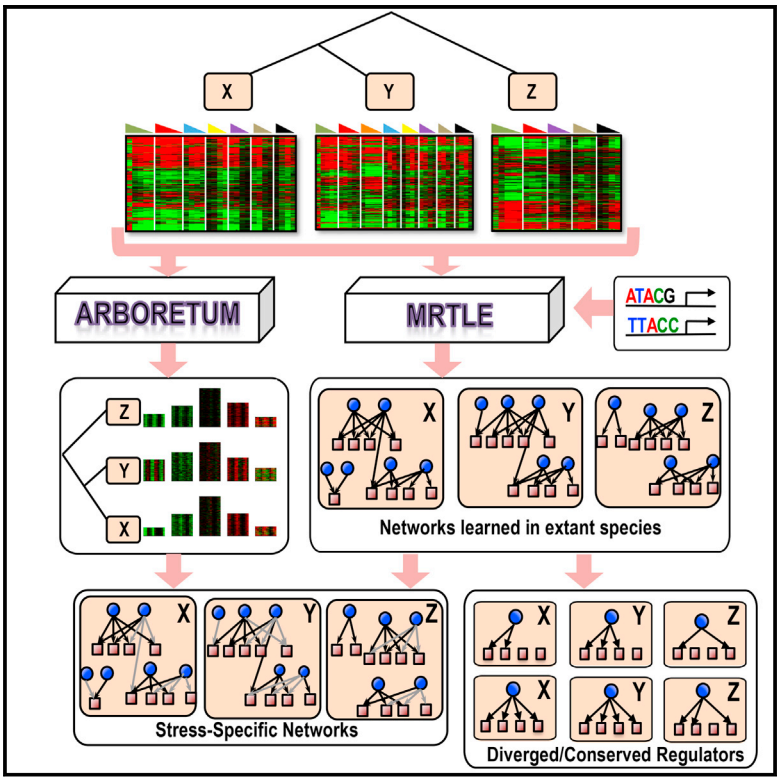
Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies

Article

Cell Systems

Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies

Graphical Abstract



Authors

Christopher Koch, Jay Konieczka, Toni Delorey, ..., Erin K. O'Shea, Aviv Regev, Sushmita Roy

Correspondence

sroy@biostat.wisc.edu

In Brief

A new computational method for genome-scale regulatory network inference that uses phylogenetic structure, sequence-specific motifs, and transcriptomes from diverse species. Here, it is used to study the evolution of stress-specific regulatory networks in six ascomycete yeasts.

Highlights

- Integrating phylogeny, motifs, and expression improves regulatory network inference
- A phylogenetic framework allows genome-scale network inference in non-model species
- Comparative analyses of predicted networks identifies properties of network evolution
- Stress response networks of non-model organisms are inferred and validated

Koch et al., 2017, Cell Systems *4*, 543–558
May 24, 2017
http://dx.doi.org/10.1016/j.cels.2017.04.010

Cell Systems
Article

Inference and Evolutionary Analysis of Genome-Scale Regulatory Networks in Large Phylogenies

Christopher Koch,[1,13] Jay Konieczka,[2,13] Toni Delorey,[2] Ana Lyons,[3] Amanda Socha,[4] Kathleen Davis,[5] Sara A. Knaack,[6] Dawn Thompson,[2] Erin K. O'Shea,[7,8,9,10] Aviv Regev,[2,11] and Sushmita Roy[6,12,14,*]
[1]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[4]Dartmouth College, Biology Department, Hanover, NH 03755, USA
[5]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA
[6]Wisconsin Institute for Discovery, 330 North Orchard Street, Madison, WI 53715, USA
[7]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
[8]Howard Hughes Medical Institute
[9]Faculty of Arts and Sciences Center for Systems Biology
[10]Department of Molecular and Cellular Biology
Harvard University, Northwest Laboratory, Cambridge, MA 02138, USA
[11]Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA
[12]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI 53792, USA
[13]These authors contributed equally
[14]Lead Contact
*Correspondence: sroy@biostat.wisc.edu
http://dx.doi.org/10.1016/j.cels.2017.04.010

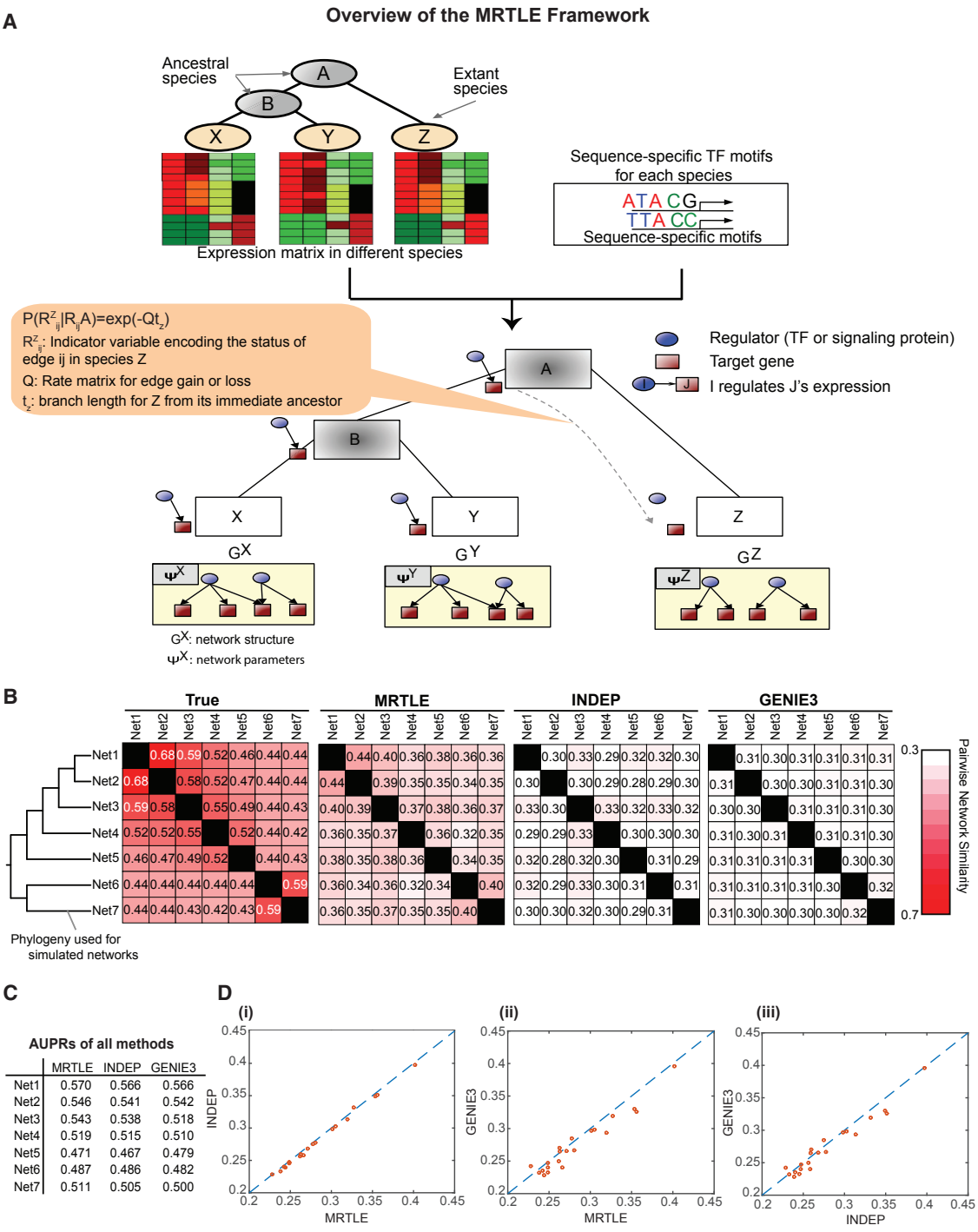
SUMMARY

Changes in transcriptional regulatory networks can significantly contribute to species evolution and adaptation. However, identification of genome-scale regulatory networks is an open challenge, especially in non-model organisms. Here, we introduce multi-species regulatory network learning (MRTLE), a computational approach that uses phylogenetic structure, sequence-specific motifs, and transcriptomic data, to infer the regulatory networks in different species. Using simulated data from known networks and transcriptomic data from six divergent yeasts, we demonstrate that MRTLE predicts networks with greater accuracy than existing methods because it incorporates phylogenetic information. We used MRTLE to infer the structure of the transcriptional networks that control the osmotic stress responses of divergent, non-model yeast species and then validated our predictions experimentally. Interrogating these networks reveals that gene duplication promotes network divergence across evolution. Taken together, our approach facilitates study of regulatory network evolutionary dynamics across multiple poorly studied species.

INTRODUCTION

Transcriptional regulatory networks are key components of cellular information processing and transmit upstream signals to affect downstream context-specific expression patterns. Such networks are defined by connections of regulators such as transcription factors and signaling proteins to target genes (Kim et al., 2009). Changes in transcriptional regulatory networks have been repeatedly shown to contribute to phenotypic diversity of organisms (King and Wilson, 1975; Romero et al., 2012; Carroll, 2000; Wittkopp, 2007). However, our understanding of how regulatory networks evolve and affect complex phenotypes has been limited to a handful of transcription factors in a few species (Borneman et al., 2007; Tuch et al., 2008; Schmidt et al., 2010; Odom et al., 2007). An improved understanding of regulatory network evolution requires a systematic framework for both mapping global regulatory networks in multiple species as well as comparing the networks across species.

While significant effort has been invested in identifying regulatory networks in individual model organisms such as *Saccharomyces cerevisiae* (Hughes and de Boer, 2013; Harbison et al., 2004; Macisaac et al., 2006) and *Escherichia coli* (Faith et al., 2007), an open challenge is to identify these networks in newly sequenced species and compare networks across species. Recently, several comparative functional genomic studies have measured genome-wide mRNA levels in multiple species (Brawand et al., 2011, 2014; Thompson et al., 2013). These quantitative datasets serve as "readouts" of the network state and provide the opportunity to comprehensively study how regulatory networks convert environmental signals into species-specific phenotypes and change globally across species. However, there are two major challenges that need to be overcome. First, most successful network reconstructions have used hundreds of samples, whereas the available data for each species in a comparative study is restricted to a few dozen samples. Second, to understand the role of regulatory network evolution on species evolution, regulatory networks need to be inferred for a complex phylogeny consisting of a sufficiently large number of species.



	MRTLE	INDEP	GENIE3
Net1	0.570	0.566	0.566
Net2	0.546	0.541	0.542
Net3	0.543	0.538	0.518
Net4	0.519	0.515	0.510
Net5	0.471	0.467	0.479
Net6	0.487	0.486	0.482
Net7	0.511	0.505	0.500

Figure 1. Overview of the MRTLE Learning Algorithm and Results on Simulated Data

(A) The MRTLE algorithm takes as input a phylogenetic tree relating multiple extant species, expression data for each extant species, and optionally sequence-specific transcription factor binding motifs for each species. MRTLE uses the phylogenetic tree and motif instances as prior knowledge and outputs multiple regulatory networks, one for each species. Each regulatory network specifies the directed connections among regulatory proteins such as transcription factors (blue filled circles) to target genes (red filled squares). To capture the evolutionary dynamics of regulatory edge gain and loss, MRTLE uses a phylogenetic prior that is parameterized by a continuous-time Markov chain. Each branch on the tree can have different gain and loss rates depending upon the branch length (e.g., t_z for species Z) and an overall gain and loss rate of regulatory connections specified in the rate matrix Q. R_{ij}^Z denotes the state of the edge between regulator i and target gene j in species Z.

(B) Pairwise similarities measured by F-score for the simulated ground truth (True) set of seven networks, Net1-Net7, and the inferred sets of networks using two baseline methods that do not incorporate any phylogenetic information (INDEP, GENIE3), and MRTLE that uses the phylogenetic tree of the considered species during network inference.

(legend continued on next page)

Incorporating the phylogenetic structure enables us to account for the inherent relatedness of species based on their DNA sequence composition, to trace the evolution of individual regulatory connections (edges) at different points on the phylogeny, and to compare the relative contribution of sequence and network divergence to phenotypic divergence. A large phylogeny is important to be able to systematically observe patterns of conservation and divergence and to study different factors such as gene duplication that can contribute to regulatory network divergence. Existing approaches to infer regulatory networks for multiple species have either not attempted to explicitly model the phylogeny of the species involved (Penfold et al., 2015; Joshi et al., 2014) or their applications have been restricted to two or three species (Xie et al., 2011; Penfold et al., 2015). Extending such approaches to infer genome-scale networks for a large phylogeny with complex orthologies can be computationally expensive. While a number of studies have compared gene expression profiles across multiple species (Bergmann et al., 2003; Ihmels et al., 2005; Kristiansson et al., 2013; Roy et al., 2013b), these approaches typically identify gene modules that are conserved or diverged across species and do not provide fine-grained regulatory network connectivity information. Such information is critical to identify specific regulatory connections that evolution must have made and broken as the species diverged.

In this paper, we develop a probabilistic graphical model-based method, multi-species regulatory network learning (MRTLE), that uses a phylogenetic framework to infer regulatory networks in multiple species simultaneously. In MRTLE, the regulatory network of each species is modeled as a probabilistic graphical model (Friedman, 2004), and the phylogenetic information is incorporated by specifying a prior probability distribution over edge gain and loss from the ancestral to extant species. We use the ascomycete yeasts as a model system to study the evolution of regulatory networks and validate MRTLE using simulations, available reconstructions of the yeast *S. cerevisiae* network, and available chromatin immunoprecipitation (ChIP)-chip-based transcription factor (TF) datasets in other species. MRTLE reconstructs networks better than approaches that do not incorporate phylogenetic information, while also inferring networks that diverge in a manner consistent with the phylogeny of the species involved. We use our inferred networks to identify regulators with evolutionarily conserved roles in stress-related repression and induction across ascomycete yeasts. In total, our computational framework of simultaneously inferring regulatory networks for multiple species and assessing regulatory network divergence enables a systematic study of the evolution of gene regulatory networks in a complex phylogeny.

RESULTS

Inference and Analysis of Regulatory Networks in Multiple Species Using MRTLE

We developed a multi-species network inference algorithm called MRTLE that imposes a phylogenetically motivated prior distribution on a set of graphs, each graph describing the regulatory network of a species (Figure 1A; STAR Methods). The prior distribution encodes the belief that regulatory networks diverge according to the phylogeny, that is, the regulatory networks of species that are phylogenetically closer are likely to be more similar. This probability is in turn described over individual edge states (R^z_{ij} in Figure 1A), for a given species Z as a function of its state in A, the immediate ancestor of Z, $P(R^z_{ij}|R^A_{ij})$. This is modeled as a continuous-time Markov process parameterized by the rate matrix Q, which specifies the rates at which we expect regulators to gain or lose targets per unit time, and the branch length t_z, which specifies the divergence time between species Z and its immediate ancestor A. This parameterization allows for the probability of edge gain and loss to be branch specific (Hobolth and Jensen, 2005; Garber et al., 2009; Habib et al., 2012). Each regulatory network is modeled by a dependency network, a special type of a probabilistic graphical model (PGM) (Friedman, 2004). A PGM has two components: the graph structure (Figure 1A, G_X, G_Y, G_Z) and parametric functions (Figure 1A, ψ_X, ψ_Y, ψ_Z). The nodes in the graph correspond to random variables and encode the expression levels of a gene. The graph structure specifies the regulators of each gene, while the parameters of the graph specify how the regulator levels determine the output expression level.

