Title
EVOLUTION OF FUNGAL TRANSCRIPTION CIRCUITS

Author
Tuch, Brian

Publication Date
2008-02-27

Peer reviewed|Thesis/dissertation

Evolution of Fungal Transcription Circuits

by

Brian B. Tuch

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOLOGICAL AND MEDICAL INFORMATICS

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

ACKNOWLEDGEMENTS

I dedicate this thesis to Geraldine Kim, the one and only Bonobo, and to my parents, whose support made getting to graduate school and getting through graduate school possible. Now that I am officially a "doctor", you are free to stop worrying about me.

I am also extremely grateful to Matt Jacobson and Patsy Babbitt for providing the strong guidance that pointed me in the right direction at the outset of graduate school; to Chris Voigt, for helping me to discover the field of research I so greatly enjoy; to Chris Baker, Richard Bennett, Lauren Booth, Christina Chaivorapol, Jeff Chuang, Brad Green, Aaron Hernday, Oliver Homann, Lisa Kohn, Quinn Mitrovich, Suzanne Noble, Rebecca Zordan and members of the Li and Johnson labs for providing invaluable discussions, inordinate patience, tutelage and companionship; to Annie Tsong for the same, in addition to an immensely enjoyable and rewarding collaboration; to my thesis committee, Hiten Madhani and Joe DeRisi, for offering a unique perspective, new ideas and opinions (lots of opinions); and, finally, to Hao and Sandy for nurturing my ideas and ambitions and providing me with exactly the environment I needed to successfully pursue these.

The text of Chapters 1 and 4 (Introduction and Conclusions) of this thesis is based on material found in a review written by Alexander Johnson, Hao Li and Brian Tuch:

Tuch BB, Li H & Johnson, AD. The evolution of eukaryotic transcription circuits. Science (In review, 2008).

The text of Chapter 2 of this thesis is a reprint, with minor modifications, of material that appears in:

Tsong AE*, Tuch BB*, Li H & Johnson, AD. Evolution of alternative transcriptional circuits with identical logic. Nature 443, 415-20 (2006).

* authors contributed equally

AET did the experimental and BBT the computational work. Both contributed to the development of ideas and to the writing of the manuscript.

The text of Chapter 3 of this thesis is a reprint of material that appears in:

Tuch BB*, Galgoczy DJ*, Hernday AD, Li H & Johnson AD. The evolution of combinatorial gene regulation in fungi. PLoS Biol 6(2): e38 (2008).

* authors contributed equally

DJG did the experimental work, with contributions from BBT. BBT performed all computational analyses and wrote the manuscript.

The text of Appendix 1 is a manuscript that was written by Hao Li and Brian Tuch in 2004. The manuscript was not published, but is included here in the hopes that it may be of some use to others. The remaining appendices were written expressly for this thesis and contain unpublished results that I am hopeful will, one day, contribute to more complete works.

EVOLUTION OF FUNGAL TRANSCRIPTION CIRCUITS

Brian Tuch

ABSTRACT

The gradual rewiring of transcriptional circuits over evolutionary timescales is a major source of the diversity of life on the planet. Studies in animals have shown how seemingly small changes in gene regulation can have large effects on morphology and physiology and how selective pressures can act on these changes. The underlying principle in these studies is that gene regulation is modular—changes can be made to the expression of a gene at one place and time, without affecting the expression of that gene at other places and times. Genome-wide studies in single cell yeasts, including those described here, have uncovered evidence of massive transcriptional rewiring, indicating that even closely related species regulate their genes using surprisingly different circuitries. The work described in this thesis begins to suggest some general principles guiding the evolution of transcription circuits. Mechanisms by which large sets of co-expressed genes can be rewired (without disrupting co-expression) are proposed and combinatorial regulation is implicated as a catalyst for change in transcriptional networks.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 2:

Chapter 3:

Appendix 1:

LIST OF FIGURES

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Appendix 1:

Appendix 2:

Appendix 3:

Chapter 1

Introduction

Over thirty years ago, an influential paper by King and Wilson was published[1]. Using tools that now seem remarkably unwieldy, the authors compared proteins and nucleic acids from human with those from chimpanzee and concluded, "their macromolecules are so alike that regulatory mutations may account for their biological differences." The completion and comparison of genome sequences from many organisms (including human and chimpanzee) has provided overwhelming support for the importance of regulatory changes in the evolution of organisms; it is clear, for example, that organism complexity does not scale in a simple way with gene number or content[2].

Much of the regulatory circuitry of an organism depends upon *cis*-regulatory sequences and the sequence-specific DNA binding proteins (also known as transcription factors) that recognize them and regulate transcription[3-5]. Transcription of each gene in a eukaryotic organism is controlled by a collection of *cis*-regulatory sequences that are typically positioned in proximity to the gene (sometimes very close, sometimes spread over hundreds of thousands of base pairs). The collection of *cis*-regulatory sequences associated with each gene specifies the time and place in the organism that the gene is to be transcribed. This code is read by the transcription factors, which recognize these sequences and which themselves are typically expressed or active only at particular times and places in the life of the organism. It is the combination of active transcription factors present at a particular location and time that selects, via interaction with *cis*-regulatory sequences, those genes to be transcribed. Of course, there are many additional steps in transcription and there is more to gene regulation than just transcription[6]; yet the *cis*-

regulatory sequences and the proteins that recognize them form an important tier of gene regulation, and many important evolutionary insights have converged upon this tier.

Over the past three decades, studies of eukaryotic gene transcription have revealed several general properties of *cis*-regulatory sequences and transcription factors that are especially important for understanding their roles in evolution. *cis*-regulatory sequences are typically short (generally 5 to 10 base pairs long) and degenerate (many similar sequences confer equivalent transcription factor binding). Additionally, their positions, relative to the gene whose transcription they control, can be variable. Different *cis*-regulatory sequences are often found in close proximity to each other, and transcription factors often bind cooperatively to these adjacent sites. This cooperative binding is a form of combinatorial control—the use of multiple, rather than single, transcription factors to control expression of a gene. Strings of *cis*-regulatory sequences often appear to be arranged in "modules", each directing expression of the gene to a particular part of the organism at a specified time. Many genes are controlled by several such modules, which often act independently (Figure 1). As a result, the destruction of one module by mutation may eliminate expression of the gene in the time and place specified by that module, but may not affect expression of the gene elsewhere in the organism.

The transcription factors are also modular and, in the laboratory, segments from different transcription factors can be recombined to produce novel types of regulation. Moreover point mutations can alter their DNA-binding specificity, their interactions with other proteins, and their influence (activating or repressing) on transcription. Because many of

the crucial protein-protein interactions made by transcription factors are relatively weak and non-specific, even small changes to them can have large effects on gene regulation.

Based in part on these considerations, it was predicted that losses and gains of *cis*-regulatory sequences (by mutation or recombination), as well as changes in transcription factors, would likely be major sources of evolutionary novelty[2, 5, 7-9]. Many specific examples from a variety of sources have confirmed and extended this basic idea. I shall begin by describing some studies that investigate changes in the regulation of single genes. These studies are often able to link alterations in gene regulation to specific phenotypic changes that may have provided a selective advantage to organisms acquiring the alteration. Examples such as these, largely from metazoan organisms, served as the inspiration for our studies of yeasts. Our genome-wide studies and those of others, described briefly at the end of this introduction and more fully in Chapters 2 to 4, have revealed the modes and magnitude of circuit rewiring that have occurred over long timescales.

Changes in transcriptional regulation of single genes.

Some of the most important insights into transcription circuit rewiring begin with an observable difference between species (or between different members of the same species) and work backwards to the underlying cause. Several excellent and comprehensive reviews of this work have recently appeared[10, 11] and just a few examples are cited below.

A fascinating story begins with the observation that only certain populations of humans can digest lactose as adults. This property, termed lactase persistence, has been traced to specific nucleotide changes in the cis-regulatory sequences controlling the *LCT* gene[12-14]. The *LCT* gene product is an enzyme produced in the small intestine that breaks down lactose; in nearly all mammals and most humans, it is expressed during weaning and is shut off in adults. The "mutant" *LCT cis*-regulatory sequences of lactose-persistent humans apparently allow the enzyme to be synthesized at appreciable levels in the adult intestine. It has been proposed that the domestication of cattle some 10,000 years ago (and the consequent ready availability of milk to adults) provided the selective pressure for these alleles to have spread through pastoral populations. European and African populations of lactose persistent humans carry different mutations in this *cis*-regulatory region, providing a striking example of convergent evolution and neatly demonstrating how a small molecular change in the regulation of a single human gene can have a large adaptive consequence.

In the three spine stickleback (a bony fish), heritable variations in pelvic spines[15] and in pigmentation[16] (phenotypes related to predator escape) have been traced to *cis*-regulatory sequence variants that control the expression of two key developmental regulators. These studies are especially attractive because marine populations of the stickleback have been repeatedly isolated in fresh water lakes, allowing opportunities to observe independent adaptations to different environments, including different types of predators. Here too

convergence is observed: preexisting alleles conferring reduced pelvic structures and pigmentation were independently selected in isolated populations.

Another influential example is found in the repeated loss and gain of wing pigmentation spots in the *Drosophila* lineage[17, 18]. This gain-loss pattern has given rise to a range of modern species, some with and some without spots. Formation of a wingspot requires the spot-specific expression of the yellow protein, which is required to form the dark pigment. A loss and a gain of *yellow* expression in the wingspot position have been traced to changes in the *cis*-regulatory sequences that control *yellow* transcription. As is true for the other examples discussed above, these changes do not appreciably alter the expression of the gene at other times and in other places in the animal, reflecting the modularity of *cis*-regulatory sequences (Figure 1).

Another example from flies, one which makes an entirely different point, is based on the *cis*-regulatory module that directs *evenskipped* (*eve*) expression to its "stripe 2" position in the developing fly larva[19, 20]. When the stripe 2 modules from *D. melanogaster* and *D. pseudoobscura* were compared, numerous differences were noted, including losses/gains of *cis*-regulatory sequences and major differences in their spacing. Yet the two modules (but not hybrids between them) each directed eve expression to the proper place when introduced into *D. melanogaster*. Thus stabilizing selection apparently preserved the function of the *eve* stripe 2 module, while allowing numerous rewiring changes to take place within it. Similarly, in results described here, we see that entire sets of genes can be rewired, apparently without losing their co-expression.

The examples thus far emphasize the importance of changes in *cis*-regulatory sequences at single genes. Changes in transcription factors are also crucial in the evolutionary rewiring of transcription circuits, although the consequences of these changes may be more complex and therefore more difficult to associate with single traits. Members of the nuclear steroid receptor family of transcription factors, for example, have undergone many independent duplications and divergences, providing a whole range of gene expression patterns, each controlled by a different ligand (or combination of ligands)[21]. Similarly, the MADS Box family of transcription factors, which plays key roles in plant development, has greatly expanded via duplication in this lineage. There is growing evidence that with the expansion of this family in plants, its members have taken on novel and shifted roles in plant development[22, 23]. While it is clear that transcription factor families often undergo lineage specific expansions, it is still unclear how often this process yields novel function, rather than just partitioning ancestral function between duplicates.

A different type of evidence for the importance of changes in transcription factors comes from studies of a Hox gene, *Ultrabithorax*[24, 25]. It has been inferred that the ancient acquisition of a short repression domain in the Ultrabithorax protein contributed to the change in limb number seen between many-limbed crustaceans and the hexapod insects. From this example, one can speculate that modification of transcription factors may be an important component of the larger morphological shifts that likely happened in metazoan evolution.

Transcriptional rewiring viewed from the genome.

A complementary approach to studying transcriptional rewiring begins with a molecular description of a transcription circuit, typically a large one involving several transcription factors and many target genes. The circuit is then compared among two or more species, and the differences and similarities observed. In contrast to the examples cited above, this strategy does not require prior knowledge of the phenotypic consequences (if any) of these changes.

This approach has been used to compare circuitry in closely related yeast, fly, and mammal species[26-33]. Typically, bioinformatics, transcriptional profiling and whole-genome chromatin immunoprecipitation (ChIP-Chip) are used, often in combination. The work presented in this thesis takes such a whole-genome approach, using yeast species as model organisms.

S. cerevisiae, long favored as a model organism for basic studies in genetics and cell biology, also serves as an excellent model for answering many evolutionary questions. Of course, a single organism will not suffice when attempting to infer the evolutionary history of a group of species. Here too, *S. cerevisiae* is unmatched; many of *S. cervisiae*'s close and distant relatives (e.g., *C. glabrata, K. lactis* and *C. albicans*) are also genetically tractable and actively studied in many labs. These advantages, combined with the incomparable functional annotation of *S. cerevisiae* and the current availability

of nearly forty fungal genome sequences, make fungal species a particularly attractive set of organisms in which to study the principles underlying evolution of gene regulation. For instance, it is now straightforward to design and order custom ORF or tiling microarrays for any yeast species with a fully sequenced genome. Furthermore, it will soon be affordable to sequence a novel yeast species in less than one week. Thus, one can begin comparing gene expression and transcription factor binding across any group of yeast species desired. However, one challenge with yeast species is identifying phenotypic changes for study, as the relevant modifications are not as easily recognized as those transforming morphology in metazoans.

Rather than seeking a particular phenotypic change to study, I took a genomic approach and studied changes in the molecular architecture of gene regulation across yeast species. This approach yielded some interesting results. In Chapter 2, I propose a model, based on experimental and informatics evidence from extant species, that explains how a set of co-expressed genes (the **a**-specific genes) have transitioned from positive control by a transcriptional activator to negative control by a transcriptional repressor, without losing proper co-expression in the process. In Chapter 3, again combining experiment and computation, I describe how a large combinatorial circuit made up of many interacting transcription regulators has evolved over the past few hundred million years. This evolution has included the massive rewiring of both protein-DNA and protein-protein interactions.

The results described here, together with the results of others, yield the following three insights: (1) a transcription factor's target genes can change rapidly, (2) the same set of co-expressed genes can be regulated by divergent mechanisms in different species, and (3) cooperative binding of transcription factors (a form of combinatorial control) may facilitate circuit rewiring. These conclusions and the avenues this research has opened are discussed at length in Chapter 4.

FIGURES

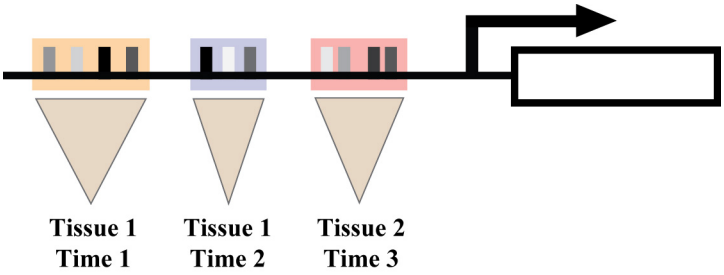


Figure 1. The modularity of gene expression control regions.

cis-regulatory elements for different transcription factors (marks colored black, white and gray) are often arranged in modules (boxes shaded orange, blue and red) which act independently to direct gene expression to a particular part of the organism at a specified time.

Chapter 2

Evolution of alternative transcriptional circuits with identical logic

ABSTRACT

Evolution of gene regulation is a major contributor to the variety of life. Here, we analyze the evolution of a combinatorial transcriptional circuit composed of sequence-specific DNA binding proteins conserved among all eukaryotes. This circuit regulates mating in the ascomycete yeast lineage. We first identify a group of mating genes that was transcriptionally regulated by an activator in a fungal ancestor, but is now transcriptionally regulated by a repressor in modern bakers' yeast. Despite this change in regulatory mechanism, the logical output of the overall circuit remains the same. By examining the regulation of mating in modern yeasts that are related to different extents, we deduce specific, sequential changes in both *cis*- and *trans*-regulatory elements that constitute the transition from positive to negative regulation. These changes suggest specific mechanisms by which fitness barriers were traversed during the transition.

INTRODUCTION

Darwinian evolution posits that natural selection, acting on heritable, random, "successive, slight variations" in organisms over billions of years, can result in novel biological features.[34] While recent work has revealed that biological novelty is often attributable to changes in transcriptional regulation,[2, 35-38] detailed analyses of such changes are often limited to a subset of the *cis*- or *trans*-elements involved.[18, 24-26, 39-41] Here, we present a step-by-step analysis of evolution in a combinatorial transcriptional circuit which regulates mating in multiple yeast species of the ascomycete lineage.

Mating type in the yeasts *Saccharomyces cerevisiae* and *Candida albicans* is controlled by a segment of DNA called the MAT locus.[30, 42, 43] The MAT locus exists in two versions, MAT**a** and MATα, each of which encodes unique sequence-specific DNA binding proteins that direct an extensive program of gene transcription. Cells which express only the MAT**a**- or MATα-encoded DNA binding proteins are **a** cells and α cells respectively, and are specialized for mating. **a** cells express the **a**-specific genes (**a**sgs), required for **a** cells to mate with α cells. Likewise, α cells express the α-specific genes (αsgs). The third cell-type, **a**/α, is formed when an **a** cell mates with an α cell. These cells do not mate, because the **a**sgs and αsgs are turned off (Figure 1a).

While this strategy is the same for both *S. cerevisiae* and *C. albicans,* the molecular details differ in a remarkable way.[30] In *S. cerevisiae*, the **a**sgs are on by default, and are repressed in α and **a**/α cells by a homeodomain protein (α2) encoded by MATα. In *C.*

albicans, however, the **a**sgs are off by default, and are activated in **a** cells by an HMG-domain protein (a2), encoded by MAT**a** (Figure 1a). Both molecular mechanisms give the same logical output: **a**sgs are expressed only in **a** cells. As the a2-activation mode is found over a broad phylogenetic range of yeasts, this strategy most likely represents the ancestral state (Figure 1b).[30, 44-48] In contrast, the a2 gene was recently lost in the *S. cerevisiae* lineage, which now uses the α2-repressing mode of **a**sg regulation,[49] suggesting that α2-mediated repression of **a**sgs is a recent innovation.

The evolutionary transition from positive to negative regulation of the **a**sgs has necessarily included at least two steps: 1) **a**sg expression becoming independent of the activator a2, and 2) **a**sgs coming under negative control of α2. In this work, we use experimental and informatic approaches to identify multiple changes in *cis*- and *trans*-elements that underlie these steps; we also infer the order in which these steps likely occurred.

RESULTS

Identification of ancestral a-specific genes

To understand how α2 came to repress the **a**sgs in *S. cerevisiae*, we first sought the ancestral *cis*-element responsible for positive regulation of **a**sgs by a2. We reasoned that extant yeasts which retain the ancestral regulatory logic, such as *C. albicans*, may also have retained *cis*-elements close to the ancestral form. *C. albicans*, a fungal pathogen of humans, last shared a common ancestor with S. cerevisiae 200-800 million years ago.[30, 50, 51]

We first experimentally identified the **a**sgs in *C. albicans* by comparing the transcriptional profiles of pheromone induced **a**-cells to that of pheromone induced α-cells (Figure 2; for experimental details, see Methods and Figure S1).[30, 52, 53] This comparison revealed a group of six genes induced only in **a** strains (Figure 2c). Below, we show that the gene *STE2* is also **a**-specific. Of these seven total genes, four have orthologs previously classified as **a**-specific in *S. cerevisiae* (*ASG7, BAR1, STE2,* and *STE6*), suggesting that they were **a**-specific in the common ancestor of *S. cerevisiae* and *C. albicans*.

Identification of the DNA regulatory sequence that activates asgs in *C. albicans*

To identify *cis*-elements involved in activation by a2, we submitted *C. albicans* **a**sg promoters (1000 bp) to MEME.[54] In the promoters of six **a**sgs, we found a regulatory element with several distinctive features (Figure 3a). First, at 26 bp long, the element is

unusually specified for a eukaryotic *cis*-acting sequence. Second, the sequence contains a region closely resembling the binding site of Mcm1, a MADS box sequence-specific DNA binding protein that is expressed equally in all three mating types, and is required for both **a**sg and αsg regulation in *S. cerevisiae*. The Mcm1 residues that contact DNA[55, 56] are fully conserved between *C. albicans* and *S. cerevisiae*, strongly implicating this region of the element as a binding site for Mcm1 in *C. albicans.* Third, the putative Mcm1 site in *C. albicans* **a**sg promoters lies adjacent to a motif of the consensus sequence CATTGTC (Figure 3a). The spacing between this motif and the Mcm1 site is invariantly 4 bp. This motif is similar to demonstrated binding sites for a2 orthologs in *S. pombe* and *Neurospora crassa*, and to the α2 monomer site of *S. cerevisiae* (see Figure 3b).[47, 57-59]

Experimental validation of the *C. albicans* asg regulatory sequence

To test whether the motif upstream of *C. albicans* **a**sgs is functional, we fused a wildtype or a mutant fragment of the *STE2* promoter to a GFP reporter (Figure 3c).[60] In the mutant promoter, the conserved motif was mutated from CATTGTC to CATAATC, a change predicted to destroy the a2 binding site. The wildtype promoter activated GFP on exposure to α-factor (Figure 3c), while the mutant promoter showed no induction of GFP (Figure 3d), demonstrating that this *cis*-element is required for a2-dependent activation of **a**sgs.

Analysis of *cis*- asg regulation across species

For ancestral **a**sgs to undergo the transition from positive to negative regulation, a2-bound *cis*-elements were likely lost, while α2-bound elements must have been gained. To investigate when this transition occurred, we first inferred a phylogeny of 16 yeast species whose genomes have been sequenced, then identified orthologs of *C. albicans* and *S. cerevisiae* **a**sgs in all 16 yeasts (Figure 4b, Methods).[61, 62] Position specific scoring matrices (PSSMs) constructed from the *S. cerevisiae* or *C. albicans* **a**sg operators (Figure 4a) were used to scan promoters of each **a**sg ortholog. Maximum $\log_{10}$-odds scores are shown (Figure 4c-d).

S. cerevisiae-like **a**sg operators (an Mcm1 site flanked by two α2 binding sites) were clearly found in orthologous promoters of organisms as far diverged as *S. castellii*. Past *S. castellii*, the presence of an *S. cerevisiae*-like **a**sg operator was diminished, though present in some *C. glabrata, K. lactis, E. gossypii,* and *K. waltii* promoters (Figure 4c). The *C. albicans* PSSM yielded a nearly converse pattern (Figure 4d). Organisms that branch with *C. albicans* have *C. albicans*-like **a**sg operators (an Mcm1 site flanked by a single a2 site); however, this matrix recovered no significant matches in species close to *S. cerevisiae*, correlating with the loss of a2.[49] These results are unchanged by recently proposed alternate phylogenetic topologies.[36]

Identification of the asg operator in the *K. lactis* branch

Neither the *C. albicans* nor the *S. cerevisiae* matrices elicited strong matches in the *K. lactis*-branch yeasts, which share a more recent common ancestor with *S. cerevisiae* than does *C. albicans* (Figure 4b). To independently determine whether this lineage has a

unique **a**sg operator, we submitted promoters of the ancestral **a**sg orthologs (*ASG7, BAR1, STE2,* and *STE6)* from the *K. lactis* branch yeasts to MEME. The top scoring hit was a DNA motif having features in common with both the *S. cerevisiae* and *C. albicans* **a**sg operators, suggesting a transitional form. As in *C. albicans*, this motif contains an Mcm1 site flanked by an a2 site on one side. However, it is also defined on the opposite side, resembling the tripartite operator structure of the *S. cerevisiae* operator. This additional sequence information is similar to both the *S. cerevisiae* α2 and the *C. albicans* a2 site consensus binding sequences; moreover, the spacing from the Mcm1 binding sequence is also similar to that found in *S. cerevisiae* and *C. albicans* **a**sg operators (Figure 4e). An independent clustering analysis of putative **a**sg operators further suggests a transitional form in the *K. lactis* branch (Figure S2).

Because of low genome sequence coverage[63] we did not systematically incorporate the yeast *S. kluyveri*, which branches near *K. lactis* and retains a2,[26, 49, 64, 65] into our studies. However, the available sequences of **a**sg promoters from *S. kluyveri* also contain operators similar to those of the *K. lactis* branch (not shown), suggesting that transitional forms of the operator exist in this species as well.

Emergence of the α2-Mcm1 interaction

Repression of the **a**sgs in *S. cerevisiae* requires a cooperative interaction between the *trans*-factors α2 and Mcm1. To determine when this interaction arose, we aligned orthologs of α2 and Mcm1 across multiple yeast species, then searched for conservation of the interaction interface (Figure 5a-b).[55, 56, 66] The region of Mcm1 known to contact

α2 is highly conserved across all species analyzed (Figure 5a). Many proteins besides α2 contact this region, so the high degree of conservation is not surprising. In contrast, the portion of α2 that contacts Mcm1 varies considerably across yeasts (Figure 5b). A critical 9-residue "linker" region required for the *S. cerevisiae* α2-Mcm1 interaction[66] is highly conserved from *S. cerevisiae* to *C. glabrata,* and is also somewhat conserved in *K. lactis* and *S. kluyveri*; however, this region shows no conservation in yeasts that branch with *C. albicans*, consistent with observations that α2 is not involved in **a**sg expression in *C. albicans* (Figure 1a).[30]

Structural homology modelling of *K. lactis* α2 and Mcm1 using the *S. cerevisiae* crystal structure[56] as a template reveals that, despite several substitutions, the α2-Mcm1 interaction interfaces in *K. lactis* are fully compatible (Figure 5c)[67]; thus, the appearance of the α2-Mcm1 interaction coincides with the emergence of the tripartite, *S. cerevisiae*-like **a**sg operator in the *K. lactis* branch (Figure 4), suggesting that *K. lactis* **a**sg operators are bound by α2-Mcm1. We also know that *K. lactis* a2 is required for wildtype levels of **a** type mating (A.E.T., unpublished work). Taken together, our data suggest that *K. lactis* **a**sgs are controlled by both α2 and a2 through one of three possible scenarios: (1) some operators are bound exclusively by a2-Mcm1 and others are bound exclusively by α2-Mcm1, (2) hybrid operators are bound by both a2-Mcm1 and α2-Mcm1, or (3) a combination of (1) and (2).

DISCUSSION

In this work, we identify a group of genes (the **a**sgs) that was positively regulated in an ancestral yeast, but is negatively regulated in modern *S. cerevisiae*. Orthologs of these genes are required for sexual differentiation in fungal lineages proposed to span up to 1.3 billion years of evolution.[51, 68] We identify specific changes in *cis*- and *trans*-elements that underlie the two critical steps in this transition: 1) **a**sg expression becoming independent of the activator a2, and 2) **a**sg expression coming under negative control of α2. The nature of these changes provide a plausible explanation for how fitness barriers were overcome during the regulatory transition, both in terms of smaller-scale challenges of evolving individual protein-protein and protein-DNA interactions, as well as the larger-scale challenge of maintaining appropriate **a**sg regulation throughout the transition.

Independence of asg expression from the activator a2

During the transition from positive to negative regulation of the **a**sgs, **a**sg expression became independent of the activator a2. We have shown that the transcriptional regulator Mcm1 was present at ancestral **a**sg promoters as a co-activator with a2; in *S. cerevisiae,* Mcm1 is also present at **a**sg promoters, serving as both an activator (on its own) and a co-repressor (with α2). In *S. cerevisiae,* high A/T content surrounding the Mcm1 binding site allows Mcm1 to function without a cofactor.[69] Thus, a simple increase in the A/T content surrounding the ancestral Mcm1 binding site could "tune up" existing Mcm1 activity such that it no longer requires the cofactor a2 to activate transcription. Consistent with this idea, the A/T content flanking Mcm1 sites in *S. cerevisiae* **a**sg

operators is far higher than that flanking Mcm1 sites in *C. albicans* **a**sg operators (Figure 4a).

Establishment of asg repression by α2

On its own, an increase in A/T content flanking the Mcm1 site would lead to inappropriate constitutive activation of the **a**sgs, since Mcm1 is expressed equally in all cell-types. However, **a**sg regulation could be maintained if this increase were accompanied by evolution of α2-mediated repression. Indeed, this is precisely what we observe: *cis*- and *trans*-changes, signifying the emergence of α2-mediated repression in the *K. lactis* branch, accompany the increase in A/T content surrounding the Mcm1 site (Figures 4e, 5). Prior involvement of Mcm1 in **a**sg regulation likely assisted in evolution of α2-mediated repression by increasing the number of surfaces available for α2-promoter interaction to include both protein and DNA.

The similarity of the a2 binding site (CATTGTC) to the α2 binding site (CATGT), in both sequence and spacing from the Mcm1 site, no doubt contributed to the evolution of the *S. cerevisiae* **a**sg operator (Figures 3, 4); a small change to the *cis*-element could convert it from an a2 to an α2 recognition sequence. The similarity of the sites is particularly striking, given that a2 and α2 belong to different protein families (the HMG and homeodomain families, respectively).

Ordering the pathway

An important clue as to the order in which individual *cis*- and *trans*- changes occurred comes from *K. lactis* and *S. kluyveri*. Both yeasts have retained a2 at their MAT**a** loci, yet in both yeasts an α2-Mcm1 interaction interface and a tripartite **a**sg operator similar to the *S. cerevisiae* α2-Mcm1 binding site have emerged. By examining the data in a phylogenetic context (Figure 6d), we can tentatively define the succession of events leading to repression of **a**sgs in modern *S. cerevisiae* as follows: a2-Mcm1 activated **a**sgs in an ancestor (Figure 6a,d). Subsequently, the α2-Mcm1 protein interaction evolved, coincident with evolution of an α2 site and a strengthening of the Mcm1 binding site in the **a**sg operator (Figure 6b,d). After the divergence of *K. lactis*, the α2-Mcm1 *cis*-operator specificity and A/T content were increased, and a2 was lost, completing the hand-off from positive to negative control (Figure 6c,d). A crucial feature of this model is that **a**sgs are appropriately regulated throughout each stage of circuit evolution, a condition made possible by the continued presence of Mcm1. Intriguingly, both the loss of a2 and the conversion of **a**sg regulation to an exclusively negative regulatory scheme coincide with a whole-genome duplication.[70, 71] The evolution of the **a**sg regulatory circuit may have been facilitated in part by greater flexibility in **a**sg regulation conferred by duplication of its component *cis*- and *trans*-elements.

Conclusion

Our analysis demonstrates how a concerted series of subtle changes in *cis*- and *trans*-elements can lead to a profound evolutionary change in the wiring of a combinatorial circuit. These changes include: 1) "tuning up" of a binding site for a ubiquitous activator, making gene expression independent of a cell-type specific activator, 2) a small change in

an existing DNA binding site, converting its recognition from one protein to an unrelated protein, 3) a small change in the amino acid sequence of a sequence specific DNA binding protein, allowing it to bind DNA cooperatively with a second protein. Significantly, the coordinated optimization of protein-DNA and protein-protein interactions we have described allows regulation of the target genes to be maintained throughout a major evolutionary transition. Because the proteins that have participated in this transition represent several highly conserved and prominent protein families, including the MADS box family (Mcm1), the HMG-domain family (a2), and the Homeodomain family (α2), the types of changes we have documented will likely apply to other examples of transcriptional circuit evolution.

METHODS

Strain construction

The pheromone **a**-factor has not yet been identified in *C. albicans*. In order to compare the pheromone response of **a** cells to that of α cells, we "fooled" α-cells into responding to α-factor by ectopically expressing α-factor receptor (strain ATY497), a strategy previously employed in *S. cerevisiae*.[52] Constructs and primers used are listed in Supplementary Materials.

α-factor induction

Strains were grown to OD_{600} 1.0 in YEPD+55μg/ml adenine, then induced with 10 μg/ml α-factor from a stock dissolved in either DMSO or water. Sample preparation and microarrays were previously described.[53] All microarray data are available online.

Yeast phylogeny

Briefly, groups of orthologous genes (See Supplementary Materials) with one and only one representative from each of the 16 yeasts were multiply aligned with ClustalW,[72] then concatenated yielding a single alignment. A maximum-likelihood species tree was inferred from this alignment using the TREE-PUZZLE algorithm.[73] Trees with identical topologies were also generated using additional algorithms (see Supplementary Materials).

Structural Modeling

The K. lactis α2-Mcm1 interaction was modelled using the Protein Local Optimization Program, by Matthew P. Jacobson, (http://francisco.compbio.ucsf.edu/~jacobson/), using the crystal structure of the *S. cerevisiae* α2-Mcm1 complex (PDB ID: 1MNM; Tan S 1998) as a template.

ACKNOWLEDGEMENTS

We are grateful to Matt Jacobson for valuable advice provided in modelling the *K. lactis* α2 and Mcm1 structures. We also wish to thank Stefan Ästrøm for providing unpublished *K. lactis* strains and advice on their handling, Peter Sudbury for providing the GFP reporter construct, and Mike Lorentz and Gerald Fink for the collaboration that produced the DNA microarrays used in this paper, the Broad Institute (http://www.broad.mit.edu/annotation/fungi/fgi/), the Sanger Center (http://www.sanger.ac.uk/Projects/Fungi/), and the Pathogen Sequencing Unit at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/sequencing/Candida/dubliniensis/) for making sequence data available. Mark Ptashne provided valuable comments on the manuscript. We thank Bethann Hromatka for overseeing microarray printing, and Richard Bennett for microarray data and valuable discussions. Rebecca Zordan, Andrew Uhl, Matt Lohse, Mathew Miller, Randy Wu, Christina Chaivorapol, and other members of the Johnson and Li labs provided many useful discussions. This work was supported by grants from the NIH (R01 GM37049 and R01 AI49187) to A.D.J.. A.E.T was supported by a Howard Hughes Medical Institute Predoctoral Fellowship. B.B.T. is a NSF Predoctoral Fellow. B.B.T. and H.L. acknowledge partial support from a Packard Fellowship in Science and Engineering (to H.L.) and NIH grant GM070808.

FIGURES

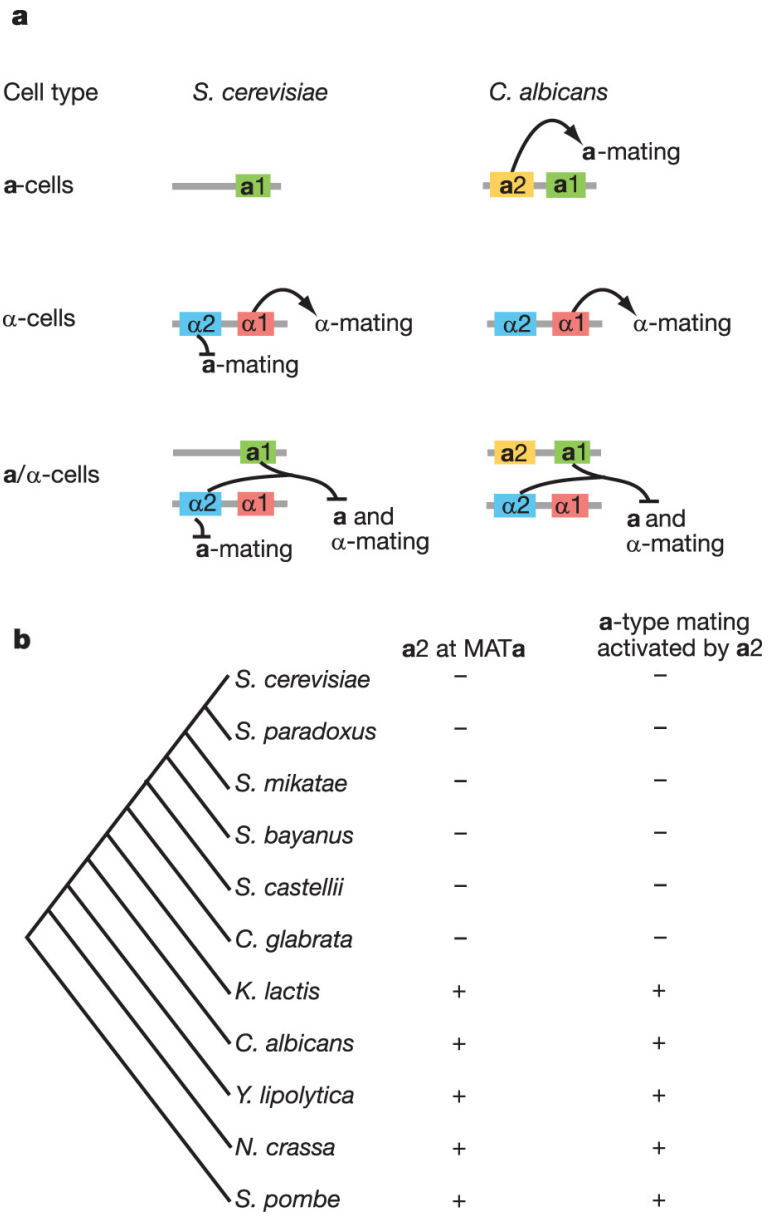


Figure 1. a-type mating is negatively regulated in modern *S. cerevisiae*, but was positively regulated in its ancestor.

(a) *S. cerevisiae* and *C. albicans* transcribe their genes according to one of three programs, which produce the **a**, a, and **a**/α cells. The particular cell-type produced is

determined by the MAT locus, which encodes sequence-specific DNA-binding proteins (colored blocks). Regulation of **a**-type mating differs substantially between *S. cerevisiae* and *C. albicans*. In *S. cerevisiae,* **a**-type mating is repressed in α cells by α2. In *C. albicans*, **a**-type mating is activated in **a** cells by a2. In both organisms, **a**-cells mate with α-cells to form **a**/α cells, which cannot mate.

(b) a2 is an activator of **a**-type mating over a broad phylogenetic range of yeasts.[30, 45-48, 74] In *S. cerevisiae* and close relatives, a2 is missing and α2 has taken over regulation of the **a**sgs.[49]

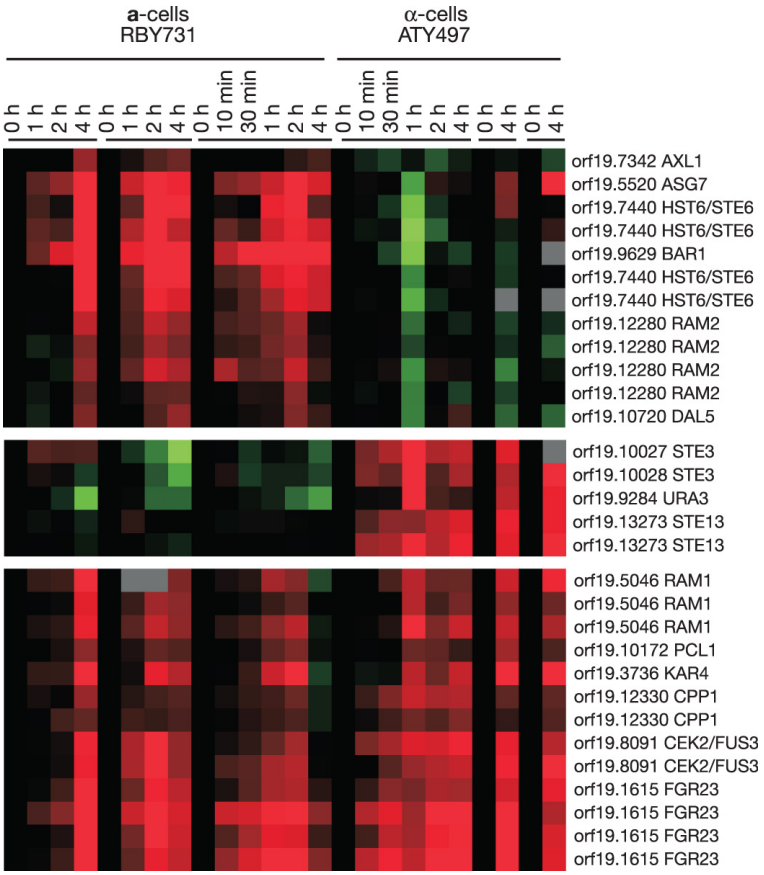


Figure 2. Identification of a-specific genes in *C. albicans.*

Pheromone induction profiles of **a** cells (RBY731) and α cells (ATY497) in six pheromone induction time-courses are compared. Top: genes upregulated only in **a** cells. Middle: genes upregulated only in α cells. *URA3* is induced because it is under control of the *STE2* promoter. Bottom: subset of genes upregulated in both **a** and α cells. The first two time-courses were previously published by Bennett et al.[53]

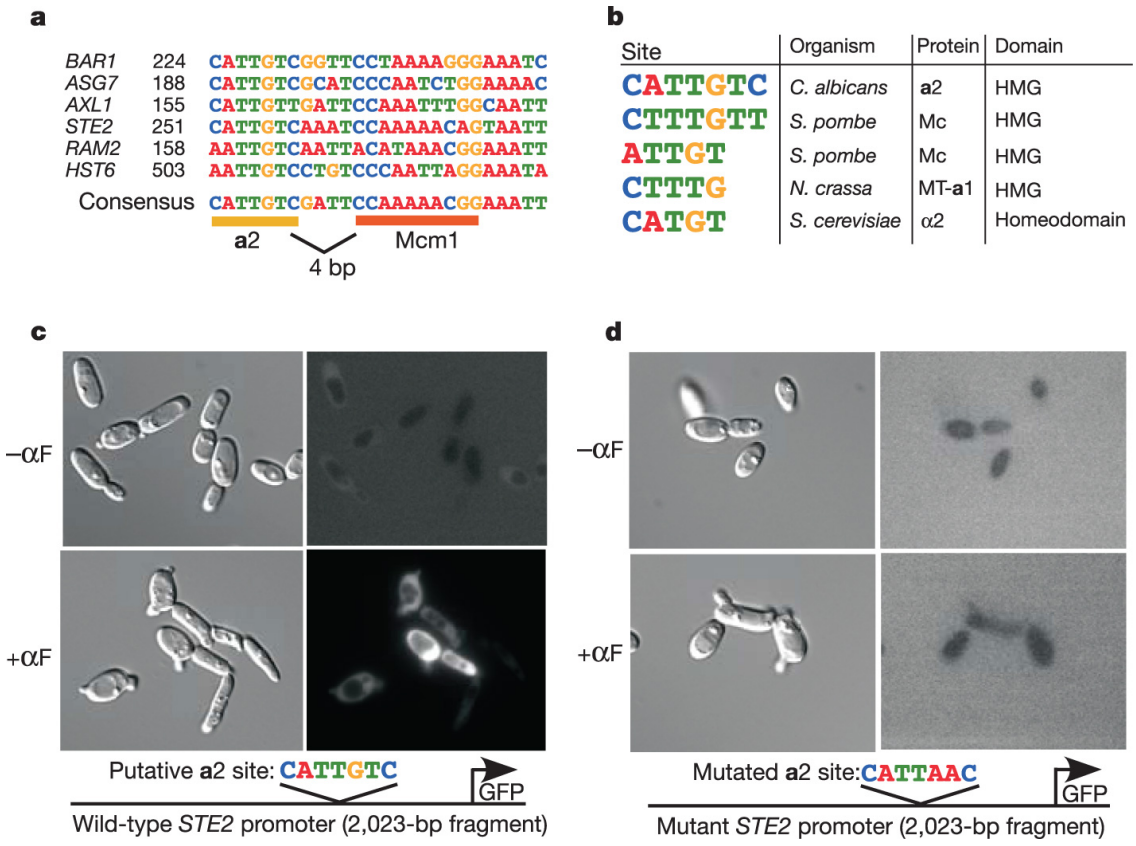


Figure 3. Identification and validation of the *C. albicans* asg operator.

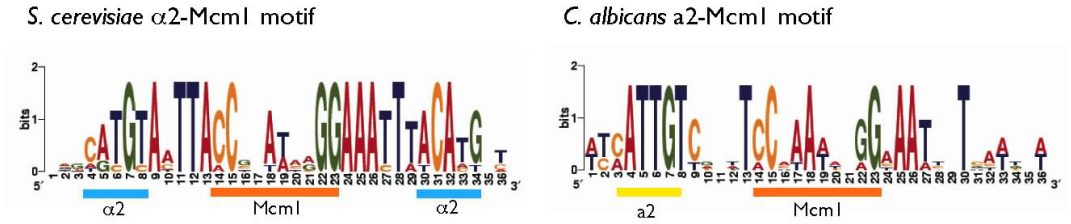
(a) 1000 bp of each *C. albicans* **a**sg promoter were submitted to MEME.[54] The motif shown was present in 6 **a**sg promoters. Distance from the translation start site is indicated. The element contains a conserved 7-bp site (yellow) and a putative Mcm1 binding site (orange), separated by 4 bp.

(b) The 7-bp motif is similar to binding sites of a2 orthologs from *N. crassa* and from *S. pombe*, as well as α2 from *S. cerevisiae*.[47, 57-59] *S. pombe* MatMc binds the two indicated sites equally.[57]

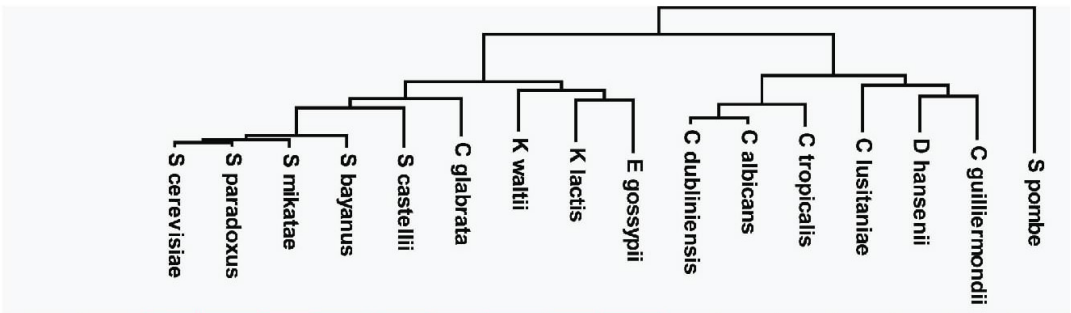
(c,d) A wildtype (c) or mutant (d) 2023-bp fragment of the *STE2* promoter was fused to a GFP reporter and integrated at the RP10 locus of *C. albicans*.[60] Top panels: uninduced

cells. Bottom panels: α-factor induction. Only the wildtype *STE2* promoter activates GFP expression (bottom right panels).

a.



b.



c. *S. cerevisiae* PSSM

YDR461W	MFA1	7.40	8.05	6.10	7.08	6.68	0.97	0.65	2.94	4.57							
YNL145W	MFA2	6.87	7.08	5.48	6.32	6.59											
YGL032C	AGA2	7.79	6.81	6.43	6.39	5.32	1.60	-1.35	1.75	4.43							
YJL170C	ASG7	7.08	-2.11	7.08	7.00	3.13		-0.55	-0.65	2.30	-1.38	-1.38	-0.84	1.51	-0.68		
YIL015W	BAR1	8.54	7.45	7.43	7.76	5.64	3.45	0.39	1.63	0.82	-0.35	-1.52	-0.71		-0.02		
YFL026W	STE2	7.86	6.45	7.86	5.83	5.19	-0.31	0.09	1.47	5.61	-1.76	-0.32	-0.11	-1.60	-1.80	-1.12	2.07
YKL209C	STE6	7.78	7.26	6.27	6.40	2.17	1.96	-1.28	0.32	0.21	0.11	0.41		-2.39	-1.76	1.88	0.64
YKL019W	RAM2	-0.65	-0.84	-0.96	-1.27	-1.59	-1.71	-1.72	-1.11	-1.15	0.98	-0.20	1.83	0.51	0.86	-1.26	-1.26
YPR122W	AXL1	0.38	0.99	0.82	-0.28	3.09	-1.38	-1.87	-1.16	-1.44	-1.35	-0.45	-0.07	-0.94	-1.65	0.65	0.34

d. *C. albicans* PSSM

YDR461W	MFA1	-0.52	-1.16	-0.16	-0.97	0.63	0.05	1.27	0.63	1.61							
YNL145W	MFA2	-0.18	-0.29	-0.41	0.11	-0.47											
YGL032C	AGA2	-0.36	-0.66	-0.49	0.04	-1.46	-0.54	-0.85	1.20	1.47							
YJL170C	ASG7	-0.32	-0.75	-0.33	-0.10	0.33		-0.37	-0.32	2.10	3.24	5.52	3.13	0.97	-0.27		
YIL015W	BAR1	-0.92	-0.79	-1.18	-0.38	0.42	0.07	-0.61	0.81	0.58	5.99	6.44	0.14		0.29		
YFL026W	STE2	-0.61	-1.13	-0.51	0.81	-0.52	-1.01	1.41	0.48	4.55	3.71	4.19	3.50	1.58	2.07	1.86	0.72
YKL209C	STE6	0.59	0.66	0.29	-0.75	0.75	1.52	0.78	0.19	1.12	4.05	4.82		-0.10	1.50	4.37	-0.42
YKL019W	RAM2	-0.43	0.49	0.07	-0.30	-0.52	-1.49	-0.53	-0.76	-0.46	4.13	4.56	1.37	-0.62	-0.57	0.32	-0.97
YPR122W	AXL1	-0.24	-0.43	-0.84	1.22	-0.25	-1.09	0.43	1.00	-0.26	4.61	4.89	0.09	-0.02	0.26	2.73	-0.64

e.

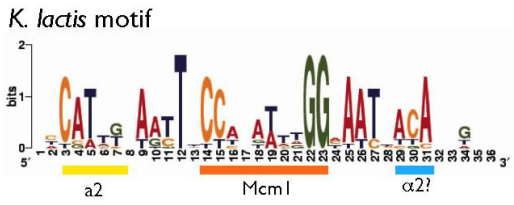


Figure 4. Analysis of *cis*- asg regulation across species.

(a) *S. cerevisiae* α2-Mcm1 and *C. albicans* a2-Mcm1 position specific scoring matrices (PSSMs) were derived from the seven *S. cerevisiae* **a**sg operators, or six *C. albicans* **a**sg operators.

(b) A phylogeny of 16 sequenced yeasts was inferred using methods similar to those of Rokas et al.[61] **a**sg ortholog promoters were scanned with the *S. cerevisiae* PSSM (c) or *C. albicans* PSSM (d). Maximum $\log_{10}$-odds scores are shown. Darker shades of red indicate stronger matches.

(e) Promoters from the *K. waltii, K. lactis,* and *E. gossypii* orthologs of *ASG7, BAR1, STE2,* and *STE6* were pooled and submitted to MEME.[54] The recovered motif has elements of both the *S. cerevisiae* and *C. albicans* **a**sg operators: an a2-like site resembles that of *C. albicans*, while the tripartite structure resembles the *S. cerevisiae* operator.

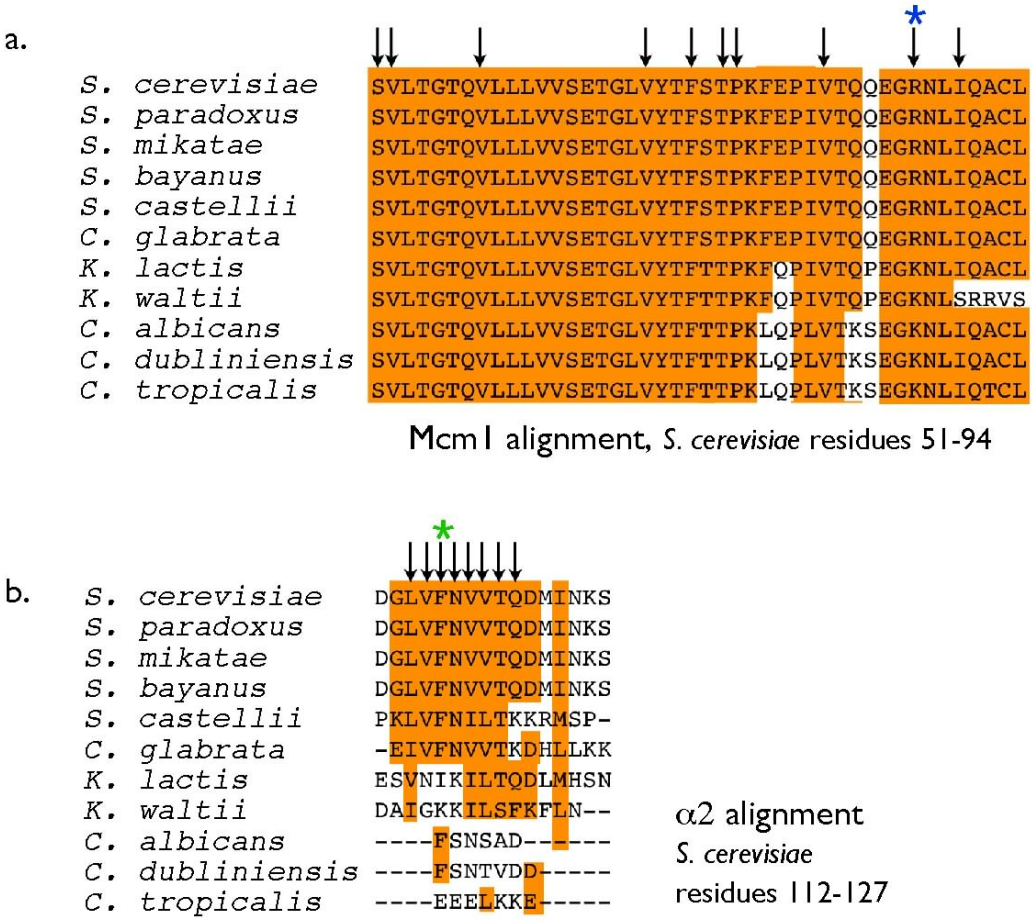


Figure 5. Evolution of the α2-Mcm1 interaction.

(a) Mcm1 from multiple species are aligned. Arrows denote residues of Mcm1 which contact a2 in *S. cerevisiae.*[55, 56]

(b) α2 proteins from multiple species are aligned. Arrows indicate residues of α2 that contact Mcm1, and are required for α2-Mcm1 repression.[56, 66] This region is well-conserved out to *C. glabrata* with *K. lactis* and *S. kluyveri* α2 also showing significant conservation.

(c) The *K. lactis* α2-Mcm1 complex was modelled using the crystal structure of the *S. cerevisiae* α2-Mcm1 complex (PDB ID: 1MNM; Tan S 1998) as a template. Left: *S. cerevisiae* α2 linker region and Mcm1 interface. Mcm1-Arg87 (*blue) and α2-Phe116 (*green) form a favorable pi-stacking interaction. Right: *K. lactis* model. The Arg87-Phe116 interaction is not present, suggesting that the *K. lactis* interaction is weaker than that of *S. cerevisiae*.

