Molecular Systems Biology 8; Article number 619; doi:10.1038/msb.2012.50
Citation: ***Molecular Systems Biology*** **8:619**
© 2012 EMBO and Macmillan Publishers Limited All rights reserved 1744-4292/12
www.molecularsystemsbiology.com

molecular systems biology

A functional selection model explains evolutionary robustness despite plasticity in regulatory networks

Naomi Habib[1,2,3,7], Ilan Wapinski[4,5,7], Hanah Margalit[2,*], Aviv Regev[5,6,*] and Nir Friedman[1,3,*]

[1] School of Computer Science and Engineering, Hebrew University, Jerusalem, Israel, [2] Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University, Jerusalem, Israel, [3] Alexander Silberman Institute of Life Sciences, Hebrew University, Jerusalem, Israel, [4] Department of Systems Biology, Harvard Medical School, Boston, MA, USA, [5] Broad Institute, 7 Cambridge Center, Cambridge, MA, USA and [6] Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA
[7]These authors contributed equally to this work
* Corresponding authors. H Margalit, Department of Microbiology and Molecular Genetics, IMRIC, Faculty of Medicine, Hebrew University, Jerusalem 91120, Israel. Tel.: +972 2 675 8614; Fax: +972 2 675 7529; E-mail: hanahm@ekmd.huji.ac.il. or A Regev, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA. Tel.: +1 617 714 7021; Fax: +1 617 714 8956; E-mail: aregev@broad.mit.edu or N Friedman, School of Computer Science and Engineering, Hebrew University, Givat Ram Campus, Jerusalem 91904, Israel. Tel.: +972 2 549 4557; Fax: +972 2 658 5261; E-mail: nir@cs.huji.ac.il

Received 20.7.12; accepted 29.8.12

Evolutionary rewiring of regulatory networks is an important source of diversity among species. Previous evidence suggested substantial divergence of regulatory networks across species. However, systematically assessing the extent of this plasticity and its functional implications has been challenging due to limited experimental data and the noisy nature of computational predictions. Here, we introduce a novel approach to study *cis*-regulatory evolution, and use it to trace the regulatory history of 88 DNA motifs of transcription factors across 23 *Ascomycota* fungi. While motifs are conserved, we find a pervasive gain and loss in the regulation of their target genes. Despite this turnover, the biological processes associated with a motif are generally conserved. We explain these trends using a model with a strong selection to conserve the overall function of a transcription factor, and a much weaker selection over the specific genes it targets. The model also accounts for the turnover of bound targets measured experimentally across species in yeasts and mammals. Thus, selective pressures on regulatory networks mostly tolerate local rewiring, and may allow for subtle fine-tuning of gene regulation during evolution.

Molecular Systems Biology 8:619; published online 23 October 2012; doi:10.1038/msb.2012.50
Subject Categories: metabolic and regulatory networks; computational methods
Keywords: bioinformatics; computational methods; evolution; regulation; transcription

Introduction

Changes in gene regulation have been postulated to play a key role in generating the vast phenotypic diversity observed across species (King and Wilson, 1975; Wittkopp *et al*, 2004; Prud'homme *et al*, 2007). Evolutionary rewiring of regulatory networks is often driven by the gain or loss of *cis*-regulatory elements in gene promoters bound by sequence-specific transcription factors or chromatin remodelers (Khaitovich *et al*, 2006; Wilson *et al*, 2008; Wittkopp *et al*, 2008; Tirosh *et al*, 2009), as observed for specific factors in yeasts (Borneman *et al*, 2007; Doniger and Fay, 2007; Tuch *et al*, 2008), flies (Moses *et al*, 2006; Bradley *et al*, 2010), and mammals (Odom *et al*, 2007; Schmidt *et al*, 2010). Changes in *cis*-regulatory elements in genes' promoters are associated with fine-grained regulatory 'tinkering' of individual genes (Borneman *et al*, 2007; Lavoie *et al*, 2010). For example, the individual target genes of the yeast regulatory factor Mcm1 have diverged significantly between three related species (Tuch *et al*, 2008), while maintaining its regulation of the cell cycle and mating in all three species. On the other hand, there are cases where changes in *cis*-regulatory elements lead to dramatic rewiring of the regulation of entire gene modules (Hogues *et al*, 2008; Tuch *et al*, 2008). For example, the transcription of ribosomal protein encoding genes in yeasts is regulated by distinct transcription factors in *Candida albicans* (Tbf1 and Cbf1) and *Saccharomyces cerevisiae* (Rap1), primarily through changes in *cis*-regulatory elements in promoter regions (Tanay *et al*, 2005; Hogues *et al*, 2008). Although such individual examples are instructive, they represent only anecdotal evidence of the role that *cis*-regulatory divergence plays across evolution. It is thus of great interest to quantitatively and qualitatively assess the extent of *cis*-regulatory plasticity of different regulatory DNA motifs and their associated target genes and its functional implications.

For example, consider the yeast transcription factor Gcn4. According to the occurrences of its predicted binding sites in gene promoters (motif targets), we estimate that Gcn4 potentially regulates an average of 470 target genes in each of the 20 *Ascomycota* fungi in which we can identify it (see below; Materials and methods). In each of these species, the set of predicted Gcn4 target genes is enriched for genes encoding enzymes involved in amino-acid metabolism, consistent with Gcn4's known role in the model organisms *S. cerevisiae* and *C. albicans* (Martchenko *et al*, 2007). Naively, this observation suggests that Gcn4 would mostly target orthologous genes across the different species. To our surprise, however, the orthologs of only 8% of all the Gcn4 targets in *S. cerevisiae* are also targeted by Gcn4 in all five closest relatives of *S. cerevisiae* in this study. Furthermore, for only 8 of the 307 genes targeted by Gcn4 in *S. cerevisiae* are the orthologs across 16 or more species also targeted by Gcn4.

Such an analysis is interesting but suffers from several potential caveats. First, computational predictions of regulatory targets based on occurrences of DNA motifs in gene promoters are noisy (MacIsaac *et al*, 2006; Hannenhalli, 2008; Zhu *et al*, 2009) and have only limited reliability, as compared with experimentally measured binding (Capaldi *et al*, 2008). However, directly measuring the binding of a host of transcription factors across dozens of species and in all relevant conditions is still very challenging. Second, there is a lack of known regulatory DNA motifs in non-model organisms, and determining targets in one species with motifs from another (possibly distant) species can lead to both false positive and negative predictions.

To address these challenges, we developed CladeoScope (Figure 1; Materials and methods)—an unbiased phylogenetic approach to reconstruct *cis*-regulatory evolution. CladeoScope reconstructs regulatory networks, associating DNA motifs with putative target genes across species in a phylogeny. It starts with experimentally determined DNA motifs from a model organism, and computationally adapts them to each species. To control for noisy predictions in individual species, CladeoScope predicts reliable targets in ancestral genomes based on phylogenetic support from multiple extant species. We used CladeoScope to trace the evolutionary history of regulatory interactions of 88 regulatory DNA motifs associated with transcription factors (or groups of paralogous factors) across 23 *Ascomycota* fungi, spanning >300 million years of evolution, showing that most have conserved their cognate motifs over large evolutionary distances. While most motifs show widespread gain and loss of individual target genes, the biological processes associated with them are typically highly conserved. We reconcile these trends by a model that assumes a strong selection to conserve the overall function of a motif, and a much weaker selection for its specific target genes. The model is consistent with the number of highly conserved target genes and with observed turnover of bound target genes as determined by protein–DNA binding experiments across species, thus revealing a unifying principle of *cis*-regulatory evolution.

Results

CladeoScope: a framework for reconstructing *cis*-regulatory evolution

We developed CladeoScope (Figure 1), a computational framework for reconstructing *cis*-regulatory networks and their evolution across a phylogeny of species. CladeoScope relies on two assumptions. First, we assume that the binding specificities of transcription factors, represented as DNA motifs, are largely conserved, even when their specific target genes and functional roles may have substantially diverged (Wapinski *et al*, 2007; Tuch *et al*, 2008; Schmidt *et al*, 2010). We therefore initiate our reconstruction with DNA motifs of known transcription factors that have been experimentally determined, but without any further assumptions about conservation of their individual targets or their global functional roles. We do allow for relatively small changes in binding affinities across evolution, and thus refine those motifs in a species-specific manner (see below). Second, although predicting the target genes for a motif (motif targets) across the genome is prone to errors (Hannenhalli, 2008), we assume that targets that are conserved across several related species within a monophyletic clade provide a reliable and conservative estimate for the targets in the ancestor of the clade. Thus, for each motif associated with a known transcription factor in *S. cerevisiae* (e.g., Gcn4), CladeoScope finds its ancestral target genes in various ancestors in the phylogeny. A gene is considered to be targeted by a motif in the ancestor of a clade of species only if evolutionary analysis of the orthologous targets across the species in the clade indicated that the ancestral gene (Wapinski *et al*, 2007) of that clade was a target of the motif (see Materials and methods). CladeoScope then compares between the ancestral targets of different clades, allowing us to reliably track evolutionary changes across the phylum by considering the evolutionary changes between clades while filtering out spurious targets within a clade.

CladeoScope consists of four steps (Figure 1B): In step 1—*Initialization*—CladeoScope is initialized with known DNA motifs (position weight matrices (PWM)) from one or more model organisms in the phylogeny. It uses these initial motifs to find a set of provisional target genes for each initial motif in each species, according to the motif's occurrences in a gene's promoter. We do not require these provisional target sets to be evolutionarily conserved. In step 2—*Species-Specific Motifs*—CladeoScope takes each initial motif and its provisional target sets, and learns species-specific motifs and targets in an iterative manner. In step 3—*Network Refinement*—CladeoScope uses a parsimony-based algorithm to reconstruct the set of each motif's ancestral targets for the last common ancestor (LCA) of each clade in the phylogeny (Figure 2). These inferred ancestral targets within a clade are considered reliable (Figure 2; Supplementary Figure 1). In step 4—*Filtration*—CladeoScope filters motifs and target genes based on their phylogenetic conservation. In particular, we define a motif as detectable in an ancestor and in each of its descendant extant species if the number of the targets in the ancestor and in each extant species is statistically significant (see details below). The algorithm iterates between steps 3 and 4 until it converges. CladeoScope's output includes for each motif, its weight matrix in each species, the ancestors and extant species in which it is detectable, and the targets in each ancestor.

Parsimonious phylogenetic filtering of motifs and targets

To infer the ancestral motif targets in step 3, CladeoScope traces motif-target relations across orthologous loci. This is done separately for each ancestral gene at each ancestral position in the tree (Figure 2). To determine if an ancestral

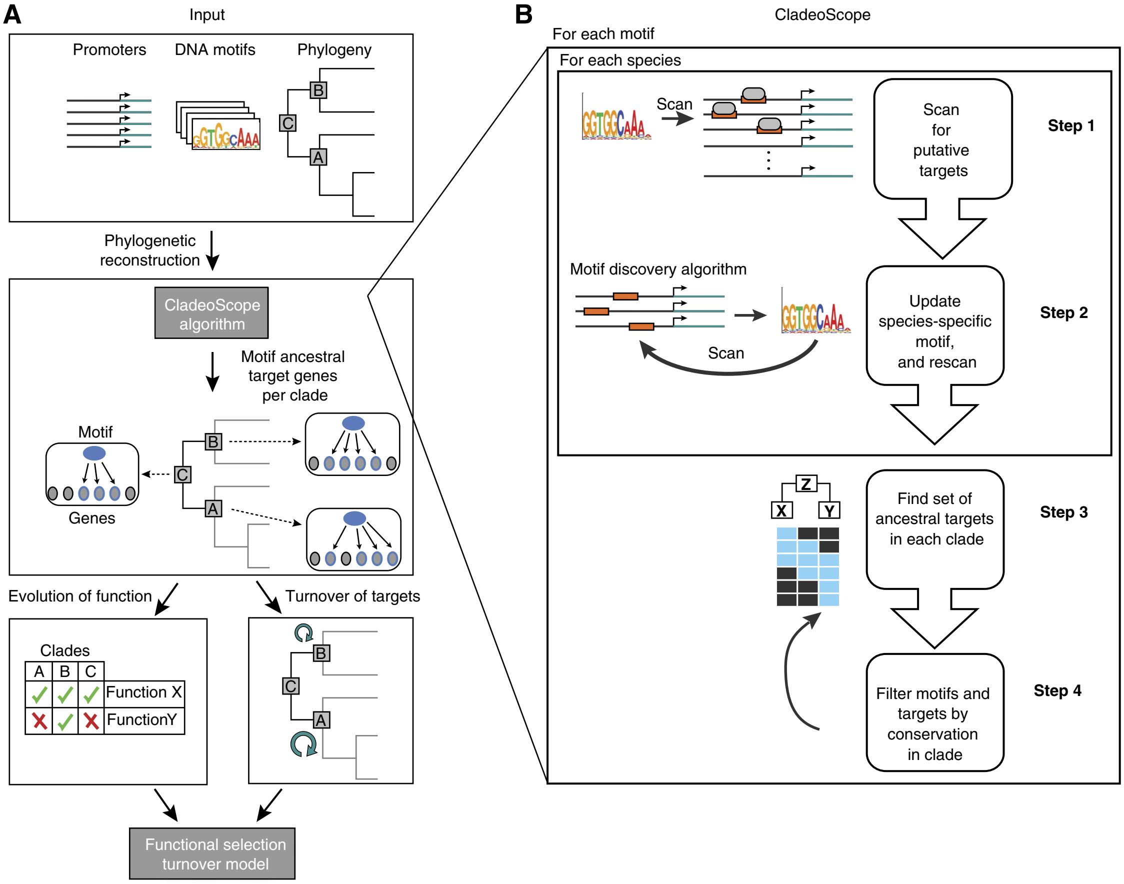
A
Input
Promoters
DNA motifs
Phylogeny
B
C
A
Phylogenetic reconstruction
CladeoScope algorithm
Motif ancestral target genes per clade
Motif
Genes
Evolution of function
Turnover of targets
Clades
A B C
Function X
FunctionY
Functional selection turnover model
B
CladeoScope
For each motif
For each species
Scan
Scan for putative targets
Step 1
Motif discovery algorithm
Scan
Update species-specific motif, and rescan
Step 2
Z
X
Y
Find set of ancestral targets in each clade
Step 3
Filter motifs and targets by conservation in clade
Step 4