MRTLE takes as input expression data from k different species, a phylogenetic tree with branch lengths, gene orthology relationships including those arising from gene duplications, and rate parameters for regulatory edge loss and gain (STAR Methods; Figure 1A). The output of MRTLE is k networks, one for each species. The prior is flexible and can integrate species-specific regulatory information such as sequence-specific motifs. The prior probability of a regulatory interaction between a target gene and a regulator depends upon both per-species prior regulatory information (e.g., presence of sequence-specific motifs if available) and the phylogenetic prior (STAR Methods).

Since the majority of real regulatory network connections remain undiscovered, especially in non-model organisms, we first used simulations to assess our approach. The goal of the simulation is to ask if the observed expression data from multiple species are generated from phylogenetically divergent networks, does a method such as MRTLE perform better than other methods? In our simulation, regulatory networks for seven extant species were evolved from an ancestral network using a phylogenetic tree (Figure 1B), followed by generation of simulated expression data at the extant species (STAR Methods). We compared MRTLE with two baseline approaches, INDEP and GENIE3 (Huynh-Thu et al., 2010), that performed network inference in each species independently (STAR Methods). INDEP is similar to MRTLE except it did not use a phylogenetic prior. GENIE3 was shown to have state-of-the-art performance in network inference problems (Huynh-Thu et al., 2010; Marbach et al., 2012). Three criteria were used for evaluation: (1) do the inferred regulatory networks exhibit phylogenetic patterns of conservation that are similar to the true regulatory networks, (2) how

(C) Area under the precision-recall curve (AUPR) values comparing networks inferred by MRTLE, INDEP, and GENIE3 with the seven simulated ground truth networks. The greater the AUPR the better the method.
(D) Comparison of AUPR between (i) MRTLE and INDEP, (ii) MRTLE and GENIE3, (iii) INDEP and GENIE3, when considering only true and predicted conserved edges between species pairs.

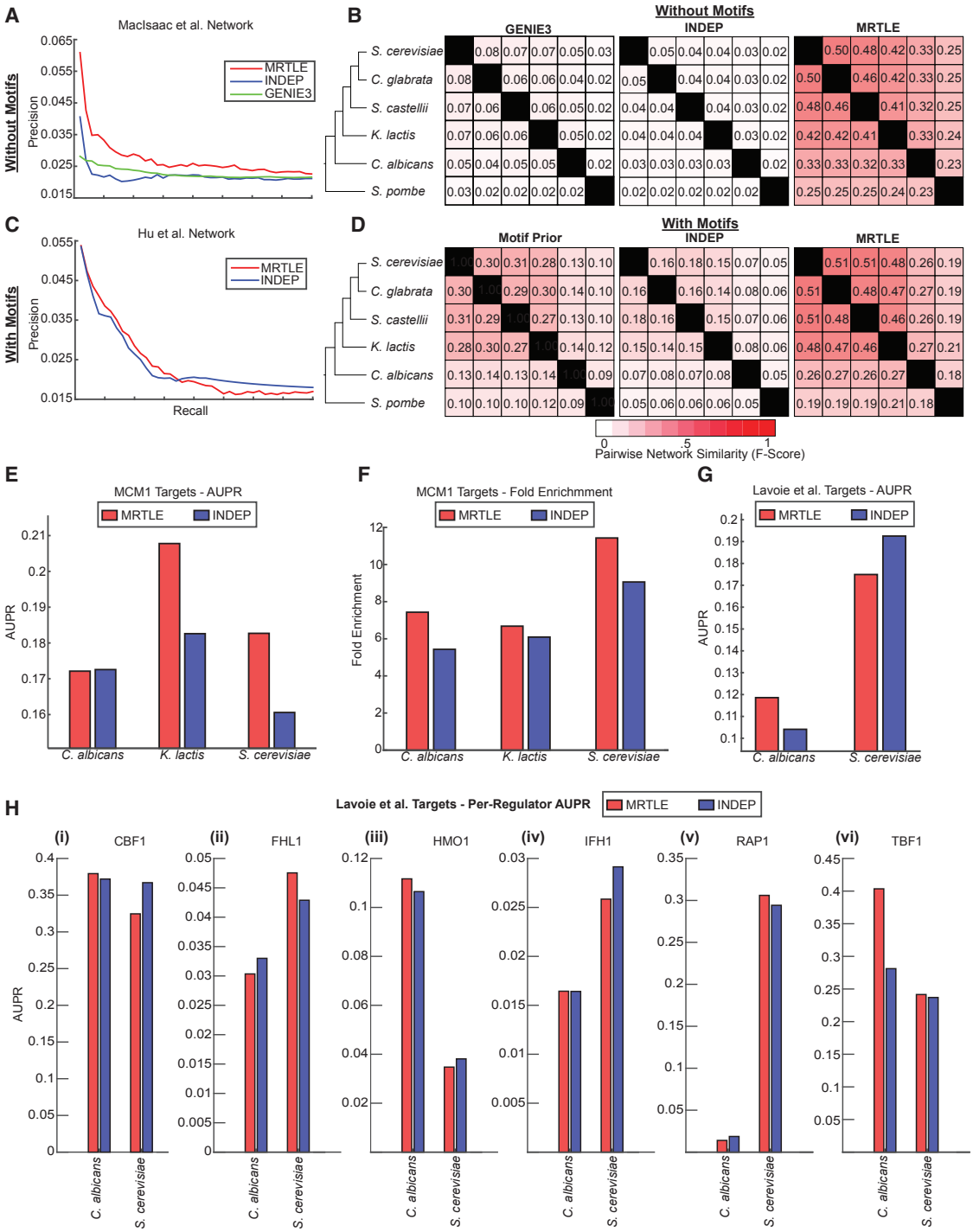


Figure 2. Assessing Inferred Networks on the Ascomycete Yeast Phylogeny

(A) Precision-recall curves for MRTLE, INDEP, and GENIE3 without motifs assessing the agreement of inferred networks to an *S. cerevisiae* gold standard network derived from ChIP-chip experiments.

(B) Pairwise similarities measured by F-score for the networks inferred by GENIE3, INDEP, and MRTLE when motifs were withheld, for six yeast species.

(C) Precision-recall curves for MRTLE and INDEP when motifs were included, assessing the agreement of the inferred networks using an *S. cerevisiae* transcription factor (TF) knockout network from Hu et al. (2007) as the gold standard.

(D) Pairwise similarities measured by F-score for the networks inferred by INDEP and MRTLE and the prior motif network for six yeast species.

(E) AUPR values assessing MRTLE and INDEP at recovering ChIP-chip targets of the TF, MCM1, in three species, *C. albicans*, *K. lactis*, and *S. cerevisiae*.

(F) Fold enrichment of MCM1 ChIP-chip targets in MCM1's inferred target set by MRTLE or INDEP in the 30,000 most confident edges from each method.

(legend continued on next page)

well do the methods recover edges from the ground truth network, and (3) how well do the methods recover those edges that are conserved.

For (1), we computed the F-score-based similarity (STAR Methods) for each pair of species' true networks, and compared this with the F-score for all pairs of inferred networks. Inclusion of the phylogenetic prior greatly aids in recovering a pattern of network similarity that agrees with the true pattern of conservation and divergence (Figure 1B). For example, when using MRTLE, inferred networks Net1 and Net2 are more similar to each other (F-score, 0.44), than Net1 is to Net7 (F-score, 0.36). Similarly, Net6 and Net7 are more similar to each other (F-score, 0.40) than they are to any of the other species. This is in agreement with the observed trend in the ground truth networks. In contrast, both INDEP and GENIE3 substantially underestimated the similarity between all pairs of networks, and their inferred networks did not exhibit a strong phylogenetic pattern of conservation, but rather appeared uniformly similar to each other. For (2), we used edge precision and recall curves and the area under the precision-recall curve (AUPR). Overall, MRTLE outperforms both INDEP and GENIE3 (Figures 1C and S1), achieving a higher AUPR than GENIE3 in six of the seven networks and a higher AUPR than INDEP on all seven networks. Although the differences in AUPR are small, they are significant when comparing MRTLE against the other two methods (t test, $p < 0.05$). GENIE3 and INDEP are comparable in performance with no significant difference in performance, with INDEP tending to have higher AUPRs than GENIE3. For (3), we considered only true conserved edges between pairs of species and again assessed the methods' accuracies in terms of AUPR. We found that MRTLE is generally better at recovering edges that are evolutionarily conserved compared with INDEP (Figure 1D, i) and GENIE3 (Figure 1D, ii). Furthermore, INDEP was better than GENIE3 at recovering true conserved edges (Figure 1D, iii).

Our simulation results show that if the observed data are generated from networks that share an evolutionary history, an approach such as MRTLE that uses phylogenetic information can more effectively learn regulatory networks across multiple species. Having established the utility of MRTLE on simulated datasets, we next compared MRTLE, GENIE3, and INDEP for inferring regulatory networks from real expression data from six yeast species (STAR Methods): *Saccharomyces cerevisiae, Candida glabrata, Saccharomyces castellii, Candida albicans, Kluyveromyces lactis,* and *Schizosaccharomyces pombe*. These datasets measure genome-wide transcriptome states in different stress conditions: glucose depletion, heat shock, oxidative stress, and osmotic stress. Glucose depletion, heat shock, and oxidative stress datasets were previously published (Thompson et al., 2013; Roy et al., 2013b; Wapinski et al., 2007), while osmotic stress was generated as part of this study. As potential regulators, we included ~500 genes that have known DNA-binding roles in *S. cerevisiae*, as well as genes whose protein products are known to bind RNA (Table S1). We used the species tree branch lengths and gain and loss rate parameters inferred by Habib et al. (2012) to specify the probabilities of edge loss and gain in MRTLE. We first assessed all three network inference methods without making use of sequence-specific motif priors. This enabled us to compare against GENIE3, which does not incorporate priors, and also to assess the broader, future applicability of MRTLE to species phylogenies for which such information may not be available. To evaluate the inferred networks, we used criteria similar to the simulation setting. To compute precision-recall curves, we used a ChIP-chip-based regulatory network in *S. cerevisiae*, which has been a gold standard in the field (Macisaac et al. (2006)). MRTLE outperforms INDEP and GENIE3, achieving higher precision at the same recall (Figure 2A). When comparing the phylogenetic pattern of conservation, we observe that MRTLE-inferred networks diverge in a pattern consistent with the phylogeny (Figure 2B). In contrast, the networks inferred by INDEP and GENIE3 display extreme divergence. Furthermore, the extent of conservation in MRTLE networks is more consistent with observed conservation of ChIP-chip-based binding profiles (Tuch et al., 2008) than either INDEP or GENIE3 (Figure S2). Overall, these results suggest that using phylogenetic information as prior can enable a more accurate reconstruction of a regulatory network, and the absence of a phylogenetic prior leads to an overestimation of the divergence in the species' networks. Since GENIE3 did not have a significantly different performance than INDEP and does not incorporate sequence-specific motifs, our subsequent results include only INDEP as the baseline.

Having established that MRTLE is able to outperform methods that do not use phylogenetic priors (e.g., INDEP and GENIE3) when neither method has access to sequence-specific motifs, we next evaluated MRTLE when given valuable sequence-based regulatory information. We used species-specific motifs from Habib et al. (2012) as additional priors on the graphs. We could not use the gold standard network of Macisaac et al. (2006), because it used evolutionarily conservation as an additional filter to define TF target edges, and our motif priors were also defined using an evolutionary signature (Habib et al., 2012). As an alternative gold standard, we used an *S. cerevisiae* regulatory network from Hu et al. (2007) obtained by systematically deleting regulators and analyzing the downstream effects on expression (Hu et al., 2007) (STAR Methods). Using this gold standard, we found MRTLE to outperform INDEP in edge recovery (Figure 2C). Notably, MRTLE outperforms INDEP at low recall (high precision) thresholds, suggesting that those regulatory edges supported by expression, evolutionary conservation, and a motif instance are more likely to be functional than those supported only by expression and a motif instance.

Next, we examined the networks inferred by MRTLE and INDEP to assess whether they diverge in a manner consistent with the phylogeny. Since the degree and pattern of network similarity is dependent upon the similarity in the motif networks used as priors in addition to the expression data, we also

(G) AUPR values assessing MRTLE and INDEP at recovering edges of regulatory networks consisting of ChIP-chip targets of six different TFs in *C. albicans* and *S. cerevisiae*.

(H) AUPR values for each TF assessing MRTLE and INDEP at recovering ChIP-chip targets of each of the six different TFs in *C. albicans* and *S. cerevisiae*. The ground truth in (H) is the same as in (G) but presented at the per-TF level.

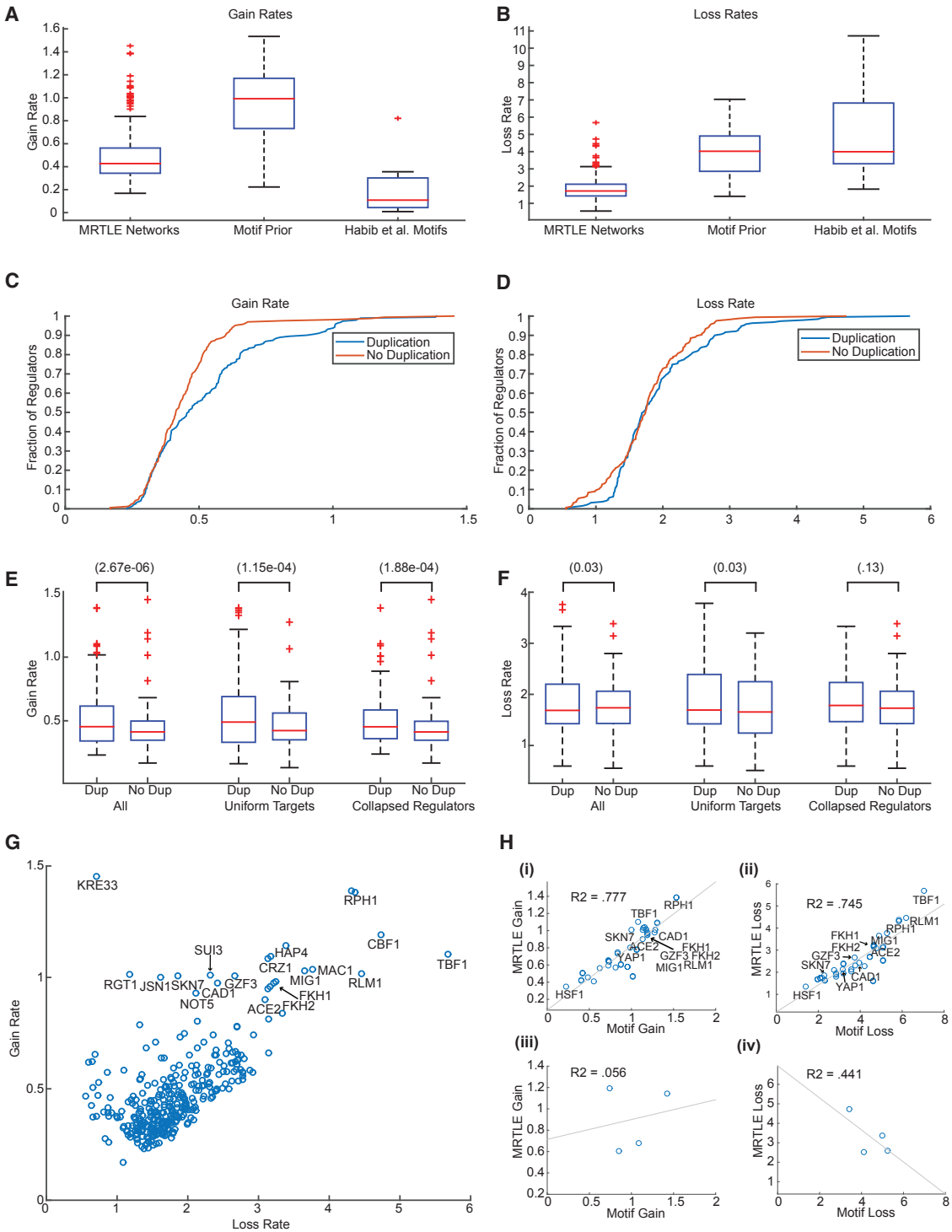


Figure 3. Assessing Rates of Target Gain and Loss for Regulators in MRTLE-Inferred Networks and Motif Networks

(A and B) Boxplots of gain (A) and loss (B) rates calculated for the MRTLE-inferred networks, the rates calculated for the motifs used as prior knowledge in the MRTLE framework, and the rates calculated for the motifs used by Habib et al. (2012). For the MRTLE networks, rates were calculated using the top approximately 50,000 edges in each species' network.