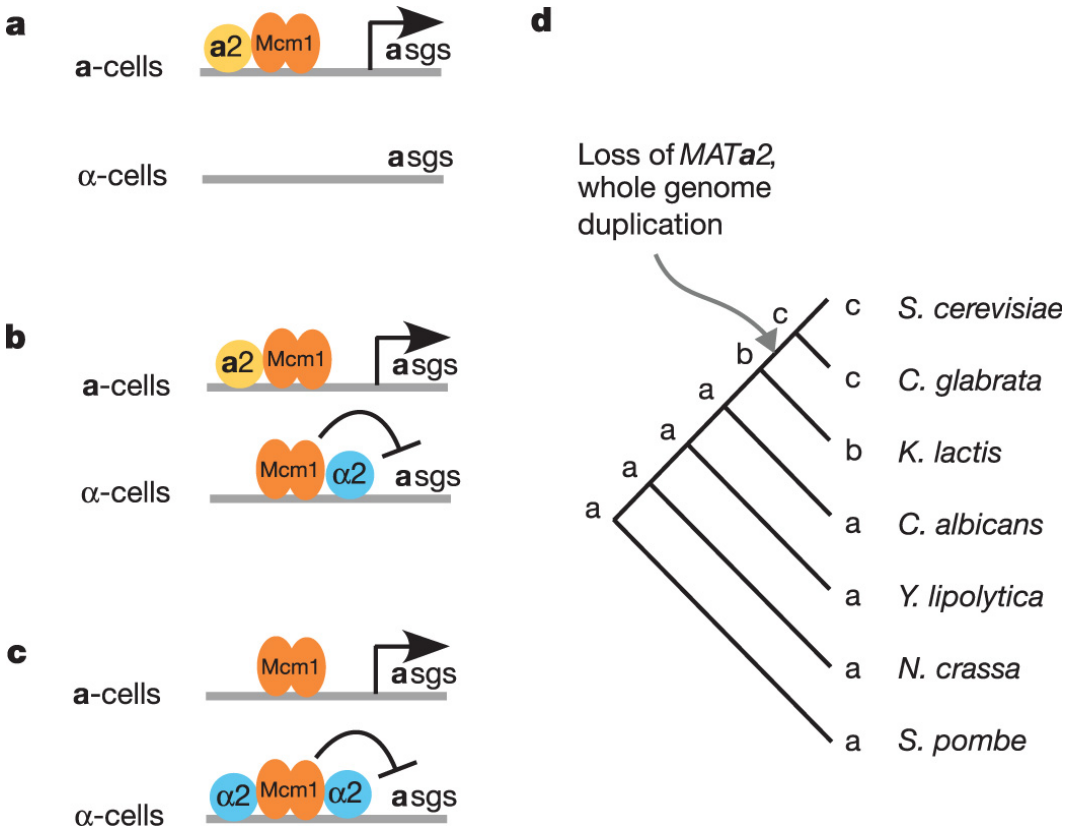


Figure 6. Ordering the changes in *cis*- and *trans*- regulatory elements.

(a) In an ancestral yeast, a2-Mcm1 activated **a**sgs in **a** cells. This scheme persists in modern *C. albicans*.

(b) *cis*- and *trans*-elements in the *K. lactis* branch suggest that **a**sgs are positively regulated by a2-Mcm1 in **a** cells and negatively regulated by α2-Mcm1 in α cells.

(c) In modern *S. cerevisiae,* **a**sgs are activated by Mcm1 in **a** cells and repressed by α2-Mcm1 in α cells.

(d) Regulatory schemes in parts a, b, and c are mapped onto extant species and ancestral nodes. *C. albicans*-*S. pombe* most closely resemble "a",[30, 44-46] while *K. lactis* fits "b," and *S. cerevisiae*-*C. glabrata* fit "c." The most parsimonious evolutionary scenario maps "a" as the ancestral state. "b" is transitional, first appearing in the ancestor of *K. lactis*

and *S. cerevisiae*. “c” is most derived, appearing in the ancestor of *C. glabrata* and *S. cerevisiae*.

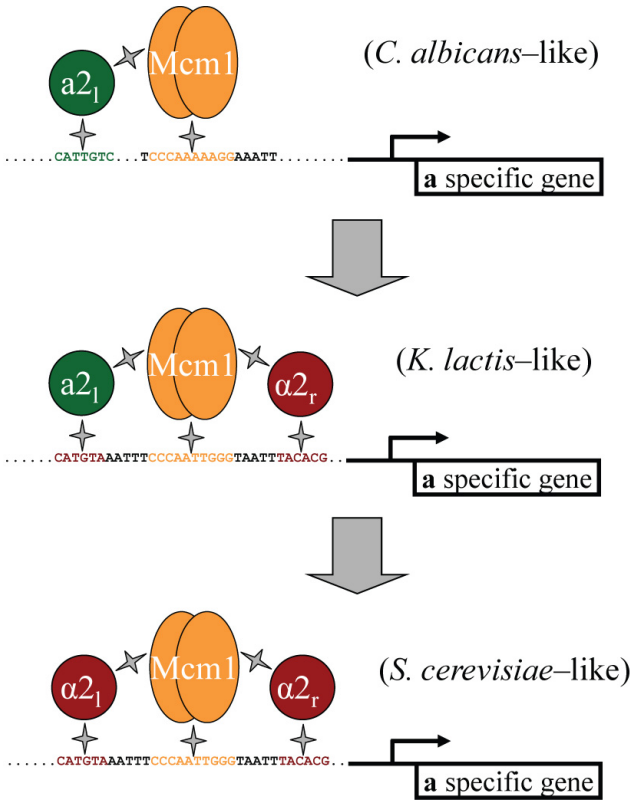


Figure 7. An alternative depiction of the transition from positive to negative regulation of asgs.

In presentations, I often use this alternative depiction of the regulatory transition we inferred to take place at the **a**sgs. I include it here for posterity. Note: for the *K. lactis*-like intermediate, a2 and α2 are predicted to act with Mcm1 in **a** cells and α cells, respectively.

SUPPLEMENTARY METHODS

Strain construction

To express *STE2* (α-factor receptor) in *C. albicans* α cells, we integrated *STE2* under the promoter of the α-specific gene *STE3,* [30] yielding strains ATY496 and 497. The *STE2* ORF was amplified using primers 5'-aacattgagctccgatcaacaaccagtattcc-3' and 5'-tgtgaccgcggttcaatgcctgtaccgtggc-3'. The product was cut with *SacI* and *SacII* and ligated into pDDB57[75] cut with *SacI* and *SacII*, yielding pAT103. A fragment of pAT103 containing *STE2* and *URA3* was amplified using primers 5'-TTAATACCTTAGCATACATAGAGAACTTTATTTTGGCTTCTTAATAATATTTA AAGCAAAGTTTTCCCAGTCACGACGTT-3' and 5'-TAGACTTGTTTTTTTTTATTTATTATATTTTTCATCAACAAGTAACAGGATAC GATGGATGTGGAATTGTGAGCGGATA-3'. This PCR product was purified and transformed into MMY563, an α strain deleted for *MFα*[53] Integration of *STE2* under the *STE3* promoter was confirmed by PCR. The 5' junction was tested using 5'-TTGAGGTATTTGCTGGTGCC-3' and 5'- gttgttgatcgGAGCTCccc-3'. The 3' junction was confirmed with primers 5' TCGTGGCCATACTAACGCCC-3' and 5'-ggcttattatgacacctgg-3'.

Validation of the asg operator

To validate the *C. albicans* **a**sg element, we used wildtype or mutant *STE2* promoters to drive expression of GFP. A 2023-bp region of the *STE2* promoter containing the putative a2-Mcm1 site was amplified using the primers 5'-

ATTACGCGCGGATCCAAGCTTcagattagaagtcaaaagcgctttccactacc-3' and ATO379 5'-TTTACGCAAGGATCCtgaagaaagataaatgaaagaggaatactgg-3'. This product was digested with *BamHI* and cloned into pRSC4b digested with *BamHI*.[60] The pRSC4b plasmid encodes a version of GFP codon-optimized for *C. albicans*.[76] the *URA3* marker, and the *RP10* gene with an engineered *StuI* site. The resulting plasmid, pAT199, was cut with *StuI* and transformed into the **a**-strain RRY15, yielding the strain ATY631. Integration at the *RP10* locus was confirmed by PCR. The 5' integration event was tested with primers 5'- GATTTTGTTACAGCGTAACCAGTGCG-3' and 5'-GATTTATGAAAGTTTTCAGCTCTAGTCACG-3'. The 3' integration event was tested with primers 5'- GGAATGTGGTAGTCGATATTCAGGG-3' and 5'-ccaactccaggtgtcataataagccaatc-3'.

To create the mutation in the putative a2 site, we used the Quikchange Multi-Site Directed Mutagenesis kit (Stratagene) with primer 5'-ttttttgcagcaatCATTaaCAAATCCAAAAACAGTAATTttcc-3' to mutagenize pAT199. The mutagenized plasmid, pAT205, was integrated at the *RP10* locus of RRY15 as described above, yielding the strain ATY634. Strains were switched the opaque phase as previously described,[77] then induced with 10 μg/ml α-factor (stock concentration of 10 mg/ml, dissolved in 10% DMSO).

Preparation of cultures and cDNA for microarray experiments.

α-factor inductions and microarray protocols were described previously in detail,[53] with the following modification: here, strains were grown to OD1.0 in YEPD+55μg/ml

adenine, then induced with 10 μg/ml α-factor, using either a stock concentration of 10 mg/ml dissolved in 10% DMSO or 100 μg/ml dissolved in water in order to separate effects of DMSO and α-factor.

Phylogeny reconstruction

A robust phylogeny of the 16 yeast species was inferred using methods similar to Rokas et al.[61] This method requires identification of orthologous genes. At the time of this report, 4 of the 16 yeast genomes were not yet annotated; ORFs for these four genomes were annotated by translating genomic nucleotide sequence in all six reading frames and retaining any contiguous amino acid sequence longer than 100 residues (i.e. no stop codon). To build the phylogeny, groups of orthologous genes containing one and only one representative from each of the 16 yeasts were chosen at a stringent branch length cutoff (0.4; see *Orthologous ORF mapping* below). This yielded 139 groups of orthologous genes that, showing no evidence for deletion or duplication events, were more likely than other groups of orthologous genes to preserve the underlying speciation signal. The 16 sequences within each group were then multiply aligned with ClustalW.[72] The resulting 139 alignments were concatenated and columns containing gaps were dropped, producing a single alignment with 37,963 columns. Finally, a maximum likelihood species tree was estimated employing the TREE-PUZZLE algorithm with default parameters (VT substitution model).[73] Demonstrating the robustness of this inference, a tree with identical topology and similar branch lengths was generated when either the maximum likelihood algorithm PHYML[78] (using the WAG substitution model as recommended by ProtTest[79]) or the neighbor-joining method of ClustalW[72] was

applied to the same dataset (not shown). Concerned that the whole genome duplication (WGD) may have affected our prediction of orthologs within the lineage including *S. cerevisiae, S. castellii* and *C. glabrata*,[80] we filtered 9 groups of orthologous genes (from our alignment of 139) that are affected by the phenomenon of differential gene loss following the WGD. The reduced set of 130 alignments was submitted for the same analysis as the previous set (TREE-PUZZLE, PHYML, etc.) and trees with the same topology and similar branch lengths to those seen with the previous alignment were produced.

Orthologous ORF mapping

Reciprocal best BLAST hit (RBBH) is often used in comparative genomics studies to pair orthologous genes between species.for example, see [26] However, in its unmodified form this method can not take advantage of the added information offered by multiple sequence alignment. Furthermore, RBBH does not treat the one-to-many and many-to-many evolutionary relationships that can arise due to gene duplication and which have been shown to be particularly prominent within the yeast lineage under study.[70, 81] Because our primary interest is to examine the evolution of the regulation of genes that can be traced to a single gene in an ancestral organism, we devised the following method for mapping orthologs. We ran PSI-BLAST for each *S. cerevisiae* ORF "query" sequence against a single database containing all ORF sequences from each of the 16 fungal species, employing an E-value cutoff of 10^{-5} and the Smith-Waterman alignment option.[82] The sequences returned by PSI-BLAST were then multiply aligned with ClustalW (using the fast alignment option) and a neighbor joining tree (NJ) was inferred,

again using ClustalW.[72] Finally, the resulting NJ tree was traversed to extract a set of orthologous genes in the following manner: Start at the leaf node for the query sequence and ascend the tree, incrementing a level counter for each node ascended. At each internal node descend. If a leaf node is reached, the gene is from a species not yet seen at a lower level, and the branch length traversed is less than a cutoff (1.0), then add that gene to the set of orthologous genes. This procedure was repeated for each *S. cerevisiae* sequence, resulting in a 16 species many-to-many ortholog map.

Ortholog mapping in this automated fashion was employed for the purposes of reconstructing the fungal phylogeny (see *Phylogeny reconstruction*). Due to the small number of **a**sgs (7 **a**sgs in *S. cerevisiae* and 6 **a**sgs in *C. albicans)*, we were able to incorporate additional biological information in predicting orthologs. Manual mapping maximized the accuracy and coverage of our ortholog mapping. For example, *MFA1* and *MFA2*, two **a**sgs from *S. cerevisiae*, are less than 40 amino acids long and were therefore not annotated as ORFs in several of the fungal sequencing projects. Using TBLASTN we identified putative *MFA1/MFA2* orthologs and added them to our sequence database and ortholog map. The *C. albicans* ortholog to *S. cerevisiae BAR1*, which encodes an **a**-specific aspartyl protease, could not be identified on the basis of protein sequence data alone. However, orf19.9629 belonging to the family of aspartyl proteases clustered with the *C. albicans* **a**sgs in the microarray experiments (Figure 2c) and had an a2-Mcm1 motif in its promoter region (Figure 3a). Additionally, we have experimental evidence that this ORF is a functional homolog (R.J. Bennett, personal communication). On this basis, we could confidently assign this *C. albicans* ORF as the ortholog to *S. cerevisiae*

BAR1 and in turn assign ORFs from other *Candida* species to the *BAR1* ortholog group as well. Finally, within the clade of organisms descendent from the common ancestor of *S. cerevisiae* and *K. lactis* we verified our **a**sg ortholog mapping using synteny information provided by the Yeast Gene Order Brower.[62]

Alignment of α2 and Mcm1 orthologs

Putative orthologs of *S. cerevisiae* Mcm1 from the 15 other fungal genomes were taken directly from our 16 species ortholog map. Because MATα2 contains an intron that is sometimes not properly annotated, putative orthologs had to be selected more carefully. For each genome we undertook an iterative process of TBLASTN query, nucleotide sequence extraction, splice site prediction and translation that yielded MATα2 orthologs from each species except *E. gossypii*, *C. lusitaniae*, *D. hansenii*, *C. guillermondii* and *S. pombe*. It is likely that the MATα2 ortholog was not found in these species either because the MATα mating-type locus was not sequenced (e.g., an **a** strain of *E. gossypii* was used for genome sequencing) or because the gene is simply not present in that genome (e.g., *C. lusitaniae* and *D. hansenii*, where α2 is not encoded by the α locus). Although the *K. delphensis* genome has not been sequenced, the MATα2 gene has been sequenced[49] and was therefore added to our set of MATα2 orthologs. *K. delphensis* branches with *C. glabrata*. The translated Mcm1 and MATα2 sequences were then aligned with T-COFFEE[83] using default parameters.

Homology mapping of *K. lactis* sequences onto *S. cerevisiae* a2-Mcm1 structure

We modeled the *K. lactis* α2-Mcm1 complex with the Protein Local Optimization Program (PLOP, Matt Jacobson) using the crystal structure of the *S. cerevisiae* α2-Mcm1 complex (PDB ID: 1MNM; Tan S 1998) as a template. First, pairwise sequence alignments of α2 and Mcm1 (between *K. lactis* and *S. cerevisiae*) were extracted from their respective multiple alignments (see *Alignment of* α*2 and Mcm1 orthologs*). The *K. lactis* α2 and Mcm1 chains were then modeled independently with PLOP utilizing both the pairwise alignments and the *S. cerevisiae* α2-Mcm1 complex. After modeling, the chains were recombined and amino acid side chains were optimized.

Microarray data, ortholog alignments, species trees, and homology models are available online at http://genome.ucsf.edu/asg_evolution/.

SUPPLEMENTARY TABLES

Table S1

C. albicans a2-Mcm1 Position Specific Scoring Matrices

ALPHABET= ACGT
probability matrix: alength= 4

0.400	0.100	0.100	0.400
0.100	0.300	0.100	0.500
0.300	0.500	0.100	0.100
0.700	0.100	0.100	0.100
0.100	0.100	0.100	0.700
0.100	0.100	0.100	0.700
0.100	0.100	0.700	0.100
0.100	0.100	0.100	0.700
0.100	0.600	0.100	0.200
0.300	0.200	0.400	0.100
0.400	0.200	0.200	0.200
0.300	0.100	0.200	0.400
0.100	0.100	0.100	0.700
0.200	0.600	0.100	0.100
0.100	0.700	0.100	0.100
0.400	0.300	0.100	0.200
0.600	0.100	0.100	0.200
0.700	0.100	0.100	0.100
0.400	0.100	0.100	0.400
0.400	0.200	0.100	0.300
0.200	0.300	0.200	0.300
0.200	0.100	0.600	0.100
0.100	0.100	0.700	0.100
0.500	0.200	0.100	0.200
0.700	0.100	0.100	0.100
0.700	0.100	0.100	0.100
0.200	0.100	0.100	0.600
0.200	0.300	0.100	0.400
0.400	0.200	0.200	0.200
0.100	0.100	0.100	0.700
0.100	0.400	0.300	0.200
0.500	0.200	0.100	0.200
0.400	0.100	0.100	0.400
0.100	0.200	0.300	0.400
0.300	0.300	0.200	0.200
0.400	0.100	0.100	0.400

Table S2:

S. cerevisiae α2-Mcm1 Position Specific Scoring Matrices

6 bp spacing:

ALPHABET= ACGT
probability matrix: alength= 4
0.250 0.250 0.167 0.333
0.333 0.083 0.333 0.250
0.167 0.167 0.500 0.167
0.167 0.583 0.083 0.167
0.583 0.083 0.250 0.083
0.083 0.167 0.083 0.667
0.083 0.083 0.750 0.083
0.083 0.167 0.083 0.667
0.750 0.083 0.083 0.083
0.500 0.250 0.083 0.167
0.083 0.083 0.083 0.750
0.083 0.083 0.083 0.750
0.750 0.083 0.083 0.083
0.167 0.667 0.083 0.083
0.083 0.750 0.083 0.083
0.083 0.250 0.417 0.250
0.417 0.167 0.167 0.250
0.667 0.083 0.083 0.167
0.333 0.083 0.083 0.500
0.417 0.250 0.083 0.250
0.417 0.167 0.333 0.083
0.083 0.083 0.750 0.083
0.083 0.083 0.750 0.083
0.750 0.083 0.083 0.083
0.750 0.083 0.083 0.083
0.750 0.083 0.083 0.083
0.083 0.167 0.083 0.667
0.083 0.083 0.083 0.750
0.200 0.100 0.100 0.600
0.667 0.083 0.083 0.167
0.083 0.750 0.083 0.083
0.750 0.083 0.083 0.083
0.167 0.167 0.083 0.583
0.083 0.083 0.667 0.167
0.167 0.167 0.333 0.333
0.167 0.250 0.083 0.500

5 bp spacing:

ALPHABET= ACGT
probability matrix: alength= 4
0.250 0.250 0.167 0.333
0.333 0.083 0.333 0.250
0.167 0.167 0.500 0.167
0.167 0.583 0.083 0.167
0.583 0.083 0.250 0.083
0.083 0.167 0.083 0.667
0.083 0.083 0.750 0.083
0.083 0.167 0.083 0.667
0.750 0.083 0.083 0.083
0.500 0.250 0.083 0.167
0.083 0.083 0.083 0.750
0.083 0.083 0.083 0.750
0.750 0.083 0.083 0.083
0.167 0.667 0.083 0.083
0.083 0.750 0.083 0.083
0.083 0.250 0.417 0.250
0.417 0.167 0.167 0.250
0.667 0.083 0.083 0.167
0.333 0.083 0.083 0.500
0.417 0.250 0.083 0.250
0.417 0.167 0.333 0.083
0.083 0.083 0.750 0.083
0.083 0.083 0.750 0.083
0.750 0.083 0.083 0.083
0.750 0.083 0.083 0.083
0.750 0.083 0.083 0.083
0.083 0.167 0.083 0.667
0.083 0.083 0.083 0.750
0.667 0.083 0.083 0.167
0.083 0.750 0.083 0.083
0.750 0.083 0.083 0.083
0.167 0.167 0.083 0.583
0.083 0.083 0.667 0.167
0.167 0.167 0.333 0.333
0.167 0.250 0.083 0.500
0.417 0.083 0.250 0.250

SUPPLEMENTARY FIGURES

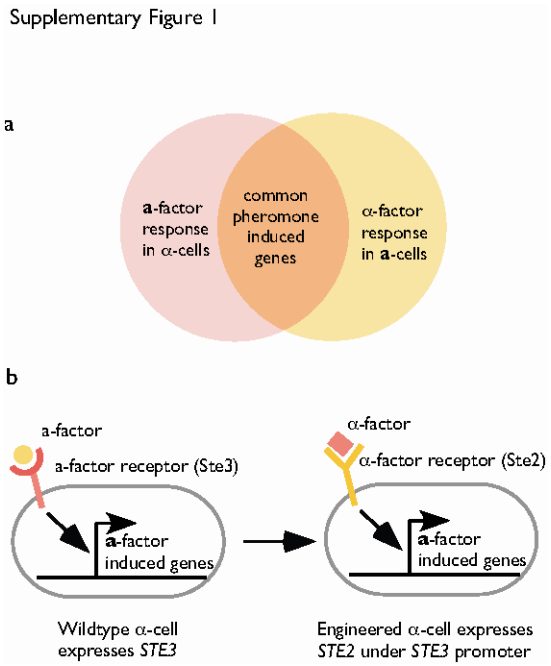


Figure S1. Identification of *C. albicans* asgs.

(a) Detectable **a**-specific gene expression in *C. albicans* **a** cells requires exposure to α-factor, the pheromone from the opposite mating-type.[30] However, the genes induced by α-factor include not only the **a**sgs, but also a large group of general pheromone response genes.[53] Parsing the **a**sgs from the general pheromone response genes can be accomplished by comparing the transcriptional profile of **a** cells induced by α-factor to that of α cells induced by **a**-factor.

(b) Because **a**-factor has not been identified in *C. albicans*, we "fooled" α-cells into responding to α-factor by ectopically expressing the α-factor receptor *STE2*, a strategy described previously in *S. cerevisiae*.[52]

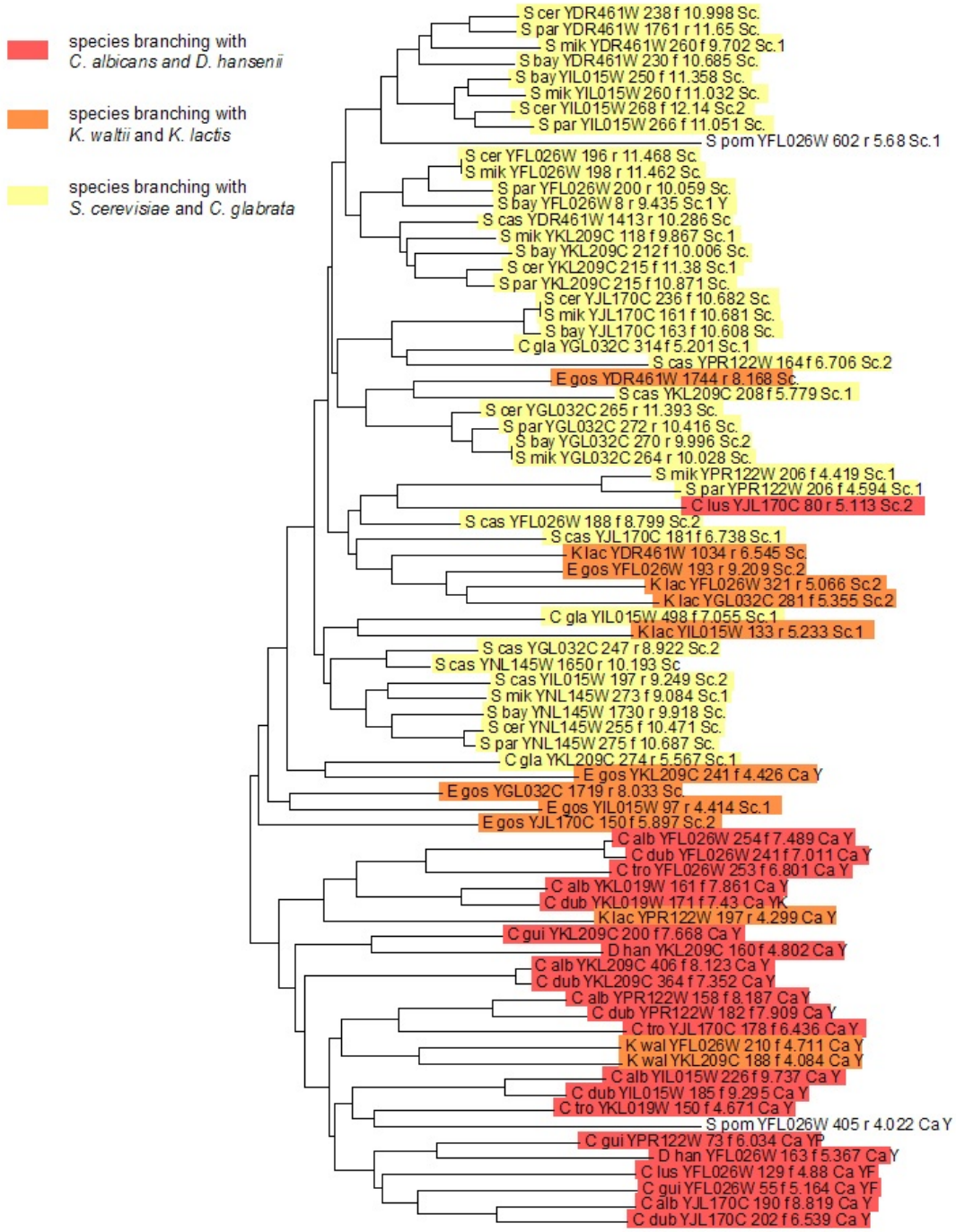


Figure S2. Clustering asg operators.

An alternative approach to determining the number of distinct operator groups is to extract the putative operator elements upstream of each **a**sg ortholog in the 16 species and to cluster an alignment of these sequences. We scored all subsequences in each **a**sg

ortholog promoter (2kb upstream of the translational start) with either the *S. cerevisiae* α2-Mcm1-like PSSM or the *C. albicans* a2-Mcm1-like PSSM (as in Figure 4d) and chose to add to our alignment the max scoring subsequence upstream of each **a**sg ortholog if it either received a $\log_{10}$-odds score greater than 5.0 or received a $\log_{10}$-odds score greater than 3.0 and was found within 600bp of the translational start. This set of operator sequences was then hand-aligned by adding a single gap beteen putative α2 and Mcm1 sites when it improved the alignment (it has been shown that in *S. cerevisiae* the TGT of the α2 site can be placed either 5 or 6bp from the CC of the Mcm1 site[84]). A neighbor-joining tree was then constructed from this alignment with ClustalW (see Supplementary Figure 1). The point of this procedure was to cluster sequences by similarity, as constructing a phylogeny is not possible given that some sequences in the set are unrelated. The clustering clearly indicates two distinct groups, the α2-Mcm1-like operators fall into one group (note that this group contains all operator elements from the *S. cerevisiae* to *C. glabrata* branch) and the a2-Mcm1-like operators fall into another (note that this group contains all but one operator element from the *C. albicans* to *D. hansenii* branch). Operators from the *K. lactis* to *K. waltii* branch tend to co-cluster with both the a2-Mcm1-like group and the α2-Mcm1-like group, adding confidence to our model that these operators signify a hybrid of the *S.cerevisiae* and *C. albicans* regulatory modes.

Chapter 3

The evolution of combinatorial gene regulation in fungi

ABSTRACT

It is widely suspected that gene regulatory networks are highly plastic. The rapid turnover of transcription factor binding sites has been predicted on theoretical grounds and has been experimentally demonstrated in closely related species. Here we combine experimental approaches with comparative genomics, to focus on the role of combinatorial control in the evolution of a large transcriptional circuit in the fungal lineage. Our study centers on Mcm1, a transcriptional regulator that, in combination with five cofactors, binds roughly 4% of the genes in *S. cerevisiae* and regulates processes ranging from the cell-cycle to mating. In *K. lactis* and *C. albicans*, two other hemiascomycetes, we find that the Mcm1 combinatorial circuits are substantially different. This massive rewiring of the Mcm1 circuitry has involved both substantial gain and loss of targets in ancient combinatorial circuits as well as the formation of new combinatorial interactions. We have dissected the gains and losses on the global level into subsets of functionally and temporally related changes. One particularly dramatic change is the acquisition of Mcm1 binding sites in close proximity to Rap1 binding sites at seventy ribosomal protein genes in the *K. lactis* lineage. Another intriguing and very recent gain occurs in the *C. albicans* lineage, where Mcm1 is found to bind in combination with the regulator Wor1 at many genes which function in processes associated with adaptation to the human host, including the white-opaque epigenetic switch. The large turnover of Mcm1 binding sites and the evolution of new Mcm1-cofactor interactions illuminate in sharp detail the rapid evolution of combinatorial transcription networks.

INTRODUCTION

The recent genome sequencing and annotation of the major model organisms established that organismal complexity does not scale in a simple way with gene count. This discordance is consistent with earlier proposals that "tinkering" with gene regulation may be a particularly powerful mode of evolution[1, 85, 86]. In principle, changes in when and where, and thereby in what combinations, genes are expressed can help to explain changes in organismal complexity over longer timescales. Over shorter timescales the contributions of changes in gene regulation to phenotypic variation have been clearly demonstrated[5, 10]. For example, small changes in gene regulation underlie the gain and loss of wing spots in *Drosophila* species[18] and armor in stickleback fish[15].

The plasticity of gene regulatory networks is of interest because it presumably relates directly to the ability of these networks to generate phenotypic novelty[87]. The potential for rapid turnover (gains and losses) of transcription factor binding sites was predicted on theoretical grounds[88-90] and was supported by comparisons of *cis* regulatory sequence both within and between species[33, 91-93]. Recently, experimental localization of four transcription factors across the mouse and human genomes revealed that binding sites have diverged appreciably between these two species[29]. Analogous experiments performed on two transcription factors from closely related yeasts led to similar conclusions[28], though in this case it was not clear how the differences in binding related to gains and losses of *cis*-acting sequences.

The ascomycete lineage, which includes the model yeast *S. cerevisiae*, serves as a powerful framework for investigating the general impact of regulatory evolution because several of its members are particularly easy to study experimentally. These include the model yeast *S. cerevisiae*, the dairy yeast *K. lactis*, and the human pathogen *C. albicans*. *S. cerevisiae* and *K. lactis* diverged more recently than either did from *C. albicans*; the divergence of *S. cerevisiae* and *C. albicans* is thought to have occurred on the order of 300 million years ago[94]. To date only a handful of comparative gene regulation studies have been carried out in fungi. These include a few large scale analyses of changes in gene expression[95] and *cis* regulatory motifs[26, 96] as well as some smaller scale studies[27, 31, 39] focusing on sets of co-regulated genes. While the whole-network studies have generally uncovered an abundance of divergence, the smaller scale studies have characterized this divergence in greater detail or provided mechanistic insight into transcriptional rewiring.

Here we take an approach intermediate in scale and attempt to characterize the evolution of a large combinatorial circuit comprised of the MADS-box transcriptional regulator Mcm1 and each of its cofactors. Mcm1 has been intensively studied in *S. cerevisiae* and, in most cases, it is found as a homodimer that binds DNA cooperatively with other sequence-specific DNA binding cofactors to regulate sets of genes termed regulons. Five regulons have been identified in *S. cerevisiae* where Mcm1 acts in combination with a second transcriptional regulator. Mcm1 joins with (1) MATα2 to turn off the **a**-specific genes (**a**sgs), (2) MATα1 to turn on the α-specific genes (αsgs), (3) Fkh2 and Ndd1 to activate G2/M-specific genes, (4) Yox1 to repress the M/G1-specific genes and (5) Arg80

and Arg81 to either repress or activate the arginine metabolic genes[97]. Because Mcm1 itself is not generally regulated, it is typically the regulation of its cofactors that produces the effect of differential gene regulation at each of the Mcm1-cofactor regulons[97]. For example, it is the regulated binding of Mcm1's cofactor Yox1 that leads to the M/G1 specific expression of genes in the Mcm1-Yox1 regulon[98]. At these Mcm1-cofactor regulons, Mcm1 is thought to increase specificity through added protein-DNA and protein-protein interactions[58].

Previously we showed Mcm1 to be at the center of a rewiring event that led to replacement of one cofactor (MATa2) with another (MATα2)[31]. In principle, the free energy gain contributed by the interaction between Mcm1 and its flanking cofactor could catalyze evolutionary change by expanding the space of *cis*-regulatory sequences that yield appropriate gene regulation. For instance, mutations that strengthen Mcm1's interaction with its cofactor or with DNA can compensate for mutations to the cofactor-DNA interaction, thereby expanding the possibilities for cross-reaction with a new DNA binding protein. This idea bears at least a formal similarity to the neutral networks in RNA sequence space studied by Fontana and colleagues[99]. Because Mcm1 participates in many combinatorial interactions in *S. cerevisiae* and because it regulates a large number of genes, we felt Mcm1 provided a particularly strong entry point to study the evolution of combinatorial networks.

To study this problem we performed ChIP-Chip (chromatin immunoprecipitation, analyzed genome-wide using microarrays) on Mcm1 in three species (*S. cerevisiae, K.*

lactis and *C. albicans*) and combined this data with informatics analyses across 32 fungal species. We find that all five Mcm1-cofactor regulons currently characterized in *S. cerevisiae* are present at least in limited form in *K. lactis* and *C. albicans*, suggesting an ancient origin of these regulons. Although the Mcm1-cofactor interaction is typically conserved and a small set of core target genes remains part of the regulon in each species, most of these regulons have undergone substantial divergence through gain and loss of *cis*-acting sequences. On the global level, substantial gain and loss of Mcm1 binding sites is also evident. Although some of this, as discussed above, is due to target genes moving in and out of existing regulons, much of it is due to the evolution of entirely new Mcm1-cofactor regulons. We highlight two specific instances in which combinatorial regulation by Mcm1 and a cofactor is gained; in one case we observe large-scale convergent evolution of regulation at the ribosomal genes and in the other we describe a very recent gain of regulation that was likely shaped by the selective pressures of the human host. The picture that emerges from this study is one of massive transcriptional rewiring in species that span approximately the same range of protein sequence divergence as human, fish and sea squirt[100, 101]. This rewiring consists of both rapid turnover of *cis*-acting sequence and the formation of new combinations of regulatory proteins.

RESULTS

Mcm1 binds upstream of approximately 4% of genes in *S. cerevisiae* and approximately 12% of genes in *K. lactis* and *C. albicans*.

Mcm1 was chromatin immunoprecipitated (ChIP-ed) from *S. cerevisiae*, *K. lactis* and *C. albicans* **a** cells using peptide antibodies custom designed for the Mcm1 ortholog of each species. To maximize the detection of Mcm1 binding, each strain was grown under two different conditions known to stimulate binding of Mcm1: YEPD medium and pheromone inducing medium with α pheromone (details in the Supporting Text). Immunoprecipitate and whole cell extract samples were competitively hybridized to custom designed Agilent microarrays that tile 60mer probes at a median spacing of 66, 59 and 79 bp across the genomes of *S. cerevisiae*, *K. lactis* and *C. albicans* respectively (Figure S1). For each species/condition the ChIP-Chip was performed twice and the two biological replicates were combined in downstream data processing. Data were processed by a variety of methods and it was determined empirically that the Joint Binding Deconvolution (JBD) algorithm[102] provides the best combination of consistency across species and accuracy on a test set of previously characterized *S. cerevisiae* binding sites (see the Supp. Text). Complete ChIP profiles for all experiments can be viewed at: http://genome.ucsf.edu/mcm1_evolution/.

The majority of regions which JBD called as bound by Mcm1 contained at least one instance of the well-characterized Mcm1 binding motif[103, 104]. We therefore decided to incorporate motif information into our final criteria for Mcm1 bound segments. *De novo*

motif finding by MEME[54] on a set of high confidence bound regions from JBD yields Mcm1 binding site motifs that are roughly the same in each species (Figure 1); the motif deduced *de novo* from *S. cerevisiae* closely resembles previously described Mcm1 recognition sequences. In *C. albicans* there was a large subset of bound regions without a canonical Mcm1 motif. These regions are largely explained by the appearance of a non-canonical motif (Figure 1), discussed later.

Parameter cutoffs for JBD statistics and the motif p-value were chosen that correctly call 85% (28 of 33 genes) of our *S. cerevisiae* test set as positives while also calling an additional 219 of 5769 genes as bound. Details regarding test set selection are provided in the Supp. Text along with receiver operator characteristic plots (Figure S8) evaluating a variety of parameter value choices and a discussion of false positive rates. These same cutoffs used for the *S. cerevisiae* data yield 626 of 5327 genes bound in *K. lactis* and 761 of 6090 genes bound in *C. albicans* (gene lists in Table S1). For these and all subsequent calculations, Mcm1 targets from the two growth conditions examined have been pooled.

Genes bound in any one species are only moderately likely to be regulated in one of the other two species.

After defining Mcm1 targets in each species, we sought to evaluate the overlap of these targets between species. We mapped orthologs using an existing algorithm[31], which was modified to reduce directional bias (Supporting Text, "Mapping orthologous gene sets"), on an updated database of ORF sequences from 32 fully sequenced genomes (Supp. Table 2).

Genes bound by Mcm1 in each species A were then "mapped to" one of the other two species B via our ortholog map. The number of genes "mapped from" A and also found to be in the Mcm1 bound gene set of B was counted and is displayed as a fraction of the total genes bound in species A that can be mapped to species B (Figure 2a). Note the lack of symmetry; comparing the Mcm1 bound gene set of species A to that of B is not the same as comparing the bound gene set of species B to that of A because of the different total number of Mcm1 bound genes in the different species. Overlap p-values were also calculated for each species pair employing the hypergeometric distribution (Figure 2b).

There is significant overlap in Mcm1 targeted genes between each pair of species ($p<10^{-3}$). As might be expected, conservation is strongest between the two more closely related species, *S. cerevisiae* and *K. lactis*, with 42% of mapped *S. cerevisiae* Mcm1 targets also bound by *K. lactis* Mcm1. However, as the lower frequency (16%) of *K. lactis* Mcm1 mapped targets bound by *S. cerevisiae* Mcm1 indicates, the *K. lactis* Mcm1 target set is much larger. Interestingly, the *C. albicans* Mcm1 target set overlaps more significantly with the *K. lactis* Mcm1 set than it does with the *S. cerevisiae* Mcm1 set, indicating that a sizeable fraction of the extra genes bound by *K. lactis* Mcm1 are shared with *C. albicans* Mcm1 and are therefore likely to have been lost as Mcm1 targets on the branch leading to *S. cerevisiae* (see next section).

For simplicity we have focused here on only those genes that can be mapped in a 1:1 fashion between species. However, similar results are obtained when genes with more

complex interspecies mappings (e.g. 2:1) are included. To rule out the possibility that our results were biased by the exact parameters chosen, we repeated the analysis with a variety of parameter choices and obtained similar results (Figure S9).

Mcm1 binding site turnover is extensive, with sizeable gain and loss rates.

In order to assess the prevalence of gain and loss of Mcm1 binding sites across the three species phylogeny we constructed a model with nine parameters: four gain rates and four loss rates corresponding to each of the four branches of the rooted tree and a single parameter representing the probability of an Mcm1 binding site at the root of the tree (Figure 2c). We take as our dataset the Mcm1 binding occurrence patterns at each of the 2766 genes that can be mapped between *S. cerevisiae*, *K. lactis*, and *C. albicans* in a 1:1:1 fashion via our ortholog mapping. There are eight such patterns, e.g. the pattern "101" for hypothetical gene X indicates an Mcm1 binding site is present upstream of gene X in *S. cerevisiae* and *C. albicans*, but not in *K. lactis*. We devised a modified maximum likelihood algorithm to infer the gain and loss rates on each branch of the three species phylogeny. A more thorough description of this procedure is given in the Supp. Text.

The results show a high degree of binding site turnover on all branches of the tree. For example, we estimate that the last common ancestor of *S. cerevisiae* and *K. lactis* had Mcm1 binding sites at 156 genes. Since divergence Mcm1 binding sites were gained at 44 genes and lost at 109 genes in the *S. cerevisiae* lineage. Likewise, Mcm1 binding sites were gained at 128 genes and lost at 38 genes in the *K. lactis* lineage. Thus, present day

S. cerevisiae and *K. lactis* have only 38 Mcm1 targeted genes in common. We do not believe this analysis is biased by any systematic failures to detect Mcm1 binding sites in our ChIP experiments – either through experimental biases or because the growth conditions chosen did not promote Mcm1 binding. In cases where Mcm1 is bound upstream of a gene in one species but not in the other two species, the Mcm1 motif is generally not present in those other two species as well (Supporting Text, "Mcm1 DNA motifs are not present at genes that are not bound"). In the sections that follow we will further dissect the changes in combinatorial regulation that give rise to the conservation and divergence summarized in Figure 2c.

There is a small conserved core of "ancestral Mcm1 bound genes".

If we consider just the subset of genes that has a 1:1:1 mapping in our ortholog table, only twelve genes (~13% of the genes bound in *S. cerevisiae*) are part of the Mcm1 circuit in all three species (Figure 3). If gene duplications are allowed, the number of genes in *S. cerevisiae* with at least one "ortholog" bound in *K. lactis* and *C. albicans* is 45 (~18% of the genes bound in *S. cerevisiae*). Presumably this conserved set of target genes reflects a conserved role played by Mcm1 in the common ancestor as well as in the three modern species.

The set of ancestral Mcm1-bound genes is clearly enriched for genes regulated by the cell-cycle (Figure 3, shaded orange) and mating-type (Figure 3, shaded blue). The latter is confirmation of results from our previous study[31] describing the conservation of membership within the **a**-specific gene regulon despite the dramatic switch from positive

regulation by MATa2 to negative regulation by MATα2. In *S. cerevisiae* the cell-cycle genes listed are regulated by the Mcm1 cofactors Fkh2/Ndd1 and Yox1. The conservation of these genes as targets of Mcm1 prompted us to inquire whether combinatorial control by Mcm1 and each of its *S. cerevisiae* cofactors was also conserved since *S. cerevisiae*, *K. lactis* and *C. albicans* diverged from a common ancestor.

Most Mcm1-cofactor interactions observed in modern *S. cerevisiae* emerged early, but their target genes have changed dramatically.

The Mcm1-cofactor regulons of *S. cerevisiae* were mapped to Mcm1-bound regions in *K. lactis* or *C. albicans*, and motif finding was performed to identify *cis*-regulatory elements controlling the orthologous regulons (details presented in the Supp. Text). The results of this analysis (Figure 4a,b) demonstrate that most known Mcm1-cofactor interactions from *S. cerevisiae* are present in *K. lactis* and *C. albicans* and are therefore likely of ancient origin. In the description that follows we first compare the *cis* regulatory motifs of the more closely related *S. cerevisiae* and *K. lactis* and then compare these to the motifs of the more divergent *C. albicans*. We note that in this paper we use the term "interaction" to refer to both demonstrated protein-protein interactions as well as those inferred from the co-occurrence in *cis* of two or more regulatory motifs. One caveat of this approach is that co-occurrence of motifs can arise from cooperative as well as competitive binding of two transcription factors. However, we think the latter is unlikely for most cases documented in this work, because the spacing of the motifs tends to be highly constrained and non-overlapping, a feature typically observed for cooperative binding with Mcm1.

In general the *cis* regulatory elements of the *K. lactis* and *S. cerevisiae* Mcm1-cofactor regulons are similar, suggesting that the corresponding Mcm1-cofactor interactions have changed little since these two lineages split. Notable exceptions are the changes seen at **a**sgs discussed previously[31] and the apparent added Fkh2 specificity flanking several of the Yox1-Mcm1 sites in *K. lactis* (Figure 4b). Although the latter is seen in at least a few genes in *S. cerevisiae*[98], this Yox1-Mcm1-Fkh2 architecture appears much more prominent in *K. lactis*.

Comparison of *K. lactis* and *S. cerevisiae* to the more divergent *C. albicans* reveals that a number of changes to *cis* regulatory motifs have occurred over longer timescales. At the Fkh2-Mcm1 regulon there is a shift in the placement of the Fkh2 site relative to the Mcm1 site by one base pair which occurs across the entire regulon. We note that species with the tighter Fkh2-Mcm1 spacing have clear orthologs to Ndd1, a protein which in *S. cerevisiae* binds the Fkh2-Mcm1 complex periodically thereby driving G2/M-specific expression[105], while those with the lengthened spacing do not. It is not known how the Fkh2-Mcm1 complex of *C. albicans* would function to drive G2/M-specific gene expression without an Ndd1 ortholog, although this altered spacing may provide a clue. It is also noteworthy that Fkh2 is related to another protein, Fkh1, derived from the yeast whole genome duplication event[62], meaning that these two genes found in *S. cerevisiae* map to a single gene in *K. lactis* and *C. albicans*. It is known that Fkh2 binds DNA cooperatively with Mcm1, but that Fkh1 does not[106]. Given the evidence for the Fkh2-Mcm1 motif in *K. lactis* and *C. albicans*, we infer that this interaction is ancestral to the

species under study and that after duplication only Fkh2 retained the ability to bind cooperatively with Mcm1.

The *cis*-regulatory motif at the MATα1-Mcm1 regulon has clearly changed as well, indicating that MATα1, despite its obvious conservation, recognizes distinct DNA motifs in different species. However, the altered MATα1 motif observed in *C. albicans* is not necessarily incompatible with the *S. cerevisiae* protein, a surmise based on previous mutagenesis studies[107]. Despite this change in motif, experimental evidence indicates that MATα1's function as an activator of αsgs is the same in *S. cerevisiae* and *C. albicans*[30].

Once it was determined that most Mcm1-cofactor pairings are conserved across the species we examined, we then determined to what extent the set of genes in their corresponding regulons was also conserved. The motif matrices for each of the Mcm1-cofactor pairs were employed to score the entire set of Mcm1 bound sequences in each species and thus to define the members of each Mcm1-cofactor regulon in each species (see the Supp. Text). We find that the number of targets in each regulon is roughly the same across the three species, but the precise set of members is not. However, within each regulon there is a small, core set of conserved genes (Figure 4c). For example, the Fkh2-Mcm1 regulon consists of roughly twenty genes in each species, but only three genes are part of the regulon in all three species. Previously we showed that for the **a**sg regulon at least, this core is conserved throughout the yeasts spanning the lineage of *S. cerevisiae* and *C. albicans*[31]. A similar promoter sequence analysis with the Mcm1-Fkh2 matrices supports a conserved core within this regulon as well (unpublished data). For

example, the promoters of BUD4 and CDC20 have strong matches to the Fkh2-Mcm1 matrix in most species within the lineage spanning *S. cerevisiae* and *C. albicans*. Thus, turnover within these regulons is not a purely stochastic process, but rather is constrained in some respects by purifying selection.

Lineage specific gain and loss of Mcm1-cofactor interactions is also evident.

As summarized in Figure 2c, this study revealed many specific instances of gains and losses of Mcm1 regulation across the ascomycete lineage. The large number of changes seen at the global level, however, can not be fully accounted for by binding site turnover within the ancestral Mcm1-cofactor regulons alone (Figure 4c). In the following section we highlight three examples of large-scale rewiring events, chosen for their particular clarity and their relevance to well-developed systems.

Mcm1 and Rap1 binding sites at ribosomal genes in K. lactis. There are 378 genes bound by Mcm1 in *K. lactis*, but not in *S. cerevisiae* or *C. albicans*. 59 of these are annotated as constituents of the cytosolic ribosome in *S. cerevisiae* ($p<10^{-45}$). In total 70 of the 101 genes annotated as cytosolic ribosomal genes are bound by Mcm1 in *K. lactis*. A closer examination reveals that the 70 ribosomal genes bound by Mcm1 encode for structural constituents of the small or large subunits, whereas the other 31 genes tend to encode for translational accessory proteins such as the acetyltransferase Nat5 and the mRNA decapping factor Pat1.

Since, of the three species we studied, only *K. lactis* has Mcm1 sites at its ribosomal genes, we examined a broader range of fungi to determine with greater resolution whether this pattern likely results from gains in the *K. lactis* lineage or losses in the *S. cerevisiae* and *C. albicans* lineages. To do so we mapped the 162 cytosolic ribosomal genes (GO:0005830) from *S. cerevisiae* to 31 other fully sequenced fungal genomes and then performed *de novo* motif finding on the promoters of these genes (500 bp upstream of the translational start) with MEME[54].

To our surprise, motifs resembling that of Mcm1 were found at ribosomal genes in several species, *C. glabrata, K. lactis* and *Y. lipolytica*, which do not cluster phylogentically. Furthermore, a motif resembling Mcm1, plus an unknown cofactor, was found in the branch spanning *A. nidulans to H. capsulatum* (Figure 5a). To verify that presence of the Mcm1-like motifs at ribosomal genes was limited to just *C. glabrata, K. lactis, Y. lipolytica* and the *A. nidulans* branch, we scored the ribosomal gene promoters (1 kb upstream of the translational start) of each species with the Mcm1 motif matrices (Figure 5b). Indeed, evidence for Mcm1-like motifs at ribosomal genes is limited to just the aforementioned species.

Formally, we can not rule out the possibility that Mcm1 may bind indirectly to ribosomal gene promoters in species which we have not performed ChIP. However the changes in *cis* acting sequence are striking and imply, at the very least, a change in mechanism. We also cannot formally rule out the possibility that a smaller than statistically significant fraction of the ribosomal genes is regulated by Mcm1 in some other species. However,

given that the ribosome plays such an essential role in the cell and that even small reductions in expression of a single ribosomal gene relative to the others can lead to substantial slowing of growth rate[108], the latter seems unlikely as well.

If we suppose that loss of established Mcm1 regulation of the ribosomal genes is just as costly as gaining Mcm1 regulation of ribosomal genes, then the evolution of Mcm1 at ribosomal genes is most parsimoniously explained by four independent gains. The next most parsimonious scenario is three gains and two losses. If we posit a single gain of regulation, then at least five losses must occur as well.

Our discovery of Mcm1 at the ribosomal genes in *K. lactis* (but absent from the orthologous genes of *S. cerevisiae* and *C. albican*s) prompted us to search for a possible cofactor. The same MEME search that identified the Mcm1 motif at the ribosomal gene promoters of *K. lactis* also identified a *cis* regulatory motif that is similar in sequence to that recognized by Rap1 in *S. cerevisiae*[109] (Figure 5c). Indeed, it was shown previously that the Rap1-like motif is present at ribosomal gene promoters in *S. cerevisiae*, *K. lactis* and closely related yeasts and thus it was inferred that this motif was present at ribosomal genes in the last common ancestor of *S. cerevisiae* and *K. lactis*[26]. By searching the cytosolic ribosomal gene promoters of *K. lactis* for the presence of maximal scoring Mcm1 and Rap1 motifs ($\log_{10}$-odds scores > 2.0), we find that the newly discovered Mcm1 sites are semi-strictly positioned at a median 54 bp downstream (with respect to the ORF) of Rap1 sites (Figure 5d). Although the distance constraint is not as strict as those typically seen for other Mcm1 cofactors (Figure 4b), we believe it is likely that

Rap1 is a newly discovered Mcm1 cofactor in *K. lactis*. To summarize, it seems likely that in the *K. lactis* lineage Mcm1 binding sites were gained at 70 ribosomal genes and that a combinatorial interaction between Mcm1 and a pre-existing ribosomal regulator, Rap1, was formed.

Mcm1, Arg80 and Arg81 binding sites at arginine metabolic genes. One of the more prominent aspects of the loss-gain diagram of Figure 2c is the relatively higher rate of loss on the branch leading to *S. cerevisiae*. This finding is consistent with the results of the pairwise comparison, which suggested the existence of a set of genes conserved between *K. lactis* and *C. albicans*, but lost on the branch to *S. cerevisiae*. The set of genes with an Mcm1 binding site in *K. lactis* and *C. albicans*, but lacking sites in *S. cerevisiae* totals 58 (in *S. cerevisiae*) and is enriched for arginine metabolic genes (GO:0006525; N=5; $p<10^{-6}$).

Mcm1 has a duplicate in *S. cerevisiae*, Arg80, which arose after the divergence of *K. lactis* and *S. cerevisiae*. Our observations are most consistent with a model whereby Mcm1's ancestral role, collaborating with the Mcm1 cofactor Arg81 in arginine metabolism was, at least in part, handed off to its duplicate Arg80. Although previous *in vitro* work demonstrated that Mcm1 and Arg80 form heterodimers at operator sequences found upstream of arginine metabolic genes in *S. cerevisiae*[110], our Mcm1 ChIPs and the Mcm1 and Arg80 ChIPs performed by others[111] suggest that *in vivo* these dimers might more typically consist of two molecules of Arg80. Based on our identification of Mcm1 binding at arginine metabolic genes in *K. lactis* and *C. albicans* (Figure 4b), Mcm1's role

interacting with Arg81 at arginine genes is inferred to be ancient, having evolved prior to the divergence of *S. cerevisiae* and *C. albicans*. The timing of the handoff to Arg80 is coincident with not only the whole genome duplication, but also with the switch from a putatively hybrid (positive and negative) mode of **a**sg regulation by MATa2 and MATα2 to a purely negative mode by MATα2, and with roughly 50% of all substitutions in the DNA binding domain of Mcm1 (see alignment in Figure 6). It is plausible that this handoff of some arginine regulon function to an Mcm1 duplicate "freed up" the surface of Mcm1, allowing for the strengthening of an interaction between Mcm1 and MATα2[31].