Figure 1 The CladeoScope method. (**A**) Analysis overview. We use the CladeoScope algorithm that takes as input DNA motifs, promoter sequences, and the species phylogeny (top box) to reconstruct regulatory networks (i.e., ancestral motif-target genes) at each ancestral position (clade) of the phylogeny (middle box). A gene is considered as a putative target of the motif if its promoter contains an occurrence of the DNA motif (a binding site), and the ancestral motif targets are inferred by phylogenetic reconstruction. We use these networks to study both the turnover (gain and loss) of target genes associated with the motif across the phylogeny (bottom right box), and the evolution of functions associated with the motif (bottom left box) in each ancestral position of the phylogeny, where the function is determined by the functional annotation of its motif targets. We then built an evolutionary model that explains both trends. (**B**) The CladeoScope method. Shown is a flowchart of the input to CladeoScope (top) and its three consecutive steps: step 1—Initialization using known DNA motifs from a model organism and promoter sequences of other species. For each motif, we find putative sets of motif-containing target genes in the other genomes; step 2—Learning species-specific motifs and targets; and step 3—Network refinement, definition of detectable motifs and sets of ancestral targets per clade. step 4—Filtration of motif and target genes based on their phylogenetic conservation.

gene is a motif target, CladeoScope uses a parsimonious phylogenetic reconstruction approach to minimize the number of target gain and loss events (Fitch, 1971). This reconstruction explicitly considers each gene paralog derived from the same ancestor by duplication, and distinguishes a lost gene from a present gene that is not a target (see Materials and methods).

Phylogenetic filtering addresses both noisy predictions of target genes as well as DNA motifs that are 'non-functional' in a species or a clade (i.e., no longer act as a functional regulatory element bound by a cognate transcription factor). CladeoScope tests each motif in each species independently, based on the overlap between the motif's putative target genes in that species and the motif's ancestral targets in any relevant ancestor. Only motifs where the overlap is statistically significant (hypergeometric *P*-value <0.001, see Materials and methods) are termed 'detectable' in the species. Since filtering the motifs and the reconstruction of ancestral targets are dependent, our algorithm iterates between both steps. If any insignificant motifs are found in the clade (step 4), the most insignificant one is removed, and CladeoScope returns to step 3. After convergence, CladeoScope filters the motifs at the clade level, requiring that the number of inferred targets for a motif in the clade's ancestor is statistically significant (empirical *P*-value computed by 1000 reconstructions of ancestral targets for random sets of motif targets of the same size for each species, see Materials and methods).

CladeoScope is robust to noise in target prediction and to different parameters

Using simulated data we confirmed that CladeoScope is highly robust to noise in target prediction for individual species and to other input variations. To assess robustness, we used hundreds of simulated evolved motif-target sets, where each simulation varied the extent and type of noise in target prediction, the size of the ancestral target set, the degree of target turnover and the topology of the species tree (960 different combinations of parameters, see Materials and methods, Supplementary Note 1). For example, when 30%

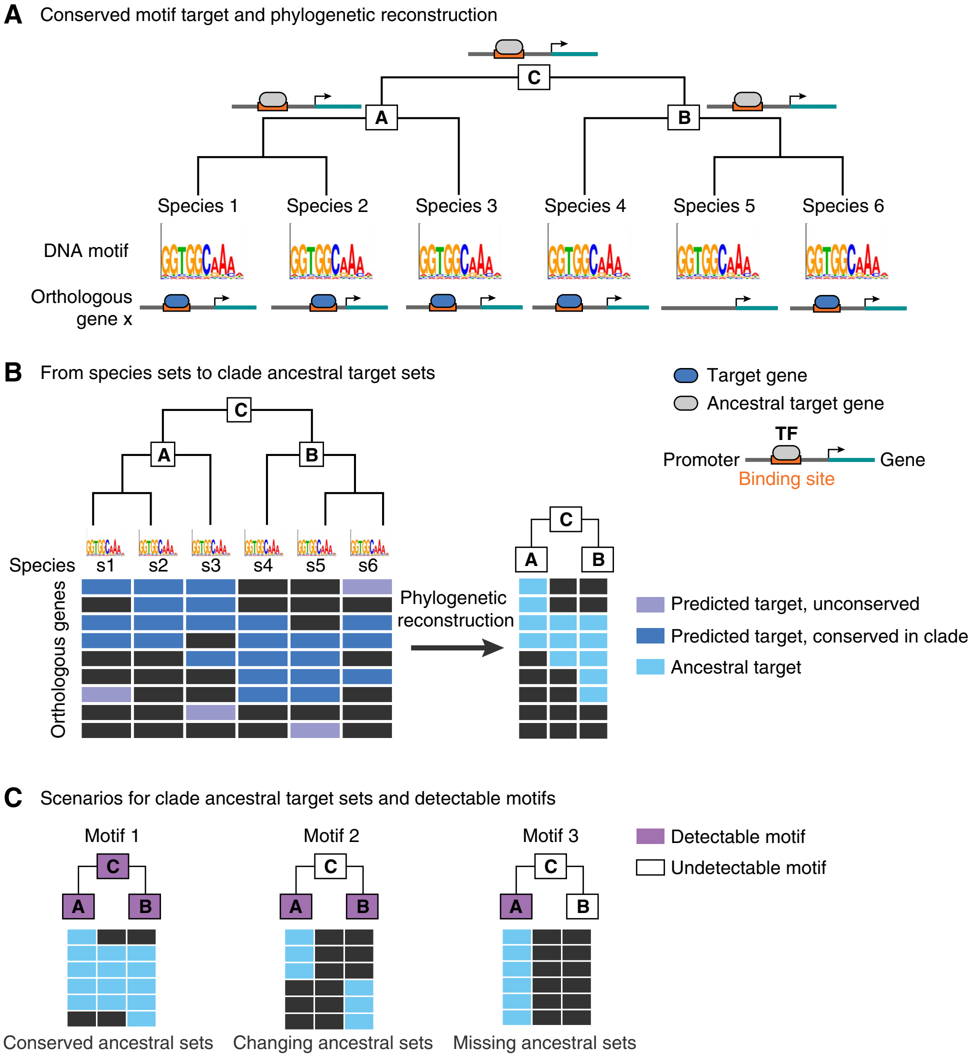

Figure 2 Principles of phylogenetic reconstruction of regulatory history. (**A**) Phylogenetic reconstruction of motif-target genes. Given a set of DNA motifs, representing the binding specificities of a transcription factor (blue oval, bottom) in different species (sequence logos), and their motif-target genes, we reconstruct the parsimonious ancestral regulatory state in each internal node (gene cartoons at internal nodes, A, B, C, squares). In this cartoon example, the gene has orthologs in species 1–6, but there is no binding site associated with the motif in species 5, and we reconstruct an ancestral target in species A, B, and C. (**B**) Deriving sets of ancestral targets per clade from target genes in each species. Given all motif-target genes (rows in left matrix; blue and purple—predicted target conserved or not-conserved respectively; black—not a target) for each motif (sequence logo, columns) in each species, we reconstruct all the ancestral targets for each gene as in (A). The resulting set of ancestral targets for each clade (right matrix, ancestral species A, B, C in columns, ancestral targets genes in rows, light blue—ancestral target gene; black—not an ancestral target). (**C**) Illustrative examples of sets of ancestral targets and detectable motifs. Shown are several possible evolutionary scenarios. In all cases, clades (A, B, C) in columns; purple—detectable motif in clade; white—not detectable; target genes in rows; light blue—ancestral target gene; black—not an ancestral target. In 'conserved ancestral sets' (left), a motif has statistically significant sets of ancestral targets (i.e., is detectable) in all three clades, and the targets are highly conserved. In 'changing ancestral sets' (center), a motif has statistically significant sets of ancestral targets in clades A and B, but these are not conserved between the two clades, and are hence missing in the ancestral clade C. In 'missing ancestral sets' (right), a motif has a significant set of ancestral targets (i.e., is detectable) only in clade A, and not in the other clades.

of the true targets were removed from the set of target genes provided to CladeoScope, CladeoScope has >85% sensitivity (percent predicted targets among true targets), and when 80% false targets were added in each species, CladeoScope has >80% specificity (percent true targets among predicted targets) (Supplementary Figure 1; Supplementary Note 1).

CladeoScope's predictions are also highly robust to variation in its various parameters (Supplementary Note 1). For example, varying the threshold for the significance of a motif in a species between 10^{-5} to 5×10^{-2} had little or no effect on the number of ancestral targets reconstructed per clade. Similarly, varying the threshold for conservation of a motif in a clade between 0.05 and 0.001 had little impact on the number of significant motifs per clade. Thus, evolutionary conservation within a clade—rather than parameter fine-tuning—is the main determinant of CladeoScope's results and performance.

To examine the possibility that our relatively strict motif-target detection threshold excludes weak, yet functional, binding sites, we compared the score distribution of functional

