# Extensive DNA-binding specificity divergence of a conserved transcription regulator

Christopher R. Baker[a], Brian B. Tuch[a,b,c,1], and Alexander D. Johnson[a,b,2]

Departments of [a]Biochemistry and Biophysics and [b]Microbiology and Immunology, University of California, San Francisco, CA 94143-2200; and [c]Genome Analysis Unit, Amgen, South San Francisco, CA 94080

Edited by Mark Ptashne, Memorial Sloan–Kettering Cancer Center, New York, NY, and approved March 18, 2011 (received for review December 20, 2010)

**The DNA sequence recognized by a transcription regulator can be conserved across large evolutionary distances. For example, it is known that many homologous regulators in yeasts and mammals can recognize the same (or closely related) DNA sequences. In contrast to this paradigm, we describe a case in which the DNA-binding specificity of a transcription regulator has changed so extensively (and over a much smaller evolutionary distance) that its *cis*-regulatory sequence appears unrelated in different species. Bioinformatic, genetic, and biochemical approaches were used to document and analyze a major change in the DNA-binding specificity of Matα1, a regulator of cell-type specification in ascomycete fungi. Despite this change, Matα1 controls the same core set of genes in the hemiascomycetes because its DNA recognition site has evolved with it, preserving the protein-DNA interaction but significantly changing its molecular details. Matα1 and its recognition sequence diverged most dramatically in the common ancestor of the CTG-clade (*Candida albicans, Candida lusitaniae,* and related species), apparently without the aid of a gene duplication event. Our findings suggest that DNA-binding specificity divergence between orthologous transcription regulators may be more prevalent than previously thought and that seemingly unrelated *cis*-regulatory sequences can nonetheless be homologous. These findings have important implications for understanding transcriptional network evolution and for the bioinformatic analysis of regulatory circuits.**

transcription regulation | DNA-binding protein | transcription factor | evolution of gene expression

The importance of changes in the DNA-binding specificity of orthologous transcription regulators to the evolution of transcriptional networks is an open question. Several lines of evidence have been used to argue that divergence in transcription regulator DNA-binding specificity occurs infrequently. These arguments include the amino acid conservation of transcription regulator DNA-binding domains (1), the potentially pleiotropic nature of alterations to transcription regulator DNA-binding specificity (2), and the conservation of function across large evolutionary distances for certain transcription regulators (3, 4). Several cases of drift in the transcription regulator DNA-binding specificity have been documented across species, but the changes were limited to a small number of amino acid positions and the *cis*-regulatory sequence remained similar across species (5, 6). Here, we show that the DNA-binding specificity of a deeply conserved transcription regulator (Matα1) can change so extensively that its *cis*-regulatory sequence in different species appears unrelated as assessed by bioinformatic criteria.

In the model yeast *Saccharomyces cerevisiae,* the HMG DNA-binding domain transcription regulator Matα1 activates a set of genes involved in cell-type (mating-type) specification, known as the α-specific genes (αsgs). Matα1 associates with αsg promoters through direct sequence-specific DNA binding aided by a protein-protein interaction with a second sequence-specific DNA-binding protein, Mcm1 (7, 8). This basic form of αsg regulation appears to be conserved in the pathogenic yeast *Candida albicans*, which is estimated to have diverged between 100 and 300 Mya from the lineage that gave rise to *S. cerevisiae* (9). For example, deletion of the Matα1 ortholog in *C. albicans* results in a loss of αsg expression, and the *C. albicans* Mcm1 ortholog has been shown to bind αsg promoters (10, 11). Despite the overall similarity of the regulatory scheme, the *cis*-regulatory DNA sequences that regulate the αsgs have diverged substantially between the two yeasts (11). Here, we demonstrate that the source of this divergence is the extensive evolution of Matα1 DNA-binding specificity.

## Results

**Significant Divergence of the αsg *cis*-Regulatory Sequence Between *C. albicans* and *S. cerevisiae*.** To computationally demonstrate the divergence of the αsg *cis*-regulatory DNA sequences between *C. albicans* and *S. cerevisiae*, position-specific scoring matrices (PSSMs) for αsg *cis*-regulatory sequences were determined for the *S. cerevisiae* and *C. albicans* clades (Fig. 1*A*). For this study, we define the *S. cerevisiae* clade as encompassing *S. cerevisiae, Saccharomyces bayanus, Saccharomyces mikatae,* and *Saccharomyces paradoxus* (12) and the *C. albicans* clade as *C. albicans, Candida tropicalis,* and *Candida dubliniensis* (13). The extent of divergence between the two PSSMs was then measured, revealing significant differences between the αsg *cis*-regulatory sequences of the *C. albicans* and *S. cerevisiae* clades (Fig. 1*B*). Although the Mcm1-binding site was strongly conserved between the two clades ($E = 0.0016$; *Materials and Methods*), the adjacent sequence (known to be recognized by Matα1 in *S. cerevisiae*) was not conserved ($E > 1{,}200$). Instead, the *C. albicans* clade appeared to have a different binding site in the same position.

At least three models can be invoked to explain this divergence. In the first model, "regulatory protein substitution," a transcription regulator other than Matα1, recognizes the motif adjacent to the Mcm1 site within the *C. albicans* αsg *cis*-regulatory sequence. According to this model, the synthesis of this other transcription regulator would depend on Matα1, thereby preserving the regulatory logic (14). In the second model, "binding specificity divergence," the binding specificity of Matα1 would have coevolved with its binding site to such an extent that the two binding sites no longer appear related by a standard criterion. In the third model, the Matα1 protein would possess a relaxed specificity enabling it to recognize both *cis*-regulatory sequences.

