figure 1 B**]. In such cases the CM could be seen as a cofactor of transcription.

The budding yeast, *Saccharomyces cerevisiae*, is an excellent organism to model eukaryotic transcription regulation. It is one of the most intensively explored eukaryotic organisms, particularly in molecular biology. Most new technologies are implemented first on this yeast and as such, vast amount of information has accumulated regarding its gene functions and their transcriptional networks. In this model organism several interactions between CMs and TFs have been studied in detail [7-9]. We set out to systematically exploring the nature of these interactions.

In this thesis we have assembled a large compendium of gene expression experiments in which various CMs were deleted or genetically manipulated. Using a statistical approach, we have carried out a systematic search for TF-CM pairs that function in concert. We show that our compendium allows a system level overview of the effect of chromatin on transcription and also pinpoints specific TF-CM interplays. We test our method on known examples and shed light on the regulation of such interactions. In addition, we uncover many novel potential TF-CM interactions, which may provide new insights into the mechanism of chromatin structure mediated regulation.

2 Results

2.1 The CM compendium

Two types of high-throughput data set were used in this thesis. To obtain a complete list of cohorts and their respective regulating TFs we used the comprehensive dataset of Harbison et al. [1] (ChIP-on-chip, see Introduction),. In their dataset, a statistical model was applied on the raw signals of the array and a measure of significance (p-value) was given for each TF and gene. This p-value reflects the level of confidence for finding the particular TF bound at a particular gene's promoter. Harbison et al. carried out location analysis for 204 proteins presumed to have affinity to the DNA and thus that could function as transcription factors. The location experiment were conducted in YPD (Yeast extract Peptone Dextrose), normal rich medium. For 84 TFs the location analysis was carried out also in at least one more condition. The conditions were chosen to resemble to environments for which the TF is expected to become active, as for some TFs the appearance of a regulated cohort is dependent on the condition in question. For example, Msn2 and Msn4 together regulate the main general stress response; however, in rich medium their activity is not expected. Thus, we would not expect to see a strong binding to the promoters of their cohort in rich medium, and only in conditions of stress (such as extreme heat) their cohorts should be notable. We selected for each of the analyzed TFs, the group of genes it binds to (its cohort), by applying a strict binding threshold of p-value < 0.001. According to the original publication, this ensures a low level of false positives (<8 %) [1].

The second data set, a gene expression compendium, was gathered from the literature, and contains experiments carried out with yeast strains in which particular CMs were deleted or genetically modified to loose their catalytic capability [**Supplementary Table A**]. This compendium, consisting of 170 gene expression profiles taken from 26 different publications, covers more than 60 potential interacting CMs. CMs usually operate in large complexes of proteins, and in a manner that is not fully understood share many components, even between complexes of opposite biochemical activity. The compendium is comprehensive for CM complexes it covers, having at least one member (usually the catalytic one) of most of the known CM complexes in *S. cerevisiae*. Such complexes function as histone acetyl transferases (HATs: the NuA4, HAT1 and SAGA complexes) and histone deacetylases (HDACs: the RPD3, HDA1 and SET3 complexes),

respectively, adding and removing acetyl groups from conserved lysine residues on the histone tails. Other complexes function as ATP dependent chromatin remodelers (the SWI/SNF, SWR1, INO80, ISWI and RSC complexes). For example, RSC complex is presumed to generate negative supercoiled DNA [65] and the SWR complex substitutes H2A histone by its variant H2A.Z (Htz1) in the nucleosome, giving the nucleosome different characteristics [21]. Yet for most of the remodelers, although clearly having important functions in the cell (e.g SWI/SNF), their biochemical function is still not clear. In addition to the two main sets of complexes described above, the compendium also encompass histone methyltransferases complexes (the COMPASS complex), and other chromatin-affecting and co-factors such as Spt10, Sir proteins, TBP, etc. [**Figure 2 A**].

The compendium described above is the first attempt to collectively build a resource that can be a starting point in any analysis of the involvement of CMs in transcription. The work described in this thesis relates to that part of transcription that is regulated by the transcription factors, but other angles of research could be easily implemented.

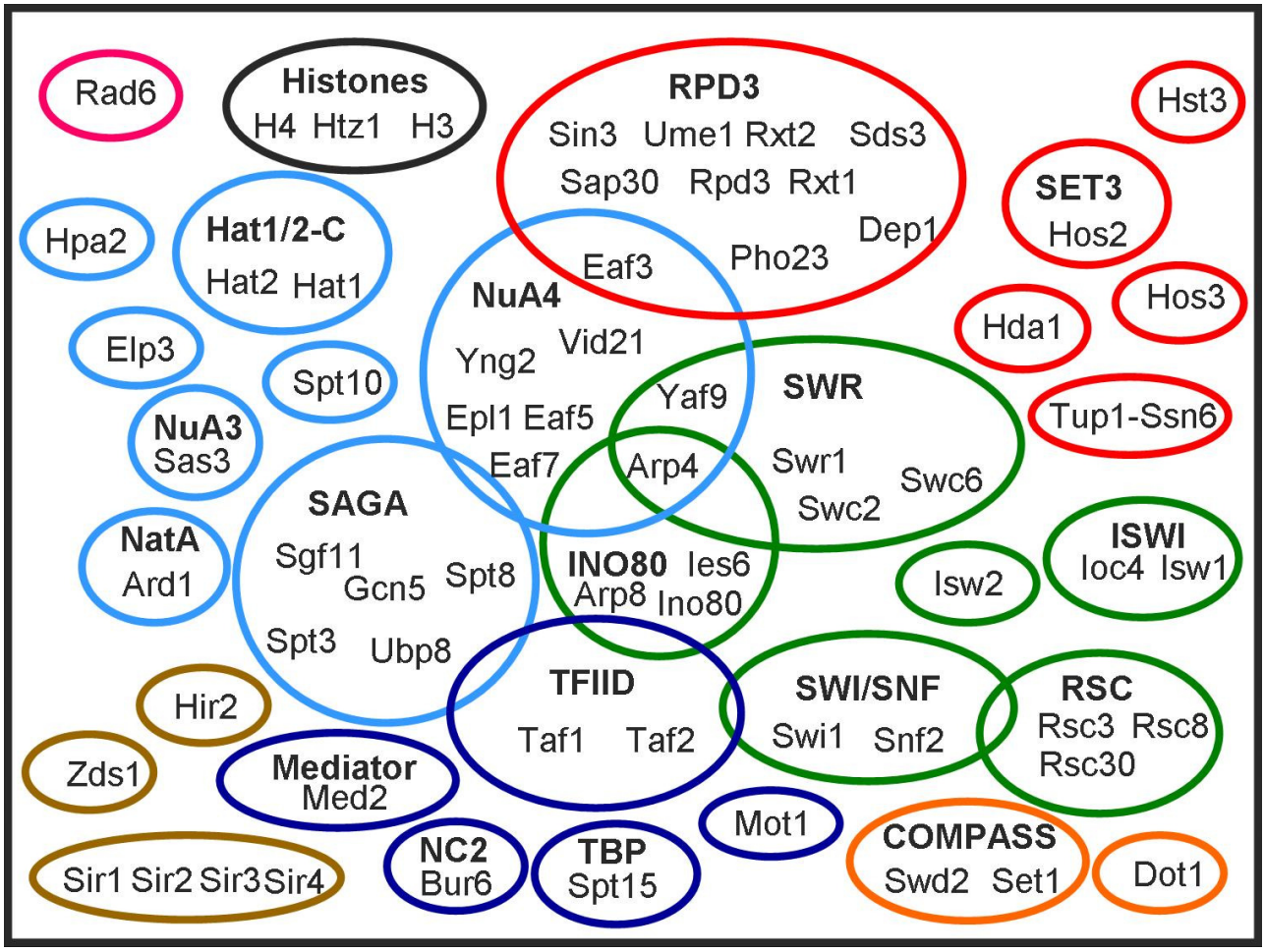


Figure 2A: The CM gene expression compendium. *The expression profiles available in the compendium. Each of the listed CMs has one or more profiles in the compendium, created by a genetic alteration of the CM in the yeast genome. CMs that belong to the same complex are circumscribed by an oval, with the complex name in bold. Colors indicate the CM's proposed biochemical activity: HATs in light blue, HDACs in red, methyltransferases in orange, Ubiquitin-conjugating enzymes in magenta, chromatin remodelers in green, TAF-related factors in dark blue, silencing factors in brown and Histone subunits in black. The full references of the studies are available in Supplementary Table A.*

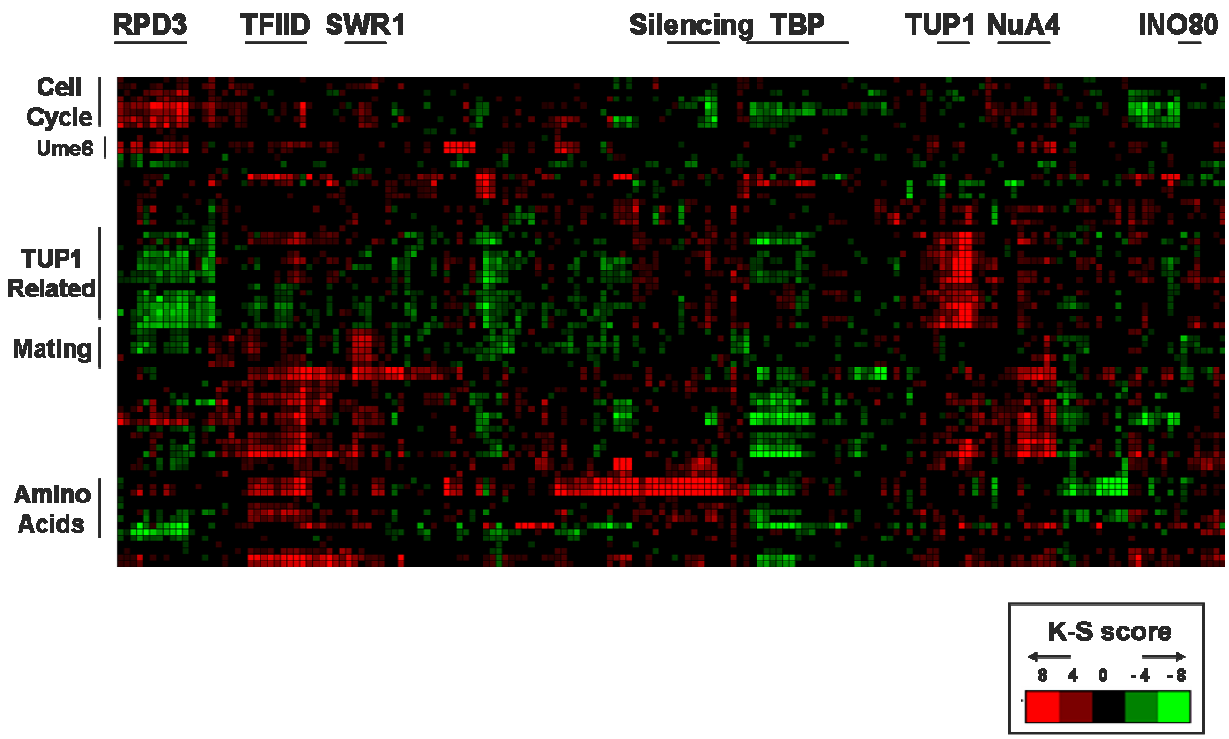


Figure 2B: The CM gene expression compendium. *B) Clustering of the compendium. Rows represent TF cohorts and columns represent conditions. Colors indicate CM-cohort K-S scores. To obtain a global view of the TF-CM interaction landscape, we hierarchically clustered the cohorts and conditions according to their K-S scores (positive scores in red and negative in green). Groups of functionally related TFs (ordinate) and functionally related conditions (abscissa) are marked. Using the global view we can see that mutations affecting general repressors, like Tup1, show a global activation of most cohorts while mutations in general activators, such as the TATA Binding Protein (TBP), exhibit the opposite effect. The detailed hierarchical clustering solution is available in Supplementary Figure 1.*

2.2 The model

Our goal was to investigate whether there are selective interactions between a TF and any of the CMs in our compendium. The rationale was as follows: Mutations affecting a particular CM are expected to have a broad effect on gene expression, affecting many genes. If, however, activation or repression of genes by a specific TF depends particularly on the activity of a certain CM, we expect to see that mutations in the CM cause a preferential effect on expression of the TF target genes [**Figure 1 B**]. To test whether the regulation by a particular TF is affected by deletion of a CM, we partitioned the gene expression profile of a CM-mutated strain into two groups: the TF cohort (the genes bound by the TF and thus directly regulated by it) and the control group, consisting of the rest of the genes in the genome. If no particular interaction (direct or indirect) exists between the CM and the TF, we expect to observe the same distribution of gene expression levels in both groups. In contrast, if the TF and CM cooperate in controlling the expression of a subset of genes, deletion of the CM should cause a differential change in expression of these genes [**Figure 1 B**].