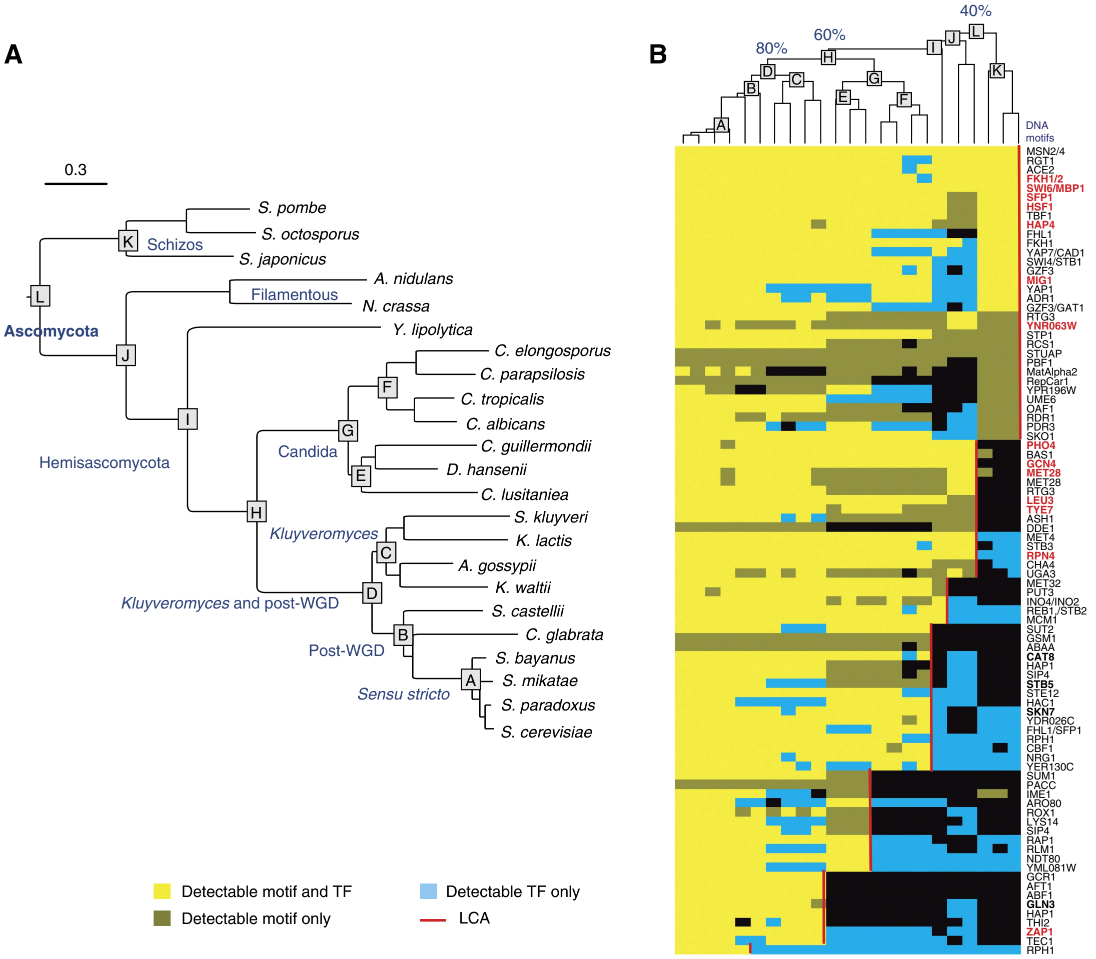

Figure 3 Motif detectability corresponds to the phylogenetic profile of the cognate transcription factor. (**A**) The phylogenetic tree for species in this study. Shown is the phylogenetic tree of the 23 *Ascomycota* species in this study (see Materials and methods). A–L: clades in which ancestral target sets are defined; clade names are denoted next to their letter in dark blue. (**B**) Motif detection and transcription factor presence across the species. Shown are 88 motifs (rows) across 23 species (columns) along with a phylogenetic tree (as in a, but not shown to scale). Yellow—motif is detectable and transcription factor is conserved; yellow–gray—the motif is detectable but the transcription factor is not conserved; blue—the motif is not detectable but the transcription factor is conserved, black—the motif is not detectable and the transcription factor is absent. The fraction of the motifs inferred to be detectable up to clades D, H, and L is marked on top of the respective clades. Red line denotes the most ancestral clade in the species tree where a motif is detectable. Motif names in red denote motifs that are further discussed in the text.

but weak binding sites to non-functional sites. We identify potential candidates for weak functional sites as ones with conserved target genes in the sister species within the same clade, which are classified as non-target ('lost') in the reference species. Indeed, in 85% of the cases we tested, such 'lost' targets have a distribution of scores similar to genes that are not targets throughout the clade. Hence, lowering the threshold would not have increased our sensitivity to such weak sites (Supplementary Note 2). Nonetheless, as an additional validation, we tested the main findings using a lower threshold for motif-targets detection and found our results to be robust (Supplementary Notes 1 and 3).

Finally, as a negative control, we provided CladeoScope an input set of randomly generated motifs. Although in each species we do find targets for such motifs, CladeoScope's phylogenetic filtering found that these motifs are not conserved (Supplementary Note 1). The only exception is in the closely related *sensu-stricto Saccharomyces*, where intergenic sequences have not yet had enough time to acquire sufficient mutations. We therefore do not report motifs found to be conserved only in this clade.

Systematic reconstruction of the regulatory history of 23 *Ascomycota* species

We applied CladeoScope to 88 DNA motifs associated with known transcription factors or groups of paralogous factors from *S. cerevisiae* (MacIsaac *et al*, 2006; Matys *et al*, 2006; Zhu *et al*, 2009) across 23 *Ascomycota* species, defining motif-target genes in 12 clades (A–L, Figure 3A; Supplementary website). As points for reconstruction of ancestral targets we chose clades with a large evolutionary distance between them and relatively small distances within each (Figure 3A). These clades include: the *sensu-stricto Saccharomyces* (four species, clade A), the *Kluyveromyces* (four species, clade C), the

Candida (seven species, clade G), the *Hemiascomycota* (*Kluyveromyces*, *Saccharomyces*, *Candida* clades, and *Yarrowia lipolytica*, 18 species, clade I), and the full *Ascomycota* clade (23 species, clade L). The resulting ancestral network contains 190 689 reliable motif-target connections (conserved in at least one clade), compared with 996 476 connections prior to phylogenetic filtering. For example, of the 307 predicted Gcn4 targets in *S. cerevisiae*, 195 pass our phylogenetic filter.

To assess CladeoScope's performance in this phylogeny, we compared its predicted targets to those measured *in vivo* by chromatin immunoprecipitation (ChIP) in *S. cerevisiae* (MacIsaac *et al*, 2006) and four other species (Borneman *et al*, 2007; Tuch *et al*, 2008) (Supplementary Note 1). In most cases, using CladeoScope's in-clade conservation increases the precision of the predicted motif targets. For instance, for the Cbf1 motif, CladeoScope reaches 80% precision rate and 50% sensitivity using the ancestral motif targets in clade A, compared with 55% and 10%, respectively, in the predicted motif targets in *S. cerevisiae* that are not conserved (Supplementary Note 1). These improved predictions are consistent for different thresholds for motif-targets detection in each species (Supplementary Note 1).

Regulatory motifs are detectable across large evolutionary distances

For most regulatory DNA motifs we could detect ancestral target genes within clades across the phylogeny (Figure 3B; Supplementary Table 1). This is consistent with our assumptions that transcription factors retain their binding specificities and that many of their target genes are conserved in closely related species. For example, ∼83% of the motifs were detectable in clade D (*Kluyveromyces* and post-whole genome duplication (post-WGD) clades) and ∼68% were detectable in clade H, including in species as remote from each other as *S. cerevisiae* and *C. albicans*. The latter include motifs involved in central metabolic and cellular processes (Figure 3B, red highlights), such as Gcn4 (amino-acid biosynthesis), Rpn4 (proteasome), and Mig1 (glucose repression). In all, 39% of motifs were detectable up to the LCA of the entire *Ascomycota* phylum (clade L), including those involved in cell-cycle regulation (Fkh1, Swi6-MBP1, Figure 3B) and stress response factors (Hsf1, STRE, Figure 3B, red highlights). The number of motifs detectable across the phylogeny is particularly remarkable given the substantial evolutionary distances, the large intra-species divergence within the *Schizosaccharomyces* (Rhind *et al*, 2011), and the fact that as many as 25% (102 of 392) of the transcription factors in *S. cerevisiae* do not have a clearly identifiable ortholog in *Schizosaccharomyces pombe* (Wapinski *et al*, 2007).

The phylogenetic profiles of transcription factors largely correspond to the detectability of their cognate motifs, supporting our reconstruction. In most cases (73%), detectable motifs and factors are co-conserved (Figure 3B): when a motif is detectable in a species, the ortholog of its known cognate factor is present in the same species, and vice versa. The few cases where there is discrepancy are due to either evolutionary innovations or limitations in ortholog mapping or motif detection (Supplementary Note 2).

Rapid target turnover for conserved motifs during evolution

To assess changes during the evolution of regulatory networks, we first calculated the amount of turnover events for each of the 88 regulatory motifs as the number of target genes gained or lost at each clade since its direct ancestral clade. Overall, there is an extensive and rapid turnover of motif-target genes. This high turnover of targets is apparent even for broadly conserved motifs with ancient ancestral targets, such as Gcn4 and Fkh1 (Figure 4A and B). For example, less than half of the targets of Gcn4 in clade D (the LCA of pre- and post-WGD species) remained as Gcn4 targets in its two daughter clades B (post-WGD species) and C (pre-WGD, *Kluyveromyces* species). This plasticity at the clade level is consistent with our initial analysis of Gcn4's target turnover at the species level.

For many of the regulatory motifs (72%), the targets are substantially changed at a specific point in the phylogeny. For example, the Mig1 motif, involved in glucose repression in *S. cerevisiae* (Nehlin and Ronne, 1990), is detectable in species across the phylum (up to the LCA, clade L), including a set of ancestral targets in clade D (*Kluyveromyces* and post-WGD, spanning *S. cerevisiae* and *Kluyveromyces lactis*) and in clade G (*Candida*), but with no statistically significant set of shared ancestral targets between these two clades (Figure 4C). Thus, although the motif likely existed in their shared ancestor (clade H), its targets have diverged significantly between the two descendant clades, precluding reconstruction of the ancestral state. This suggests substantial plasticity in the targets associated with many regulatory DNA motifs.

To quantify the extent of plasticity of motif targets, we developed a model of motif-targets turnover, which handles the gains and losses of a target gene as a stochastic continuous-time Markov process (see Materials and methods). This model is akin to standard models of sequence character evolution (Felsenstein, 1981). We found that motif targets are globally gained and lost at fast rates (Supplementary Figure 2), with a median loss rate per target of 5.2 losses/tU (time unit) and a median gain rate per target of 0.24 gains/tU (Supplementary Table 2; Supplementary Figure 2, see Materials and methods). This discrepancy in the rates is due to differences in the pool of targets versus non-targets in the genome. The typical gain rate is 'lower' than the loss rate since it is calculated as a fraction of a larger number of non-target genes (∼4000), whereas the loss rate is calculated out of ∼100 ancestral target genes.

An instructive measure for the target turnover rates is the number of targets we expect to be retained at different branch lengths, computed by averaging simulations over the expected gain and loss rates of all regulatory motifs (Figure 5A, see Materials and methods). Turnover rates vary substantially among individual motifs. For example, the Hsf1 (heat shock factor) motif exhibits low rates of target gain and loss (Figure 5B), while variants of the CACGTG motif (bound by Pho4, Tye7, and Met28) have very high turnover rates (Figure 5C). On average we found that only 7% of a given motif's targets in the *sensu-stricto* clade (clade A, Figure 3A) are expected to be conserved in the LCA of the phylogeny (clade L, Figure 3A), and only 16% of the targets were conserved since the LCA with the *Candida* clade (clade H, Figure 3A).

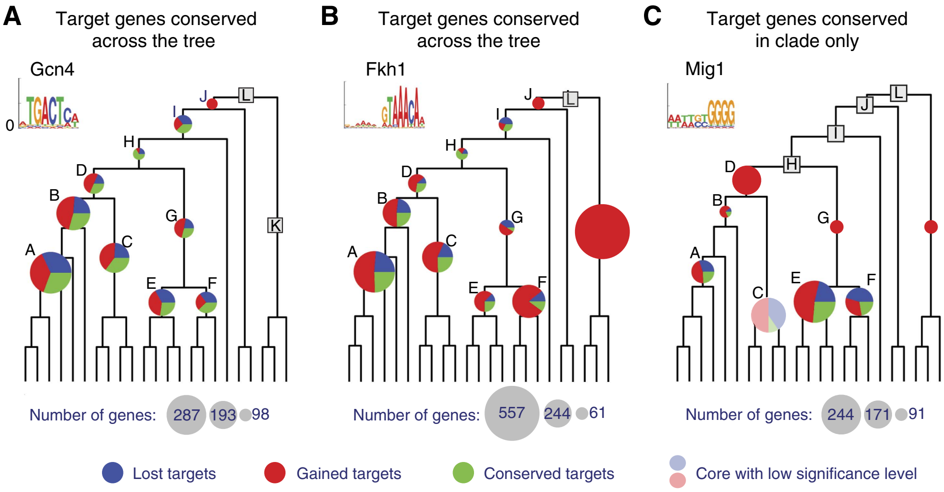

Figure 4 Turnover of motif targets across clades. (**A**–**C**) Comparison between the sets of ancestral targets of a clade and its immediate ancestral clade. Examples are shown for the targets of the Gcn4 motif (A, conservation across all clades despite turnover), the Fkh1 motif (B, motif is detectable in all species and clades, with no ancestral sets in the LCA), and the Mig1 motif (C, complete turnover between clades D and G). Pie charts at internal nodes reflect fractions of conserved (green), gained (red), and lost (blue) targets compared with the immediate ancestral clade; circle area is scaled to the number of target genes in the ancestral set (only clades with ancestral sets have charts, transparent chart indicates a borderline statistical significance of the ancestral set).