***C. albicans* Matα1 Activates Transcription by Binding to the *C. albicans* αsg *cis*-Regulatory Sequences.** To distinguish between these possibilities, we ectopically expressed *C. albicans* Matα1 in *S. cerevisiae* MAT**a** cells (which lack *S. cerevisiae MATα1*) and

Author contributions: C.R.B., B.B.T., and A.D.J. designed research; C.R.B. performed research; C.R.B. and B.B.T. analyzed data; and C.R.B., B.B.T., and A.D.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession nos. AEAS00000000 (*Kluyveromyces aestuarii*) and AEAV00000000 (*Kluyveromyces wickerhamii*)].

[1]Present address: Genome Analysis Unit, Amgen, South San Francisco, CA 94080.

[2]To whom correspondence should be addressed. E-mail: ajohnson@cgl.ucsf.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1019177108/-/DCSupplemental.

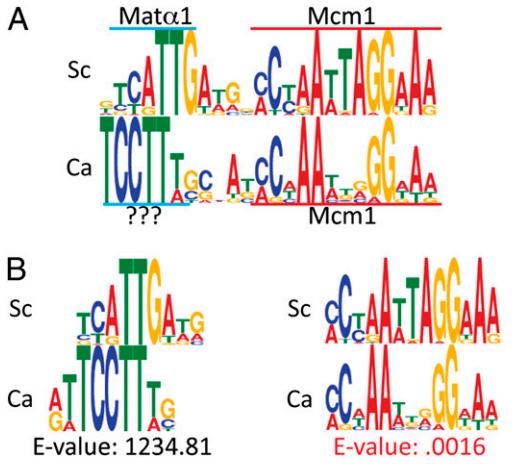


**Fig. 1.** Significant divergence of the αsg *cis*-regulatory sequence between *C. albicans* and *S. cerevisiae*. (*A*) PSSM for the *S. cerevisiae* clade αsg *cis*-regulatory sequence (Sc) was derived using MEME from 27 sequences identified in either the promoters of known *S. cerevisiae* αsgs (42) or the promoters of the orthologous genes in *S. mikatae, S. paradoxus,* and *S. bayanus*. The PSSM for the *C. albicans* clade αsg *cis*-regulatory sequence (Ca) was derived using MEME from 12 sequences that originated from either *C. albicans* αsg promoter sequences (10) or promoters of the orthologous genes in *C. tropicalis* and *C. dubliniensis*. (*B*) Alignments of the *S. cerevisiae* Matα1 motif to the unknown motif within the *C. albicans* αsg *cis*-regulatory sequence (*Left*) and the αsg Mcm1 motif from *S. cerevisiae* and *C. albicans* (*Right*). Motif alignments and *E* values were calculated using MochiView (30), which quantifies similarities between motifs by using an algorithm derived from Gupta et al. (32).

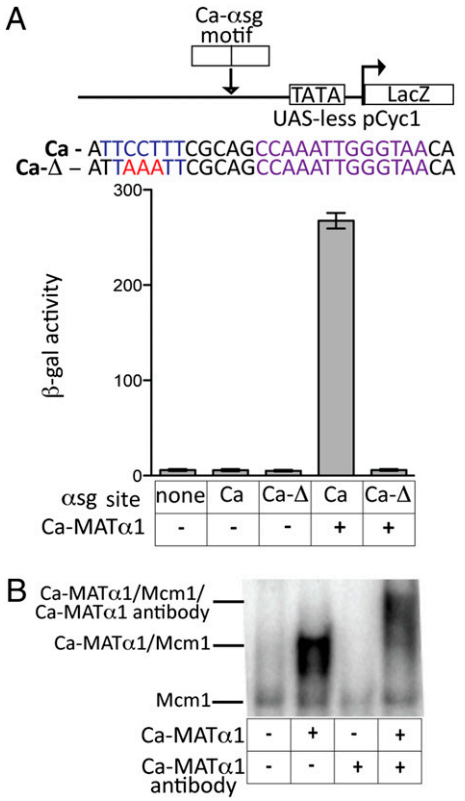


**Fig. 2.** *C. albicans* (Ca) Matα1 activates transcription by binding to the *C. albicans* αsg *cis*-regulatory sequences. (*A*) A *C. albicans* αsg *cis*-regulatory sequence taken from the α-mating pheromone gene was inserted into a basal promoter construct upstream of a β-gal reporter (pLG669Z). The same *C. albicans* αsg *cis*-regulatory sequence was also mutated to alter the residues at the position where Matα1 binds to the *S. cerevisiae cis*-regulatory sequence (Ca-Δ). These constructs were introduced into *S. cerevisiae* MAT**a** cells (MAT**a** cells lack *S. cerevisiae MATα1*). In the two right lanes, strains also contain a 415-translation elongation factor promoter (TEF) plasmid modified to express a codon-changed *C. albicans* Matα1 (the codon changes were necessary because *C. albicans* decodes the CUG codon as serine and most other species, including *S. cerevisiae*, decode it as a leucine). Reporter activity was monitored using β-galactosidase assays. For each sample, $n = 5$ and error bars represent SE. (*B*) Electrophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts. The labeled oligonucleotide used in this experiment was the *C. albicans* αsg *cis*-regulatory sequence described in *A*. Extracts were prepared from an *S. cerevisiae* MAT**a** strain containing a galactose-inducible copy of the codon-changed *C. albicans* Matα1. Each lane contains 5 mg of protein from cell extracts. Galactose induction was performed overnight on samples in lanes 2 and 4 (lanes 1 and 3 are grown in glucose, turning off *C. albicans* Matα1 expression). In lanes 3 and 4, an N-terminal peptide antibody against *C. albicans* Matα1 (Bethyl Laboratories) was used to confirm that DNA-binding activity was attributable to the *C. albicans* Matα1 protein.