2.3 The statistical test

To evaluate the difference in the distribution of gene expression values in the two groups (the cohort and the control) we used the Kolmogorov-Smirnov (K-S) statistical test [10]. This test is appropriate for two main reasons:

1) It is a non-parametric test. Due to its non-parametric nature, the test is robust and does not require any linear normalization of the expression values. This feature is imperative when dealing with heterogeneous sources and thus, it is suitable to our dataset which contains diverse expression profiles originating from many studies. Another benefit that comes from this feature is that no threshold is needed to be affixed. Thus, no data is lost because of arbitrary thresholds (e.g. two-folded expression of genes for the definition of activation) and even the slightest trends can be observed.

2) It provides an exact p-value. Under the assumption of gene independence (which is not always true, but nevertheless many studies make it), the test provides a p-value for the KS value measured (the statistic). The KS statistic, plainly speaking, is the maximal percentile difference between the two distributions over all possible expression values. As such, the p-value indicates the

statistical significance of the difference between the two distributions (see Methods). This feature is also imperative when testing so many hypotheses (each TF and CM is a hypothesis test for their interaction), as the alternative simulation procedure to evaluate the significance of an interaction is too computational causative.

The log of the K-S p-value provides a measure of the discrepancy in expression of the TF cohort from the rest of the genes when the CM gene is mutated. In order to indicate the direction of discrepancy, we have added to it a positive or negative sign and used it as a score to rank the CM-TF interactions. Positive scores indicate that the TF cohort is activated in the particular CM experiment, whereas negative scores imply reduced expression of the TF cohort (see Methods). Hence, the K-S score expresses both the direction and significance of the disparity between the two distributions.

2.4 Ume6 regulation

We first tested our method on the well-characterized example of the TF Ume6, a central regulator of early meiotic genes, which is known to regulate its cohort through interactions with CMs [7]. During vegetative growth, binding of Ume6 upstream of specific early meiotic genes facilitates the recruitment of the RPD3 complex (an HDAC) and ISW2 complex (an ATP dependent chromatin remodeler) [11,12]. RPD3 complex was shown to remove acetyl groups from histones H3 and H4 [13]. ISW2 complex is presumed to have the ability to slide nucleosomes along the DNA thus alternating between open and closed chromatin states. It is not clear what is the sequence of event that leads to the recruitment of the two CMs, whether it is RPD3 or ISW2 that arrives first and recruits the other, but the hypoacetylation by RPD3 complex and the chromatin remodeling by ISW2 is presumed to create a condensed chromatin structure that prevents gene expression [11]. Thus, Ume6 keeps its cohort genes in a silent state, halting their function by preventing their expression. During entry to meiosis, Ume6 preferentially interacts with the activator Ime1. This alternative interaction releases the CMs, which promote expression of the meiosis -related cohort [14].

The Ume6 cohort, as defined by Harbison et al. [1], consists of 131 genes. Due to its large size, we can deduce with high statistical confidence the activity level of the Ume6 repressor in the entire CM gene expression compendium. As expected, deletion of *UME6* leads to a significant shift

in the expression pattern of the Ume6 cohort [**Table 1**]. In particular, a strong effect was seen in an experiment carried out by Fazzio et al. [11]: While deletion of UME6 did not lead to a general shift in gene expression, the Ume6 cohort exhibited a strong de-repression that was articulated in increase of their expression (K-S score = 12.72) [**Figure 3 A**].

Ume6 acts as a repressor only through its ability to recruit the RPD3 complex and the Isw2 chromatin remodeler to its binding location [11]. According to this dogma not only a deletion of *UME6* but also a double deletion of *ISW2* and *RPD3* should de-repress all of Ume6-regulated genes. For the synergistic cooperation of Isw2 and Rpd3, a single deletion of *ISW2* or of *RPD3* should result in a partial de-repression of the Ume6 cohort. Our results show exactly this effect: while the Ume6 cohort exhibited a significant activation in an experiment carried out with the doubly deleted *isw2*Δ *rpd3*Δ strain [**Figure 3 B**], a less significant effect was seen for a strain individually deleted for *RPD3* and no effect was observed in the *ISW2*-deleted strain [**Table 1**]. It is important to note is that the activation of the Ume6 cohort could in principle result from an indirect repression of Ume6 itself by each of the CMs. The expression level of the *UME6* gene was monitored in each of the experiments. No repression of *UME6* was observed in the single deleted strains; in the doubly deleted strain, *UME6* even showed a ~4-fold up-regulation, probably an attempt to compensate for the mis-regulation of its cohort (**Supplementary Table B**). Examination of the de-repressed genes (all genes with expression Z-score > 1, see Methods) from the Ume6 cohort in both the *ume6*Δ and the *isw2*Δ *rpd3*Δ experiments reveals a significant overlap (hyper-geometric $p < 4.6*10\text{-}7$) [**Figure 4 A**]. The similar effect observed in both experiments points to the common mechanism of regulation by Ume6 and Isw2 with Rpd3.

Publication	KS Score	Condition
Fazzio *et al.* [19]	**13.02**	*ume6*Δ
Sabet *et al.* [15]	**11.29**	*rpd3*Δ *H3*Δ*N* vs. *H3*Δ*N*
Bernstein *et al.* [59]	**7.54**	*ume6*Δ
Fazzio *et al.* [19]	**6.7**	*isw2*Δ *rpd3*Δ
Bernstein *et al.* [59]	**5.48**	*rpd3*Δ
Fazzio *et al.* [19]	-0.36	*isw2*Δ

Table 1: Response of the Ume6 cohort in various CM knockout experiments. *Six selected gene expression experiments taken from the CM compendium are listed. The K-S score of the Ume6 cohort's disparity from the rest of the yeast genes is presented (see Methods). A significant disparity is defined as scores with absolute value above 5.41 (Bonferroni corrected p-value < 0.05). The experiments were chosen to test the mutual contribution of Isw2 and Rpd3 on the Ume6 cohort. The doubly deleted strain shows a stronger effect on the Ume6 cohort compared to the corresponding singly deleted strains.*

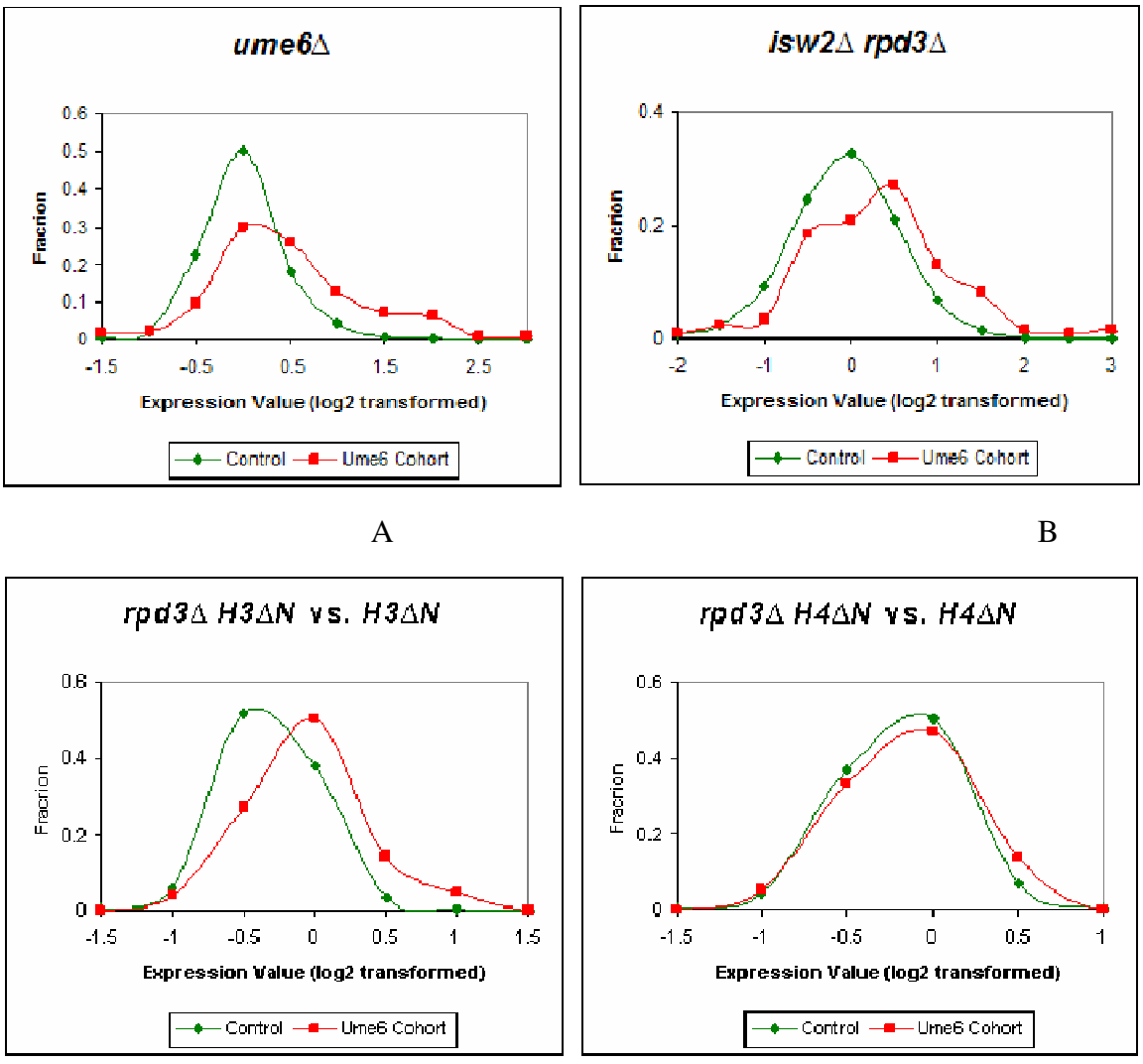


Figure 3: Distribution of expression values for the Ume6 cohort in various CM knockout experiments. *Distributions of expression levels (log2 transformed) are presented for the Ume6 cohort and the control group (rest of the genes). Red: the Ume6 cohort. Green: the control group. A) Strain deleted for UME6 [11]. B) Strain doubly deleted for ISW2 and RPD3 [11]. C) Strain deleted for RPD3 along with a deleted N-terminus of histone H3 compared to an isogenic strain carrying only the histone mutation [15].D) Strain deleted for RPD3 and a deleted N-terminus of histone H4 compared to an isogenic strain carrying only the histone mutation [15].*