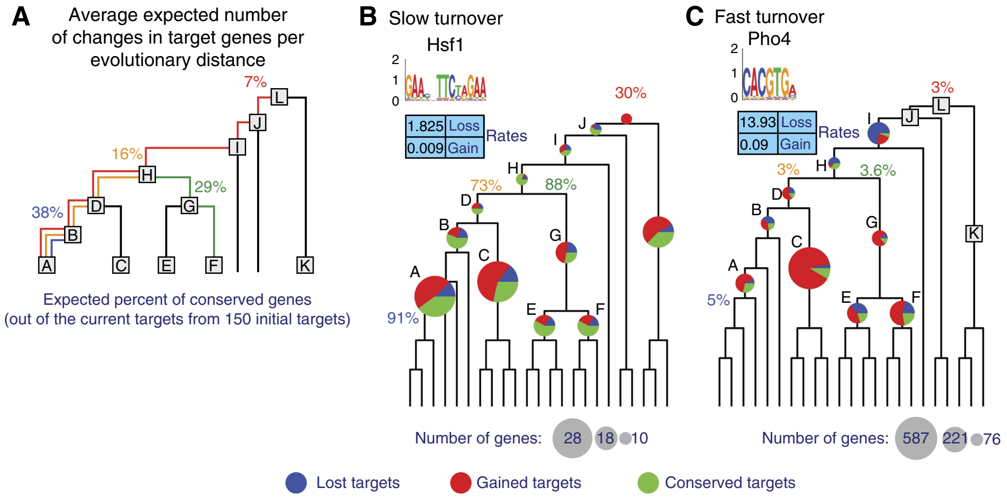

Figure 5 Gain and loss rates of motif targets. (**A**) Average expected fraction of conserved targets at different evolutionary distances, across all regulatory motifs, based on the targets turnover rates computed for each motif separately. The number shown is the fraction of extant targets expected to be derived from an ancestral target, assuming 150 ancestral targets at different phylogenetic distances. (**B**, **C**) Turnover rates for motifs with high turnover rates (Hsf1; B) and low turnover rates (Pho4, CACGTG; C). For each motif, shown are the turnover rates for gain and loss of a target (table), the fractions of conserved (green), gained (red), and lost (blue) targets (pie charts, as in Figure 4), and the expected number of conserved targets computed by the rates (%).

Associating DNA motifs with regulatory functions

To assess the functional implications of target turnover, we next associated each motif in each clade with a regulatory function, based on the functional categories to which its targets in the clade belong. The substantial redundancy between functional annotations can lead to many overlapping ‘functions’, which are hard to compare across clades (Supplementary Note 3). We therefore developed a method to create functional modules that contain genes that share functional annotations and are all ancestral targets of the same regulatory motif (see Materials and methods, Supplementary Figure 3a). For example, consistent with our initial analysis, Gcn4 targets in each clade are associated only with the amino-acid metabolism module (Figure 6A). This module includes several overlapping gene sets, such as amino-acid biosynthetic process (Ashburner *et al*, 2000), amino-acid metabolism (Segal *et al*, 2003), amino-acid nitrogen metabolism (Segal *et al*, 2003), or pyridoxal phosphate binding (Ashburner *et al*, 2000). Notably, each motif can be associated with one or more such modules in each clade, and possibly with different modules in different clades (Supplementary Tables 2 and 3; Supplementary website).

Compared with direct enrichment of individual gene sets, functional modules are a more concise and non-redundant representation that can be easily compared across the

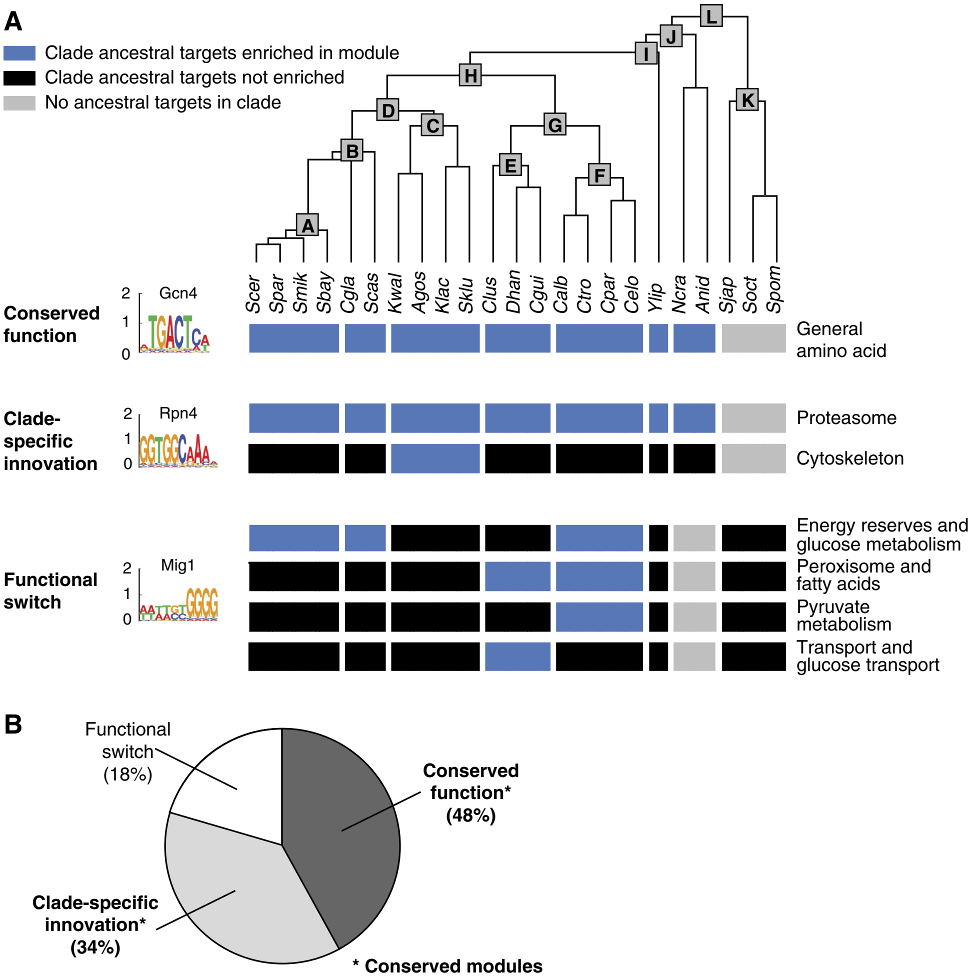

Figure 6 Functional conservation and innovation of DNA motifs across clades. (**A**) Examples of functional conservation and innovation patterns. In each case, the enrichment of motif-target genes with different functional modules is shown across the clades (blue: targets enriched in module; black: not enriched; gray: no ancestral targets in clade), demonstrating functional conservation of the Gcn4 motif (top), clade-specific innovation of the Rpn4 motif (middle), and functional switch of the Mig1 motif (bottom). Additional examples are shown in Supplementary Figures 3 and 4. (**B**) Distribution of functional conservation patterns for *cis*-regulatory motifs. Pie chart of the fractions of motifs associated with complete functional conservation (dark gray), clade-specific innovation (light gray), or a functional switch (white).

phylogeny. In addition, they are more robust to the choices of detection threshold, to the threshold over enrichment of functional categories with motif targets, and to the threshold for merging functional categories (Supplementary Note 3; Supplementary Table 7). Furthermore, supporting our procedure and CladeoScope's predictions, our functional assignments are consistent with known functions of the associated transcription factors in *S.cerevisaie, C. albicans*, and *S. pombe*, for most motifs with a known function (75%) (with another 12% of the motifs with a partial match; Supplementary Note 3; Supplementary Table 8).

Extensive functional conservation of regulatory DNA motifs

We observed functional conservation for a large fraction of the regulatory DNA motifs. In all, 48% of motifs are associated only with the same functions in all clades in which the motif is detectable, even across large phylogenetic distances. Examples include the Gcn4 motif (Figure 6A), the Hsf1 motif with a heat shock module (Supplementary Figure 3b), and the Mbp1 motif with cell-cycle and DNA replication modules (Supplementary Figure 3b). Furthermore, although in other cases the motif might gain or lose an association to functional modules during evolution, 82% of all the motifs have at least one conserved function across all clades (Figure 6B).

Innovations through expansion and switch of functions

In some cases, turnover of target genes does contribute to evolutionary innovation, by either expanding or switching the scope of functions ascribed to a regulatory DNA motif (Figure 6A). For 34% of the motifs, we observed *clade-specific expansion*: a motif gains a new function in a specific clade in addition to maintaining its ancestral function. In such cases, the motif is identified in genes from the same functional module(s) in all clades where it was detected, and is also associated with an additional module unique to a specific clade.

We find various innovations in different clades (Figure 6A; Supplementary Figure 4). For example, the Rpn4 motif is associated with the proteasomal module in all clades (Mannhaupt and Feldmann, 2007), while in clade C (the *Kluyvermyces* species) it is also identified in genes of a cytoskeletal module (Figure 6A). There are several cases of highly conserved motifs exhibiting innovations in the remote *Schizosaccharomyces* clade (K). For example, the cell-cycle motif Fkh1 regulates genes involved in meiosis specifically in this clade, and the Hap4 motif associated with oxidative phosphorylation in all clades also regulates extracellular matrix and GTP binding genes in the *Schizosaccharomyces* species (Supplementary Figure 4a). This latter example involves a regulatory switch, as the regulation of GTP binding genes in all other clades is regulated by the Sfp1 motif (Supplementary Figure 4b).

For 18% of the regulatory motifs we observed a functional switch between clades: the same motif has target genes from distinct functional modules in different clades, thus losing one function while gaining another. For example, the Mig1 motif is associated in the *Candida* (G) clade with modules such as peroxisome and fatty acid metabolism, whereas in the *Kluyveromyces* (C), the 'post-WGD' (B), and the *Schizosaccharomyces* (K) clades it is associated with other carbon metabolism modules (Figure 6A). An additional example is the motif bound by the factor Ynr063w (Zhu *et al*, 2009). This motif is associated with general metabolic processes in all clades where it is detected, but switches its specific function: it is associated with the TCA cycle in the *Candida* clades (E–G), glycolysis in *Schizosaccharomyces* clade (K), but with the peroxisome and aerobic metabolism in the 'post-WGD' clade (A–B) (Supplementary Figure 4a).

Conservation of regulatory function despite high turnover rate of targets

The observations of substantial target turnover and extensive functional conservation are seemingly contradictory. One possible way to reconcile this contradiction would be if the rapid turnover of motif targets is mainly restricted to motifs that exhibit functional changes, but not to those with conserved functions. However, we find rapid target turnover for most regulatory DNA motifs, including those associated with conserved functional modules (Supplementary Tables 2 and 4), such as Gcn4.

Moreover, we observed extensive turnover of motif targets within the functional modules themselves. Specifically, in 80% of modules associated with the same motif in more than one clade, we observed substantial turnover of the motif targets between those clades (Figure 7A). On average, 62% of a module's genes are associated with the regulatory motif in only a minority of the relevant clades. For example, the Fkh1 motif is consistently associated with a cell-cycle regulation module across the entire phylum (12 clades), but its individual targets substantially turnover, with ~90% of genes detected as Fkh1 targets in only one or two clades (Figure 7B).

A functional selection turnover model

The observed conservation of regulatory function despite high target turnover suggests that the global functional roles associated with a regulatory motif are under stronger selection than the individual regulatory interactions. To formalize this notion we propose the *Functional Selection Turnover Model*, where selective pressure acts differentially to conserve motif-target relations within the same biological process (compared with outside of the process), but not particular target genes within that process (Figure 7C).

To test this hypothesis, we used a likelihood ratio test (LRT) to compare two alternative evolutionary models (see Materials and methods): (1) a 'neutral' turnover model, where targets are gained and lost at the same rates regardless of the functional module to which they belong and (2) a 'module-specific' turnover model (described above), where turnover rates—both gain and loss—are different for targets in the functional module compared with those outside. We applied this test to all functional modules in all associated clades (a total of 745 tests).

In nearly all cases (96%, 715 tests), target turnover is significantly constrained by the genes' function (*P*-value <0.05 after Bonferroni correction, Figure 7D). Most notably, the probability to gain an additional target gene within the same functional module is typically at least two orders of magnitude higher than the probability to gain a new target from genes outside of the module (Supplementary Table 4). Thus, gain and loss of target genes are highly constrained by their function, resulting in conservation of the motif's functional role despite turnover at individual sites. These results are not sensitive to the choice of parameters used in the process of target prediction or in defining functional modules, and hence are not an artifact of specific threshold choices made in our computational analysis (Supplementary Note 3).

The functional selection turnover model explains the number of highly conserved targets

Against the backdrop of rapid turnover, some motif targets remain highly conserved. For example, 25 of the Gcn4 targets have Gcn4 binding motifs in their promoters in every clade (out of an average of 130 Gcn4 targets per clade). Such conservation may reflect an important specific function of these particular genes; alternatively, a few conserved genes may be expected by chance, given the functionally constrained turnover rate of the motif and the size of the functional module. To distinguish between these possibilities, we performed simulations to estimate the probability of the observed number of highly conserved targets under our functional selection model (see Materials and methods), assuming a differential turnover rate for the targets, based on their function but not based on their individual identity. We examined 20 regulatory DNA motifs that have ancestral targets broadly conserved across clades, such as Gcn4 (Figure 7E), Mbp1, and Rpn4 (Supplementary Table 5).