(C and D) Cumulative distribution function plots of gain (C) and loss (D) rates calculated using the MRTLE-inferred networks at confidence thresholds amounting to approximately 50,000 edges for regulators with duplication (blue) and without duplications (red).

(E and F) Boxplots of gain (E) and loss (F) rates for regulators in the MRTLE networks when considering all regulators and all targets (left; All), all regulators and targets without duplications (middle; Uniform Targets), and duplicated regulators collapsed into a single average regulator with all targets (right; Collapsed

(legend continued on next page)

estimated the similarity of all pairs of motif networks (Figure 2D, Motif Prior). The motif prior networks exhibited stronger evolutionary conservation compared with networks learned from INDEP (Figure 2D). As was the case for simulated data (Figure 1B) and for real expression data alone (Figure 2B), the networks learned by MRTLE exhibit stronger evolutionary conservation than those learned by INDEP and diverge in a pattern consistent with the phylogeny (Figure 2D). As in the no motif case, the observed conservation levels for MRTLE with motifs agrees more with previous studies (Figure S2; Tuch et al., 2008). The similarity scores for INDEP networks increased relative to the scores when not using motifs, consistent with the hypothesis that the motif prior constrains the inferred networks to be more conserved than expression alone (Figures 2B and 2D). The similarity scores for the MRTLE networks were comparable with and without motifs, suggesting that MRTLE is robust to the prior inputs.

Although large-scale knockout and ChIP-chip networks are not available in non-model organisms, a handful of TFs have been studied across multiple species (Tuch et al., 2008; Lavoie et al., 2010) using ChIP-chip experiments. In particular, Tuch et al. (2008) measured binding gene targets of the TF, MCM1, in *S. cerevisiae*, *K. lactis*, and *C. albicans*. Lavoie et al. (2010) measured targets of CBF1, HMO1, FHL1, IFH1, and RAP1 in *S. cerevisiae* and *C. albicans*. We used these two ChIP-chip datasets to test the ability of MRTLE and INDEP with motifs to recover these targets. On the MCM1 datasets, MRTLE outperforms INDEP in *K. lactis* and *S. cerevisiae*, and performs comparably in *C. albicans* (Figure 2E). As an additional evaluation measure, we calculated the fold enrichment of the ChIP-chip MCM1 targets in the predicted MCM1 targets among the top ~30,000 edges (Figure 2F; STAR Methods). Although the predicted targets from both methods were enriched for ChIP-chip MCM1 targets, MRTLE achieved a higher fold enrichment than INDEP in all three species. We combined the predicted targets of all TFs studied by Lavoie et al. (2010) into a single network and found MRTLE to significantly outperform INDEP for *C. albicans* (Figure 2G). However, MRTLE was outperformed on *S. cerevisiae* (Figure 2G). To gain insight into the lower performance of MRTLE on this *S. cerevisiae* network, we analyzed our predictions per TF (Figures 2H and S3). In *S. cerevisiae*, MRTLE outperformed INDEP on RAP1 and TBF1, and it was outperformed for CBF1 (Figures 2H and S3). Both methods had low AUPRs on HMO1, IFH1, and FHL1, likely due to the small number of targets. It is likely that CBF1's targets diverge substantially across species giving no additional advantage with MRTLE, or, it is possible that the current CBF1 target set is incomplete. Future experiments combining ChIP-chip experiments with TF knockout are needed to examine this property. Taken together, MRTLE was more effective than INDEP at recovering ChIP-based regulatory edges in non-model organisms, demonstrating that a phylogenetic prior-based framework is beneficial for non-model organisms as well.

The genome-wide regulatory networks for these six species enable us to more systematically study factors driving regulatory network evolution. For example, estimated rates of gain and loss of edges can provide insights into the relative importance of these two types of network changes in regulatory network divergence. Previously, Habib et al. (2012) assessed gain and loss rates of computationally inferred binding sites of individual TFs. Using a similar framework to Habib et al., we computed gain and loss rates of targets for each regulator (TFs and signaling proteins; STAR Methods, Table S2). We find loss rates to be higher (1.84 ± 0.67) than gain rates (0.48 ± 0.20). A similar trend was observed with the rates from Habib et al. (loss rate of 4.91 ± 2.35 and gain rate of 0.17 ± 0.17), as well as in our recalculations of the rates using motif instances only (loss rate 3.92 ± 1.31, gain rate 0.94 ± 0.31). Our results show that regulatory networks evolve by losing edges more rapidly than by gaining edges, and this property is true for both purely sequence-based networks and MRTLE-inferred networks. Although the same trends are observed in all three sources of rates, rates inferred using MRTLE networks were significantly different from the rates inferred from Habib et al. (2012) or the rates obtained in the prior networks. In particular, regulators in the MRTLE network have a relatively lower loss rate (mean 1.84), compared with the loss rate (mean 4.91) estimated by Habib et al. MRTLE gain (Figure 3A) and loss rates (Figure 3B) are also lower than those estimated directly on the motifs used as priors. The significant differences in the rates from Habib et al.'s prior networks and the MRTLE-inferred networks, suggest that the MRTLE-inferred networks represent the output of integrating expression and sequence-specific motifs.

Duplication of TFs can significantly contribute to regulatory network divergence (Pougach et al., 2014; Voordeckers et al., 2015). We next asked if regulators with duplications differ in their rates of gain and loss compared with regulators without duplications. We find that regulators with duplications have significantly higher edge gain rates (Kolmogorov-Smirnov [KS] test, $p < 1 \times 10^{-6}$, Figure 3C) compared with regulators without duplications. Such regulators also tend to lose edges more than those without duplications, but the trend is less pronounced (KS test, $p < 0.04$, Figure 3D). We repeated the rate calculations using targets with uniform orthology, and collapsing duplicated regulators into a single orthogroup by taking the average rate, and found similar results (Figures 3E and 3F, STAR Methods). In addition, we calculated the rates at various confidence thresholds and found the results to be robust to the threshold used (Figure S4).

We identified 19 regulators that had a significantly higher rate of edge gain (>2 SD from mean; Table S3). These regulators were associated with diverse processes including stress response

Regulators). Rates are computed using the top 50,000 edge set. p values from KS tests are given in parentheses, testing the hypothesis that regulators with duplications have higher gain (E) or loss (F) rates than regulators without duplications.

(G) Each point represents a regulator, with the x coordinate specifying the regulator's loss rate and the y coordinate specifying its gain rate. Outlier regulators with high gain rates (2 SD above the mean) are noted.

(H) Comparison of MRTLE and motif prior rates of target gain (i, iii) and loss (ii, iv) for each regulator and its targets, including only those regulators from orthogroups with at least one duplication (i, ii) and from orthogroups without duplications (iii, iv). Each point represents a specific regulator, with the x coordinate specifying the gain/loss rate of the regulator's motif-based targets and the y coordinate specifying the gain/loss rate of the regulator's MRTLE-based targets from the top 50,000 edge set.

(SKN7, CRZ1, CAD1, RLM1), response to nutrients (MIG1, GZF3, CBF1, HAP4), cell cycle (FKH1, FKH2, ACE2), RNA binding (SUI3, JSN1, NOT5), and chromatin organization (CBF1, FKH1, FKH2, RPH1, TBF1). Regulators with high gain rates tend to also have high loss rates (Pearson's correlation of 0.66), but this pattern was defied by KRE33, which had one of the slowest loss rates (1.68 SD below mean) despite having the highest gain rate (4.77 SD above mean; Figure 3G). KRE33 is involved in ribosomal biogenesis, a process that has been shown to be inherently tied to species lifestyle in the ascomycete lineage (Thompson et al., 2013), and KRE33 might be an important factor in regulatory divergence in this phylogeny. Although the majority of these regulators were from orthogroups that had a duplication, four of the regulators (CBF1, KRE33, HAP4, SUI3) were from orthogroups that did not have duplications. Such regulators tend to be associated with response to stress and chemical stimuli, suggesting that such processes may be subject to multiple forces of evolutionary turnover, including gene duplication.

Recently, Pougach et al. (2014) showed that sequence affinity of paralogous TFs diverges after duplication, which can influence regulatory network rewiring. To investigate the role of sequence affinity divergence on the overall edge gain rate, we correlated the MRTLE gain and loss rates to the motif gain and loss rates. We found a strong correlation between rates calculated using MRTLE networks or the motif networks (Figure 3H, i, ii) for TFs from duplicated families. This correlation was negative or weak for TFs from families with no duplications (Figure 3H, iii, iv), although we had many fewer TFs that had motifs and came from non-duplicating families. This suggests that sequence divergence can contribute to network divergence of TFs from duplicated gene families. For two of the TF families, we had sequence motifs for both paralogs: SKN7, HSF1 and YAP1, CAD1. The difference in MRTLE gain rates was much greater for the SKN7, HSF1 pair compared with the YAP1, CAD1 pair (Figure 3H, i). Interestingly, SKN7 and HSF1 had very different sequence affinities (Figure S5) compared with YAP1 and CAD1. These results are consistent with published studies of regulatory divergence of individual TFs (Pougach et al., 2014) and offer preliminary evidence that sequence divergence could explain, in part, the greater tendency to gain targets. Taken together, our inferred networks enabled us to quantitatively assess regulatory network evolution and predict regulators that contribute to regulatory network divergence more than others. Such regulators tend to come from regulator families with duplications or are implicated in stress response.

Evolution of the Osmotic Stress Response Regulatory Network

To gain insight into how changes in regulatory networks can affect complex phenotypes, we used MRTLE-inferred regulatory networks to study response to osmotic stress across six Ascomycota species. Response to environmental stress is a major driving force in the evolution of new phenotypic traits (Hiyama et al., 2012; Hoffmann and Willi, 2008), especially in unicellular organisms (Gasch, 2007). Our current understanding of the regulatory network in response to stress is strongly biased to *S. cerevisiae*, and we understand little about its structure and function in other species. To address this gap, we first measured using microarrays, genome-wide gene expression profiles under osmotic stress in six species. We then identified stress-specific transcriptional modules using a multi-species module inference algorithm, Arboretum (Roy et al., 2013b). Application of Arboretum to our osmotic stress response (OSR)-specific expression data identified five modules ranging from the most repressed genes (module 1) to the most induced genes (module 5, Figure 4A). We then inferred OSR-specific networks by filtering the original MRTLE inferred networks to keep only those edges that connected targets and regulators within the same OSR module (STAR Methods). We refer to this approach of inferring context-specific expression networks as MRTLE + Arboretum. To assess the accuracy of our inferred context-specific regulatory network edges, we performed miSeq expression profiling in knockout strains of two regulators, MSN2/4 and SKO1, under osmotic stress (Figures 4B–4D; STAR Methods). MSN2/4 is a general stress response regulator (Gasch et al., 2000), and SKO1 is an OSR-specific regulator. Both of these regulators coordinate with the protein kinase, HOG1, to control OSR in *S. cerevisiae* (Capaldi et al., 2008). We compared our predicted targets against the miSeq data in two ways. First, we asked whether the expression of MRTLE and MRTLE + Arboretum inferred targets of these two TFs was significantly different based on a KS test, from non-targets under osmotic stress (Figures 4B and 4C). Second, we used LIMMA to define targets of these mutants in each species (Figure 4D; STAR Methods) (Smyth et al., 2005). Based on the KS test, both MRTLE and MRTLE + Arboretum targets are significantly repressed in the MSN2/4 knockout in *S. cerevisiae* compared with wild-type, which suggests that our predicted regulatory connections are valid. We did not find significant differences for the knockout of the ortholog of MSN2/4 in the two other species, *C. albicans* and *Schizosaccharomyces pombe*. The lack of significant differences in these species is consistent with previous observations where MSN2/4 does not play a significant role in general stress response (Nicholls et al., 2004; Chen et al., 2008; Sanso et al., 2008). In particular, the *C. albicans* MSN2/4 homologs, MNL1 and MSN4, do not play a role in general stress response (Nicholls et al., 2004). Only MNL1 is required for adaptation to weak acid stress (Ramsdale et al., 2008). For SKO1, we found a significant downregulation of targets in *C. albicans* and a significant, albeit reduced, effect in *S. cerevisiae*.