Mcm1 and Wor1 binding sites at white-opaque genes in the C. albicans lineage. As mentioned previously, the Mcm1 bound sequences of *C. albicans* contain a second "non-canonical" *cis*-regulatory motif (Figure 1) that strongly correlates with Mcm1 occupancy at roughly 127 genes that lack a strong match to the canonical Mcm1 motif ($\log_{10}$-odds non-canonical motif score > 4.0 and $\log_{10}$-odds canonical motif score < 3.0). To rule out possible cross-hybridization of our Mcm1 antibody to another DNA-binding protein, we repeated the ChIP of Mcm1 in *C. albicans* (1) in the same strain with an antibody raised against a peptide from the N-terminus of Mcm1 (rather than the C-terminus as before) and (2) in a myc-tagged Mcm1 strain[112] using an antibody to the myc-epitope. Both ChIPs were hybridized to *C. albicans* tiling arrays (normalized to whole cell extract DNA), and both results validate the enrichment of Mcm1 seen at the non-canonical motif (unpublished data). Furthermore, at the promoters of these genes the non-canonical motif tends to be centered with respect to the peak of Mcm1 ChIP enrichment (unpublished data), suggesting either direct binding of Mcm1 to this motif or tight interaction of Mcm1

with another transcriptional regulator that recognizes this motif. The non-canonical motif is absent from the Mcm1 bound regions of *S. cerevisiae* and *K. lactis*, and the non-canonical Mcm1 bound genes of *C. albicans* are generally not bound by Mcm1 in either *S. cerevisiae* and *K. lactis*.

Among the 110 Mcm1-bound *C. albicans* genes with very strong non-canonical motif scores ($\log_{10}$-odds > 4.5) are several genes annotated with functions in cell adhesion (N=9; GO:0007155; $p<10^{-5}$), biofilm formation (N=7; GO:0042710; $p<10^{-5}$) and regulation of white-opaque switching (Wor1/orf19.4884, Efg1/orf19.610 and Wor2/orf19.5992)[113, 114]. These three processes are important for *C. albicans* to interact with its mammalian host.

To determine when Mcm1 regulation at the non-canonical binding site arose, we mapped the 110 Mcm1-bound genes with very strong non-canonical motif scores to orthologs in each of the other 31 species and scored the promoters of these ORFs (2 kb upstream of the translational start) for presence of the non-canonical Mcm1 binding motif (Figure 7d). The presence of the non-canonical Mcm1 motif at these genes is clearly limited to *C. albicans* and *C. dubliniensis* (a very closely related human pathogen) suggesting that either the non-canonical regulatory motif arose just prior to the *C. albicans—C. dubliniensis* split or that it evolved earlier and has just recently moved to this set of genes. That the non-canonical motif was not seen at the *S. cerevisiae* and *K. lactis* Mcm1-bound genes increases our confidence that the gain of this non-canonical regulatory motif was very recent. By way of comparison, when we mapped the genes

bound by Mcm1 at the canonical motif in *C. albicans* to the other species, one sees clear evidence for the canonical Mcm1 motif in species of the *D. hansenii* branch and (with somewhat lowered confidence) in species as far diverged as *S. bayanus* and *K. lactis*. This observation suggests that the presence of a *cis* regulatory element in only two very closely related species is unusual, and thus further increases our confidence that the non-canonical motif is recently evolved.

The role of the non-canonical Mcm1 binding site in *C. albicans* white-opaque switching bears further scrutiny, as the regulatory circuit behind this epigenetic phenomenon has been studied intensively. Briefly, *C. albicans* forms two distinctive types of cells, white and opaque, which differ in their appearance[115] (Figure 7a-c), the genes they express[30], their mating behavior[77] and interaction with host sub-environments[116, 117]. Both states are heritably maintained for many generations and switching between them occurs at low frequency (~$1/10^4$ cell generations). A master regulator of white-opaque switching, Wor1, has been identified[118-120] and shown to bind many white- and opaque-specific genes[113].

Comparison of the Mcm1 and Wor1 ChIPs in opaque cells reveals a striking overlap of Mcm1 and Wor1 binding in the upstream regions of all known critical regulators of white-opaque switching, including WOR1 itself (Figure 7e). Genome-wide, 36 of the 110 genes with non-canonical Mcm1 binding sites are also bound by Wor1 (33%; hypergeometric $p<10^{-24}$), suggesting an interaction between the two proteins. These results indicate the intimate involvement of Mcm1 and the non-canonical Mcm1 motif in

white-opaque switching and raise the possibility that the evolution of this motif played an important role in the acquisition of white-opaque switching and other interactions with the host by the *C. albicans* lineage.

DISCUSSION

In this work we have tracked the evolution of combinatorial gene regulation by the highly conserved transcriptional regulator Mcm1 and each of its known cofactors across the ascomycete fungal lineage. Our analysis shows that the genes regulated by Mcm1 have changed considerably over the evolutionary timescales represented by this lineage; our results reveal many more differences than similarities in the Mcm1 circuitry. Regulation by Mcm1 is more pervasive in *K. lactis* and *C. albicans*, where 12% of all genes are bound, than in *S. cerevisiae* where 4% of genes are bound. The fraction of genes shared as targets between all three species is very low (13-18%), and we have demonstrated that this is due to both substantial gain and loss of Mcm1 binding sites along each branch of this phylogeny (Figure 2b). The extensive amount of gain and loss observed is consistent with recent studies in mammals[29] and closely related yeasts[28] and suggests three possibilities: (1) there is a richness of selective advantages offered in the dynamic rewiring of gene regulatory networks, (2) there are a large number of neutral alternatives to gene regulation by Mcm1, or (3) selection on gene expression is weak. The latter possibility seems at odds with other observations such as the large fraction of genes devoted to transcriptional regulation in *S. cerevisiae* (~3%), the greater than expected number of transcriptional regulators retained after the whole genome duplication (~6% vs. ~3%), and the considerable conservation found in many *S. cerevisiae* promoters[121, 122]. Additionally, the fact that many of the Mcm1 sites are enriched at functionally related genes and often found tandem with cofactor motifs argues strongly against the hypothesis that a large number of these sites are fortuitous and non-functional. Gauging the relative

contributions of selection versus neutral drift on the gene regulatory networks will be an exciting challenge for future research[123].

Despite the highly dynamic nature of evolution of Mcm1 regulation, we find evidence that most Mcm1-cofactor interactions characterized in *S. cerevisiae* are also present in *K. lactis* and *C. albicans* (Figure 4b). Although the Mcm1-cofactor pairings are conserved, the set of genes that each regulates has diverged considerably across species. Nonetheless, each Mcm1-cofactor pair targets a small core of genes conserved as part of the regulon. These regulon cores are enriched for genes functioning in the cell-cycle and mating. Thus it would seem that Mcm1's role in these processes evolved prior to the split of the species we have chosen to study. Nevertheless, even at these conserved regulons there are many species-specific differences. For example, across an entire regulon the spacing between Fkh2 and Mcm1 binding sites has changed in *S. cerevisiae* and *K. lactis* relative to *C. albicans*, as have the DNA recognition sequences of MATα1. This latter observation is particularly interesting because it suggests that the specificity of MATα1 has evolved without an accompanying gene duplication.

In addition to the conservation of Mcm1-cofactor interactions associated with cell-cycle and mating, we see the evolution of new Mcm1-cofactor regulons. For example, Mcm1 binding sites are gained at the majority of ribosomal genes in *K. lactis* in close proximity to binding motifs for another transcription factor, Rap1 (Figure 5c,d). The evolution of ribosomal gene regulation has been studied previously[26], but a role for Mcm1 was not discussed. Our new results support the idea, first proposed by Tanay et al.[26], that while

the protein sequence of this critical macromolecular machine has remained nearly constant, its regulation has undergone substantial diversification in yeasts. What is perhaps most surprising is our finding that the set of species which contain Mcm1 binding motifs upstream of ribosomal genes (Figure 5a,b; *C. glabrata, K. lactis, Y. lipolytica* and the *A. nidulans* lineage) do not cluster phylogenetically. From this we inferred that Mcm1 binding at ribosomal genes likely evolved on four separate occasions. If further genome sequencing continues to support this result, this will serve as the largest instance of convergent regulatory evolution yet reported. The relatively sudden appearance of Mcm1 binding sites in close proximity to Rap1 sites at roughly 70 ribosomal genes in *K. lactis* raises another important question. Can the commonly accepted mutational processes, such as point mutation and recombination, support this scale of concomitant changes – or must some alternative mechanism for moving promoters around the genome be invoked[124, 125]? One can argue that, without a redundant mechanism in place, loss or gain of Mcm1 regulation of even a single gene means potentially under-expressing one component of a macromolecular complex that is very sensitive to such changes in expression[108]. With further sequencing and characterization of Mcm1's functional role at the ribosomal genes, it may become clear how such a massive regulatory change can take place at a set of genes encoding such highly conserved, tightly regulated and essential proteins.

In *C. albicans* we identified the presence of Mcm1 at a non-canonical motif upstream of roughly 110 genes. The non-canonical motif differs significantly from the canonical Mcm1 motif (Figure 1), although in both cases GC-rich regions flank an AT-rich core.

To our knowledge no MADS-box domain has ever been shown to bind a sequence this far diverged from the canonical Mcm1 motif. Even so, we find that non-canonical motifs tend to be centered with respect to peaks of ChIP-Chip enrichment and thus conclude that Mcm1 either binds this motif directly with some unknown cofactor or some unknown transcriptional regulator recognizes this motif and interacts strongly with Mcm1. The set of genes at which Mcm1 binds the non-canonical motif is enriched for processes such as adhesion and contains three of four known regulators of the white-opaque phenotypic switch[113]. The white-opaque switch is of considerable interest because the white and opaque states are heritable and because the two states are thought to allow adaptation to different niches within a human host[116, 117]. In this vein the evolution of regulation associated with the switch deserves special attention too, as the changes seen here represent the first gene regulatory changes to be associated with a heritable biological process and one of only a few instances implicated to play an adaptive role in fungal biology[126]. The results of our comparative analysis of 32 yeast species demonstrate that Mcm1 binding at the non-canonical motif is found only in two very closely related species, *C. albicans* and *C. dubliniensis*, and thus likely arose only very recently (Figure 7d). Moreover white-opaque switching has been described only in these two species[127], which are both pathogens of humans. Thus, the evidence so far suggests that the white-opaque switch may have arisen just prior to the divergence of *C. albicans* and *C. dubliniensis* and that the emergence of the non-canonical Mcm1 motif at white-opaque regulators was crucial to this development. Alternatively, the white-opaque switch may have arisen earlier and the addition of Mcm1 regulation may have refined it in some way, affecting heritability, for example.

The picture that emerges from this study is one of massive transcriptional rewiring in species that span approximately the same range of divergence as human, fish and sea squirt[100, 101]. Mcm1 regulates hundreds of genes in *S. cerevisiae*, *K. lactis* and *C. albicans*, but less that 20% of Mcm1-target gene connections are preserved in all three species. The differences arise from target genes moving in and out of ancient Mcm1-cofactor regulons, but also from the formation of new Mcm1-cofactor interactions and the loss of ancient ones. Taken together with our previous work[31], we have now provided evidence for the gain of three interactions: Mcm1 with MATα2, Mcm1 with Rap1 and Mcm1 with Wor1. We have also described loss of an interaction between Mcm1 and MATa2 and the loss of an interaction between Mcm1 and Arg81 that was preserved in an Mcm1 duplicate. In attempting to judge the relative contributions of combinatorial control *per se* to the evolution of transcriptional circuits, we acknowledge that the ideal "control" datasets do not exist. For example, data collected from a large non-combinatorial circuit (should one even exist) over several species would allow an objective assessment of the special contribution of combinatorial control to circuit evolution. Nonetheless, our results provide experimental and informatic support for the idea that combinatorial networks are highly evolvable[4, 8, 128, 129], and perhaps more importantly, document specific mechanisms by which one large combinatorial circuit has evolved.

METHODS

Detailed methods can be found in the Supporting Text.

ACKNOWLEDGMENTS

We thank C. Baker, V. Chubukov, O. Homann, S. Liang, H. Madhani, A. Tsong, R. Zordan and members of the Johnson and Li labs for helpful discussions. We also thank J. Felsenstein for suggestions pertaining to the inference of binding site gains/losses and S. Rupp and G. Sprague for contributing reagents. We are grateful to the Broad Institute, the Sanger Center, Génolevures, the Joint Genome Institute, Cécile Fairhead, K. Wolfe and K. Byrne for making genome sequence data available.

FIGURES

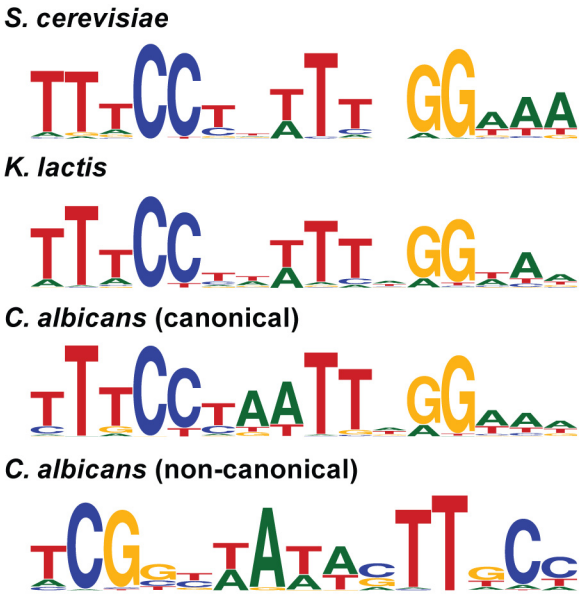


Figure 1. Mcm1 *cis* regulatory motifs in three species.

The four *cis* regulatory motifs identified by searching a high confidence set of Mcm1 bound regions in the indicated species. In *C. albicans* a non-canonical motif was found in addition to the canonical Mcm1 motif.

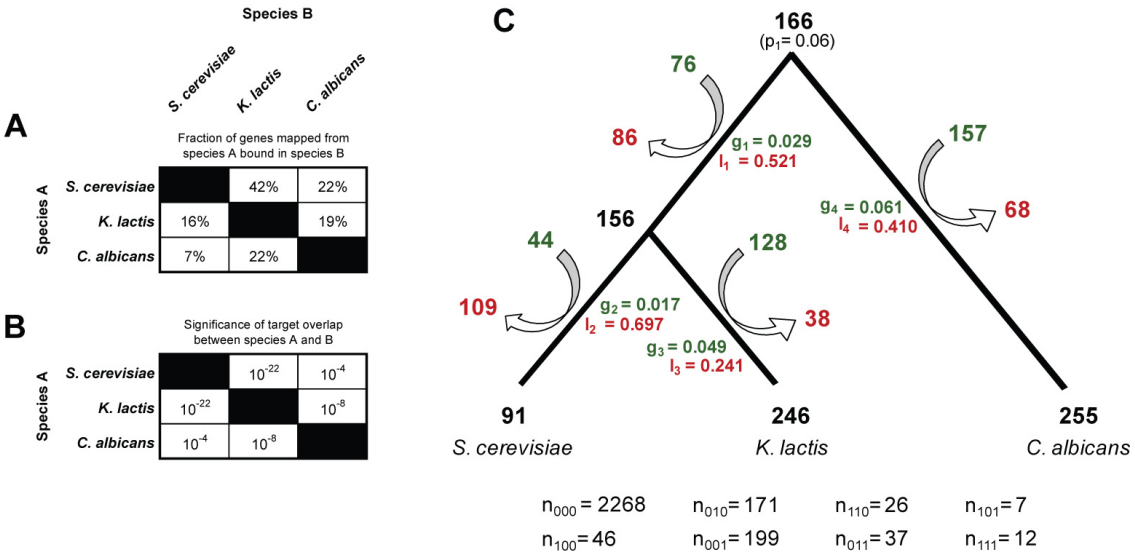


Figure 2. Comparison of Mcm1-bound target genes in three species.

(a-b) Mcm1 targeted gene sets are compared in a pairwise fashion between species.

(a) The number of genes mapped from species A and also found to be in the Mcm1 bound gene set of species B, as a fraction of the total genes bound in species A that can be mapped to species B. (b) The significance (hypergeometric p-value) of each pairwise overlap.

(c) The inference of gain and loss rates (green and red respectively) along each branch of the rooted three species phylogeny. The inferred number of genes added and removed from the Mcm1 regulon is listed at the top and bottom of an arrow flanking each branch. The total counts for each of the eight possible occurrence patterns used as input to the inference algorithm are presented below the tree.

Gene Name	Cell-cycle regulated?	Mcm1 Cofactor	Description
Bud4	G2/M	Fkh2	Involved in bud-site selection and required for axial budding pattern
Cdc20	G2/M	Fkh2	Cell-cycle regulated activator of anaphase-promoting complex/cyclosome
Chs2	G2/M	Fkh2	Required for the synthesis of chitin in the primary septum during cytokinesis
Cdc6	M/G1	Yox1	Component of the pre-replicative complex required for DNA replication
Ase1	G2/M	Yox1	Microtubule-associated protein required for spindle elongation
Cln3	G2/M	?	G1 cyclin, activates Cdc28p kinase to promote the G1 to S phase transition
Cdc5	G2/M	?	Polo-like kinase found at bud neck, nucleus and SPBs; functions in cytokinesis
Ste2	G2/M	α2	Receptor for α-factor pheromone required for mating between haploid cells
Ste6	G2/M	α2	ATP-binding cassette (ABC) transporter required for the export of a-factor
Bar1	No	α2	Cleaves α factor allowing recovery from α-factor-induced cell cycle arrest
Asg7	No	α2	Regulates pheromone signaling from Ste4p and its relocalization within the cell
Yhl008c	No	?	Localizes to the vacuole

Figure 3. The ancestral Mcm1 bound genes.

These twelve genes are targets of Mcm1 in all three species. For each gene the cell-cycle phase of increased expression[130] (if applicable), the relevant Mcm1 cofactor (if known) and a brief functional annotation is listed. Cell-cycle and mating-type regulated genes are shaded orange and blue, respectively.

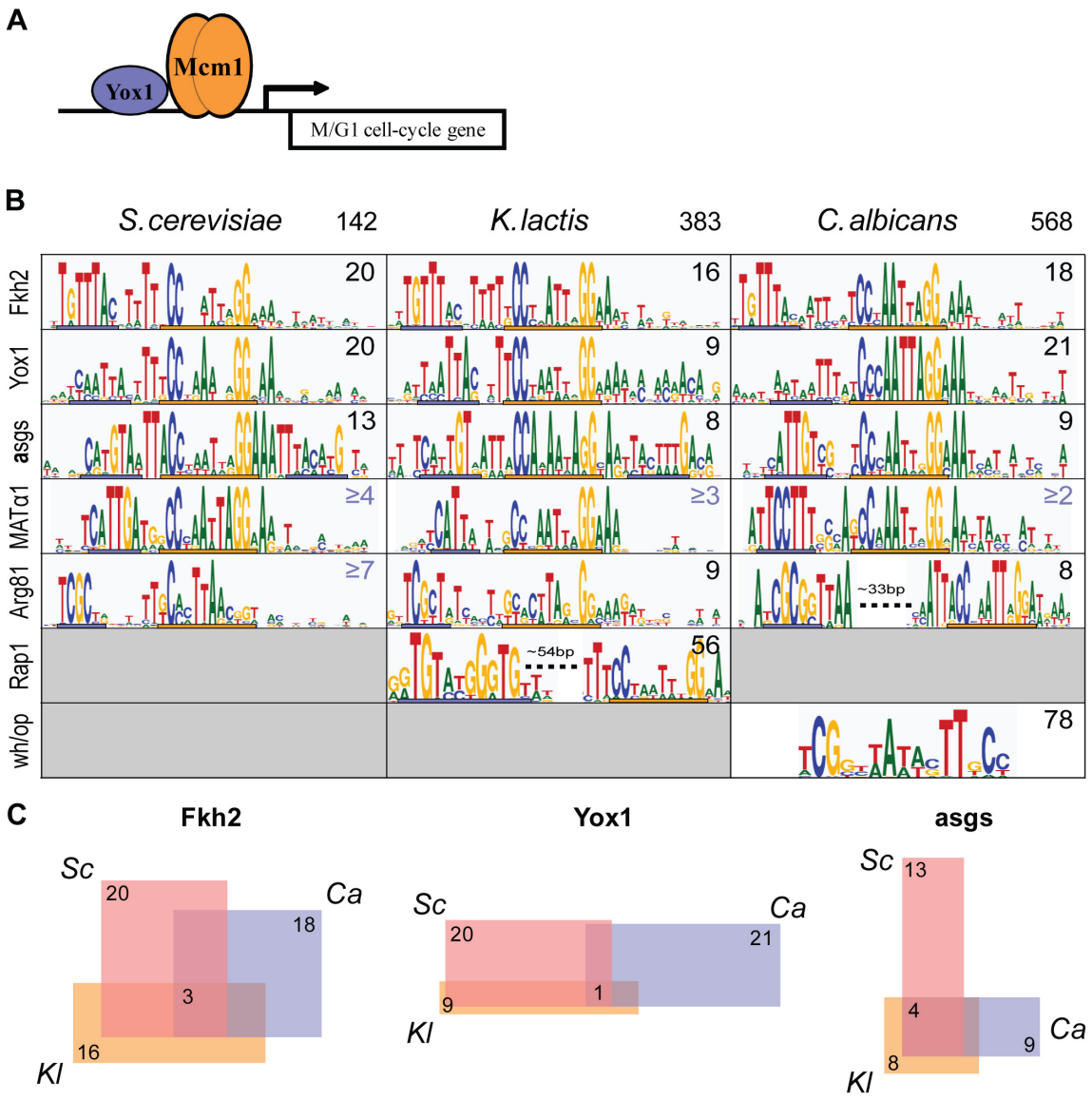


Figure 4. Comparison of Mcm1-cofactor regulons across species.

(a) An example schematic of the Mcm1 homodimer and its cofactor, Yox1, binding in close proximity upstream of an M/G1-specific cell-cycle gene.

(b) Mcm1 associated c*is* regulatory motifs discovered across the three species in this work. Each row of the table specifies an Mcm1-cofactor regulon and each column a species. The total number of Mcm1 bound regions in each species is listed in the header

row. The number of Mcm1 bound regions assigned to each Mcm1-cofactor regulon in each species is listed in the upper right corner of each cell of the table; numbers colored black are based on Mcm1 ChIP data, while those in blue are not and are therefore more tentative. Mcm1 binds or is predicted to bind the consensus sequence denoted by the orange bar in each cell. The known or predicted cofactor motif is denoted by a blue bar in each cell. Motif graphics were generated with WebLogo [131]

(c) The three-way overlap of target genes in the Fkh2-Mcm1, Yox1-Mcm1 and **a**sg (Mcm1-a2 or Mcm1-α2) regulons in the three species (Sc = *S. cerevisiae*, Kl = *K. lactis* and Ca = *C. albicans*).

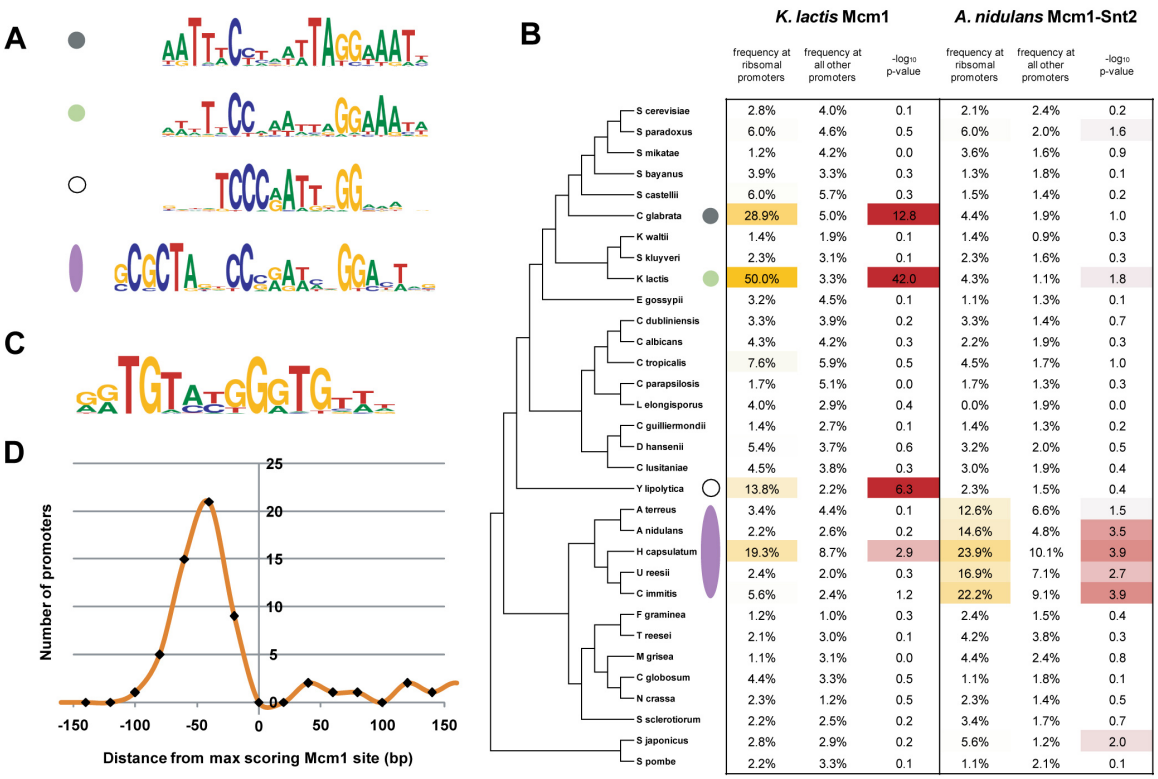


Figure 5. Evolution of Mcm1 binding sites at ribosomal genes in the ascomycete lineage.

(a-b) Convergent evolution of Mcm1 motifs at ribosomal genes.

(a) Four Mcm1-like *cis* regulatory motifs discovered in a MEME search of the ribosomal gene promoters of 32 fully-sequenced ascomycete genomes. The motifs were discovered in the species indicated by the colored circles and oval in b. The Mcm1-like motif of the *A. nidulans* branch has a tandem cofactor motif which is nearly identical to that derived from Snt2 ChIP-Chip experiments in *S. cerevisiae*[111]; we therefore predict that the Snt2 orthologs of the *A. nidulans* lineage are the Mcm1 cofactors at the ribosomal genes of this lineage.

(b) The Mcm1 motifs from *K. lactis* (green circle) and the *A. nidulans* lineage (lavender oval) were employed to score ribosomal promoters across the ascomycete lineage and thus to verify that presence of the Mcm1 motifs is limited to the four lineages in which Mcm1-like motifs were found *de novo* by MEME. The significance of motif enrichment at the ribosomal promoters of each species was determined by comparison to genome-wide background frequencies of occurrence using the binomial distribution. See the Supp. Text for description of the ascomycete phylogeny reconstruction[31, 61].

(c) An additional motif similar to that recognized by Rap1 in *S. cerevisiae* was discovered in the MEME search of *K. lactis* ribosomal promoters.

(d) In *K. lactis*, the positioning of Rap1-like motif instances is constrained relative to Mcm1 motif instances.

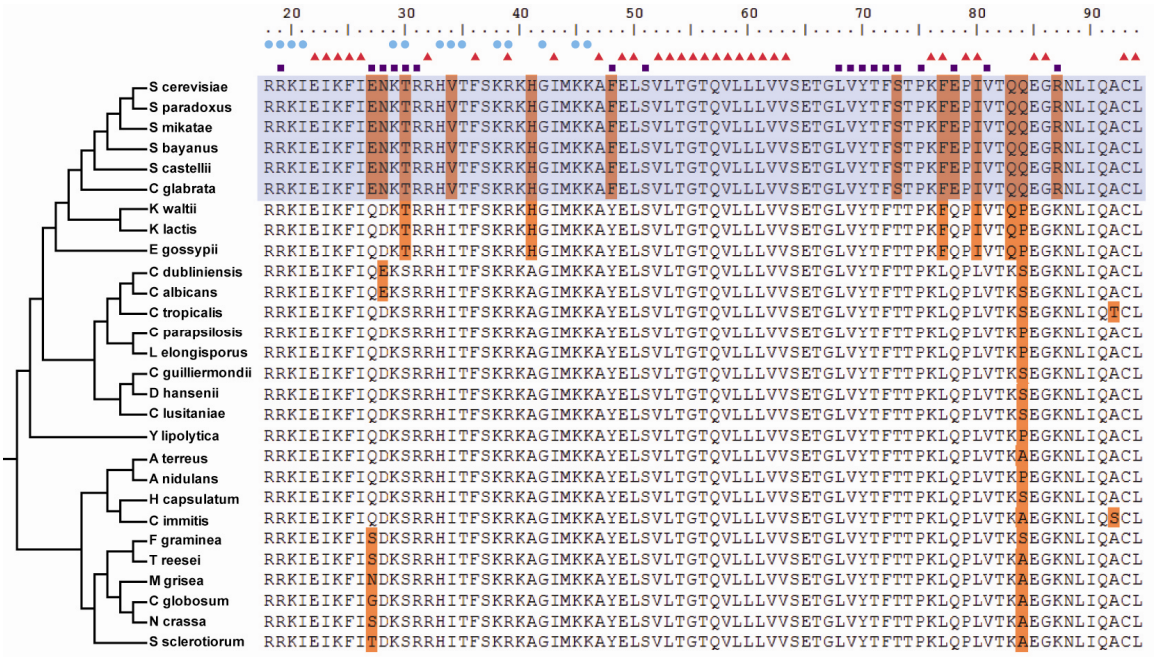


Figure 6. Substitutions within the MADS box domain of Mcm1.

There are a few substitutions to the MADS box domain of Mcm1 (orange) against a background of strong conservation (white) within the hemiascomycete and euascomycete lineages. The shaded box indicates Mcm1 orthologs from species which also have an Mcm1 duplicate (named Arg80 in *S. cerevisiae*). Mcm1 residues forming contacts with α2, Mcm1 or DNA in the crystal structure of the α2-Mcm1-DNA ternary complex (PDB ID: 1mnm) are indicated above the alignment with squares, triangles and circles, respectively. Note the strong correlation between those species having substitutions at the α2 interacting residues, those species with an Mcm1 duplicate and those species thought to be employing a purely negative mode of **a**sg regulation by Mcm1 and α2[31].

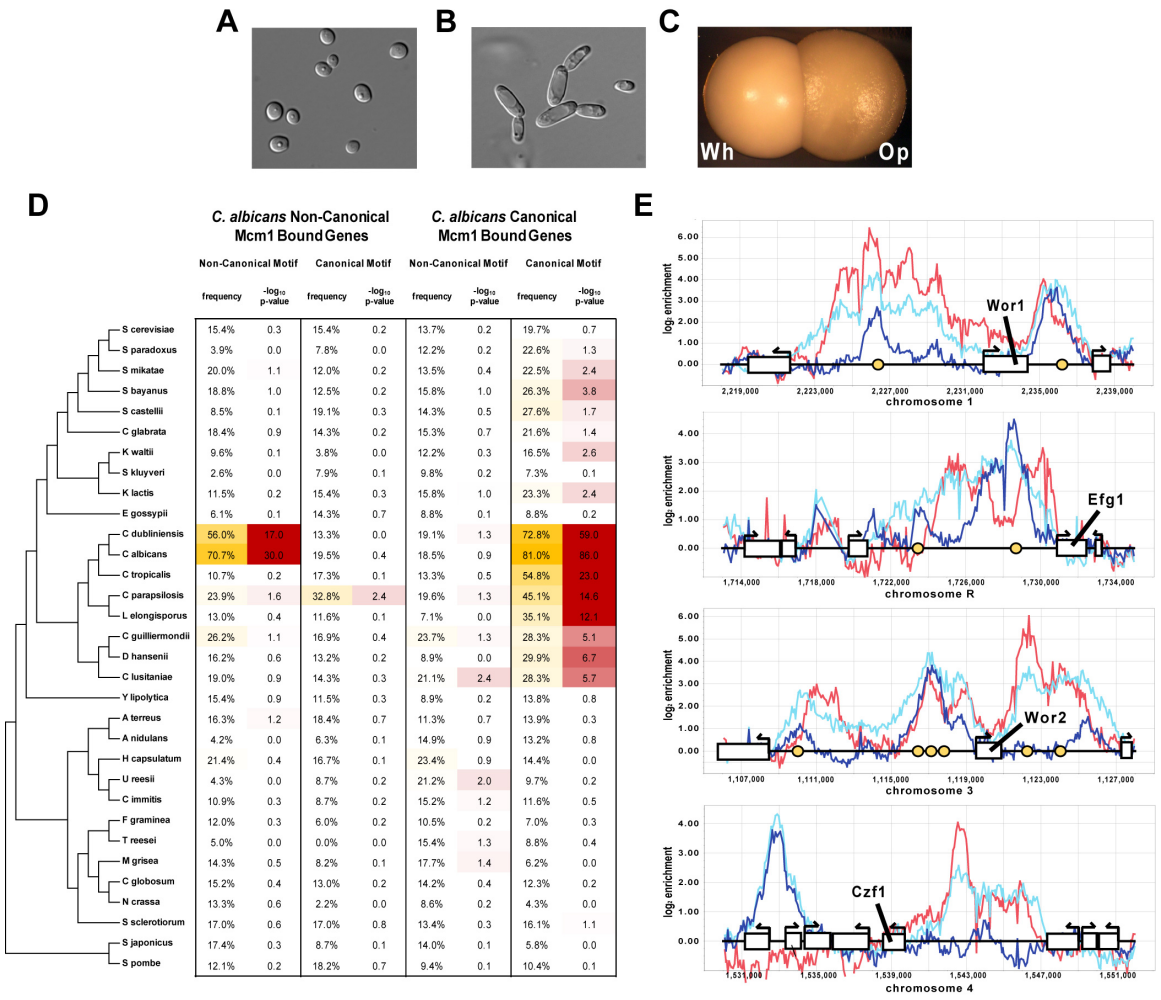


Figure 7. Recent evolution of non-canonical Mcm1 binding sites at white-opaque genes.

(a-c) *C. albicans* cell types. (a) White cells. (b) Opaque cells. (c) A white colony (Wh) and an opaque colony (Op).

(d) The non-canonical and canonical Mcm1 motif matrices of *C. albicans* (Figure 1) were employed to score promoters for two sets of genes (genes where Mcm1 is found at the non-canonical motif in *C. albicans* and genes where Mcm1 is found at the canonical motif in *C. albicans*) across the ascomycete lineage. The significance of motif

enrichment at the two mapped gene sets of each species was determined by comparison to genome-wide background frequencies of occurrence using the binomial distribution.
(e) ChIP-Chip profiles for Mcm1 and Wor1 in regions flanking four key regulators of the white-opaque switch[113]. Blue, teal and red lines indicate the Mcm1 ChIP of white cells, Mcm1 ChIP of opaque cells and Wor1 ChIP of opaque cells, respectively. Yellow circles indicate a non-canonical Mcm1 motif.

SUPPLEMENTARY METHODS

Tiling array design

We designed three custom ChIP-Chip arrays by tiling 181,900 probes of length 60 bp across the:

(1) 12.1 Mb of sequence included in the GenBank release of the *S. cerevisiae* genome dated 5/12/2006 (downloaded 5/2006)

(2) 10.7 Mb of sequence included in the GenBank release of the *K. lactis* genome dated 7/30/2004 (downloaded 5/2006)

(3) 14.3 Mb of sequence included in what, at the time of design, was the most recent build of the *C. albicans* genome (Assembly 20 from Andre Nantel 4/2006).

Rather than choose probes spaced uniformly, an effort was made to optimize four characteristics of the oligo set (uniqueness, GC content, self-annealing and sequence complexity), while still maintaining probe spacing that is close to uniform. We used a previously developed algorithm, ArrayOligoSelector[132], to score all possible probes for each of these four characteristics. These scores, as well as a penalty for a too-long or too-short distance to the neighboring probe, were integrated into a single score via a weighting scheme. Due to the lack of studies which systematically explore the importance of uniform spacing and each of the four probe characteristics on ChIP-chip quality, we chose weights based largely on intelligent guessing. Stronger weights were given to probe spacing, GC content and uniqueness than to self-annealing and sequence

complexity. A Monte Carlo optimization was employed to search for the highest scoring probe set.

The highest scoring probe sets found for *S. cerevisiae, K. lactis* and *C. albicans* had median probe spacing (measured from oligo start to oligo start) of 66 bp, 59 bp, and 79 bp respectively (Figure S1). The uniqueness of the probe set chosen for each genome (pink curve) was significantly higher than that for all possible probes in the genome (red curve). A higher uniqueness score indicates a more unique probe that is less likely to be affected by cross-hybridization. For *C. albicans*, whereas only 27% of all possible probes are completely unique (uniqueness score of 0.0), 40% of the probes in our probe set are completely unique. Note that the tail of this distribution (uniqueness score < 35) has been truncated in the graph. Finally, the GC content of the chosen oligos (pink curve) was much improved over the genome as a whole (red curve). While the average was kept similar to that of the whole genome, the variance was considerably reduced.

Mcm1 ChIP-Chip experiments

*Strains and Media***.** The following strains were used in our ChIP-Chips of the three species:

Species	**Genotype**	**Strain Id**
S. cerevisiae	S288c *MATa* prototroph	yDG765
K. lactis	*MATa lysA1 trp1 leu2 metA1 uraA1*	SAY45[133]
C. albicans	SC5314*: ura3::imm434/ ura3::imm434 iro1::imm434/ iro1::imm434 mtlα1Δ::hisG-URA3-hisG mtlα2Δ::hisG*	RRY8 (white isolate) yDG914 (opaque isolate of RRY8)

C. albicans	ade2::hisG/ade2::hisG ura3::imm434/ura3::imm434 ENO1/eno1::ENO1-tetR-ScHAP4AD-3xHA-ADE2 pTR(97t)-CaMCM1-Myc-URA3/camcm1::FRT	MRcan42[112]

For growth in YEPD, cells were grown at 30°C overnight to an OD_{600} of 0.4. For growth in α-factor, cells were treated as needed for efficient pheromone response in the three species as follows. The *S. cerevisiae* strain *(yDG765)* was grown in YEPD at 30°C overnight to an OD600 of 0.4, synthetic α-factor (Sigma-Aldrich) was added to a final concentration of 50nM (in water), and treated cells were incubated for 30min, shaking. For *K. lactis (SAY45)*, cells were grown overnight at 30°C in SCD medium supplemented with 500µg/ml leucine to an OD600 of 0.9. Cells were then pelleted, washed twice in sterile water, resuspended in 200ml SCD medium lacking phosphate and supplemented with 500µg/ml leucine at an OD600 of 0.25, and incubated 6hr at 30°C. 165µl 13-mer α-factor (WSWITLRPGQPIF; Genemed Synthesis, So. San Francisco, Ca; 10mg/ml in 10% DMSO) was added and cells were grown for 4hr at 30°C. For *C. albicans (yDG914)*, five opaque colonies were taken from a synthetic complete with 2% dextrose and 100µg/ml uridine (SCD + Urd) plate and grown in YEPD for ≈16h at 25°C to an OD_{600} of 0.4. Cells were pelleted, washed twice in sterile water, resuspended in 200ml SpiderM medium[134] at an OD600 of 0.4. Cells were then treated with 13mer α-factor[53] at a final concentration of 10µg/ml (from 10mg/ml stock in 10% DMSO), and incubated 4hr at 25°C.

Chromatin Immunoprecipitation. For chromatin immunoprecipitation, formaldehyde (37%) was added to a 1% final concentration. Treated cultures were shaken to mix and

incubated for 15min at room temperature. 2.5M glycine was added to a final concentration of 125mM, and treated cultures were mixed and incubated 5min at room temperature. Cells were pelleted at 3,000 x g for 5min at 4°C and washed twice with 100ml 4°C TBS (20mM TrisHCl, pH7.6/150mM NaCl). Spheroplasting and ChIP were carried out as previously described[113], with modifications. Cell pellets were resuspended in 20 ml Buffer Z + β-ME (1 M sorbitol, 50 mM Tris-Cl [pH 7.4], 10 mM β-ME) and cells were vortexed. *S. cerevisiae*, *K. lactis*, or *C. albicans* cell suspensions were lysed using 500µl of zymolyase (5mg/ml in Buffer Z) or 2µl or 20µl of lyticase (Sigma, MO, United States) solution (2 mg/ml in Buffer Z), respectively. Cell suspensions were incubated for 30min (*S. cerevisiae* and *K. lactis*) or 15min (*C. albicans*) at 30°C. Spheroplasted cells were then spun at 3,000 × g, 10 min, at 4°C and resuspended in 500µl 4°C lysis buffer (50 mM HEPES-KOH, pH 7.5,140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate) with protease inhibitors. All subsequent ChIP and wash steps were done at RT with 4°C buffers. DNA was sheared by sonication 10 times for 10 seconds at power setting 2 on a Branson 450 microtip sonicator (Danbury, CT), incubating on ice for 2 minutes between sonication pulses. Extracts were clarified by centrifugation. 50µl of extract were set aside as ChIP input material. For chromatin IPs, 450 µl lysis buffer was added to 50µl extract, and the appropriate Mcm1 antibody was added in the following quantities: 12µl antibody serum raised against a *S. cerevisiae* Mcm1 peptide[135]; 15µl affinity purified antibody raised against a C-terminal peptide from *K. lactis* Mcm1 (Bethyl Laboratories, Montgomery, TX); or 5µl affinity purified antibody raised against N- or C- terminal peptides from *C. albicans* Mcm1 (Bethyl Laboratories, Montgomery, TX). Extract plus antibody was incubated 2hr at 4°C, with agitation. 50µl

of a 50% suspension of protein A-Sepharose Fast-Flow beads (Sigma, St. Louis, MO) in lysis buffer was added and incubated 1.5hr at 4°C, with agitation. The beads were pelleted 1min at 3,000 × g. After removal of the supernatant, the beads were washed with a series of buffers for five minutes for each wash: twice in lysis buffer, twice in high salt lysis buffer (50 mM HEPES-KOH, pH 7.5, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate), twice in wash buffer (10 mM Tris-HCl [pH 8.0], 250 mM LiCl, 0.5% NP-40, 0.5% sodium deoxycholate, 1mM EDTA), and once in TE (10 mM Tris, 1 mM EDTA [pH 8.0]). After the last wash, 100µl of elution buffer (50 mM Tris/HCl [pH 8.0], 10 mM EDTA, 1% SDS)) was added to each sample, and the beads were incubated at 65 °C for 15 min. The beads were spun for 1 min at 10,000 × g, and the supernatant was removed and retained. A second elution was carried out with 150µl elution buffer 2 (TE, 0.67% SDS) and eluates from the two elution steps were combined. For the ChIP input material set aside earlier, 200µl TE, 1%SDS was added. ChIP and input samples were incubated overnight at 65°C to reverse crosslinks. 250µl proteinase K solution (TE, 20µg/ml glycogen, 400µg/ml Proteinase K) was added to each sample, and samples were incubated at 37°C for 2h. Samples were extracted once with 450µl Tris buffer-saturated phenol/chloroform/isoamyl alcohol solution (25:24:1). 55µl 4M LiCl and 1ml 100% ethanol (4°C) was added and the DNA was precipitated on ice for 1hr. The DNA was pelleted by centrifugation at 14,000 × g for 15 min at 4 °C, washed once with cold 75% ethanol, and allowed to air dry. Samples were resuspended in 25µl TE containing 100µg/ml RNaseA and incubated 1hr at 37°C.

For the verification of Mcm1 binding at the non-canonical motif in the myc-tagged Mcm1 strain (MRcan42) the protocol above was used with the following modifications.

Cells were lysed by bead mixing rather than by spheroplasting. After washing with TBS, cell pellets were resuspended in 700µl ice-cold lysis buffer with protease inhibitors and 500µl of 0.5mm glass beads was added. This mix was placed in an Eppendorf mixer (part #5432) for 2h at 4°C. Chromatin was sheared by sonication, as before, but in this case with a Bioruptor (Wolf Laboratories, Manchester, UK) for 15min (30s on, 60s off) on the medium setting. Here the IP was carried out with anti-myc antibody 9E10 (Invitrogen, Carlsbad, CA) and protein G-Sepharose Fast-Flow beads (Sigma, St. Louis, MO).

DNA amplification and labeling. ChIP-enriched DNA was amplified and fluorescence labeled as described[130]. Labeled DNA for each channel was combined and hybridized to arrays in Agilent hybridization chambers for 40 hours at 65°C, according to protocols supplied by Agilent (Agilent Technologies, Santa Clara, CA). Arrays were then washed and scanned, using an Axon Instruments Genepix 4000A scanner.

Identification of binding events in ChIP-Chip data

We evaluated several approaches for calling Mcm1 binding sites from the ChIP-Chip data in three species. In the end we determined that the Joint Binding Deconvolution[102] (JBD) algorithm provides the best combination of consistency across species and accuracy on a test set of previously characterized *S. cerevisiae* binding sites. In what follows we compare two methods for defining binding sites based on ChIP-Chip data:

1) A software package from Agilent called Chip Analytics v 1.3 (CA). This software first applies the single array error model (SAEM; first introduced by

Roberts et al.[136] for analysis of gene expression microarray data and subsequently used for ChIP-Chip analysis by Ren et al.[137]) to calculate an enrichment statistic (X_{bar}), which is a normalized comparison of signal in the Immuno-Precipitate (IP) channel to signal in the Whole Cell Extract (WCE) channel for every probe on the array. The distribution of X_{bar} values is then fit with a Gaussian by taking the mean and standard deviation over the entire distribution. Next, a p-value is assigned to each probe based on the placement of that probe's enrichment statistic on the fitted Gaussian distribution. "Segments" (regions likely to be bound by the IP-ed protein) are then called by a peak identification heuristic called the "Whitehead Per-Array Neighbourhood Model v1.0", which looks for neighboring probes with p-values below a threshold. Here we use the program's default parameters with the exception of the parameter specifying "maximum distance (in bp) for two probes to be considered as neighbors", which we set to 500bp rather than 1000bp. In testing this algorithm we vary only a single parameter, the p-value threshold for inclusion of a probe in a segment.

2) A software package from the Fraenkel Group called Joint Binding Deconvolution (JBD). The package takes a somewhat more sophisticated approach to the problem and is described in detail elsewhere[102]. Briefly, JBD treats the observed enrichment ratios as a convolution of unobserved discrete binding events and an "influence" function derived from the distribution of DNA shear lengths. JBD attempts to deconvolute these, producing probabilities of binding at or below the resolution of the tiling array. Here

again we use the program's default settings varying two parameters $p_{binding}$ and $\sum(p_{binding} * strength_{binding})$ to define bound regions. Prior to processing with JBD we perform a global loess normalization for each experimental replicate using Goulphar[138] (exact options are foreground=0, do.bgcorr=1, do.saturating=1, saturating=55000). A requirement of the JBD algorithm is an estimate of the influence function for each experimental replicate. We estimate this influence function from the data as described later in "Estimating influence functions for JBD".

We compared the two methods (CA and JBD) over a broad range of parameter choices on the two ChIP-Chip datasets from *S. cerevisiae* ("YPD" and "alphaF"). As shown in the receiver operating characteristic (ROC) plot below (Figure S2), both methods perform similarly on our test set of previously characterized Mcm1 bound genes from *S. cerevisiae* (see "A test set of Mcm1 regulated *S. cerevisiae* genes"). For example, there are some parameter settings for each algorithm that call ~94% of test set genes bound while also calling only ~4% of all genes bound in S. *cerevisiae*. In the absence of a test set of non-bound genes, we think this is the most appropriate way to evaluate performance of the algorithm.

These results suggest that the two algorithms perform equally well. As one might expect then, genes called as bound outside of the test set are also very similar between the algorithms. With this result we decided to move forward with an analysis on all three species using the much less computationally intensive CA algorithm.

Unfortunately, the success seen with the Chip Analytics algorithm in *S. cerevisiae* was not reproduced on the data from *K. lactis* and *C. albicans*. Due to the lack of published experiments, it is not feasible to put together test sets for *K. lactis* and *C. albicans*. Nevertheless, problems became apparent in our attempts to choose a single p-value threshold across species, as shown in the plot below (Figure S3). Here we compare the gene sets resulting from a variety of enrichment p-value parameters both within the *S. cerevisiae* test set (left axis; silver and black bars) and across the three genomes (right axis; pink, purple and blue lines). Our expectation was that lower values for this threshold would result in proportionally smaller sets of genes called as bound. However, a strange behavior is found especially for the *K. lactis* and *C. albicans* alphaF experiments, where lowering of the enrichment p-value threshold results in a disproportionate loss of Mcm1 targets compared to the loss in other experiments (e.g. compare the drop in "fraction of the genes in genome" bound by Mcm1 resulting from a drop in p-value threshold from 10^{-2} to 10^{-3} in the *C. albicans* YPD and alphaF experiments). This behavior suggests that the calculation of these p-values may be flawed.

Examination of the distributions of the enrichment statistic (X_{bar}) calculated from the SAEM reveals the underlying problem (Figure S4). For *S. cerevisiae* the assumption made by the SAEM that most probes on the array are not enriched is valid and the corresponding SAEM Gaussian fit (Figure S4a blue curve) is reasonable. However, this assumption is clearly not valid for *K. lactis* and *C. albicans* where a Gaussian distribution

for a subset of the probes is evident, but where a sizeable portion of the X_{bar} distribution is found in the tail (Figure S4b,c). As expected, the Gaussian fit provided by Chip Analytics (estimated from the mean and standard deviation of all X_{bar} values) for these data is highly suspect. That SAEM is problematic on datasets where >5% of probes show enrichment was previously recognized[139]. Because the shape of the X_{bar} distribution varies so much between experiments, we can not rely on this method to obtain p-values.

We attempted to remedy this problem by performing a least-squares fit of the data to a Gaussian in which the overall weight was not fixed at 1.0 (this method is referred to hereafter as "CA_FIX"). In other words, a fit to the following equation:

$$\frac{\alpha}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

While this approach does give an improved fit to the non-enriched portion of the data (Figure S4a,b,c red lines), strange behavior is still evident when applied across experiments (Figure S5). Although on the surface it would appear that we have corrected the scaling problem seen in Figure S3, examining plots of the ChIP enrichment for the YPD and alphaF experiments across the *C. albicans* genome reveals a new problem. For these two experiments CA_FIX predicts roughly equal numbers of bound genes across all p-value thresholds chosen. However, in a quick visual scan across the genome it is apparent that there are roughly 50-100% more Mcm1 binding events in the *C. albicans* alphaF experiment than in the *C. albicans* YPD experiment (plots online at http://genome.ucsf.edu/mcm1_evolution/). Perhaps there is no simple remedy to the SAEM approach when a large fraction of probes on the array is enriched.

In principle, the fatter tail exhibited in the X_{bar} distributions from the *K. lactis* and *C. albicans* alphaF experiments can be attributed to either of two sources: (1) a larger number of Mcm1 binding events or (2) a longer/wider DNA shear length distribution. A visual scan of the enrichment data plotted along the chromosomes suggests that there is indeed a larger number of Mcm1 binding events in the *K. lactis* and *C. albicans* alphaF experiments than in the other experiments. Comparison of the estimated influence functions for each experiment (Figure S6; also see "Supplementary Methods – Estimation of influence functions for JBD") and examination of the whole-cell extract DNA lengths on a gel (not shown), indicate that the shearing of DNA is probably not as complete in *K. lactis* and *C. albicans* alphaF experiments as it is in some of the other experiments. Therefore, in our experiments both effects are likely contributing to the substantial fraction of probes enriched in the IP.

Because our data are ill-suited for the SAEM analysis of CA, we turn to JBD, which performs equally well on the *S. cerevisiae* test set, but also has the advantage that it directly accounts for the variability of DNA shear distributions through its use of influence functions. As the plot below (Figure S7) shows, JBD gives us the scaling we expect across different parameter values, while also agreeing with our expectations from visual inspection of the data plotted across the genome (e.g. that there are roughly 50-100% more binding events in the *K. lactis* alphaF experiment than in the *K. lactis* YPD experiment).

Integration of motif information and the final Mcm1-bound segment calls

Mcm1 binds a well defined *cis* regulatory motif in *S. cerevisiae*[103, 104]. *De novo* motif finding with MEME[54] on sequences predicted by JBD to be bound with high confidence ($p_{binding} > 0.9$ and $\sum[p_{binding}*strength_{binding}] > 2.0$) gives Mcm1 motifs that are roughly the same in each species (see Figure 1). We decided to integrate this motif information into our Mcm1-bound segment calls in the following manner. The motif matrices found by MEME for the YPD ChIP-Chip experiments for each species were used to score overlapping 1 kb windows across the corresponding genome, taking the sum of all the odds ratios against the matrix in each window as the "motif score" for a window. The distribution of $\log_{10}$ motif scores was fit with a Gaussian, taking the mean and the variance of $\log_{10}$ motif scores across the genome. Motif p-values for each genomic window were calculated on this Gaussian distribution. For *C. albicans*, where we find Mcm1 associating with two different *cis* regulatory motifs (Figure 1), we modified the above method so that each 1kb window is scored with both the canonical and non-canonical Mcm1 motif matrices and the odds ratios of each across the window are summed.

In our final Mcm1-bound segment calls we integrated four parameters (one which defines the contribution of Mcm1 sequence motifs and three that define the contributions of ChIP-Chip enrichment):

(1) motif p-value; this parameter, on the scale 0.0 to 1.0, determines the weight given to the presence of Mcm1 motifs within the segment and can be overridden by parameter (4).

(2) $p_{binding}$; this parameter, on the scale 0.0 to 1.0, determines the minimum probability of binding as output by JBD.

(3) $\sum(p_{binding}*strength_{binding})$; this parameter, on the scale 0.0 to ∞, determines the minimum "sum of probability of binding times strength of binding" as output by JBD.

(4) $\sum(p_{binding}*strength_{binding})$ override; if the $\sum(p_{binding}*strength_{binding})$ is larger than this parameter then the region is called as Mcm1-bound regardless of motif p-value.

Parameters (2) and (3) were suggested by the developers of JBD as a reasonable way to define bound regions[102]. The fourth parameter was added when we observed that occasionally strong Mcm1 enrichment is unaccompanied by an Mcm1 motif, possibly due to errors in genome sequencing in the bound region or possibly due to recruitment of Mcm1 to promoters by other transcription factors.