assessed its ability to activate transcription from a *C. albcians* αsg *cis*-regulatory sequence. We observed strong transcriptional activation by the *C. albicans* Matα1 that depended on the presence of the sequence adjacent to the Mcm1 site (Fig. 2*A*), as well as the Mcm1 site itself (Fig. S1). These results indicate that *C. albicans* Matα1 can activate transcription by binding directly to the *C. albicans* αsg *cis*-regulatory sequence. To confirm this observation, we expressed high levels of *C. albicans* Matα1 in *S. cerevisiae* MAT**a** cells and showed by electrophoretic mobility gel shift assays on cell extracts that *C. albicans* Matα1 bound a *C. albicans* αsg *cis*-regulatory sequence; incubation of the sample with a *C. albicans* Matα1 peptide antibody resulted in a supershift (Fig. 2*B*). Taken together, these results rule out the protein-substitution model.

**Extensive DNA-Binding Specificity Divergence of the Matα1 Protein.** We next addressed whether the lack of similarity between *S. cerevisiae* and *C. albicans* Matα1-binding sites reflected a true difference in the DNA-binding specificity between the two orthologs, as opposed to a relaxed Matα1 DNA-binding specificity that allows for the recognition of both sequences. We measured the ability of the *S. cerevisiae* and *C. albicans* Matα1 proteins to activate transcription from both the *S. cerevisiae* and *C. albicans* αsg *cis*-regulatory sequences and found that Matα1 efficiently activated transcription only from the αsg *cis*-regulatory sequence of its own species (Fig. 3*A*). These findings were verified by electrophoretic gel shift assays using *S. cerevisiae* cell extracts containing either ectopically expressed *S. cerevisiae* Matα1 or *C. albicans* Matα1 (Fig. 3*B*).

The experiments described above were performed using the *cis*-regulatory sequences from a particular α-specific gene (α-mating pheromone gene), but the same results were obtained for another set of *cis*-regulatory sequences taken from the promoters of another α-specific gene (mating **a**-factor receptor gene) (Fig. S2). Additional constructs ruled out the possibility that small differences in the Mcm1-binding site could be contributing to species specificity of Matα1 binding (Fig. 3*C*). Taken together, these experiments lead to the conclusion that the Matα1 protein has undergone a substantial change in its DNA-binding specificity.

**DNA-Binding Specificity of the *C. albicans* Matα1 Protein Evolved After the Divergence of *S. cerevisiae* and *C. albicans*** When in the evolutionary history of the hemiascomycetes did the change in Matα1 DNA-binding specificity occur? To address this question, orthologs of the *S. cerevisiae* and *C. albicans* α-specific genes were identified across all available genome-sequenced yeasts. This analysis includes two newly sequenced fungal genomes: *Kluyveromyces wickerhamii* and *Kluyveromyces aestuarii* (*Materials and Methods*). When an unambiguous ortholog could be identified, it was then determined (using PSSMs) whether an *S. cerevisiae*-like or *C. albicans*-like αsg *cis*-regulatory sequence was present in the orthologous αsg promoters. The *S. cerevisiae*-like