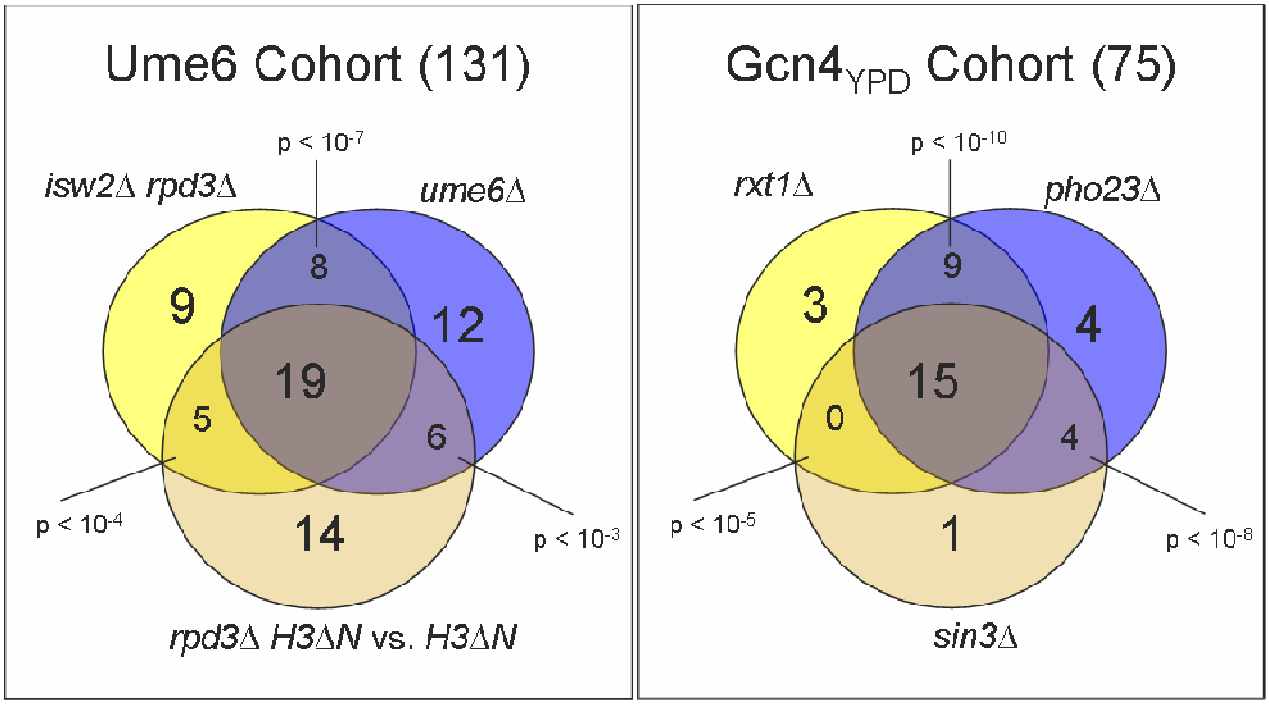
2.4.1 Expanding the analysis of Ume6 regulation

Reassured by the ability of our methodology to expose the well-characterized contribution of Isw2 and Rpd3 to Ume6 regulation, we carried out a systematic exploration of the Ume6 cohort in the entire CM compendium. This exploration allowed us to uncover novel characteristics of Ume6 regulation.

Sabet et al. [15] explored the relationship between the transcription regulation by Rpd3 and the amino termini of histones H3 and H4. Since the deletion of the N-terminal domain of histones prevents their regulation by most ATP-independent CMs, strains were constructed carrying mutant versions of either histone H3 or histone H4, in which the N-terminus of the protein was deleted (*H3ΔN* and *H4ΔN,* respectively). To test whether Rpd3 has an effect on gene expression independent of H3, the *H3ΔN* strain for which *RPD3* was also deleted was compared to the isogenic *H3ΔN* strain. This experiment showed a highly significant and specific disparity in the expression of the Ume6 cohort [**Figure 3 C**]. As in the previous experiments the activation of the Ume6 cohort was not a consequence of the repression of Ume6 itself (Ume6 expression log2 value of 0.7). The activated genes from the Ume6 cohort (Z score > 1) in this experiment share a significant overlap with those de-repressed in the strain deleted for *UME6*, as well as with the strain doubly deleted for *ISW2* and *RPD3* (hyper-geometric $p < 10\text{-}3$ and $p < 10\text{-}4$, respectively) [**Figure 4 A**]. The commonality of affected genes in the *H3ΔN* strain experiment with the *ISW2-RPD3* and *UME6* strains points to a shared mechanism of regulation.

Interestingly, in the parallel experiment carried out with *H4ΔN* no effect was observed [**Figure 3 D**]. *In vitro* studies have implicated both the H3 and H4 histones in the binding of Isw2 to nucleosomes [16-19]. The additive effect of the *RPD3* deletion to the H3 mutation, as opposed to the H4 mutation, suggests that histone H4, but not H3, is likely to work with Rpd3. In addition, the similar effects obtained in the *rpd3Δ* strain lacking the N-terminus of histone H3 and in the *rpd3Δ* strain lacking *ISW2* suggest that H3 tails play a central role in the recruitment of Isw2 by Ume6.

Hence, in the case of the intricate transcription regulation by Ume6 our method enabled the discovery of known Ume6 CM co-factors solely by exploring the behavior of the Ume6 cohort in various experiments. Our results also shed new light on Isw2 participation in the Ume6 repression mechanism and support the involvement of histone H3 N-terminus in the regulation of expression by Isw2.



A B

Figure 4: Overlap in altered cohort genes. *Level of overlap between altered cohorts in various gene expression experiments (see Methods). A) Overlap in de-repressed Ume6 cohort genes in three experiments. Out of 131 Ume6 cohort genes, 45 showed a notable induction (Z-score > 1) in a UME6 deleted strain [11], 41 in a doubly deleted ISW2 RPD3 strain [11] and 44 in strain deleted for RPD3 along with a deleted N-terminus of histone H3 compared to an isogenic strain carrying only the histone mutation [15]. The significance of the overlap between each pair of strains is indicated (hyper-geometric p-value). B) Overlap in activated Gcn4YPD cohort genes in three experiments. Out of 75 Gcn4YPD cohort genes, 32 showed a notable induction in a PHO23 deleted strain, 27 in a RXT1 deleted strain and 20 in a SIN3 deleted strain [25].*

2.5 Systematic exploration of all CM-TF interactions

To reduce the number of hypothesis testing, only TF cohorts with a sufficiently large number of genes were taken for analysis. Out of the 204 TFs analyzed by Harbison *et al.* [1], 49 generate cohorts large enough; out of these 19 were analyzed in more than one environment. In total we were able to analyze 75 cohorts (see Methods). The behavior of each of these cohorts was tested against the entire compendium. Our test generated 4645 TF-CM pairs with a K-S p-value < 0.05 and, after Bonferroni correction for multiple testing, 531 significant pairs remained (|K-S score| > 5.41, see Methods) [**Supplementary Table B**].

The significant pairs were obtained from 55 different cohorts (defined for 35 TFs) and 129 gene expression experiments, covering most of the complexes known to participate in chromatin structure regulation [**Figure 2**]. In total we obtained 287 unique pairs of TF-CM [**Supplementary Table F**] giving a first comprehensive picture of the TF contribution in chromatin structure regulation in a eukaryote.

The average number of significant pairs for each TF cohort is 9.6 and it is 4.1 for each CM profile. Some TFs define cohorts that behave more promiscuously; the Hap4 cohort, for example, shows significant disparity in 26 experiments, which associate it to 16 different CMs, as opposed to the Reb1 cohort that shows disparity in only one experiment. Several factors may determine this behavior: (1) Better quality of the location analysis results may lead to a better definition of the cohort. The reduction of noise from the cohort will enable to detect more subtle trends. (2) Cohorts containing a larger number of genes usually provide higher statistical significance. As described above many TFs were excluded from the analysis as their cohorts were too small, eluding many potential interactions. (3) The biological activity of the TF should be relevant to both the location analysis and the gene expression experiment. For the identification of a TF that is active only in particular conditions, we would need to have a location analysis result in this condition to focus on its cohort, but also we would need a gene expression profiling of a deleted strain of its interacting CM in that same condition.

Similarly, some CM mutants lead to a preferential change of expression in several cohorts (e.g.: 18 cohorts (from 11 TFs) with significant deviation in the *ssn6*Δ experiment [20] and only one cohort for the *vps72*Δ experiment [21]). The number of significant cohorts per CM is effected by: (1) as for the TF, the quality of the expression profiling and the condition for which the

experiment was carried on. Some CM might function in particular environments. (2) The level of robustness in the system, where the effect should vary due to the level of the CM importance. CM complexes share many component and overlap in biochemical function (more then 4 complexes that are considered HDACs or HATs). Catalytic members of the CM complexes are expected to have, in general, a more widespread effect.

A global view of the compendium and its interplay with the TF cohorts is obtained by dual-hierarchical clustering of CMs and of TFs according to similarity of their K-S score profiles across all experimental conditions and cohorts [**Figure 2 B**]. This procedure enables, on one hand, the visualization of common trends of different cohorts in response to all the CM perturbations, and on the other hand, the detection of CMs with similar specificity according to their effect on the cohorts. The resulting representation shows that TF cohorts are grouped according to various biological processes: cell cycle, amino acid biosynthesis, mating and more. The inclusion of two TFs in the same group is sometimes due to a high level of overlap between their cohorts, but in many cases reflects common CM-mediated mechanisms of regulation. Our results suggest that the genome is organized along functional similarities; also we show that cohorts that are involved in common biological processes are affected by similar CMs. In the case of cell cycle progression, for example, TFs affecting different stages are nonetheless grouped together, implying a common interplay with CMs. Interestingly, the well characterized TF Ume6 (see the section above) is placed in the hierarchical clustering near the cell cycle TFs, although the Ume6 cohort shares almost no gene with the cell cycle cohorts. Inspection of the global view indeed reveals that, like Ume6, all the cell cycle TFs show a relative induction in various experiments in which the *RPD3* complex members were mutated [**Figure 2 B**].

When clustering the CMs according to the similarity of their K-S scores in each cohort, well-defined complexes are grouped together. For example experiments with strains deleted for members of the NuA4 complex (Eaf3, Eaf5, Eaf7, Yng2, Vid21, Epl1 and Arp4) were hierarchically clustered along with Rsc8 (RSC) and Isw1 (ISWI) (both shown to interact physically and genetically with the NuA4 complex [21-23]). This is remarkable, taking into consideration that the data were derived from experiments carried out with yeast of different genetic backgrounds and using different experimental protocols (e.g., NuA4 experiments were taken from three different publications [21,24,25]). Similarly, whenever a CM deletion was analyzed independently in two laboratories, the results cluster nicely together. In addition, factors that are known to act as global

activators/repressors, like the TATA Binding Protein or the Tup1 repressor, manifest a global effect on the genome, reflected by the joint induction/repression of many of the cohorts [**Figure 2 B**]. In the disturbance of the TATA binding protein (see section ahead) this comprehensive down regulation, which is defined by comparing the TF cohort to the rest of the genes, seems paradoxical - if all cohorts are down, then who is up? But these results point to the fact that genes that have a strong binding of TF in their promoters, are more dependent on the TATA binding protein mechanism of transcription induction. On the other hand, deletion of *TUP1* (see section ahead), which is a know repressor of genes, manifest a comprehensive de-repression of the TFs that are know to be regulated by its repression mechanism.

2.6 CM-TF interaction results

Our analysis reveals many novel putative TF-CM interactions. In the previous section we described the overall CM-TF interaction landscape. In this section we focus on several interesting cases where a mutation in a specific CM has a significant effect on a TF cohort. The full table of results is available as **Supplementary Table B**.

2.6.1 Gcn4 as a ***repressor*** of amino acid biosynthetic genes

The Gcn4 TF activates many genes under conditions of amino acid (AA) starvation (reviewed in [26]). In accordance with the positive role of Gcn4, its cohort was strongly repressed in the expression profile of a *gcn4*Δ strain [20] and strongly activated in a strain over-expressing *GCN4* [29]. Initiation of transcription by Gcn4 was shown to be dependent on many co-activators [9], including the CMs SWI/SNF and SAGA. These chromatin modification complexes (a chromatin remodeler and a histone acetyltransferase, respectively) are recruited by Gcn4 in response to AA starvation [27,28], as such Gcn4 is a good candidate for the exploration of other cooperation with chromatin modification complexes.