For all motifs tested, we could not reject the null hypothesis that the observed number of highly conserved targets is consistent with the overall turnover rates according to the Functional Selection Turnover Model ($P\geqslant 0.5$). Thus, even the number of highly conserved targets is consistent with selection at the module level rather than selection towards the individual function of each gene within the module.

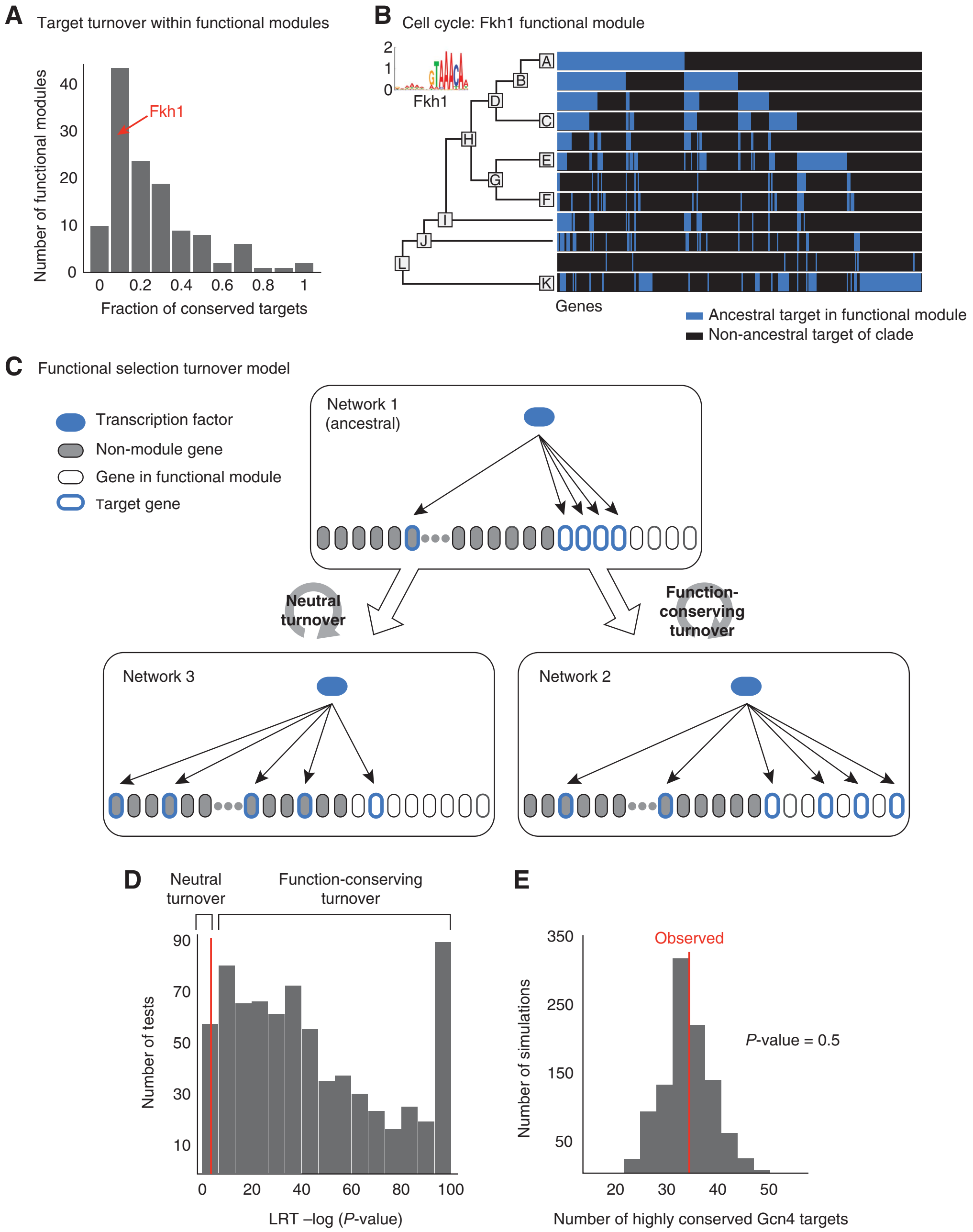

Figure 7 Target turnover and the functional selection turnover model. (**A**) Extent of target turnover within functional modules. The distribution of the percent of conserved targets (*x*-axis, defined as targets of a motif in the majority of the clades associated with the functional module), for all functional modules with targets in at least two clades. The bin pertaining to the Fkh1 cell-cycle module is marked with a red arrow. (**B**) Fkh1 target turnover within the cell-cycle module. Target genes (rows) for the Fkh1 motif within the cell-cycle module across the clades (columns). Blue: target in a clade; black: non-ancestral target in the clade. (**C**) Functional Selection Turnover Model. Cartoon illustration of the model (white: gene in functional module; gray: gene not in module; blue border: target gene of motif; blue node: motif/transcription factor) with two alternative scenarios for target genes turnover from the ancestral Network 1 (top). Both scenarios (bottom) show extensive turnover of target genes. The functional selection scenario (Network 2, bottom right) has selection on the genes' function, as reflected by the module to which they belong, but not on individual targets, and leads to enrichment of targets within the module along with turnover of individual targets. The module-neutral scenario (Network 3, bottom left) has random turnover of targets and thus leads to loss of target enrichment within the module genes. (**D**) Testing the functional selection model for *cis*-regulatory site turnover. The distribution of the log *P*-value of the LRT between the two models for target turnover, a module-neutral turnover model (H_0, Figure 3C, bottom left) and a functional selection turnover model (H_1, Figure 3C, bottom right), for 745 functional modules and each of their associated clades. Red line—threshold of *P*-value 0.05 after the Bonferoni multiple hypothesis correction for rejecting the H_0 model. (**E**) The number of highly conserved Gcn4 target genes is as expected given the functional selection turnover model. The distribution of the number of expected highly conserved Gcn4 targets from 1000 simulations, according to the functional selection turnover model. The observed number of Gcn4 targets conserved up to clade H is 35 (red line), with an empirical *P*-value $\geqslant 0.5$. Thus, we cannot reject the null hypothesis that the number of highly conserved genes is as expected by the functional selection turnover model.

The functional selection turnover model is consistent with transcription factor binding data measured across yeast and mammalian species

To examine the generality of our results, we tested whether they hold at the level of individual species as well as clades, when targets are determined experimentally rather than computationally. We thus examined published *in vivo* transcription factor binding data (from ChIP-chip or ChIP-seq experiments) (Borneman *et al*, 2007; Tuch *et al*, 2008; Schmidt *et al*, 2010). Recent functional studies of transcription factor binding to DNA reported substantial divergence in the bound targets of conserved transcription factors in *Ascomycota* yeast species (Borneman *et al*, 2007; Tuch *et al*, 2008) and between mammalian species (Schmidt *et al*, 2010). These include Mcm1 binding measured across three relatively distant species (*S. cerevisiae*, *K. lactis*, and *C. albicans*) (Tuch *et al*, 2008), Ste12 and Tec1 binding in three closely related *Saccharomyces* species (*S. cerevisiae*, *Saccharomyces mikatae*, and *Saccharomyces bayanus*) (Borneman *et al*, 2007), and HNF4α measured across three mammalian species (human, mouse, and dog) (Schmidt *et al*, 2010).

Consistent with our *cis*-regulatory analysis, the binding profiles of all four factors demonstrate high turnover of targets within conserved functional modules (see Materials and methods, Supplementary Table 6; Figure 8), in addition to some species-specific innovations. Applying the two tests described above, we find that the Functional Selection Turnover Model fits the binding data of these four factors in all species ($P<10^{-12}$), and that the number of highly conserved targets of these factors is as expected by the model ($P>0.2$). Notably, in mammals the results are not sensitive to the specific threshold for associating an upstream binding site with a target gene (see Materials and methods). Overall, this analysis demonstrates the generality of our findings at different evolutionary distances, measurement methods (sequence analysis and ChIP assays), phylogenetic resolution (species and clades), and group (yeast and mammals).

Discussion

We applied a novel approach for reconstructing *cis*-regulation to 23 *Ascomycota* species to study the evolution of their *cis*-regulatory networks. Using this approach, we systematically identified *cis*-regulatory interactions for 88 known regulatory DNA motifs across the 23 species, their conserved target genes in each clade and their functional annotations. We exploited this resource to study the regulatory history of

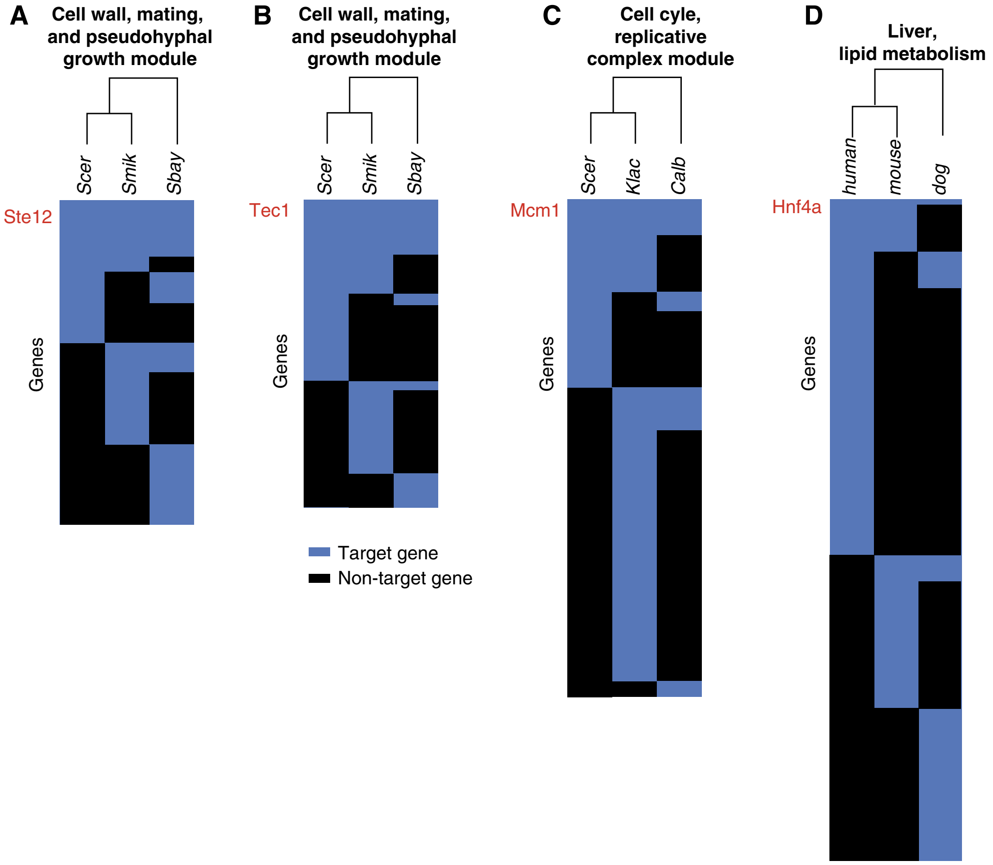

Figure 8 Turnover of target genes within functional modules from experimentally measured binding profiles in yeasts and mammals. The target genes (rows) for each species (columns), associated with a conserved functional module of different transcription factors: (**A**) Ste12 (Borneman *et al*, 2007) in yeasts, (**B**) Tec1 (Borneman *et al*, 2007) in yeasts, and (**C**) Mcm1 (Tuch *et al*, 2008) in yeasts (**D**) HNF4α (Schmidt *et al*, 2010) in mammals. Blue: target in a species; black: non-target in the species.

specific transcription factors and to reach general principles of regulatory evolution. In addition, this constitutes a rich public resource (http://www.compbio.cs.huji.ac.il/OrthoMotifs/) that will facilitate future studies of regulatory evolution of individual clades or species, including human and plant pathogens.

We find and quantify a pervasive gain and loss of motif targets at high evolutionary rates. The high turnover rates that we estimate from data for all motifs (average ~7% expected conserved targets from the LCA of the phylum, Figure 5A) are reflected by the small number of highly conserved targets, and by the complete switch of targets for 72% of motifs in at least one point in the phylogeny. These high turnover rates generalize previous binding studies of few individual transcription factors across three yeast species (Borneman *et al*, 2007; Tuch *et al*, 2008), four flies (Moses *et al*, 2006; Bradley *et al*, 2010) and five mammals (Schmidt *et al*, 2010).