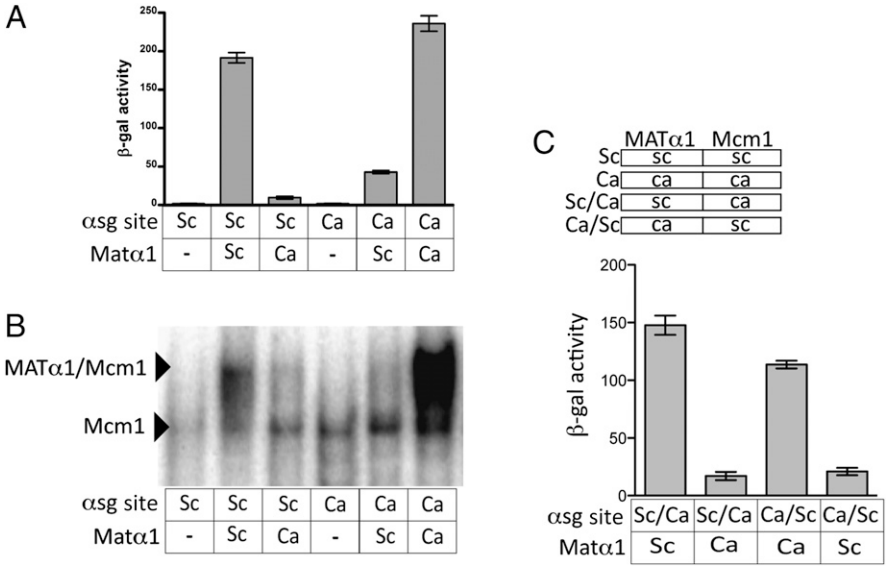


**Fig. 3.** Extensive DNA-binding specificity divergence of the Matα1 protein. (*A*) αsg *cis*-regulatory sequence of the promoter for the α-mating pheromone from *C. albicans* (Ca) or from *S. cerevisiae* (Sc) was inserted into a basal promoter construct (pLG669z). These constructs were introduced into *S. cerevisiae* MATα Δ*matα1* cells along with a 415-TEF plasmid modified to express *S. cerevisiae MATα1* (columns 2 and 5) or a 415-TEF plasmid modified to express the codon-changed *C. albicans MATα1* (columns 3 and 6). Reporter activity was monitored using β-galactosidase assays. For each sample, $n = 5$ and error bars represent SE. (*B*) Electrophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts. The labeled oligonucleotide used in this experiment was either the *C. albicans* αsg *cis*-regulatory sequence (lanes 4–6) or *S. cerevisiae* αsg *cis*-regulatory sequence (lanes 1–3), both of which are described in *A*. Extracts were prepared from either *S. cerevisiae* MAT**a** cells containing a galactose-inducible copy of *C. albicans MATα1* or *S. cerevisiae* MATα cells containing a galactose-inducible copy of the *S. cerevisiae MATα1* (p415GAL). Galactose induction was performed overnight on samples in lanes 2, 3, 5, and 6 (lanes 1 and 4 are grown in glucose). Each lane contains 5 mg of protein from cell extracts. (*C*) To create the Ca/Sc hybrid construct, the Matα1-binding site from the Ca reporter construct was used to replace the Matα1-binding site in the Sc reporter construct. To create the Sc/Ca hybrid construct, the Matα1-binding *Saccharomyces* site from the Sc reporter construct was used to replace the Matα1-binding site in the Ca reporter construct. Reporter activity was monitored using β-galactosidase assays.

*cis*-regulatory sequence appears to be present as early as the common ancestor of *S. cerevisiae* and *Kluyveromyces lactis* (Fig. 4*A*), a result that was experimentally corroborated using the *K. lactis* Matα1 protein (Fig. S3).

The *C. albicans*-like sequence appears to be largely conserved across the CTG-clade (e.g., *C. albicans, Debaryomyces hansenii*). Proceeding outward along the phylogenetic tree, we found matches to the *S. cerevisiae cis*-regulatory sequence in the filamentous fungi (e.g., *Aspergillus terreus*, *Sclerotinia sclerotiorum*), an outgroup to both the *Candida* and *Saccharomyces* lineages. In fact, the filamentous fungi αsg *cis*-regulatory sequence (derived from the promoters of all identifiable orthologs to either *C. albicans* or *S. cerevisaie* αsgs) closely resembles the *S. cerevisiae* clade αsg *cis*-regulatory sequence (Fig. 4*B*). This analysis indicates that the common ancestor to *S. cerevisiae*, *C. albicans*, and the filamentous fungi may have had a Matα1 DNA-binding specificity similar to that of the modern *S. cerevisiae* protein and that the binding specificity of the modern *C. albicans* Matα1 changed along the evolutionary path to the common ancestor of the CTG-clade. We tested this hypothesis directly by moving an αsg *cis*-regulatory sequence from a filamentous fungus (*Uncinocarpus reesii*) into *S. cerevisiae* (15). Expression was efficiently activated from this sequence by the *S. cerevisiae* Matα1 and only weakly activated by the *C. albicans* Matα1 (Fig. 4*C*), consistent with the idea that the ancestral Matα1 protein possessed an *S. cerevisiae*-like DNA-binding specificity and that the most dramatic specificity change occurred in the common ancestor of the CTG-clade.

Even within the CTG-clade, however, the Matα1 DNA-binding specificity did not remain constant. *Candida lusitaniae* showed significant differences from *C. albicans* in its *cis*-regulatory sequences (Fig. 4*A*). In addition, the HMG DNA-binding domain of the *C. lusitaniae* Matα1 is the most divergent amino-acid sequence among the CTG-clade Matα1 orthologs (Fig. S4). To test whether these differences have consequences, we ectopically expressed *C. lusitaniae* Matα1 in *S. cerevisiae* and determined whether it could activate transcription from *cis*-regulatory sequences from *C. lusitaniae*, *S. cerevisiae*, or *C. albicans*. Matα1 from *C. lusitaniae* efficiently activated transcription only from its own species *cis*-regulatory sequence (Fig. 5*B*). This result indicates that Matα1 DNA-binding specificity has undergone additional changes within the CTG-clade. We also note that the αsg *cis*-regulatory sequence in *Yarrowia lipolytica* does not resemble the *C. albicans* or *S. cerevisiae* PSSM, suggesting yet another specificity change within that lineage (Fig. 4 and Fig. S5).

## Discussion

We have combined bioinformatic, genetic, and biochemical experiments to demonstrate a substantial change in the DNA-binding specificity of a deeply conserved transcription regulator. Matα1 (an HMG domain protein) and its recognition sequence appear to have diverged substantially across the ascomycete lineage. The most dramatic changes likely occurred in the common ancestor of the CTG-clade (e.g., *C. albicans, D. hansenii*). One manifestation of this change is that the DNA sequences recognized by Matα1 from *C. albicans* appear unrelated to those recognized by its *S. cerevisiae* ortholog. The divergence of Matα1 DNA-binding specificity is not limited to a single phylogenetic branch point, indicating that the divergence of Matα1 DNA-binding specificity has occurred multiple times.

**Insights into Transcription Regulator DNA-Binding Specificity Divergence.** Several examples of transcription regulator DNA-binding specificity evolution have been linked to gene duplications (16, 17), which are hypothesized to permit drift in DNA-binding specificity by relaxing negative selection (18). The evolution of Matα1 DNA-binding specificity demonstrates that DNA-binding specificity can extensively diverge apparently in the absence of gene duplication. Matα1 orthologs can be easily traced through-