As an activator of many AA biosynthesis pathways, the mechanism by which Gcn4 promotes the transcription of its cohort when starved for an AA is tightly and complexly regulated. This feature makes Gcn4 a good example of a context-dependent transcription factor. As described above, the cohort defined by each location analysis experiment is highly dependent on the experimental conditions. Exploring cohorts defined for the same TF under different conditions

assists us in the study of the TF's regulatory program. Harbison *et al.* [1] defined the Gcn4 cohort in an experiment carried out in rich medium ($Gcn4_{YPD}$), but also in cells exposed to sulfometuron methyl (SM), an inhibitor of several AA biosynthesis pathways ($Gcn4_{SM}$). The $Gcn4_{SM}$ cohort is larger and consists of 189 genes, but interestingly the $Gcn4_{YPD}$ cohort, which consists of only 75 genes, is a subset of the SM cohort [1]. These results indicate that Gcn4 binds to its core cohort under optimal growth conditions, and not only after AA deprivation. The reason for the binding of Gcn4 to its core cohort is not clear and might point to a function Gcn4 maintains even in its non active state in rich media.

In an experiment done by Keogh *et al.* [25] the RPD3 complex was thoroughly analyzed using biochemical and genetic tools, among others co-immunoprecipitation of each member of RPD3 complex. The authors defined two distinct RPD3 complexes, RPD3(L) and RPD3(S), which share a core of three proteins: Rpd3, Sin3 and Ume1. Eaf3 and Rco1 uniquely belong to the RPD3(S) small complex, whereas Pho23, Rxt1 and Rxt2 are specific to the larger RPD3(L) complex. Surprisingly, our results show a clear activation of the Gcn4 cohort when subunits of the large RPD3(L) complex are deleted [**Table 2**]. Gcn4 activation was not due to activation of Gcn4 itself (expression levels of *GCN4* are available in **Supplementary table B**). Moreover, the activated genes in each RPD3(L)-deleted strain experiment were highly overlapping [**Figure 4 B**] emphasizing the essential contribution of the RPD3(L) complex, and not a particular member, to the regulation by Gcn4. Interestingly, when subunits of the small RPD3(S) complex were deleted, the cohort showed no disparity from the rest of the genes [**Table 2**]. And accordingly core member that belong to both complexes show a milder effect. Thus, the RPD3(S) results provide an appropriate control and show that, in addition to the specific linkage of the RPD3(S) to Set2 methyl-transferase [25], the two complexes have also functionally divergent roles in the regulation by Gcn4. Probably in the affinity specification to other factors, such as Gcn4.

The gene expression experiments carried out by Keogh *et al*. were not done in AA -limiting conditions but in rich medium. However, the Rpd3 effect can be seen on all Gcn4 cohorts. The additional targets, available in the $Gcn4_{SM}$ cohort, preserve the described trend and even exhibit stronger activation in the experiments carried out with RPD3(L) deleted members, which strengthen the significance of the result. The specificity of this result is strengthened by the fact

that, like the $Gcn4_{YPD}$ cohort, in the RPD3(S) deleted members the additional targets match the control distribution [**Table 2**].

The effect described above points to a wide participation of the RPD3 complex in the regulation by Gcn4, an effect that is seen even on weak targets of Gcn4 in rich medium. Gcn4 has been shown to use SAGA, a histone acetyl transferase, to activate its cohort [9]. Our results point to the opposite biochemical reaction, removal of acetyl groups from histones, performed by the RPD3 HDAC complex, as a mechanism that can maintain its target genes in an inactive state. Functional analysis on the activated genes in the experiments in which RPD3(L) members were deleted reveals an over-representation of arginine biosynthesis genes (all 8 genes involved in arginine biosynthesis present in the $Gcn4_{YPD}$ cohort show increased expression, $p<0.001$). Thus, our results suggest that Rpd3 and Gcn4 act as *negative* regulators of the arginine biosynthesis pathway under optimal growth conditions.

$Gcn4_{SM}$ KS score	**$Gcn4_{YPD}$ KS score**	**Constituent of**	**Condition**
17.34	**7.73**	RPD3(L)	*pho23*Δ
13.01	**5.92**	RPD3(L)	*rxt1*Δ
6.69	4.71	RPD3(L)	*rxt2*Δ
10.15	4.05	Core Complex	*sin3*Δ
9.16	2.8	Core Complex	*rpd3*Δ
6.42	2.71	Core Complex	*ume1*Δ
0.004	0.06	RPD3(S)	*eaf3*Δ
0.040	0.022	RPD3(S)	*rco1*Δ

Table 2: K-S scores for the Gcn4 cohorts in RPD3C deleted members. Two cohorts were defined, one in rich medium ($Gcn4_{YPD}$) and another in AA limiting medium ($Gcn4_{SM}$). The table presents the K-S scores of each cohort in strains deleted for various RPD3 complex members [25]. The RPD3 complex contains two alternative sub-complexes, RPD3(L) and RPD3(S), which share the core Rpd3-Sin3-Ume1 proteins. Expression profiles were obtained for deleted members of both complexes. Significant K-S scores are highlighted ($p<0.05$, corrected for multiple testing). A significant activation of $Gcn4_{YPD}$ cohort is notable specifically in the RPD3(L) deleted members strains only. The activation becomes stronger for the extended $Gcn4_{SM}$ cohort, and covers also the deleted core complex strains.

2.6.2 Regulation of Yap6 through repression by Tup1

Having tested our methodology on the well characterized example of Gcn4, we asked whether novel interactions could be revealed for other TFs. In particular, a lot can be learned about a TF by examining its interaction with CMs, so we decided to focus on the TF Yap6. Yap6 has sequence similarity to AP-1 [30] and has been linked to lithium and sodium resistance [31], other than that very little is known about the Yap6.

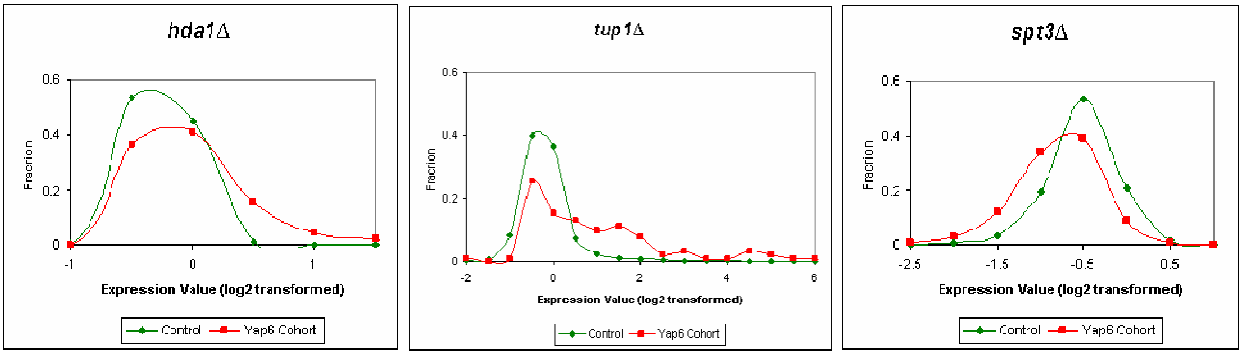
Examination of the behavior of the Yap6 cohort against the entire compendium reveals a range of potential interactions with various CMs [**Supplementary Table B**], which is surprising given the anonymity of Yap6. A good example of such interaction is a significant activation of the Yap6 cohort in a strain deleted for *HDA1* (K-S score = 9.03) [**Figure 5 A**]. Hda1 is the catalytic member of the *HDA1* HDAC complex known to be involved in gene expression and silencing [32]. An interesting feature of Hda1, among other, is its participation in the repression mechanism of Tup1 [33]. This example is interesting since Tup1 is an example of a repressor that acts as a mediator between TFs and CMs. Tup1 has the ability to recruit CMs to confer repressed chromatin structure [34]. Since Hda1 is one of the Tup1 -recruited CMs we were interested in the relation between Tup1 and Yap6. To test whether Yap6 works through Tup1 we examined the Yap6 cohort behavior in a gene expression experiment carried out in a strain deleted for *TUP1* [20]. Indeed the Yap6 cohort exhibits a strong activation in the *tup1*Δ strain experiment (K-S score =18.6) [**Figure 5 B**], implying that Tup1 indeed participates in the regulation by Yap6. This activation was also found to be even stronger than in the *hda1*Δ strain experiment which strongly suggesting that Hda1, although fundamental, is not unique in the repression mechanism of Yap6 cohort by Tup1 and that other CMs might participate as well [34]. Another confirmation that indeed Tup1 and Hda1 repress these genes by a common mechanism, is the high level of overlap between the activated genes (Z score > 1) in both experiments (23 genes, hyper-geometric $p < 0.003$) [**Supplementary Figure 2**].

A brief exploration of characteristics and function of the genes affected in both *HDA1* and *TUP1* deleted strains, reveals that they are mostly subtelomeric (15 out of 23 genes; $p < 10^{-10}$) and are highly enriched for members of the hexose-transport family (5 genes, $p < 0.001$). Thus, our results clearly indicate a role for Yap6 in the regulation of sugar transport that, surprisingly, is

affected by Tup1 and Hda1, and not by the CMs usually implied in silencing of subtelomeric genes, such as the Sir proteins and Set1/Isw1 [35].

Following the Gcn4 example described in the previous section, we went on to search for CMs that affect the Yap6 cohort in a manner opposite to that of Tup1-Hda1. We found that the Yap6 cohort is significantly down regulated in a strain deleted for *SPT3* (K-S score = 7.9) [**Figure 5 C**], a key member of the SAGA complex [36]. SAGA is a well-characterized HAT complex that acts as a global inducer [37]. Interestingly, although Spt3 is a SAGA member and was shown to be required for the recruitment of the TATA-Binding-Protein (TBP) to various SAGA-regulated genes [38,39], no effect on the expression of the Yap6 cohort was observed in mutants deleted for either *GCN5* (SAGA's catalytic subunit) or in strains carrying various mutations in TBP (data not shown). Many of the SAGA complex components can also be found in a different complex, named SAGA-Like complex (SILK), which also acts as an inducer of genes [40]. Spt3 was previously shown to regulate genes through SILK in a manner that does not require SAGA's HAT activity [41], and this kind of mechanism is suggested by our results as well. Thus, our results uncover a collaboration between Yap6 and Spt3 that is independent of *GCN5*, suggesting the existence of an uncharacterized interactor that provides HAT activity.

Analysis of the genes of Yap6 that are *repressed* (Z score > 1, see Methods) in *SPT3* deletion demonstrate an extensive overlap with those *activated* in strains deleted for *HDA1* (14 genes, $p < 0.009$) [**Supplementary Figure 2**]. The high overlap between the genes suggests an acetylation homeostasis achieved by the Tup1-Hda1 and Spt3 -related HAT activities.



A B C

Figure 5: Distribution of expression values for the Yap6 cohort in various CM knockout experiments. *Distributions of expression levels (log2 transformed) are presented for the Yap6 cohort and the control group (the rest of the genes). Legends are as in Figure 3. A) strain deleted for HDA1 [58]. B) Strain deleted for TUP1 [20]. C) Strain deleted for SPT3 [37].*

2.6.3 TBP -dependent Transcription Factors

As described above, CMs interact with TFs to regulate gene expression. The same principle should be applicable to additional proteins that, like the CMs, have a wide influence on transcription. The TATA Binding Protein (TBP), a central activator of transcription, is such a factor. The TBP regulates gene expression by binding AT-rich sequences called TATA boxes, affecting transcription of most of the genome, and collaborates with co-factors, many of which are CMs. Among the TBP co-factors we can find Mot1 (SWI/SNF like), Spt3 (HAT), Taf1 (HAT) and the inhibitor NC2 (reviewed in [42]).

To explore possible interactions between TFs and TBP and to check the involvement of each of the TBP co-factors in that regulation, we employed our method on the gene expression data set generated by Chitikila *et al.* [43]. As TBP is an essential component of the cell a gene expression profiling of a strain deleted for TBP is not possible. Chitikila *et al.* overcome this problem by creating strains mutated for various component of the TBP. By over-expressing the TBP mutants in the cell they managed to modify the activity of the TBP. Taf1 contains a domain called TANDI, which mimics the TATA box and competitively inhibits the TBP interaction with the TATA box [43]. Another TBP inhibition mechanism is through TBP self dimerization.