A large number of parameter choices were sampled and the resulting *S. cerevisiae* test set accuracies for a subset of these choices (where $p_{binding} \geq 0.2$) are plotted in Figure S8 (left axis; silver and black bars). Additionally, the fraction of genes in each genome called as bound was reported for each experimental condition (right axis; pink, purple and blue lines). In the end we chose the following parameter cutoffs to define bound regions in each species:

motif p-value	0.1
$p_{binding}$	0.2
$\sum(p_{binding}*strength_{binding})$	0.5

$\sum(p_{binding}*strength_{binding})$ override 2.0

Estimating the rate of false positive Mcm1 binding site calls

Although the JBD algorithm performs more consistently on our Mcm1 datasets (as explained above), one deficiency of this program is its inability to estimate false positive rates. The authors of the JBD algorithm suggest two ways of estimating false positive rates (http://cgs.csail.mit.edu/jbd/signif.html). The first relies on a set of regions where it is known *a priori* that no binding events occur; this we do not have. The second method relies on scrambling the data for each probe with respect to the chromosomal coordinate, which we think is not a particularly accurate way of estimating false positives because it destroys much of the long range correlation structure of ChIP-Chip experiments that would tend to give rise to false positives in the first place.

Assuming that only those regions with both ChIP-Chip enrichment and Mcm1 binding site motifs represent the bona fide *cis*-acting sequences, we estimate that using the ChIP-Chip data alone yields false positive rates between 11 and 36%. This can be calculated from Figure S8 by comparing the "fraction of genes in the genome" bound when using our chosen parameters, where the motif p-value cutoff is 0.1, to those when the motif p-value cutoff is 1.0 (i.e., when motif matches are not considered). Because our motif p-values are derived by fitting the genome-wide motif score distribution to a Gaussian (explained in the previous section), applying our motif p-value cutoff of 0.1 to a set of randomly chosen genomic regions would reduce the number of regions by ~90%.

With the added requirement of an Mcm1 binding site motif under the ChIP-Chip peaks, we believe our false positive rates are likely much lower. For experiments in *S. cerevisiae*, our false positive rate *before* integrating the motif information can be estimated using our test set as follows:

$$PP_{after} = FP_{before} \times EFPR + TP_{before} \times ETPR = FP_{before} \times EFPR + (PP_{before} - FP_{before}) \times ETPR$$

where PP is the number of predicted positives and FP and TP are the numbers of false and true positives, respectively. An upper estimate of the rate at which false positives pass the motif filter (EFPR) is 0.1. The rate at which true positives pass the motif filter (ETPR) is estimated by the fraction of our test set remaining after motif filtering (28/31 = 0.9). Solving for FP_{before}, gives an estimate of ~7 false positives *before* filtering and thus < 1 (7 * 0.1) false positive *after* filtering.

Estimation of influence functions for JBD

An influence function is used by JBD to specify "the expected relative probe intensity as a function of distance from a binding event"[102]. In their paper Qi et al. derive the influence function from the distribution of shear fragment lengths. Here we estimate influence functions for each experiment by averaging the relative enrichment as a function of distance for the 50 strongest, idealized peaks in each experiment (Figure S6). Specifically, we start by sorting IP/WCE ratios for all probes (normalized with Goulphar as described above). We then move in descending order through the list and for each candidate probe look in the genomic region ±3kb, defining the candidate peak probe as having ratio $\left(\frac{IP}{WCE}\right)_0$ with position d=0. If (1) this flanking region does not contain a

probe that was previously annotated as part of another peak and (2) this region contains one probe with at least half the enrichment of the peak probe:

$$\text{i.e., } \frac{\left(\frac{IP}{WCE}\right)_x}{\left(\frac{IP}{WCE}\right)_0} \geq 0.5 \text{ for at least one probe x}$$

and (3) the $\left(\frac{IP}{WCE}\right)_x$ ratios continually decrease at more distant probes:

$$\text{ie., } \frac{\left(\frac{IP}{WCE}\right)_x}{\left(\frac{IP}{WCE}\right)_0} \text{ can not be greater than } 0.4 + \frac{\left(\frac{IP}{WCE}\right)_{\min at\ d \leq x}}{\left(\frac{IP}{WCE}\right)_0}$$

then the region is taken to be a peak and the relative enrichment levels, $\frac{\left(\frac{IP}{WCE}\right)_x}{\left(\frac{IP}{WCE}\right)_0}$, are recorded for each probe at distance x from the peak probe 0. The second criterion filters "peaks" consisting of a single probe (a.k.a. "blips") and the third criterion is a heuristic that attempts to filter regions in which there is more than one binding event. The algorithm terminates when it has recorded relative enrichment levels for 50 idealized peaks. The relative enrichments at each distance x are averaged across the peaks. As the resulting influence function is somewhat rough and typically lacking data for many distances x, we smooth it by replacing each relative enrichment level at distance x with the average of all relative enrichment levels in the range x ± 50 bp. The influence function for each of our experiments is plotted in Figure S6.

Inference of the 32 species fungal phylogeny

A robust phylogeny of the 32 yeast species was inferred using the methods similar to those previously developed[31, 61]. To build the phylogeny, orthologous gene sets (see "Mapping orthologous gene sets") containing one and only one representative from each of the 32 yeasts were chosen at a stringent branch length cutoff (0.5; see "Mapping orthologous gene sets" below). Of the resulting 122 orthologous gene sets, 22 are affected by the phenomenon of differential gene loss following the WGD and were filtered[62]. This yielded 100 orthologous gene sets that, showing no evidence for deletion or duplication events, were more likely than other orthologous gene sets to preserve the underlying speciation signal. The 32 sequences within each set were then multiply aligned with ClustalW[72]. The resulting 100 alignments were concatenated and columns containing gaps were dropped, producing a single alignment with 19,989 columns. Finally, a maximum likelihood species tree was estimated employing the TREE-PUZZLE algorithm with default parameters (VT substitution model)[73]. Demonstrating the robustness of this inference, a tree with identical topology and similar branch lengths was generated when the neighbor-joining method of ClustalW[72] was applied to the same dataset (not shown). The maximum likelihood algorithm PHYML[78] (using the WAG substitution model as recommended by ProtTest[79]) also yields a similar tree that differs only in that *E. gossypii* branches with *K. lactis* rather than just prior to the whole clade that spans *S. cerevisiae* to *K. lactis*. If this alternate tree is correct it only serves to strengthen our argument for four independent gains of Mcm1 regulation at ribosomal genes. The alternate topology does not affect any of the other arguments we present in the paper. The alternate placement of *C.glabrata* (swapped with *S. castellii*) recently

proposed by Scannell et al.[80] also serves to strengthen our argument for several independent gains of ribosomal gene regulation by Mcm1.

Mapping orthologous gene sets (OGSs)

Here again we use a method similar to that which we used previously[31]. We ran PSI-BLAST for each *S. cerevisiae* ORF "query" sequence against a single database containing all ORF sequences from each of the 32 fungal species, employing an E-value cutoff of 10^{-5} and the Smith-Waterman alignment option[82]. The sequences returned by PSI-BLAST were then multiply aligned with MUSCLE (setting the maxhours parameter to 0.5 and maxiters to 2 if the sequence database was greater than 50,000 residues in length)[140] and a neighbor joining tree (NJ) was inferred, using ClustalW[72]. Finally, the resulting NJ tree was traversed to extract an orthologous gene set (OGS) in the following manner: Start at the leaf node for the query sequence and ascend the tree, incrementing a level counter for each node ascended. At each internal node descend. If a leaf node is reached, the gene is from a species not yet seen at a lower level, and the branch length traversed is less than a cutoff (1.0), then add that gene to the OGS. This procedure was repeated for each *S. cerevisiae* sequence, resulting in a 32 species many-to-many ortholog map. For the purposes of generating Figure 2a,b and the section entitled "Genes bound in any one species are only moderately..." we attempted to reduce the species bias of this approach. For these analyses a second OGS map was used in which additional OGSs were generated using each of the ORF sequences from *K. lactis* and *C. albicans* not already present in an OGS as a query sequence. Because our goal in the results section entitled, "Mcm1 binding at a non-canonical motif upstream…", was to examine the

presence of the non-canonical Mcm1 motif found in *C. albicans* at the orthologous promoters of other species, an ortholog map was built using the same method as above, but with the set of all *C. albicans* (rather than *S. cerevisiae*) ORF sequences serving as the query database.

As previously[31] we found that because the number of **a**sgs (7 in *S. cerevisiae* and 6 in *C. albicans*) and αsgs (5 in *S. cerevisiae*) is small, a more careful ortholog mapping of these genes benefits downstream analyses of promoter sequences. For example, *MFA1* and *MFA2*, two **a**sgs from *S. cerevisiae*, are less than 40 amino acids long and were therefore not annotated as ORFs in several of the fungal sequencing projects. Using TBLASTN we identified putative *MFA1/MFA2* orthologs and added them to our ortholog map.

Robustness of results to parameter choices

In the Results section entitled "Genes bound in any one species are only moderately likely…" we claim that the results of our pairwise comparison of Mcm1 target gene sets between species are robust to the exact parameters chosen to define Mcm1 binding sites. In Figure S9 we support that claim with the results of the pairwise comparison of Figure 2a-b for a variety of parameter choices.

Inference of Mcm1 binding site gain and loss rates

In order to assess the prevalence of gain and loss of Mcm1 binding sites across the three species phylogeny we constructed a 4 branch model with 9 parameters: 4 gain rates ($g_{1\text{-}4}$) and 4 loss rates ($l_{1\text{-}4}$) corresponding to each of the 4 branches of the rooted tree and a

single parameter, p_1, representing the probability of an Mcm1 binding site at the root of the tree (Figures 2c and S10). We take as our dataset the Mcm1 binding occurrence patterns at each of the 2766 genes that can be mapped between *S. cerevisiae*, *K. lactis*, and *C. albicans* in a 1:1:1 fashion via our ortholog mapping. There are eight such patterns, e.g. the pattern "101" for hypothetical gene X indicates an Mcm1 binding site is present upstream of gene X in *S. cerevisiae* and *C. albicans*, but not in *K. lactis*. The total counts for each occurrence pattern can be found in Figure 2c.

We use maximum likelihood approach to estimate the parameters. Since there are nine independent parameters and only seven independent occurrence patterns (eight patterns with the normalization that the sum equals the total number of genes), there is degeneracy in the solution. In fact, there is a two dimensional space of solutions with equal maximum likelihood, i.e., they all perfectly fit the observed patterns. It can be shown that the four branch rooted tree model is mathematically equivalent (in terms of the statistics of the occurrence patterns) to a three branch star model, where the center is the *S. cerevisiae–K. lactis* ancestor (node B in Figure S10), with branches 2 and 3 leading to *S. cerevisiae* and *K. lactis* and "branch" * leading to *C. albicans*. For the 3 branch model, there is a unique maximum likelihood solution, thus the parameters for branches 2 and 3 are fixed. The parameters for the * branch and the probability of an Mcm1 binding site at B are also fixed, and are related to the variable parameters for branch 1 and 4 and the ancestor in the 4 branch model, allowing us to explicitly find all possible solutions:

$$p_B = p(B=1) = p_1(1-l_1)+(1-p_1)g_1$$
$$g_* = \frac{p(B=0,C=1)}{p(B=0)} = \frac{p_1 l_1(1-l_4)+(1-p_1)(1-g_1)g_4}{p_1 l_1+(1-p_1)(1-g_1)}$$
$$l_* = \frac{p(B=1,C=0)}{p(B=1)} = \frac{p_1(1-l_1)l_4+(1-p_1)g_1(1-g_4)}{p_1(1-l_1)+(1-p_1)g_1}$$

Solving the above equations for l_4 and g_4 and substituting $p_1 = \frac{p_B - g_1}{1-l_1-g_1}$ and $g_1 = \frac{p_B - p_1(1-l_1)}{1-p_1}$ gives:

$$l_4 = \frac{-l_* p_B + g_1\left(1+g_*(p_B-1)+(l_*-1)p_B\right)}{g_1 - p_B}$$
$$g_4 = \frac{-l_1(l_*-1)\,p_B + g_*(l_1+p_B-l_1\,p_B-1)}{-1+l_1+p_B}$$

Substituting our previously calculated 3-branch values for g_*, l_* and p_B gives the following equations in which the degeneracy of the four branch model is now clearly evident:

$$l_4 = \frac{-0.03772+0.90781\,g_1}{g_1-0.0563}$$
$$g_4 = \frac{0.05503\,l_1-0.0736}{l_1-0.9437}$$

To estimate these four rates (l_1, g_1, l_4 and g_4) and the probability of an Mcm1 binding site at the root of the tree (p_1), we chose the solution with minimal distance to all other equivalent ML solutions in this two dimensional space. We believe this is the most reasonable method for averaging with the constraint that the parameters chosen actually represent one of the ML solutions to the problem:

parameter	value	parameter	value
p_1	0.06		
g_1	0.029	l_1	0.521
g_2	0.017	l_2	0.697
g_3	0.049	l_3	0.241
g_4	0.061	l_4	0.410

Mcm1 DNA motifs are not present at genes that are not bound

Our argument that Mcm1 binding site turnover rates are high hinges on the completeness and reliability of our ChIP-Chip data. Here we attempt to demonstrate that genes which are not bound by Mcm1 in species A, but have an ortholog bound in another species B, also do not have evidence for Mcm1 *cis* regulatory elements in their promoters in species A. We find that 34% of genes bound by Mcm1 only in *S. cerevisiae* have Mcm1 motifs with $\log_{10}$-odds > 2.8 in *K. lactis* and *C. albicans*, a frequency roughly equivalent to the background rates of occurrence of 40% and 30% respectively (binomial $p > 0.8$ and $p > 0.3$ respectively). For *S. cerevisiae*, we find Mcm1 motifs at more than 67% of these promoters (employing the same $\log_{10}$-odds cutoff) as compared to a background rate of roughly 41% (binomial $p < 10^{-5}$). Similar results are obtained when genes bound by Mcm1 only in *K. lactis* and only in *C. albicans* are examined in the other species.

Mapping Mcm1-cofactor regulons across species

*Yox1, Fkh2, **a**-specific genes (**a**sgs).* Yox1 and Fkh2 regulons were identified in *S. cerevisiae* using Mcm1 ChIP-chip, cofactor ChIP-chip and cell-cycle gene expression data. Specifically, genes were taken to be part of the regulon if they were bound by

Mcm1 in our ChIP-Chip experiments, bound by the cofactor in Harbison et al.'s ChIP-Chip experiments[111] and cell-cycle oscillating in Spellman et al.'s gene expression experiments[130]. For **a**sgs, the union of previously defined regulons in *S. cerevisiae* and *C. albicans* was used[30, 31, 141]. The **a**sg regulons of *S. cerevisiae* and *C. albicans* are based on ChIP-Chip, mating-type gene expression data and promoter sequence scoring against MATα2-Mcm1 and MATa2-Mcm1 motif matrices.

Regulons defined in *S. cerevisiae* were mapped to *K. lactis* and *C. albicans* via our standard ortholog map (see "Mapping orthologous gene sets"). In each species, bound segments (±250bp) flanking genes in the mapped regulon and promoters (600bp upstream of the translational start) of orthologous genes in closely related species were scored for the presence of a single Mcm1 binding site sequence with log-odds score greater than 2.0. Here we use the same position weight matrices previously used to define the bound segments (Figure 1).

- For *S. cerevisiae*, orthologous promoters from *S. mikatae, S. bayanus, S. castellii and C. glabrata* were used.
- For *K. lactis*, orthologous promoters from *K. waltii, S. kluyveri and E gossypii* were used.
- For *C. albicans*, orthologous promoters from *C. dubliniensis, C. tropicalis, C. parapsilosis* and *L elongisporus* were used.

The resulting putatively Mcm1 bound subsequences and flanking sequence (to a final length of 40, centered on the Mcm1 motif) were submitted to MEME with a min length parameter of 25, a max length parameter of 40 and a target frequency of 0.5. The choice

of the target frequency parameter is somewhat arbitrary, but is based on the notion that not all submitted sequences are expected to be true operators within the regulon. The resulting multi-species position weight matrix (with 0.25 pseudocounts added for each nucleotide in each column) was then used to score all the bound segments in that species. A new, single-species weight matrix was then generated from only those sequences within the species which score below a p-value threshold (see below for calculation of p-value and choice of threshold; again, 0.25 pseudocounts were added for each nucleotide in each column of the weight matrix). The bound segments within the species were then rescored with the single-species weight matrix and final regulon membership was determined based on thresholding on the single-species weight matrix p-value.

For the purposes of standardizing cutoffs across species/regulons, sequence scores genome-wide for a given Mcm1-cofactor weight matrix were fit to a Gaussian so that the p-values could be calculated for each score. A p-value cutoff of 10^{-7} was employed in defining members of the regulon in both the first (using the multi-species weight matrix) and second (using the single-species weight matrix) passes. This parameter was chosen to maximize the number of known **a**sg regulon members in *S. cerevisiae* and *C. albicans*[31]; with a cutoff of 10^{-7} we achieve a 93% sensitivity, correctly identifying the *S. cerevisiae* recombinational enhancer and 12 of 13 known **a**sgs as members of the regulon (*C. albicans* Ram2 is missed because it lacks Mcm1 enrichment in our experiments), while also achieving near perfect specificity (only 6 bound regions flank ORFs that were not previously implicated as **a**sgs: orf19.171/orf19.172, orf19.2308,

orf19.7380/orf19.7381 from *C. albicans*, and Cdc6/Elo1, Cdc47, Mcm3 from *S. cerevisiae*).

MATα1. The α-specific gene regulon was defined as the union of previously defined MATα1 regulons from *S. cerevisiae* and *C. albicans*. These regulons are based on ChIP-chip and mating-type gene expression data[30, 141]. The procedure is the same as for the Yox1/Fkh2/**a**sgs regulons, but with two modifications because αsgs are not expected to be bound by Mcm1 in our ChIPs of **a** cells and because αsg motifs are often found in multiple copies upstream of target genes:

(1) Instead of restricting the motif search to Mcm1 bound segments, we used the promoters of all αsg orthologs.

(2) The MEME search is carried out on all subsequences scoring greater than 2.0, rather than just the max scoring subsequence greater than 2.0.

Arg81. A couple complications arose in the analysis of the Arg81 regulon. First, there was some question as to whether Mcm1 is really bound upstream of arginine metabolic genes with Arg80/81 *in vivo*. Our ChIPs indicate little to no enrichment of Mcm1 at genes typically cited as members of this regulon. This issue is discussed in greater length in the Results section. Second, it appears as if the strict positioning of the Mcm1/Arg80 *cis* regulatory site relative to Arg81 found in *S. cerevisiae* (see Figure 4b) is not found in *C. albicans*. For these reasons we took a somewhat different approach to mapping this regulon. We mapped genes encoding enzymes in the metabolic neighborhood of arginine (YPL111W, YOL058W, YJL088W, YER069W, YOL140W, YLR438W, YJL071W,

YHR018C, YMR062C, YOR303W and YJR109C) to orthologs in *K. lactis* and *C. albicans*. We then performed a MEME search on the promoters (500bp upstream of the translational start) for these genes and the promoters of orthologous genes in closely related species:

- For *S. cerevisiae*, orthologous promoters from *S. bayanus, S. castellii and C. glabrata* were used.
- For *K. lactis*, orthologous promoters from *K. waltii, S. kluyveri and E gossypii* were used.
- For *C. albicans*, orthologous promoters from *C. dubliniensis,* and *C. tropicalis* were used.

In *S. cerevisiae*, the expected Mcm1/Arg80-Arg81 motif was found (Figure 4b), but we could not use this motif matrix to score Mcm1 bound segments because Mcm1 does not appear to bind at this regulon in our *S. cerevisiae* ChIP-Chips. In *K. lactis* a similar motif was found. As with motifs found at the other regulons, we fit the genome-wide distribution of log-odds scores for this motif matrix with a Gaussian and then calculated p-values for the maximum scoring motif match at each *K. lactis* Mcm1 bound segment. A p-value cutoff of 10^{-6} was employed to define regulon membership, yielding 9 segments and 17 genes (including Arg1/3/8, Car1/2 and Gap1) as members. Finally, in *C. albicans* an Mcm1-like motif and Arg81-like motif were found in separate MEME rounds (Figure 4b). This suggests that while Arg81 and Mcm1 regulate arginine metabolic genes together in *C. albicans*, they do so via a mechanism that allows for relaxed spacing between the two transcription factors. We fit each genome-wide distribution of log-odds scores for these two motif matrices with a Gaussian and then

calculated p-values for the maximum scoring motif match at each *C. albicans* Mcm1 bound segment. We examined the 13 segments which matched both matrices with p-value < 10^{-5} and found that in 8 cases the two motif matches were spaced 15 to 39bp from each other. Thus these 8 segments, which flank the genes Arg1/3/4, Car1 and Cpa1/2, define the Mcm1-Arg81 regulon in *C. albicans*.

Yhp1. We could not identify an *S. cerevisiae* regulon from existing data.

A test set of Mcm1 regulated *S. cerevisiae* genes

A test set of previously characterized Mcm1 regulated genes was compiled from primary and secondary sources (Table S3). Secondary sources include:

YPD (Yeast Proteome Database; https://www.proteome.com/proteome/Retriever/)

SCPD (*S. cerevisiae* Promoter Database; http://rulai.cshl.edu/SCPD/)

TRANSFAC (http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi)

SUPPLEMENTARY TABLES

Table S1.

Lists of Mcm1-bound genes in each species.

S. cerevisiae	*K. lactis*	*C. albicans*
YAL022C	KLLA0A00242g	orf19.1011
YAL023C	KLLA0A00264g	orf19.1012
YAL038W	KLLA0A00418g	orf19.1048
YAL039C	KLLA0A00484g	orf19.1062
YAL040C	KLLA0A00506g	orf19.1066
YAR018C	KLLA0A00572g	orf19.1067
YBL001C	KLLA0A00594g	orf19.1070
YBL092W	KLLA0A00616g	orf19.1075
YBL093C	KLLA0A01199g	orf19.1078
YBR036C	KLLA0A02453g	orf19.1080
YBR037C	KLLA0A02475g	orf19.1093
YBR038W	KLLA0A02497g	orf19.1105.2
YBR066C	KLLA0A02541g	orf19.1105.3
YBR067C	KLLA0A02629g	orf19.1106
YBR077C	KLLA0A02849g	orf19.111
YBR078W	KLLA0A02871g	orf19.1120
YBR091C	KLLA0A02893g	orf19.1137
YBR092C	KLLA0A03025g	orf19.1139
YBR138C	KLLA0A03069g	orf19.1146
YBR139W	KLLA0A03091g	orf19.1148
YBR157C	KLLA0A03179g	orf19.1168
YBR158W	KLLA0A03201g	orf19.1169
YBR202W	KLLA0A03223g	orf19.118
YCL024W	KLLA0A04059g	orf19.121
YCL025C	KLLA0A04081g	orf19.122
YCL054W	KLLA0A04213g	orf19.1223
YCR024C-A	KLLA0A04235g	orf19.1224
YCR024C-B	KLLA0A04609g	orf19.1225
YCR065W	KLLA0A05346g	orf19.1234
YDL037C	KLLA0A05368g	orf19.1238
YDL227C	KLLA0A05500g	orf19.1239
YDR032C	KLLA0A05522g	orf19.1240
YDR033W	KLLA0A05687g	orf19.1257
YDR077W	KLLA0A05700g	orf19.1258

YDR084C	KLLA0A06303g	orf19.1268
YDR085C	KLLA0A06325g	orf19.1270
YDR132C	KLLA0A06336g	orf19.1277
YDR146C	KLLA0A06468g	orf19.1285
YDR147W	KLLA0A06556g	orf19.1286
YDR190C	KLLA0A06578g	orf19.1307
YDR191W	KLLA0A06886g	orf19.1311
YDR308C	KLLA0A07018g	orf19.1313
YDR309C	KLLA0A07040g	orf19.1321
YDR389W	KLLA0A07150g	orf19.1334
YDR451C	KLLA0A07172g	orf19.1358
YDR452W	KLLA0A07194g	orf19.1362
YDR461W	KLLA0A07216g	orf19.1363
YDR462W	KLLA0A07227g	orf19.1364
YDR506C	KLLA0A08602g	orf19.1365
YDR507C	KLLA0A08624g	orf19.1368
YDR524C-B	KLLA0A09009g	orf19.1369
YDR525W-A	KLLA0A09031g	orf19.1370
YDR528W	KLLA0A09053g	orf19.1401
YEL001C	KLLA0A09075g	orf19.1402
YEL032W	KLLA0A09097g	orf19.1415
YEL040W	KLLA0A09163g	orf19.1446
YEL044W	KLLA0A09185g	orf19.1473
YEL046C	KLLA0A10483g	orf19.1488
YEL047C	KLLA0A10505g	orf19.1490
YER001W	KLLA0A11110g	orf19.1497
YER110C	KLLA0A11374g	orf19.1499
YER111C	KLLA0A11396g	orf19.1505
YER112W	KLLA0A11418g	orf19.1522
YER149C	KLLA0A11704g	orf19.1535
YER150W	KLLA0A11726g	orf19.1536
YER158C	KLLA0A11748g	orf19.1539
YER159C	KLLA0B00671g	orf19.1541
YFL014W	KLLA0B00693g	orf19.1562
YFL016C	KLLA0B01474g	orf19.1582
YFL023W	KLLA0B01496g	orf19.1585
YFL024C	KLLA0B01562g	orf19.1598
YFL025C	KLLA0B01584g	orf19.1599
YFL026W	KLLA0B01606g	orf19.1604
YFL027C	KLLA0B01980g	orf19.1617
YFL028C	KLLA0B02002g	orf19.1618
YGL006W-A	KLLA0B02541g	orf19.1619
YGL007C-A	KLLA0B02563g	orf19.1621

YGL008C	KLLA0B02585g	orf19.1625
YGL021W	KLLA0B02717g	orf19.1626
YGL032C	KLLA0B02893g	orf19.1671
YGL116W	KLLA0B02915g	orf19.1687
YGL201C	KLLA0B02937g	orf19.1690
YGR014W	KLLA0B02959g	orf19.1702
YGR041W	KLLA0B03091g	orf19.1704
YGR047C	KLLA0B03113g	orf19.1708
YGR048W	KLLA0B03586g	orf19.1709
YGR077C	KLLA0B03608g	orf19.171
YGR078C	KLLA0B03630g	orf19.1720
YGR079W	KLLA0B03652g	orf19.1721
YGR085C	KLLA0B04664g	orf19.1743
YGR086C	KLLA0B04686g	orf19.1747
YGR092W	KLLA0B04774g	orf19.1748
YGR106C	KLLA0B04796g	orf19.1763
YGR108W	KLLA0B05038g	orf19.1764
YGR143W	KLLA0B05060g	orf19.177
YGR188C	KLLA0B05225g	orf19.1778
YGR189C	KLLA0B05247g	orf19.1779
YGR191W	KLLA0B05269g	orf19.1789.1
YGR229C	KLLA0B05291g	orf19.1793
YGR230W	KLLA0B05742g	orf19.1800
YGR279C	KLLA0B05786g	orf19.1801
YHL008C	KLLA0B05918g	orf19.1821
YHL009C	KLLA0B05951g	orf19.1835
YHL025W	KLLA0B05973g	orf19.1836
YHL026C	KLLA0B06138g	orf19.1842
YHL028W	KLLA0B06193g	orf19.1843
YHL029C	KLLA0B07370g	orf19.1867
YHR004C	KLLA0B07392g	orf19.1890
YHR005C	KLLA0B07447g	orf19.1891
YHR005C-A	KLLA0B07592g	orf19.1893
YHR006W	KLLA0B07601g	orf19.1906
YHR022C	KLLA0B07623g	orf19.1907
YHR022C-A	KLLA0B07645g	orf19.1934
YHR023W	KLLA0B07909g	orf19.1935
YHR098C	KLLA0B08151g	orf19.1944
YHR099W	KLLA0B08173g	orf19.1948
YHR149C	KLLA0B08195g	orf19.1957
YHR150W	KLLA0B08800g	orf19.1958
YHR151C	KLLA0B08822g	orf19.1959
YHR152W	KLLA0B08998g	orf19.1960

YIL015W	KLLA0B09724g	orf19.1961
YIL069C	KLLA0B09746g	orf19.1963
YIL070C	KLLA0B10010g	orf19.1964
YIL076W	KLLA0B10032g	orf19.1978
YIL077C	KLLA0B10076g	orf19.201
YIL106W	KLLA0B10098g	orf19.2059
YIL107C	KLLA0B10120g	orf19.206
YIL122W	KLLA0B10351g	orf19.2060
YIL123W	KLLA0B10373g	orf19.2077
YIL158W	KLLA0B11055g	orf19.2082
YJL051W	KLLA0B11231g	orf19.2084
YJL079C	KLLA0B11253g	orf19.215
YJL115W	KLLA0B11495g	orf19.216
YJL116C	KLLA0B11517g	orf19.2169
YJL127C-B	KLLA0B11594g	orf19.2170
YJL157C	KLLA0B11616g	orf19.2179
YJL158C	KLLA0B12056g	orf19.218
YJL159W	KLLA0B12958g	orf19.220
YJL160C	KLLA0B12980g	orf19.2238
YJL170C	KLLA0B13211g	orf19.2247
YJL171C	KLLA0B13233g	orf19.2253
YJL194W	KLLA0B13321g	orf19.2308
YJL196C	KLLA0B13343g	orf19.2332
YJR090C	KLLA0B13365g	orf19.2333
YJR091C	KLLA0B13387g	orf19.2356
YJR092W	KLLA0B13409g	orf19.24
YJR094C	KLLA0B13431g	orf19.2451
YJR094W-A	KLLA0B13838g	orf19.2452
YKL032C	KLLA0B13860g	orf19.2459
YKL104C	KLLA0B14234g	orf19.2460
YKL105C	KLLA0B14256g	orf19.250
YKL163W	KLLA0B14498g	orf19.251
YKL164C	KLLA0B14817g	orf19.2517
YKL185W	KLLA0B14861g	orf19.2638
YKL186C	KLLA0B14883g	orf19.2639
YKL208W	KLLA0B14949g	orf19.2652
YKL209C	KLLA0C00352g	orf19.2653
YKR041W	KLLA0C00374g	orf19.2654
YKR042W	KLLA0C00671g	orf19.2672
YKR065C	KLLA0C00693g	orf19.2685
YKR066C	KLLA0C00935g	orf19.2686
YKR067W	KLLA0C00957g	orf19.2690
YKR097W	KLLA0C01650g	orf19.2691

YLR034C	KLLA0C01848g	orf19.271
YLR083C	KLLA0C01870g	orf19.2723
YLR084C	KLLA0C02233g	orf19.2726
YLR110C	KLLA0C02255g	orf19.2747
YLR113W	KLLA0C02343g	orf19.2757
YLR130C	KLLA0C02365g	orf19.2758
YLR131C	KLLA0C02937g	orf19.2765
YLR154C-G	KLLA0C03069g	orf19.2766
YLR189C	KLLA0C03091g	orf19.2767
YLR190W	KLLA0C03113g	orf19.2787
YLR254C	KLLA0C03179g	orf19.2788
YLR256W	KLLA0C03410g	orf19.2809
YLR272C	KLLA0C03432g	orf19.2810
YLR273C	KLLA0C03454g	orf19.2813
YLR274W	KLLA0C03564g	orf19.2822
YLR332W	KLLA0C03586g	orf19.2823
YLR342W	KLLA0C03960g	orf19.2831
YML027W	KLLA0C03982g	orf19.2832
YML052W	KLLA0C04015g	orf19.2833
YML053C	KLLA0C04037g	orf19.2870
YML054C-A	KLLA0C04103g	orf19.2871
YML057W	KLLA0C04125g	orf19.2881
YML058W	KLLA0C04213g	orf19.2882
YML058W-A	KLLA0C04796g	orf19.2892
YML059C	KLLA0C04809g	orf19.2903
YML119W	KLLA0C04818g	orf19.2905
YML120C	KLLA0C05016g	orf19.2929
YMR001C	KLLA0C05038g	orf19.2941
YMR001C-A	KLLA0C05060g	orf19.2942
YMR002W	KLLA0C06094g	orf19.2943.5
YMR031C	KLLA0C06116g	orf19.2952
YMR032W	KLLA0C06138g	orf19.2953
YMR121C	KLLA0C06677g	orf19.2954
YMR122W-A	KLLA0C06699g	orf19.2962
YMR123W	KLLA0C06721g	orf19.2990
YMR199W	KLLA0C07755g	orf19.2991
YMR252C	KLLA0C08173g	orf19.3003
YMR253C	KLLA0C08195g	orf19.301
YMR255W	KLLA0C08217g	orf19.3010.1
YMR305C	KLLA0C08283g	orf19.302
YMR306W	KLLA0C08371g	orf19.305
YNL053W	KLLA0C08866g	orf19.3105
YNL056W	KLLA0C08888g	orf19.3127

YNL058C	KLLA0C11407g	orf19.3130
YNL059C	KLLA0C11429g	orf19.3131
YNL145W	KLLA0C11495g	orf19.3133
YNL146C-A	KLLA0C11517g	orf19.3134
YNL146W	KLLA0C12133g	orf19.3149
YNL190W	KLLA0C12177g	orf19.3152
YNL289W	KLLA0C12199g	orf19.3193
YNL298W	KLLA0C12309g	orf19.3195
YNL327W	KLLA0C12551g	orf19.3221
YNL328C	KLLA0C12573g	orf19.3222
YNL329C	KLLA0C13013g	orf19.3234
YNR028W	KLLA0C13035g	orf19.3234.1
YNR061C	KLLA0C13057g	orf19.3261
YNR062C	KLLA0C13277g	orf19.3264
YNR063W	KLLA0C13519g	orf19.3268
YOL011W	KLLA0C13541g	orf19.3269
YOL012C	KLLA0C14047g	orf19.3302
YOR022C	KLLA0C14069g	orf19.3304
YOR023C	KLLA0C14091g	orf19.3305
YOR025W	KLLA0C14454g	orf19.3328
YOR058C	KLLA0C14762g	orf19.3336
YOR066W	KLLA0C15433g	orf19.334
YOR245C	KLLA0C15455g	orf19.335
YOR246C	KLLA0C16005g	orf19.3374
YOR247W	KLLA0C16357g	orf19.3392
YOR313C	KLLA0C16423g	orf19.3393
YOR315W	KLLA0C16445g	orf19.34
YOR342C	KLLA0C16467g	orf19.3406
YOR344C	KLLA0C16489g	orf19.3413
YOR346W	KLLA0C16511g	orf19.3414
YPL187W	KLLA0C16874g	orf19.3415
YPL242C	KLLA0C17226g	orf19.3417
YPL255W	KLLA0C17600g	orf19.3418
YPL256C	KLLA0C18216g	orf19.3433
YPR112C	KLLA0C18513g	orf19.3434
YPR113W	KLLA0C18546g	orf19.344
YPR119W	KLLA0C18909g	orf19.345
YPR156C	KLLA0C18931g	orf19.3455
YPR157W	KLLA0C19129g	orf19.3456
YPR194C	KLLA0C19151g	orf19.3499
YPR196W	KLLA0C19184g	orf19.35
	KLLA0C19206g	orf19.3501
	KLLA0C19250g	orf19.3527

KLLA0C19316g	orf19.3528
KLLA0C19338g	orf19.3529
KLLA0C19360g	orf19.3568
KLLA0C19382g	orf19.3569
KLLA0C19404g	orf19.3577.1
KLLA0C19437g	orf19.3578
KLLA0D00506g	orf19.3579
KLLA0D00528g	orf19.3586
KLLA0D01474g	orf19.3589
KLLA0D01507g	orf19.3590
KLLA0D02310g	orf19.3618
KLLA0D02332g	orf19.3642
KLLA0D02354g	orf19.3643
KLLA0D02420g	orf19.3646
KLLA0D02442g	orf19.3668
KLLA0D02970g	orf19.3670
KLLA0D02992g	orf19.3671
KLLA0D03014g	orf19.3672
KLLA0D03388g	orf19.3674
KLLA0D03410g	orf19.3675
KLLA0D03542g	orf19.3719
KLLA0D03608g	orf19.3720
KLLA0D04059g	orf19.3721
KLLA0D04158g	orf19.3722
KLLA0D04180g	orf19.3733
KLLA0D04202g	orf19.3734
KLLA0D05115g	orf19.3751
KLLA0D05159g	orf19.3752
KLLA0D05181g	orf19.3757
KLLA0D05247g	orf19.3764
KLLA0D05269g	orf19.3770
KLLA0D05291g	orf19.3793
KLLA0D05621g	orf19.3794
KLLA0D05643g	orf19.3799
KLLA0D05951g	orf19.3802
KLLA0D06919g	orf19.3803
KLLA0D06941g	orf19.3804
KLLA0D06963g	orf19.3845
KLLA0D07405g	orf19.3846
KLLA0D07766g	orf19.3868
KLLA0D07788g	orf19.3869
KLLA0D07810g	orf19.3884
KLLA0D07832g	orf19.3885

KLLA0D07854g	orf19.3893
KLLA0D08283g	orf19.3895
KLLA0D08305g	orf19.3897
KLLA0D08327g	orf19.3932
KLLA0D08602g	orf19.3934
KLLA0D08624g	orf19.3935
KLLA0D09240g	orf19.3936
KLLA0D09262g	orf19.3941
KLLA0D09548g	orf19.3944
KLLA0D09559g	orf19.3945
KLLA0D09581g	orf19.3968
KLLA0D09977g	orf19.3974
KLLA0D09999g	orf19.3981
KLLA0D10021g	orf19.3982
KLLA0D10197g	orf19.3997
KLLA0D10219g	orf19.4000
KLLA0D10505g	orf19.4002
KLLA0D10527g	orf19.403
KLLA0D10549g	orf19.4056
KLLA0D10637g	orf19.4059
KLLA0D10659g	orf19.4060
KLLA0D11022g	orf19.4063
KLLA0D11198g	orf19.4064
KLLA0D11220g	orf19.4066
KLLA0D11550g	orf19.4069
KLLA0D11572g	orf19.4070
KLLA0D11594g	orf19.4072
KLLA0D11660g	orf19.4145
KLLA0D12320g	orf19.4146
KLLA0D12342g	orf19.4147
KLLA0D12518g	orf19.4148
KLLA0D12540g	orf19.4149.1
KLLA0D12584g	orf19.4166
KLLA0D12606g	orf19.4167
KLLA0D12628g	orf19.4220
KLLA0D12672g	orf19.4221
KLLA0D12694g	orf19.4222
KLLA0D12914g	orf19.4231
KLLA0D13002g	orf19.4232
KLLA0D13816g	orf19.4233
KLLA0D14058g	orf19.4245
KLLA0D14080g	orf19.4246
KLLA0D14091g	orf19.4250

KLLA0D14113g	orf19.4251
KLLA0D14905g	orf19.4255
KLLA0D15246g	orf19.4257
KLLA0D15268g	orf19.4273
KLLA0D15378g	orf19.4274
KLLA0D15400g	orf19.4279
KLLA0D15543g	orf19.4280
KLLA0D15565g	orf19.430
KLLA0D15631g	orf19.4304
KLLA0D15653g	orf19.4308
KLLA0D15895g	orf19.4309
KLLA0D15917g	orf19.431
KLLA0D16005g	orf19.4315
KLLA0D16027g	orf19.4316
KLLA0D16049g	orf19.4318
KLLA0D16071g	orf19.432
KLLA0D16434g	orf19.4320
KLLA0D16456g	orf19.4321
KLLA0D16588g	orf19.4322
KLLA0D16962g	orf19.4342
KLLA0D17952g	orf19.4349
KLLA0D17974g	orf19.4353
KLLA0D17996g	orf19.4354
KLLA0D18304g	orf19.4375.1
KLLA0D18502g	orf19.4376
KLLA0D18535g	orf19.4390
KLLA0D19162g	orf19.4404
KLLA0D19184g	orf19.4405
KLLA0D19470g	orf19.4424
KLLA0D19492g	orf19.4426
KLLA0D19943g	orf19.4427
KLLA0E00484g	orf19.4438
KLLA0E00506g	orf19.4456
KLLA0E00682g	orf19.4459
KLLA0E00704g	orf19.4461
KLLA0E01001g	orf19.4463
KLLA0E01023g	orf19.4475
KLLA0E01694g	orf19.4476
KLLA0E01716g	orf19.4477
KLLA0E02618g	orf19.4478
KLLA0E02772g	orf19.450
KLLA0E02794g	orf19.451
KLLA0E03146g	orf19.4527

KLLA0E03168g	orf19.4528
KLLA0E03333g	orf19.4531
KLLA0E03355g	orf19.454
KLLA0E03377g	orf19.4555
KLLA0E03597g	orf19.4565
KLLA0E03619g	orf19.4579
KLLA0E03663g	orf19.4590
KLLA0E03751g	orf19.4591
KLLA0E05071g	orf19.4592
KLLA0E05093g	orf19.4593
KLLA0E05588g	orf19.4599
KLLA0E05610g	orf19.4600
KLLA0E05852g	orf19.4623.3
KLLA0E05874g	orf19.4629
KLLA0E05896g	orf19.4630
KLLA0E05918g	orf19.4631
KLLA0E05962g	orf19.4649
KLLA0E05984g	orf19.4651
KLLA0E06809g	orf19.4653
KLLA0E06831g	orf19.467
KLLA0E06963g	orf19.4688
KLLA0E06985g	orf19.4748
KLLA0E07007g	orf19.4749
KLLA0E07095g	orf19.4752
KLLA0E07458g	orf19.4753
KLLA0E07502g	orf19.4769
KLLA0E07524g	orf19.4775
KLLA0E08151g	orf19.4776
KLLA0E09075g	orf19.4781
KLLA0E09790g	orf19.4783
KLLA0E09878g	orf19.4784
KLLA0E09900g	orf19.4798
KLLA0E10813g	orf19.4799
KLLA0E10835g	orf19.4818
KLLA0E10857g	orf19.4833
KLLA0E10945g	orf19.4857
KLLA0E10967g	orf19.4858
KLLA0E10989g	orf19.4867
KLLA0E11704g	orf19.4869
KLLA0E12221g	orf19.4883
KLLA0E12243g	orf19.4884
KLLA0E12265g	orf19.4885
KLLA0E12287g	orf19.4890

KLLA0E12353g	orf19.4892
KLLA0E12375g	orf19.4900
KLLA0E12397g	orf19.491
KLLA0E12419g	orf19.4913
KLLA0E12441g	orf19.4914
KLLA0E12463g	orf19.4927
KLLA0E12507g	orf19.4933
KLLA0E12529g	orf19.4934
KLLA0E12551g	orf19.4936
KLLA0E12947g	orf19.4952
KLLA0E12969g	orf19.4960
KLLA0E12991g	orf19.4961
KLLA0E13277g	orf19.4972
KLLA0E13409g	orf19.4975
KLLA0E13431g	orf19.4976
KLLA0E13926g	orf19.4991
KLLA0E13948g	orf19.4993
KLLA0E14256g	orf19.5017
KLLA0E14432g	orf19.5019
KLLA0E14454g	orf19.5023
KLLA0E14938g	orf19.5032
KLLA0E14960g	orf19.5037
KLLA0E14982g	orf19.5038
KLLA0E15620g	orf19.508
KLLA0E15642g	orf19.5094
KLLA0E15840g	orf19.510
KLLA0E15862g	orf19.5102
KLLA0E16214g	orf19.511
KLLA0E16313g	orf19.5124
KLLA0E16335g	orf19.5131
KLLA0E17193g	orf19.5132
KLLA0E17391g	orf19.5169
KLLA0E17413g	orf19.5170
KLLA0E18436g	orf19.5171
KLLA0E18645g	orf19.5188
KLLA0E18678g	orf19.5203
KLLA0E19173g	orf19.5242
KLLA0E19855g	orf19.5248
KLLA0E20295g	orf19.5249
KLLA0E20317g	orf19.5267
KLLA0E20339g	orf19.5274
KLLA0E20449g	orf19.5299
KLLA0E20515g	orf19.5300

KLLA0E20537g	orf19.5314
KLLA0E20559g	orf19.532
KLLA0E20845g	orf19.533
KLLA0E20867g	orf19.535
KLLA0E20889g	orf19.5380
KLLA0E21153g	orf19.5384
KLLA0E21736g	orf19.5437
KLLA0E21758g	orf19.5438
KLLA0E22011g	orf19.5493
KLLA0E22055g	orf19.5495
KLLA0E22077g	orf19.5501
KLLA0E22099g	orf19.5513
KLLA0E22176g	orf19.5514
KLLA0E22198g	orf19.5519
KLLA0E22330g	orf19.5520
KLLA0E23705g	orf19.5521
KLLA0E23727g	orf19.5531
KLLA0E23749g	orf19.5532
KLLA0E23760g	orf19.5536
KLLA0E23826g	orf19.5537
KLLA0E23848g	orf19.5539
KLLA0E23892g	orf19.5555
KLLA0E24090g	orf19.5556
KLLA0F00352g	orf19.556
KLLA0F00594g	orf19.557
KLLA0F00616g	orf19.5572
KLLA0F00638g	orf19.5573
KLLA0F00682g	orf19.5574
KLLA0F00704g	orf19.5602
KLLA0F00726g	orf19.5604
KLLA0F01210g	orf19.5610
KLLA0F01364g	orf19.5620
KLLA0F01386g	orf19.5629
KLLA0F01408g	orf19.5650
KLLA0F01595g	orf19.5651
KLLA0F01903g	orf19.5663
KLLA0F02299g	orf19.5664
KLLA0F02750g	orf19.5711
KLLA0F02772g	orf19.5716
KLLA0F02816g	orf19.5717
KLLA0F02838g	orf19.5727
KLLA0F03146g	orf19.5728
KLLA0F03168g	orf19.5729

KLLA0F04015g	orf19.5741
KLLA0F04235g	orf19.575
KLLA0F04257g	orf19.5757
KLLA0F04279g	orf19.5758
KLLA0F04433g	orf19.5760
KLLA0F04477g	orf19.5770
KLLA0F04499g	orf19.5784
KLLA0F04840g	orf19.5798
KLLA0F05247g	orf19.5799
KLLA0F05555g	orf19.5801
KLLA0F06006g	orf19.5838
KLLA0F06028g	orf19.5844
KLLA0F06831g	orf19.5845
KLLA0F06853g	orf19.5854
KLLA0F07051g	orf19.5862
KLLA0F07073g	orf19.5901
KLLA0F07579g	orf19.5902
KLLA0F07601g	orf19.5903
KLLA0F07623g	orf19.5906
KLLA0F07843g	orf19.5908
KLLA0F07865g	orf19.5910
KLLA0F08151g	orf19.5911
KLLA0F08162g	orf19.5912
KLLA0F08228g	orf19.5960
KLLA0F08261g	orf19.5962
KLLA0F08635g	orf19.5963
KLLA0F08657g	orf19.5974
KLLA0F08679g	orf19.5975
KLLA0F08701g	orf19.5991
KLLA0F09691g	orf19.5992
KLLA0F09713g	orf19.5994
KLLA0F09790g	orf19.5999
KLLA0F09812g	orf19.6000
KLLA0F10043g	orf19.6003
KLLA0F10065g	orf19.6007
KLLA0F10087g	orf19.6008
KLLA0F10285g	orf19.6010
KLLA0F10307g	orf19.6017
KLLA0F10769g	orf19.6021
KLLA0F10791g	orf19.6022
KLLA0F11682g	orf19.6027
KLLA0F11704g	orf19.6028
KLLA0F11726g	orf19.6053

KLLA0F12584g	orf19.6054
KLLA0F12606g	orf19.6055
KLLA0F12782g	orf19.6086
KLLA0F12892g	orf19.6090
KLLA0F12914g	orf19.6091
KLLA0F13134g	orf19.6092
KLLA0F13156g	orf19.6094
KLLA0F13178g	orf19.6096
KLLA0F13288g	orf19.610
KLLA0F13310g	orf19.6118
KLLA0F13332g	orf19.6119
KLLA0F13398g	orf19.6121
KLLA0F13684g	orf19.6124
KLLA0F14366g	orf19.6139
KLLA0F14377g	orf19.6141
KLLA0F15906g	orf19.6160
KLLA0F15928g	orf19.6163
KLLA0F16324g	orf19.6175
KLLA0F16346g	orf19.6176
KLLA0F16577g	orf19.6177
KLLA0F16599g	orf19.6178
KLLA0F16665g	orf19.6196
KLLA0F16709g	orf19.6197
KLLA0F16753g	orf19.6200
KLLA0F16775g	orf19.6202
KLLA0F16797g	orf19.6224
KLLA0F16907g	orf19.6293
KLLA0F17160g	orf19.6305
KLLA0F17182g	orf19.6306
KLLA0F17369g	orf19.6307
KLLA0F17391g	orf19.6310
KLLA0F17589g	orf19.6312
KLLA0F17776g	orf19.6328
KLLA0F17798g	orf19.6329
KLLA0F18018g	orf19.6336
KLLA0F18040g	orf19.6536
KLLA0F18084g	orf19.655
KLLA0F18106g	orf19.656
KLLA0F18194g	orf19.6585
KLLA0F18216g	orf19.6586
KLLA0F18524g	orf19.6592
KLLA0F18546g	orf19.6594
KLLA0F18865g	orf19.6596

KLLA0F18887g	orf19.660
KLLA0F18975g	orf19.661
KLLA0F19019g	orf19.663
KLLA0F20031g	orf19.6640
KLLA0F20053g	orf19.6641
KLLA0F20614g	orf19.6642
KLLA0F20702g	orf19.6656
KLLA0F20724g	orf19.6659
KLLA0F20988g	orf19.6678
KLLA0F21010g	orf19.6679
KLLA0F22066g	orf19.6680
KLLA0F22088g	orf19.6689
KLLA0F22154g	orf19.670.2
KLLA0F22649g	orf19.671
KLLA0F22671g	orf19.6713
KLLA0F22682g	orf19.6715
KLLA0F22792g	orf19.6734
KLLA0F23111g	orf19.6736
KLLA0F23353g	orf19.6737
KLLA0F23375g	orf19.675
KLLA0F24882g	orf19.6760
KLLA0F24904g	orf19.6763
KLLA0F25102g	orf19.6771
KLLA0F25520g	orf19.6773
KLLA0F25542g	orf19.6784
KLLA0F25784g	orf19.6785
KLLA0F26268g	orf19.6786
KLLA0F26400g	orf19.6789
KLLA0F26422g	orf19.6790
KLLA0F27225g	orf19.6805
KLLA0F27423g	orf19.6817
KLLA0F27445g	orf19.6818
KLLA0F27995g	orf19.683
	orf19.6834.10
	orf19.6844
	orf19.6852
	orf19.686
	orf19.6863
	orf19.6864
	orf19.687
	orf19.6874
	orf19.6889
	orf19.6922

orf19.6941
orf19.6942
orf19.6944
orf19.695
orf19.6950
orf19.696
orf19.6968
orf19.698
orf19.6983
orf19.6984
orf19.6986
orf19.699
orf19.6996
orf19.6998
orf19.700
orf19.701
orf19.7024
orf19.7025
orf19.7030
orf19.7049
orf19.7054
orf19.7055
orf19.7068
orf19.7069
orf19.7077
orf19.7078
orf19.715
orf19.7151
orf19.7152
orf19.7153
orf19.716
orf19.7203
orf19.7204
orf19.721
orf19.7218
orf19.7219
orf19.723
orf19.7231
orf19.7247
orf19.7250
orf19.7251
orf19.7252
orf19.728

orf19.7305
orf19.7306
orf19.7308
orf19.7312
orf19.7332
orf19.7336
orf19.7337
orf19.7341
orf19.7342
orf19.7350
orf19.7362
orf19.7363
orf19.7372
orf19.7374
orf19.7377
orf19.7380
orf19.7381
orf19.7382
orf19.740
orf19.7409
orf19.7409.1
orf19.7436
orf19.744
orf19.7440
orf19.7448
orf19.745
orf19.7468
orf19.7469
orf19.7489
orf19.7502
orf19.7521
orf19.7522
orf19.7539
orf19.7539.1
orf19.7547
orf19.7548
orf19.7550
orf19.7551
orf19.7554
orf19.7555
orf19.7566
orf19.7585
orf19.7586

orf19.7592
orf19.7600
orf19.7601
orf19.761
orf19.762
orf19.7648
orf19.767
orf19.7676
orf19.7678
orf19.778
orf19.8
orf19.801
orf19.802
orf19.804
orf19.813
orf19.814
orf19.815
orf19.828
orf19.829
orf19.84
orf19.85
orf19.850
orf19.851
orf19.86
orf19.867
orf19.868
orf19.893
orf19.9
orf19.90
orf19.909
orf19.932
orf19.933
orf19.935
orf19.938
orf19.948
orf19.949
orf19.951
orf19.986
orf19.997

Table S2.

List of genomes used in this work.

Species	Source
S. cerevisiae	[142]
S. paradoxus	[121]
S. mikatae	[121]
S. bayanus	[121]
S. castellii	[143]
C. glabrata	[142]
K. waltii	[121]
S. kluyveri	[143]
K. lactis	[142]
E. gossypii	[142]
C. dubliniensis	[144]
C. albicans	[145, 146]
C. tropicalis	[147]
C. parapsilosis	[144]
L. elongisporus	[147]
C. guilliermondii	[147]
D. hansenii	[142]
C. lusitaniae	[147]
Y. lipolytica	[142]
A. terreus	[147]
A. nidulans	[147]
H. capsulatum	[147]
U. reesii	[147]
C. immitis	[147]
F. graminearum	[147]
T. reesei	[148]
M. grisea	[147]
C. globosum	[147]
N. crassa	[147]
S. sclerotiorum	[147]
S. japonicus	[147]
S. pombe	[142]

Table S3.

A test set of Mcm1 regulated *S. cerevisiae* genes.