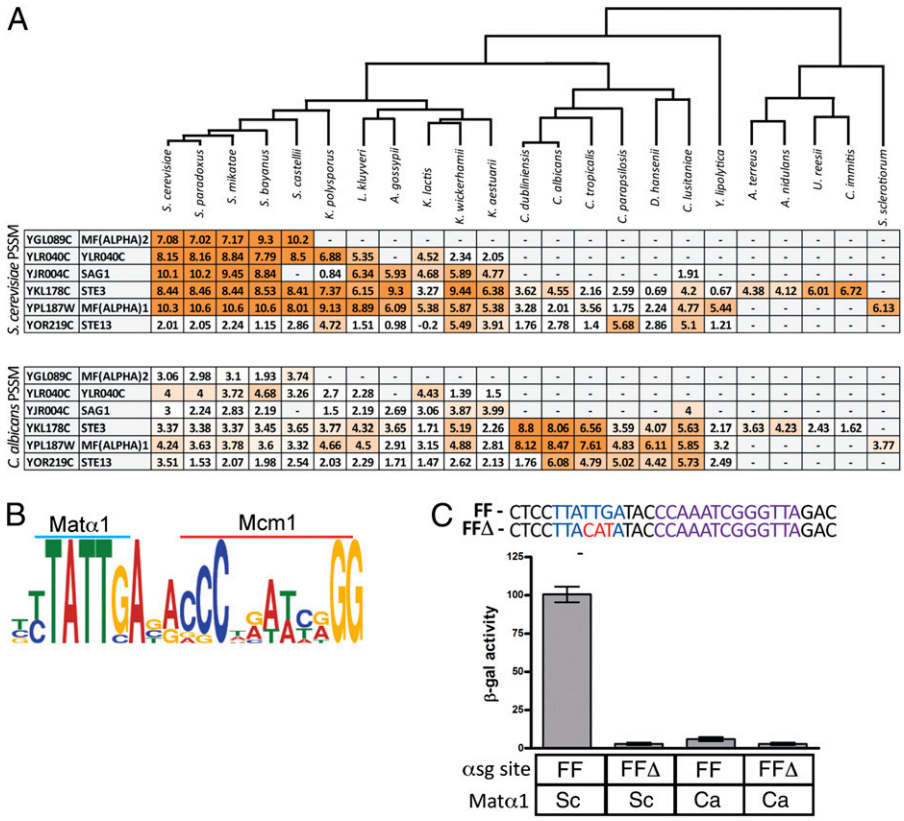


**Fig. 4.** DNA-binding specificity of the *C. albicans* Matα1 protein evolved after the divergence of *S. cerevisiae* and *C. albicans*. (*A*) Orthologs of the *S. cerevisiae* and *C. albicans* αsgs were mapped across 38 genome-sequenced yeasts (10, 11, 13, 28, 46, 48). Where a clear ortholog could be detected, the promoters of these orthologs were scanned with either the *S. cerevisiae* or *C. albicans* clade αsg *cis*-regulatory sequence PSSM (created as described in Fig. 1*A*). Maximum $\log_{10}$ odds scores are shown. Darker shades of orange indicate a stronger match to the PSSM. One-to-one orthologs become more difficult to detect with greater evolution distance, hence, the small number of orthologs identified in the filamentous fungi (e.g., *S. sclerotiorum*, *A. terreus*). (*B*) PSSM for the filamentous fungi αsg *cis*-regulatory sequence was derived using MEME from nine sequences identified in the promoters of αsg orthologs in the filamentous fungi species *U. reesii, C. immitis, F. graminea, A. terreus, A. nidulans,* and *S. sclerotiorum*. (*C*) Putative αsg *cis*-regulatory sequence from the promoter of the *STE3* ortholog in the filamentous fungi species *U. reesii* (FF) was placed into the basal promoter construct (pLG669z). The same construct was mutated at the position of the putative Matα1 motif (FFΔ). *S. cerevisiae* Matα1 was supplied by the endogenous copy within a MATα strain (columns 1 and 2), and *C. albicans* Matα1 was supplied from expression of p415TEF within an *S. cerevisiae* MATa strain (columns 3 and 4). Reporter activity was monitored using β-galactosidase assays. For each sample, $n = 5$ and error bars represent SE.

out the yeasts because of their conserved synteny within the MAT locus and their conserved protein sequence (Fig. S4). Orthology mapping of Matα1 (*Materials and Methods*) across 38 genome-sequenced yeasts detected only a single unique Matα1 ortholog in all species in which the MAT locus has been sequenced. In contrast to examples of specificity changes between paralogs, Matα1 DNA-binding specificity divergence is not limited to a single phylogenetic branch point. Instead, Matα1 DNA-binding specificity appears to have diverged at several different points, indicating that DNA-binding specificity divergence between orthologous regulators can be a continuous process.

Despite this change in DNA-binding specificity, the Matα1 transcription regulator retains the same core function in both *S. cerevisiae* and *C. albicans*—activation of the αsgs. The conservation of function despite changes in DNA-binding specificity has been previously reported for other transcription regulators [e.g., Rpn4 (5), Yap1 (6)]. In these cases, however, the changes in DNA-binding specificity were subtle and likely resulted from limited coevolution of protein and DNA. We propose that the divergence of Matα1 DNA-binding specificity also represents a case of coevolution with its recognition sequence. If so, the overall change likely occurred in a stepwise fashion, perhaps the end result of numerous independent changes similar in magnitude to the DNA-binding specificity divergence between the *C. albicans* and *C. lusitaniae* Matα1. Consistent with this idea, the HMG DNA-binding domain of the *S. cerevisiae* and *C. albicans* Matα1 has undergone substantial divergence (Fig. S4).

We note that most fungi have approximately five αsgs; although this is not a large regulon, its conserved size indicates that the evolution of the Matα1-DNA interaction occurred across a set of target genes rather than across a single gene. In addition, the interaction of Matα1 with its cofactor Mcm1 also appears to be conserved between *S. cerevisiae* and *C. albicans*. This conserved protein-protein interaction could have facilitated the evolution of Matα1 by helping to "hold it in place" while its protein-DNA interaction slowly changed.