In order to characterize inhibition mechanisms, Chitikila *et al.* thus created mutations that affected TBP dimerization (TBPd), interaction with Taf1 through deletion of the TANDI region (DeltaT) or interaction with NC2 through a mutation in the NC2 –binding region (NC2). The NC2 complex and Taf1 are considered inhibitors of the TBP transcription induction [42]. NC2 acts by competitively inhibiting the TBP association to TFIIA and TFIIB [43].

Over-expression of the TBPd mutations leads to a preference in the use of the non-dimerizing mutated TBP. As such this loss of dimerization leads to a reduced functional capability [43]. The loss of function attributed to the TBPd mutation allowed us to use it in our analysis to search for TBP-dependent TFs. The other mutants were used to investigate the regulatory contribution of NC2 and Taf1.

Our results support the generally positive regulatory function of the TBP: a clear reduction in gene expression of many cohorts was observed [**Figure 2 B**]. Among the TBP -dependent TFs we focused on Hap1, Skn7 and Swi4, three TFs that illustrate different mechanisms for their TBP regulation interaction and also the contribution of each of the cofactors NC2/Taf1 [**Table 3**].

KS score	**WT**	**TBPd**	**DeltaT**	**NC2**	**TBPd DeltaT**
Hap1	-1.18	**-6.83**	1.44	**8.21**	**-6.45**
Skn7	-2.67	**-12.36**	**5.81**	4.15	**-7.47**
Swi4	0.21	**-7.95**	0.73	3.5	**-6.55**

Table 3: K-S scores in experiments disrupting various TBP interactions. *The K-S scores for the cohorts of the TFs Skn7, Swi4 and Hap4 were computed based on expression profiles carried out for over-expressed TBP mutants [43]. WT: an empty vector. TBPd: disruption of TBP dimerization by the TBP mutation V161E. DeltaT: disruption of TBP-Taf1 interaction using a strain with a TAND I deleted form of TAF1. NC2: disruption of TBP-NC2 interaction using the TBP mutation F182V. While none of the cohorts exhibit a significant activity in the WT profile experiments, in the TBP dimerization disruption, a significant repression is notable. The significant K-S scores are highlighted (Bonferroni corrected p-value < 0.05).*

Although TATA box-containing genes comprise only ~20% of the yeast genome [47], an analysis of the distribution of TATA-box occupancy (**Supplementary Table D**) shows that, as expected, the TBP-dependent cohorts are highly enriched (~40%) for TATA box-containing genes (Hap1, Skn7 and Swi4 cohorts with hyper-geometric $p < 10^{-14}$, $p < 10^{-20}$ and $p < 10^{-9}$, respectively) which illustrates the essential contribution of the TBP in the transcription regulation of these cohort. As noted before, for each of these TFs, expression level by itself was not sufficient to explain the proposed trend of its cohort (TF expression levels are available in **Supplementary Table B**).

Hap1 is a TF with roles in the cellular response to heme and oxygen [44]. Its cohort, consisting of 141 genes, is significantly repressed in a strain carrying the TBPd mutations (K-S score = -6.45). These results are an indication that Hap1 is dependent on TBP to induce its genes. A deletion of the TANDI region of *TAF1* (DeltaT) has no effect on Hap1 cohort which points to a TBP induction mechanism that is independent for Taf1. Interestingly, mutations that affect NC2 binding caused a strong increase in the expression of the Hap1 cohort (K-S score = 8.21). As stated above, NC2 is a cofactor of the TBP that acts as an inhibitor of the TBP regulation. The strong de-repression of the Hap1 cohort gives NC2 a strong contribution to the repression mechanism through Hap1. We can conclude that Hap1 is a good example of a transcription factor that promotes the transcription of its target genes by TBP recruitment but uses the NC2 complex to regulate these genes in the opposite manner.

Another distinctive example of TBP dependent regulation is that of Skn7. Skn7 is a TF associated with various stress responses, in particular with the oxidative stress response [45]. Its cohort consists of 187 genes, and like Hap1 cohort, exhibits a strong de-activation in strains carrying the TBPd mutations (K-S score = -12.36). However, unlike Hap1, the Skn7 cohort also exhibits a significant induction in the DeltaT strain (K-S score = 5.81) and to a lesser extent also in the strain defective in NC2 interaction (K-S score = 4.16). The activation of Skn7 cohort in the two deletions shows that both cofactors, Taf1 and NC2, participate in the regulation of Skn7. Unlike NC2 contribution to Hap1 regulation, the de-repression of Skn7 cohort is less significant in both deletions. The less significant effect of Taf1 and NC2 on the Skn7 cohort could arise from a third cofactor that has higher contribution to the repression mechanism of Skn7 cohort, or alternatively can be due to a complementary repression by the two mechanisms, each with its own repression targets. Unfortunately it is hard to test the proposed hypotheses as no expression profiling is

available for either a deletion of other TBP cofactors or the double mutated strain, DeltaT and NC2. In spite the lack of additional information, very little overlap is observed among genes affected in each of the DeltaT and NC2 mutations [**Supplementary Figure 2**], which gives support to the complementary repression mechanism. We can conclude that Skn7 is an example of a transcription factor that promotes the transcription of its target genes by TBP recruitment and uses both cofactors, NC2 and Taf1 to regulate its genes in the opposite manner.

The last example in this thesis that illustrates a regulation mechanism that is dependent on TBP is that of Swi4. Swi4 is a central cell cycle TF that together with Swi6 promotes transcription of late G1 genes [46]. The Swi4 cohort, consisting of 156 genes, is also significantly repressed upon mutation in the TBP dimerization domain (K-S score = -7.95), but unlike Hap1 and Skn7 its cohort depends neither on Taf1 nor on the NC2 repressor (K-S scores 0.73 and 3.5 respectively). Thus, in the case of Swi4, if there is a repression mechanism that works through the TBP, it is conferred by factors other than the ones tested here (Taf1, NC2).

From the examples above another principle can be learned about the regulation mechanism of the TBP. In all the experiments carried out in strains lacking both TBP dimerization and the TANDI region (TBPd-DeltaT), a strong reduction of expression is observed, similar to the one seen in strains affected for dimerization only [**Table 3**]. This epistatic effect of the TBP destabilizing mutation points to a need for a functional TBP in the Taf1-mediated regulation.

Thus, our analysis shows that the TBP plays a central role in the regulation carried out by several TFs. Furthermore, by analyzing a data set [43] created originally to explore the participation of Taf1 and NC2 in the TBP regulation, we were able to analyze not only TBP dependency, but also the contribution level of each of the TBP co-factors.

3 Discussion

Chromatin organization plays a central role in many biological mechanisms, and particularly in transcription. Although many factors were found to participate in the regulation of the chromatin structure, to date there has been no systematic study of their global contribution to transcription. In this work, using a compendium of genome-wide profiles of strains defective in CM activity, we lay the infrastructure to the study of the contribution of the CMs to transcription and transcription regulation through their interactions with TFs. We show that this approach is able to detect cooperation between a TF and CMs even when complex combinatorial regulation is involved. Our systematic analysis of all available TF cohorts against the large gene expression compendium provides the first comprehensive picture in a eukaryote of the complex regulation by TFs in the context of chromatin organization. We have shown that our method is robust enough to detect novel regulation mechanisms of well-characterized TFs (e.g., Ume6, Gcn4), as well as to characterize regulation features of uncharacterized TFs, such as Yap6. Furthermore our method is applicable even to general factors, such as Tup1 and TBP. Note that our test cannot distinguish between direct and indirect CM-TF interaction. The difficulty in separating direct effects from indirect ones is prevalent in many studies on gene regulation networks [48-51].

In the sequel we refer to some limitations of our approach and suggest directions for future work.

3.1 Expanding the CM compendium

The gene expression profiles collected in this work cover a comprehensive compendium of CM complexes in the yeast *S. cerevisiae*, by containing at least one member of each of the known yeast CM complexes. In the example of Gcn4 and RPD3 complex illustrated above, different functional attributes could be assigned to the RPD3 complex only due to the available extensive profiling of each of the RPD3 complex members. As additional profiles are accumulating in the public repositories, the current compendium could be expanded to include comprehensive characterization of other CM complexes. The addition of other CM related profiling will allow us to understand at a finer resolution the contribution of each CM complex in the regulation of transcription.

Another aspect that is expected to be better understood with the addition of other CM related profiles is the mechanism by which environmental conditions lead to differential gene expression. The majority of the profiles in the current compendium were measured under standard growth conditions (e.g., rich medium). By using our method on transcriptional profiles obtained in other environments we can start to investigate this question. As the number of CMs analyzed under many environmental conditions grows, we expect to obtain insights into the complex mechanisms that control environmental responses.

3.2 Improving the statistical model

Our analysis used Kolmogorov-Smirnov (K-S) analysis to test whether a set of genes is over- or under-expressed in a given gene expression profile. This K-S test was found to be robust for our data sets, and helped to reveal known CM-TF interaction along with novel CM-TF interactions. One shortcoming of the K-S analysis is that it is more sensitive for deviation of the target set in the middle of the distribution. Enrichment analysis is very popular in the field of functional genomics and some groups have used a variant of the K-S to compensate for the K-S limitations (e.g. GSEA [66]). Recent work even compared various enrichment analysis tools, and although it was shown that K-S analysis is the most sensitive tool, the authors emphasize the benefit of combining results from more than one statistical test in the analysis [67]. It would be interesting to improve our statistical prediction by combining the results from other tools (e.g. the Wilcoxon rank sum test).

Another aspect in our analysis that will probably improve our predictions is a better definition of the TF cohort. Various studies have shown that when binding is binarized by taking a simple binding p-value cutoff in ChIP-on-chip data (e.g., $p<0.001$, as was used by our methodology and by many others), a lot of valuable information is overlooked [68,69]. A more flexible cutoff on the data set or even a regression approach might help in exploring TFs that have weak binding specificities. Also, as precipitation methodologies improve and better characterization of TF cohorts is being generated, our methodology could be put to a better use.

3.3 Working with different organisms

In recent years it has become evident that chromatin modifications are involved in many important biological processes in higher eukaryotes. Since modifications on histones are correlated

with transcription, these chromatin modifications are commonly studied in the context of transcriptional regulation. However, it is clear that additional processes, related to DNA, , such as repair, replication and recombination, are affected and regulated by histone modifications [70]. Recent studies even link histone modifications to other central mechanisms such as RNA interference and DNA methylation [71]. It is not surprising, then, to find CMs over-expressed and mutated in many cancer cells [72]. In fact, inhibitors of deacetylases are now in phase I and II clinical trials [73].

As many CMs are evolutionarily conserved [52], similar mechanisms of regulation are expected to be observed in higher eukaryotes. The accumulation of data sets similar to those used in our analysis in higher eukaryotes [53,54] will allow the application of our methodology on those organisms. We believe that this kind of analysis will help in understanding whether the regulatory functions unveiled in yeast are also conserved in higher eukaryotes, and will also provide insights into the overall global regulatory mechanisms that underlie many central processes.

Our understanding of transcription regulation has undergone several transformations over the last decade, and the emerging picture is very complex. Alternative splicing, RNA-based regulation and chromatin organization are today recognized as central regulatory mechanisms of gene expression. Still, our understanding of each of these processes is incomplete. We hope that the proposed methodology will be able to shed light on the effects of chromatin modifications on transcription factors and on transcription regulation in general.

4 Methods

4.1 K-S analysis

Given two samples of values, the Kolmogorov-Smirnov (K-S) test [10] is designed to examine whether they have the same value distribution. The main advantage of this test is that it makes no assumption on the distributions from which the samples originated. This is important when dealing with expression profiles from different sources.