The LIMMA-based analysis confirmed our observations. At $p < 0.05$, we found 117 MSN2/4 targets in *S. cerevisiae* and 159 SKO1 targets in *C. albicans*. LIMMA identified relatively fewer targets (14) for *S. cerevisiae* SKO1, and therefore we excluded it from this analysis. After removing genes from these sets that were not in the dataset used by MRTLE, we were left with 114 targets of MSN2/4 in *S. cerevisiae* and 149 targets of SKO1 in *C. albicans*. Our MRTLE + Arboretum approach yielded 311 predicted targets of MSN2/4, 31 of which were among the 114 LIMMA targets, representing a 4.2-fold enrichment (hypergeometric test, $p < 1.2 \times 10^{-12}$, Figure 4D). In contrast, the original MRTLE *S. cerevisiae* network predicted 891 MSN2/4 targets, 50 of which overlapped with the LIMMA results, representing a 2.4-fold enrichment ($p < 1^{-10}$). Similarly for *C. albicans* SKO1, MRTLE alone predicted 334 targets, 21 of which overlapped with LIMMA targets (2.3-fold enrichment, $p < 1.7 \times 10^{-4}$). In contrast, 6 of MRTLE + Arboretum's 40 SKO1 targets overlapped with LIMMA resulting in a higher fold enrichment (5.6-fold

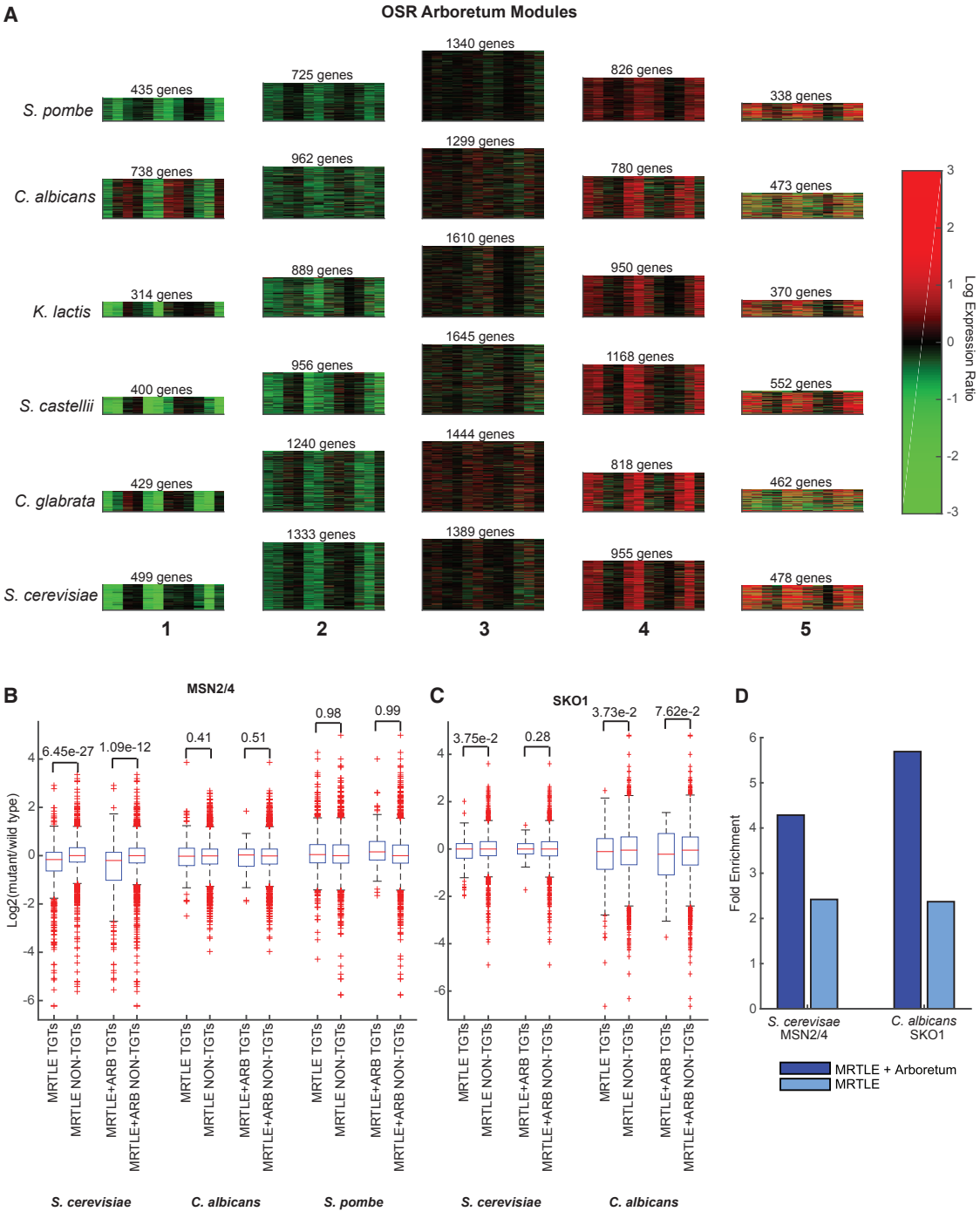
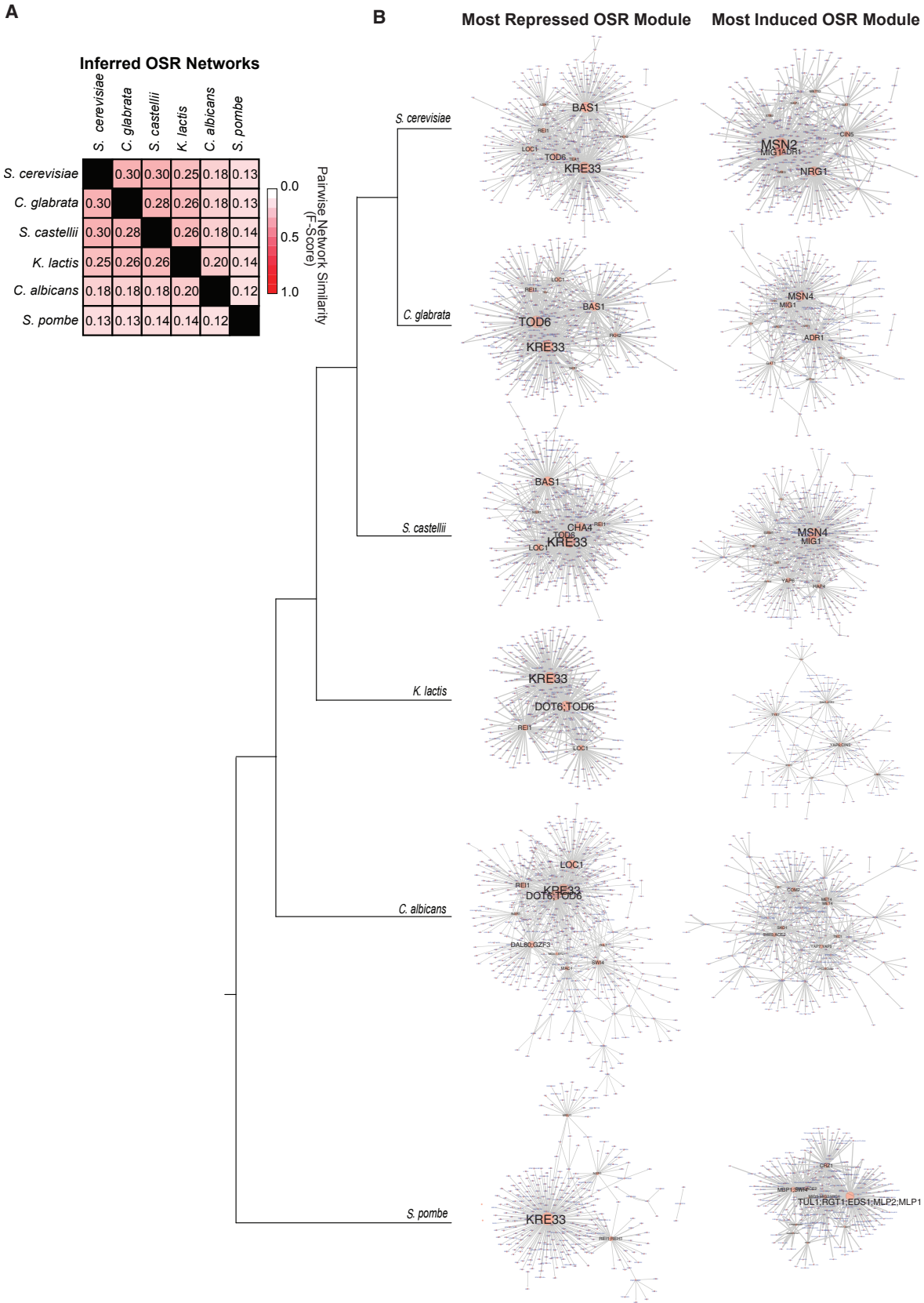


Figure 4. Osmotic Stress Response Module Assessment

(A) Expression heatmaps for each of five inferred OSR-specific expression modules, ranging from most repressed (left) to most induced (right). The height of each heat bar is proportional to the genes in each module.

(B and C) Boxplots comparing differential expression under osmotic stress response (OSR) for predicted targets and non-targets of MSN2/4 (B) and SKO1 (C), as predicted by MRTLE and an approach that combines MRTLE with modular filtering (MRTLE + Arboretum). Targets inferred with the MRTLE + Arboretum approach are those targets inferred by MRTLE with the additional constraint that a target must be present in the same OSR-specific module as its regulator. Each plot shows the log2 ratio of expression in knockout over wild-type of the specified regulator's targets. p Values from KS tests are given for each pair of comparisons, testing the hypothesis that the predicted targets have decreased expression after knockout relative to the non-targets, implying that the knocked out TF has an activating role under salt stress.

(D) Fold enrichment of LIMMA-based targets of MSN2/4 in *S. cerevisiae* and SKO1 in *C. albicans*. Targets were called with LIMMA and fold enrichment.



(legend on next page)

enrichment, $p < 5.8 \times 10^{-4}$). These analyses suggest that the MRTLE + Arboretum approach can greatly improve the accuracy of stress-specific regulatory network learning.

To assess the overall extent of conservation in our complete OSR-specific networks, we calculated the F score similarity between networks of each species pair (Figure 5A). We found a significant phylogenetic pattern, although the extent of conservation was lower than what we observed before (Figure 2D). We then examined the portions of our OSR-specific networks spanning the most repressed and most induced modules, and identified conserved regulators acting as hubs in each case (Figure 5B). In the repressed module, KRE33 remained a conserved hub across all species. BAS1 acted as a repressor in the three most recently diverged species, *S. castellii*, *C. glabrata*, and *S. cerevisiae*, while TOD6 acted as a repressor in all species except *Schizosaccharomyces pombe*, for which no ortholog exists. In the induced modules, we found MSN2/4 as a hub in the most recently diverged species (Table S4). Intriguingly, we found COM2 (MNL1 in *C. albicans*), which belongs to the MSN2/4 family, as a hub in *C. albicans*. In the other species we found the YAP family of TFs and cell-cycle regulators (SWI5, SWI4, MBP1) to act as hubs. In *Schizosaccharomyces pombe* glucose regulators were predicted as the strongest hub followed by the cell-cycle-related regulators. These regulatory networks thus predict several regulators that have not been associated with stress response in these species that can be followed up with future validation studies.

While the structure of the network specifies which regulators regulate which genes, the function of a network specifies how the regulator drives the expression of its targets. A regulator can regulate expression by acting as an activator or repressor of expression. Do regulator roles of activation and repression change across species and to what extent do such changes depend upon the stress? To address these questions, we examined the regulator-module relationships in the OSR and heat shock response (HSR) data (Table S5) (Roy et al., 2013b).

We used two measures to assess a regulator's activating or repressive role. The first measure used the significance of enrichment of a regulator's targets in the activating versus repressive module (STAR Methods). Our second measure compared the expression of the targets for each time point in the repressed or induced module. Our enrichment-based analysis identified several notable regulators with a conserved association with repression in response to osmotic stress, such as KRE33, NSR1, SFP1, LOC1, REH1/REI1, and CHA4/TEA1 (Figure 6A, Table S5). Interestingly, the majority of the conserved, repressed regulators are associated with ribosomal biogenesis, which is repressed in species under stress. Regulators with conserved activating roles across all six species included the MSN2/4 family, the SKN7/HSF1 family, and AFT1/AFT2. Most of these regulators have general or specific stress-related functions. Our second analysis focused on regulators with targets in both activating and repressive modules. This was a complementary measure, which recapitulated regulators from our enrichment-based measure and also identified several additional candidates of regulator divergence (typically in one or two species, Figure 6B). This included cell-cycle regulators such as FKH1/2 and MBP1/SWI4, stress regulators (CRZ1), chromatin remodelers (GIS1, RPH1), and HAP4. Notably, several of these regulators were also associated with higher gain rates, suggesting that regulator expression divergence might be associated with the tendency of the regulators to gain or lose edges. However, additional datasets would be needed to more fully understand this phenomenon. Overall, regulators tended to not change signs between species from activating to repressive or vice versa.

To examine the generality of this observation, we compared the OSR regulator signs with those in HSR (Figure 7). The majority of the regulators had similar associations in these stresses, with stress-related regulators such as MSN2/4 exhibiting a conserved activation and ribosomal biogenesis regulators exhibiting a conserved repression across species. However, some notable differences were uncovered, including a pronounced inductive role under heat stress in all species for HSP60, which is known to have a regulatory role post heat stress. Consistent with its role in the *S. cerevisiae* OSR, SKO1 also exhibited a conserved role of upregulation in all species except *Schizosaccharomyces pombe*, and showed no significant association in HSR. Examples of regulators that changed their association with expression modules between stresses were observed primarily in a species-specific manner. In particular, PHO4 and TYE7 were associated with repression in heat shock in *C. albicans* (Figure 7) but did not have a significant association in *C. albicans* in osmotic stress (Figure 6). In summary, regulator associations with module expression are generally conserved across species for particular stresses. Regulator-module associations change their sign between stresses, but these changes are rare and happen in a species- and clade-specific manner.

DISCUSSION

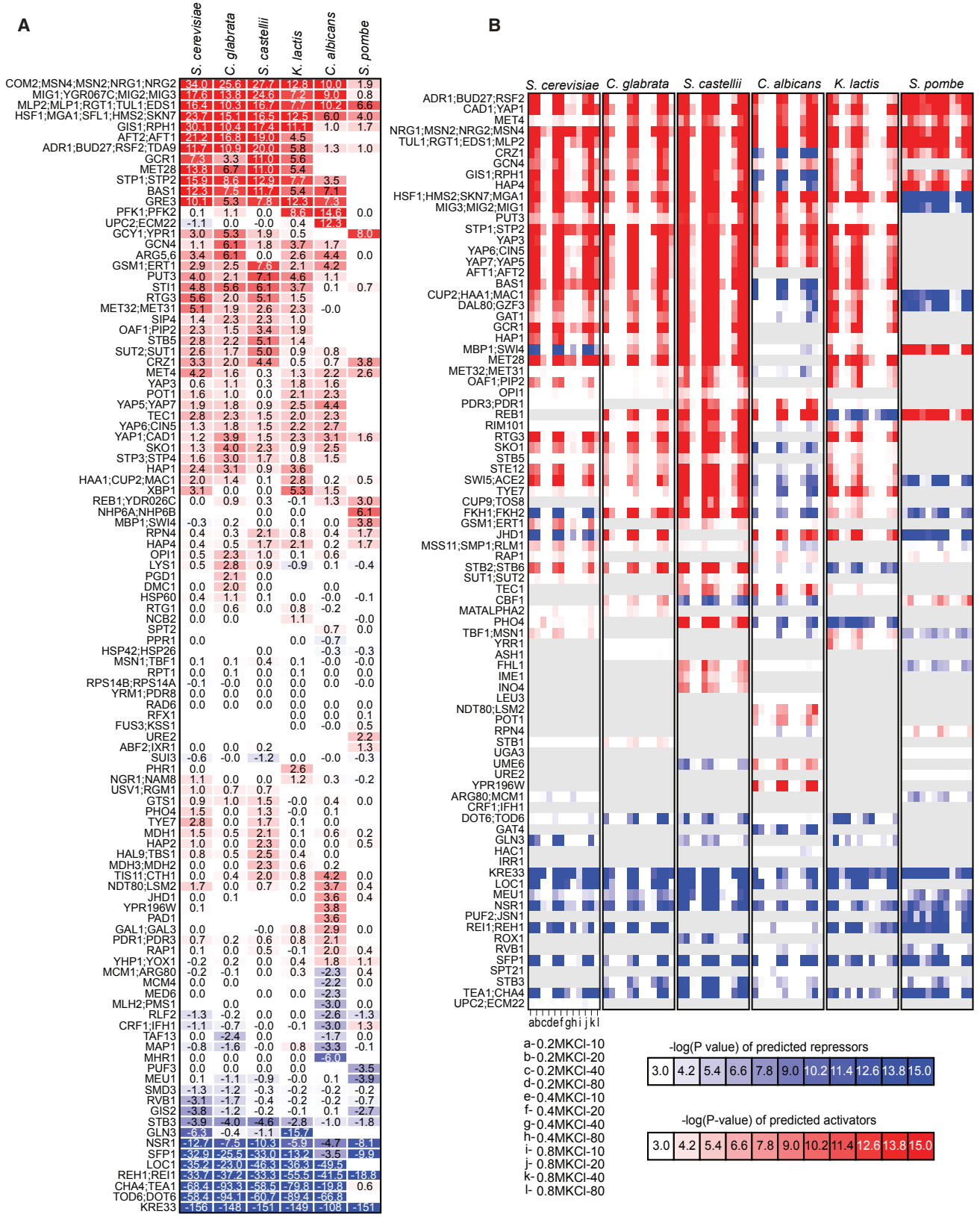
A comparative framework for regulatory networks can provide insights into principles of gene regulation (Garfield and Wray, 2010; Li and Johnson, 2010; Wohlbach et al., 2009), as well as inform better learning of network structure (Penfold et al., 2015; Thompson et al., 2015). Here, we have presented our algorithm MRTLE for inferring regulatory networks for multiple species related by a known phylogeny. MRTLE makes use of a known phylogenetic tree to explicitly model evolutionary rates of regulatory edge gain and loss and can additionally incorporate sequence-specific motifs to identify regulatory networks in a complex phylogeny. Furthermore, MRTLE is able to incorporate complex many-to-many orthology relationships arising from gene duplications, which are known to play a crucial role in

Figure 5. MRTLE + Arboretum Inferred Osmotic Stress Response Networks in Six Ascomycete Yeast Species

(A) Conservation of the inferred OSR networks for each species measured by F score.