**Missing Examples of DNA-Binding Specificity Divergence.** How widespread are major evolutionary changes in DNA-binding specificity by transcription regulators? There are surprisingly few documented examples of extensive DNA-binding specificity divergence between orthologs or paralogs, a fact that has been used to argue that DNA-binding specificity evolution is uncommon in transcriptional networks. However, there is an unintended experimental bias against detecting instances of transcription regulator divergence (19). There are many reasons why a regulator from one species might not function in another species; hence, these observations are rarely pursued and often left unpublished. As a result, examples of functional conserva-

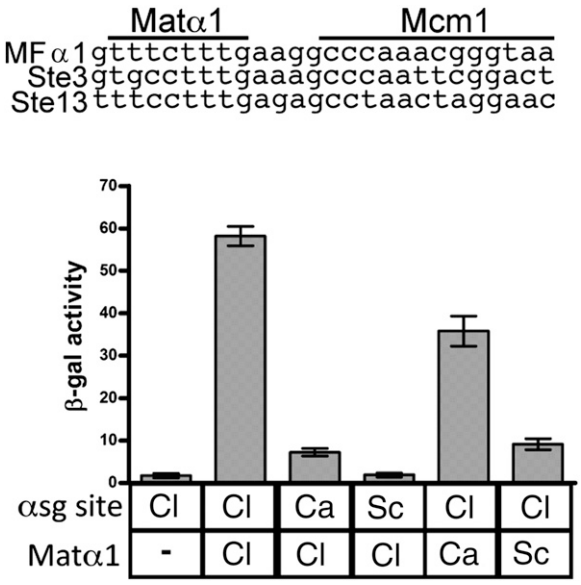


**Fig. 5.** Matα1 DNA-binding specificity has continued to diverge within the CTG-clade. Three putative αsg *cis*-regulatory sequences were identified by MEME in the promoters of *C. lusitaniae* αsg orthologs. The αsg *cis*-regulatory sequence of the promoter for the α-mating pheromone (*MFα1*) from *C. lusitaniae* (Cl) was inserted into a basal promoter construct (pLG669z), and the *C. lusitaniae* Matα1 was expressed from a 415-TEF plasmid. Plasmids were transformed into an *S. cerevisiae* MATα Δ*matα1* strain. Reporter activity was monitored using β-galactosidase assays. For each sample, $n = 5$ and error bars represent SE.

tion between orthologous transcription regulators may be overrepresented in the literature (20–22). For these reasons, we suggest that evolutionary changes in the DNA-binding specificity of transcriptional regulators, as documented here, may be more common than previously assumed.

The example of Matα1 DNA-binding specificity evolution has implications for bioinformatic approaches to transcriptional circuit evolution. If the only data available were the divergent *cis*-regulatory motifs, it would not be possible to distinguish between the three models described in the introduction (transcription regulator substitution, evolution of DNA-binding specificity, and relaxed DNA-binding specificity) and the observation could easily be misinterpreted. Furthermore, Matα1 DNA-binding specificity evolution demonstrates that orthologous transcription regulators can bind *cis*-regulatory sequences that appear unrelated by computational methods. This finding underscores a significant limitation of bioinformatic approaches to studying transcriptional networks that assume limited transcriptional regulator DNA-binding specificity divergence between species (23–25).

**Evolution of the Mating-Type Regulatory Circuitry and Speciation.** The evolution of Matα1 DNA-binding specificity is consistent with a network drift model of transcriptional network evolution (26). In other words, the coevolution of *cis*-regulatory sequences and transcription regulator DNA-binding specificity may have provided no specific adaptive advantage. However, it has been noted that compensatory mutations in developmental pathways could drive speciation events through the creation of Dobhanskzy–Mueller incompatibilities (27). Efficient mating in both *S. cerevisiae* and *C. albicans* requires the expression of the αsgs (7, 10), and a disruption in the Matα1-DNA interaction would produce a sterile phenotype. Therefore, a mating event between an individual that had experienced Matα1/*cis*-regulatory motif compensatory evolution and an individual that had not would produce a high fraction of infertile progeny. Thus, in the absence of spatial isolation of species, coevolution of the mating regulator Matα1 and its DNA-binding sites may have contributed to speciation.

## Materials and Methods

**PSSMs and Motif Alignments.** The PSSM for the *C. albicans*, *K. lactis*, and *S. cerevisiae* clade αsg *cis*-regulatory sequences was derived by performing multiple em for motif elicitation (MEME) (28) on 12, 15, and 27 sequences, respectively (sequence sets are provided in Table S1). The PSSM for the filamentous fungi αsg *cis*-regulatory sequences was derived by performing MEME from 9 sequences identified in the promoters of αsg orthologs in the filamentous fungi species *U. reesii, Cociddes immitis, Fosterella graminea, A. terreus, Aspergillus nidulans,* and *Sclerotinia sclerotiorum* (15, 29). Promoter sequences from closely related species were pooled to increase the number of sequences submitted to MEME, thereby yielding more accurate PSSMs (under the assumption that species so closely related would not experience drastic changes in DNA-binding specificity between orthologous regulators). No close relatives of *Y. lipolytica* have been genome-sequenced (30); therefore, our set of αsg orthologs for this branch was quite small (four orthologous genes). Hence, the PSSM built from 6 putative αsg *cis*-regulatory sequences identified in *Y. lipolytica* is not as information-rich as the other PSSMs presented in this work (Fig. S5). Motif alignments were computed using the motif comparison utility in MochiView (31). MochiView relies on an algorithm derived from Gupta et al. (32) to perform motif alignments. The algorithm maximizes the similarity score between two motifs and then derives an *E* value from this similarity score by screening a PSSM library to determine how often this similarity score would occur by chance. The PSSM libraries that are compiled in MochiView to increase the accuracy of *E* values for motif alignments are JASPAR (33), SwissRegulon (34), Gasch/Eisen (5), Badis/Hughes (35), MotifVoter (36), MacIsaac (37), and Zhu (38).