For each value v the K-S test measures the difference in the fraction of genes that have an expression value lower than v between the control and the cohort samples. The K-S statistic is defined to be the maximum absolute value of that difference.

In the case of the null hypothesis (the two samples originate from the same distribution) the distribution of the statistic can be calculated and a p-value $\text{K-S}_{\text{p-value}}$ can be assigned to the disparity between the two samples [10].

The K-S score is defined as: $\text{K-S}_{\text{score}} = -\log_{10}(\text{K-S}_{\text{p-value}})$ if the statistic is positive and $\log_{10}(\text{K-S}_{\text{p-value}})$ otherwise. Hence, the absolute value of the $\text{K-S}_{\text{score}}$ indicates significance of the disparity, and its sign indicates the direction of the disparity: a positive sign shows that the cohort genes tend to have higher values than the rest of the genes.

4.2 Yeast genome

6646 Yeast ORFs were retrieved from Saccharomyces Genome Database (www.yeastgenome.org) (version July 2005). To avoid cross-hybridization biases in the gene expression and location data set, 103 ORFs, containing mitochondrial genes and short dubious ORFs were ignored in the analysis [**Supplementary Table C**].

4.3 Data preparation

170 gene expression profiles obtained with strains mutated for various CMs were collected from 26 publications. The complete list of publications and experiments is available in **Supplementary Table A**. Data were downloaded from papers' web supplements. Normalization was done as in [55].

TF-DNA binding profiles were obtained from Harbison et al. [1]. A p-value cutoff of 0.001 was used to define the set of genes bound by a particular TF (the TF cohort).

To account for the strong correlated response of the ribosomal genes [55] in most experiments, all TFs that were found to be significantly enriched ($p<0.001$) in ribosome related GO terms were excluded from the analysis (**Supplementary Table E**). Our analysis used the remaining 75 cohorts, containing at least 50 genes, that were originated from 49 TF tested in different environments.

4.4 Altered gene groups and their overlap test

A gene is considered altered in a gene expression experiment if its Z-score is greater than 1. Given a gene expression experiment *E* with average μ and SD σ, and a TF cohort S (the TF target gene group), the elevated cohort genes are defined as: $TF_E = \{g \text{ in } S | E(g) > \mu + \sigma\}$ while the set of declining genes is defined as: $\{g \text{ in } S | E(g) < \mu - \sigma\}$.

Given two altered (elevated or reduced) sub-groups S_1 and S_2 from S, the significance of their overlap is calculated using the hyper-geometric distribution, where S is considered as the samples pool.

4.5 Annotation enrichment

All GO annotations were taken from the Gene Ontology database [56] (version July 2005). Annotation enrichments were obtained using the TANGO program [57]. TANGO finds GO terms that are enriched with the target set in study. The strength of TANGO is to provide a p-value for that enrichment using simulation of random samplings.

4.6 Hierarchical clustering

Hierarchical clustering of the cohorts and the experimental conditions based on the significant K-S scores matrix (all $|\text{K-S scores}| > 1.3$) was carried out using the EXPANDER analysis and visualization tool (Version 3.0) [57].

References

1. Harbison, C.T. et al. Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99-104 (2004).

2. Kurdistani, S.K. & Grunstein, M. Histone acetylation and deacetylation in yeast. Nat Rev Mol Cell Biol 4, 276-84 (2003).

3. Tsukiyama, T. The in vivo functions of ATP-dependent chromatin-remodelling factors. Nat Rev Mol Cell Biol 3, 422-9 (2002).

4. Eberharter, A. & Becker, P.B. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. EMBO Rep 3, 224-9 (2002).

5. Jenuwein, T. & Allis, C.D. Translating the histone code. Science 293, 1074-80 (2001).

6. Kouzarides, T. Histone methylation in transcriptional control. Curr Opin Genet Dev 12, 198-209 (2002).

7. Kadosh, D. & Struhl, K. Repression by Ume6 involves recruitment of a complex containing Sin3 corepressor and Rpd3 histone deacetylase to target promoters. Cell 89, 365-71 (1997).

8. Watson, A.D. et al. Ssn6-Tup1 interacts with class I histone deacetylases required for repression. Genes Dev 14, 2737-44 (2000).

9. Swanson, M.J. et al. A multiplicity of coactivators is required by Gcn4p at individual promoters in vivo. Mol Cell Biol 23, 2800-20 (2003).

10. Stephens. Use of the kolmogorov-smirnov, cram-vvon mises and related statistics without extensive tables. . J. R. Statist. Soc. 32, 115–122 (1970).

11. Fazzio, T.G. et al. Widespread collaboration of Isw2 and Sin3-Rpd3 chromatin remodeling complexes in transcriptional repression. Mol Cell Biol 21, 6450-60 (2001).

12. Goldmark, J.P., Fazzio, T.G., Estep, P.W., Church, G.M. & Tsukiyama, T. The Isw2 chromatin remodeling complex represses early meiotic genes upon recruitment by Ume6p. Cell 103, 423-33 (2000).

13. Robyr, D. et al. Microarray deacetylation maps determine genome-wide functions for yeast histone deacetylases. Cell 109, 437-46 (2002).

14. Rubin-Bejerano, I., Mandel, S., Robzyk, K. & Kassir, Y. Induction of meiosis in Saccharomyces cerevisiae depends on conversion of the transcriptional represssor Ume6 to a positive regulator by its regulated association with the transcriptional activator Ime1. Mol Cell Biol 16, 2518-26 (1996).

15. Sabet, N., Volo, S., Yu, C., Madigan, J.P. & Morse, R.H. Genome-wide analysis of the relationship between transcriptional regulation by Rpd3p and the histone H3 and H4 amino termini in budding yeast. Mol Cell Biol 24, 8823-33 (2004).

16. Santos-Rosa, H. et al. Methylation of histone H3 K4 mediates association of the Isw1p ATPase with chromatin. Mol Cell 12, 1325-32 (2003).

17. Kagalwala, M.N., Glaus, B.J., Dang, W., Zofall, M. & Bartholomew, B. Topography of the ISW2-nucleosome complex: insights into nucleosome spacing and chromatin remodeling. Embo J 23, 2092-104 (2004).

18. Clapier, C.R., Langst, G., Corona, D.F., Becker, P.B. & Nightingale, K.P. Critical role for the histone H4 N terminus in nucleosome remodeling by ISWI. Mol Cell Biol 21, 875-83 (2001).

19. Fazzio, T.G., Gelbart, M.E. & Tsukiyama, T. Two distinct mechanisms of chromatin interaction by the Isw2 chromatin remodeling complex in vivo. Mol Cell Biol 25, 9165-74 (2005).

20. Hughes, T.R. et al. Functional discovery via a compendium of expression profiles. Cell 102, 109-26 (2000).

21. Krogan, N.J. et al. Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4. Proc Natl Acad Sci U S A 101, 13513-8 (2004).

22. Gavin, A.C. et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631-6 (2006).

23. Krogan, N.J. et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440, 637-43 (2006).

24. Mnaimneh, S. et al. Exploration of essential gene functions via titratable promoter alleles. Cell 118, 31-44 (2004).

25. Keogh, M.C. et al. Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex. Cell 123, 593-605 (2005).

26. Hinnebusch, A.G. Translational regulation of GCN4 and the general amino acid control of yeast. Annu Rev Microbiol 59, 407-50 (2005).

27. Kuo, M.H., vom Baur, E., Struhl, K. & Allis, C.D. Gcn4 activator targets Gcn5 histone acetyltransferase to specific promoters independently of transcription. Mol Cell 6, 1309-20 (2000).

28. Natarajan, K., Jackson, B.M., Zhou, H., Winston, F. & Hinnebusch, A.G. Transcriptional activation by Gcn4p involves independent interactions with the SWI/SNF complex and the SRB/mediator. Mol Cell 4, 657-64 (1999).

29. Natarajan, K. et al. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast. Mol Cell Biol 21, 4347-68 (2001).

30. Fernandes, L., Rodrigues-Pousada, C. & Struhl, K. Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. Mol Cell Biol 17, 6982-93 (1997).

31. Mendizabal, I., Rios, G., Mulet, J.M., Serrano, R. & de Larrinoa, I.F. Yeast putative transcription factors involved in salt tolerance. FEBS Lett 425, 323-8 (1998).

32. Rundlett, S.E. et al. HDA1 and RPD3 are members of distinct yeast histone deacetylase complexes that regulate silencing and transcription. Proc Natl Acad Sci U S A 93, 14503-8 (1996).

33. Wu, J., Suka, N., Carlson, M. & Grunstein, M. TUP1 utilizes histone H3/H2B-specific HDA1 deacetylase to repress gene activity in yeast. Mol Cell 7, 117-26 (2001).

34. Smith, R.L. & Johnson, A.D. Turning genes off by Ssn6-Tup1: a conserved system of transcriptional repression in eukaryotes. Trends Biochem Sci 25, 325-30 (2000).

35. Santos-Rosa, H., Bannister, A.J., Dehe, P.M., Geli, V. & Kouzarides, T. Methylation of H3 lysine 4 at euchromatin promotes Sir3p association with heterochromatin. J Biol Chem 279, 47506-12 (2004).

36. Grant, P.A. et al. Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex. Genes Dev 11, 1640-50 (1997).

37. Huisinga, K.L. & Pugh, B.F. A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae. Mol Cell 13, 573-85 (2004).

38. Dudley, A.M., Rougeulle, C. & Winston, F. The Spt components of SAGA facilitate TBP binding to a promoter at a post-activator-binding step in vivo. Genes Dev 13, 2940-5 (1999).

39. Bhaumik, S.R. & Green, M.R. Differential requirement of SAGA components for recruitment of TATA-box-binding protein to promoters in vivo. Mol Cell Biol 22, 7365-71 (2002).

40. Sterner, D.E., Belotserkovskaya, R. & Berger, S.L. SALSA, a variant of yeast SAGA, contains truncated Spt7, which correlates with activated transcription. Proc Natl Acad Sci U S A 99, 11622-7 (2002).

41. Belotserkovskaya, R. et al. Inhibition of TATA-binding protein function by SAGA subunits Spt3 and Spt8 at Gcn4-activated promoters. Mol Cell Biol 20, 634-47 (2000).

42. Pugh, B.F. Control of gene expression through regulation of the TATA-binding protein. Gene 255, 1-14 (2000).

43. Chitikila, C., Huisinga, K.L., Irvin, J.D., Basehoar, A.D. & Pugh, B.F. Interplay of TBP inhibitors in global transcriptional control. Mol Cell 10, 871-82 (2002).

44. Zhang, L. & Hach, A. Molecular mechanism of heme signaling in yeast: the transcriptional activator Hap1 serves as the key mediator. Cell Mol Life Sci 56, 415-26 (1999).

45. Raitt, D.C. et al. The Skn7 response regulator of Saccharomyces cerevisiae interacts with Hsf1 in vivo and is required for the induction of heat shock genes by oxidative stress. Mol Biol Cell 11, 2335-47 (2000).

46. Bahler, J. Cell-cycle control of gene expression in budding and fission yeast. Annu Rev Genet 39, 69-94 (2005).

47. Basehoar, A.D., Zanton, S.J. & Pugh, B.F. Identification and distinct regulation of yeast TATA box-containing genes. Cell 116, 699-709 (2004).

48. Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. & Young, R.A. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput, 437-49 (2002).

49. Gardner, T.S., di Bernardo, D., Lorenz, D. & Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102-5 (2003).

50. Segal, E. et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet 34, 166-76 (2003).

51. Workman, C.T. et al. A systems approach to mapping DNA damage response pathways. Science 312, 1054-9 (2006).

52. Khorasanizadeh, S. The nucleosome: from genomic organization to genomic regulation. Cell 116, 259-72 (2004).

53. Boyer, L.A. et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell 122, 947-56 (2005).

54. Le Brigand, K. et al. An open-access long oligonucleotide microarray resource for analysis of the human and mouse transcriptomes. Nucleic Acids Res 34, e87 (2006).

55. Tanay, A., Steinfeld, I., Kupiec, M. & Shamir, R. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. Mol Syst Biol 1, 2005 0002 (2005).