Gene	ORF Id	Source DB	Literature Source
ACE2	YLR131C	--	[149]
AGA1	YNR044W	YPD	[150]
AGA2	YGL032C		[151]
ARG1	YOL058W	--	[152]
ARG5,6	YER069W	YPD	[153]
ASE1	YOR058C	--	[154]
ASG7	YJL170C	YPD	[151]
BAR1	YIL015W	YPD	[141, 151]
BUD4	YJR092W	--	[155]
CAR1	YPL111W	YPD	[153]
CAR2	YLR438W	YPD	[153]
CCP1	YKR066C	SCPD	[156]
CDC20	YGL116W	--	[98]
CDC46	YLR274W	YPD	[157]
CDC47	YBR202W	YPD	[157]
CDC5	YMR001C	YPD	[158]
CDC6	YJL194W	YPD	[157]
CLB1	YGR108W	YPD	[158]
CLB2	YPR119W	YPD	[158]
CLN3	YAL040C	TRANSFAC	[159]
FAR1	YJL157C	SCPD	[160]
HSP150	YJL159W	SCPD	[156]
MCM3	YEL032W	YPD	[98]
MFA1	YDR461W	SCPD	[161]
MFA2	YNL145W	--	[161]
PCK1	YKR097W	SCPD	[156]
PIS1	YPR113W	SCPD	[156]
PMA1	YGL008C	YPD	[156]
SPS4	YOR313C	--	[106]
STE2	YFL026W	YPD	[141, 151]
STE6	YKL209C	YPD	[162]
SWI4	YER111C	YPD	[157]
SWI5	YDR146C	YPD	[158]

SUPPLEMENTARY FIGURES

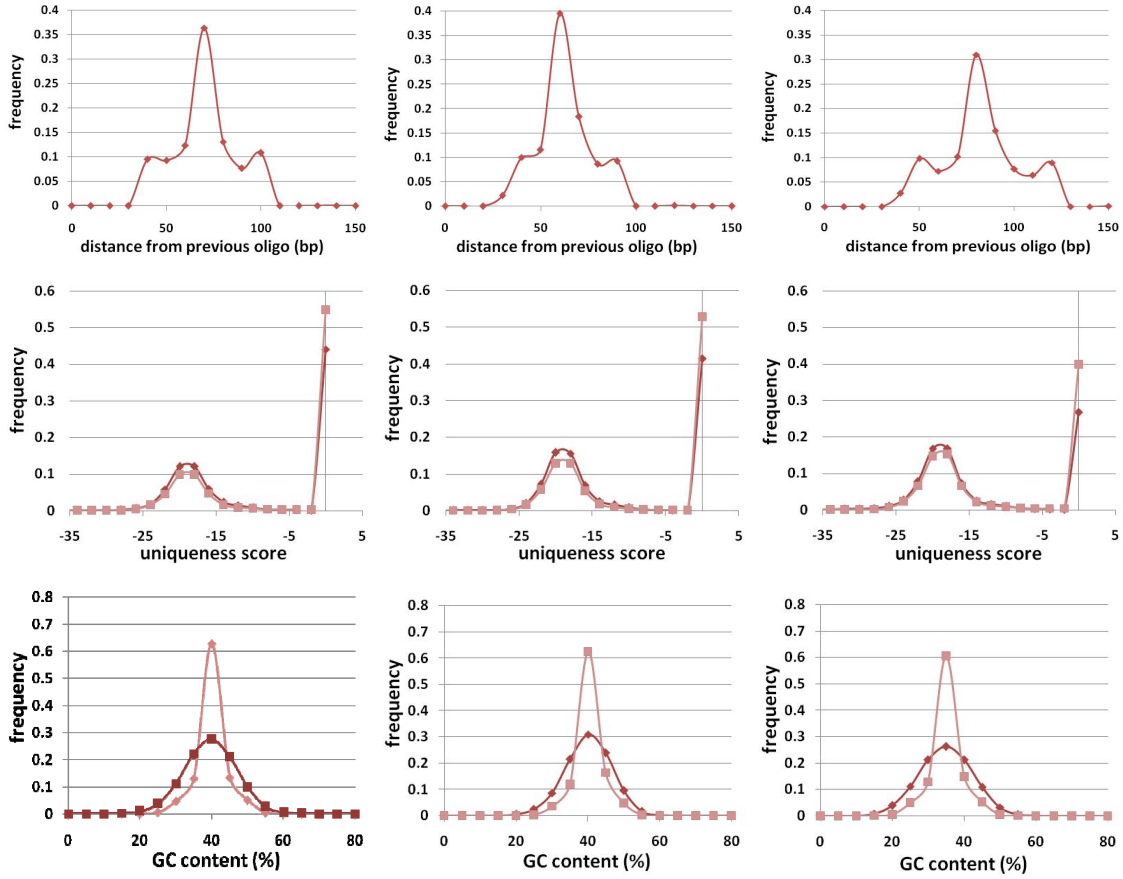


Figure S1. Evaluation of tiling array design.

Columns 1-3 contain plots for the *S. cerevisiae, K. lactis* and *C. albicans* tiling array designs, respectively. See Supplementary Methods for description.

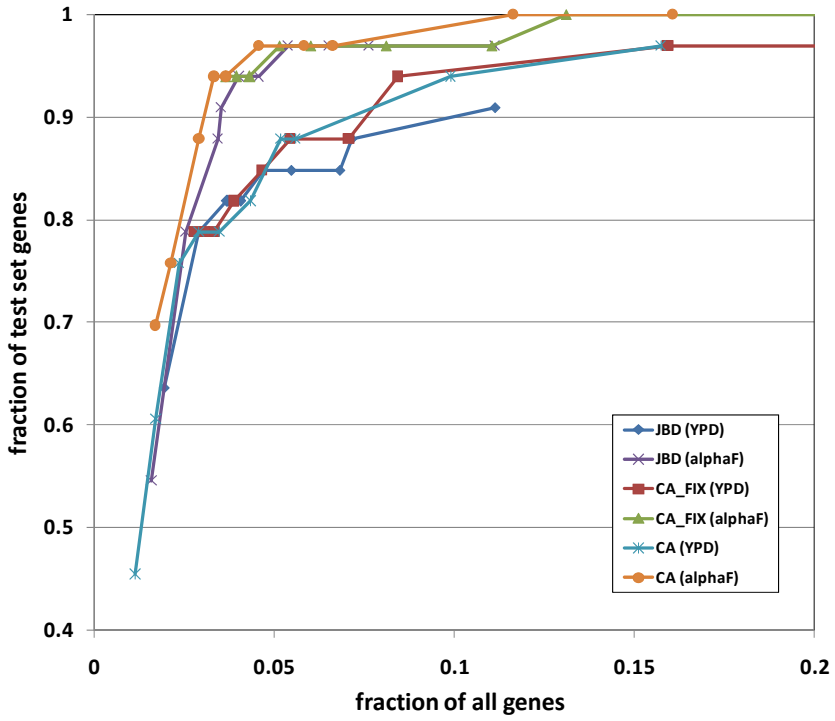


Figure S2. Comparison of the performance of ChIP Analytics (CA) and Joint Binding Deconvolution (JBD) on *S. cerevisiae* ChIP-Chip data.

Receiver Operator Characteristic (ROC) plots for the three analysis methods (CA, CA_FIX and JBD) on the ChIP-Chips of *S. cerevisiae* Mcm1 under two growth conditions (YPD and α-factor). See Supplementary Methods for further description.

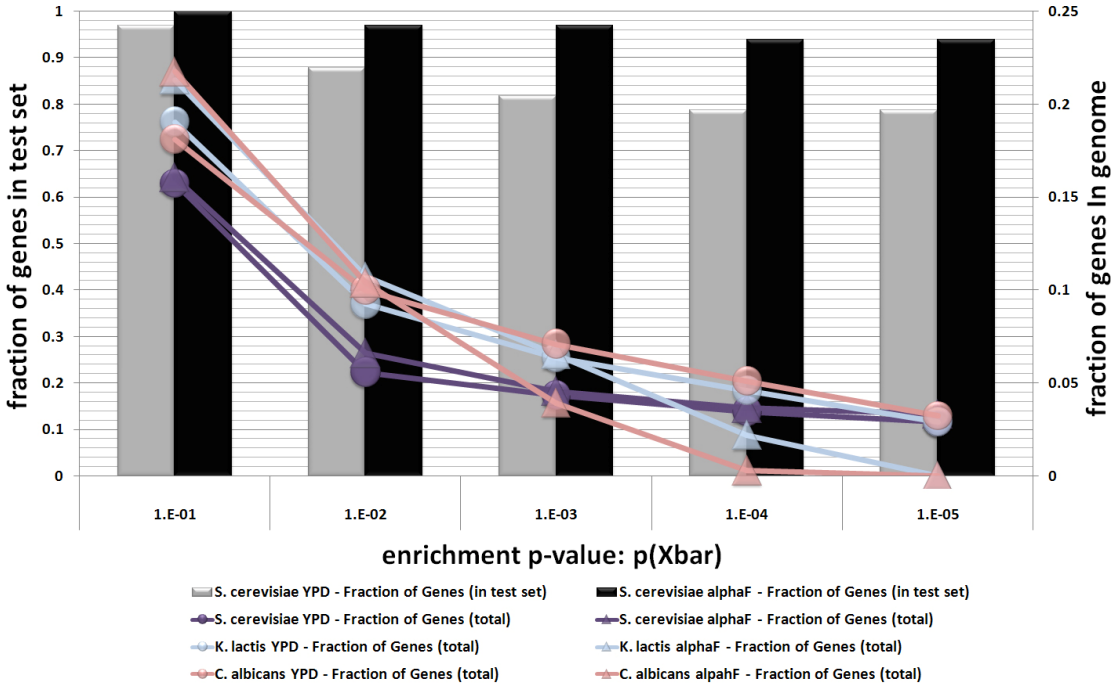


Figure S3. Results of ChIP Analytics (CA) on the ChIP-Chip data sets from all three species.

The enrichment p-value cutoff was varied (X axis) and the resulting number of bound genes called is recorded, both as a fraction of all test set genes in *S. cerevisiae* (left Y axis; silver and black bars) and as a fraction of all genes in each of the three genomes (right Y axis; pink, purple and blue lines).

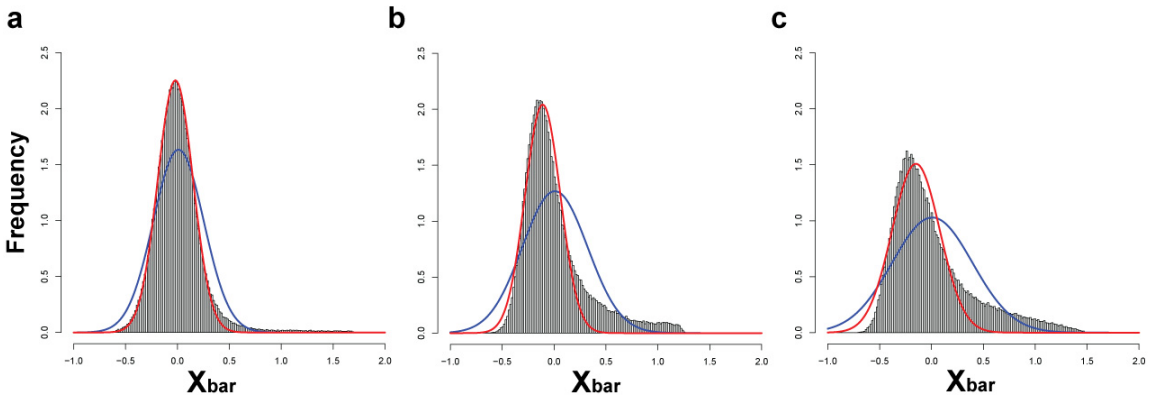


Figure S4. Distributions of the enrichment statistic (X_{bar}).

ChIP Analytics X_{bar} distributions for **a**, *S. cerevisiae*, **b**, *K. lactis* and **c**, *C. albicans* alphaF ChIP-Chip experiments. The blue line is the ChIP Analytics (CA) Gaussian fit and the red line is our attempt at an improved Gaussian fit (CA_FIX).

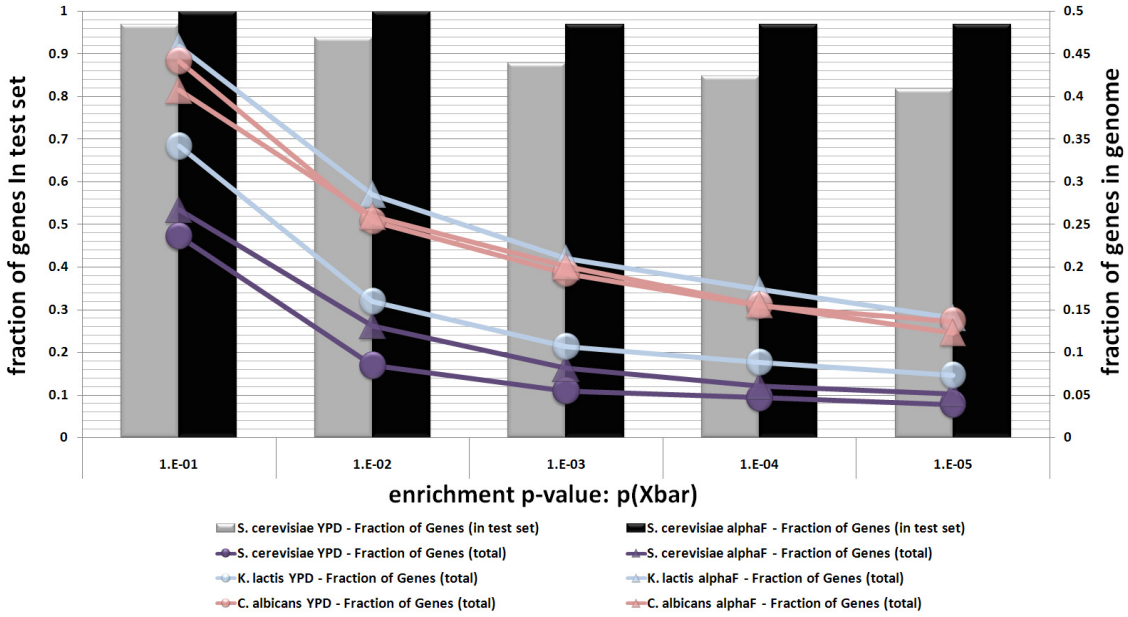


Figure S5. Results of the modified ChIP Analytics (CA_FIX) on the ChIP-Chip data sets from all three species.

The modified enrichment p-value cutoff was varied (X axis) and the resulting number of bound genes called is recorded, both as a fraction of all test set genes in *S. cerevisiae* (left Y axis; silver and black bars) and as a fraction of all genes in each of the three genomes (right Y axis; pink, purple and blue lines).

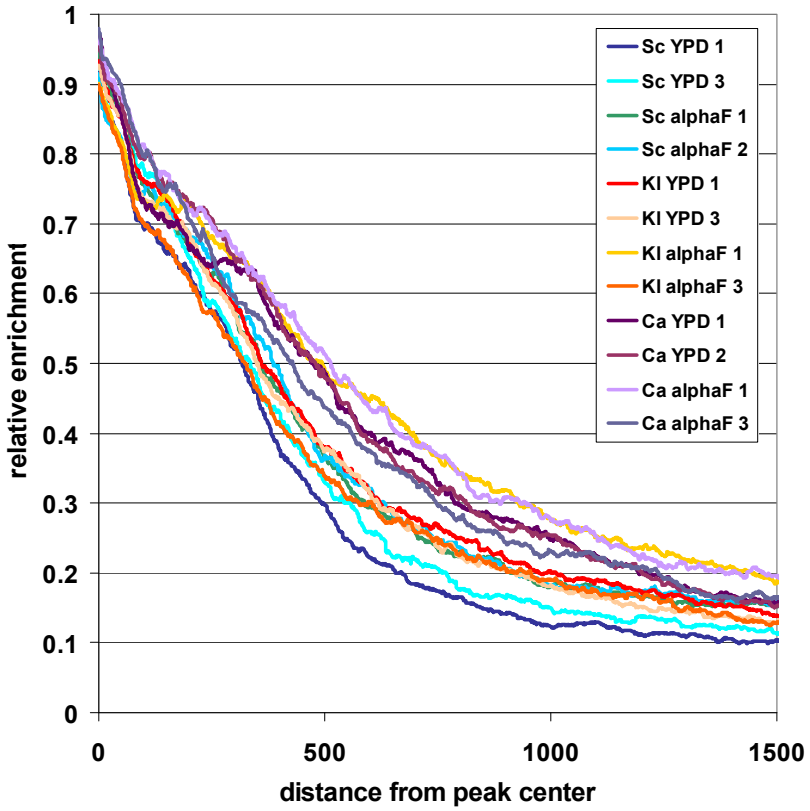


Figure S6. Estimated influence functions for each experiment.

For each experiment, we estimate an influence function as the average of the relative enrichment as a function of distance from the 50 strongest, idealized peaks in each experiment. Sc = S. cerevisiae, Kl = K. lactis and Ca = C. albicans

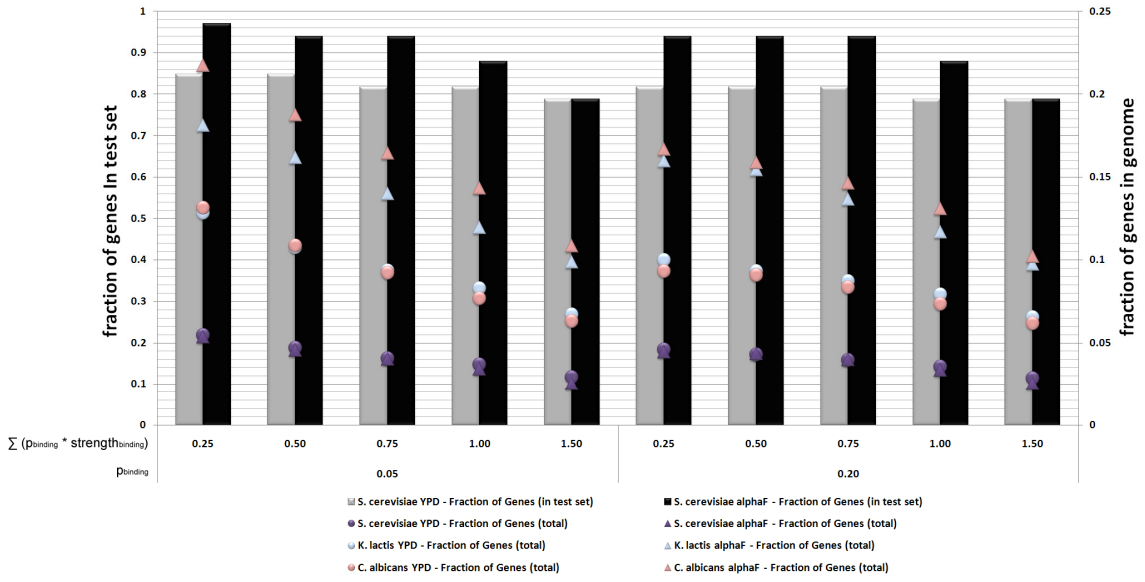


Figure S7. Results of Joint Binding Deconvolution (JBD) on the ChIP-Chip data sets from all three species.

The cutoffs for JBD statistics ($p_{binding}$ and $\sum[p_{binding}*strength_{binding}]$) were varied (X axis) and the resulting number of bound genes called is recorded, both as a fraction of all test set genes in *S. cerevisiae* (left Y axis; silver and black bars) and as a fraction of all genes in each of the three genomes (right Y axis; pink, purple and blue lines).

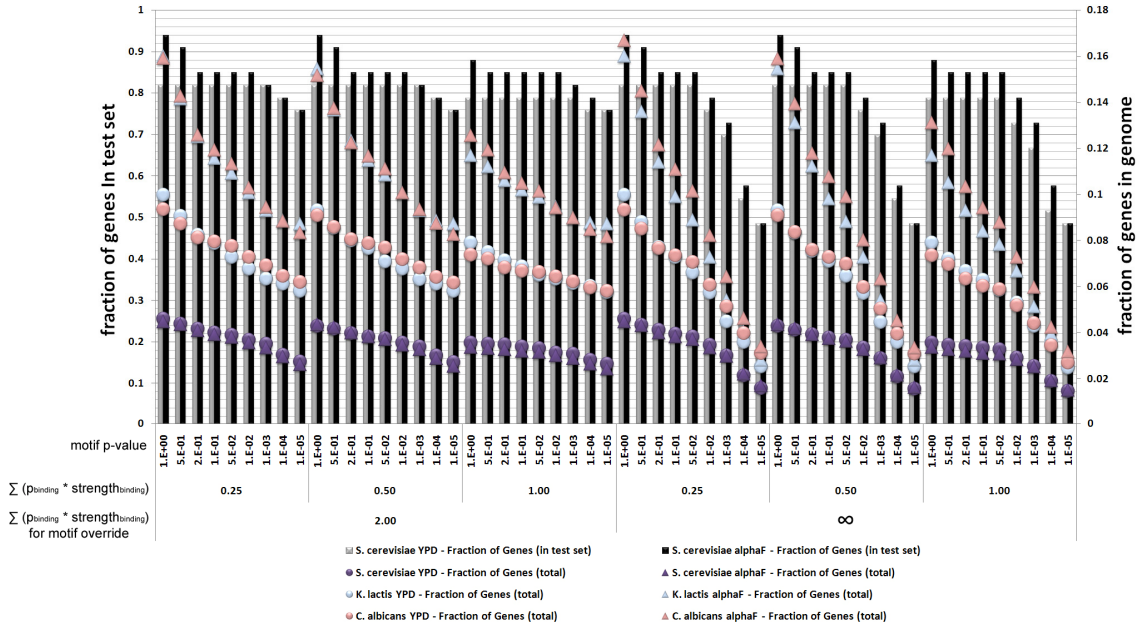


Figure S8. Results of Joint Binding Deconvolution (JBD) integrated with motif information on the ChIP-Chip data sets from all three species.

The cutoffs for the motif p-value and the JBD statistics ($\sum[p_{binding}*strength_{binding}]$ and $\sum[p_{binding}*strength_{binding}]$ for motif override) were varied (X axis) and the resulting number of bound genes called is recorded, both as a fraction of all test set genes in *S. cerevisiae* (left Y axis; silver and black bars) and as a fraction of all genes in each of the three genomes (right Y axis; pink, purple and blue lines). Here the cutoff for $p_{binding}$ is 0.2.

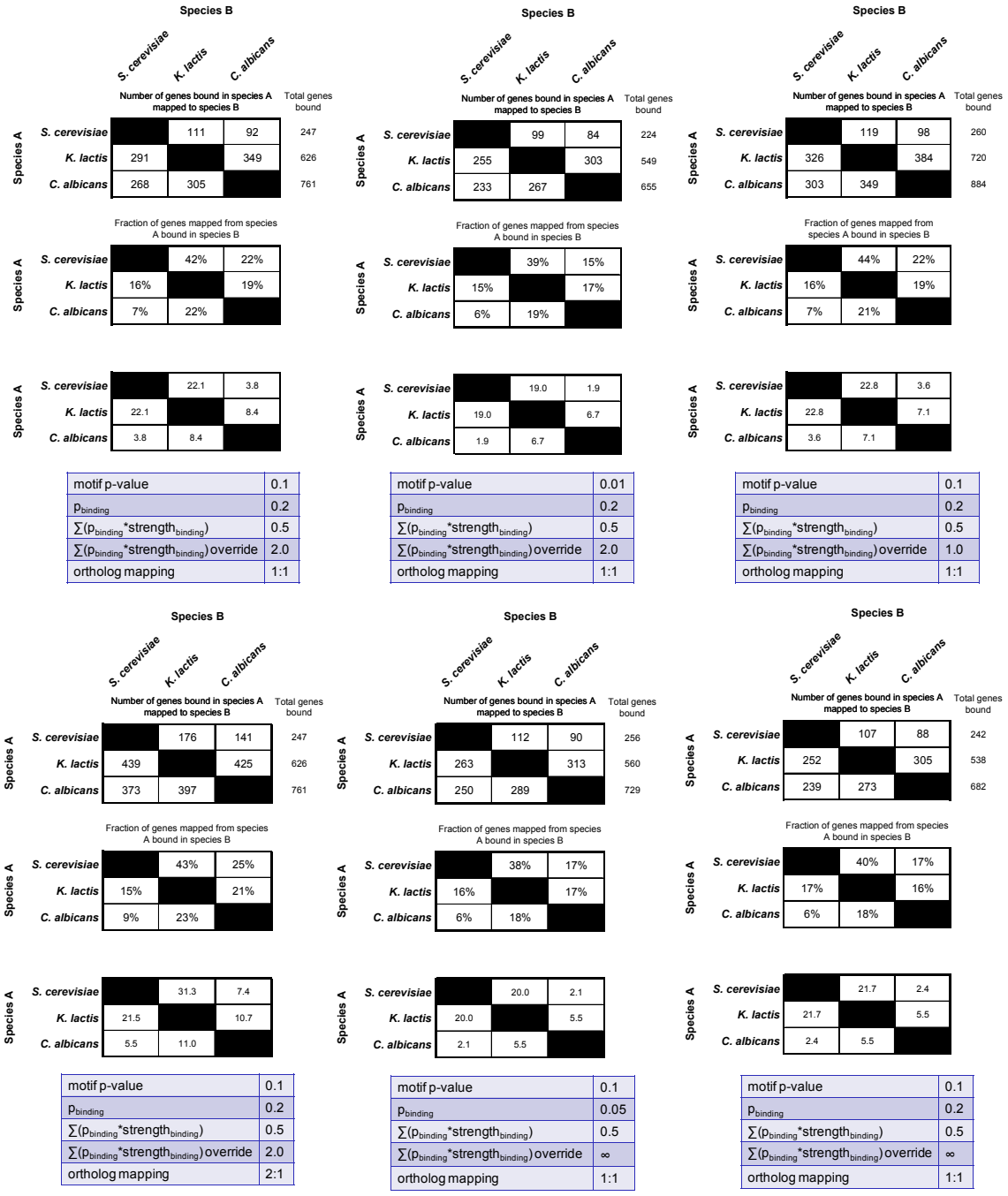


Figure S9. Robustness of pairwise species comparison results to parameter choices. Cutoffs for the four parameters which define the set of genes called as Mcm1 bound in each species were varied (shown in each blue table) and the results of the pairwise species comparison (described in detail in the Results section) were recomputed. The

first 3x3 table in each column indicates the number of genes bound by Mcm1 in each species A that can be mapped to one of the other two species B in a 1:1 manner. The second table indicates the number of genes mapped from A and also found to be in the Mcm1 bound gene set of B, as a fraction of the total genes bound in species A that can be mapped to species B. The third table indicates the significance (hypergeometric p-value) of each pairwise overlap.

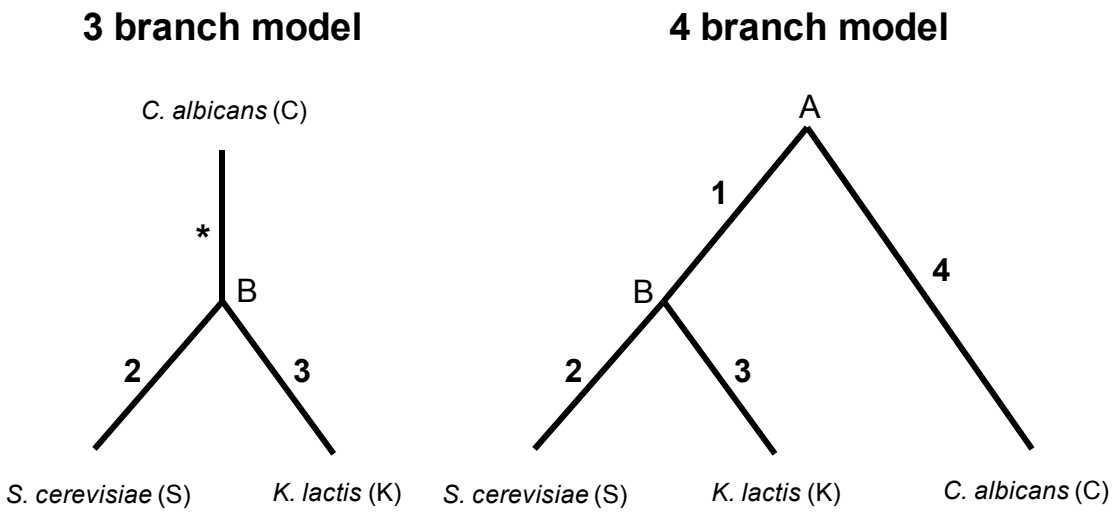


Figure S10. The three branch (star) and four branch, rooted three species tree models.

Chapter 4

Conclusions

The results presented in this thesis contribute to a growing field of enquiry into the evolution of gene regulation. While this field began by trying to explain how and to what extent changes in gene regulation have impacted the evolution of animal form ("evo-devo", see Chapter 1), it has quickly blossomed into a broader field with many actively pursued questions: How do changes in regulatory mechanism sweep through a large set of co-regulated genes[26, 31, 32]? How prevalent is non-adaptive circuit drift[123, 163]? Are gene expression differences between strains and species more often explained by changes in *cis*-regulatory sequences, in transcription factors, or concerted changes in both[41, 164, 165]? How many of the phenotypic differences among individuals of the same species are due to changes in gene regulation[15, 16]? Do special mechanisms promote loss and gain of *cis*-regulatory sequence[125, 166]?

My contributions to this growing field derive from taking a genomic approach, whereby I have inferred the evolutionary history of extinct yeast species from the results of experiments and comparative sequence analysis performed in extant yeast species. This work produced a model (proposed in Chapter 2), that explains how a set of co-expressed genes (the **a**-specific genes) has transitioned from positive control by a transcriptional activator to negative control by a transcriptional repressor, without losing proper co-expression in the process. Extending this basic approach to an expanded portion of the transcriptional network, I describe how a large combinatorial circuit made up of many interacting transcriptional regulators has evolved over the past few hundred million years (see Chapter 3). The history of this circuit has included the massive rewiring of both

protein-DNA and protein-protein interactions. The results described here, together with the results of others, yield the following three insights.

I. A transcription factor's target genes can change rapidly.

Although there are clear examples in which a transcription factor maintains a direct linkage with the same target gene (via the same or nearly the same *cis*-regulatory sequence) over long evolutionary times (see Appendix 1), these may be relatively rare compared with the large number of changes. For example, recent ChIP-Chip studies of four liver-specific transcription factors (FOXA2, HNF1A, HNF4A and HNF6) across 4,000 genes in mouse and human hepatocytes found that less than two thirds of genes are conserved as targets of each transcription factor[29]. A similar study of two transcription factors (Ste12 and Tec1) in three closely related yeast species (20 million years divergence) estimated that approximately 30% of the transcription factor-target gene connections seen in one species were preserved in the other two[28]. Only some of these differences could be attributed to loss and gain of *cis*-regulatory sequences, so it remains to be seen what other types of molecular changes contribute to this divergence. Although these studies clearly show evidence for a sizeable amount of turnover in transcription factor binding sites, a bit of caution should be taken in accepting the conclusion that so few sites are conserved between such closely related species. Results from failed ChIP-Chip experiments or ones in which the enrichment threshold for binding is set too low (allowing for a large number of false positives), would also be expected to give little

overlap between species. It seems this possibility was insufficiently explored in the Ste12/Tec1 study[28].

Our study examining combinatorial circuitry involving the transcription factor Mcm1 and its cofactors across three highly divergent yeasts (~300 million years divergence) also found evidence of massive rewiring[32] (Chapter 3). Only about 15% of the direct Mcm1 target gene interactions of *S. cerevisiae* were preserved in both *K. lactis* and *C. albicans*. As one might expect, conservation of targets was higher between the more closely related *S. cerevisiae* and *K. lactis* than between either of these species and *C. albicans*. We found that regions with significant enrichment in the ChIP-Chip experiments were very likely to contain an Mcm1 recognition sequence, making interpretation of the gain and loss of binding more straightforward than in the previous two studies. Mcm1 binds cooperatively to DNA with a set of cofactors to regulate many genes in each species, and the extensive rewiring observed was traced to high rates of gain and loss of *cis*-regulatory sequences as well as to the formation of new Mcm1-cofactor combinations and the breaking of old ones. The protein-protein interactions between Mcm1 and its cofactors are relatively weak; thus, it is conceptually simple to imagine how new partnerships could arise and how they might occasionally lead to the gain of binding at many new target genes.

II. The same set of co-expressed genes can be regulated by divergent mechanisms in different species.

As described in Chapter 1, stabilizing selection can maintain the expression pattern of a single gene, while still allowing for considerable "developmental drift" in the underlying mechanism of regulation[19, 20, 167]. Studies in yeast have extended this idea, uncovering examples where entire groups of co-expressed genes remain co-regulated in different species, while the relevant transcription factors and their cognate *cis*-regulatory sequences have changed. An important point, emphasized by many genomics studies, is that genes functioning in the same macromolecular complex or even just in the same process are often regulated as a group in response to changes in the environment (i.e., they are "co-expressed"). For example, in *S. cerevisiae* the presence of galactose induces transcription of genes that produce galactose-metabolizing enzymes via the transcriptional activator Gal4. In *C. albicans* (which last shared a common ancestor with *S. cerevisiae* some 300 MYA) the same enzymes are induced by galactose, but the Gal4 ortholog seems to have no role in this process; instead these genes appear to be controlled by *cis*-regulatory sequences recognized by a different transcription factor, and Gal4 ortholog regulates glycolytic enzymes[27].

In Chapter 2, I describe a similar example in the mating-type regulation of fungi: in the lineage leading to modern *S. cerevisiae*, regulation of the co-expressed **a**-specific genes (on in **a** cells and off in α cells) was handed off from a transcriptional activator (MATa2, an HMG-domain protein) to a transcriptional repressor (MATα2, a homeodomain

protein)[31]. Because the activator and repressor are expressed in opposite cell types, the overall logic of the circuit is conserved. The mating-type regulation system has been characterized extensively over the past three decades, providing us the data necessary to trace, down to the amino acid and nucleotide levels, the changes to protein-protein and protein-DNA interactions that likely underlie this transition.

These types of handoffs, in which the control of a set of genes is transferred from one regulator to another, may have occurred through an intermediate state in which the target genes came under dual regulation, thus preserving some form of control throughout the transition (Figure 1). Transition through a redundant intermediate has also been suggested for changes in the regulation of ribosomal genes in fungi (depicted in Figure 1, left pathway)[26].

It is not yet clear whether the rewiring of these co-expressed gene sets provides any advantage to the organism, as the overall regulatory pattern of the target genes seems, at least on the surface, to have remained constant. It is possible that these rewiring events have led to a quantitatively different induction of the galactose-metabolizing, ribosomal or **a**-specific genes[168], one that may have been positively selected at inception. It is also entirely possible that the rewiring of these gene sets yielded truly identical gene expression profiles or, at least, expression profiles that confer equal fitness to the organism and which could have evolved neutrally. It has been argued that many examples of transcriptional rewiring are not adaptive at all, but may simply reflect genetic drift[123, 163]. Whether or not adaptive evolution underlies these large-scale

changes, these examples clearly show the extreme degree to which transcriptional networks are plastic.

III. Cooperative binding of transcription factors (a form of combinatorial control) may facilitate circuit changes.

In its simplest form, the occupancy of two cooperatively binding proteins, A and B, on DNA is dependent on the concentration of each protein, the strength of each protein-DNA interaction, and the net favorable interaction between the two proteins. Because the system is cooperative, a decrease in any one of these parameters can, in principle, be compensated by a gain in any other. This would allow significant shifts in the relative contribution of each component to the overall energetics without destroying the regulation; this flexibility, in turn, could catalyze regulatory change. For example, the *cis*-regulatory sequence of B could drift away from consensus if the A-B interaction were sufficiently favorable (Figure 1, right path). This drift could produce a weak *cis*-regulatory sequence for a third transcription factor, C, whose expression might overlap that of B. If the A-C interaction was then strengthened by point mutation, the regulation of the gene would have changed from A-B to A-C through a series of small steps, none of which would destroy regulation of the gene. This single scenario is but one of many that is made possible by cooperative binding. If the number of cooperative components is increased, then the possibilities for "movement" in that system are multiplied.

Several studies have provided experimental support for these ideas. For example, in Chapter 2, I present evidence for a fungal mating circuit change that roughly follows the scenerio presented above (where A=Mcm1, B=MATa2, and C=MATα2)[31]. My analysis of the entire Mcm1-associated combinatorial network, in Chapter 3, indicates that gain and loss of combinatorial interactions may be relatively common. These two works, together, have now provided evidence for the gain of three interactions: Mcm1 with MATα2, Mcm1 with Rap1 and Mcm1 with Wor1. These studies have also shown loss of an interaction between Mcm1 and MATa2 and the loss of an interaction between Mcm1 and Arg81 that was preserved in an Mcm1 duplicate. The gain of a combinatorial interaction can be associated with the gain of many new target genes, as is likely the case with the interaction gained between Mcm1 and Rap1 at seventy ribosomal genes (Chapter 3, Figure 5). Or alternatively, the gain of a combinatorial interaction can be associated with relatively little change in the set of target genes, as seen with the interaction gained between Mcm1 and MATα2 at **a**-specific genes (Chapter 2, Figure 4; however, also see Appendix 2).

Further evidence, for the role of combinatorial interactions in facilitating circuit evolution, comes from a whole-network analysis of *S. cerevisiae*'s transcriptional circuitry[169]. Here, the authors found a strong correlation between the number of transcription factors that regulate a gene and the fuzziness (departure from consensus) of the *cis*-regulatory sequences present at that gene. This fuzziness may reflect the cooperative binding of transcription factors to DNA or simply that, with multiple factors independently regulating a gene, the importance of any one is relaxed.

It has been argued, by Zuckerkandl, that the type of neutral (or nearly neutral) changes permitted by cooperative assembly of transcription factors on DNA has catalyzed the formation of complex regulatory circuits[129]. While his conjectures do not align exactly with the conclusions and proposals made here, it is clear that Zuckerkandl anticipated the possibility that neutral networks in gene regulation space facilitate the evolution of novelty and complexity. Analogously, a theory relating neutral sequence networks to the evolvability and robustness of RNA structure has been developed, and, in this case ideas and relationships have been more formally examined[99].

Combinatorial regulation and cooperativity may be especially important for changes in the regulation of entire sets of co-expressed genes (as opposed to single genes). Here the dilemma is: how can changes in regulatory mechanism sweep through a large set of co-expressed genes? In addition to the pathways of Figure 1, an even simpler scenario is imaginable: the gain of a protein-protein interaction between transcription factors may "jumpstart" the rewiring of an entire set of genes at which one factor is already present (Figure 2). The development of a protein-protein interaction between Mcm1 and Rap1 is one way to explain how Mcm1 initially evolved binding at seventy ribosomal genes in the *K. lactis* lineage[32] (Chapter 3), without disrupting the co-expression of these genes. Afterwards the new circuit could be improved, target gene by target gene, through the gradual formation of optimal *cis* regulatory sequence. One concern with this model is that the interaction between Mcm1 and Rap1 is likely to be symmetric, and thus, after the gain of a Rap1-Mcm1 interaction, Rap1 might also be found at Mcm1's preexisting target

genes. Alternatively, it is possible that co-expression of the ribosomal genes was not maintained in the transition from regulation without Mcm1 to regulation with Mcm1. For instance, if the addition of Mcm1 regulation to these genes had the effect of repression in a condition where ribosome production was disadvantageous (e.g., a new stress condition), then the addition of Mcm1 regulation to *each* of the seventy ribosomal genes provides an additional benefit—namely, that of not wastefully producing that particular ribosomal gene transcript.

Future Directions

My studies of the Mcm1-associated transcriptional networks across yeast species have led to at least as many questions as they have answers.

How does transcription factor specificity evolve?

We observed that the putative Mcm1-α1 recognition sequences upstream of α-specific genes in *Candida* species are substantially different from those upstream of α-specific genes in *Saccharomyces* species. This suggests the readily testable hypothesis that the specificity of MATα1 has changed in this lineage. There is no evidence for a duplication of MATα1 in this same lineage, so the change in specificity likely occurred without duplication. If this is the case, it is interesting to consider whether the small size of this regulon (~3 target genes across hemiascomycetes) was required for such a change to take place. It is also interesting to consider whether such a change was facilitated by a

cooperative interaction with Mcm1, and to what extent combinatorial interactions can facilitate such changes. In other words, could the specificity of a regulator that binds 10 or 100 times as many genes also change substantially?

How do transcription factors diversify after duplication?

I outlined, in Chapter 3, the duplication of Mcm1 and the correlated changes that occurred to the sequence and structure of Mcm1, the control mechanism of **a**-specific genes and the control mechanism of arginine metabolic genes. Specifically, it appears that after the duplication of Mcm1, both sub- and neo-functionalization occurred. One paralog, Arg80, apparently kept all or part of its ancestral role in regulating arginine metabolic genes, while the other paralog, Mcm1, maintained control of all the other ancestral regulons. Consistent with Mcm1 maintaining most of the ancestral regulatory roles, it is clear, from sequence alignments of Mcm1 and Arg80 orthologs, that Arg80 is the more derived/divergent of the two paralogs. Across the hemiascomycete lineage only 14 substitutions to the MADS box domain of Mcm1 are observed, and, of these, 7 apparently occur on the branch where the duplication is inferred to have happened (Chapter 3, Figure 6). Amazingly, 6 of these 7 substitutions occur at residues which contact MATα2 in the Mcm1-MATα2 crystal structure (altogether 19 of 95 residues in Mcm1's MADS box domain contact MATα2). I homology-modeled the Mcm1-MATα2 structure of *K. lactis* (a species that diverged before duplication of Mcm1), and from this model speculated that the interaction between Mcm1 and MATα2 was strengthened at the time of duplication. One more correlated event is also of relevance; the branch on which

duplication occurs is the same one on which MATa2, the positive regulator of **a**-specific genes is lost.

In Chapter 3, I did not draw very strong conclusions from these observations because there are some unresolved discrepancies between our *in vivo* data and the *in vitro* data of others, regarding the control mechanism at the arginine regulon. However, if one could sort out these issues, then the Mcm1 duplication event could be an excellent system for studying the diversification of transcriptional regulators following duplication. Due to the relatively small number of substitutions occurring after duplication and the high degree of functional characterization of these residues, one might even imagine mapping out the entire post-duplication evolutionary pathway (substitution-by-substitution). In a recent paper[126], Hittinger et al. study the regulatory neo-functionalization of GAL1/3 after duplication. The authors investigate changes to the promoter regions of each duplicate that have led to a re-optimization of expressions levels since duplication. So the focus there is on upstream changes (primarily changes in *cis*), whereas here one has the opportunity to study the downstream effects of transcription factor duplication and diversification.

For anyone considering this challenge, it is worth noting that both *S. castellii* and *C. glabrata* have three recent paralogs of Mcm1. So it is likely that Mcm1 was actually triplicated on the branch leading to *S. cerevisiae*, *S. castellii* and *C. glabrata*, and that on the subsequent branch to the *Saccharomyces sensu stricto* species, one of these three paralogs was lost (see Appendix 3, Figure 9).

How does a developmental switch evolve?

In Chapter 3, I discussed our discovery that Mcm1 binding sites have very recently been gained at over one hundred genes in the *C. albicans* lineage. This change is of note, not only for its very recent occurrence, but also because many of the genes at which Mcm1 is bound are relevant to interactions with the human host, and at these genes Mcm1 is found binding a non-canonical *cis* regulatory motif (Chapter 3, Figures 1 and 7). Mcm1 binding at the non-canonical motif occurs upstream of several genes functioning in biofilm formation, as well as three out of four known regulators of the white-opaque developmental switch. Given the very recent appearance of Mcm1 at these genes, I began to wonder what impact Mcm1 regulation was having on these processes, and whether the gains of such regulation have adapted *C. albicans* to its human host. More generally, I began wondering how a developmental switch evolves and whether the white-opaque switch had evolved recently enough that one might still be able to trace its origins with extant organisms. The details of this ongoing endeavor are presented in Appendix 3. Although the results so far look promising, the evolution of this developmental switch and other linked developmental processes (e.g., hyphal growth) may be quite complex.

How does mating-type come to control non-mating processes?

Mating-type in *S. cerevisiae* can be regarded as a simple model for cellular differentiation[170]. In *S. cerevisiae*, **a** and α cells are differentiated to a limited extent; most genes differentially expressed between the two cell types are directly related to the process of mating (e.g., pheromone and pheromone receptors). It is interesting to consider the extent to which mating-type might expand to control processes not directly related to mating. For example, in *C. albicans* it was discovered that mating-type controls the white-opaque switch and the white-opaque switch, in turn, controls mating [30, 77, 118].

In *K. lactis*, we have begun to uncover evidence that MATα2 is regulating a large number of metabolic genes (see Appendix 2). MATα2 is apparently combining with MATa1 in **a**α cells to regulate genes functioning in carbohydrate metabolism. MATα2 is also apparently combining with Mcm1 in **a**α cells to regulate many other genes, including those functioning in "energy derivation by oxidation", fermentation and phosphate transport (Note: the word "apparently" is used because the definitive knockout controls have not yet been done). Although MATα2 is bound at these genes according to our ChIP-Chip assay, it is still unclear whether these genes are differentially expressed in the different cell types and whether MATα2 regulation is also active in α cells (though we do know that MATα2's transcript is expressed there). Related to this was my discovery that mating-type switching in *K. lactis* is media dependent (switching occurs on SD media, but not YEPD) and unidirectional (from **a** to α on SD). Is it possible that mating-type has come to control carbon metabolism and that different yeast cell types have radically different metabolic profiles?

Concluding Remarks

In conclusion, the results of my studies, as well as those of others, have revealed that transcription circuits rapidly rewire. Several distinct molecular mechanisms have been described, including proposals for how large sets of co-expressed genes can be rewired. In some cases the adaptive consequences of circuit rewiring are clear, though in most cases we can not yet say whether the force driving change is selection or drift. In coming years we should expect to uncover more of the general principles that underlie circuit rewiring and to further characterize the specific roles rewiring has played in the history of life.

FIGURES

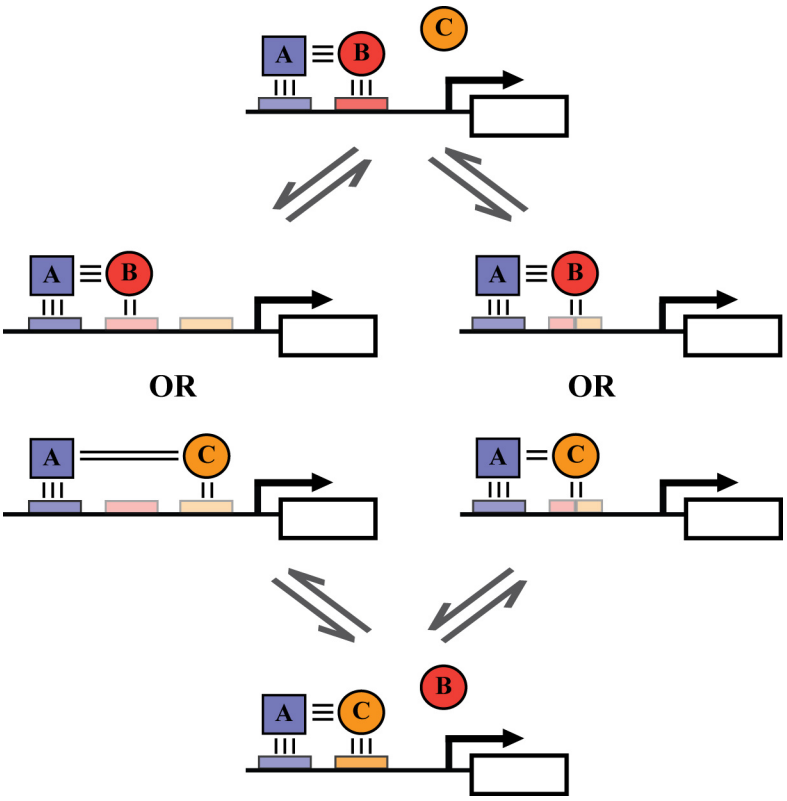


Figure 1. Pathways to the rewiring of combinatorial circuitry.

Illustrated are two of the many possible pathways by which regulation of a gene (or set of genes) can transition from control by the transcription factors A and B to that of A and C. In both pathways an intermediate stage exists in which regulators B and C may act redundantly. Small black lines represent protein-protein and protein-DNA interactions, the number of these indicating the strength of the favorable interaction. At any given time, each gene within a co-expressed set may have different control states (B only, C only, or B and C). The left pathway may be the route by which ribosomal genes and galactose-metabolizing genes were rewired in fungi[26, 27]. The right pathway is the likely route by which **a**-specific genes were rewired, also in fungi[31].

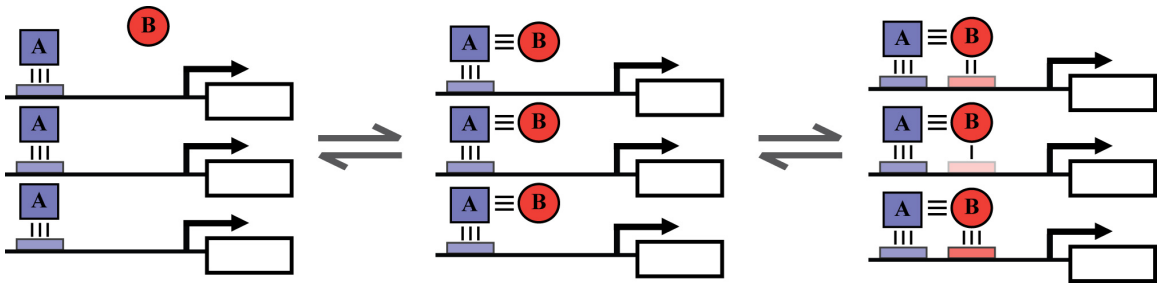


Figure 2. A plausible pathway to the concurrent rewiring of a large set of genes. According to this scenario an interaction is acquired between transcription factors A and B, after which interactions between B and DNA are optimized gene-by-gene.

Appendix 1

Mapping analogous regulatory elements across highly divergent species by translation

ABSTRACT

The growing wealth of genome sequence, gene expression and transcription factor binding data in diverse species provides us with the means to compare the *cis*-regulatory systems of one organism to another. Such comparisons will allow functional annotation of uncharacterized regulatory systems in one organism based on the knowledge of others, and may also shed light on how such systems evolve. Here, exploring an analogy between language and genomes, we develop an algorithm named "Metamorphosis" to identify analogous regulatory elements in diverse genomes. The algorithm is based on the assumption that regulatory elements with analogous function in different species (or words with the same meaning in different languages) will tend to occur in similar contexts, even if they do not share similar sequence (or spelling). When applied to German and English versions of Kafka's "The Metamorphosis", the algorithm successfully translates English and German words. When applied to the promoter sequences of genes oscillating with the cell cycle or genes responding to heat shock, the algorithm successfully translates transcription factor binding motifs with analogous function between *S. cerevisiae* and highly divergent species, including *C. albicans, S. pombe, D. melanogaster* and *H. sapiens*.

INTRODUCTION

The identification of orthologous genes in highly divergent species such as yeast, worm, fly and human is a well-studied problem[171, 172], and its solution allows for the transfer of protein function annotations from one species to another[173, 174]. In contrast, the identification, across divergent species, of "analogous" cis-regulatory elements--elements which regulate a similar set of genes within the context of a similar biological process– has been attempted only quite recently[26, 96]. Solutions to this problem are clearly valuable, as they allow one to more fully exploit existing knowledge about a regulatory system in one species when trying to characterize its counterpart in another species. The notable slow progress on this problem can be attributed to the difficulty it poses. Due to both the rapid evolution of regulatory systems and fast divergence of promoter sequences[18, 19, 30], analogous regulatory elements are, in general, not expected to share sequence similarity, and thus comparative genome methods relying on the conservation of such elements[63, 121, 175, 176] may not be applicable.

Several plausible scenarios can lead to analogous regulation without sequence similarity. One possibility is divergent evolution, whereby a regulator and its cognate *cis* regulatory elements co-evolve. For sufficiently long divergence time, regulatory sequences could evolve beyond recognition while ancestral regulatory relationships, between a transcription factor and its targets, are maintained. We refer to such "analogous" motifs, related by descent from a common ancestor, as "orthologous". A second scenario is convergent evolution, whereby divergent organisms discover novel solutions to a similar

regulatory problem, such that orthologous genes are dynamically expressed in a comparable manner, but by means of non-orthologous regulators and thus non-orthologous binding motifs.

The observation that patterns of gene expression for orthologous genes are often correlated in species as divergent as yeast, worm, fly and human[177, 178] strongly suggests that analogous, and perhaps orthologous, regulatory mechanisms underlie many of the biological processes common to divergent species. For example, McCarroll et al. established that several categories of genes, such as genes functioning in mitochondrial energy generation, are coordinately regulated during ageing in worms and flies. A number of studies have also demonstrated that similar biological processes, such as cell division cycle in yeast and mammals, are regulated by analogous mechanisms[179], even if sequence similarity between the corresponding regulators and *cis* regulatory elements is not detectable. Together these studies suggest that analogous regulation in divergent species may be prevalent and that relationships between a regulator and its target genes can be more persistent than regulatory sequences themselves.

In this study, we develop an approach to map analogous regulatory elements between divergent species by exploiting the persistence of regulatory relationships. We reason that similar regulatory relationships imply the co-occurrence of analogous regulatory elements in the promoter regions of orthologous genes (Figure 1). To better illustrate the idea, we employ an analogy with language, wherein promoters correspond to paragraphs and regulatory motifs correspond to words. Analogous elements are then, different

spellings of a word with the same meaning in two different languages. The language analogy was employed previously to aid in the development of MobyDick, an algorithm designed to identify putative *cis* regulatory motifs. MobyDick was so named because of its ability to successfully build a dictionary of English words from a version of the novel **Moby Dick** in which all punctuation and spacing was removed[180]. Here, utilizing the concept of co-occurrence, we extend the analogy between language and genome sequence by building translation tables for words with the same meaning in different languages and binding motifs with similar function in divergent species (Figure 1). The assumption we make is that although analogous words may not share a similar spelling and analogous motifs may not share a similar sequence, their pattern of occurrence across paragraphs and upstream of genes will be better preserved.

Using English and German versions of **The Metamorphosis** we demonstrate that the algorithm can successfully build a German-English translation table without any prior knowledge of the true English and German dictionaries. This provides an initial validation of our approach. We then apply the translation algorithm to genes oscillating with the cell cycle in *S. cerevisiae*[130] and their orthologs in *C. albicans, S. pombe, D. melanogaster* and *H sapiens*, organisms thought to have diverged from one another on the billion year timescale[181]. In doing so we discover several orthologous and analogous binding sites previously demonstrated as important for cell cycle regulation in these organisms. Finally, we turn to the patterns of gene expression associated with heat shock response[182], again in *S. cerevisiae, C. albicans, S. pombe, D. melanogaster* and *H*

sapiens. Here, once more, we successfully validate our approach with existing data and also identify other motif associations not previously characterized.

RESULTS

Translations between German and English using two versions of The Metamorphosis

In order to establish a proof of concept for our algorithm we applied it to two versions of Kafka's short story **The Metamorphosis –** the original, German version of the text and an expert-translated, English version of the text. The German version contains 3742 unique words and the English version contains 3053 unique words. All punctuation and spacing were removed from the texts and a dictionary was built independently for each version using MobyDick[180]. The resulting English dictionary contains 2668 predicted words approximately 35% of which are exact matches to words from the English text prior to removal of spacing, and the German dictionary contains 2353 predicted words of which approximately 34% are exact matches to words from the German text prior to removal of spacing.