**Cloning.** Primers used in this study are included in Table S2. Because of several CUG codons in the HMG DNA-binding domain of *C. albicans MATα1*, we had the gene codon-optimized by DNA 2.0 for expression in *S. cerevisiae.* Each species' *MATα1* was cloned into the 415-translation elongation factor promoter (TEF) CEN/ARS plasmid and sequenced to check for mutations (39). The level of ectopic expression from these plasmids was insufficient to detect a gel shift. Therefore, each *MATα1* was cloned into the inducible, high-expression, 415-GAL 2μ plasmid (40). To study αsg *cis*-regulatory sequences, 42-bp regions centered around the putative αsg *cis*-regulatory sequences for α-mating pheromone gene (except for the filamentous fungi sequence; because of the absence of a clear α-mating pheromone gene ortholog, a sequence from the promoter of mating **a**-factor receptor gene was used instead) were cloned into the UAS-less Cyc1 reporter construct pLG699Z (41) using *Xho*1. Correct orientation relative to the transcriptional start site for the αsg *cis*-regulatory sequences within our pLG669z-derivatives was confirmed by PCR and sequencing.

**Strain Construction.** *S. cerevisiae* strains used and generated in this study are presented in Table S3. β-galactosidase experiments were either performed in *S. cerevisiae* W303 MAT**a** cells or *S. cerevisiae* EG123 MATα Δ*matα1* strains (42). Gel shift experiments were performed using cell extracts from strains built in the *S. cerevisiae* W303 background.

**β-Galactosidase Assays.** β-galactosidase assays were performed using a standard protocol (41). Strains were grown in SD-Ura-Lue media to maintain selection for both plasmids. For each strain, five colonies were grown overnight, diluted back, and allowed to reach log phase. Cells were harvested and permeabilized, and activation assays were performed. The data provided throughout any figure are from the same day.

**Electrophoretic Mobility Shift Assays.** Yeast strains were grown overnight in either glucose or galactose medium (in both media types, selection was maintained for the plasmid marker), depending on whether ectopic expression of Matα1 was desired. Harvested cells were of an $OD_{600}$ between 0.75 and 1.0. *S. cerevisiae* pellets were resuspended in 100 mM Tris (pH 8), 200 mM NaCl, 1 mM EDTA, 10 mM $MgCl_2$, 10 mM β-mercaptoethanol, 20% (vol/vol) glycerol, and Roche Complete protease inhibitors (one tablet per 10 mL). Extracts were lysed by sonification and then cleared by centrifugation at 12,000 × *g* for 20 min, yielding ~10 mg/mL total protein. Electrophoretic mobility gel shift assays were performed using *S. cerevisiae* cell extracts as described by Keleher et al. (43). The αsg *cis*-regulatory sequence oligonucleotide probes were labeled with $^{32}P$ γ-ATP using T4 polynucleotide kinase. Binding conditions were 50 mM Tris (pH 8), 100 mM NaCl, 10% (vol/vol) glycerol, 5 mM MgCl2, 5 mM β-mercaptoethanol, 50 μg/mL Poly(dI-dC)

(limits nonspecific protein/DNA binding), and 1.2 μM labeled oligonucleotide. Antibody supershifts were accomplished using a Matα1 N-terminal peptide antibody (antigenic sequence MGNKKKTRKTVPKEFISLC; Bethyl Antibodies). For a 20-μL protein/DNA-binding reaction, 0.5 μL of a 1:100 dilution of immune serum was sufficient to induce supershifts.

**Orthology Mapping.** Orthology mapping was performed as described by Tsong et al. (44). *S. cerevisiae* and *C. albicans* αsg protein sequences were used to "query" a single database containing all ORF sequences from 38 fungal species using PSI-BLAST (45), utilizing an *E* value cutoff of $10^{-5}$ and the Smith–Waterman alignment option. The sequences returned by PSI-BLAST were then multiply aligned with multiple sequence comparison by log (MUSCLE), and a neighbor joining (NJ) tree was inferred, again using ClustalW (46). Finally, the resulting NJ tree was traversed to extract a set of orthologous genes.

**Genome Sequencing.** To improve our ability to detect *cis*-regulatory sequences in *K. lactis* using phylogenetic footprinting (47), the genomes of the two close relatives of the *K. lactis* [*K. aestuarii* (American Type Culture Collection 18862) and *K. wickerhamii* (UCD 54-210)] were sequenced. *K. aestuarii* was sequenced to an estimated coverage of 14× coverage and *K. wickerhamii* was sequenced to an estimated coverage of 12× coverage on a 454 platform at the Washington University Genome Sequencing Center. The Washington University Genome Sequencing Center used the assembly algorithm Newbler in early 2008 to assemble the 454 reads into contigs. This level of sequencing was insufficient to assemble complete chromosomes but was sufficient to extract information about αsg orthologs in these species. For *K. wickerhamii,* after assembly, the number of long contigs (>500 bp) was 510 and the number of short contigs (>100 bp) was 953. For *K. aestuarii*, the number of long contigs (>500 bp) was 336 and the number of short contigs (>100 bp) was 682. The sequence will be available through the Johnson laboratory Web site, along with ORF calls, and is currently available through GenBank as a whole-genome shotgun sequencing project data [GenBank accession nos. AEAS00000000 (*K. aestuarii*) and AEAV00000000 (*K. wickerhamii*)].