56. Harris, M.A. et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32, D258-61 (2004).

57. Shamir, R. et al. EXPANDER--an integrative program suite for microarray data analysis. BMC Bioinformatics 6, 232 (2005).

58. Bernstein, B.E. et al. Methylation of histone H3 Lys 4 in coding regions of active genes. Proc Natl Acad Sci U S A 99, 8695-700 (2002).

59. Bernstein, B.E., Tong, J.K. & Schreiber, S.L. Genomewide studies of histone deacetylase function in yeast. Proc Natl Acad Sci U S A 97, 13708-13 (2000).

60. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science ;270(5235):467-70 (1995).

61. Nonaka G, Blankschien M, Herman C, Gross CA, Rhodius VA. Regulon and promoter analysis of the E. coli heat-shock factor, sigma32, reveals a multifaceted cellular response to heat stress. Genes Dev. 20(13):1776-89 (2006)

62. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 9(12):3273-97 (1998)

63. Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, Pienta KJ, Rubin MA, Chinnaiyan AM. Delineation of prognostic biomarkers in prostate cancer. Nature. 412(6849):822-6 (2001)

64. Bakkenist CJ, Kastan MB. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. Nature. 421(6922):499-506 (2003)

65. Lia G, Praly E, Ferreira H, Stockdale C, Tse-Dinh YC, Dunlap D, Croquette V, Bensimon D, Owen-Hughes T. Direct Observation of DNA Distortion by the RSC Complex, Molecular Cell 21 (3): 417-425 (2006)

66. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 102(43):15545-50 (2005)

67. Levine DM, Haynor DR, Castle JC, Stepaniants SB, Pellegrini M, Mao M, Johnson JM. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome Biol. 7(10):R93 (2006)

68. Eden E, Lipson D, Yogev S, Yakhini Z. Discovering motifs in ranked lists of DNA sequences. PLoS Comput Biol. 3(3):e39 (2007)

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

70. Muegge K. Modifications of histone cores and tails in V(D)J recombination Genome Biol. 4(4):211 (2003)

71. Bachman KE, Park BH, Rhee I, Rajagopalan H, Herman JG, Baylin SB, Kinzler KW, Vogelstein B. Histone modifications and silencing prior to DNA methylation of a tumor suppressor gene. Cancer Cell. 3(1):89-95 (2003)

72. Wolffe AP. Chromatin remodeling: why it is important in cancer. Oncogene. 20(24):2988-90 (2001)

73. Reid T, Valone F, Lipera W, Irwin D, Paroly W, Natale R, Sreedharan S, Keer H, Lum B, Scappaticci F, Bhatnagar A. Phase II trial of the histone deacetylase inhibitor pivaloyloxymethyl butyrate (Pivanex, AN-9) in advanced non-small cell lung cancer. Lung Cancer. 45(3):381-6 (2004)

74. Israel Steinfeld, Ron Shamir & Martin Kupiec. A genome-wide analysis in Saccharomyces cerevisiae demonstrates the influence of chromatin modifiers on transcription. *Nature Genetics* **39**, 303 - 309 (2007)

Appendix: Supplements

Supplementary Table A: The publications from which the gene expression data were obtained

The CM gene expression compendium is composed from 26 assays taken from different publication. This table refers to the relevant publications.

Name	Pubmed ID	Full Name
Sudarsanam et al.	10725359	Whole genome expression analysis of snf/swi mutants of S.cerevisiae
Geisberg et al.	11283253	yeast NC2 associates with the RNA polymerase II pre-initiation complex and selectively affects transcription in vivo
Angus-Hill et al.	11336698	rsc3/rsc30 zinc cluster dimmer
Krogan et al.	15353583	Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4
Fazzio et al.	11533234	Widespread collaboration of Isw2 and Sin3-Rpd3 Chromatin Remodeling Complexes in Transcriptional Repression
Meneghini et al.	12628191	Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin
Mizuguch et al.	14645854	ATP-driven exchange of histone H2A.Z variant catalyzed by SWR1 chromatin remodeling complex
Bernstein et al.	11095743	Genome-wide studies of histone deacetylase function in yeast
Bernstein et al.	12060701	Methylation of histone H3 Lys 4 in coding regions of active genes.
Chitikila et al.	12419230	Interplay of TBP inhibitors in global transcriptional control
Martin et al.	15280228	Redundant Roles for histone H3 N-terminal lysine residues in subtelomeric gene repression in Saccharomyces cerevisiae.
Sabet et al.	15456858	Genome-wide analysis of the relationship between transcriptional regulation by Rpd3p and the histone H3 and H4 amino termini in budding yeast.
Keogh et al.	16286008	Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex.
Huisinga et al.	14992726	A genome-wide housekeeping role for TFIID and a highly regulated stress-related role for SAGA in Saccharomyces cerevisiae.
Ingvarsdottir et al.	15657441	H2B ubiquitin protease Ubp8 and Sgf11 constitute a discrete functional module within the Saccharomyces cerevisiae SAGA complex.
Orlandi et al.	14623890	Transcriptional profiling of ubp10 null mutant reveals altered subtelomeric gene expression and insurgence of oxidative stress response
Rosaleny et al.	16023114	Yeast HAT1 and HAT2 deletions have different life-span and transcriptome phenotypes
Eriksson et al.	16199888	Global regulation by the yeast Spt10 protein is mediated through chromatin structure and the histone upstream activating sequence elements.
Xu et al.	15882620	Acetylation in histone h3 globular domain regulates gene expression in yeast.
Attikum et al.	15607975	Recruitment of the INO80 Complex by H2A Phosphorylation Links ATP-Dependent Chromatin Remodeling with DNA Double-Strand Break Repair.
Kaeberlein et al.	15126388	Saccharomyces cerevisiae SSD1-V confers longevity by Sir2p-independent mechanism.
Boa et al.	12845608	Saccharomyces cerevisiae Set1p is a methyltransferase specific for lysine 4 of histone H3 and is required for efficient gene expression.
Hughes et al.	10929718	Functional discovery via a compendium of expression profiles.
Mnaimneh et al.	15242642	Exploration of Essential Gene functions via Titratable Promoter Alleles.
Dasgupta et al.	11880621	Mot1 activates and represses transcription by direct, ATPase-dependent mechanisms.

Durant et al.	16537921	Genome-wide relationships between TAF1 and histone acetyltransferases in Saccharomyces cerevisiae.

[1]***Supplementary Table B: All KS scores.***

Supplementary Table C: All the ORFs excluded from our analysis. *To avoid cross-hybridization biases in the gene expression and location data set, 103 ORFs, containing mitochondrial genes and short dubious ORFs were ignored in the analysis.*

COS2	YHL050C	YMR325W
COS7	YBL111C	YLL025W
COS4	YHR218W	YIR041W
COS12	YFL066C	YBL108C-A
COS6	YLR464W	Q0010
COS8	YEL076C	Q0017
COS5	YFL068W	Q0032
COS9	YEL076W-C	COX1
COS3	YLR463C	AI1
COS1	YHL049C	AI2
COS10	YBL112C	AI3
PAU7	YFL067W	AI4
PAU3	YPR203W	AI5_ALPHA
PAU2	YFL065C	AI5_BETA
PAU5	YLR465C	AAP1
PAU1	YPR202W	ATP6
PAU4	YFL064C	Q0092
PAU6	YEL075C	COB
YRF1-1	YER189W	BI2
YRF1-2	YLR462W	BI3
YRF1-3	DAN1	BI4
YRF1-4	DAN2	OLI1
YRF1-5	DAN3	VAR1
YRF1-6	DAN4	Q0142
YRF1-7	YGR294W	Q0143
YOR396W	YHL046C	Q0144
YML133C	YIL176C	SCEI
YLL066C	YDR542W	Q0182
YIL177C	YLL064C	COX2
YJL225C	YAL068C	Q0255
YEL077C	YGL261C	COX3
YLL067C	YOL161C	Q0297
YHR219W	YKL224C	HXT12
YPR204W	YOR394W	SDC25
YBL113C	YPL282C	

[1] Long supplements are not included in the thesis and are available via the *Nature Genetics* supplemental information page http://www.nature.com/ng/journal/v39/n3/suppinfo/ng1965_S1.html

Supplementary Table D: The full cohort-TATA box occupancy level.

The total number of genes that are targeted by a TF is 3698 (see the paper methods section). According to Basehoar et al. 764 of the above mentioned, TF targeted, genes contain a TATA box (~20%). This Table presents the TATA box frequency for each of the TF cohorts analyzed in our work.

Cohort Name	Cohort Size	Number of Genes with TATA box in Cohort	% Genes with TATA box in Cohort	Log (hyper-geometric p)
SKN7_H2O2Lo	187	73	0.39	-20.22
NRG1_H2O2Hi	126	51	0.4	-15.88
CIN5_YPD	150	57	0.38	-15.05
HAP1_YPD	141	54	0.38	-14.63
SUT1_YPD	69	30	0.43	-11.62
SOK2_BUT14	73	31	0.42	-11.39
HSF1_H2O2Lo	102	39	0.38	-10.99
DAL81_RAPA	95	37	0.38	-10.97
HSF1_H2O2Hi	125	45	0.36	-10.69
MSN2_H2O2Hi	79	32	0.4	-10.59
CIN5_H2O2Lo	127	45	0.35	-10.26
SWI4_YPD	156	52	0.33	-9.77
YAP6_YPD	91	33	0.36	-8.46
PHD1_BUT90	106	37	0.34	-8.43
GLN3_RAPA	68	26	0.38	-7.9
GCN4_SM	189	57	0.3	-7.73
RLM1_YPD	55	22	0.4	-7.61
SKN7_H2O2Hi	99	34	0.34	-7.58
NDD1_YPD	85	30	0.35	-7.37
SWI6_YPD	153	47	0.3	-7.11
SWI5_YPD	102	34	0.33	-7.03
ASH1_BUT14	51	20	0.39	-6.79
MSN4_H2O2Hi	70	25	0.35	-6.6
XBP1_H2O2Lo	68	24	0.35	-6.24
NRG1_YPD	72	25	0.34	-6.2
YAP7_H2O2Lo	152	45	0.29	-6.16
GCR2_SM	54	20	0.37	-6.05
PHD1_YPD	67	23	0.34	-5.69
STE12_BUT14	128	38	0.29	-5.55
YAP7_H2O2Hi	141	41	0.29	-5.49
GCN4_RAPA	160	45	0.28	-5.27
STE12_YPD	54	19	0.35	-5.25
RIM101_H2O2Lo	54	19	0.35	-5.25
FKH2_H2O2Hi	106	32	0.3	-5.21
STE12_BUT90	78	25	0.32	-5.15
YAP6_H2O2Lo	59	20	0.33	-5.03
GCN4_YPD	75	24	0.32	-5
MBP1_YPD	130	37	0.28	-4.85

SKN7_YPD	67	21	0.31	-4.38
CIN5_H2O2Hi	80	24	0.3	-4.3
YAP1_YPD	72	22	0.3	-4.26
UME6_YPD	131	35	0.26	-3.95
CBF1_SM	279	67	0.24	-3.84
HAP4_YPD	67	20	0.29	-3.83
MBP1_H2O2Hi	133	35	0.26	-3.79
UME6_H2O2Hi	102	28	0.27	-3.78
YAP6_H2O2Hi	79	22	0.27	-3.47
PUT3_H2O2Lo	88	22	0.25	-2.8
STE12_Alpha	115	26	0.22	-2.54
FKH2_YPD	121	27	0.22	-2.52
OAF1_YPD	59	15	0.25	-2.51
AFT2_H2O2Lo	98	22	0.22	-2.42
MCM1_Alpha	106	23	0.21	-2.38
RCS1_H2O2Hi	52	13	0.25	-2.34
DIG1_BUT14	63	15	0.23	-2.32
DAL82_SM	55	13	0.23	-2.21
ROX1_YPD	67	15	0.22	-2.2
SMP1_YPD	77	16	0.2	-2.18
TYE7_YPD	56	13	0.23	-2.17
RTG3_RAPA	52	12	0.23	-2.12
AFT2_H2O2Hi	61	13	0.21	-2.08
DAL82_RAPA	56	12	0.21	-2.05
FHL1_YPD	188	13	0.06	0
FKH1_YPD	142	25	0.17	0
MCM1_YPD	77	14	0.18	0
MGA1_YPD	63	5	0.07	0
PDR1_YPD	68	11	0.16	0
PHO4_YPD	72	10	0.13	0
RAP1_YPD	162	23	0.14	0
REB1_YPD	146	15	0.1	0
SUM1_YPD	59	9	0.15	0
YAP5_YPD	56	7	0.12	0
FHL1_SM	204	20	0.09	0
STP1_SM	68	10	0.14	0
FHL1_RAPA	191	14	0.07	0
MSN4_RAPA	54	7	0.12	0
REB1_H2O2Hi	59	7	0.11	0
MBP1_H2O2Lo	58	8	0.13	0
RCS1_H2O2Lo	260	29	0.11	0
RPN4_H2O2Lo	100	14	0.14	0
DIG1_Alpha	60	12	0.2	0

Supplementary Table E: All cohorts found to be enriched in ribosomal related GO terms using TANGO.