The two texts were then mapped via corresponding paragraphs. There are 96 pairs of paragraphs in total. Predicted words from the English and German dictionaries were then paired exhaustively and evaluated as putative translations based on their co-occurrence across paired paragraphs. The twenty most significant putative translations are listed in Table 1. Of the top twenty predicted translations, 12 are exactly correct and 6 are missing just one or two letters from one or both of the words in the pair, but remain recognizable as correct translations. The words in the remaining two translations are also related, but in a less straightforward manner.

To assess the quality of translations on a larger scale and to gauge the algorithm's ability to separate signal from noise, we applied a permutation test as described in Methods. Figure 2 shows a comparison of the distribution of word pairs with the lowest 10,000 overlap p-values, computed using either the true paragraph pairings or the permuted pairings. The plot indicates a clear separation of signal from noise and suggests a method for choosing meaningful word pairs based on tolerance for noise.

Upon viewing the ranked list of predicted translations we wondered whether there was enrichment for "true" words (i.e., words exactly matching those words in the corresponding text prior to removal of spacing) at the top rankings. One would expect this is to be the case if the algorithm successfully selects for correct translations. Indeed this is the case (Table 2). Whereas the original English and German dictionaries are only 35% and 34% accurate the top 200 ranked word pairings are 2 fold more accurate (73% and 66%, respectively).

Translations associated with the cell cycle between *S. cerevisiae* and *C. albicans, S. pombe*, *D. melanogaster* and *H. sapiens*

Having demonstrated how Metamorphosis can translate words between languages, we turned to the more difficult translation of binding motifs. Because the process is ancient and the experimental data abundant, we chose the cell cycle as the basis of our first attempts at translation. We selected *S. cerevisiae* as one of the two species to compare because transcriptional regulation in this species has been extensively studied[183, 184]. The

cell cycle is an appropriate choice of biological process not only for the detailed experimental characterization that exists[179], but also for the recent microarray experiments surveying genome-wide mRNA abundance over the course of the cell cycle[130]. The other species paired with *S. cerevisiae (C. albicans, S. pombe*, *D. melanogaster* and *H. sapiens)* were chosen for their varying degrees of divergence and the availability of extensive experimental characterization.

Here we define our book as the set of promoter sequences for genes which oscillate with the cell cycle in *S. cerevisiae*, as determined by whole genome expression profiling[130]. There are 800 such genes, which were then mapped via orthologous relationships to each of the four other species. Genes which could not be mapped from *S. cerevisiae* to any one of the other species were subsequently removed from the analysis of that particular species pair. The net effect is that the size of the *S. cerevisiae* book varies somewhat, depending on the species with which it is paired. The resulting book pairs contain 408 (*S. cerevisiae—C. albicans*), 507 (*S. cerevisiae—S. pombe*), 145 (*S. cerevisiae—D. melanogaster*), and 156 (*S. cerevisiae—H. sapiens*) orthologous gene pairs respectively. The text of these books contain, for each selected gene, the 600 nucleotides found immediately upstream of the translational start site in that particular species.

As before, we employed MobyDick to build dictionaries of presumptive binding motifs independently for each set of upstream sequences. We also applied a clustering algorithm to each dictionary to group very similar sequences, which likely represent variants of the same binding motif. With language, this step was unnecessary as the

spelling of words tends to be more rigidly defined. The resulting dictionaries range in size from 80 to 186 clusters of motifs.

Predicted clusters of motifs from each pair of dictionaries (e.g. *S. cerevisiae* and *C. albicans)* were then paired exhaustively and evaluated as putative translations based on their co-occurrence across orthologous promoter pairs. As with the language translation example we randomly permuted the pairing of orthologs to assess the signal to noise ratio, this time accounting for the many-to-many relationships that can exist between orthologs (Methods) (Figure 3a-d).

As one might expect, we found that the degree of separation between distributions is dependent on the relatedness of the two species compared. For the comparisons between *S. cerevisiae* and *C. albicans* and between *S. cerevisiae* and *S. pombe,* there are many putative translations which are more significant than any of the motif pairs generated from permuted data. For comparisons between *S. cerevisiae*, and the more diverged *D. melanogaster* and *H. sapiens*, the separation between actual and permuted motif pairs is less apparent. While they may not carry strong statistical support due to the high level of noise generated, it is nevertheless informative to examine the top ranking motif pairs, as we still expect the algorithm will rank true pairs at the top of the list.

Instead of describing these pairs exhaustively, we will focus on a few of the top ranking translations for which some prior experimental characterization exists (Figure 3e). The complete data set, containing the exhaustive pairing of motif clusters and related statistics

for each pairwise species comparison, is available as Supplementary Tables S2-5 (http://genome.ucsf.edu/metamorphosis/). In the summary that follows we will generally discuss only those translations with p-values outside the noise range. Occasionally translations between *S. cerevisiae* and *D. melanogaster* or *S. cerevisiae* and *H. sapiens* with p-values falling within the noise range will be noted if there are compelling reasons to do so (e.g. if a translation ranks highly and the motifs either have strong enrichment statistics or some experimental characterization).

In the first row of Figure 3e we note our first biological validation: CGCGT (*S. cerevisiae*) to TCGCGTCGCS (*S. pombe*). CGCGT is the binding site for the Mbp1-Swi6 (MBF) complex in *S. cerevisiae*[185], termed the MCB box, and TCGCGTCGCS appears to be the dimer version of this binding site for the Mbp1-Swi6 ortholog in *S. pombe*[186]. Furthermore, Mbp1-Swi6 is a key cell cycle regulator of the G1-to-S phase transition in both species[186, 187]. Although only indirect experimental validation exists[188], the translation between *S. cerevisiae* and *C. albicans* indicates that the MCB box has also been conserved in the *C. albicans* lineage. It is encouraging to see that these motif sequences are related, even though this relatedness was not a criterion for selection and even though neutral mutations have fully saturated on this timescale.

Interestingly, the top ranking translations for the MCB box in *D. melanogaster* and *H. sapiens* look neither like the MCB box in the yeast lineage (ACGCGT) nor like one another (ATCGATAG and CGCCGCG, respectively). However, there is convincing evidence that the *D. melanogaster* motif, ATCGATAG, is fully functional and governs a

similar regulatory process. This motif is identical to those bound by Dref[189], a transcription factor required for normal DNA replication during the *D. melanogaster* cell cycle[190]. Among the known targets of Dref is E2f, another transcriptional regulator of the cell cycle in *D. melanogaster*[191]. The motif from *H. sapiens*, CGCCGCG, is more difficult to interpret, but is found in the flanking regions of promoter sequence bound by P53 and NF-Y[192, 193], factors with relevance to cell proliferation in *H. sapiens*[194]. Intriguingly, Yun et al. showed that P53 exerts its repressive effects on Cdc2 via a mechanism that requires NF-Y. Perhaps the CGCCGCG motif is the binding site of an as yet unidentified factor in the process of cell proliferation or apoptosis in *H. sapiens*.

Consistent with its involvement in the G1-to-S transition, the sets of genes in the overlap (see Methods) for each of these MCB box translations are significantly enriched ($p < 10^{-6}$) for at least one, and typically several, of the following biological processes or components: DNA replication (GO:0006260), DNA metabolism (GO:0006259), chromatin assembly or disassembly (GO:0006333), the chromosome (GO:0005694) and the replisome (GO:0030894).

While there is insufficient data to validate each of these putative cell cycle translations, a number of motifs identified by Metamorphosis correspond to the sequence specificity of experimentally characterized transcription factors. For example, one of the *S. cerevisiae* motifs identified by Metamorphosis, GTTTACT, is a close match to a binding site[106] and an exact match to the consensus sequence[184] for a complex formed between Mcm1 and Fkh2, which is critical to regulation of the G2-to-M phase transition[158, 195]. This motif

was paired with the *D. melanogaster* motif, CAGCACTG, characterized as the Sry-β binding site in *D. melanogaster*. Aside from the fact that it is differentially expressed during embryonic development[196], little is known about the regulatory processes governed by Sry-β. We find five genes in the overlap (RFA1, DUN1, ERV25, CDC54 and YTH1), but no clear theme is evident.

CACGTG, the characterized binding motif for both Pho4 and Cbf1 in *S. cerevisiae* [197, 198], is paired with GTTGGT in *S. pombe*. While the latter motif has not been characterized in *S. pombe*, in *S. cerevisiae* it flanks the core of one non-canonical Pho4 binding site (cacGTTGGTgc)[197]. For this translation we find strong enrichment for sulfate assimilation genes (GO:0003993; $p < 10^{-8}$) in the overlapping gene set. Cbf1 is typically associated with the regulation of methionine biosynthesis genes via the CACGTG motif, but O'Connell et al. has also demonstrated that Pho4 can functionally complement for Cbf1 in this regard, thus suggesting a link between phosphate and sulfate regulation[199]. Although Pho4 and Cbf1 are both members of the helix-loop-helix family of transcription factors, only Cbf1 has a clear ortholog in *S. pombe* (SPAC3F10.12c). The connection between methionine concentrations and regulation of progression through the cell cycle is known[200], but many details remain to be elucidated.

Translations associated with heat shock between *S. cerevisiae* and *C. albicans, S. pombe*, *D. melanogaster* and *H. sapiens*

To demonstrate the general applicability of Metamorphosis, we next considered the response to heat shock[201]. Heat shock is one of several stimuli that evoke the

environmental stress response (ESR) in *S. cerevisiae*, a response in which hundreds of genes associated with the ribosome, RNA metabolism and nucleotide synthesis (i.e., general growth processes) are repressed and hundreds of genes thought to help maintain homeostasis are induced[182]. Given its existence in many forms of life, this response may employ analogous regulatory mechanisms.

Whereas our book for the analysis of cell cycle contained promoter sequences for oscillating genes from *S. cerevisiae*, here we define our book as the promoter sequences for the 1301 genes that are positively or negatively regulated greater than two-fold 20 minutes after heat transfer from 30°C to 37°C[201]. Genes were then mapped from *S. cerevisiae* to the four other species via the pairwise ortholog table for each. Genes which could not be mapped from *S. cerevisiae* to any one of the other species were subsequently removed from the analysis of that particular species pair. The text of the *S. cerevisiae*, *C. albicans, S. pombe*, *D. melanogaster* and *H. sapiens* books contain, for each selected gene, the 600 nucleotides found immediately upstream of the translational start site in that particular species.

MobyDick was employed to build dictionaries of presumptive binding motifs independently for each book of sequences and a clustering algorithm was applied to each dictionary to group very similar sequences. The resulting dictionaries range in size from 88 to 213 clusters. Predicted clusters of motifs from each pair of dictionaries were then paired exhaustively and evaluated as putative translations based on their co-occurrence across ortholog pairs. As before, we randomly permuted the pairing of orthologs to

assess the signal to noise ratio (Methods). The distributions of overlap p-values for actual and permuted orthologous promoter pairs separate to varying degrees depending on the two species compared (Figure 4a-d).

As with the cell cycle analysis, we have chosen only a few of the most interesting translations to discuss (Figure 4e), attaching the exhaustive pairing of motif clusters as Supplementary Tables S6-9 (http://genome.ucsf.edu/metamorphosis/). We find that in each of the four pairwise comparisons the highest ranking or second highest ranking translation has as its *S. cerevisiae* motif, SAYCCRTACA, the characterized binding site for Rap1 (Figure 4)[109]. As one of its several functions, Rap1 activates ribosomal genes via this motif [202, 203]. Consistent with this is the finding that for all such translations, the genes in the overlap (see Methods) are significantly enriched ($p < 10^{-7}$) for processes such as protein biosynthesis (GO:0006412) or components such as the cytosolic ribosome (GO:0005830). Interestingly, the motifs with which the *S. cerevisiae* Rap1 motif is paired do not resemble the *S. cerevisiae* Rap1 motif. In *C. albicans* we find that it is paired with AGCCCTAA and in *S. pombe* with two experimentally characterized ribosomal motifs [204, 205], ARCAGTCACAG and ACCCTACCCTAG, the latter sharing a 5bp core with the *C. albicans* motif. In *D. melanogaster,* the *S. cerevisiae* Rap1 motif is paired with the Dref motif[189], CTATCGATAGTT, previously encountered in our cell cycle analysis where it was paired with the MCB box. Whereas the cell cycle translation was supported by experimental evidence, the connection between Dref and ribosomal protein regulation has not yet been established. Finally, in *H. sapiens* the Rap1 motif is paired with a palindromic motif, TCTCGCGAGA, which is required for activation of the

ARF3 gene[206]. While the relevant transcription factor remains unknown, Haun et al. proposed that this motif may serve as the binding site for a substitute activator at TATA-less promoters. Interestingly, Rap1 acts as one component of an activator which recruits TFIID to TATA-less ribosomal promoters in *S. cerevisiae*[207]. Taken together, these results suggest this motif is the *H. sapiens* equivalent of the Rap1 binding site.

A second *S. cerevisiae* motif, TGCGATGAGCTRA, contains a GATGAG core sequence known to be vital to the regulation of at least one ribosomal biosynthesis gene[208], and is paired with a similar motif from *C. albicans*, GATGAGATGAG. The GATGAG motif from *C. albicans* also translates to a different motif from *S. cerevisiae*, AAAATTTTTCA. The importance of this motif in regulation of *S. cerevisiae* ribosome biosynthesis has also been established[208]. As expected, the overlapping gene set for each of these translations is enriched for processes such as ribosome biogenesis (GO:0042254; $p < 10^{-31}$) and components such as the nucleolus (GO:0005730; $p < 10^{-22}$). A truncated variant of the *S. cerevisiae* motif, TTTTTCA, is translated to an *S. pombe* motif, ACAGTCACA, which is, as mentioned already, in turn paired with the *S. cerevisiae* Rap1 motif. The overlapping gene set for this translation is strongly enriched for protein biosynthesis genes (GO:0006412; $p < 10^{-43}$) and genes of the cytosolic ribosome (GO:0005830; $p < 10^{-77}$), but is also enriched, albeit more weakly, for ribosomal subunit assembly genes (GO:0042257; $p < 10^{-5}$) and ribosome biogenesis genes (GO:0042254; $p < 0.02$). That the translations form a network, rather than a simple one-to-one map, may be a result of combinatorial control of ribosome-related genes by multiple motifs.

A more detailed evolutionary analysis of ribosomal gene regulation within the ascomycete fungal lineage was recently published[26]. For the most part, our findings are consistent with these results. For example, we each find an AGCCCTAA motif upstream of the ribosomal genes in *C. albicans* and ACCCTACCCTA and CAGTCACA motifs upstream of the ribosomal genes in *S. pombe*. In the promoters of *C. albicans* ribosomal biogenesis genes, we each find a GATGAG-containing motif. In contrast to Tanay et al., who treat the ribosomal regulon and the ribosomal biogenesis regulon as distict entities, we observe that the CAGTCACA motif found upstream of ribosomal genes in *S. pombe* is also present at ribosomal biogenesis genes in this species.

DISCUSSION

We have described a method for uncovering orthologous and analogous relationships among *cis* regulatory elements from highly divergent organisms. This method does not require any prior knowledge of the elements and does not rely on sequence similarity when pairing elements from different species. Using this method, we predicted translations relevant to the cell cycles of *S. cerevisiae, C. albicans, S. pombe, D. melanogaster* and *H. sapiens*, successfully translating the known Mbp1-Swi6 binding motif from *S. cerevisiae* to orthologous motifs in *C. albicans* and *S. pombe* and to analogous motifs in *D. melanogaster* (Dref) and *H. sapiens* (the P53 and NF-Y related motif). Our translations make many experimentally testable predictions. For example, we predict that a motif from *S. pombe* that may be functionally equivalent to the Cbf1 binding site in *S. cerevisiae*. In our analysis of heat shock, translations related to the regulation of ribosomal proteins and rRNA processing (both associated with the ESR) were prominent. For example, we discovered motifs in *C. albicans, S. pombe, D. melanogaster* and *H. sapiens* that may fulfill a role in ribosomal regulation similar to that of Rap1 in *S. cerevisiae.* The pairing of several dissimilar motifs in this analysis hints at a relatively high degree of evolutionary plasticity underlying the ribosomal regulon[26].

As an initial validation for our translation procedure we extended the language analogy previously set forth in the original description of MobyDick. MobyDick is capable of building a *de novo* English dictionary using the text of **Moby Dick** with spacing and punctuation removed. The same algorithm is also quite accomplished at building

dictionaries of binding motifs using a large set of promoter sequences for co-regulated genes [180, 209, 210]. In each case it assumes that the text is generated by concatenation of words drawn from a probabilistic dictionary. Here we extended the analogy between language and genome sequence, by exploring the evolution of each. Using English and German versions of **The Metamorphosis** we found that functionally equivalent words in two diverged languages can be paired simply by assessing their occurrence across paragraphs. Similarly, orthologous and analogous binding motifs in divergent genomes, such as human, fly and yeast, can be paired by assessing their occurrence upstream of genes. In our language translations we found pairings, such as "Family" and "Familie", which, due to their related spelling, we infer are likely to share a common origin (i.e. they are orthologous words). Similarly, in the biological analysis we found motif pairs, such as *S. cerevisiae* CGCGT and *C. albicans* CGCGT, also related by descent from a common ancestor. Although we can not be certain, other translations in which the two words are spelled quite differently, such as "head" and "kopf", seem to be related through a distinct evolutionary mechanism. One possible scenario is that the ancestors to "head" and "kopf" were once used interchangeably, but that one word fell out of favor and was subsequently lost from each of the languages ancestral to English and German. Similarly, some of the ribosomal motifs, discovered in our analysis of heat shock, appear to be related by analogy rather than orthology. It will be interesting to further explore the analogy between genome evolution and language evolution.

In designing the translation algorithm, we assumed that while the nucleotide sequence of analogous motifs may not be similar, their pattern of occurrence upstream of orthologous

genes should show stronger correspondence. Our approach differs from that of Gasch et al. and Tanay et al. in two respects[26, 96]. First, there is no need to define many clusters of tightly co-regulated or functionally related genes. In fact, the algorithm is quite successful at building a translation table from the promoter sequences of a broad class of many genes (ranging in this study from ~100 to ~1000). Second, we have introduced an additional statistic, the overlap p-value, with which to assess the significance of a motif pairing between species. This measure is independent of the enrichment statistics employed by others and serves as a rigorous criterion for assessing the quality of a translation. Recently, Elemento et al. developed a method for identifying regulatory elements, also based on their cross-species co-occurrence patterns. While the assessment criterion is similar to ours, their method is primarily aimed at discovering regulatory elements that are fully conserved and its applicability is limited to closely related species (e.g. *S. cerevisiae* and *S. bayanus*)[176].

That our methodology seems to correctly identify motif translations indicates that our basic assumption about the conservation of motif occurrence patterns is valid. However, this approach relies on conservation of a particular type of regulatory structure and would therefore not be applicable to all regulons. For instance, recent work in *C. albicans* has revealed that some regulons are actually quite plastic, with only a small fraction of a transcription factor's targets conserved[30, 53]. In the future, it will be interesting to determine the extent to which transcription factor binding specificity and regulon structure have evolved and the precise constraints that guide each mode of evolution.

METHODS

Texts

The German and English versions of The Metamorphosis were downloaded from The Kafka Project website (http://www.kafka.org/).

Promoter sequences

For *S. cerevisiae, S. pombe, C. albicans and D. melanogaster*, sequences 600bp immediately upstream of the translational start site of all putative open reading frames (ORFs) were extracted from the genomic sequence provided by NCBI (for *S. pombe* and *S. cerevisiae*), Stanford Genome Technology Center (for *C. albicans*), and FlyBase (for *D. melanogaster*). For *H. sapiens*, sequences 600bp immediately upstream of the transcriptional (when known) or translational start site of all putative open reading frames (ORFs) in *H. sapiens* were extracted from the 2000bp sequences provided by the UCSC Genome Browser website.

We recognize that transcription factor binding sites in *D. melanogaster* and *H. sapiens*, such as those within developmental enhancers, often occur well upstream of the first 600bp or even in introns and other downstream sequence. However, Xie et al. recently established that conserved motifs in *H. sapiens* are strongly biased to the few hundred base pairs upstream of the transcriptional start site[211]. For the purposes of isolating the strongest signal we focus on this region.

Ortholog mappings

For *S. cerevisiae to S. pombe*, a human curated many-to-many ortholog map was graciously provided by Valerie Wood at the Sanger Institute's *S. pombe* Genome Project. The version we used was distributed on 4/3/2005 and is available from her upon request (val@sanger.ac.uk). For *S. cerevisiae to H. sapiens and S. cerevisiae to D. melanogaster*, computer generated one-to-one maps were produced via reciprocal best BLAST hit (RBBH) as described previously [177].

For *S. cerevisiae to C. albicans*, we were unable to locate a human curated ortholog map. However, we believed that due to the evolutionary nearness of the two species and the wealth of available, closely related genomes that we could improve on the standard RBBH methodology. We first compiled a protein sequence database for seven species (*S. cerevisiae, C. glabrata, K. lactis, E. gossypii, C. albicans, D. hansenii,* and *S. pombe*). All sequences were retrieved from the NCBI website, except for *C. albicans*, which was retrieved from the Stanford Genome Technology Center, and *C. glabrata*, *K. lactis*, and *D. hansenii*, which were retrieved from the Génolevures website[100]. We then ran PSI-BLAST for each *S. cerevisiae* query sequence against the compiled database, employing an E-value cutoff of 10^{-5} and the Smith-Waterman alignment option[82]. The sequences returned by PSI-BLAST were then multiply aligned with ClustalW (using the fast alignment option) and a neighbor joining (NJ) tree was inferred, again using ClustalW[72]. Finally, the resulting NJ tree was traversed to extract a set of orthologous genes in the following manner: Start at the leaf node for the query sequence and ascend the tree, incrementing a level counter for each node ascended. At each internal node descend. If a

leaf node is reached, the gene is from a species not yet seen at a lower level, and the branch length traversed is less than a cutoff (1.0), then add that gene to the set of orthologous genes. This procedure is repeated for each *S. cerevisiae* sequence. The resulting seven species many-to-many ortholog map (Table S10, http://genome.ucsf.edu/metamorphosis/) can then be reduced to a two species many-to-many ortholog map (Table S11, http://genome.ucsf.edu/metamorphosis/).

Filtering repetitive nucleotide sequences

Prior to building dictionaries of motifs, books containing promoter nucleotide sequence were filtered for repetitive sequence using Reputer with maxreplen parameter equal to 15 and minfraglen equal to 50[212].

Building dictionaries of putative words and binding motifs

Moby Dick was used to build dictionaries of putative words (for the language example) and binding motifs (for the biological examples) with default parameters[180]. For the language example, the appropriate 26 and 30 letter English and German alphabets were substituted for the default 4 letter nucleotide alphabet. Words (or motifs) that occurred too frequently (on average, more than one time per paragraph (or promoter)) or had a low quality factor (i.e., N / Xi statistic less than 0.1) were filtered. Filtering these words (motifs) serves to reduce noise during exhaustive word pairing (see "Assessment of Noise" section).

For the biological examples, putative motifs were clustered using a single parameter (nucleotide substitution rate = 0.5) scoring function and the CAST algorithm [213]. All reverse complement motifs not already present in each cluster were added.

For the purposes of compactly displaying motif sequences within Figures 3 and 4, we condensed several independent translations into a single pair of consensus motifs when motif sequences on each end of a translation were clearly extended or degenerate versions of motif sequences on each end of another translation. For each compacted set of translations we displayed only the rank and overlap p value for the lowest ranking member of the set. As an example, the following independent translations for heat shock between *S. cerevisiae* and *C. albicans*:

Rank	*S. cerevisiae* motif	*C. albicans* motif	-$\log_{10}$ (overlap p value)
1	aaaatttt---	gagatgag---	14.69
2	aaaatttt---	--gatgag---	10.94
3	aaaatttt---	----tgagatg	6.45
8	--aatttttca ---atttttca	--gatgagat-	5.49

were compacted into a single translation in Figure 4:

Rank	*S. cerevisiae* motif	*C. albicans* motif	-$\log_{10}$ (overlap p value)
1	aaaatttttca	gagatgagatg	14.69

Translation algorithm (Metamorphosis)

Words (or motifs) from the dictionaries built for two versions of the same book (e.g., **The Metamorphosis** in German and in English or e.g., sequences upstream of genes differentially expressed in response to heat shock in yeast and human), were paired

exhaustively and evaluated for their co-occurrence across paragraphs (or promoter sequences). See Figure 1. The significance of a word pair and therefore our confidence in it as a translation was assessed via its overlap p-value on the hypergeometric distribution:

$$p(O \geq o) = \sum_{O=o}^{\min(m,n)} \frac{\binom{n}{O}\binom{N-n}{m-O}}{\binom{N}{m}}$$

N = set of all paragraph (promoter) pairs

n = set of paragraphs (promoters) in the English (Yeast) text with the English word (Yeast motif)

m = set of paragraphs (promoters) in the German (Human) text with the German word (Human motif)

o = overlap of n and m as paired by corresponding paragraph (or orthologous promoter sequence)

Assessment of noise

We applied a permutation test to assess the effect that multiple non-independent tests (i.e. exhaustive word pairing) would have on our overlap p-value calculations. For each pair of words evaluated for co-occurrence across paragraphs we also evaluated a pair in which the true paragraph pairings were randomly permuted. For the language case this is straightforward, as the relationship between paragraphs in different texts is one-to-one. However, for the biological case the relationship between orthologous genes can be many-to-many (i.e., for the *S. cerevisiae* to *C. albicans* and the *S. cerevisiae* to *S. pombe* mappings), and it was therefore necessary to account for this mapping structure when permuting. This was accomplished by grouping each set of paralogs (i.e., genes within

one species which map to the same gene or group of genes in the other species) into a single block and then permuting the pairing of blocks, rather than the pairing of individual genes. As would be expected for multiple independent tests, we found that the permutation procedure yielded minimal permuted p-values generally equal to 1 / (# of word pairings). Thus, the overlap p-values, corrected for multiple independent testing, give an accurate estimate of the statistical significance of word pairings.

Functional bias of genes in the overlap

Another measure of the biological relevance of our translations is whether or not the translated motifs occur upstream of genes that take part in a common physiologic process or together form a common cellular component. To assess the functional relatedness of genes in our translations we applied Gene Ontology (GO) analysis as described previously[214]. For each translation, *S. cerevisiae* genes were chosen for GO analysis if their promoter contained at least one copy of the *S. cerevisiae* motif and the orthologous promoter from the second species contained at least one copy of the motif from that species. We refer to this set variously as "genes in the overlap" or "the overlapping gene set." All GO analysis p values referred to in this paper are corrected for multiple testing.

Enrichment statistics

Enrichment statistics measure the overrepresentation of a motif within a set of particular upstream sequences relative to the set of all upstream sequences in a given genome using the Poisson distribution. The set of all upstream sequences was used to estimate the per nucleotide rate at which the motif occurs. An enrichment p value was then computed on

the Poisson distribution based on the genome-wide rate of occurrence of a motif and the number of times it occurs within the particular subset of upstream sequences.

ACKNOWLEDGMENTS

The authors thank J. Chuang and J. DeRisi for useful discussions and comments on the manuscript and C-S. Chin and K. Kechris for computational support. B.T. is supported by an NSF graduate research fellowship and H.L. by an NIH grant, a Packard Fellowship in Science and Engineering and the Chinese National Science Foundation.

TABLES

Table 1.

The 20 most significant English-German translations derived from our analysis of **The Metamorphosis**.

English Word	German Word	Overlap $-\log_{10}$ p value	Correct Translation?	
mother	mutt	25.1	Mostly	Mutter
father	vater	22.4	Exactly	
sister	wester	21.3	Mostly	Schwester
sister	hwester	21.3	Mostly	Schwester
rents	eltern	18.2	Mostly	parents
family	familie	18.1	Exactly	
room	zimmer	17.3	Exactly	
door	tür	17.1	Exactly	
head	kopf	15.7	Exactly	
sister	diesc	15.5	Partially	die Schwester = the sister
almost	fast	15.1	Exactly	
legs	bein	14.7	Exactly	
window	fenst	14.5	Mostly	Fenster
gregor	gregor	14.3	Exactly	
chief	prokur	14.3	Partially	Prokurist = chief clerk
said	sagte	14.0	Exactly	
hands	hände	13.8	Exactly	
samsa	samsa	13.7	Exactly	
chair	sessel	13.6	Exactly	
morn	morg	12.9	Mostly	morning / Morgen

Table 2.

Enrichment for true English and German words amongst the top N ranked translations.

Top N Ranked Pairs	English	German
200	72.5%	66.0%
1000	61.5%	55.5%
5000	51.1%	48.2%
10000	46.0%	44.1%
6277804 (all)	34.7%	34.0%

FIGURES

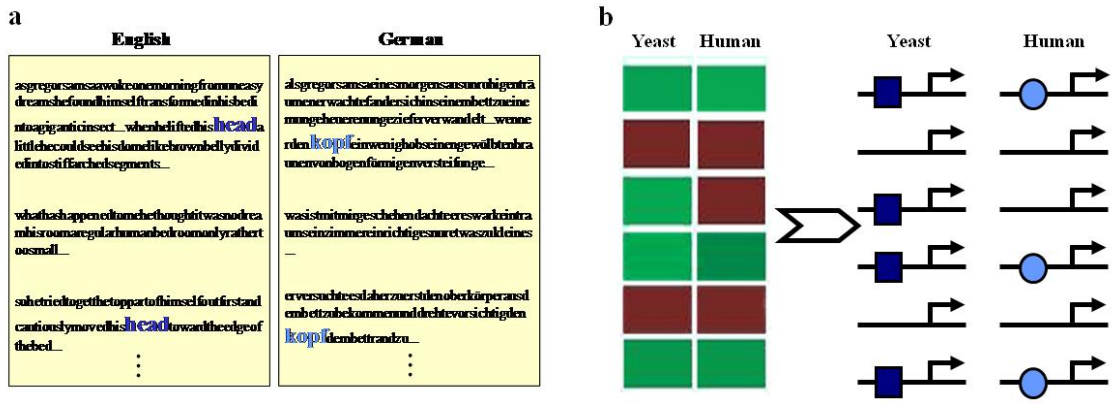


Figure 1. Overview of the translation methodology.

(a) Words from a text in two divergent languages can be translated (e.g. the English word "head" to German equivalent "kopf") by assessing their occurrence across corresponding paragraphs from the two texts.

(b) Binding sites regulating the expression of orthologous genes in two divergent species can be "translated" by assessing their occurrence upstream of those orthologous genes.

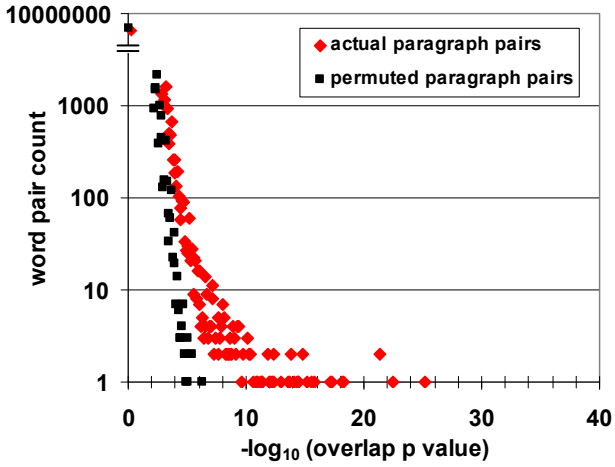
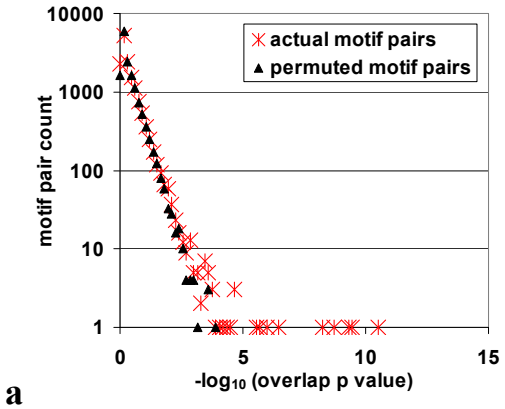
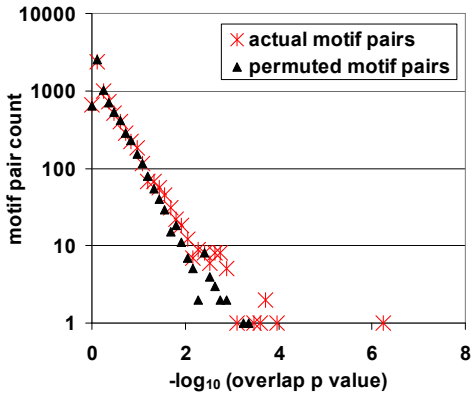


Figure 2. Assessment of signal-to-noise levels in the translation of words between English and German using *The Metamorphosis*.

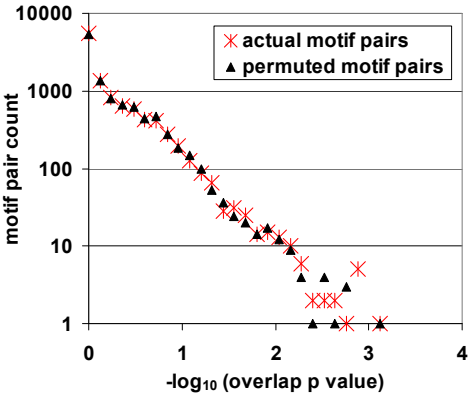
Distributions are shown for word pairs with the 10,000 lowest overlap p-values, computed using either the actual paragraph pairings (red diamonds) or permuted paragraph pairings (black squares).



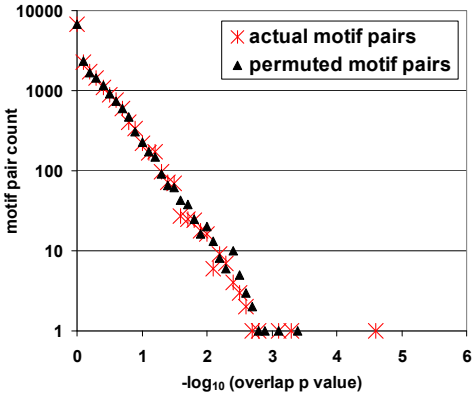
a



b

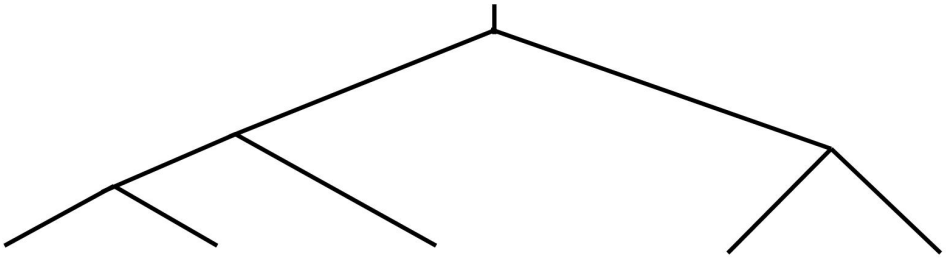


c



d

e

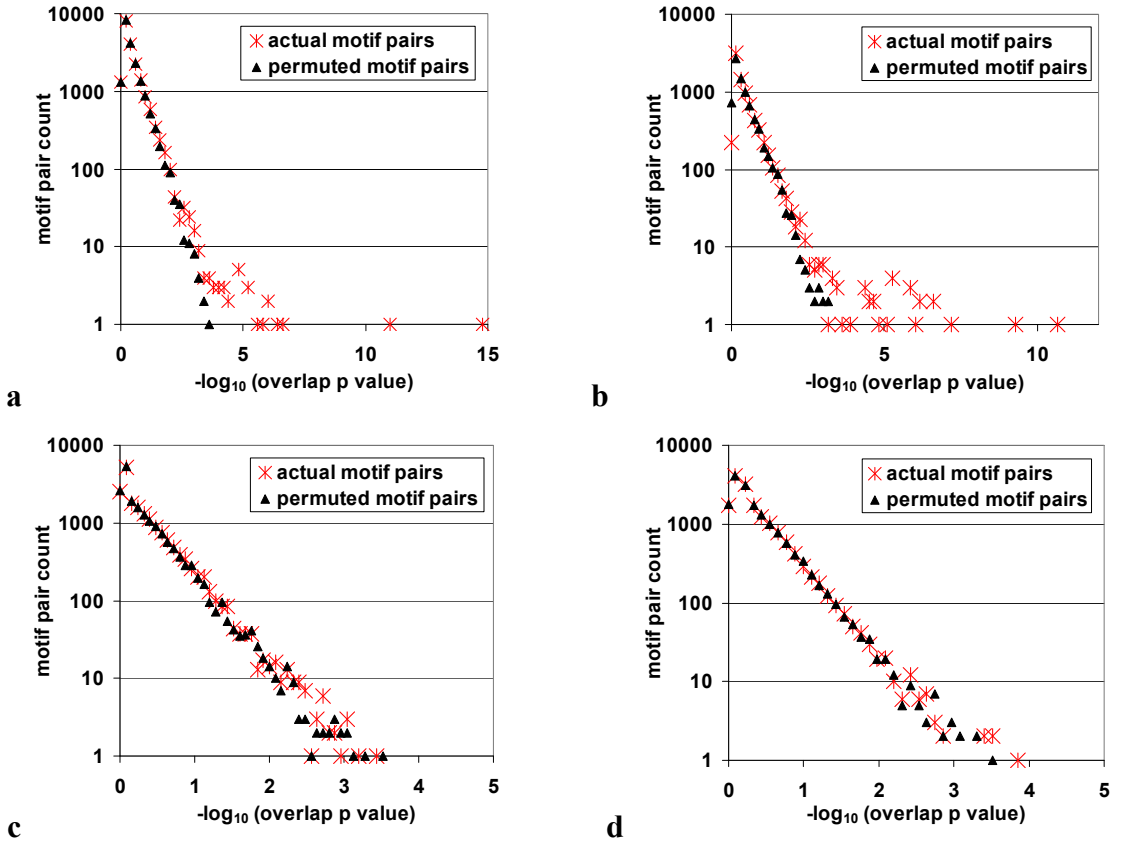


S. cerevisiae		*C. albicans*			*S. pombe*			*D. melanogaster*			*H. Sapiens*		
Motif	Factor	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)
cgcgt aaacgcgt acgcgwaa	MBF (Swi6 + Mbp1)	cgcgt	MBF	1 (10.44)	tcgcgtcgcs	MBF	1 (6.17)	accgcagc		7 (2.69)	cgccgcg	P53 NF-Y	11 (2.20)
		cgcg	MBF	2 (9.35)	cgtcgc	MBF^{Sc}	3 (3.70)	atcgatag	Dref	19 (2.22)			
ccctttt	Mcm1	tctat		10 (4.62)									
tctttct		tttcc	Mcm1^{Sc}	11 (4.60)									
gtttact	Mcm1 + Fkh2							cagcactg	Sry-β	3 (2.77)			
tatatat tatgtat tattgata	TBP (Spt15)	acacaaaa	Sum1^{Sc} Ndt80^{Sc}	15 (4.10)	tcgtcccta		4 (3.65)				ccaatc	NF-Y HiNF-B others	8 (2.47)
tctctt	Cup2				ccgttcca	Cup2^{Sc}	2 (3.88)						
cacgtg	Pho4 Cbf1				gttggt	Pho4^{Sc}	5 (3.59)						

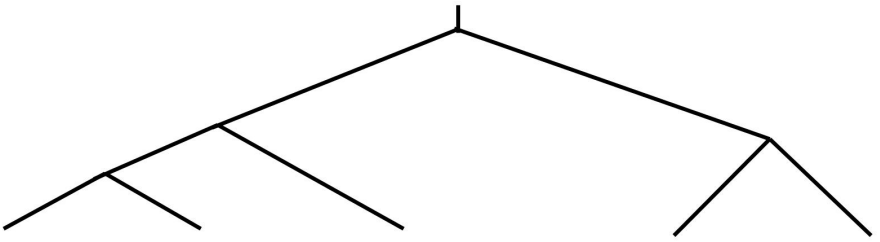
Figure 3. Summary of translations associated with the cell cycle between *S. cerevisiae* and *C. albicans, S. pombe, D. melanogaster* and *H. sapiens*.

(a-d) Assessment of signal-to-noise levels in the translation of motifs between (a) *S. cerevisiae* and *C. albicans,* (b) *S. cerevisiae* and *S. pombe,* (c) *S. cerevisiae* and *D. melanogaster* and (d) *S. cerevisiae* and *H. sapiens*. In each, distributions of overlap p-values for motif cluster pairs computed using either the actual paragraph pairings (red stars) or permuted paragraph pairings (black traingles) are shown.

(e) Table compiling selected motif pairs generated from our exhaustive analysis. For each *S. cerevisiae* motif listed in the leftmost column, if it exists, a translated motif, translation rank, and $-\log_{10}$ overlap p-value (OPV) is listed in the appropriate species column. Motif cells are shaded to denote the enrichment p-value for the most significantly enriched motif listed within the cell: red $p < 10^{-5}$, orange $p < 10^{-3}$, yellow $p < 10^{-1.5}$, and blue $p < 10^{0}$. When a motif was previously associated through experiment with a transcription factor it is recorded in the Factor column. ^{Sc} indicates that the study cited undertook experimental characterization of the motif in *S. cerevisiae*, rather than the species in which the motif was discovered. Please see Methods for a description of how the consensus motifs were derived.



e



S. cerevisiae		*C. albicans*			*S. pombe*			*D. melanogaster*			*H. Sapiens*		
Motif	Factor	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)	Motif	Factor	Rank (OPV)
sacccataca catccgtaca	Rap1	agccctaa		17 (4.35)	arcagtcacag	ribosomal	1 (10.50)	ctatcg	Dref	1 (3.41)	tctcgcgaga	?	2 (3.41)
					accctaccctag	ribosomal	7 (6.05)	tatcgatagtt	Dref	9 (2.78)			
aaaatttttca tttttca	ribosome biogenesis	gagatgagatg		1 (14.69)	acagtcaca	ribosomal	15 (5.16)						
tgcgatgagctra	ribosome biogenesis	gatgagatgag		4 (6.23)									

Figure 4. Summary of translations associated with heat shock between *S. cerevisiae* and *C. albicans, S. pombe, D. melanogaster* and *H. sapiens*.

(a-d) Assessment of signal-to-noise levels in the translation of motifs associated with heat shock between (a) *S. cerevisiae* and *C. albicans,* (b) *S. cerevisiae* and *S. pombe,* (c) *S. cerevisiae* and *D. melanogaster* and (d) *S. cerevisiae* and *H. sapiens*. In each, distributions of overlap p-values for motif cluster pairs computed using either the actual paragraph pairings (red stars) or permuted paragraph pairings (black traingles) are shown. (e) The table compiles selected motif pairs generated from our exhaustive analysis. For each *S. cerevisiae* motif listed in the leftmost column, if it exists, a translated motif, translation rank, and $-\log_{10}$ overlap p value (OPV) is listed in the appropriate species column. Motif cells are shaded to denote the enrichment p-value for the most significantly enriched motif listed within the cell: red $p < 10^{-5}$. When a motif was previously associated through experiment with a transcription factor it is recorded it in the Factor column. ^{Sc} indicates that the study cited undertook experimental characterization of the motif in *S. cerevisiae*, rather than the species in which the motif was discovered. Please see Methods for a description of how the consensus motifs were derived.

Appendix 2

Mating-type regulation in *K. lactis*

INTRODUCTION

In Chapter 2, I detailed our discovery that the pathway from positive to negative regulation of **a**-specific genes (**a**sgs) likely involved transition through a hybrid state in which both forms of regulation were employed. I predicted that this hybrid state was retained in modern-day *K. lactis*. I therefore attempted to test this prediction by experimentally characterizing the mechanism of **a**sg control in *K. lactis*. I also undertook a more general enquiry into the mating-type circuitry of *K. lactis*, in collaboration with Lauren Booth and Dave Galgoczy. All ChIP-Chips presented here were performed by Dave and subsequently analyzed by me. Gene expression profiling was performed by Lauren and me, with Lauren taking on the majority of the experimental work. All other experimental and computational work documented here is my own. This enquiry is not yet complete and is currently in the very capable hands of Lauren.

RESULTS

The regulatory transition at a-specific genes may also involve the gain of Ste12 regulation

After publishing the work describing a transition from positive to negative control of **a**-specific genes[31] (Chapter 2), I realized it might also be interesting to look for differences in Ste12 regulation of the **a**sgs. Ste12, the master regulator of the response to pheromone, is known to positively regulate **a**sgs (in addition to αsgs and a large set of general pheromone response genes) in *S. cerevisiae*. It is not known whether this regulation of **a**sgs exists in other yeast species as well. Though, it is known that the general pheromone activated genes (up-regulated in both **a** and α cells) of *C. albicans* also contain a Ste12-like response element (TGTTTSA) in their promoters[134]. I counted instances of the Ste12 motif (TGTTTSA and its reverse complement) in promoters of **a**sgs from each of three clades (Figure 1a; all promoters within a clade were pooled): those putatively implementing the purely negative form of **a**sg control (spanning *S. cerevisiae* to *C. glabrata*), the hybrid form of **a**sg control (spanning *K. lactis* to *K. waltii*) and the purely positive form of **a**sg control (spanning *C. albicans* to *D. hansenii*). Apparently, Ste12 motifs are enriched only at **a**sg promoters in those species implementing the purely negative form of **a**sg control (*S. cerevisiae* to *C. glabrata*). Given that the Ste12 motif is not likely to have diverged within the hemiascomycetes, this suggests that Ste12 activation of **a**sgs likely arose just prior to the divergence of *S. cerevisiae* and *C. glabrata*, concurrent with the putative switch from the hybrid to the "negative" mode of **a**sg control and the loss of MATa2. A simple test of this hypothesis

would be to perform gene expression profiling of *K. lactis* and *C. albicans* strains with Ste12 deleted. If Ste12 does not regulate **a**sgs in these two species, then we would expect to see no change in the expression of **a**sgs, which is the opposite of what was seen in *S. cerevisiae* (Figure 1b). This new analysis suggests a more complex transition from the hybrid to the "negative" form of **a**sg regulation. Specifically, these new data imply that loss of the activator MATa2 may have been compensated for by gain of positive regulation of **a**sgs by Ste12 (Figure 1c). Thus, the "purely negative" form of **a**sg control in *S. cerevisiae*, may actually be thought of as another hybrid form of control (with activation by Ste12 and repression by MATα2).

The putative hybrid form of a-specific gene control may differ by gene

One aspect of the putative hybrid form of **a**sg regulation (of the *K. lactis* lineage) that was not emphasized in our publication[31] is the heterogeneity of these **a**sg operators. A close examination of an alignment of **a**sg operators from the *K. lactis* lineage reveals that while some clearly have motif information specified on both sides of the Mcm1 motif (i.e., both sides closely match the consensus), others have information specified on only a single side (Figure 2b). Could this signify that some **a**sgs are regulated by just MATa2 or just MATα2 and that others are regulated by both? Furthermore, this feature (information specified on one side versus two) shows some conservation across orthologous **a**sgs (Figure 2c). For example, the **a**sg operators upstream of STE2 conform to the consensus sequence on both sides of their Mcm1 motifs in all four species of the *K. lactis* lineage. Given the great divergence of these four species, this suggests that there

may be some functional difference between operators with one flanking site and those with two, and that this functionality is under purifying selection at STE2.

Engineering *K. lactis* α cells to respond to α-factor

One strain, which was very useful for decoding the mating-type circuitry of *C. albicans* and which I felt was likely to be useful for this same endeavor in *K. lactis*, is a MATα that has its **a**-factor receptor gene (STE3) swapped for a gene encoding the α-factor receptor (STE2) (Figure 3a). The cells of this strain are then α cells that produce a pheromone response to α-factor, alleviating the need to synthesize or work with **a**-factor. I constructed this strain by first deleting MFα from yBT16 (MATα nej-), producing strain yBT19. I then transformed yBT19 with a construct containing STE2, URA3 and sequence targeting the construct to the STE3 locus. This produced strain yBT26, a MATα that responds to α-factor. At four hours of exposure to α-factor over 90% of cells grown in our standard *K. lactis* pheromone-response conditions (see Chapter 3, Supporting Methods) show a shmoo morphology (Figure 3b, bottom); the shmoo morphology resembles that seen when **a** cells respond to α-factor (Figure 3b, top). At 17 hours, engineered α cells (yBT26) continue to respond (Figure 3b, note the very long shmoos), whereas **a** cells do not (Figure 3b, note what looks to be retracting shmoos). This may be because **a** cells produce Bar1 (or at least we know that this transcript is up-regulated in **a** cells responding to pheromone), an **a**sg encoding a protease that degrades α-factor in *S. cerevisiae*, whereas α cells do not.

a1-α2 regulates different genes in *K. lactis* than in *S. cerevisiae* and *C. albicans*

In *S. cerevisiae*, MATa1 and MATα2 encode transcription factors which combine to repress "haploid-specific genes" in an **a**α diploid. The same is apparently true for the MATa1 and MATα2 orthologs of *C. albicans*, except in this case the term "haploid-specific genes" is a misnomer, because there is no known haploid state in *C. albicans*, rather these should be termed "MAT-homozygous-specific genes" (or something similarly cumbersome). For simplicity, I will refer to genes regulated by the a1-α2 heterodimer as "a1-α2 regulated genes." The a1-α2 regulated genes have been defined in *S. cerevisiae* and *C. albicans* through a combination of ChIP-Chip, gene expression profiling and sequence analysis[30, 120, 141] (see Figure 4a). There is considerable divergence in the a1-α2 regulated gene sets of *S. cerevisiae* and *C. albicans*. However, the pheromone response MAPK pathway is repressed by a1-α2 in both species (Figure 4b). While it is clear that both species repress the pheromone response pathway in the **a**α state, the exact components of this pathway targeted differs between the two species.

To determine the targets of a1-α2 in *K. lactis*, ChIP-Chip of a1 and α2 was performed on **a**α cells. Globally, a1 and α2 bind many of the same regions (Figure 5a). Surprisingly, these 58 regions have neighboring genes that are enriched with functions in carbohydrate metabolism. Although these two transcription factors apparently bind to many overlapping regions, only ten of these regions show evidence for the canonical a1-α2 *cis* regulatory motif (shown in Figure 5a; found by a MEME search of the 58 regions). It is unclear whether a1 and α2 binding to the remaining 48 regions occurs through some means distinct from that seen in *S. cerevisiae*, or whether perhaps there is some cross-

hybridization or technical issues common to both ChIPs. ChIP-Chip of a1 and α2 in strains deleted for one or both genes should help to clarify what is happening here.

The set of thirteen genes flanking the ten regions with both a1-α2 binding and the canonical a1-α2 *cis* regulatory motif are shown in Figure 5b. This set of a1-α2 regulated genes is almost entirely divergent from that seen in *S. cerevisiae* and *C. albicans*. The single gene in common with *S. cerevisiae* is RME1, which encodes a repressor of entry into meiosis. The repression by a1-α2 of this repressor serves to restrict meiosis to **a**α cells in *S. cerevisiae*; one could infer that this is also true for *K. lactis*. Genes encoding transcription factors (e.g. SWI5, TEC1) and DNA repair enzymes (e.g., RAD54, RAD19) make up a substantial fraction of the a1-α2 regulated genes in *K. lactis* (Figure 5b).

Genes encoding components of the pheromone response pathway are noticeably absent from this set of thirteen a1-α2 regulated genes. I wondered whether this was related to *K. lactis*'s apparent preference for a haploid lifestyle. Yeast species are generally thought to prefer either the haploid (**a** or α) or the diploid (**a**α) state. For example, wild *S. cerevisiae* haploid strains grown in rich media will mate to form diploids, which will then reproduce asexually. Thus *S. cerevisiae* is said to prefer the diploid lifestyle. Since, *S. cerevisiae* spends considerable time as a diploid (**a**α) it may be advantageous to keep the pheromone response pathway tightly down-regulated, thus avoiding improper mating activation, crosstalk with other MAPK pathways and/or the wasteful production of mRNA. One would imagine that these advantages would not be as great (or perhaps they would even

be detrimental) in a species where the haploid state (**a** or α) is preferred and the diploid state (**a**α) is only transitory (as in *K. lactis*).

Although ploidy preference is admittedly somewhat vaguely defined, I did find an interesting correlation between lack of repression of the pheromone response pathway and preference for the haploid lifestyle (Figure 6). It would be interesting to determine whether a1-α2 regulation controls ploidy preference in some manner, or whether this preference is encoded elsewhere and a1-α2 regulation of the pheromone response pathway is a downstream effect of this choice. Ploidy preference is apparently a very plastic trait in yeast species (Figure 6).

The picture emerging from a1-α2 binding data, now gathered in three species, indicates that a1-α2's target set evolves rapidly. I extended this analysis across a larger set of yeast species, by mapping the a1-α2 regulated genes of *S. cerevisiae, K. lactis*, and *C. albicans* to other species with fully-sequenced genomes, extracting promoter sequence and performing a motif search with an a1-α2 position-specific scoring matrix (PSSM). The results, displayed in Figure 6, are consistent with the idea that the set of genes, targeted by a1-α2, evolves rapidly; genes with a strong a1-α2 PSSM score in their promoter in one species do not often show strong a1-α2 PSSM scores in orthologous regions from more than one or two other species. Clear exceptions to this rule include RME1 and genes encoding components of the pheromone response pathway (discussed above).

α2 and Mcm1 bind at many of the same regions in *K. lactis*

The effort to decipher the mating-type circuitry of *K. lactis* began with the goal of validating the hybrid form of **a**sg regulation and thereby validating our model of a hybrid form of **a**sg control acting as a transition state between the positive and negative forms of control. MATa2's role as a positive regulator of **a**sgs was validated by experiments in which deletion of MATa2 was shown to result in defective **a**-type mating (A. Tsong, unpublished results). My attempts to validate MATα2's role as a negative regulator of **a**sgs have focused on ChIP experiments of α2 in α cells. The results of these experiments have been somewhat ambiguous (data not shown), but with regards to a possible interaction between MATα2 and Mcm1 in *K. lactis*, a ChIP-Chip of α2 in **a**α cells, performed by D. Galgoczy has been very informative. Genome-wide Mcm1 and α2 bind 107 of the same regions (Figure 5c), which is roughly half the total number of regions bound by each transcription factor individually. This strongly suggests that Mcm1 and α2 interact, as predicted by my previous work[31]. What's surprising is that they are not found interacting at **a**sgs, as we had also predicted. If it turns out to be true that Mcm1 and α2 interact at many genes (these experiments need to be repeated and ideally performed in α cells in a range of conditions), but not at **a**sgs, then one of two revised scenarios for the evolution of **a**sg control is possible: (1) The Mcm1-α2 interaction evolved outside the **a**sg regulon and then moved inward, or (2) The Mcm1-α2 interaction evolved at the **a**sg regulon (as originally proposed) and then moved outward on the *K. lactis* branch.