(B) Networks spanning the most repressed and most induced OSR modules. Node size is proportional to node degree. Networks were constructed at the gene level rather than the orthogroup level, but nodes are labeled with *S. cerevisiae* orthology names for species other than *S. cerevisiae*. Nodes with many *S. cerevisiae* orthologs were truncated due to space considerations.

Regulator Osmotic Stress Response (OSR) Associations



(legend on next page)

regulatory network evolution (Voordeckers et al., 2015; Teichmann and Babu, 2004; Perez et al., 2014).

By leveraging data from related species within a phylogenetic framework, MRTLE is able to outperform methods that do not make use of evolutionary information (INDEP, GENIE3), in both simulated and real data settings. By favoring networks that are more phylogenetically coherent, MRTLE is able to recover the conserved parts of regulatory networks more accurately than methods that do not incorporate the phylogeny. MRTLE can accurately learn regulatory networks even when the sample size of expression data is small as demonstrated by our cross-species ChIP-chip comparisons. These results suggest that MRTLE can be an effective tool for inferring regulatory networks in non-model organisms, for which data are just becoming available and little is known about their regulatory networks. Computationally inferred high-confidence regulatory interactions could be critical for prioritizing ChIP-seq and regulator perturbation experiments needed to understand the regulatory networks in these poorly characterized species.

Inferring genome-wide regulatory networks in a large set of species enabled us to perform several systematic analyses to study regulatory network evolution. One of the properties that we discovered was the relatively higher rates of target gain and loss in regulators with duplications versus regulators without duplications. Notable exceptions were a few stress-related regulators that exhibited high rates of turnover but did not have duplications. Consistent with previous work (Pougach et al., 2014), we find that the MRTLE rates of TFs in duplicated families are more correlated to the sequence-derived rates, suggesting that sequence affinity divergence can facilitate TFs with duplications to diverge. However, additional experimental data measuring sequence affinity of individual members for a larger number of families are needed to more robustly examine this property. The MRTLE framework also enabled us to compare, for the first time to our knowledge, global transcriptional networks for a specific stress. We found that patterns of functional divergence of a regulator-module relationship were typically gradual and included a change from down- or upregulation to no significant association with a module. While some regulators changed their association across different stresses, most of the divergence in association is likely to occur gradually through fine-tuning of expression.

MRTLE can be extended in several directions. A particular challenge to employing MRTLE is setting the prior probability of an edge gain or loss for each branch. In this paper, we used previously established motif gain and loss rates as a proxy for regulatory edge gain and loss rates. While this yielded good performance in our setting, a different approach may be necessary in phylogenies where motif turnover rates are not available. In addition, one could incorporate variable gain and loss rates for each putative regulator, making use of prior information about each regulator's gain and loss rates. MRTLE's reliance on the high predictive power of a target's mRNA level based on the expression of TFs makes it difficult to discover potential regulatory roles of genes such as HOG1, which is known to be important in the OSR in *S. cerevisiae*. Integrating regulator activity levels that are less dependent on gene expression levels is another future extension to MRTLE. Another direction of future work is to extend our simulation to model the evolution of sequence-specific motifs, together with the network evolution model, to enable a more controlled study of the role of sequence and expression evolution in regulatory network evolution. In summary, MRTLE represents a powerful framework to infer and compare regulatory networks on a genome-wide scale in a complex phylogeny, and should enable furthering our understanding of regulatory network evolution and its impact on how species interact and adapt to environmental changes.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

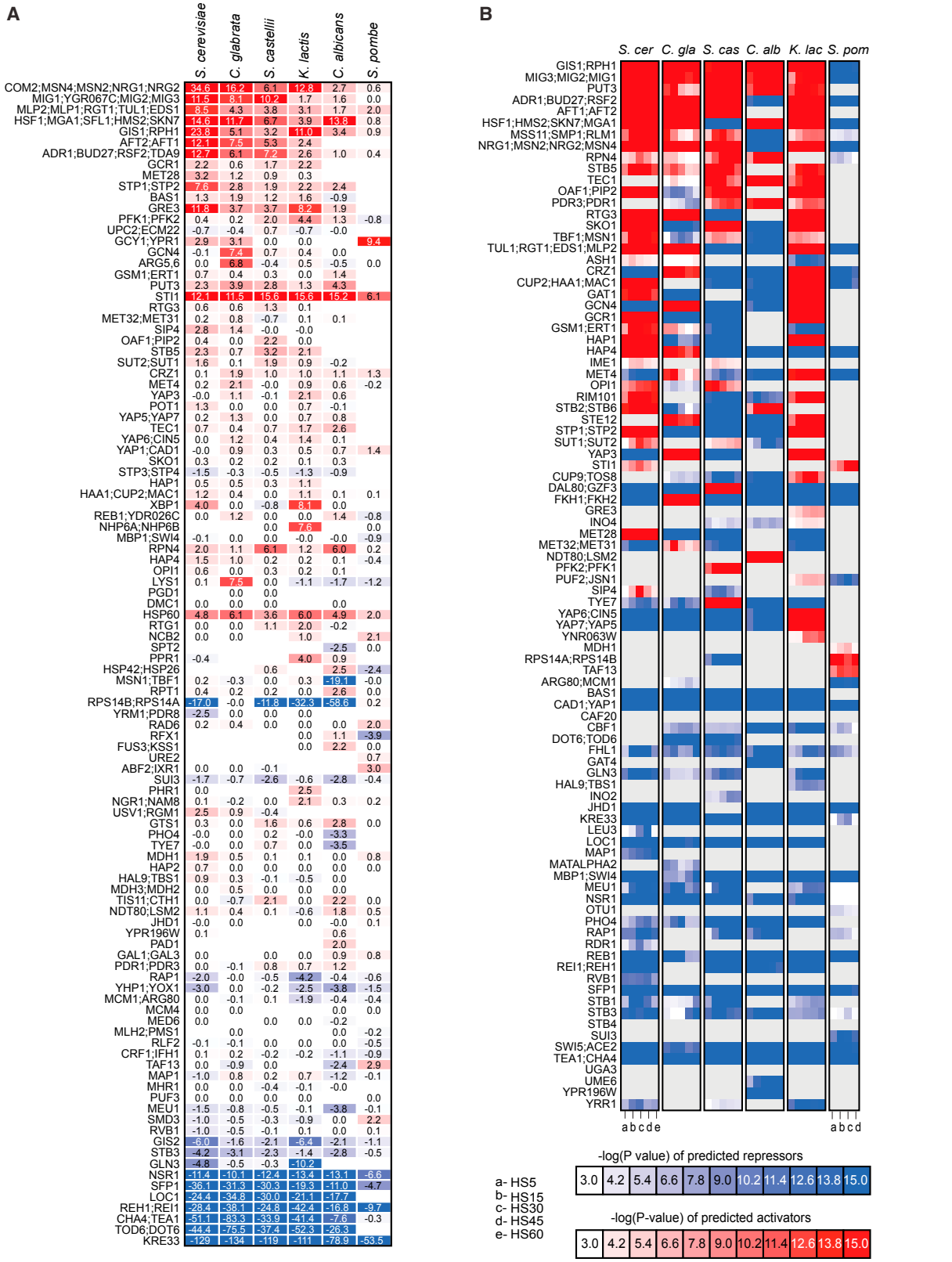
- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Osmotic Stress Response Gene Expression Profiling
- METHOD DETAILS
  - Probabilistic Framework for Phylogeny-Aware Regulatory Network Learning for Multiple Species: MRTLE
  - Details of the Baseline Algorithms Compared
  - Datasets
  - Evaluation of Learned Networks
  - Experiments on Simulated Data
  - Evaluation in Real Expression Setting
  - Evaluation Metrics
  - Inference of Stress-Specific Regulatory Networks for Multiple Species
  - Assessing a Regulator's Role as Repressive or Activating
  - Defining Targets of Selected Regulators based on LIMMA
  - Estimation of Gain and Loss Rates in MRTLE Inferred Network
- DATA AND SOFTWARE AVAILABILITY

Figure 6. Comparative Analysis of Regulator Association to Osmotic Stress Response Expression Levels

(A) Shown are regulator-module association scores computed using the most repressed and most induced OSR modules. Each association score represents the difference of the negative log(p value) from two hypergeometric tests, one for the most induced and one for the most repressed module. Positive scores (red) represent a stronger association with the most induced module compared with the repressed module, while negative scores (blue) represent a stronger association with the repressed module compared with the induced module. Blank scores represent a species for which a regulator was not present due to a gene loss event, or for which no targets were predicted in the top 30,000 edges in MRTLE. Regulators for which all species had low scores for both stresses (absolute value <2) are excluded from the figure (see Table S5 for all regulators).

(B) Shown is the negative log(p value) from a t test comparing the expression levels of targets of a regulator in the induced versus repressed module for each experimental condition (time point and stress signal) for OSR. The intensity of red or blue in each entry is proportional to −log(p value) (see color scale at the bottom). Regulators with more targets in the induced module than the repressed module are considered as activators and use the red color map. Regulators with more targets in the repressed module than the induced module are considered as repressors and use the blue color map.

Regulator Heat Stress Response (HSR) Associations



(legend on next page)

SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and five tables and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2017.04.010.

AUTHOR CONTRIBUTIONS

Conceived and Designed Study: S.R., A.R., E.K.O., J.K., and D.A.T. Algorithm Development and Implementation: S.R. and C.K. Computational Experiments: S.R., C.K., and S.K. Expression Time-Series and Validation Experiments: J.K., A.S., A.L., T.D., K.D., and D.A.T. Writing: S.R., C.K., J.K., and A.L. All authors provided feedback on the manuscript.

ACKNOWLEDGMENTS

This research was possible through generous support from NSF (NSF DBI: 1350677), the Sloan Foundation, the McDonnell Foundation (S.R), from HHMI (E.K.O), from NIH (R01CA119176-01, DP1OD003958-01), Broad Institute, HHMI, Burrough Wellcome Fund Career Award at the Scientific Interface and the Sloan Foundation (A.R), and by an NIH Ruth L. Kirschstein National Research Service Award (J.K). A.R. is an SAB member of Thermo Fisher Scientific, a consultant with the Driver Group, and an SAB member of Syros Pharmaceuticals.

Received: September 23, 2016
Revised: February 20, 2017
Accepted: April 26, 2017
Published: May 24, 2017

REFERENCES

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Similarities and differences in genome-wide expression data of six organisms. PLoS Biol. *2*, E9.

Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. Science *317*, 815–819.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). The evolution of gene expression levels in mammalian organs. Nature *478*, 343–348.

Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W.W., Bezault, E., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature *513*, 375–381.

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

Carroll, S.B. (2000). Endless forms: the evolution of gene regulation and morphological diversity. Cell *101*, 577–580.

Chen, D., Wilkinson, C.R., Watt, S., Penkett, C.J., Toone, W.M., Jones, N., and Bahler, J.U.R. (2008). Multiple pathways differentially regulate global oxidative stress responses in fission yeast. Mol. Biol. Cell *19*, 308–317.

Davis, J., and Goadrich, M. (2006). The Relationship between Precision-Recall and ROC Curves (ACM Press), pp. 233–240.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J., and Gardner, T.S. (2007). Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. PLoS Biol. *5*, E8.

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

Fisher, S., Barry, A., Abreu, J., Minie, B., Nolan, J., Delorey, T.M., Young, G., Fennell, T.J., Allen, A., Ambrogio, L., et al. (2011). A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. *12*, R1.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. Science *303*, 799–805.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J. Comput. Biol. *7*, 601–620.

Garber, M., Guttman, M., Clamp, M., Zody, M.C., Friedman, N., and Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. Bioinformatics *25*, i54–i62.

Garfield, D.A., and Wray, G.A. (2010). The evolution of gene regulatory interactions. BioScience *60*, 15–23.

Gasch, A.P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. Yeast *24*, 961–976.

Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D., and Brown, P.O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell *11*, 4241–4257.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018.

Habib, N., Wapinski, I., Margalit, H., Regev, A., and Friedman, N. (2012). A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. Mol. Syst. Biol. *8*, 619.

Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., MacIsaac, K.D., Danford, T.W., Hannett, N.M., Tagne, J.-B., Reynolds, D.B., Yoo, J., et al. (2004). Transcriptional regulatory code of a eukaryotic genome. Nature *431*, 99–104.

Heckerman, D., Chickering, D.M., Meek, C., Rounthwaite, R., and Kadie, C. (2001). Dependency networks for inference, collaborative filtering, and data visualization. J. Mach. Learn. Res. *1*, 49–75.

Hiyama, A., Taira, W., and Otaki, J.M. (2012). Color-pattern evolution in response to environmental stress in butterflies. Front. Genet. *3*, 15.

Hobolth, A., and Jensen, J.L. (2005). Statistical inference in evolutionary models of DNA sequences via the EM algorithm. Stat. Appl. Genet. Mol. Biol. *4*, http://dx.doi.org/10.2202/1544-6115.1127.

Hoffmann, A.A., and Willi, Y. (2008). Detecting genetic responses to environmental change. Nat. Rev. Genet. *9*, 421–432.

Homann, O.R., Dea, J., Noble, S.M., and Johnson, A.D. (2009). A phenotypic profile of the candida albicans regulatory network. PLoS Genet. *5*, 1–12.

Hu, Z., Killion, P.J., and Iyer, V.R. (2007). Genetic reconstruction of a functional transcriptional regulatory network. Nat. Genet. *39*, 683–687.

Hughes, T.R., and de Boer, C.G. (2013). Mapping yeast transcriptional networks. Genetics *195*, 9–36.

Figure 7. Comparative Analysis of Regulator Association to Heat Stress Response Expression Levels

(A) Shown are regulator-module association scores computed using the most repressed and most induced HSR modules. Each association score represents the difference of the negative log(p value) from two hypergeometric tests, one for the most induced and one for the most repressed module. Positive scores (red) represent a stronger association with the most induced module compared with the repressed module, while negative scores (blue) represent a stronger association with the repressed module compared with the induced module. Blank scores represent a species for which a regulator was not present due to a gene loss event, or for which no targets were predicted in the top 30,000 edges in MRTLE. Regulators for which all species had low scores for both stresses (absolute value <2) are excluded from the figure (see Table S5 for all regulators).