The seemingly contradictory trends of a broad conservation of the functions associated with a motif, and the pervasive gain and loss of the motif in individual targets within the module, are reconciled by our proposed Functional Selection Turnover Model, implying that it is a general principle of regulatory evolution. Such conservation of a transcription factor's cellular function but high turnover of its individual targets had been previously indirectly implicated in the comparison of cell-cycle genes between two yeast species (*S. cerevisiae* and *S. pombe*) (de Lichtenberg *et al*, 2007) and for liver-specific transcription factors across five vertebrates (Schmidt *et al*, 2010). Our analysis suggests that it is a broad and general phenomenon and our model shows that it can be explained by a strong selection to conserve the function of the motif, but a weaker selection over the specific target genes within this function. This evolutionary model accounts for patterns of turnover from direct measurements of transcription factors across individual species in yeast and mammals, suggesting that the same principle applies at different evolutionary distances, measurement methods, phylogenetic resolution (clades and species), and remote phyla.

There are several alternative potential explanations for the observed conservation of regulatory function despite high target turnover. First, determination of transcription factor targets based on *cis*-regulatory elements (rather than on limited experimental ChIP data) is challenging and noisy (Gasch *et al*, 2004; Tanay *et al*, 2005; Wohlbach *et al*, 2009), resulting in many false positives and false negatives, which may lead to low overlap in target genes between species. To exclude this option, we compared evolutionary conserved target genes at the clade level instead of target genes for individual species. Second, the transcription factor may target additional genes in some species, thus expanding the scope of functions it regulates, as has been previously shown in yeasts for various factors, such as Mcm1 (Tuch *et al*, 2008). Although we detect such expansions in several cases (Figure 5; Supplementary Figure 4), we detect high turnover within the large majority of functional modules (Figure 7), including highly conserved modules.

Finally, the high degree of target turnover within a module may be facilitated by the fact that many target genes are co-regulated within dense overlapping regulons (Alon, 2007), where multiple factors have overlapping roles. In 'single input modules' (Alon, 2007) (e.g., the galactose utilization pathway in yeast), all the genes in a module are co-regulated by one factor, and we expect strong target conservation. Conversely, in a Dense Overlapping Regulon (e.g., Ribosomal Protein genes), multiple transcription factors regulate the module's genes, and are partly redundant, such that loss of one regulator might be compensated for by another. For example, amino-acid metabolism genes are commonly regulated by Gcn4 and Leu3, with loss of the regulation by one transcription factor compensated for through gain of regulation by the other (Tanay *et al*, 2005; Tsong *et al*, 2006; Hogues *et al*, 2008; Tuch *et al*, 2008; Wohlbach *et al*, 2009; Weirauch and Hughes, 2010) (Supplementary Figure 5b and c). This would be consistent with the conserved co-expression of many functional modules in yeast (Tanay *et al*, 2005; Hogues *et al*, 2008) and mammals (Odom *et al*, 2007). More broadly, transcriptional regulation is only one of many regulatory layers, and control of one or a few members of a complex or pathway may determine the activity level of the whole complex (de Lichtenberg *et al*, 2007). Thus, a transcription factor may influence the activity of a cellular process by targeting a few genes, and loss of regulation of one target can be compensated by gain of regulation by another transcription factor.

Overall, the function-centered model of targets turnover provides an important insight into the use of conservation as a filter for functional elements in comparative genomics studies (such as ChIP experiments that rely on evolutionary conservation to filter out noise in transcription factor target genes (Harbison *et al*, 2004)). Moreover, by taking a function rather than a gene-centered view of *cis*-regulatory evolution, our findings suggest that selection forces are more permissive than has been previously assumed. At the module and transcription factor levels, although turnover within a module may not affect the overall regulatory role of a factor, it may allow for more subtle fine-tuning of gene regulation, facilitating adaptation while controlling against dramatic changes in phenotype.

Materials and methods

CladeoScope algorithm: phylogenetic reconstruction of *cis*-regulatory networks

The CladeoScope algorithm reconstructs *cis*-regulatory networks across species: It learns species-specific DNA motifs (including in species lacking any functional annotations and known motifs), using prior knowledge about known PWMs in a model organism, and computationally adapting them to each species (see species-specific motifs). For each motif it then assigns a set of ancestral target genes in the LCA of each clade of species across the phylogeny (Figure 1), inferred using a maximal parsimonious phylogenetic reconstruction (see Phylogenetic reconstruction of ancestral targets).

Initially, a gene is predicted to be targeted by a regulatory DNA motif in a species if it contains a binding site of the motif in its promoter (see Motif scanning for putative targets). These predicted targets are used in the reconstruction to find the ancestral set of target genes. In addition, after the reconstruction of ancestral targets, we use them as an input to the motif refinement per species, and choose the optimal motif (see species-specific motifs). The DNA motifs in each species and the ancestral targets in each clade are filtered based on evolutionary conservation within clades of species by their statistical significance (see Phylogenetic filter for noisy motifs and statistical significance). The resulting resource of species-specific motifs, ancestral target sets and functional modules per clade of species are available for download at our Supplementary website: http://www.compbio.cs.huji.ac.il/OrthoMotifs.

Pseudocode:

```
Caldesosope (PWM, promoter sequences, gene trees):
1. Find provisional motif targets in each species by scanning promoters.
2. Learn species-specific motifs:
    a Apply motif discovery algorithm on the provisional target set initialized by the PWM.
    b Rescan for putative targets using new motif.
    c Repeat steps a–b using the provisional targets defined in b.
3. Reconstruct ancestral motif targets for each clade in the phylogeny from the provisional species sets using maximum parsimonious dynamic programming algorithm.
4. Repeat step 2 starting with the ancestral targets in each clade, and choose the motif conforming to the higher enrichment threshold.
5. Filter motifs:
    d Filter motifs in each species by enrichment of motif targets with ancestral targets.
    e If any motifs are removed go back to step (3).
    f Filter motifs in the clade level based on statistical significance of the number of ancestral targets.
```

We now describe the procedure involved in each step of this psuedocode.

Motif scanning for putative targets (steps 1 and 2b)

To identify the putative targets of a motif in the genome, we score each gene's promoter by summing over all possible positions of the promoter on both strands (as in Tanay, 2006), taking into account the nucleotide background distribution in the promoters of the relevant genome:

$$\text{score} = \log \sum_{i=0}^{N-M+1} \left(\prod_{j=1}^{M} \frac{P_{\text{PWM}^j}(n_{i+j})}{P_{\text{BG}}(n_{i+j})} \prod_{j=1}^{M} \frac{P_{\text{r}-\text{PWM}^j}(n_{i+j})}{P_{\text{BG}}(n_{i+j})} \right)$$

Where N is the length of the promoter; M is the length of the motif; n_i is the nucleotide at the i'th position of the promoter; P_{BG} is the background distribution of nucleotides in all promoters of the genome; P^{j}_{PWM} is the probability vector for nucleotides in position j of the motif and similarly $P^{j}_{\text{r-PWM}}$ for the reverse motif (equivalent to searching the reverse strand of the DNA). We considered the 600 base region upstream of each gene's ATG as its promoter, truncating this region whenever it overlapped a neighboring gene.

We define the target set of the motif as those genes whose promoters have a score above a threshold $T = 0.8^*$ (mean of 20 highest scoring promoters in the genome). The threshold and scanning method where determined by optimizing the precision rate and sensitivity of predictions of *in vivo* transcription factor target genes from ChIP-chip (Harbison *et al*, 2004) assays in *S. cerevisiae* of two different transcription factors: Hsf1 and Rpn4 (Supplementary Note 4). Arguably, this might bias our choices to levels of binding that are significantly detectable by these assays. However, perturbation analysis of this threshold show that our result are mostly robust to this choice (Supplementary Notes 1–3). Since this score is relative to each genome and to each motif, we exclude motifs that do not have any occurrences in the genome by filtering out motifs whose highest score in the genome is $<50\%$ of the maximum possible score of this motif in a single location. Additionally, we removed motifs from the collection if the number of inferred targets was >1500 (the upper bound was chosen to exceed the maximal number of promoters bound by any transcription factor in *S. cerevisiae*, as measured by ChIP-chip (Harbison *et al*, 2004)).

Species-specific motifs (step 2)

Our underlying assumption in this step is that the binding specificities of transcription factors, represented as DNA motifs, are largely conserved, even when their specific target genes and functional roles may have substantially diverged (Wapinski *et al*, 2007; Tuch *et al*, 2008; Schmidt *et al*, 2010). We therefore initiate our reconstruction with DNA motifs for known transcription factors that have been experimentally determined in model organisms. CladeoScope uses the MEME motif discovery algorithm (Bailey and Elkan, 1994) on the promoters of the putative targets of each initial motif in each species, with the initial motif's consensus sequence as the initialization point to the algorithm. MEME is parameterized to identify motifs on either strand, of length within two bases from the input consensus, and it is given the species-specific nucleotide distributions as background models for learning. Of the top two motifs reported by MEME, the highest scoring motif in this species is then used to rescan the species' genome and to identify a revised set of targets. CladeoScope repeats this process of motif discovery and rescanning to refine the motif and its target set once; typically the motif is not altered after the first iteration. To allow more variation in the motifs, we repeat the process, initializing the refinement with the conserved ancestral targets. This allows us to find motifs in species where the first iteration did not succeed. CladeoScope chooses the motif with the highest enrichment score between these iterations.

Phylogenetic reconstruction of ancestral targets (step 3)

To infer the ancestral motif targets we trace regulatory events across orthologous loci. CladeoScope handles genes derived from a common ancestor gene in the root of the phylogeny as related ('orthogroup' in the terminology of Wapinski *et al*, 2007), and defines an orthogroup as a target of a motif if it is predicted as a target in at least one of the ancestors in the phylogeny. The reconstruction is done using maximum parsimony (Fitch, 1971) to minimize the number of target gain and loss events along the branches of the tree. This is done separately for any potential ancestral target gene by a dynamic programming algorithm. The inputs to the algorithm are (1) the phylogenetic gene trees for each set of orthologous genes (Wapinski *et al*, 2007) and (2) a binary classification denoting whether each gene in the tree is a predicted target of the motif in each species. This reconstruction accounts for gene duplications and losses, distinguishing a lost gene from a present gene that is not a target, and operates independently on each paralogous lineage following gene duplication events, by utilizing gene trees when reconstructing ancestral targets. Given that most of these species have diverged sufficiently to lose sequence similarity at the promoters, paralogs will not necessarily be co-targets of a motif due to spurious conservation of their promoters. Thus, paralogs can be in different motif-target sets in the CladeoScope output.

Phylogenetic filter for noisy motifs and statistical significance (step 4)

We use phylogenetic conservation within a clade of species as a filter for noisy predictions of target genes (described above) as well as the DNA motifs themselves. CladeoScope filters the motifs for each species independently by enrichment of their putative target genes in that species with ancestral targets in any relevant clade (hypergeometric P-value <0.001). Since filtering the motifs (step 4) and the reconstruction of ancestral targets (step 3) are dependent, we solve this problem by iterating between the two steps. If any insignificant motifs are found in the clade, the most insignificant one is removed, and CladeoScope returns to step 3 of reconstructing the ancestral targets. This filtration per species allows for a motif to be detected in a clade of species, although it is not functional in a single species but is functional in all other species in the clade. We then filter the motifs at each clade, requiring that the number of inferred targets for a motif in the clade's ancestor be statistically significant (P-value $\leqslant 0.005$) against the null hypothesis that the targets predicted in the individual species are independent. We compute an empirical P-value by simulating target sets of the relevant size for each species in the clade, and reconstructing ancestral targets from these random sets. This process is repeated 1000 times to estimate the probability of getting a set of ancestral targets of a certain size or larger by chance. A motif is detectable in a clade if it has a statistically significant (P-value <0.005)

set of ancestral targets in the clade. Finally, we exclude motifs that are found to be significant only in clade A (*sensu-stricto*), since in this clade promoter sequences have not evolved enough for random occurrences of a motif have a non-trivial chance to be conserved.

Validating CladeoScope's performance using synthetic data

We generated simulated target sets in extant species for evaluating CladeoScope's robustness (Supplementary Figure 1) by evolving targets from an ancestral set of targets using turnover (gain and loss) rates of target genes, with several variations. First, we used two types of noise factors: (1) the proportion of erroneous target genes relative to the species true motif targets (false positives, ranging between and 0% and 200% of erroneous targets within each species set) and (2) the proportion of missing (true) targets not included in the species motif target set (false negatives, ranging between 0% and 60% of removed targets from each original species set). Second, we varied the size of the ancestral target set. Using ancestral motif targets in clade A (Figure 3A): 22 targets of Hsf1, 198 targets of Mbp1, 297 targets of Fkh1. Third, we varied the degree of targets turnover (the fast turnover is the average frequency measured in clade E (Figure 3A) over all motifs: Fast $F_{\text{gain}}=0.002$ $F_{\text{loss}}=0.3$, Medium $F_{\text{gain}}=0.0002$ $F_{\text{loss}}=0.03$, Slow $F_{\text{gain}}=0.00002$ $F_{\text{loss}}=0.003$). Fourth, we estimated the gain and loss frequencies for each of the three motifs in each relevant species directly from the data (as in the LRT described below) (Supplementary Note 1). Fifth, we used two topologies of the species tree: the topology in the *sensu-stricto* clade (clade A, Figure 3A), and the asymmetrical topology in the *Candida* clade (clade E, Figure 3A). Overall, we considered 960 combinations of these parameters. For each set of parameters, we executed CladeoScope and calculated sensitivity and specificity measures averaged over 100 independent simulations.