Mating-type switching is unidirectional and media-dependent in *K. lactis*

When doing a routine check of the mating-type of my strains, I noticed something interesting (Figure 7): MATa strains grown on SD plates at 30°C for a few days can switch mating-type to MATα, but the reverse is not true and the switching from MATa to MATα does not occur on YEPD. It is interesting that mating-type switching would be condition-dependent and biased in its direction. The biased direction may stem from the fact that, unlike in *S. cerevisiae*, one silent MAT locus (MATα) is on the same chromosome as the expressed MAT locus and the other (MATa) is not. The conditional dependence of the switch suggests that one of the players in the switch is differentially regulated by these media conditions. In *S. cerevisiae*, the obvious guess for such a player would be HO, the endonuclease that catalyzes mating-type switching by generating a double-stranded DNA break. However, *K. lactis* does not have an HO ortholog and therefore it is unclear what might be catalyzing mating-type switching on SD media.

Expression of mating-type transcription factors in *K. lactis*

In collaboration with L. Booth, *K. lactis* **a** (yBT15), α (yBT16), engineered α (yBT26), and **a**α (yDG957) strains were profiled for gene expression by microarray under three growth conditions: YEPD, SD +AA +Uri +Leu (0.5g/L) $-PO_4$, and SD +AA +Uri +Leu (0.5g/L) $-PO_4$ + α-factor (full dataset not shown). Analysis of the large dataset produced by these experiments (and especially integration with ChIP-Chip and sequence data) is still underway. Here I present just the gene expression levels of the five genes on the MAT locus and one other regulator of mating (Figure 8; MATa1, MATa2, MATα1, MATα2, MATα3 and Ste12) because these data help to illuminate mating-type regulation

in *K. lactis*. When grown in YEPD, *K. lactis* cell types show only very modest differentiation: the MATa genes are up roughly twofold in the **a** cells compared to the α cells, the MATα genes are up roughly twofold in the α cells compared to the **a** cells, and even less induction occurs in the **a**α. When grown in the PO_4 starvation media, the MAT locus genes are further up-regulated in their respective cell types: genes of the MATa locus are up-regulated another two- to eight-fold in **a** and **a**α cells relative to α cells, and, similarly, genes of the MATα locus are up-regulated two- to eight-fold in α and **a**α cells relative to **a** cells. The addition of α-factor to the PO_4 starvation medium has no effect on **a**α cells, but serves to up-regulate genes of the MATa and MATα loci another two-fold in **a** and α cells respectively.

Taken together, these data indicate that *K. lactis* can up-regulate its expressed (non-silenced) MAT locus in response to PO_4 starvation and in response to α-factor. The fact that α-factor does not affect MAT expression in **a**α cells, whereas it does impact **a** and α cells, suggests that these two forms of regulation are independent. I do not have a prediction for the mechanism of PO_4 starvation response of the MAT locus. However, one candidate for governing the response to α-factor is Ste12, the master transcription factor of pheromone response in *S. cerevisiae*. As can be seen in Figure 8, Ste12 transcriptional up-regulation recapitulates the up-regulation of the MAT locus genes seen in **a** and α cells, but not **a**α cells, in response to pheromone. Given that we did not find evidence for binding of the a1-α2 heterodimer at genes of the pheromone response pathway in **a**α cells (at least in YEPD; detailed above), it is unclear what mechanism, if any, is employed in **a**α cells to prevent pheromone response.

A rough model of the *K. lactis* sexual cycle

From the data presented in this appendix, one can synthesize a very rough model of the *K. lactis* sexual cycle, one that differs substantially from its *S. cerevisiae* and *C. albicans* counterparts (Figure 9). Moving forward, a couple interesting aspects of this model appear to be (1) asymmetric mating-type switching (from **a** to α), which occurs in conditions that overlap those required for mating, and (2) the possibility that **a** and α cells differ substantially in their metabolic profiles.

FIGURES

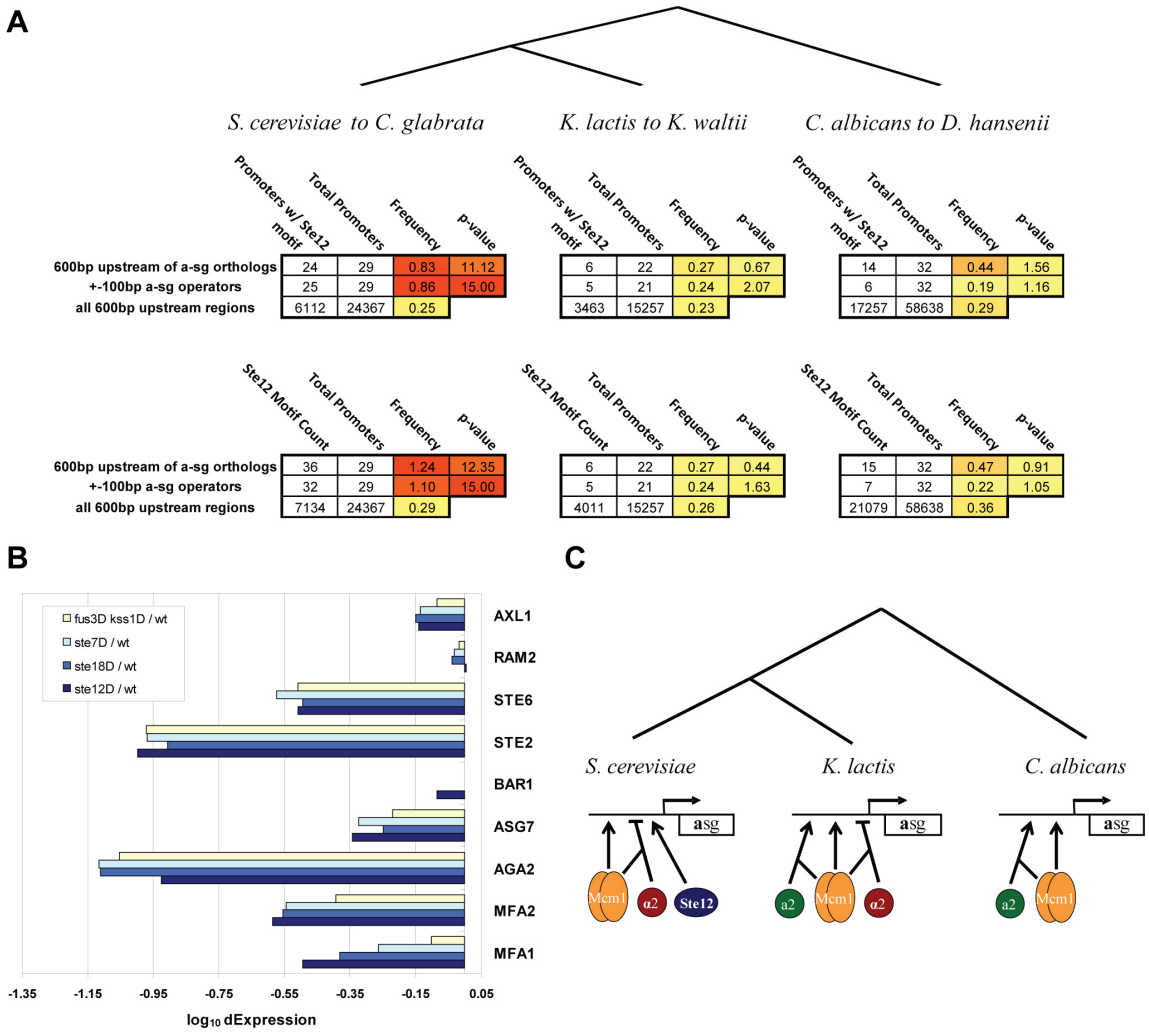


Figure 1. The regulatory transition at a-specific genes may also involve the gain of Ste12 binding sites.

(a) Ste12 motifs (TGTTTSA) were counted upstream of **a**-specific genes and the frequency of motif occurrence was compared to the background frequency (at all 600bp upstream regions). Significant enrichment for Ste12 motifs at **a**-specific genes was found in species within the branch spanning *S. cereivisiae* and *C. glabrata*, but not outside this

branch, suggesting that the rewiring of **a**-specific genes also included the gain of Ste12 binding sites. Ste12's binding motif is thought to be conserved in this lineage, as judged by motif analysis on the promoters of *C. albicans* pheromone response genes.

(b) *S. cereivisiae* strains in which Ste12 or signaling proteins upstream of Ste12 in the pheromone response pathway (e.g., Ste7) are knocked out show decreased levels of **a**-specific gene expression. This effect, which is likely a direct one, varies in magnitude for each of the **a**-specific genes. The data presented here are from Roberts et al.[136]

(c) A revised model, taking into account this new data, of the three regulatory architectures controlling **a**-specific genes across yeast species. This new model implies that loss of regulation by the activator MATa2 at **a**-specific genes occurred concomitantly with the gain of regulation by another activator, Ste12.

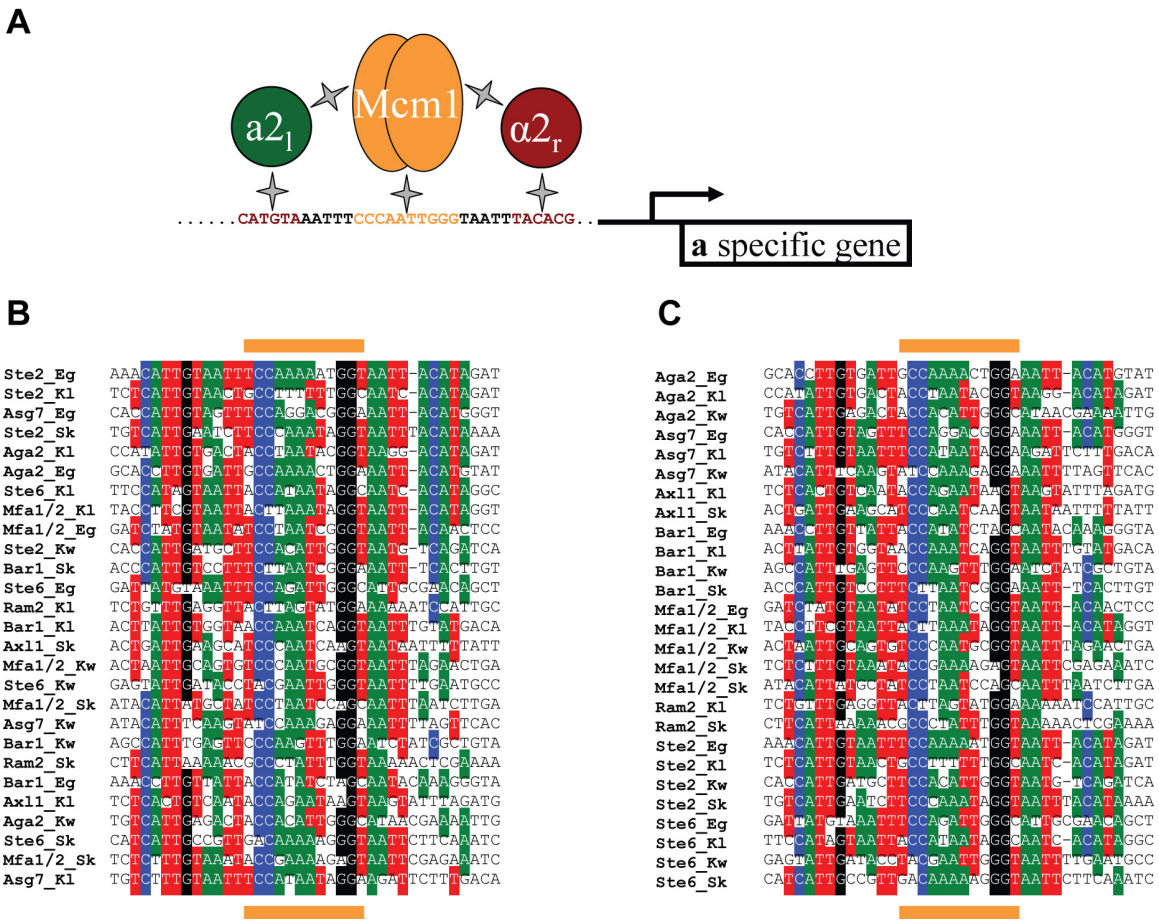


Figure 2. The putative hybrid form of a-specific gene control may differ by gene.

(a) A schematic of the hybrid form of **a**-specific gene control by MATa2 and MATα2. Note: MATa2 and MATα2 are only predicted to be active in **a** and α cells, respectively. (b-c) Alignments of the predicted **a**-specific gene operators of species predicted to implement the hybrid form of **a**-specific gene control (Kl = *K. lactis*, Kw = *K. waltii*, Eg = *E. gosyppii*, Sk = *S. kluyveri*). The alignments are sorted by similarity in (b) and by downstream **a**-specific gene in (c). It is clear from (b) that some **a**-specific gene operators have motifs defined on both sides of the Mcm1 motif, while others have a motif

defined only on one side. As seen in (c), this feature of these **a**-specific gene operators is apparently dependent on the identity of the downstream **a**-specific gene.

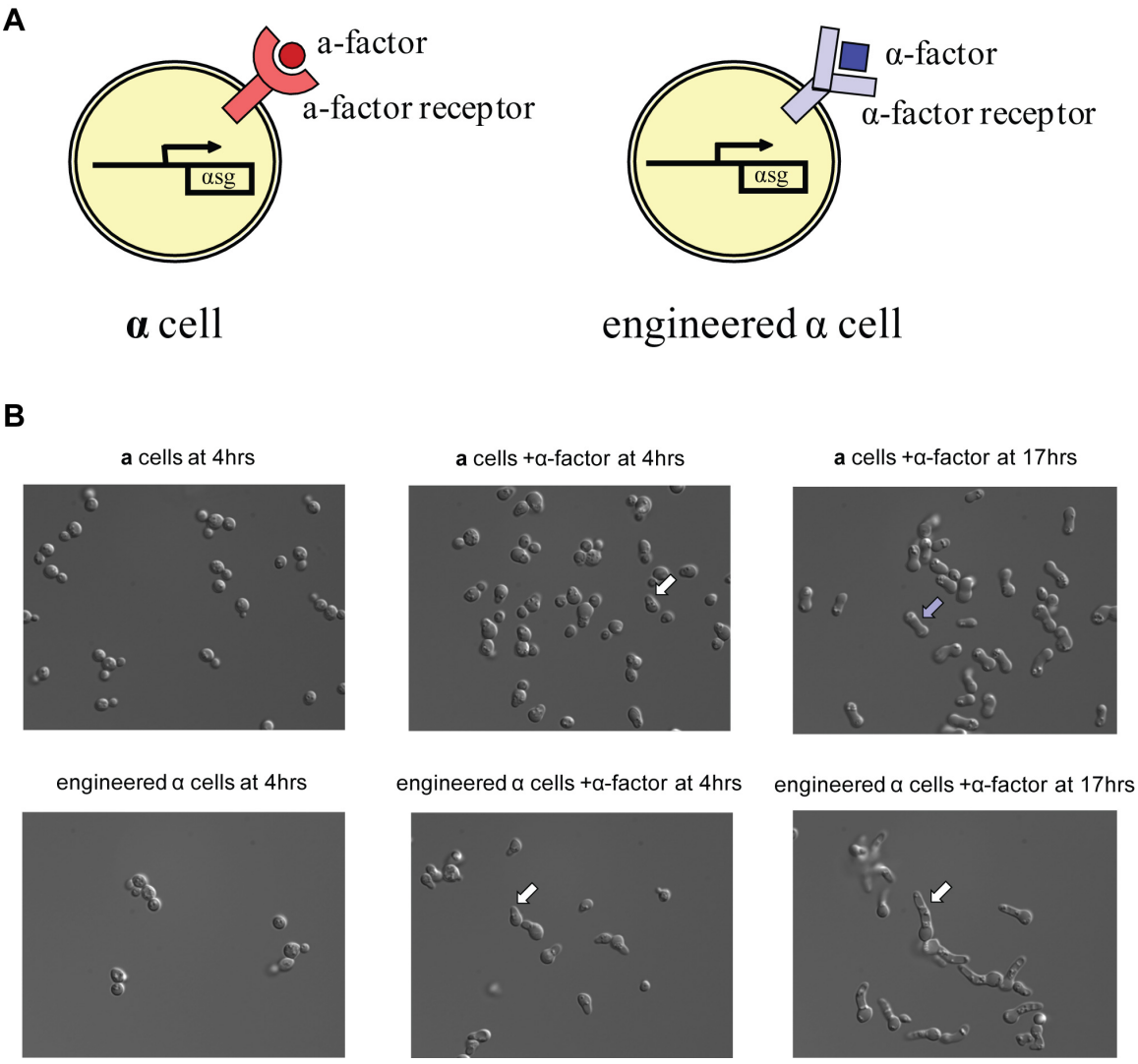


Figure 3. Engineering *K. lactis* α cells to respond to α-factor.

(a) Wild-type α cells express **a**-factor receptor (Ste3) and thus normally respond to **a**-factor. Because **a**-factor is generally expensive to synthesize and difficult to work with, I engineered α cells that respond to α-factor instead (yBT26). The trick, used previously in *C. albicans*, is to replace the endogenous **a**-factor receptor gene (STE3) with the gene encoding the α-factor receptor (STE2), while keeping the **a**-factor receptor gene's α-

specific gene promoter intact. The gene encoding α-factor (MFα) is also deleted from this strain, preventing self activation, which may be lethal.

(b) Wild-type **a** cells (yBT15) and engineered α cells (yBT26) responding to α-factor. Strains were grown overnight in SD +AA +Uri +Leu (0.5g/L) at 30°C with shaking. Cells were washed (2x with H_2O) and resuspended in PO_4 starvation media (SD +AA +Uri +0.5g/L Leu –PO_4) for 6 hours. α-factor is then added (8.3 μg/ml of culture) and imaging is performed 4 and 17 hours later. Note that, at 17 hours, **a** cells have begun to adapt to α-factor by retracting their shmoos, while engineered α cells continue to respond, perhaps because they do not express Bar1, the enzyme which cleaves α-factor. Wild-type α cells do not respond morphologically to α-factor (not shown).

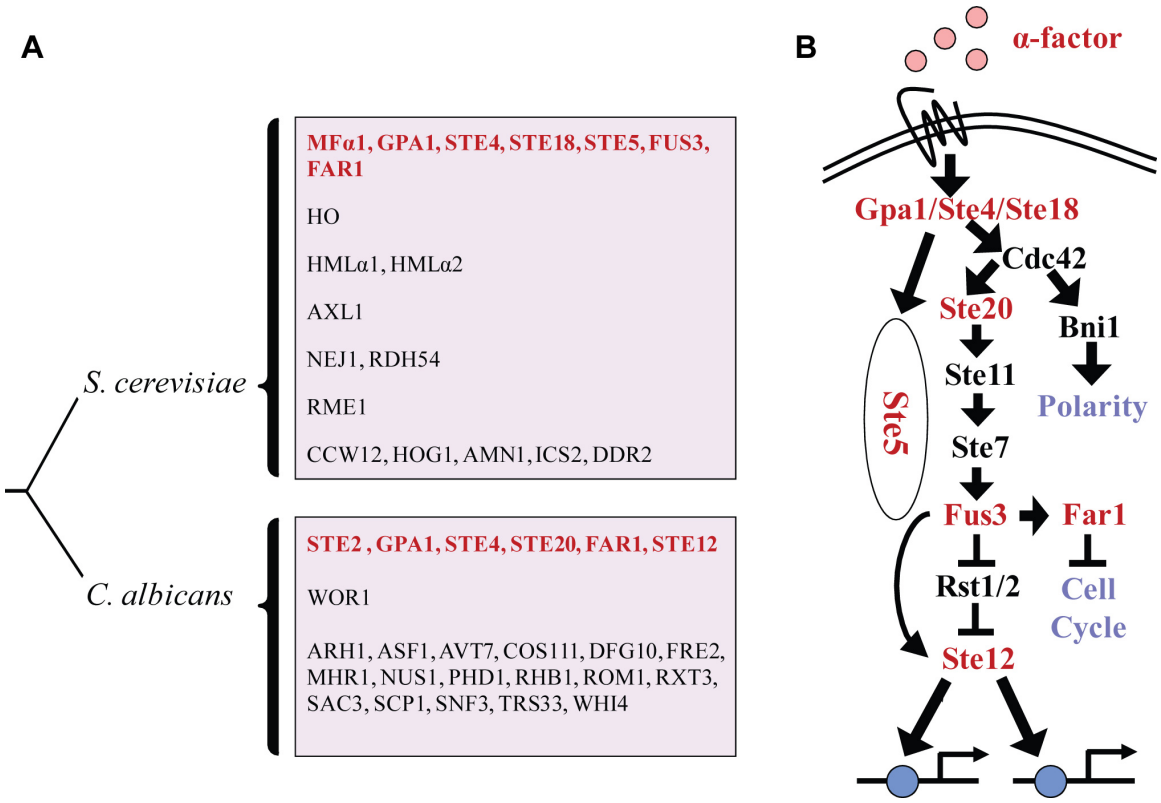
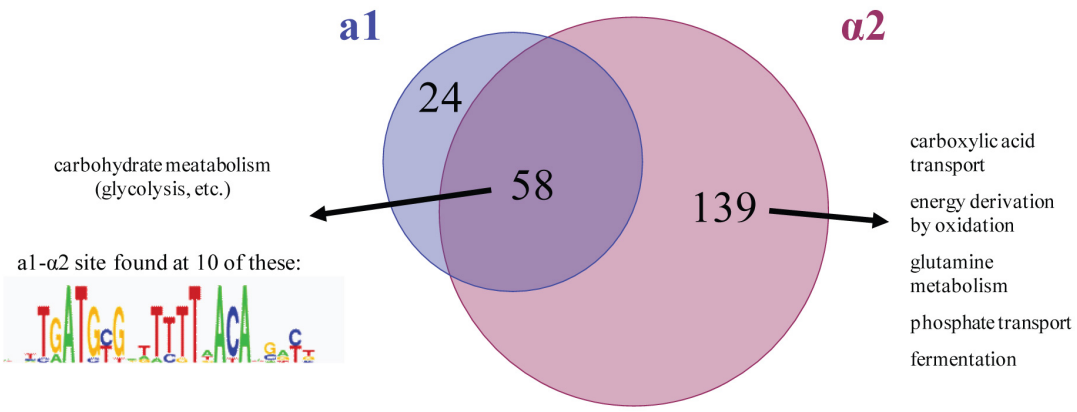


Figure 4. a1-α2 regulated genes in *S. cerevisiae* and *C. albicans*..

(a) A list of a1-α2 regulated genes in *S. cerevisiae* and *C. albicans* compiled from the literature[30, 120, 141].

(b) Schematic of the pheromone response pathway of *S. cerevisiae*. Elements of the pathway directly regulated by a1-α2 in either *S. cerevisiae* or *C. albicans* are colored red. Note that while many elements of this pathway are repressed by a1-α2 in each of the species, the overlapping set of targets is relatively small.

A



B

Swi5	Transcription factor that activates genes expressed in G1 phase and at the G1/M boundary
Tec1	Transcription factor required for haploid invasive and diploid pseudohyphal growth
Rme1	Transcription factor that prevents meiosis by repressing IME1 and promotes mitosis by activating CLN2
Sut1	Transcription factor involved in sterol uptake; involved in induction of hypoxic gene expression
Cyc8	General transcriptional co-repressor, acts together with Tup1p; also acts as part of a transcriptional co-activator complex that recruits the SWI/SNF and SAGA complexes to promoters
Pmc1	Vacuolar Ca2+ ATPase involved in depleting cytosol of Ca2+ ions; prevents growth inhibition by activation of calcineurin in the presence of elevated concentrations of calcium
Pdc6	Minor isoform of pyruvate decarboxylase, key enzyme in alcoholic fermentation
Rad54	DNA-dependent ATPase, stimulates strand exchange by modifying the topology of double-stranded DNA; involved in the recombinational repair of double-strand breaks in DNA during vegetative growth and meiosis
Rad16	Recognizes and binds damaged DNA in an ATP-dependent manner (with Rad7p) during nucleotide excision repair
Sen34	Subunit of the tRNA splicing endonuclease that contains the active site for tRNA 3' splice site cleavage
Ypr013c	
Yjl051w	Bud tip localized protein; mRNA is targeted to the bud by a She2p dependent transport system; mRNA is cell cycle regulated via Fkh2p, peaking in G2/M phase; null mutant displays increased levels of spontaneous Rad52 foci
Yjr096w	Aldehyde reductase activity

C

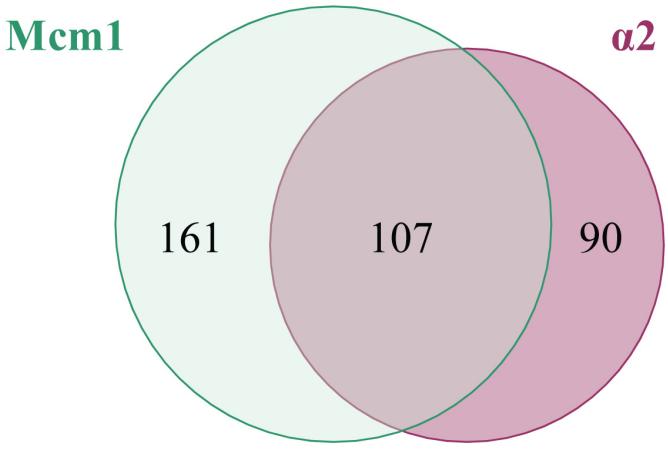


Figure 5. a1-α2 and α2-Mcm1 regulated genes in *K. lactis*.

(a) The overlap of a1- and α2-bound genes in *K. lactis*, as determined by ChIP-Chip of **a**α cells. There are 58 regions bound by both a1 and α2; the set of genes flanking these regions is enriched for functions in carbohydrate metabolism. Only ten of these a1-α2 bound regions also show strong evidence for the presence of an a1-α2 *cis* regulatory motif.

(b) The thirteen genes flanking the ten regions with both a1-α2 binding and an a1-α2 *cis* regulatory element.

(c) The overlap of Mcm1- and α2-bound genes in *K. lactis*, as determined by ChIP-Chip. There are 107 regions bound by both Mcm1 and α2. Genome-wide the Pearson correlation of log-ratios (IP/Input) for the Mcm1 and α2 ChIPs is 0.56.

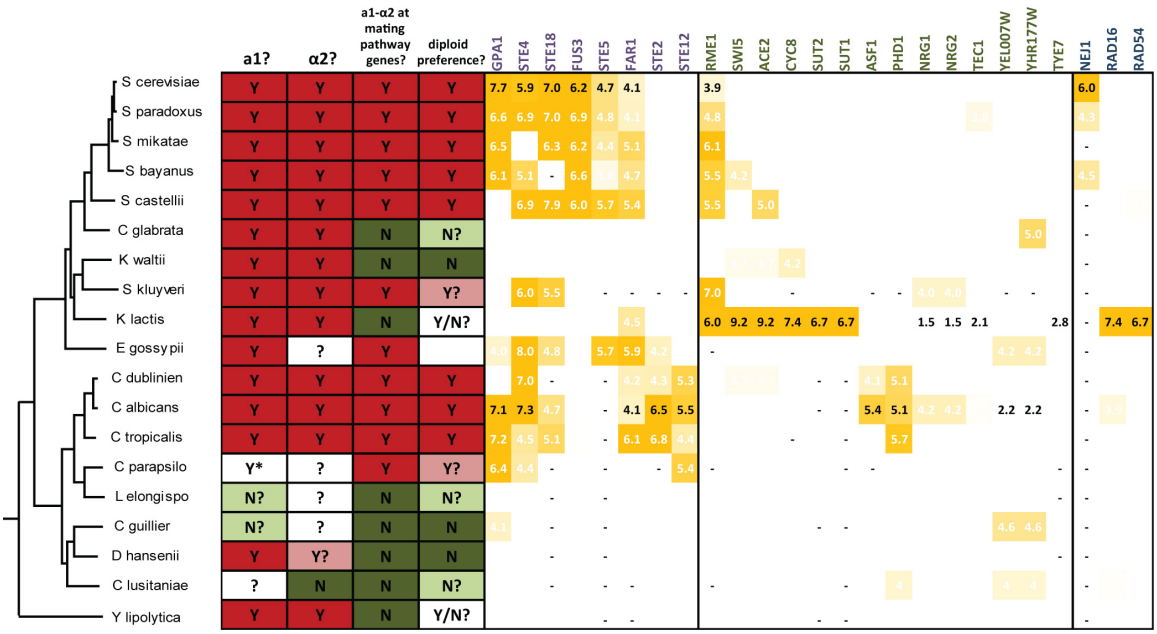


Figure 6. a1-α2 *cis* regulatory motifs and ploidy preference across yeast species.

a1-α2 regulation and ploidy preference are annotated across the hemiascomycete lineage. On the left, for each species the presence of the MATa1 and MATα2 genes at their respective MAT loci, the presence of a1-α2 motifs at the genes of the pheromone response pathway, and the ploidy preference (if known) is annotated. On the right, a subset of known a1-α2 regulated genes of *S. cerevisiae, K. lactis*, and *C. albicans* were mapped to other yeast species with fully-sequenced genomes, promoter sequences were extracted and a motif search was performed with an a1-α2 position-specific scoring matrix (PSSM). Shown are $\log_{10}$-odds scores (motif vs. background) for the 2 kb region upstream of each mapped gene (columns) in each species (rows). Cells are colored shades of yellow, the strength of the yellow indicating the strength of the match to the a1-α2 PSSM. Genes were grouped by category (1st group contains pheromone response pathway components, 2nd group contains a variety of transcriptional regulators, 3rd group

contains DNA repair enzymes). For species in which ChIP data was available (*S. cerevisiae, K. lactis*, and *C. albicans*), experimentally validated a1-α2 bound genes have their scores colored in black text (rather than white). The results shown are consistent with the idea that the set of genes that a1-α2 targets evolves rapidly; genes with a strong a1-α2 PSSM score in their promoter in one species do not often show strong a1-α2 PSSM scores in more than one or two other species.

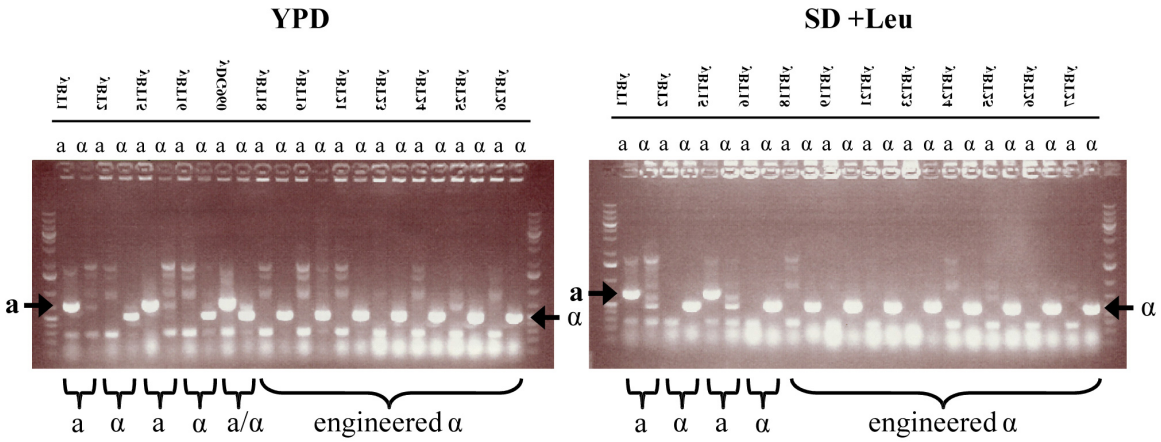


Figure 7. Mating-type switching is unidirectional and media-dependent in *K. lactis*. Colonies were streaked from -80°C to YEPD or SD +Leu plates and grown for 3 days at 30°C. (My documentation of this experiment was not good because I was not actually looking for this switching effect, but rather was intending to merely verify the mating-type of my strains.) Shown are the results of colony PCR using primers specific to the MAT**a** (denoted '**a**' along the top) or the MATα (denoted 'α' along the top) locus. Expected (i.e., starting) mating-types are listed at the bottom of the gels. The expected product lengths are shown along the periphery of the gels. Note that **a** cells can switch to α cells, but not vice versa, and only on SD +Leu media.

	YEPD					SD -PO4					SD -PO4 +alphaF		
	15 (**a**)	16 (α)	26 (α)	31 (α)	957 (**a**α)	15 (**a**)	16 (α)	26 (α)	31 (α)	957 (**a**α)	15 (**a**)	26 (α)	957 (**a**α)
MATa1	1.60	0.21	0.36	0.16	1.00	2.76	0.12	0.17	0.40	2.95	3.78	0.00	2.69
MATa2	0.76	0.30	0.27	0.00	0.19	1.93	0.38	0.42	1.07	3.29	2.76	0.07	3.24
MATalpha1	0.00	1.00	0.99	0.69	0.49	1.41	3.26	3.40	1.69	2.71	1.24	4.60	2.47
MATalpha2	0.00	0.67	0.66	0.23	0.34	0.28	1.33	1.22	0.52	2.15	0.48	2.38	2.35
MATalpha3	0.56	1.11	1.28	0.10	0.33	0.00	1.91	1.66	0.75	1.46	0.25	2.06	1.31
Ste12	0.50	0.39	0.28	0.20	0.35	2.52	2.27	1.55	1.16	1.05	3.85	4.32	0.74

Figure 8. Expression of mating-type transcription factors in *K. lactis*.

Shown above is a subset of the data produced by the gene expression profiling of *K. lactis* **a** (yBT15), α (yBT16), engineered α (yBT26), and **a**α (yDG957) strains grown in three conditions: YEPD, SD +AA +Uri +Leu (0.5g/L) $-PO_4$, and SD +AA +Uri +Leu (0.5g/L) $-PO_4$ + α-factor. Here I present relative gene expression levels for five genes on the MAT locus (MATa1, MATa2, MATα1, MATα2 and MATα3) and one other regulator of mating (Ste12). Values shown in each row (i.e., for each gene) are $\log_2$ transformed against the minimally expressed strain/condition of that row (typically one of the strains grown in YEPD).

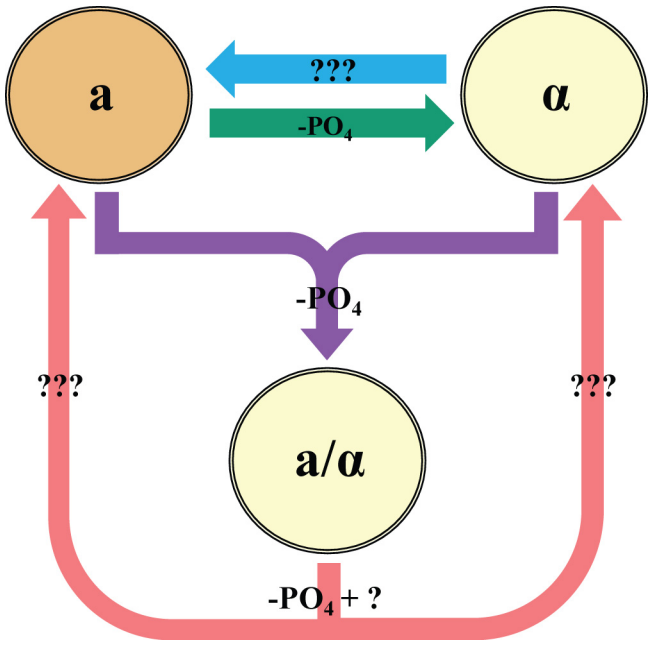


Figure 9. A rough model of the *K. lactis* sexual cycle.

A very rough model of the *K. lactis* sexual cycle (synthesized from data presented in this appendix). Mating-type switching is apparently asymmetric (only from **a** to α), and occurs in conditions that overlap those required for mating. The alternative coloring of the three cells denotes the possibility that α and **a**α cells may differ substantially from **a** cells in their metabolic profiles.

Appendix 3

Evolution of the white-opaque epigenetic switch

INTRODUCTION

In Chapter 3, I discussed our discovery that Mcm1 binding sites have very recently been gained at over one hundred genes in the *C. albicans* lineage. This change is of note, not only for its apparently very recent occurrence, but also because many of the genes at which Mcm1 is bound are relevant to host interactions, and at these genes Mcm1 is found binding a non-canonical *cis* regulatory motif (Chapter 3, Figures 1 and 7). Mcm1 was found binding the non-canonical motif at several genes functioning in biofilm formation as well as three out of four known regulators of the white-opaque developmental switch. Given the very recent appearance of Mcm1 at these genes, I began to wonder what impact Mcm1 regulation was having on these processes, and whether the gains of such regulation have adapted *C. albicans* to its human host. More generally, I began wondering how a developmental switch evolves and whether the white-opaque switch evolved recently enough that one might still be able to trace its origins with extant organisms. The details of this ongoing endeavor are presented here.

RESULTS

The white-opaque switch originated prior to the addition of Mcm1 regulation

I first asked whether the appearance of Mcm1 binding at the white-opaque regulators (and the other ~100 non-canonically bound genes) was associated with the gain of the white-opaque switch itself. We knew already that *C. albicans*' closest sequenced relative (*C. dubliniensis*) underwent the white-opaque developmental switch[127], so I acquired the next most closely related yeast with a sequenced genome, *C. tropicalis*. *C. tropicalis* does not show evidence of non-canonical Mcm1 motifs at the orthologs of the non-canonically bound genes of *C. albicans*. Naively, I predicted that *C. tropicalis* would not have this developmental switch.

If *C. tropicalis* could mate efficiently and/or respond to pheromone (e.g., by shmooing) without first undergoing the white-opaque switch, it would suggest the absence of the white-opaque switch in this species (or at least the unlinking of mating-type and switching that is found in *C. albicans*). To test this I needed to isolate *C. tropicalis* strains homozygous or hemizygous at the mating-type locus. I performed sorbose selection on **a**α strains (Johnson lab protocol), which, at least in *C. albicans*, selects for loss of the chromosome on which the mating-type locus resides. Of the three *C. tropicalis* strains I acquired and sorbose selected, one yielded both **aa** and αα isolates and another yielded just **aa** isolates (Note: it is formally possible that these strains are hemizygous for the mating-type locus/chromosome). It is somewhat surprising to me that the fitness advantage given in growth on sorbose by loss of the mating-type-locus-

containing chromosome is conserved across these yeasts (why should this be?). Using synthetic α-factor (predicted for *C. tropicalis*: KFRLTRYGWFSPN; synthesized by Genemed Synthesis), I attempted to discover a pheromone response under a variety of growth conditions (including several types of liquid media and temperatures). I also combined **aa** and αα cells in a number of different media and at several temperatures, looking for evidence of shnmooing, of zygote formation, or of streaked isolates with a MAT**a**/MATα genotype. These ventures were not successful, leaving me less confident that a "white-opaque" switch, required for efficient mating, does *not* exist in *C. tropicalis*.

I then reversed my approach and attempted to find direct evidence of a white-opaque switch. The first definitive proof of a "white-opaque" switch in *C. tropicalis* came after plating cells on blood agar at 37°C (see Figure 1). In *C. tropicalis*, unlike *C. albicans*, both **aa** and **a**α appear to undergo "white-opaque" switching, though "opaque" **aa** cells tend to be more narrow and elongated than opaque **a**α cells (compare Figures 1d and 1e). While a white-opaque switch is apparently present, the switch seems to be a bit more finicky. For example, I have not succeeded at getting back white and opaque colonies after streaking white or opaque sectors to new media. Also, the elongated morphology of opaque cells does not appear to be stable after re-suspension in liquid media. It is possible that if one was to spend more time, one could "tame" these white and opaque forms, by identifying a condition under which bi-stability is achieved. However, it is also possible that the white and opaque forms are less heritable in *C. tropicalis*. Thus, a reconsidered hypothesis about the impact of gained Mcm1 regulation on the white-

opaque switch is that the gain of Mcm1 regulation increased the strong stability and heritability of the two states seen in *C. albicans* and *C. dubliniensis*.

Increased Mcm1 expression drives formation of opaques in *C. albicans*

Mcm1 is clearly bound at several white-opaque regulators. One might then expect fluctuations in Mcm1 levels to impact the rates of white-opaque switching. I first tested this possibility by performing standard switching assays with MCM1/MCM1 and MCM1/*mcm1*Δ strains. The single knockout did not significantly affect either white to opaque or opaque to white switching rates. The double knockout strain can not be made because MCM1 is an essential gene. I next designed a Doxycycline-inducible Mcm1 ectopic expression construct and transformed it into *C. albicans* **aa** cells (yBT65). The resulting strains (yBT67a,b,c,d) form opaque sectors at 50x greater frequency when grown on SD+aa+uri supplemented with 100μg/ml of Doxycylcine than on the same media without Doxycylcine (Figure 2a). In fact, nearly every colony of yBT67 grown on inducing media has multiple opaque sectors at 10 days after plating (Figure 2b); the sectors eventually form an opaque ring around the colony. Of note, Doxycylcine apparently inhibits opaque formation in the parent strain (without ectopic expression construct; yBT65; Figure 2a), suggesting an explanation for why I rarely see large (i.e., early-forming) opaque sectors on my +Doxycylcine plates.

The non-canonical Mcm1 motif is probably bound by an unknown transcription factor

It remains to be determined whether Mcm1 truly binds the non-canonical motif or whether Mcm1 interacts with some other protein that binds this motif. However, given the divergence of this motif from the canonical MADS box motifs, it is likely there is some other regulator binding this motif. In an effort to identify this transcription factor, I scanned several databases for a *cis* regulatory element bearing similarity to the non-canonical Mcm1 motif. A close match was found in the TRANSFAC database: aaaatT**CGG**CGAAgccA**GCC**AATca is characterized as a binding site for amdS (AN4035.3), a fungal-specific Zn(2)-Cys(6) protein from *A. nidulans*. There is apparently no ortholog to this protein in *C. albicans* (Figure 3a); however, there is a fairly closely related homolog, orf19.1499. Another candidate for a protein which may bind the non-canonical motif is encoded by orf19.5729. This gene is also a member of the fungal-specific Zn(2)-Cys(6) family and shares with the other white-opaque regulators (Wor1, Wor2, Czf1 and Efg1) a common signature, namely that of binding by Mcm1, Wor1, Wor2 and Efg1 to its promoter (A. Hernday unpublished data). Additionally this gene has a very striking evolutionary history (Figure 3b); putative orthologs to orf19.5729 are apparently only present in the two species in which we find the non-canonical motif (*C. albicans* and *C. dubliniensis*).

To test whether either of these two genes is required for Mcm1 binding at the non-canonical motif, I obtained (from O. Homann, unpublished) strains in which both alleles of each gene had been knocked out in a MAT**a**/MATα background (yBT63/orf19.5729ΔΔ and yBT64/orf19.1499ΔΔ). I then performed ChIP of Mcm1 in these strains and looked for loss of binding at both canonical and non-canonical Mcm1

binding sites by qPCR (Figure 3c). There was no apparent loss of Mcm1 binding in these strains, indicating that neither of these genes is required for Mcm1 to bind the non-canonical motif.

Another proposal for the non-canonical motif binder comes from a close examination of *cis* regulatory motifs predicted (by performing MEME searches on sequences flanking genes in the arginine regulon) to be bound by Mcm1/Arg80/Arg81 across a range of fungi. The predicted Mcm1/Arg80/Arg81 motif of *C. glabrata* bears noteworthy similarity to the non-canonical Mcm1 motif from *C. albicans* (Figure 4). The requirement of orf19.4766, the *C. albicans* ortholog of Arg81, for Mcm1 binding to the non-canonical motif will be tested soon.

The evolutionary history of the five known white-opaque regulators

In a general effort to understand the origins of the white-opaque switch, I carefully analyzed the evolutionary history of each of the five known white-opaque regulators (Figures 5-9). In each case, the *C. albicans* gene of interest was BLASTed against a database of all fungal ORFs. ORFs matching with E-value less 10^{-5} were extracted from the database, multiply aligned with MUSCLE or ClustalW, and a NJ tree was inferred from this alignment using ClustalW. The resulting gene tree was inspected and duplication/loss events were mapped to the species tree (inferred in Chapter 4). Whereas Wor1, Efg1 and Mcm1 (Figures 5, 7 and 9) clearly existed prior to divergence of the ascomycetes studied here, evidence for an early origin of Wor2 and Czf1 (Figures 6 and 8) is less clear. These two Zn finger proteins could represent more recent innovations

(created through duplication and extensive divergence, for example) or it could be that their sequence is just not constrained enough to detect their true orthologs in some branches of the fungal phylogeny.

It is also very interesting that Efg1 was apparently lost on the branch leading to *C. tropicalis* (Figure 7). This means that in any comparison made between the white-opaque switches of *C. albicans* and *C. tropicalis*, at least two major rewiring events (in addition to many other possible events) should be considered: the gain of Mcm1 binding sites on the branch to *C. albicans* and *C. dubliniensis* and the loss of Efg1 on the branch to *C. tropicalis*. *C. albicans* Efg1 mutants are known to be particularly finicky with respect to the state, white or opaque, they occupy and have been remarked to appear somewhat intermediate in form between wild-type whites and opaques (A. Hernday and R. Zordan, personal communication). This may help to explain some of the troubles I've encountered in trying to tame the white-opaque switch of *C. tropicalis*.

The evolution of upstream intergenic regions at regulators of the white-opaque switch

One striking feature of the white-opaque regulators is their unusually long upstream intergenic regions in *C. albicans* (~10 Kb for Wor1, Wor2, Efg1 and Czf1). It is suspected that this feature is of key importance to their function, as white-opaque regulators tend to show enrichment across these entire regions in ChIP performed on opaque cells. While the significance of this has not yet been determined, it is still of

interest to ask when this feature of the circuit arose and whether its evolution was required for some characteristic of the white-opaque switch.

If the ORF annotations for each of the fungal genomes were extremely good, it would be a trivial task to calculate upstream intergenic lengths for all ORFs. However, this is not the case, so I have developed (and am still developing) an algorithm that predicts upstream intergenic lengths by predicting ORFs within the region upstream of a gene and calculating distances from the gene to these predicted ORFs. For each ORF the algorithm works as follows: (1) Extract 25 Kb of upstream sequence, (2) Translate all ORFs of at least 50 amino acids (starting with ATG) in all six reading frames, (3) For each putative ORF, if the length is greater than 150 amino acids then KEEP it, else if a TBLASTN against database of all fungal chromosome sequences, returns a non-self hit with E-value less than 10^{-10}, KEEP it, (4) Calculate distance to nearest upstream ORF.

The predicted upstream intergenic lengths and ATG-free lengths for orthologs of the white-opaque regulators are shown in Figure 10. It is not obvious what these results imply about the origins of the white-opaque switch. However, a couple of trends are noteworthy. Within the branch spanning *D. hansenii* and *C. lusitiniae*, both upstream intergenic and ATG-free lengths are considerably shorter than in the branch spanning *C. ablicans* and *C. parapsilosis*. Whether or not a white-opaque switch exists in species of the *D. hansenii* branch is unknown. Also of note, *C. glabrata*, a species known to undergo phenotypic switching, also has very long upstream intergenic regions at Wor1 and Efg1. Is it possible that these two transcription factors regulate phenotypic switching

in this fungal pathogen as well? Given that many of *C. glabrata*'s close relatives do not have long intergenic regions upstream of Wor1 and Efg1, is it possible that this feature evolved independently in both lineages?

FIGURES

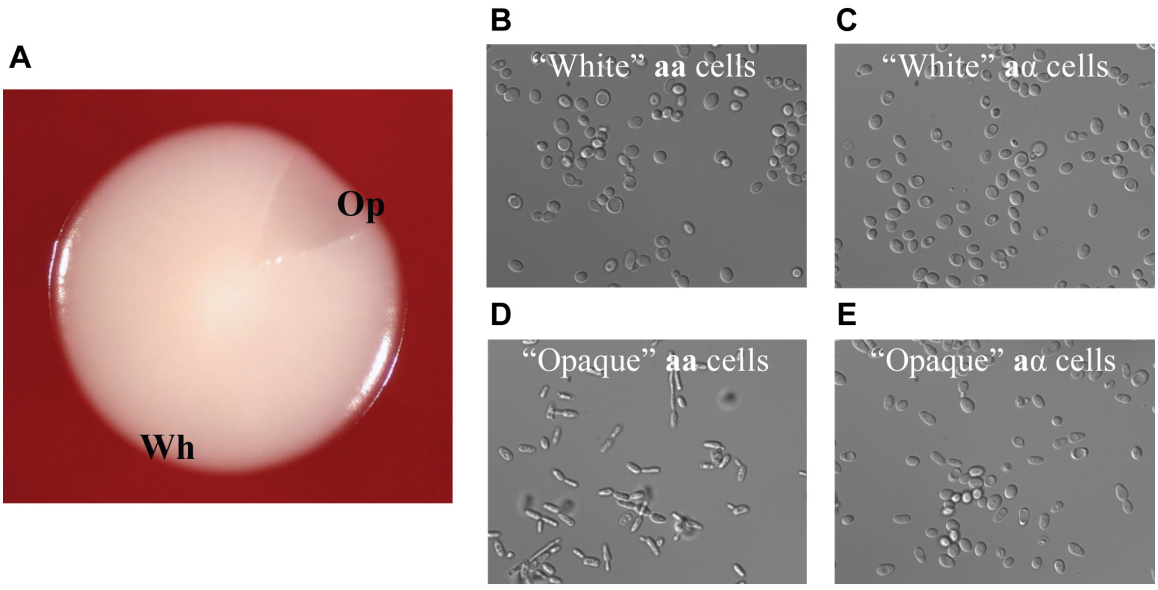


Figure 1. "White-opaque" switching discovered in *C. tropicalis*.

C. tropicalis **aa** and **a**α cells (strains yBT56/57 and yBT44, respectively) were streaked from frozen stocks to YEPD and grown O/N at 37°C. Colonies were re-suspended, diluted and plated (~100 cells per plate) to blood agar at 37°C.

(a) A white (Wh) colony with an opaque (Op) sector of **aa** cells (yBT67) at three days.

(b-e) Cells taken from white colonies or opaque sectors of **aa** and **a**α strains, as labeled.

A

	WT				pNIM1-MCM1 Isolate #1				pNIM1-MCM1 Isolate #2			
Dox	-		+		-		+		-		+	
Opaque Sector?	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Count	1926	38	1912	0	1916	34	0	1912	1922	46	0	1903
Switching Frequency	1.93%		0.00%		1.74%		100.00%		2.34%		100.00%	

B

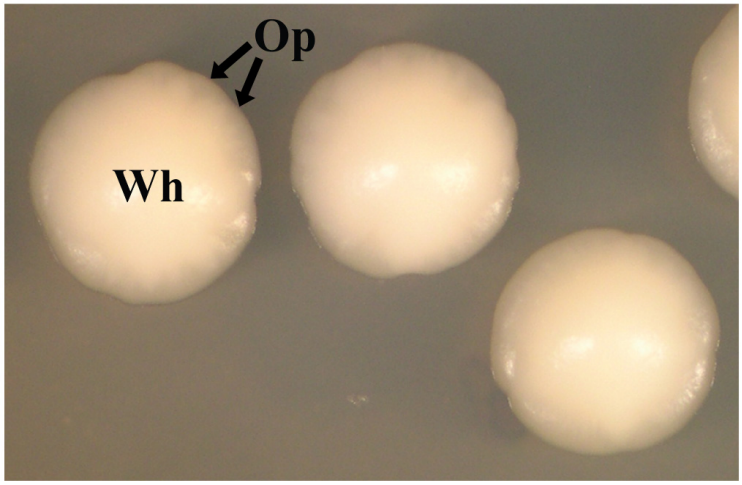


Figure 2. Increased Mcm1 expression drives formation of opaques in *C. albicans*.

A white to opaque switching assay was performed by streaking *C. albicans* **aa** cells from frozen stocks. The strain without the Mcm1 ectopic expression construct (yBT65) was streaked to YEPD+Ade+Uri and those strains with the Mcm1 ectopic expression construct (yBT67a,b) were streaked to YEPD+Ade+Uri+200µg/ml ClonNAT (selecting for the construct). Colonies were grown at RT for 5 days. For each strain, 10 white colonies were selected, mixed, diluted and plated (~60 cells/plate) to both inducing (SD+AA+Uri+100µg/ml Doxycylcine) and non-inducing (SD+AA+Uri) media.

Colonies were grown at RT and colonies with and without sectors were counted on day 11.

(a) Results of the switching assay, indicating that ectopic Mcm1 expression drives opaque formation.

(b) White colonies of the ectopically-expressing Mcm1 strain (yBT67b) at 11 days of growth on inducing media (SD+AA+Uri+100μg/ml Doxycylcine) at RT. Many small opaque sectors can be found along the circumference of the colony. These opaque sectors will eventually merge, forming an opaque ring.