(B) Shown is the negative log(p value) from a t test comparing the expression levels of targets of a regulator in the induced versus repressed module for each experimental condition (time point and stress signal) for HSR. The intensity of red or blue in each entry is proportional to −log(p value) (see heatmaps at the bottom). Regulators with more targets in the induced module than the repressed module are considered as activators and use the red color map. Regulators with more targets in the repressed module than the induced module are considered as repressors and use the blue color map.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One *5*, e12776.

Ihmels, J., Bergmann, S., Berman, J., and Barkai, N. (2005). Comparative gene expression analysis by differential clustering approach: application to the *Candida albicans* transcription program. PLoS Genet. *1*, e39.

Joshi, A., Beck, Y., and Michoel, T. (2014). Multi-species network inference improves gene regulatory network reconstruction for early embryonic development in *Drosophila*. J. Comput. Biol. *22*, 253–265.

Kim, H.D., Shay, T., O'Shea, E.K., and Regev, A. (2009). Transcriptional regulatory circuits: predicting numbers from alphabets. Science *325*, 429–432.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. Science *188*, 107–116.

Kristiansson, E., O sterlund, T., Gunnarsson, L., Arne, G., Larsson, J.G., and Nerman, O. (2013). A novel method for cross-species gene expression analysis. BMC Bioinformatics *14*, 70.

Lauritzen, S.L. (1996). Graphical Models (Oxford University Press).

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

Li, H., and Johnson, A.D. (2010). Evolution of transcription networks – lessons from yeasts. Curr. Biol. *20*, R746–R753.

Macisaac, K., Wang, T., Gordon, D.B., Gifford, D., Stormo, G., and Fraenkel, E. (2006). An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. BMC Bioinformatics *7*, 113.

Marbach, D., Costello, J.C., Küffner, R., Vega, N.M., Prill, R.J., Camacho, D.M., Allison, K.R., DREAM5 Consortium, Kellis, M., Collins, J.J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. Nat. Methods *9*, 796–804.

Markowetz, F., and Spang, R. (2007). Inferring cellular networks - a review. BMC Bioinformatics *8* (*Suppl 6*), S5.

Nicholls, S., Straffon, M., Enjalbert, B., Nantel, A.E., Macaskill, S., Whiteway, M., and Brown, A.J.P. (2004). Msn2- and Msn4-like transcription factors play no obvious roles in the stress responses of the fungal pathogen *Candida albicans*. Eukaryot. Cell *3*, 1111–1123.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W., MacIsaac, K.D., Rolfe, P.A., Conboy, C.M., Gifford, D.K., and Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat. Genet. *39*, 730–732.

Perez, J.C., Fordyce, P.M., Lohse, M.B., Hanson-Smith, V., DeRisi, J.L., and Johnson, A.D. (2014). How duplicated transcription regulators can diversify to govern the expression of nonoverlapping sets of genes. Genes Dev. *28*, 1272–1277.

Pe'er, D., Tanay, A., and Regev, A. (2006). MinReg: a scalable algorithm for learning parsimonious regulatory networks in yeast and mammals. J. Mach. Learn. Res. *7*, 167–189.

Penfold, C.A., Millar, J.B.A., and Wild, D.L. (2015). Inferring orthologous gene regulatory networks using interspecies data fusion. Bioinformatics *31*, i97–i105.

Pougach, K., Voet, A., Kondrashov, F.A., Voordeckers, K., Christiaens, J.F., Baying, B., Benes, V., Sakai, R., Aerts, J., Zhu, B., et al. (2014). Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. Nat. Commun. *5*, 4868.

Ramsdale, M., Selway, L., Stead, D., Walker, J., Yin, Z., Nicholls, S.M., Crowe, J., Sheils, E.M., and Brown, A.J.P. (2008). MNL1 regulates weak acid–induced stress responses of the fungal pathogen *Candida albicans*. Mol. Biol. Cell *19*, 4393–4403.

Romero, I.G., Ruvinsky, I., and Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. Nat. Rev. Genet. *13*, 505–516.

Roy, S., Lagree, S., Hou, Z., Thomson, J.A., Stewart, R., and Gasch, A.P. (2013a). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. PLoS Comput. Biol. *9*, e1003252.

Roy, S., Wapinski, I., Pfiffner, J., French, C., Socha, A., Konieczka, J., Habib, N., Kellis, M., Thompson, D., and Regev, A. (2013b). Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. Genome Res. *23*, 1039–1050.

Sanso, M., Gogol, M., Ayte, J., Seidel, C., and Hidalgo, E. (2008). Transcription factors Pcr1 and Atf1 have distinct roles in stress- and Sty1-dependent gene regulation. Eukaryot. Cell *7*, 826–835.

Schaffter, T., Marbach, D., and Floreano, D. (2011). GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics *27*, 2263–2270.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (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., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Siahpirani, A.F., and Roy, S. (2016). A prior-based integrative framework for functional transcriptional regulatory network inference. Nucleic Acids Res. *45*, gkw963.

Smyth, G.K.K., Michaud, J.E.L., and Scott, H.S.S. (2005). Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics *21*, 2067–2075.

Teichmann, S.A., and Babu, M.M. (2004). Gene regulatory network growth by duplication. Nat. Genet. *36*, 492–496.

Thompson, D.A., Roy, S., Chan, M., Styczynsky, M.P., Pfiffner, J., French, C., Socha, A., Thielke, A., Napolitano, S., Muller, P., et al. (2013). Evolutionary principles of modular gene regulation in yeasts. Elife *2*, e01114.

Thompson, D., Regev, A., and Roy, S. (2015). Comparative analysis of gene regulatory networks: from network reconstruction to evolution. Annu. Rev. Cell Dev. Biol. *31*, 399–428.

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H., and Johnson, A.D. (2008). The evolution of combinatorial gene regulation in fungi. PLoS Biol. *6*, e38.

Voordeckers, K., Pougach, K., and Verstrepen, K.J. (2015). How do regulatory networks evolve and expand throughout evolution? Curr. Opin. Biotechnol. *34*, 180–188.

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

Wapinski, I., Pfiffner, J., French, C., Socha, A., Thompson, D.A., and Regev, A. (2010). Gene duplication and the evolution of ribosomal protein gene regulation in yeast. Proc. Natl. Acad. Sci. USA *107*, 5505–5510.

Wittkopp, P.J. (2007). Variable gene expression in eukaryotes: a network perspective. J. Exp. Biol. *210*, 1567–1575.

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

Xie, D., Chen, C.-C., He, X., Cao, X., and Zhong, S. (2011). Towards an evolutionary model of transcription networks. PLoS Comput. Biol. *7*, e1002064.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Yeast RNA-Seq and microarray data	This Paper	GEO: GSE94628
Experimental Models: Organisms/Strains		
S. cerevisiae: W303	O'Shea Lab	W303
C. glabrata: CBS 138	O'Shea Lab	CBS 138
S. castellii: CLIB 592	O'Shea Lab	CLIB 592
K. lactis: CLIB 209	O'Shea Lab	CLIB 209
C. albicans: SC5314	O'Shea Lab	SC5314
C. albicans: BWP17	O'Shea Lab	BWP17
S. pombe: SPY73h+	O'Shea Lab	SPY73h+
C. albicans: SN152 sko1	Alexander Johnson; Homann et al. 2009 (PMID: 20041210)	SN152 sko1Δ/sko1Δ
C. albicans: SN152 mnl1	Alexander Johnson; Homann et al. 2009 (PMID: 20041210)	SN152 mnl1Δ/mnl1Δ
S. pombe: SPAC3H1.11 hsr1	Bioneer	SPAC3H1.11 hsr1Δ
S. cerevisiae: W303 msn2/4	O'Shea Lab; Capaldi et al. 2008 (PMID: 18931682)	W303 msn2Δ, msn4Δ
S. cerevisiae: W303 sko1	O'Shea Lab; Capaldi et al. 2008 (PMID: 18931682)	W303 sko1Δ
Software and Algorithms		
MRTLE	This Paper	https://bitbucket.org/roygroup/mrtle
INDEP/PGG	Siahpirani and Roy 2016	https://bitbucket.org/roygroup/pgg
GENIE3	Huynh-Thu et al. 2010	http://www.montefiore.ulg.ac.be/~huynh-thu/GENIE3.html
LIMMA	Smyth et al. 2005	http://www.bioconductor.org/packages/release/bioc/html/limma.html
Arboretum	Roy et al., 2013b	https://bitbucket.org/roygroup/arboretum_v2.0
GeneNetWeaver	Schaffter et al. 2011	http://tschaffter.ch/projects/gnw/
Precision, Recall, and AUPR Calculation Script	Davis and Goadrich 2006	http://mark.goadrich.com/programs/AUC/
Other		
MacIsaac et al. yeast ChIP-chip network	MacIsaac et al. 2006	PMID:16522208
Hu et al. yeast knockout network	Hu et al. 2007	PMID:17417638
Yeast MCM1 ChIP-chip networks	Tuch et al. 2008	PMID:1830948
Yeast multi-TF ChIP-chip networks	Lavoie et al. 2010	PMID:20231876
Motif instances	Habib et al. 2012	PMID:23089682

CONTACT FOR REAGENT AND RESOURCE SHARING

Information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Sushmita Roy, sroy@biostat.wisc.edu.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Osmotic Stress Response Gene Expression Profiling

Strains and Growth Conditions

The following wild-type strains were used for each species in the study: *S. cerevisiae* W303 (Capaldi et al., 2008), *C. glabrata* CBS 138, *S. castellii* CLIB 592, *K. lactis* CLIB 209, *C. albicans* SC5314, *C. albicans* BWP17, *S. pombe* SPY73h+. Deletion mutant strain

S. pombe HSR1 was obtained from Bioneer. *S. cerevisiae* deletion mutation strains of *MSN2/4* and *SKO1* were created on the W303 wild type strain previously described in Capaldi et al. (Capaldi et al., 2008). Deletion mutation strain for *C. albicans SKO1* and *MNL1* were previously described in Homann et al. (Homann et al., 2009). All species were grown in the following rich medium chosen to minimize cross-species variation in growth (termed BMW): yeast extract (1.5%), peptone (1%), dextrose (2%), SC amino acid mix (Sunrise Science) 2 g/L, adenine 100 mg/L, tryptophan 100 mg/L, uracil 100 mg/L (Thompson et al., 2013). For each strain, cells were plated onto BMW plates from frozen glycerol stocks. After 2 days, cells were taken from plates and re-suspended into liquid BMW and grown overnight. Approximately 100-1500ul quantities (depending on the species growth rate and timing constraints for the days experiments) of the overnight cultures were used to inoculate pre-warmed, 350 ml BMW cultures in 2L Erlenmeyer flasks in New Brunswick Scientific water bath model C76 shakers. All strains were grown at 180 rpm at 30 C except for *S. castellii*, which was grown at 25 C.

Osmotic Stress Response Profile Experiments

The OD600 was measured throughout the day to ensure culture growth was tracking as expected (Thompson et al., 2013). When samples reached a species-specific OD600, corresponding to slightly late mid-log, we transferred 150ml of the culture to each of 100ml BMW (CTRL) and 100ml BMW+KCl (EXP). Both CTRL and EXP media were pre-warmed for 40-60 minutes in the shaker prior to the experiment. EXP media was either BMW + 0.5M KCl, 1M KCl, or 2M KCl, yielding a final concentration of 0.2M KCl, 0.4M KCl, and 0.8M KCl, respectively, upon addition of the culture. In each case, CTRL media was added first, followed by the EXP media, whereupon the shaker was immediately activated to 180 rpm and the timer started simultaneously. Samples (20ml) were collected from the CTRL media + culture immediately upon activation of the shaker (T=0), then at T=10, 20, 40, and 80 minutes from the EXP media + culture. Samples were collected in 50 mL conicals filled with 30ml of 100% methanol to yield a 60/40 methanol/sample mixture. The methanol-filled tubes were stored at -80 C until ready for use. During sample collection tubes were placed in a rack in a dry-ice ethanol bath kept at approximately -40 C. Once the sample was added to the methanol, the methanol and media were separated from the cells by centrifugation and poured off. The conicals containing a cell pellet were flash frozen in liquid nitrogen and then stored at -80 C until processed for permanent storage or RNA isolation. To process, the cell pellets were washed in 5 ml of nuclease-free water and spun for 5 min at 3700 rpm at 4 C. The supernatant was discarded and the pellet re-suspended in 2 mL of RNAlater (Ambion) and transferred to 2 ml Sarstadt tubes for storage.

RNA Preparation and Labeling

Total RNA was isolated using the RNeasy Mini Kits (Qiagen) according to the provided instructions for mechanical lysis. Samples were quality controlled with the RNA 6000 Pico kit of the Bioanalyzer 2100 (Agilent). Total RNA samples were labeled with either Cy3 or Cy5 using a modification of the protocol developed by Joe DeRisi (University of California at San Francisco) and Rosetta Inpharmatics as described previously (Wapinski et al., 2010). In the case of the OSR profile experiments, the control was a pooled sample, consisting of equal quantities of 160ng RNA from each of the T= 0, 10, 20, 40, and 80 minute samples. The pool was constructed prior to the SS-III reverse transcription step, where the Agilent spike-in A (or spike-in B in the case of labeling with Cy5), could be incorporated into the reaction.

Microarray Hybridization

We used two-color Agilent 55- or 60-mer oligo-arrays in the 4 × 44 K format (four to five probes per target gene) and 8 × 15 K format (two probes per target gene). After hybridization and washing per the manufacturer's instructions, arrays were scanned using an Agilent scanner and analyzed with Agilent's Feature Extraction software (release 10.5.1.)

cDNA Synthesis for Mi-seq RNA-Sequencing Gene Expression Studies

1 ug of total RNA in a volume of 11 uL was used as input. Heat fragmentation was completed by adding 3 uL The RNA Storage Solution-Ambion (AM7000) to each sample in an Eppendorf 96 well plate (951020401, Fisher Scientific) and heating at 98°C for 30 minutes. First strand cDNA was created by adding 1uL of OligoDT to samples and heating at 70C 10min. Samples were then put immediately on ice. A mastermix of 2 uL of 10x Affinity script buffer, 0.8 uL of 25mM dNTPs, 2 uL of DTT and 1 uL of the AffinityScript RT Enzyme (AffinityScript Multiple Temperature Reverse Transcriptase, 600109) was created. 5.8 uL of the mastermix was added to each sample well and mixed. Samples were incubated at room temperature of 10 minutes in a thermocycler, followed by 1 hour at 50°C, 15 minutes at 70°C and a 4°C hold.

Second strand cDNA synthesis was completed with mRNA Second Strand Synthesis Module, E6111L. cDNA synthesis reaction was cleaned up using Agencourt AMPure XP beads (A63881, Beckman Coulter). Sample and beads were used as input to library construction; beads remain in the plate well with sample until the adapter ligation cleanup.