Assessing performance on random motifs

We created random motifs by concatenating randomly sampled positions from all known motifs from the literature (using all motifs from *S. cerevisiae*, as described below). We confirmed that the random motifs we constructed were not similar to any known motifs, comparing the random motifs to all known motifs using BLiC (Habib *et al*, 2008). For each random motif, we scanned for targets in each species (as described above), and ran CladeoScope to reconstruct the ancestral sets. We then computed an empirical *P*-value for each motif in all clades using random targets (as described above).

Assessing CladeoScope's robustness to parameters and comparison to the literature

We tested CladeoScope's robustness to variations in different parameters including (as described in Supplementary Note 1): (1) The *P*-value threshold for detection of a motif in a species—We ran the algorithm on nine different motifs across all clades, using seven different thresholds, ranging between $5e-2$ and $1e-5$, and compared the number of ancestral targets reconstructed per clade. (2) The *P*-value threshold for conservation of a motif in a clade—We tested different *P*-value thresholds ranging between 0.05 and 0.001, and compared the number of statistically significant ancestral motifs predicted per clade. (3) The motif-targets detection threshold—We tested three different thresholds (80, 75, or 70% out of the best score per motif and species). In each case, we compared the ancestral and species targets determined by CladeoScope to those from *in vivo* ChIP-chip data in *S. cerevisiae* and four other species (Supplementary Note 1).

Targets turnover rates and expected number of changes in target genes

For each motif we computed the turnover rate of its target genes based on the following model. The model treats each pair (motif, gene) as a binary character denoting whether the gene is a target of the motif or not. We model changes in this character (gain or loss events) as a stochastic continuous-time Markov process parameterized by motif-specific rates, one for gain, and another for loss. This model is akin to standard models of character evolution (Felsenstein, 1981). The rates are expressed in terms of expected number of events per time unit (tU), where a time unit corresponds to the time in which one amino-acid substitution per site is expected on average. The model assumes a constant turnover rate of targets along the phylogeny, which is reflected by two parameters for each motif: its gain rate (a) and its loss rate (b), given by the following rate matrix *R*:

$$R=\begin{pmatrix} -a & a \\ b & 1b \end{pmatrix}$$

Given this rate matrix we can compute the probabilities for target gain and loss for a given evolutionary distance *t* using the following equations:

$$P(\text{Gain}\mid t)=\frac{a}{a+b}(1-\mathrm{e}^{-(a+b)t})$$
$$P(\text{Loss}\mid t)=\frac{b}{a+b}(1-\mathrm{e}^{-(a+b)t})$$

We use a maximum-likelihood estimator to infer the parameters in *R* for each motif. The likelihood is computed based on sufficient statistics for each clade relative to its immediate ancestral clade, including the branch length (*t*), the observed number of gained (N_{gain}), lost (N_{loss}) and conserved (N_{cons}) target genes, and the probabilities described above, as:

$$\mathrm{In\,Likelihood\,(Motif)}=\sum_{b\in\text{branches}}$$
$$\left[\begin{array}{l} N^{b}_{\text{gain}}\,\mathrm{In}\,P(\text{Gain}\mid t^{b})+N^{b}_{\text{loss}}\,\mathrm{In}\,P(\text{Loss}\mid t^{b}) \\ +N^{b}_{\text{cons}}\,\mathrm{In}\,(1-P(\text{Loss}\mid t^{b}))+N^{b}_{\text{notTarget}}\,\mathrm{In}\,(1-P(\text{Gain}\mid t^{b})) \end{array}\right]$$

The maximum-likelihood estimator is found by a gradient descent algorithm using Matlab's *fminunc* function. We assume the tree topology and branch length are known (see Species phylogeny).

Annotating motifs with functional modules

To associate motifs with regulatory functions, we cluster functional gene sets together by the fraction of associated motif targets shared between them, creating sets of functional modules containing genes that share functional annotations and are all ancestral targets of the same regulatory motif in at least one clade (Supplementary Figure 3A).

The method is applied to each motif separately. As input we provide the ancestral target genes of the motif in each clade, and gene sets of functional annotations from various sources. In step 1 (*Initialization*), we identify all functional annotations enriched in each set of ancestral targets in each clade using Fisher's exact test ($P<0.01$ after correction; however, the results presented here are robust to various thresholds), and define a functional module as genes from each enriched category that are ancestral targets in any clade. In step 2 (*Merge functional modules*), we merge modules according to the fraction of associated motif targets shared between them. In this greedy procedure, we start from the most enriched module, choose another one that is most highly overlapping with it (at least 60% gene membership overlap; however, the results presented here are robust to various thresholds) and unite them into a new functional module, eliminating the two daughter modules from the collection. In step 3 (*Recalculate enrichments*), we recalculate the enrichment of the ancestral targets in each clade with this new functional module. We repeat steps 2 and 3 until no further functional modules are merged. Following the automatic assignment of modules, we manually annotated each functional module with a biologically meaningful label based on its underlying annotations.

The functional modules assigned to each motif, and the target genes assigned to each module can be found in Supplementary Table 3. Modules and genes created by different sets of parameters can be found in Supplementary Table 9.

Note that the assignment to functional modules is based on phylogenetic projection from *S. cerevisiae*, *C. albicans*, and *S. pombe* annotations. As a consequence, the function assignment often cannot distinguish between paralogs. Moreover, in a previous study of functional evolution (Wapinski *et al*, 2007), we show that when we can evaluate such divergence, most paralogs maintain the same functional category. Thus, we expect some functional modules to be enriched for paralogs (e.g., the Ribosome, due to the massive duplications of genes encoding ribosomal proteins). This, however, reflects a real phenomenon.

The classification of motifs to three functional categories

The classification of motifs to three functional categories is based on the explicit definition of each category: For *conserved* motifs, all associated modules are enriched across (the relevant) clades. *Expanded* motifs have at least one module conserved across clades, and an additional module specific for a subset of these clades. *Switched* motifs have modules enriched in different subsets of clades. We applied these definitions with one exception, in which we excluded small modules (with <5 genes uniquely classified to the module). To increase confidence in these results, we subsequently examined each classification individually, and compared the module enrichment pattern to the conservation pattern of target genes across species. We marked in Supplementary Table 2 cases where this examination raised doubts in the classification. To further ensure that these classifications are not an artifact of specific thresholds in module construction, we selected three representative parameter sets and applied this processes separately to each one of them (Supplementary Table 2). For 18 motifs (including all the ones mentioned in the text and figures), we performed a more extensive comparison of module-construction parameters (Supplementary Table 7). Notably, the examples throughout the text and figures include only robust examples across all parameters. For the analysis presented in the manuscript and in Figure 6B, we constructed a consensus classification for each motif. We examined the three classifications described above and defined the consensus/majority classification for each motif (Supplementary Table 2). In virtually all cases there was a clear-cut consensus classification.

Functional annotation resources

We used functional annotations from several sources. GO annotations (Ashburner *et al*, 2000) were assembled from the genome databases of *S. cerevisiae* (SGD), *C. albicans* (CGD), and *S. pombe* (GeneDB). Other *S. cerevisiae*-based annotations include transcription modules (Segal *et al*, 2003), MIPS (Mewes *et al*, 2011), KEGG (Kanehisa and Goto, 2000; Kanehisa *et al*, 2006), and mutant phenotypes (Hughes *et al*, 2000). Other *S. pombe*-based annotations include expression clusters (Chen *et al*, 2003). We projected each set of annotations from genes to their orthologs (Wapinski *et al*, 2007) to test gene set enrichments across all clade core-sets, as previously described (Wapinski *et al*, 2007).

Assessing robustness of functional modules and their comparison to the literature

To test the robustness of the functional modules, we applied the algorithm with different parameters and inputs, including (see Supplementary Note 3): (1) Enrichment thresholds for functional modules with motif targets (HyperGeometric *P*-value threshold ranging between: 1e − 3 and 1e − 6). (2) Threshold for merging gene sets (overlap threshold ranging between 40% and 75%). (3) Threshold for initial predictions of target genes (80% or 75% out of the bests score per motif and species).

For each set of parameters we tested several characteristics: (1) the number of modules; (2) the fit of our functional selection turnover model; (3) the classification of motifs to functional classes (functional conservation, Clade-specific innovation or Functional switch), where we examined in detail 18 motifs including those discussed specifically in the manuscript; and (4) robustness of the functional annotations of motifs by the functional modules, where we examined in detail 18 motifs including those discussed in the manuscript.

In addition, we compared the resulting functional modules to known motif and transcription factor annotations from the literature in *SGD*, *CGD* , and GeneDB.

The functional selection turnover model

We used an LRT to determine if the observed functional conservation with widespread turnover occurs by chance or according to our functional selection model. We defined two alternative hypotheses:

H_0: *Module-Neutral turnover*: Targets turnover at the same (‘neutral’) rate regardless of the functional module to which they belong (implying that the functional conservation may be a byproduct of the insufficient evolutionary distance between species).

H_1: *Functional selection*: There is selective pressure on the targets to be gained or lost within modules of genes sharing the same function. Turnover rates—both gain and loss—are different for targets in the functional module compared to those outside.

We applied the LRT to each functional module testing separately each associated clade (total of 745 tests), by computing the likelihood of the observations under each hypothesis and calculating a *P*-value (χ^2 distribution with one degree of freedom). The likelihood computations were based on maximum-likelihood estimates of gain and loss probabilities in the clade relative to its immediate ancestral clade. The required sufficient statistics are the observed number of gained (N_{gain}), lost (N_{loss}), and conserved (N_{cons}) target genes. In the functional selection model (H_1 hypothesis), we computed the gain and loss probabilities separately for genes within the module and genes outside of the module.

Description of the equations:

$$\mathrm{LRT}(\mathrm{Motif}) = ll(\mathrm{Motif} \mid \mathrm{H}_0) - ll(\mathrm{Motif} \mid \mathrm{H}_1)$$
$$ll(\mathrm{Motif} \mid \mathrm{H}_i = \sum_{b \in \mathrm{branches}} ll(b \mid \mathrm{H}_i)$$

$$ll(b \mid \mathrm{H}_0) = N_{\text{gain}} \ln P(\mathrm{Gain} \mid \mathrm{H}_0) + N_{\text{loss}} \ln P(\mathrm{Loss} \mid \mathrm{H}_0) \mid + N_{\text{cons}} \ln[1 - P(\mathrm{Loss} \mid \mathrm{H}_0)] + N_{\text{nonTarget}} \ln[1 - P(\mathrm{Gain} \mid \mathrm{H}_0)]$$

$$ll(b \mid \mathrm{H}_1) = N^{\mathrm{IN}}_{\text{gain}} \ln P(\mathrm{Gain}^{\mathrm{in}} \mid \mathrm{H}_1) + N^{\mathrm{IN}}_{\text{loss}} \mid \ln(P(\mathrm{Loss}^{\mathrm{in}} \mid \mathrm{H}_1) + N^{\mathrm{IN}}_{\text{cons}} \ln[1 - P(\mathrm{Loss}^{\mathrm{in}} \mid \mathrm{H}_1)] + N^{\mathrm{IN}}_{\text{nontarget}} \ln[1 - P(\mathrm{Gain}^{\mathrm{in}} \mid \mathrm{H}_1)] + N^{\mathrm{OUT}}_{\text{gain}} \ln P(\mathrm{Gain}^{\mathrm{out}} \mid \mathrm{H}_1) + N^{\mathrm{OUT}}_{\text{loss}} \ln P(\mathrm{Loss}^{\mathrm{out}} \mid \mathrm{H}_1) + N^{\mathrm{OUT}}_{\text{cons}} \ln[1 - P(\mathrm{Loss}^{\mathrm{out}} \mid \mathrm{H}_1)] + N^{\mathrm{OUT}}_{\text{nontarget}} \ln[1 - P(\mathrm{Gain}^{\mathrm{out}} \mid \mathrm{H}_1)]$$

The maximum-likelihood estimation for the probability of gain and loss of each motif’s target genes in the current clade C_1, compared with the immediate ancestral clade C_{p}:

$$P(\mathrm{Gain} \mid \mathrm{H}_0) = \frac{N_{\text{gain}}}{N_{\text{total}} - T_{C_p}}$$

$$P(\mathrm{Loss} \mid \mathrm{H}_0) = \frac{N_{\text{loss}}}{T_{C_p}}$$

$$P(\mathrm{Gain}^{\mathrm{IN}} \mid \mathrm{H}_1) = \frac{N^{\mathrm{IN}}_{\text{gain}}}{N^{\mathrm{IN}}_{\text{total}} - T^{\mathrm{IN}}_{C_p}}$$

$$P(\mathrm{Loss}^{\mathrm{IN}} \mid \mathrm{H}_1) = \frac{N^{\mathrm{IN}}_{\text{loss}}}{T^{\mathrm{IN}}_{C_p}}$$

Where N_{total} = total number of genes in the genome, T_{C_p} = total number of targets in clade C_{p}, IN = genes belonging to the functional module, OUT = genes not belonging to the functional module, $N^{\mathrm{IN}}_{\text{total}}$ = total number of genes in the functional module, $T^{\mathrm{IN}}_{C_p}$ = total number of genes in the functional module that are targets in clade C_{p}.