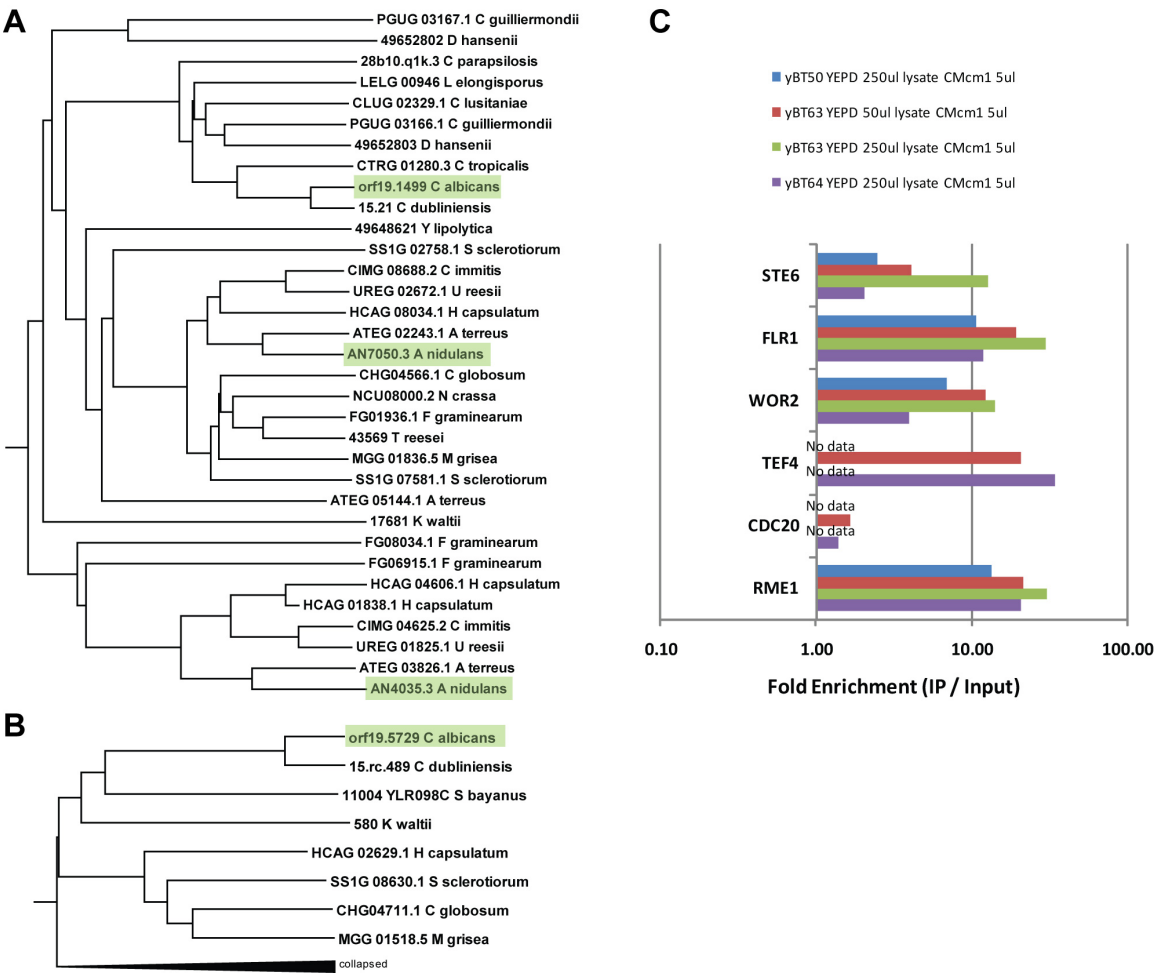


Figure 3. Two putative non-canonical Mcm1 motif binders.

(a-b) The phylogenies of two putative non-canonical Mcm1 motif binders: (a) orf19.1499 and (b) orf19.5729.

(c) Results of a ChIP of Mcm1 in WT (yBT50), orf19.5729ΔΔ (yBT63) and orf19.1499ΔΔ (yBT64) strains of *C. albicans*.

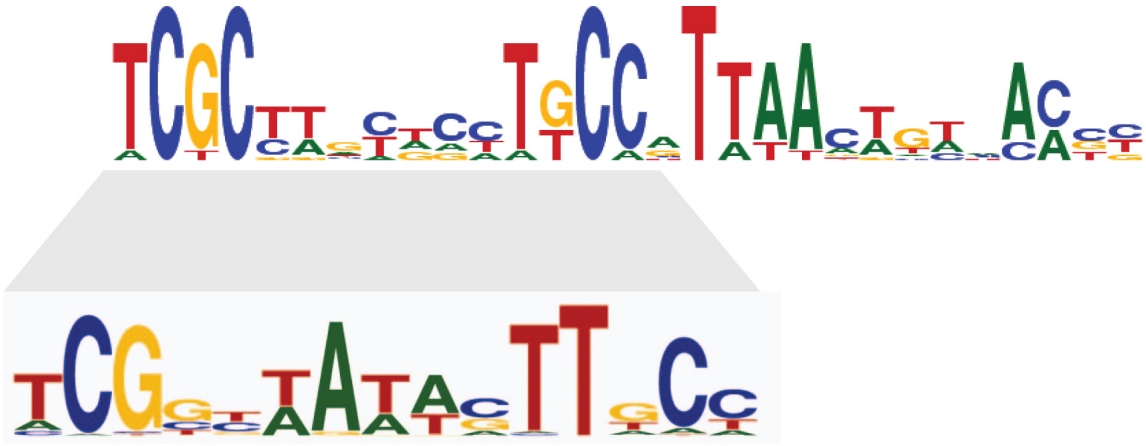


Figure 4. A *cis*-regulatory motif found at the arginine regulon of *C. glabrata* (top) is similar to the non-canonical Mcm1 motif (bottom).

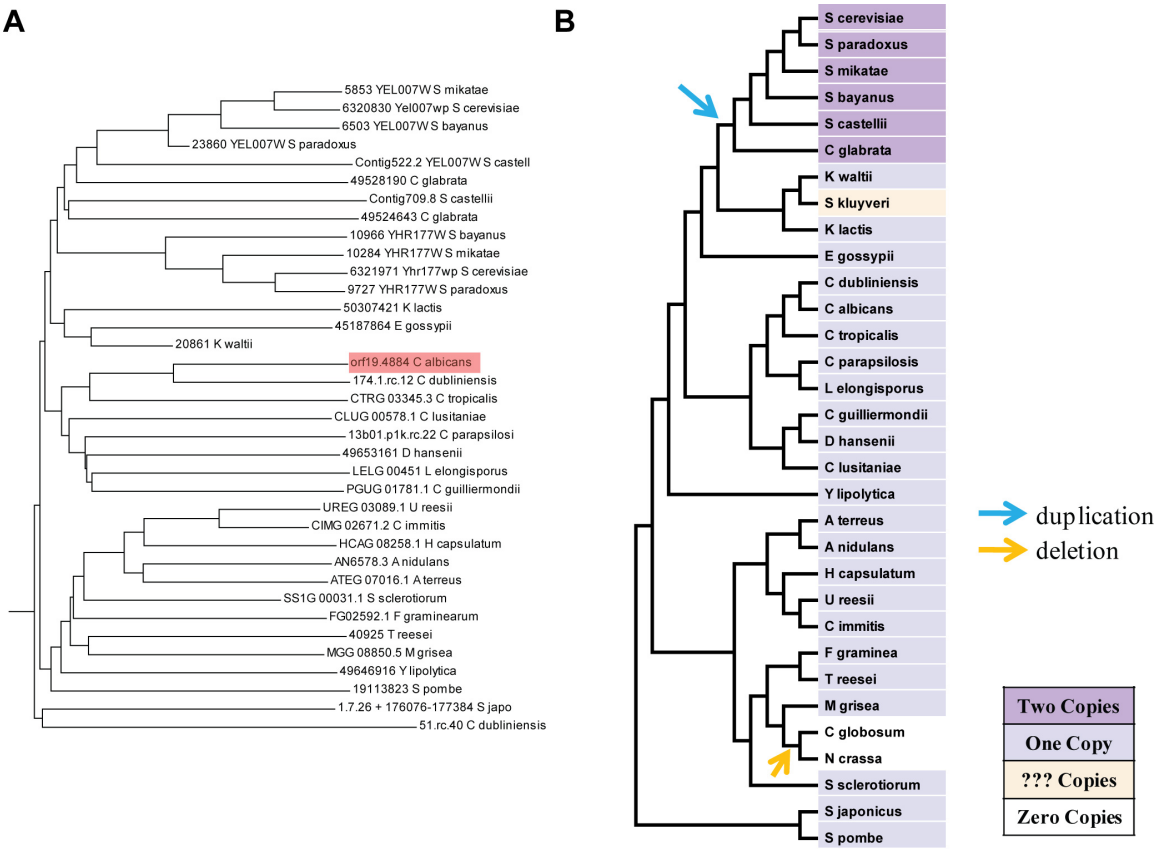


Figure 5. The evolutionary history of Wor1.

(a) A gene tree for Wor1.

(b) Wor1 duplication and deletion events mapped to the fungal species tree.

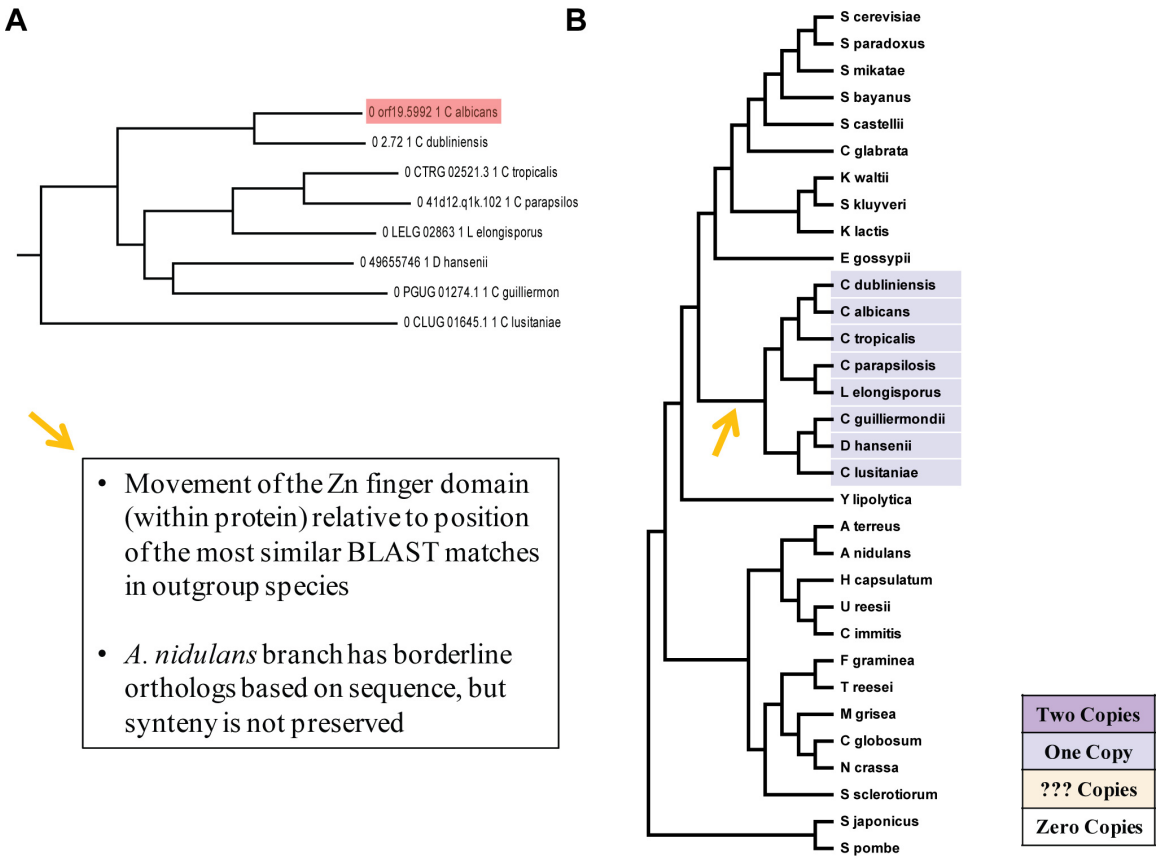


Figure 6. The evolutionary history of Wor2.

(a) A gene tree for Wor2.

(b) Wor2 evolutionary events mapped to the fungal species tree.

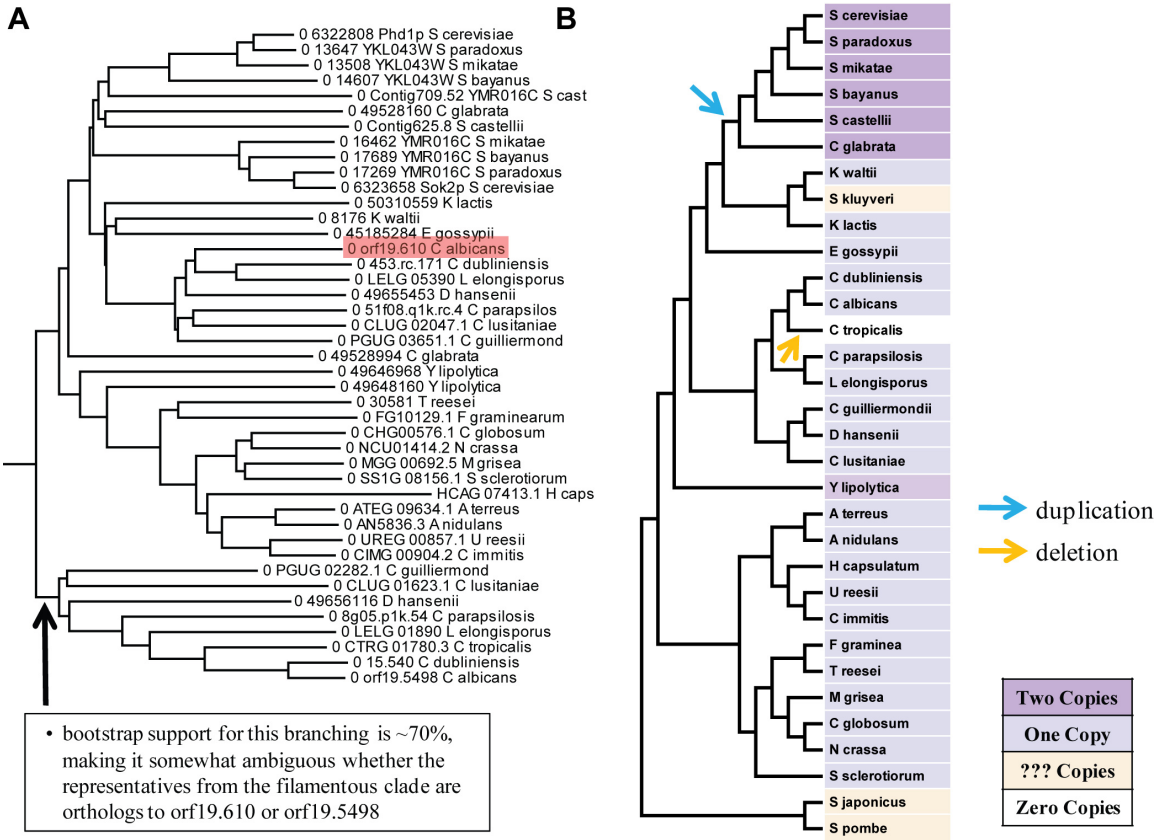


Figure 7. The evolutionary history of Efg1.

(a) A gene tree for Efg1.

(b) Efg1 duplication and deletion events mapped to the fungal species tree.

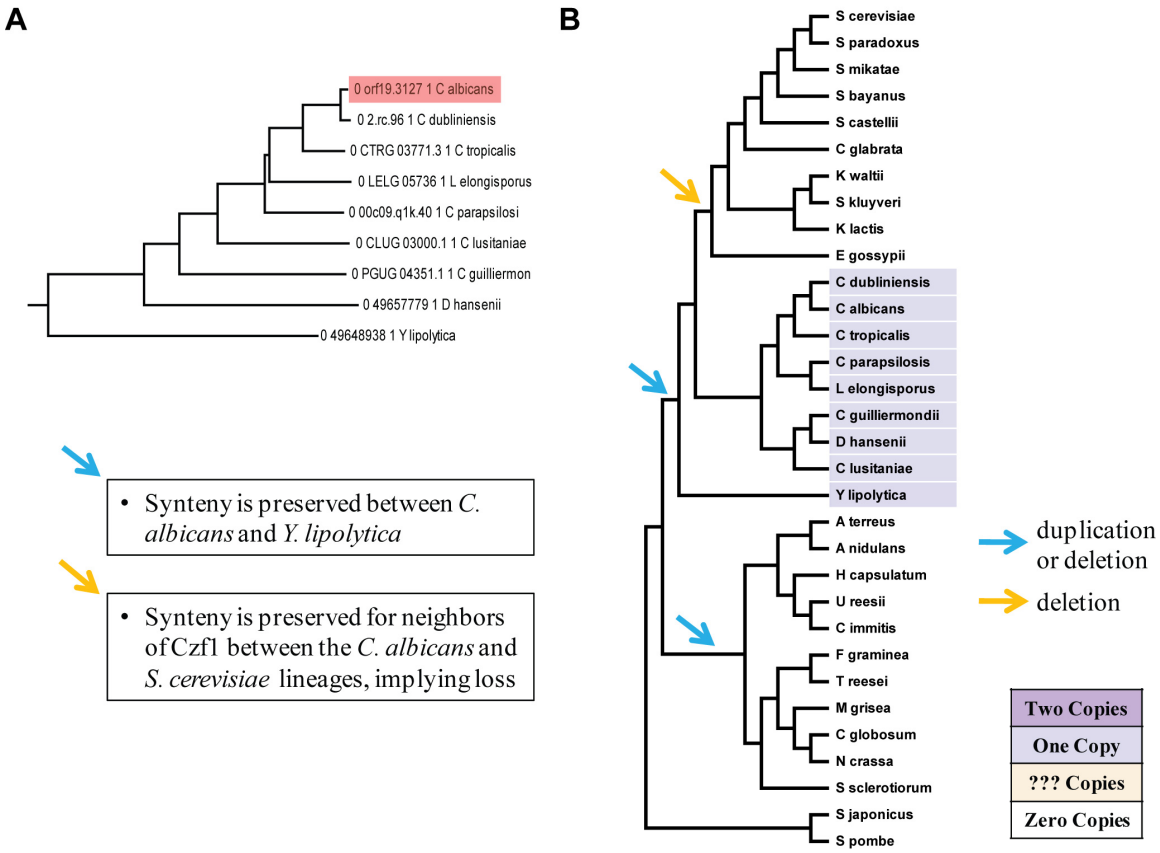


Figure 8. The evolutionary history of Czf1.

(a) A gene tree for Czf1.

(b) Czf1evolutionary events mapped to the fungal species tree.

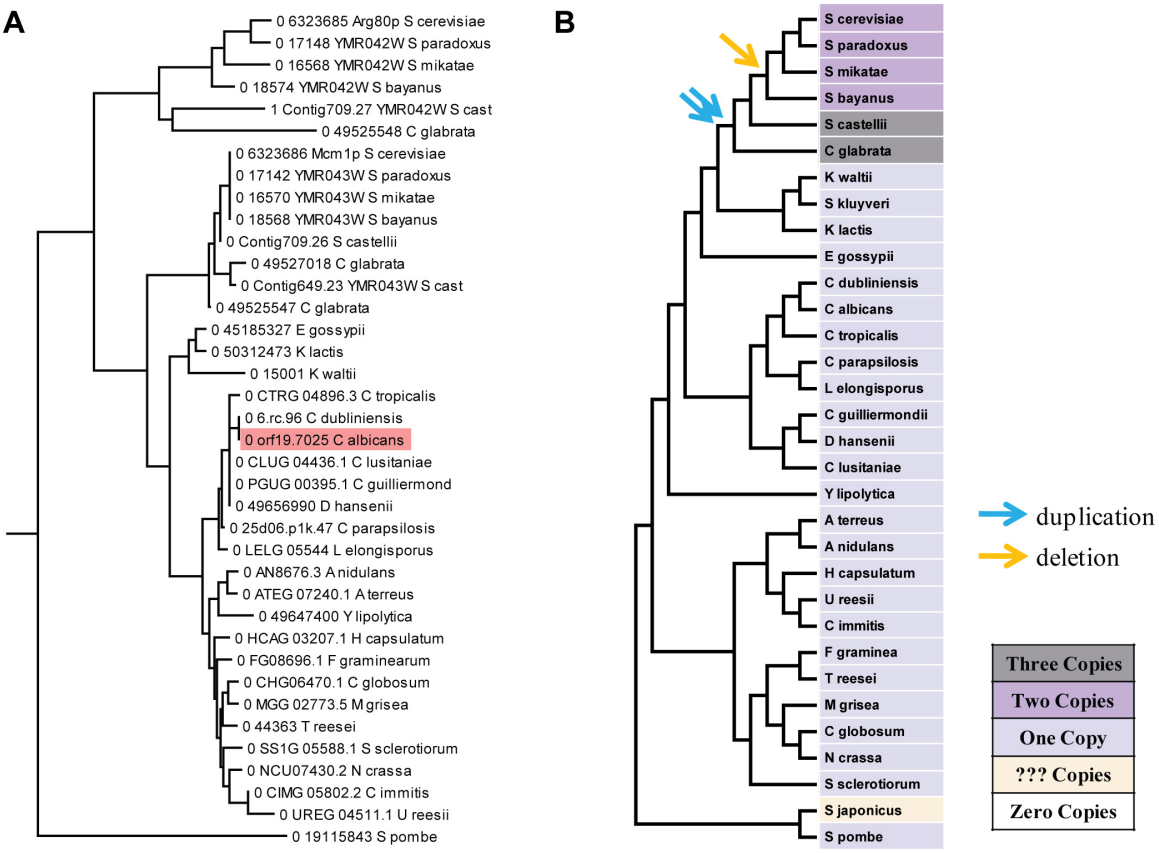


Figure 9. The evolutionary history of Mcm1.

(a) A gene tree for Mcm1.

(b) Mcm1 duplication and deletion events mapped to the fungal species tree.

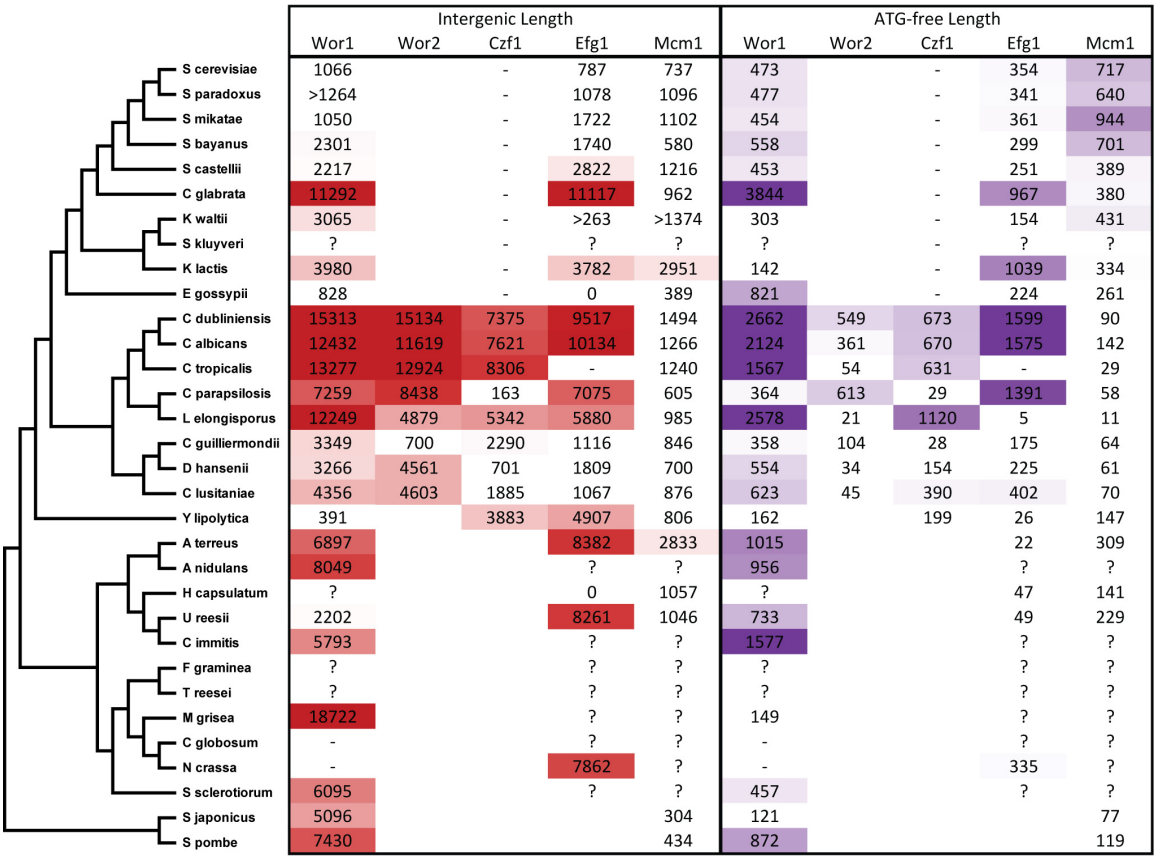


Figure 10. The evolution of upstream intergenic lengths for regulators of the white-opaque switch.

Predicted upstream intergenic (left, shaded darker red by increasing length) and upstream ATG-free region (right, shaded darker purple by increasing length) lengths are shown for each of the five known regulators of the white-opaque switch across a wide range of fungal species. These results are **preliminary** and are discussed further in the text.

References

1. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. Science 188, 107-16 (1975).
2. Levine, M. & Tjian, R. Transcription regulation and animal diversity. Nature 424, 147-51 (2003).
3. Davidson, E. H. Genomic regulatory systems : development and evolution (Academic Press, San Diego, 2001).
4. Ptashne, M. & Gann, A. Genes & signals (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2002).
5. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. From DNA to diversity : molecular genetics and the evolution of animal design (Blackwell Pub., Malden, MA, 2005).
6. Alonso, C. R. & Wilkins, A. S. The molecular elements that underlie developmental evolution. Nat Rev Genet 6, 709-15 (2005).
7. Wray, G. A. et al. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20, 1377-419 (2003).
8. Kirschner, M. & Gerhart, J. Evolvability. Proc Natl Acad Sci U S A 95, 8420-7 (1998).
9. Hsia, C. C. & McGinnis, W. Evolution of transcription factor function. Curr Opin Genet Dev 13, 199-206 (2003).
10. Wray, G. A. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8, 206-16 (2007).
11. Prud'homme, B., Gompel, N. & Carroll, S. B. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104 Suppl 1, 8605-12 (2007).
12. Tishkoff, S. A. et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39, 31-40 (2007).
13. Burger, J., Kirchner, M., Bramanti, B., Haak, W. & Thomas, M. G. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. Proc Natl Acad Sci U S A 104, 3736-41 (2007).
14. Enattah, N. S. et al. Identification of a variant associated with adult-type hypolactasia. Nat Genet 30, 233-7 (2002).
15. Shapiro, M. D. et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428, 717-23 (2004).
16. Miller, C. T. et al. cis-Regulatory Changes in Kit Ligand Expression and Parallel Evolution of Pigmentation in Sticklebacks and Humans. Cell 131, 1179-89 (2007).
17. Prud'homme, B. et al. Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440, 1050-3 (2006).
18. Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A. & Carroll, S. B. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature 433, 481-7 (2005).
19. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403, 564-7 (2000).
20. Ludwig, M. Z. et al. Functional evolution of a cis-regulatory module. PLoS Biol 3, e93 (2005).
21. Bridgham, J. T., Carroll, S. M. & Thornton, J. W. Evolution of hormone-receptor complexity by molecular exploitation. Science 312, 97-101 (2006).

22. Irish, V. F. & Litt, A. Flower development and evolution: gene duplication, diversification and redeployment. Curr Opin Genet Dev 15, 454-60 (2005).
23. Soltis, D. E. et al. The floral genome: an evolutionary history of gene duplication and shifting patterns of gene expression. Trends Plant Sci 12, 358-67 (2007).
24. Ronshaugen, M., McGinnis, N. & McGinnis, W. Hox protein mutation and macroevolution of the insect body plan. Nature 415, 914-7 (2002).
25. Galant, R. & Carroll, S. B. Evolution of a transcriptional repression domain in an insect Hox protein. Nature 415, 910-3 (2002).
26. Tanay, A., Regev, A. & Shamir, R. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci U S A 102, 7203-8 (2005).
27. Martchenko, M., Levitin, A., Hogues, H., Nantel, A. & Whiteway, M. Transcriptional rewiring of fungal galactose-metabolism circuitry. Curr Biol 17, 1007-13 (2007).
28. Borneman, A. R. et al. Divergence of transcription factor binding sites across related yeast species. Science 317, 815-9 (2007).
29. Odom, D. T. et al. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 39, 730-2 (2007).
30. Tsong, A. E., Miller, M. G., Raisner, R. M. & Johnson, A. D. Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell 115, 389-99 (2003).
31. Tsong, A. E., Tuch, B. B., Li, H. & Johnson, A. D. Evolution of alternative transcriptional circuits with identical logic. Nature 443, 415-20 (2006).
32. Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H. & Johnson, A. D. The evolution of combinatorial gene regulation in fungi. PLoS Biol (accepted, 2007).
33. Moses, A. M. et al. Large-scale turnover of functional transcription factor binding sites in Drosophila. PLoS Comput Biol 2, e130 (2006).
34. Darwin, C. R. The Origin of Species (Gramercy, New York, New York, 1859).
35. Carroll, S. B., Grenier J. K., Weatherbee, S. D. From DNA to Diversity (Blackwell Science, Inc., Malden, Massachusetts, 2001).
36. Davidson, E. H. Genomic Regulatory Systems (Academic Press, San Diego, CA, 2001).
37. Gerhart, J. & Kirschner, M. Cells, Embryos, and Evolution (Blackwell Science, Inc., Malden, Massachusetts, 1997).
38. Doebley, J. & Lukens, L. Transcriptional regulators and the evolution of plant form. Plant Cell 10, 1075-82 (1998).
39. Ihmels, J. et al. Rewiring of the yeast transcriptional network through the evolution of motif usage. Science 309, 938-40 (2005).
40. Ludwig, M. Z., Patel, N. H. & Kreitman, M. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development 125, 949-58 (1998).
41. Wittkopp, P. J., Haerum, B. K. & Clark, A. G. Evolutionary changes in cis and trans gene regulation. Nature 430, 85-8 (2004).
42. Hull, C. M., Raisner, R. M. & Johnson, A. D. Evidence for mating of the "asexual" yeast Candida albicans in a mammalian host. Science 289, 307-10 (2000).

43. Herskowitz, I., Rine, J. & Strathern, J. Mating-type determination and Mating-type interconversion in Saccharomyces cerevisiae (eds. Jones, E. W., Pringle, J. R. & Broach, J. R.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1992).
44. Staben, C. & Yanofsky, C. Neurospora crassa a mating-type region. Proc Natl Acad Sci U S A 87, 4917-21 (1990).
45. Kelly, M., Burke, J., Smith, M., Klar, A. & Beach, D. Four mating-type genes control sexual differentiation in the fission yeast. Embo J 7, 1537-47 (1988).
46. Kurischko, C., Schilhabel, M. B., Kunze, I. & Franzl, E. The MATA locus of the dimorphic yeast Yarrowia lipolytica consists of two divergently oriented genes. Mol Gen Genet 262, 180-8 (1999).
47. Philley, M. L. & Staben, C. Functional analyses of the Neurospora crassa MT a-1 mating type polypeptide. Genetics 137, 715-22 (1994).
48. Turgeon, B. G. et al. Cloning and analysis of the mating type genes from Cochliobolus heterostrophus. Mol Gen Genet 238, 270-84 (1993).
49. Butler, G. et al. Evolution of the MAT locus and its Ho endonuclease in yeast species. Proc Natl Acad Sci U S A 101, 1632-7 (2004).
50. Calderone, R. A. Candida and Candidiasis (ed. Calderone, R. A.) (ASM Press, Washington, D.C., 2002).
51. Hedges, S. B. The origin and evolution of model organisms. Nat Rev Genet 3, 838-49 (2002).
52. Bender, A. & Sprague, G. F., Jr. Yeast peptide pheromones, a-factor and alpha-factor, activate a common response mechanism in their target cells. Cell 47, 929-37 (1986).
53. Bennett, R. J., Uhl, M. A., Miller, M. G. & Johnson, A. D. Identification and characterization of a Candida albicans mating pheromone. Mol Cell Biol 23, 8189-201 (2003).
54. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2, 28-36 (1994).
55. Acton, T. B., Mead, J., Steiner, A. M. & Vershon, A. K. Scanning mutagenesis of Mcm1: residues required for DNA binding, DNA bending, and transcriptional activation by a MADS-box protein. Mol Cell Biol 20, 1-11 (2000).
56. Tan, S. & Richmond, T. J. Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. Nature 391, 660-6 (1998).
57. Kjaerulff, S., Dooijes, D., Clevers, H. & Nielsen, O. Cell differentiation by interaction of two HMG-box proteins: Mat1-Mc activates M cell-specific genes in S.pombe by recruiting the ubiquitous transcription factor Ste11 to weak binding sites. Embo J 16, 4021-33 (1997).
58. Smith, D. L. & Johnson, A. D. A molecular mechanism for combinatorial control in yeast: MCM1 protein sets the spacing and orientation of the homeodomains of an alpha 2 dimer. Cell 68, 133-42 (1992).
59. van Beest, M. et al. Sequence-specific high mobility group box factors recognize 10-12-base pair minor groove motifs. J Biol Chem 275, 27266-73 (2000).

60. Care, R. S., Trevethick, J., Binley, K. M. & Sudbery, P. E. The MET3 promoter: a new tool for Candida albicans molecular genetics. Mol Microbiol 34, 792-8 (1999).
61. Rokas, A., Williams, B. L., King, N. & Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798-804 (2003).
62. Byrne, K. P. & Wolfe, K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Res 15, 1456-61 (2005).
63. Cliften, P. et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301, 71-6 (2003).
64. Belloch, C., Querol, A., Garcia, M. D. & Barrio, E. Phylogeny of the genus Kluyveromyces inferred from the mitochondrial cytochrome-c oxidase II gene. Int J Syst Evol Microbiol 50 Pt 1, 405-16 (2000).
65. Wong, S., Butler, G. & Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc Natl Acad Sci U S A 99, 9272-7 (2002).
66. Mead, J., Zhong, H., Acton, T. B. & Vershon, A. K. The yeast alpha2 and Mcm1 proteins interact through a region similar to a motif found in homeodomain proteins of higher eukaryotes. Mol Cell Biol 16, 2135-43 (1996).
67. Jacobson, M.
68. Lengeler, K. B. et al. Signal transduction cascades regulating fungal development and virulence. Microbiol Mol Biol Rev 64, 746-85 (2000).
69. Acton, T. B., Zhong, H. & Vershon, A. K. DNA-binding specificity of Mcm1: operator mutations that alter DNA-bending and transcriptional activities by a MADS box protein. Mol Cell Biol 17, 1881-9 (1997).
70. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 428, 617-24 (2004).
71. Wolfe, K. H. & Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708-13 (1997).
72. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73, 237-44 (1988).
73. Schmidt, H. A., Strimmer, K., Vingron, M. & von Haeseler, A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18, 502-4 (2002).
74. Debuchy, R., Arnaise, S. & Lecellier, G. The mat- allele of Podospora anserina contains three regulatory genes required for the development of fertilized female organs. Mol Gen Genet 241, 667-73 (1993).
75. Wilson, R. B., Davis, D., Enloe, B. M. & Mitchell, A. P. A recyclable Candida albicans URA3 cassette for PCR product-directed gene disruptions. Yeast 16, 65-70 (2000).
76. Cormack, B. P. et al. Yeast-enhanced green fluorescent protein (yEGFP)a reporter of gene expression in Candida albicans. Microbiology 143 (Pt 2), 303-11 (1997).
77. Miller, M. G. & Johnson, A. D. White-opaque switching in Candida albicans is controlled by mating-type locus homeodomain proteins and allows efficient mating. Cell 110, 293-302 (2002).

78. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52, 696-704 (2003).
79. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21, 2104-5 (2005).
80. Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. Nature 440, 341-5 (2006).
81. Dietrich, F. S. et al. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 304, 304-7 (2004).
82. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
83. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302, 205-17 (2000).
84. Sauer, R. T., Smith, D. L. & Johnson, A. D. Flexibility of the yeast alpha 2 repressor enables it to occupy the ends of its operator, leaving the center free. Genes Dev 2, 807-16 (1988).
85. Britten, R. J. & Davidson, E. H. Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Q Rev Biol 46, 111-38 (1971).
86. Jacob, F. Evolution and tinkering. Science 196, 1161-6 (1977).
87. Mayo, A. E., Setty, Y., Shavit, S., Zaslaver, A. & Alon, U. Plasticity of the cis-regulatory input function of a gene. PLoS Biol 4, e45 (2006).
88. Stone, J. R. & Wray, G. A. Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18, 1764-70 (2001).
89. Carter, A. J. & Wagner, G. P. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc Biol Sci 269, 953-60 (2002).
90. MacArthur, S. & Brookfield, J. F. Expected rates and modes of evolution of enhancer sequences. Mol Biol Evol 21, 1064-73 (2004).
91. Rockman, M. V. & Wray, G. A. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 19, 1991-2004 (2002).
92. Dermitzakis, E. T. & Clark, A. G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol Biol Evol 19, 1114-21 (2002).
93. Doniger, S. W. & Fay, J. C. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol 3, e99 (2007).
94. Taylor, J. W. & Berbee, M. L. Dating divergences in the Fungal Tree of Life: review and new analyses. Mycologia 98, 838-49 (2006).
95. Ihmels, J., Bergmann, S., Berman, J. & Barkai, N. Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program. PLoS Genet 1, e39 (2005).
96. Gasch, A. P. et al. Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol 2, e398 (2004).
97. Messenguy, F. & Dubois, E. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. Gene 316, 1-21 (2003).

98. Pramila, T., Miles, S., GuhaThakurta, D., Jemiolo, D. & Breeden, L. L. Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. Genes Dev 16, 3034-45 (2002).
99. Fontana, W. Modelling 'evo-devo' with RNA. Bioessays 24, 1164-77 (2002).
100. Dujon, B. et al. Genome evolution in yeasts. Nature 430, 35-44 (2004).
101. Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. Trends Genet 22, 375-87 (2006).
102. Qi, Y. et al. High-resolution computational models of genome binding events. Nat Biotechnol 24, 963-70 (2006).
103. Wynne, J. & Treisman, R. SRF and MCM1 have related but distinct DNA binding specificities. Nucleic Acids Res 20, 3297-303 (1992).
104. Passmore, S., Elble, R. & Tye, B. K. A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. Genes Dev 3, 921-35 (1989).
105. Koranda, M., Schleiffer, A., Endler, L. & Ammerer, G. Forkhead-like transcription factors recruit Ndd1 to the chromatin of G2/M-specific promoters. Nature 406, 94-8 (2000).
106. Hollenhorst, P. C., Pietz, G. & Fox, C. A. Mechanisms controlling differential promoter-occupancy by the yeast forkhead proteins Fkh1p and Fkh2p: implications for regulating the cell cycle and differentiation. Genes Dev 15, 2445-56 (2001).
107. Hagen, D. C., Bruhn, L., Westby, C. A. & Sprague, G. F., Jr. Transcription of alpha-specific genes in Saccharomyces cerevisiae: DNA sequence requirements for activity of the coregulator alpha 1. Mol Cell Biol 13, 6866-75 (1993).
108. Deutschbauer, A. M. et al. Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169, 1915-25 (2005).
109. Buchman, A. R., Kimmerly, W. J., Rine, J. & Kornberg, R. D. Two DNA-binding factors recognize specific sequences at silencers, upstream activating sequences, autonomously replicating sequences, and telomeres in Saccharomyces cerevisiae. Mol Cell Biol 8, 210-25 (1988).
110. Dubois, E. & Messenguy, F. In vitro studies of the binding of the ARGR proteins to the ARG5,6 promoter. Mol Cell Biol 11, 2162-8 (1991).
111. Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99-104 (2004).
112. Rottmann, M., Dieter, S., Brunner, H. & Rupp, S. A screen in Saccharomyces cerevisiae identified CaMCM1, an essential gene in Candida albicans crucial for morphogenesis. Mol Microbiol 47, 943-59 (2003).
113. Zordan, R. E., Miller, M. G., Galgoczy, D. J., Tuch, B. B. & Johnson, A. D. Interlocking Transcriptional Feedback Loops Control White-Opaque Switching in Candida albicans. PLoS Biol 5, e256 (2007).
114. Sonneborn, A., Tebarth, B. & Ernst, J. F. Control of white-opaque phenotypic switching in Candida albicans by the Efg1p morphogenetic regulator. Infect Immun 67, 4655-60 (1999).
115. Slutsky, B. et al. "White-opaque transition": a second high-frequency switching system in Candida albicans. J Bacteriol 169, 189-97 (1987).

116. Kvaal, C. A., Srikantha, T. & Soll, D. R. Misexpression of the white-phase-specific gene WH11 in the opaque phase of Candida albicans affects switching and virulence. Infect Immun 65, 4468-75 (1997).
117. Lachke, S. A., Lockhart, S. R., Daniels, K. J. & Soll, D. R. Skin facilitates Candida albicans mating. Infect Immun 71, 4970-6 (2003).
118. Zordan, R. E., Galgoczy, D. J. & Johnson, A. D. Epigenetic properties of white-opaque switching in Candida albicans are based on a self-sustaining transcriptional feedback loop. Proc Natl Acad Sci U S A 103, 12807-12 (2006).
119. Huang, G. et al. Bistable expression of WOR1, a master regulator of white-opaque switching in Candida albicans. Proc Natl Acad Sci U S A 103, 12813-8 (2006).
120. Srikantha, T. et al. TOS9 regulates white-opaque switching in Candida albicans. Eukaryot Cell 5, 1674-87 (2006).
121. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 423, 241-54 (2003).
122. Chin, C. S., Chuang, J. H. & Li, H. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. Genome Res 15, 205-13 (2005).
123. Lynch, M. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci U S A 104 Suppl 1, 8597-604 (2007).
124. McClintock, B. The discovery and characterization of transposable elements : the collected papers of Barbara McClintock (Garland Pub., New York, 1987).
125. Shapiro, J. A. Retrotransposons and regulatory suites. Bioessays 27, 122-5 (2005).
126. Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449, 677-81 (2007).
127. Pujol, C. et al. The closely related species Candida albicans and Candida dubliniensis can mate. Eukaryot Cell 3, 1015-27 (2004).
128. Buchler, N. E., Gerland, U. & Hwa, T. On schemes of combinatorial transcription logic. Proc Natl Acad Sci U S A 100, 5136-41 (2003).
129. Zuckerkandl, E. Neutral and nonneutral mutations: the creative mix--evolution of complexity in gene interaction systems. J Mol Evol 44 Suppl 1, S2-8 (1997).
130. Spellman, P. T. et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9, 3273-97 (1998).
131. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. Genome Res 14, 1188-90 (2004).
132. Bozdech, Z. et al. Expression profiling of the schizont and trophozoite stages of Plasmodium falciparum with a long-oligonucleotide microarray. Genome Biol 4, R9 (2003).
133. Carter, S. D., Iyer, S., Xu, J., McEachern, M. J. & Astrom, S. U. The role of nonhomologous end-joining components in telomere metabolism in Kluyveromyces lactis. Genetics 175, 1035-45 (2007).
134. Bennett, R. J. & Johnson, A. D. The role of nutrient regulation and the Gpa2 protein in the mating pheromone response of C. albicans. Mol Microbiol 62, 100-19 (2006).

135. Jarvis, E. E., Clark, K. L. & Sprague, G. F., Jr. The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the MCM1 gene. Genes Dev 3, 936-45 (1989).
136. Roberts, C. J. et al. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science 287, 873-80 (2000).
137. Ren, B. et al. Genome-wide location and function of DNA binding proteins. Science 290, 2306-9 (2000).
138. Lemoine, S., Combes, F., Servant, N. & Le Crom, S. Goulphar: rapid access and expertise for standard two-color microarray normalization methods. BMC Bioinformatics 7, 467 (2006).
139. Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 83, 349-60 (2004).
140. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32, 1792-7 (2004).
141. Galgoczy, D. J. et al. Genomic dissection of the cell-type-specification circuit in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A 101, 18069-74 (2004).
142. (Genbank, ftp://ftp.ncbi.nih.gov/genomes/Fungi).
143. Langkjaer, R. B., Cliften, P. F., Johnston, M. & Piskur, J. Yeast genome duplication was followed by asynchronous differentiation of duplicated genes. Nature 421, 848-52 (2003).
144. (Sanger Centre, http://www.sanger.ac.uk/Projects/Fungi/).
145. van het Hoog, M. et al. Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol 8, R52 (2007).
146. Mitrovich, Q. M., Tuch, B. B., Guthrie, C. & Johnson, A. D. Computational and experimental approaches double the number of known introns in the pathogenic yeast Candida albicans. Genome Res 17, 492-502 (2007).
147. (Broad Fungal Genome Initiative, http://www.broad.mit.edu/annotation/fgi/).
148. (Joint Genome Institute, http://genome.jgi-psf.org/Trire2/Trire2.home.html).
149. Pic, A. et al. The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. Embo J 19, 3750-61 (2000).
150. Mead, J. et al. Interactions of the Mcm1 MADS box protein with cofactors that regulate mating in yeast. Mol Cell Biol 22, 4607-21 (2002).
151. Zhong, H., McCord, R. & Vershon, A. K. Identification of target sites of the alpha2-Mcm1 repressor complex in the yeast genome. Genome Res 9, 1040-7 (1999).
152. Yoon, S. et al. Recruitment of the ArgR/Mcm1p repressor is stimulated by the activator Gcn4p: a self-checking activation mechanism. Proc Natl Acad Sci U S A 101, 11713-8 (2004).
153. Messenguy, F., Dubois, E. & Boonchird, C. Determination of the DNA-binding sequences of ARGR proteins to arginine anabolic and catabolic promoters. Mol Cell Biol 11, 2852-63 (1991).
154. Juang, Y. L. et al. APC-mediated proteolysis of Ase1 and the morphogenesis of the mitotic spindle. Science 275, 1311-4 (1997).

155. Sanders, S. L. & Herskowitz, I. The BUD4 protein of yeast, required for axial budding, is localized to the mother/BUD neck in a cell cycle-dependent manner. J Cell Biol 134, 413-27 (1996).
156. Kuo, M. H. & Grayhack, E. A library of yeast genomic MCM1 binding sites contains genes involved in cell cycle control, cell wall and membrane structure, and metabolism. Mol Cell Biol 14, 348-59 (1994).
157. Fitch, M. J., Donato, J. J. & Tye, B. K. Mcm7, a subunit of the presumptive MCM helicase, modulates its own expression in conjunction with Mcm1. J Biol Chem 278, 25408-16 (2003).
158. Althoefer, H., Schleiffer, A., Wassmann, K., Nordheim, A. & Ammerer, G. Mcm1 is required to coordinate G2-specific transcription in Saccharomyces cerevisiae. Mol Cell Biol 15, 5917-28 (1995).
159. McInerny, C. J., Partridge, J. F., Mikesell, G. E., Creemer, D. P. & Breeden, L. L. A novel Mcm1-dependent element in the SWI4, CLN3, CDC6, and CDC47 promoters activates M/G1-specific transcription. Genes Dev 11, 1277-88 (1997).
160. Oehlen, L. J., McKinney, J. D. & Cross, F. R. Ste12 and Mcm1 regulate cell cycle-dependent transcription of FAR1. Mol Cell Biol 16, 2830-7 (1996).
161. Zhong, H. & Vershon, A. K. The yeast homeodomain protein MATalpha2 shows extended DNA binding specificity in complex with Mcm1. J Biol Chem 272, 8402-9 (1997).
162. Gavin, I. M., Kladde, M. P. & Simpson, R. T. Tup1p represses Mcm1p transcriptional activation and chromatin remodeling of an a-cell-specific gene. Embo J 19, 5875-83 (2000).
163. Lynch, M. The evolution of genetic networks by non-adaptive processes. Nat Rev Genet 8, 803-13 (2007).
164. Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A 102, 1572-7 (2005).
165. Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. Science 296, 752-5 (2002).
166. Wang, T. et al. Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc Natl Acad Sci U S A 104, 18613-8 (2007).
167. Romano, L. A. & Wray, G. A. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. Development 130, 4187-99 (2003).
168. Rokas, A. & Hittinger, C. T. Transcriptional rewiring: the proof is in the eating. Curr Biol 17, R626-8 (2007).
169. Bilu, Y. & Barkai, N. The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biol 6, R103 (2005).
170. Madhani, H. D. From a to alpha : yeast as a model for cellular differentiation (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 2007).
171. Tatusov, R. L. et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41 (2003).
172. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. Science 278, 631-7 (1997).

173. Rost, B. Enzyme function less conserved than anticipated. J Mol Biol 318, 595-608 (2002).
174. Tian, W. & Skolnick, J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333, 863-82 (2003).
175. Boffelli, D. et al. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. Science 299, 1391-4 (2003).
176. Elemento, O. & Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biol 6, R18 (2005).
177. McCarroll, S. A. et al. Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. Nat Genet 36, 197-204 (2004).
178. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249-55 (2003).
179. Breeden, L. L. Periodic transcription: a cycle within a cycle. Curr Biol 13, R31-8 (2003).
180. Bussemaker, H. J., Li, H. & Siggia, E. D. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. Proc Natl Acad Sci U S A 97, 10096-100 (2000).
181. Wang, D. Y., Kumar, S. & Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. Proc Biol Sci 266, 163-71 (1999).
182. Gasch, A. P. et al. Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11, 4241-57 (2000).
183. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics 15, 607-11 (1999).
184. Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31, 374-8 (2003).
185. Lowndes, N. F., Johnson, A. L. & Johnston, L. H. Coordination of expression of DNA synthesis genes in budding yeast by a cell-cycle regulated trans factor. Nature 350, 247-50 (1991).
186. Lowndes, N. F., McInerny, C. J., Johnson, A. L., Fantes, P. A. & Johnston, L. H. Control of DNA synthesis genes in fission yeast by the cell-cycle gene cdc10+. Nature 355, 449-53 (1992).
187. Koch, C., Moll, T., Neuberg, M., Ahorn, H. & Nasmyth, K. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science 261, 1551-7 (1993).
188. McIntosh, E. M., Looser, J., Haynes, R. H. & Pearlman, R. E. MluI site-dependent transcriptional regulation of the Candida albicans dUTPase gene. Curr Genet 26, 415-21 (1994).
189. Hirose, F., Yamaguchi, M., Handa, H., Inomata, Y. & Matsukage, A. Novel 8-base pair sequence (Drosophila DNA replication-related element) and specific binding factor involved in the expression of Drosophila genes for DNA polymerase alpha and proliferating cell nuclear antigen. J Biol Chem 268, 2092-9 (1993).
190. Hirose, F., Yamaguchi, M. & Matsukage, A. Targeted expression of the DNA binding domain of DRE-binding factor, a Drosophila transcription factor,

attenuates DNA replication of the salivary gland and eye imaginal disc. Mol Cell Biol 19, 6020-8 (1999).
191. Sawado, T. et al. The DNA replication-related element (DRE)/DRE-binding factor system is a transcriptional regulator of the Drosophila E2F gene. J Biol Chem 273, 26042-51 (1998).
192. Han, J. et al. Expression of bbc3, a pro-apoptotic BH3-only gene, is regulated by diverse cell death and survival signals. Proc Natl Acad Sci U S A 98, 11318-23 (2001).
193. Imbriano, C., Bolognese, F., Gurtner, A., Piaggio, G. & Mantovani, R. HSP-CBF is an NF-Y-dependent coactivator of the heat shock promoters CCAAT boxes. J Biol Chem 276, 26332-9 (2001).
194. Yun, J. et al. p53 negatively regulates cdc2 transcription via the CCAAT-binding NF-Y transcription factor. J Biol Chem 274, 29677-82 (1999).
195. Lydall, D., Ammerer, G. & Nasmyth, K. A new role for MCM1 in yeast: cell cycle regulation of SW15 transcription. Genes Dev 5, 2405-19 (1991).
196. Payre, F., Noselli, S., Lefrere, V. & Vincent, A. The closely related Drosophila sry beta and sry delta zinc finger proteins show differential embryonic expression and distinct patterns of binding sites on polytene chromosomes. Development 110, 141-9 (1990).
197. Torriani-Gorini, A., Yagil, E. & Silver, S. Phosphate in microorganisms : cellular and molecular biology (ASM Press, Washington, DC, 1994).
198. O'Connell, K. F., Surdin-Kerjan, Y. & Baker, R. E. Role of the Saccharomyces cerevisiae general regulatory factor CP1 in methionine biosynthetic gene transcription. Mol Cell Biol 15, 1879-88 (1995).
199. O'Connell, K. F. & Baker, R. E. Possible cross-regulation of phosphate and sulfate metabolism in Saccharomyces cerevisiae. Genetics 132, 63-73 (1992).
200. Unger, M. W. & Hartwell, L. H. Control of cell division in Saccharomyces cerevisiae by methionyl-tRNA. Proc Natl Acad Sci U S A 73, 1664-8 (1976).
201. Gasch, A. P. et al. Genomic expression responses to DNA-damaging agents and the regulatory role of the yeast ATR homolog Mec1p. Mol Biol Cell 12, 2987-3003 (2001).
202. Shore, D. & Nasmyth, K. Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. Cell 51, 721-32 (1987).
203. Moehle, C. M. & Hinnebusch, A. G. Association of RAP1 binding sites with stringent control of ribosomal protein gene transcription in Saccharomyces cerevisiae. Mol Cell Biol 11, 2723-35 (1991).
204. Witt, I., Kwart, M., Gross, T. & Kaufer, N. F. The tandem repeat AGGGTAGGGT is, in the fission yeast, a proximal activation sequence and activates basal transcription mediated by the sequence TGTGACTG. Nucleic Acids Res 23, 4296-302 (1995).
205. Witt, I., Straub, N., Kaufer, N. F. & Gross, T. The CAGTCACA box in the fission yeast Schizosaccharomyces pombe functions like a TATA element and binds a novel factor. Embo J 12, 1201-8 (1993).
206. Haun, R. S., Moss, J. & Vaughan, M. Characterization of the human ADP-ribosylation factor 3 promoter. Transcriptional regulation of a TATA-less promoter. J Biol Chem 268, 8793-800 (1993).

207. Mencia, M., Moqtaderi, Z., Geisberg, J. V., Kuras, L. & Struhl, K. Activator-specific recruitment of TFIID and regulation of ribosomal protein genes in yeast. Mol Cell 9, 823-33 (2002).
208. Wade, C., Shea, K. A., Jensen, R. V. & McAlear, M. A. EBP2 is a member of the yeast RRB regulon, a transcriptionally coregulated set of genes that are required for ribosome and rRNA biosynthesis. Mol Cell Biol 21, 8638-50 (2001).
209. Patil, C. K., Li, H. & Walter, P. Gcn4p and novel upstream activating sequences regulate targets of the unfolded protein response. PLoS Biol 2, E246 (2004).
210. Murphy, C. T. et al. Genes that act downstream of DAF-16 to influence the lifespan of Caenorhabditis elegans. Nature 424, 277-83 (2003).
211. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature 434, 338-45 (2005).
212. Kurtz, S. et al. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29, 4633-42 (2001).
213. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. J Comput Biol 6, 281-97 (1999).
214. Boyle, E. I. et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20, 3710-5 (2004).

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

February 21, 2008

Author Signature | Date