Ribosomal genes are known to be highly sensitive to any environmental change. To reduce artificial CM assignments, cohort rich in ribosomal associated genes were excluded from the analysis. This table presents only those TF cohorts that were found to be rich in ribosomal associated genes. GO annotations were taken from the Gene Ontology database (version July 2005). Annotation enrichments were obtained using the TANGO program.

TF name	GO ID	Uncorrected Hypergeometric p-value (log10)	Corrected Hypergeometric p-value (log10)	Fraction of Genes with Annotation in the Cohort	Number of Genes with Annotation in the Cohort
FHL1_YPD	GO:0005830	-156.236	-3	0.61	115
FHL1_YPD	GO:0003735	-133.941	-3	0.61	116
FHL1_YPD	GO:0005840	-124.924	-3	0.63	119
FHL1_YPD	GO:0005842	-84.1339	-3	0.33	63
FHL1_YPD	GO:0005843	-71.8443	-3	0.27	52
PDR1_YPD	GO:0005842	-9.1932	-3	0.16	11
RAP1_YPD	GO:0005830	-73.3795	-3	0.43	70
RAP1_YPD	GO:0003735	-63.4241	-3	0.43	71
RAP1_YPD	GO:0005842	-44.6747	-3	0.25	41
RAP1_YPD	GO:0042257	-9.6891	-3	0.08	13
YAP5_YPD	GO:0005830	-9.1447	-3	0.23	13
FHL1_SM	GO:0005830	-134.706	-3	0.52	108
FHL1_SM	GO:0003735	-114.827	-3	0.53	109
FHL1_SM	GO:0005843	-62.963	-3	0.24	49
SFP1_SM	GO:0005830	-39.3035	-3	0.67	31
SFP1_SM	GO:0003735	-36.6689	-3	0.69	32
FHL1_RAPA	GO:0005830	-143.568	-3	0.57	110
FHL1_RAPA	GO:0003735	-123.031	-3	0.58	111
FHL1_RAPA	GO:0005842	-81.3824	-3	0.32	62
FHL1_RAPA	GO:0005843	-62.3605	-3	0.25	48
FHL1_RAPA	GO:0042257	-14.7904	-3	0.09	18

[2]Supplementary Table F: All unique pairs of TFs and CMs taken from Supplementary Table B.

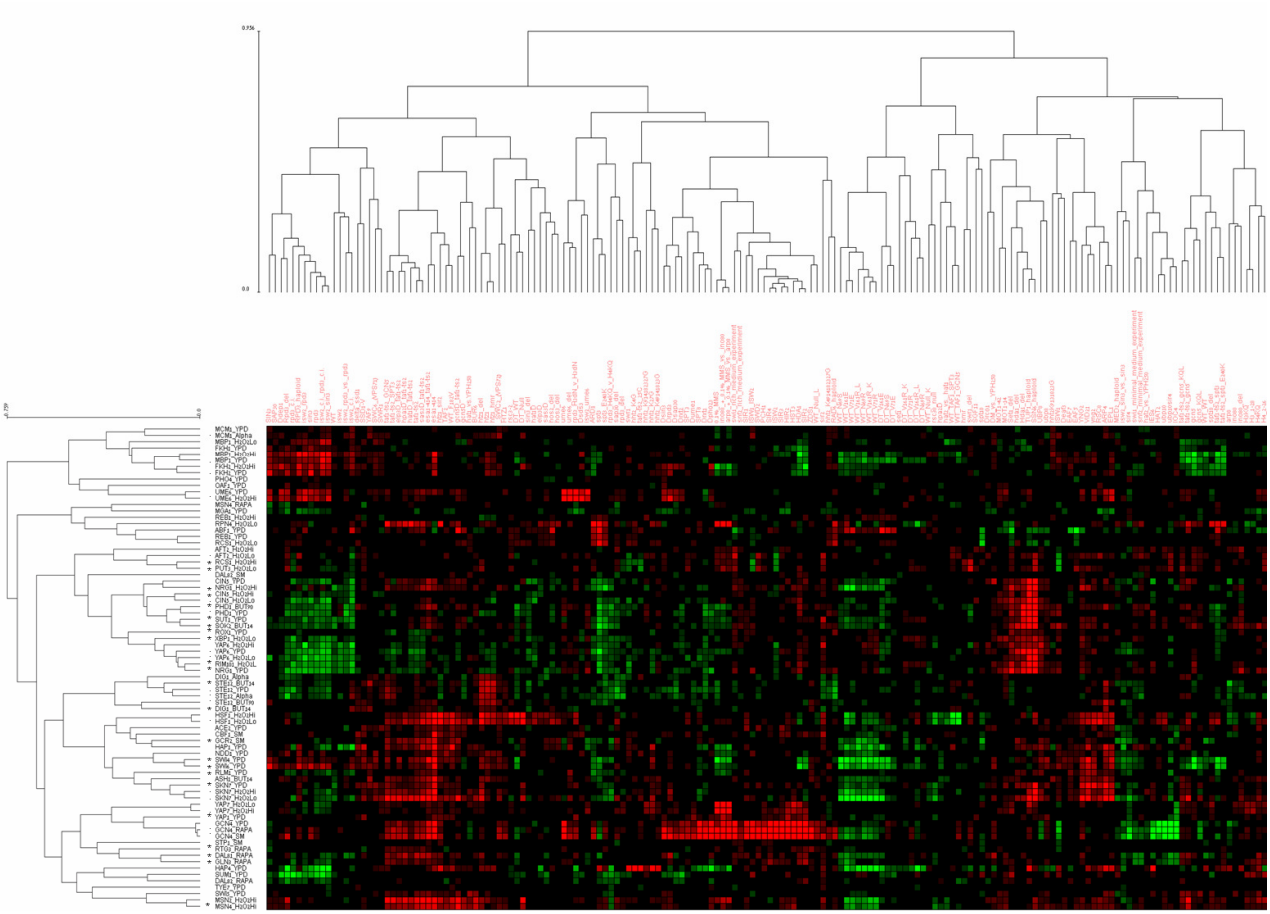
Each TF and each CM could be represented by various experiments. This table presents the unique pairing of TFs and CMs. For each pair the highest K-S score is given followed by the significant interaction numbers, taken from Supplementary Table B, that support this pairing.

[2]Supplementary Table G: All Pairs of TF cohorts that share a significant amount of genes.

For each pair of TF cohorts defined in the same condition, a hyper-geometric p was calculated on their intersection. Presented are all the intersections that were found to be significant, after bonferroni multiple correction ($p<0.01$).

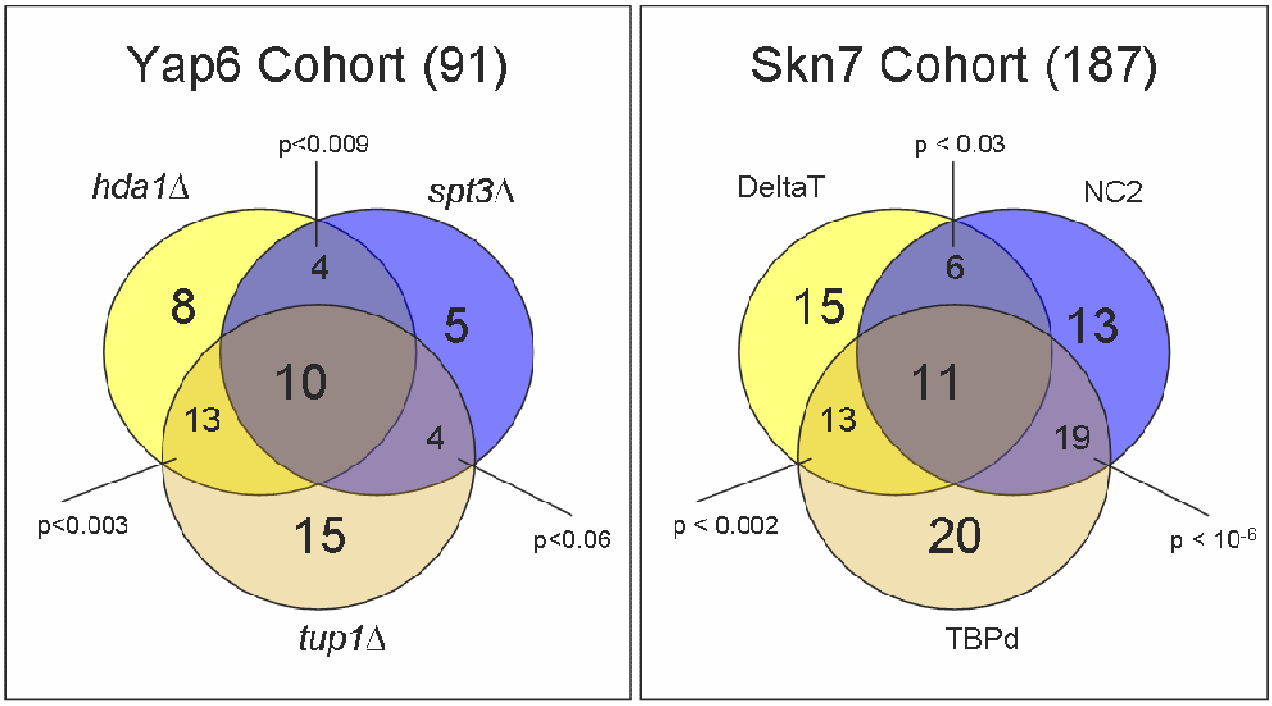
[2] Long supplements are not included in the thesis and are available via the *Nature Genetics* supplemental information page http://www.nature.com/ng/journal/v39/n3/suppinfo/ng1965_S1.html

Supplementary Figure 1: Detailed clustering of the compendium. *As in Figure 2B, rows represent TF cohorts and columns represent conditions. Colors indicate CM-cohort K-S scores. To obtain a global view of the TF-CM interaction landscape, we hierarchically clustered the cohorts and conditions according to their K-S scores (positive scores in red and negative in green). Groups of functionally related TFs (ordinate) and functionally related conditions (abscissa) are marked. TFs that share a significant number of genes with the TF immediately above them are marked with asterisks. Clustered cohorts of the same TF (under different conditions) are marked by dots.*



Supplementary Figure 2: Overlap of altered genes of the Yap6 and Skn7 cohorts in three experiments.

A) Out of the 91 genes in the Yap6 cohort, 35 showed a notable induction in a hda1D strain, 42 showed induction in a tup1D strain, and 23 a notable repression in a strain deleted for SPT3. The significance of the overlap is indicated (hyper-geometric p-value). B) Out of 187 genes in the Skn1 cohort, 45 showed a notable induction in the DeltaT strain, 49 a notable induction in the NC2 strain, and 63 a notable repression in the TBPd strain. The significance of the overlap is indicated (hyper-geometric p-value).



A B