Library Construction for Sequencing

Libraries were created using KAPA Biosystems Library Preparation Kit (KK2505 and KK8202) in an Eppendorf 96 well plate. Enzymatic reactions were cleaned up by adding AMPure XP to the sample after end repair, and leaving the beads in the sample throughout adapter ligation. 20% PEG, NaCl 2.5M was added to samples and beads for A-base and Adapter ligation cleanup, as previously described (Fisher et al., 2011). Prior to library enrichment, samples were eluted from AMPure XP beads. For library enrichment and amplification, a mastermix containing 12 uL of 5X Kapa HiFi Fidelity Buffer 2mM Mg, 1 uL of 25 mM dNTPs, 4 uL of primer mix, 1 uL of Kapa HiFi HotStart Enzyme and 2 uL of water per sample was created. 20 uL of the mastermix was added to the sample and the following PCR program was run: 98C for 45 seconds, 12 cycles of 98C for 15 seconds, 60C for 30 seconds and 72C for 30 seconds, a final extension at 72C for 1 min and a 4C hold. The library enrichment reaction was cleaned by adding 60 uL of AMPure XP beads and samples were eluted off of the AMPure beads in 15 ul of Trish-HCL (pH 8). The samples were then transferred to new plate and library quality was assessed.

Library Quality Control

Libraries were checked for quality control using Agilent High Sensitivity D5000 Screentape assay; size range for each sample was between 200-500 base pairs.

Library Sequencing

Each library was diluted to 2nM and pooled prior to sequencing. Sequencing was completed on the MiSeq platform and a 25 x 25, paired end sequencing run was completed.

METHOD DETAILS

Probabilistic Framework for Phylogeny-Aware Regulatory Network Learning for Multiple Species: MRTLE

Our multi-species network inference approach is based on a probabilistic graphical model representation of a regulatory network (Friedman, 2004; Segal et al., 2003; Friedman et al., 2000; Markowetz and Spang, 2007; Pe'er et al., 2006). Bayesian networks (Friedman et al., 2000) and dependency networks (Heckerman et al., 2001) are examples of probabilistic graphical models that have been used to represent regulatory networks. Here, we use a dependency network representation because they can be relatively easily learned from observed expression data and can capture cyclic dependencies (Huynh-Thu et al., 2010; Heckerman et al., 2001). Below, we first give a description of a probabilistic model representation of a regulatory network for a single species, followed by a description of the probabilistic priors we have employed to capture phylogenetic relationships, and then a sketch of the MRTLE algorithm.

Modeling a Regulatory Network in One Species

A probabilistic graphical model (PGM) of a regulatory network has two components: the *structure*, which specifies the regulators of a target gene, and the *parameterized functions*, which describe the sign and magnitude of the interactions of individual and combinations of regulators specifying the expression of a target gene. In PGMs, the expression level of a gene i is captured by a random variable, X_i, and a conditional probability distribution relates the expression levels of regulators to the expression level of a target gene, by specifying the probability of a target gene taking a specific expression value given the expression values of its regulators. In MRTLE, X_i and its parents are assumed to be jointly Gaussian and the conditional distribution for each X_i given its parent is a conditional Gaussian.

Extending to Multiple Regulatory Networks

Let N denote the number of species, and let G_s denote the graph associated with the s^{th} species. Let D_s denote the expression datasets associated with the s^{th} species that represent measured expression levels of both targets and regulators under multiple conditions. Given datasets, $D_1,\cdots,D_N$ and a phylogenetic tree over N species, our goal is to simultaneously infer the unknown regulatory networks $G_1,\cdots,G_N$ for all species. We use a Bayesian framework to tackle this problem and optimize the posterior probability of the graphs given the data, $P(G_1,\cdots,G_N|D_1,\cdots,D_N)$. Using Bayes rule this is proportional to $P(D_1,\cdots,D_N|G_1,\cdots,G_N)P(G_1,\cdots,G_N)$, where $P(D_1,\cdots,D_N|G_1,\cdots,G_N)$ is the data likelihood that is computed easily for each species independently, $\prod_{s=1}^{N}(PD_s|G_s)$. $P(G_1,\cdots,G_N)$ is the prior over the N graphs. To incorporate the phylogenetic similarity between species, we use a specific formulation of the multi-graph prior, which we describe below.

Phylogenetic and Species-Specific Graph Priors

$P(G_1,\cdots,G_N)$ is defined as a product of $\Gamma_1(G_1,\cdots,G_N)$, which captures the multi-species phylogenetic prior, and $\Gamma_2(G_1\cdots,G_N)$ that captures any species-specific regulatory information such as binding sites. Γ_1 and Γ_2 each define a distribution over a set of graphs, where each graph is represented by a set of edges from regulators to target genes. These are not bipartite graphs because there exist genes that act as both regulators and targets. To describe Γ_1 in more detail we make use of the concept of an orthogroup (Wapinski et al., 2007), which is defined as a set of orthologous genes. Each orthogroup contains 0 or more gene members from each species. We assume that Γ_1 decomposes as a product over sets of edges between regulator orthogroups and target orthogroups. For simplicity, we first assume that each species has one gene in each regulator orthogroup and one gene in each target orthogroup. Later, we describe how to relax this assumption. Let $\mathbf{l}_{jk}=\{l_{jk}^1,\cdots,l_{jk}^N\}$ be a binary vector for each regulator j and target gene k pair. l_{jk}^i is a binary variable capturing the state of the edge from regulator j to target k for the i^{th} species, taking a value of 0 if the edge is absent and 1 if the edge is present. We express this prior as $\Gamma_1(G_1,\cdots,G_N)=\prod_{j\to k}P(\mathbf{l}_{jk})$, which assumes that the prior decomposes as a product over the edges. $P(\mathbf{l}_{jk})$ can be efficiently computed using Felsentein's algorithm for computing the probability of discrete observations at the leaf nodes of a phylogenetic tree (Felsenstein, 1981). First, we expand $\mathbf{l}_{jk}$ to include the ancestral species at the $N-1$ intermediate points in the tree using indices $N+1$ to $2N-1$ to represent these internal points. $P(l_{jk}^1,\cdots,l_{jk}^N)$ requires us to integrate away the state of the edges at the internal nodes as $\sum_{l_{jk}^{N+1}},\ldots,\sum_{l_{jk}^{2N-1}}P(l_{jk}^1,\ldots,l_{jk}^N,l_{jk}^{N+1},\ldots,l_{jk}^{2n-1})$. Using the tree structure to make independence assumptions, we can write this as $\sum_{l_{jk}^{N+1}},\ldots,\sum_{l_{jk}^{2N-1}}P(l_{jk}^{2N-1})\prod_l P(l_{jk}^l|l_{jk}^{pa(l)})$, where $pa(l)$ denotes the immediate ancestor species of l. Hence the probability, $P(\mathbf{l}_{jk})$, can be computed efficiently using the probability of an edge state in l, given the state of the edge in the ancestor of l, $P(l_{jk}^l|l_{jk}^{pa(l)})$.

Two parameters, p_g and p_m, each taking values from zero to one, are used to determine this probability. The first, denoted p_g, represents the probability of *gaining* a regulatory edge given that the edge does *not* exist in the ancestral species. The second, denoted p_m, represents the probability of *maintaining* a regulatory edge, given its presence in the ancestral species. Setting these parameters to appropriate values is a difficult task. In our experiments on real data with six yeast species, we estimated a rate matrix using the average rate of motif binding site gain and loss from (Habib et al., 2012). We then set p_g and p_m for each branch in the phylogenetic tree based on this rate matrix and the branch length. In this regard, we used binding site gain and loss rates as a proxy for regulatory edge gain and loss rates. Thus our prior Γ_1 is parameterized by branch lengths and the two rate parameters that are multiplied to obtain the probabilities p_g and p_m. Because branch lengths vary, the probabilities p_m and p_g are modeled separately for each branch. The second part of the prior, $\Gamma_2(G_1, \cdots, G_N)$, acts in a per-species manner, and can be further decomposed as a product over species-specific graphs $\Gamma_2(G_1, \cdots, G_N) = \prod_{i=1}^{N} P(G_i)$. Each $P(G_i)$ further decomposes as a product of edges, $P(I(X_j \rightarrow X_k))$, where I is an indicator function for an edge existing between regulator j and target k. Similar to (Roy et al., 2013a), we parameterize the prior probability as a logistic function: $P(I(X_j \rightarrow X_k) = 1)$ as $\frac{1}{1+exp(\beta_0 + \beta_1 * m_{jk})}$. Here, m_{jk} specifies whether gene k has a motif in its promoter region that can be bound by regulator j. In our current implementation of the algorithm, each m_{jk} takes on a real value, proportional to the significance of an instance of j's motif found in k's promoter. These weights could be estimated using a standard motif scanning tool, for example FIMO (Grant et al., 2011). β_0 is a sparsity prior that can be used to control the extent to which the algorithm penalizes the addition of a new edge. β_1 controls the strength of the motif prior. Both β_0 and β_1 are user-tunable parameters. The addition of the motif prior enables us to select interactions that are weakly predicted by expression data, but are supported by the motif presence. Note that this framework is flexible and can easily be modified to fit a scenario where we do not have species-specific motif information (β_1=0), or in settings where additional types of prior information for an edge are present.

Score-Based Learning of Regulatory Networks

To infer graphs for all species we use a score-based approach that searches over the space of possible graphs. Because the space of possible graphs is super-exponential in the number of variables, it is not possible to find a global optima. Instead, it is typical to use heuristic search algorithms over the graph space, score each candidate graph, and select the one that corresponds to a local optima. In the multi-species setting, we need to simultaneously search over the N graphs. Specifically, the score of a current graph configuration is composed of the data likelihood, $P(D_1, \cdots, D_N | G_1, \cdots, G_N)$, as well as the graph prior, $P(G_1, \cdots, G_N)$. As described above, $P(D_1, \cdots, D_N | G_1, \cdots, G_N)$ is written as a product over the N species, $\prod_s P(D_s | G_s)$. In a dependency network we cannot easily compute the likelihood $P(D_s|G_s)$, but instead we compute a *pseudo likelihood*, which is given by the product of conditional distributions, $P(X_i^s | \mathbf{R}_{X_i}^s)$, where $\mathbf{R}_{X_i}^s$ denotes the regulator set for X_i^s in species s. We assume that each variable X_i^s and its regulators $\mathbf{R}_{X_i}^s$ are distributed according to a multi-variate Gaussian. The conditional $P(X_i^s | \mathbf{R}_{X_i}^s)$ is a conditional Gaussian distribution with mean $\mu_{X_i^s | \mathbf{R}_{X_i}^s}$ and variance $\sigma_{X_i^s | \mathbf{R}_{X_i}^s}$, estimated from the joint using Lauritzen et al (Lauritzen, 1996). Using the conditional means, $\mu_{X_i^s | \mathbf{R}_{X_i}^s}$, and variances, $\sigma_{X_i^s | \mathbf{R}_{X_i}^s}$, of a variable given its regulator set, we compute the conditional data likelihood for each variable X_i^s using data from species s. To compute the portion of the score representing the graph prior, we need to compute Γ_1 and Γ_2. As described above, Γ_1 decomposes as a product over each possible regulator-target orthogroup, and can be computed using Felsentein's algorithm (Felsenstein, 1981), while Γ_2 is computed in a species-specific manner.

Handling Non-Uniform Orthogroups

In the description so far, we have assumed that each species has exactly one gene in the regulator orthogroup, and exactly one gene in the target orthogroup. However, for most evolutionary studies, the ability to handle many-to-many mappings between species is essential. In our problem setting, when there are duplications, we need to specially handle the $\mathbf{I}_{jk}$ variable that specifies the state of edges between the regulatory orthogroup j and the target orthogroup k. If a species has more than one gene in the regulator orthogroup or target orthogroup, we consider all possible edges between the genes in the regulator orthogroup to the genes in the target orthogroup, and select the edge that has the highest improvement in score. That is, if a species l has p regulators and q targets in the j^{th} and k^{th} orthogroups, respectively, we will consider all $p \times q$ edges for that species. We set $I_{jk}^l = 1$ if any member of the j^{th} regulator orthogroup has an edge to any member of the k^{th} target orthogroup.

Computational Complexity of the MRTLE Algorithm

MRTLE uses a greedy network learning algorithm which operates on one orthogroup at a time, which can be parallelized because the priors decompose at the orthogroup level. The search decomposes into per-orthogroup regulator set estimation problems, where the orthogroup corresponds to the target gene. In each iteration, for a target orthogroup, MRTLE would search among all regulator orthogroups to find the best regulator orthogroup that would result in an overall score improvement. For a target orthogroup j and regulator orthogroup i, this score improvement is calculated based on: (a) A species-specific contribution that examines all regulators genes in the orthogroup and all target genes in the target orthogroup to find the regulator gene pair with the highest score improvement. (b) The computation of the phylogenetic prior. Operation (a) requires

$n_i^s \times m_j^s$ operations in species s with n_i^s regulators in the i^{th} orthogroup and m_j^s genes in the j^{th} orthogroup. For N species, the overall complexity for this calculation is $O(Nn_im_j)$, where n_i and m_j are the maximum number of regulator and target genes in the i^{th} and j^{th} orthogroups respectively. The second operation of computing the phylogenetic prior uses the Felsenstein algorithm that is linear in the number of species, $O(N)$. Taken together, scoring a given target and regulator orthogroup pair is therefore $O(Nn_im_j)$, which we write simply as $O(Nnm)$, with n denoting the maximum number of genes in a species in a regulator orthogroup and m denoting the maximum number of genes in the target orthogroup in a species. This search procedure is executed for all regulator orthogroups to find the best move. If R is the total number of regulator orthogroups, finding the best move takes $O(RNnm)$. Finally, the iteration of finding the next best regulator is executed at most the maximum number of pre-specified regulators a gene can have. Let this be k. Hence the overall complexity of the MRTLE algorithm is $O(kRNnm)$.

We note that the complexity of the algorithm without the phylogenetic prior, would require $O(kRNnm)$ operations as well. However, this can be parallelized across species and therefore would be faster.

Details of the Baseline Algorithms Compared

We compared MRTLE to two baseline algorithms, GENIE3 and INDEP, both of which aimed to learn a regulatory network for each species independently.

GENIE3

GENIE3 is a dependency network learning algorithm that infers the structure of the regulatory network by solving a set of individual regression problems, one per gene. Each regression problem is solved by learning tree-based ensembles (either Random Forests or Extra Trees) that represent the regulatory program of a gene. GENIE3 takes as input an expression data matrix and a set of candidate regulators and outputs a ranking of potential regulatory edges. GENIE3 was one of the best performers in the DREAM network inference challenge (Huynh-Thu et al., 2010; Marbach et al., 2012).

INDEP

The INDEP algorithm is also a dependency network learning algorithm, that infers the structure of the regulatory network by solving a set of individual linear regression problems. The INDEP algorithm uses a per-gene greedy algorithm that aims to infer the regulators of each gene one at a time and is described in more detail in Siahpirani & Roy (Siahpirani and Roy, 2016) as the Per-Gene Greedy (PGG) algorithm. Briefly if D_s represents the dataset for the s^{th} species, INDEP aims to learn the graph structure G_s by optimizing $P(G_s|D_s)$ which is proportional to $P(D_s|G_s)P(G_s)$. $P(G_s)$ is defined in the same manner as the species-specific prior of MRTLE. INDEP makes the same assumptions as MRTLE about the Gaussian distribution of the gene expression data (See Sections Modeling a regulatory network in one species and Phylogenetic and species-specific graph priors).

Datasets

We evaluated our learning algorithm on simulated data with known ground truth, as well as with real yeast expression data.