**ACKNOWLEDGMENTS.** The authors thank Oliver Homann and Xin He for sharing their knowledge of bioinformatics. The authors also thank Oliver Homann and Linet Mera for valuable comments on the manuscript. The work of the authors was supported by National Institutes of Health Grant GMO37049.

1. Wray GA, et al. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol* 20:1377–1419.
2. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci USA* 104(Suppl 1):8605–8612.
3. McGinnis N, Kuziora MA, McGinnis W (1990) Human Hox-4.2 and Drosophila deformed encode similar regulatory specificities in Drosophila embryos and larvae. *Cell* 63:969–976.
4. Halder G, Callaerts P, Gehring WJ (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science* 267:1788–1792.
5. Gasch AP, et al. (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol* 2:e398.
6. Kuo D, et al. (2010) Coevolution within a transcriptional network by compensatory *trans* and *cis* mutations. *Genome Res* 20:1672–1678.
7. Bender A, Sprague GF, Jr. (1987) MAT alpha 1 protein, a yeast transcription activator, binds synergistically with a second protein to a set of cell-type-specific genes. *Cell* 50: 681–691.
8. Jarvis EE, Clark KL, Sprague GF, Jr. (1989) The yeast transcription activator PRTF, a homolog of the mammalian serum response factor, is encoded by the MCM1 gene. *Genes Dev* 3:936–945.
9. Taylor JW, Berbee ML (2006) Dating divergences in the Fungal Tree of Life: Review and new analyses. *Mycologia* 98:838–849.
10. Tsong AE, Miller MG, Raisner RM, Johnson AD (2003) Evolution of a combinatorial transcriptional circuit: A case study in yeasts. *Cell* 115:389–399.
11. Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 6:e38.
12. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241–254.
13. Butler G, et al. (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657–662.
14. Booth LN, Tuch BB, Johnson AD (2010) Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* 468:959–963.
15. Sharpton TJ, et al. (2009) Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res* 19:1722–1731.
16. Wharton RP, Ptashne M (1985) Changing the binding specificity of a repressor by redesigning an alpha-helix. *Nature* 316:601–605.
17. Knight KL, Sauer RT (1989) DNA binding specificity of the Arc and Mnt repressors is determined by a short region of N-terminal residues. *Proc Natl Acad Sci USA* 86: 797–801.
18. Emerson RO, Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. *PLoS Genet* 5:e1000325.
19. Lynch VJ, Wagner GP (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62:2131–2154.
20. Ranganayakulu G, Elliott DA, Harvey RP, Olson EN (1998) Divergent roles for NK-2 class homeobox genes in cardiogenesis in flies and mice. *Development* 125:3037–3048.
21. Park M, et al. (1998) Differential rescue of visceral and cardiac defects in Drosophila by vertebrate tinman-related genes. *Proc Natl Acad Sci USA* 95:9366–9371.
22. Maizel A, et al. (2005) The floral regulator LEAFY evolves by substitutions in the DNA binding domain. *Science* 308:260–263.
23. Xie X, et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* 434:338–345.
24. Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLOS Comput Biol* 3:e99.
25. Li XY, et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol* 6:e27.
26. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104(Suppl 1):8597–8604.
27. Porter AH, Johnson NA (2002) Speciation despite gene flow when developmental pathways evolve. *Evolution* 56:2103–2111.
28. Bailey TA, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* (AAAI Press, Menlo Park, CA), pp 28–36.
29. Dietrich FS, et al. (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* 304:304–307.
30. Dujon B, et al. (2004) Genome evolution in yeasts. *Nature* 430:35–44.
31. Homann OR, Johnson AD (2010) MochiView: Versatile software for genome browsing and DNA motif analysis. *BMC Biol* 8:49.
32. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8:R24.
33. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: An open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32(Database issue):D91–D94.
34. Pachkov M, et al. (2006) SwissRegulon: A database of genome-wide annotations of regulatory sites. *Nucleic Acids Res* 27(Database issue):D1–D5.
35. Badis G, et al. (2009) Diversity and complexity in DNA recognition by transcription factors. *Science* 324:1720–1723.
36. Wijaya E, et al. (2008) MotifVoter: A novel ensemble method for fine-grained integration of generic motif finder. *Nucleic Acids Res* 24:2288–2295.
37. MacIsaac KD, et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7:113.
38. Zhu C, et al. (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* 19:556–566.
39. Mumberg D, Müller R, Funk M (1995) Yeast vectors for the controlled expression of heterologous proteins in different genetic backgrounds. *Gene* 156:119–122.
40. Mumberg D, Müller R, Funk M (1994) Regulatable promoters of *Saccharomyces cerevisiae*: Comparison of transcriptional activity and their use for heterologous expression. *Nucleic Acids Res* 22:5767–5768.
41. Guarente L, Ptashne M (1981) Fusion of Escherichia coli lacZ to the cytochrome c gene of Saccharomyces cerevisiae. *Proc Natl Acad Sci USA* 78:2199–2203.
42. Galgoczy DJ, et al. (2004) Genomic dissection of the cell-type-specification circuit in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci USA* 101:18069–18074.
43. Keleher CA, Passmore S, Johnson AD (1989) Yeast repressor alpha 2 binds to its operator cooperatively with yeast protein Mcm1. *Mol Cell Biol* 9:5228–5230.
44. Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443:415–420.
45. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
46. Higgins DG, Sharp PM (1988) CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* 73:237–244.
47. Cliften P, et al. (2003) Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science* 301:71–76.
48. Scannell DR, et al. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. *Proc Natl Acad Sci USA* 104:8397–8402.