Simulating the number of highly conserved targets

To test whether the number of highly conserved targets is explained by the functional selection model, we computed the probability of observing the inferred number of these targets under the functional selection model (the H_1 hypothesis defined in the LRT, above). We then simulated targets in the two subclades that share the same direct ancestral clade, and computed the overlaps between these simulated target sets. The simulations were initialized with the target set of the ancestral clade, and we simulated the targets in each subclade according to its probability of gain or loss of target genes, within and outside of the functional module (computed as described above for the LRT, using a maximum-likelihood estimator). We repeated this process 1000 times, counting the number of times in which the number of simulated ancestral targets was equal to or greater than the number observed in our data. We ran the test on motifs conserved at least up to clade H (29 motifs), and computed these empirical probabilities of the intersection between the target sets at clade D and clade G. In general, for two target sets C_1 and C_2 of clades that share an immediate ancestral clade, we computed the empirical probability for the number of genes in the intersection between the two target sets, denoted as I_p:

$$P(I \geqslant I_p) = \frac{\#\,(I_{\text{simulated}} \geqslant I_p)}{\#\,\text{simulations}}$$

Where $\#(I_{\text{simulated}} \geqslant I_p)$ is the number of simulations where $C_1 \cap C_2 \geqslant I_p$ and #simulations is the total number of simulations (1000).

Experimental transcription factor binding data

We used Ste12 and Tec1 binding in three closely related *Saccharomyces* species by ChIP-chip (Borneman *et al*, 2007); Mcm1 binding measured across three more distant yeast species by ChIP-chip (Tuch *et al*, 2008); and HNF4α binding measured across three mammalian species (human, mouse, and dog) by ChIP-seq (Schmidt *et al*, 2010). For the yeast studies, we used target genes defined in the original manuscripts. For the mammals, where regulatory elements can reside far from their target genes, we had to assign each bound regulatory element with the gene(s) it controls. We focused on binding events in the proximity of the gene, and used five alternative definitions of promoters, ranging between 1 and 5 kp upstream of the transcription start site (sequences taken from UCSC genome Browser versions hg19 (Lander *et al*, 2001), canFam2 (Lindblad-Toh *et al*, 2011), mm10 (Waterston *et al*, 2002)). The specific list of target genes changes when we modify this parameter, but the fit to the functional turnover model does not.

To find functional modules we conducted the same analysis as described above, using the target gene enrichments from the individual species instead of the targets at the clades. For the LRT and simulation tests, we conducted the same analysis as described above, but comparing targets of each individual species to the ancestral target set of all three species, defining the highly conserved targets as targets conserved in all three species. For mammals, we used gene functional annotations from MsigDB (Liberzon *et al*, 2011) (Release 3.0).

Gene, promoter annotations, and DNA motifs

We acquired the genome sequences and annotations of the 23 *Ascomycota* species from the online Fungal Orthogroups (Cherry *et al*, 1997; Wood *et al*, 2002; Cliften *et al*, 2003; Kellis *et al*, 2003, 2004; Dietrich *et al*, 2004; Dujon *et al*, 2004; Galagan *et al*, 2005; Arnaud *et al*, 2007; Butler *et al*, 2009; Rhind *et al*, 2011). Promoters were defined as the 600 bases upstream from the first codon, truncated at the neighboring coding sequence. To avoid bias in the motif discovery stage, we filter out stretches of poly-A or poly-T sequences of five bases or longer and poly-A/T sequences longer than nine bases and replaced them with poly-N of the same length.

Motifs were assembled from TRANSFAC (Matys *et al*, 2006), protein microarrays (Zhu *et al*, 2009), and previous analysis of ChIP-chip data (MacIsaac *et al*, 2006). All motifs were transformed to a PWM format (a $n \times 4$ matrix, where each i,j cell contains the count of nucleotide j in position i of the motif), and clustered (using BLiC Habib *et al*, 2008) to unite highly identical motifs.

Species phylogeny

The CladeoScope algorithm as well as the maximum-likelihood estimators described above, assume the species phylogeny is known. Thus, to reconstruct the phylogenetic relationship between the species, we first identified all the orthologous genes with exactly one copy in each of the species (Wapinski *et al*, 2007) and aligned their orthologous protein sequences. We concatenated all of these alignments to produce a meta-alignment of over 300 000 positions. We sampled 10 000 residues from this alignment, giving us an artificial protein from which we reconstructed the phylogeny using the PhyML (Guindon *et al*, 2010) software package with its default parameter settings. We repeated this process 10 times, rendering the same phylogeny at all branches except for the post-WGD clade of species, in which *Candida glabrata* and *S. castellii* were sometimes found to be inverted. Recent work (Scannell *et al*, 2006) has shown that it requires fewer genomic rearrangements to place *S. castellii* as the outgroup of this clade and that the longer branch length leading to *C. glabrata* may be due to increased selective pressure as it became a pathogenic species. Thus, we fixed the branches at this location of the tree. In order to re-estimate the branch lengths with this fixed tree topology, we repeated the same process to construct an artificial protein and ran the SEMPHY software package (Friedman *et al*, 2002) to optimize branch lengths with default parameters. We repeated this process 10 times and found branch length correlations of over 0.99 between replicates. We then averaged the branch lengths among the 10 replicates to obtain branch length estimates for the given species phylogeny.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We thank Ido Amit, Ernest Fraenkel, Tommy Kaplan, Eric Lander, Oliver Rando, Or Zuk, and the Margalit and Friedman laboratory members for helpful discussions and comments. We thank Leslie Gaffney for assistance with artwork. NH was supported by the Maydan fellowship, IW is the HHMI fellow at the Damon Runyon Cancer Research Foundation. This work was supported by HHMI, an NIH PIONEER award, an NIH R01 2R01CA119176-01, a Burroughs-Wellcome Fund Career Award at the Scientific Interface, and the Sloan Foundation (to AR), by the US-Israel Binational Science Foundation (BSF) award (to NF and AR), and by two Israel Science Foundation grants (to HM and NF) and European Research Council grant (to NF).

Author contributions: NH, IW, HM, AR, and NF conceived and designed the study; NH and IW developed and applied the CladeoScope framework; NH developed the models and performed the analyses; NH, IW, HM, AR, and NF wrote the paper.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Alon U (2007) Network motifs: theory and experimental approaches. *Nat Rev Genet* **8:** 450–461

Arnaud MB, Costanzo MC, Skrzypek MS, Shah P, Binkley G, Lane C, Miyasato SR, Sherlock G (2007) Sequence resources at the Candida Genome Database. *Nucleic Acids Res* **35:** D452–D456

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25–29

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

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M (2007) Divergence of transcription factor binding sites across related yeast species. *Science* **317:** 815–819

Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8:** e1000343

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Agrafioti I, Arnaud MB, Bates S, Brown AJ, Brunke S, Costanzo MC, Fitzpatrick DA, de Groot PW, Harris D, Hoyer LL *et al* (2009) Evolution of pathogenicity and sexual reproduction in eight Candida genomes. *Nature* **459:** 657–662

Capaldi AP, Kaplan T, Liu Y, Habib N, Regev A, Friedman N, O'Shea EK (2008) Structure and function of a transcriptional network activated by the MAPK Hog1. *Nat Genet* **40:** 1300–1306

Chen D, Toone WM, Mata J, Lyne R, Burns G, Kivinen K, Brazma A, Jones N, Bahler J (2003) Global transcriptional responses of fission yeast to environmental stress. *Mol Biol Cell* **14:** 214–229

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D (1997) Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387:** 67–73

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301:** 71–76

de Lichtenberg U, Jensen TS, Brunak S, Bork P, Jensen LJ (2007) Evolution of cell cycle control: same molecular machines, different regulation. *Cell Cycle* **6:** 1819–1825

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pöhlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P (2004) The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. *Science* **304:** 304–307

Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* **3:** e99

Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich JM, Beyne E *et al* (2004) Genome evolution in yeasts. *Nature* **430:** 35–44

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17:** 368–376

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specified tree topology. *Syst Zool* **20:** 406–416

Friedman N, Ninio M, Pe'er I, Pupko T (2002) A structural EM algorithm for phylogenetic inference. *J Comput Biol* **9:** 331–353

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Baştürkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S *et al* (2005) Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* **438:** 1105–1115

Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol* **2:** e398

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59:** 307–321

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

Hannenhalli S (2008) Eukaryotic transcription factor binding sites—modeling and integrative search methods. *Bioinformatics* **24:** 1325–1331

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104

Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, Nantel A, Whiteway M (2008) Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell* **29:** 552–562

Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J *et al* (2000) Functional discovery via a compendium of expression profiles. *Cell* **102:** 109–126

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34:** D354–D357

Kellis M, Birren BW, Lander ES (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428:** 617–624

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

Khaitovich P, Enard W, Lachmann M, Paabo S (2006) Evolution of primate gene expression. *Nat Rev Genet* **7:** 693–702

King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* **188:** 107–116

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K *et al* (2001) Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921

Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M (2010) Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol* **8:** e1000329

Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27:** 1739–1740

Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J *et al* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478:** 476–482

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics* **7:** 113

Mannhaupt G, Feldmann H (2007) Genomic evolution of the proteasome system among hemiascomycetous yeasts. *J Mol Evol* **65:** 529–540

Martchenko M, Levitin A, Whiteway M (2007) Transcriptional activation domains of the Candida albicans Gcn4p and Gal4p homologs. *Eukaryot Cell* **6:** 291–301

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34:** D108–D110

Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, Spannagl M, Mayer KF, Stümpflen V, Antonov A (2011) MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res* **39:** D220–D224

Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, Biggin MD, Eisen MB (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* **2:** e130

Nehlin JO, Ronne H (1990) Yeast MIG1 repressor is related to the mammalian early growth response and Wilms' tumour finger proteins. *EMBO J* **9:** 2891–2898

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39:** 730–732

Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci USA* **104**(Suppl 1)**:** 8605–8612

Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, Habib N, Wapinski I, Roy S, Lin MF, Heiman DI, Young SK, Furuya K, Guo Y, Pidoux A, Chen HM, Robbertse B, Goldberg JM, Aoki K, Bayne EH, Berlin AM (2011) Comparative functional genomics of the fission yeasts. *Science* **332:** 930–936

Scannell DR, Byrne KP, Gordon JL, Wong S, Wolfe KH (2006) Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* **440:** 341–345

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, Talianidis I, Flicek P, Odom DT (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328:** 1036–1040

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* **34:** 166–176

Tanay A (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16:** 962–972

Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci USA* **102:** 7203–7208

Tirosh I, Reikhav S, Levy AA, Barkai N (2009) A yeast hybrid provides insight into the evolution of gene expression regulation. *Science* **324:** 659–662

Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* **443:** 415–420

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

Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449:** 54–61

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P *et al* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562

Weirauch MT, Hughes TR (2010) Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet* **26:** 66–74

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VL, Fisher EM, Tavare S, Odom DT (2008) Species-specific transcription in mice carrying human chromosome 21. *Science* **322:** 434–438

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430:** 85–88

Wittkopp PJ, Haerum BK, Clark AG (2008) Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* **40:** 346–350

Wohlbach DJ, Thompson DA, Gasch AP, Regev A (2009) From elements to modules: regulatory evolution in Ascomycota fungi. *Curr Opin Genet Dev* **19:** 571–578

Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A *et al* (2002) The genome sequence of Schizosaccharomyces pombe. *Nature* **415:** 871–880

Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, Philippakis AA, Hu Y, De Masi F, Pacek M, Rolfs A, Murthy T, Labaer J, Bulyk ML (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19:** 556–566