Simulated Datasets

Our simulation framework made use of a simple probabilistic process of network structure evolution, which was parameterized with the probability, p_g, of gaining an edge that does not exist in the ancestral species, and the probability, p_m, of maintaining an edge that exists in the ancestral species. The simulation started from an ancestral network of 300 genes and 33 regulators and a species tree shown in Figure 1B, and evolved each possible edge down the branch of a tree until the leaves in the species tree were reached. We set $p_g = 0.2$, and $p_m = 0.8$ for this process. Once we had the network structures for each species, we used GeneNetWeaver to generate data from each species (Schaffter et al., 2011). GeneNetWeaver uses stochastic differential equations to generate expression data. Specifically, each sample in each dataset represents a steady state measurement after perturbing a node and running the system to steady state. Each dataset consisted of 300 samples.

Yeast Expression Datasets

We applied our algorithm to real expression data from six ascomycete yeast species (Thompson et al., 2013; Roy et al., 2013b; Wapinski et al., 2010), and a new osmotic stress response dataset collected in this work (GSE94628). These data measure gene expression for six species, namely *S. cerevisiae, C. glabrata, S. castellii, K. lactis, C. albicans* and *S. pombe* in four stresses: glucose depletion, heat shock, osmotic stress, and oxidative stress (oxidative stress data was not available for *S. pombe*). A total of 35 measurements were used for *C. albicans* and *C. glabrata*, 30 measurements were used for *S. cerevisiae, S. castellii*, and *K. lactis*, and 21 measurements were used for *S. pombe* for which oxidative stress data was not available. In addition to the phylogenetic priors, our study in yeast included species-specific sequence motifs identified using the Cladeoscope algorithm, developed by Habib et al (Habib et al., 2012). We learned regulatory networks using a gene set drawn from 6,547 orthogroups, which included genes with complex orthology relationships and many duplication levels. 459 of these orthogroups contained at least one potential regulator in at least one species.

Evaluation of Learned Networks

We assessed the effectiveness of network reconstruction using the MRTLE approach by comparing against two baseline approaches described above, GENIE3 and INDEP, on both simulated and real expression data.

Experiments on Simulated Data

GENIE3

We downloaded GENIE3 from http://homepages.inf.ed.ac.uk/vhuynht/software.html. We ran GENIE3 on the entire dataset of 300 samples. GENIE3's internal ensemble framework automatically generates confidence estimates on individual regulatory edges. GENIE3 has two main parameters: the number of trees, n_b, and the number of features to be used at each split, K. We tested multiple configurations for each parameter: $n_b \in \{100,500,1000,1500\}$, and $K \in \{sqrt,all\}$, where *sqrt* uses the square root of the number of regulators, while *all* will select all the regulators. For each configuration of these parameters, GENIE3 will output a confidence value of the presence of a regulatory edge for all potential edges. To select a particular configuation we used AUPR (described below in Evaluation metrics). We found the configuration of n_b=1500 and K=*all* to give the best AUPR and used the network inferred from this setting for our downstream evaluation. However, the overall performance of GENIE3 was stable across different parameter configurations (Figure S6A).

INDEP

INDEP was run within a stability selection framework, where a network was learned on one of 50 random subsamples of data containing 150 samples each. This allowed us to compute a confidence for each regulatory edge defined by the fraction of data subsets for which the edge was selected. The INDEP algorithm has two parameters that control the influence of the prior distribution: β_0 for controlling the sparsity of the inferred network and β_1 to control the influence of the sequence-specific motifs. In the simulation case β_1 was set to 0. We tried different parameter configurations of $\beta_0 \in \{-0.6,-0.8,-1.0,-1.2,-1.4,-1.6,-1.8,-2.0,-3.0,-4.0,-5.0\}$. As in GENIE3, we used AUPR to select the best setting. We found β_0 = -3.0 to give the best performance, however the overall performance of INDEP was stable across different parameter configurations (Figure S6B).

MRTLE

Similarly to INDEP, MRTLE was also run within a stability selection framework, with 50 random subsamples of the data each comprising 150 samples. MRTLE has multiple parameter configurations: p_g for controlling the probability of an edge to be gained in the child species, p_m to control the probability of maintaining an existing edge, β_0 for controlling sparsity, and β_1 for controlling the influence of the motif prior. As in the INDEP case, we set β_1=0. We tested different configurations for MRTLE p_g= {0.1,0.2,0.3,0.4}, p_m={0.7,0.8,0.9} and β_0={−0.6,−0.8,−1.0,−1.2,−1.4,−1.6,−1.8,−2.0}. We found β_0=−2.0, p_g=0.2, p_m=0.9, to give the best results, however, the performance of MRTLE is stable across different configuration settings (Figure S6C).

Evaluation in Real Expression Setting

The evaluation proceeded in the same way as in the simulation case where we tried different parameter configurations and selected the one with the highest AUPR.

GENIE3

We ran GENIE3 with different values of the features used per split, $K \in \{all,sqrt\}$, and number of trees, $n_b \in \{100,500,1000,1500,2000\}$. We selected the best configuration based on the AUPR performance on the MacIsaac gold standard available for *S. cerevisiae* (Macisaac et al., 2006). This configuration was K = *sqrt*, and n_b = 500, but GENIE3 was quite stable to different parameter configurations (Figure S7A).

INDEP

To infer the networks, we used a stability selection framework, where we divided the expression datasets into 25 equal partitions each consisting of 20 measurements of the available stress response measurements. In *S. pombe*, for which oxidative stress data was not available, we partitioned the data into subsamples consisting of 14 measurements. We then inferred networks using each of the 25 data partitions, and calculated a confidence for each regulator-target interaction for each species, by calculating the percentage of the 25 networks that each edge was present in.

We used the "with motif" case to determine the optimal parameter configurations. Specifically, we set the sparsity parameter $\beta_0 \in \{-1,-2,-3,-4,-5\}$ and the motif parameter $\beta_1 \in \{1,2,3,4,5\}$ (Figure S7B). We used AUPR computed on the Hu et al. dataset from *S. cerevisiae* (Davis and Goadrich, 2006), to determine the best setting (similar strategy as in GENIE3), and found β_0=−5.0 and β_1=5.0 to give the best results.

In the case where motifs were not used, we need only to specify the sparsity parameter, β_0. We selected β_0=−5.0, because this was the configuration that was ideal for the motif case. We checked the sensitivity of INDEP to multiple settings of β_0: $\beta_0 \in \{-1.0,-2.0,-3.0,-4.0,-5.0\}$. INDEP results were very stable across different β_0 values (Figure S7C).

MRTLE

Similarly to INDEP, we used a stability selection framework to learn regulatory networks with MRTLE. We used settings of p_g and p_m that were derived from the species tree branch lengths, and used $\beta_0 \in \{-0.8,-0.9\}$ and $\beta_1 \in \{3,4,5\}$. We ran MRTLE with these configurations using only those target orthogroups without duplications, and computed AUPR on the Hu et al. dataset. We found the AUPRs to be very stable (Figure S7D), however, β_0=−0.9 and β_1=4 gave the best AUPR, and we used this configuration in all further analyses.

Evaluation Metrics

We used different evaluation metrics to assess the quality of the inferred networks.

Area Under the Precision Recall Curve

On both simulated data (all species) and real expression data (*S. cerevisiae*), we compared the inferred networks with the true networks based on Area under the precision recall curve (AUPR) computed using the *aupr* tool from Davis and Goadrich. (Davis and Goadrich, 2006). Precision is defined as the ratio of true positives to the total number of predicted edges. Recall is defined as the ratio of the number of true positive edges to the number of true edges. To compute the precision-recall curve, we need to estimate precision and recall at different confidence thresholds for edges. For MRTLE and INDEP, we obtained these confidences using stability selection. That is, we generated random subsamples of the data, learned a network from each subsample, and computed a confidence for each edge representing the fraction of inferred networks in which the edge was present. GENIE3 has its own bootstrap procedure during the Random Forests learning procedure and directly outputs a confidence for each edge. The area under the precision-recall curve gives an overall assessment of the quality of the inferred networks.

Pattern of Phylogenetic Conservation

We assessed the quality of the inferred regulatory networks using the extent of inferred conservation and the ability to capture *phylogenetically coherent patterns of conservation* between species. A pattern of conservation is said to be phylogenetic if it obeys the phylogenetic structure, that is, networks for species that are close on the phylogeny should exhibit greater similarity than networks of species that are further apart. We used an F-score measure to assess the similarity between pairs of networks, where F-score is defined as the harmonic mean of precision, P, and recall, R, F-score $= \frac{2*P*R}{P+R}$. This required us to specify a network at a specific confidence threshold for each species. For the simulated data we picked these thresholds to obtain $\approx$ 3,000 edges. For the real data, we picked thresholds to obtain $\approx$ 30,000 edges.

In the simulation setting, since we had access to the true networks, we could additionally directly assess the extent of conservation and divergence present, and compare this to the conservation and divergence present in the inferred networks. This comparison was done by defining the predicted common edges between two inferred networks and comparing to true common edges using AUPR.

Evaluating Regulator-Target Edge Predictions in S. cerevisiae

To evaluate our networks inferred for *S. cerevisiae* when motifs were withheld, we used a ChIP-chip derived TF-target gene network from Macisaac et al., 2006, which has previously been used as a gold standard in the field (Marbach et al., 2012). When evaluating the full power of MRTLE with motifs included into its prior formulation, we used a dataset from Hu et al., which was constructed by systematically examining the genome-wide expression profile in 268 individual deletion strains, each strain representing a transcription factor (TF) (Hu et al., 2007). The regulatory network was defined using a two step approach. First, an initial network was defined as the total set of significantly differentially expressed genes in each deletion strain. Second, this network was refined using a regulatory epistasis approach in an effort to remove indirect interactions. See Hu et al (Hu et al., 2007) for details.

Evaluation of Regulator-Target Edge Predictions in Non- S. cerevisiae *Species*

The evaluation of edges in the non-*S. cerevisiae* species was done using available ChIP-chip datasets for a handful of transcription factors, namely MCM1 from (Tuch et al., 2008) and CBF1, HMO1, FHL1, IFH1, and RAP1 from (Lavoie et al., 2010). We evaluated the quality of the inferred interactions for these TFs based on AUPR and fold enrichment of ChIP-based targets in the MRTLE inferred networks. Fold enrichment is defined as the ratio of the observed over expected proportion of true edges as follows:

$$\frac{(\#\ of\ true\ positive\ targets)\ /\ (\#\ of\ predicted\ targets)}{(\#\ of\ actual\ targets)\ /\ (\#\ of\ genes\ in\ dataset)}.$$

Inference of Stress-Specific Regulatory Networks for Multiple Species

To define the regulatory network for each stress, e.g., Osmotic stress response, we used the Arboretum algorithm to first define five transcriptional modules as described in Roy et al (Roy et al., 2013b). We next filtered the MRTLE regulatory network inferred in each species using the module assignments such that an edge was removed from the network if either of the end points of the edge were in different modules, resulting in a single stress-specific network for each species. We refer to this combined approach as Arboretum+MRTLE.

Assessing a Regulator's Role as Repressive or Activating

We used two measures to assess whether a regulator acted in a repressive or activating manner. In the first, we used a Hypergeometric distribution to calculate the significance of overlap between the MRTLE inferred targets of a regulator and a transcriptional module. A regulator, r's association with a repressed or induced module was quantified based on the difference, $-\log(p-\text{value}_{ACT}) - (-\log(p-\text{value}_{REP}))$, where $p-\text{value}_{ACT}$ and $p-\text{value}_{REP}$ are the Hypergeometric test p-values obtained when testing for enrichment of r's targets in the most induced or most repressed module, respectively. Regulators with negative values for this measurement were considered to be repressive regulators, while positive values indicated an activator. In the second analysis, we directly compared the expression of the targets of a regulator in the induced and repressed modules based on a one-sided T-test for each time point. We required a regulator to have at least 5 targets in one module (e.g. most induced) and at least 2 targets in the other (e.g. most repressed) module and tested whether the targets in the induced module were significantly higher than the repressed module. Next, to assign a sign to a regulator, we used the difference in the number of targets in each module; if a

regulator had more targets in the repressive module, it was considered as a repressor, whereas, if it had more targets in the induced module, it was called an activator. This too gave a single statistic that could be used to assess if a regulator was mostly repressive or activating.

Defining Targets of Selected Regulators based on LIMMA

To validate predicted targets of key transcription factors we measured mRNA levels using miseq, and utilized the LIMMA software (Smyth et al., 2005), applying it to salt stress data sets in wild type and mutant strains of yeast. Here a wild type (Scer.WT) and an MSN2/MSN4 knockout mutant (Scer.msn2.4) *S. cerevisiae* strain were used to define the targets of MSN2/4, and a wild type and SKO1 mutant strain of *C. albicans* were used to define the targets of SKO1 in *C. albicans*. For *S. cerevisiae* two replicate RNA-seq experiments were performed for each of 4 conditions for both wild type and mutant: (1) T=0 minutes under no salt stress (BMW.T0), (2) T=20 minutes under no salt stress (BMW.T20), (3) T=0 minutes under a KCL salt stress treatment (KCL.T0), (4) T=20 minutes under a KCL salt stress treatment (KCL.T20). Using LIMMA, differentially expressed genes were called for MSN2/MSN4 in *S. cerevisiae* with the following contrast functions: (Scer.WT.KCl.T20-Scer.WT.BMW.T0)-(Scer.msn2.4.KCl.T20-Scer.msn2.4.BMW.T0). The rationale for this contrast function is that the genes that are under MSN2/4 control under osmotic stress are those whose expression changes from T0 to T20 in the wild type, but not in the MSN2/4 strain when subjected to the same stress. Similarly for *C. albicans*, the contrast function used was (Calb.WT.KCl.T20-Calb.WT.BMW.T0)-(Calb.sko1.KCl.T20-Calb.-sko1.BMW.T0). A similar contrast function was applied for the other strains as well. While targets were called for other species and regulators, in only the above two cases did MRTLE and the limma analysis both find a sufficient number of targets to allow for enrichment analyses. The LIMMA algorithm results for these two contrasts then provided us with a log-fold change and an adjusted *p*-value (*q* value) measure for the significance of differential expression of each gene. A q-value <0.05 was chosen to select targets of MSN2/MSN4 in *S. cerevisiae* and SKO1 in *C. albicans*. These target lists were then utilized in the downstream analyses. When comparing our MRTLE results to the MSN2/4 double knockout, any gene predicted to be regulated by either MSN2 or MSN4 by MRTLE was considered.

Estimation of Gain and Loss Rates in MRTLE Inferred Network

Rates of target gain and loss were calculated for each regulator orthogroup by modeling gain and loss of targets with a continuous-time Markov process, and using an expectation-maximization (EM) based approach to estimate the rates, as in Hobolth et al. and Garber et al. (Hobolth and Jensen, 2005; Garber et al., 2009). When assessing the rates in MRTLE, three separate sets of rates were calculated (Figures 3E and 3F). The first allowed for many-to-many orthology relationships within a regulator group, constructing rate matrices for each possible mapping when regulator duplications were present (Figures 3E and 3F; **All**). In the second set, we separated the effect of genes lost from the genome from the effect of regulators' targets lost. For this, we calculated rates for each regulator after removing any targets from consideration that did not have uniform, one-to-one orthology (Figures 3E and 3F; Uniform Targets). In the third set, we tested whether double-counting of certain regulators did not bias the results. To address this, we calculated the rates by taking the average rate of each of the possible orthology mappings for a regulator (Figures 3E and 3F; Collapsed Regulators).

DATA AND SOFTWARE AVAILABILITY

The MRTLE software and inferred networks are available at https://bitbucket.org/roygroup/mrtle. The expression datasets generated as part of this study are available in GEO (GSE94628).