Inferring fungal cis-regulatory networks from genome sequences

Alan M Moses 1, Jason E Stajich 2, Audrey P Gasch 3, David A Knowles 4,5

1 Cell & Systems Biology, Computer Science, University of Toronto,
2 Microbiology and Plant Pathology & Institute for Integrative Genome Biology, University of California Riverside
3 Medical Genetics & Center for Genomic Science Innovation, University of Wisconsin, Madison
4 New York Genome Center
5 Systems Biology, Computer Science, Columbia University

Abstract

Gene expression patterns are determined to a large extent by transcription factor binding to non-coding regulatory regions in the genome. However, gene expression cannot yet be systematically predicted from genome sequences, in part because non-functional matches to the sequence patterns (motifs) recognized by transcription factors (TFs) occur frequently throughout the genome. Large-scale functional genomics data for many TFs has enabled characterization of regulatory networks in experimentally accessible cells such as budding yeast. Beyond yeast, fungi are important industrial organisms and pathogens, but large-scale functional data is only sporadically available. Uncharacterized regulatory networks control key pathways and gene expression programs associated with fungal phenotypes. Here we explore a sequence-only approach to inferring regulatory networks by leveraging the 100s of genomes now available for many clades of fungi. We use gene orthology as the learning signal to infer interpretable, TF motif-based representations of non-coding regulatory regions, thus scaling comparative genomics beyond evolutionary comparisons where these regions can be aligned. We show that similarity of promoters in our motif-based representation predicts gene co-expression, and that we can infer known and novel regulatory connections in diverse fungi. Our new predictions include a pathway for recombination in *C. albicans* and pathways for mating and an RNAi immune response in *Neurospora*. Taken together, our results indicate that specific hypotheses about transcriptional regulation in fungi can be obtained for many genes from genome sequence analysis alone.

Introduction

The ability to infer transcriptional regulatory networks from genome sequence alone is a long-standing goal of the post-genomic era: it would imply a basic understanding of the cis-regulatory code that connects regulatory sequences in the genome to gene expression patterns[1,2]. Although sequence-specific transcription factors (TFs) are appreciated as key determinants of transcription initiation [3], applying TF motif-discovery and motif-scanning to non-coding regions of single genomes is plagued by false positives [4,5]. It is long appreciated that evolutionary conservation can be used to increase the signal-to-noise because functional regulatory elements are more likely to be conserved [6–8]. Indeed, analyses of whole genome alignments have identified catalogues of TF binding motifs [9–11] and in some cases have power to identify individual regulatory connections between TFs and their targets[12].

In many clades, including the fungi considered here, regulatory regions diverge at the primary sequence level, even when functional information is conserved[13–16]. This means that functionally conserved non-coding regions may be difficult to align accurately beyond the most closely related species[16–18], so that alignment-based approaches to improving the signal-

signal-to-noise of TF motif-discovery and scanning are limited. Several approaches have been developed to address this issue [17,19–22], but they do not readily scale to the large numbers of genomes currently available[23–25]. Fungal genomes represent a feasible case for regulatory network inference from sequence alone: they contain relatively short intergenic regions and introns, so automatic gene annotation pipelines are accurate, and most genes do not have multiple enhancers[26,27]. Predicting regulatory networks from sequence would have far-ranging applications. For diverse fungi, it is often not practical to perform systematic experiments because aspects of the life cycle are inaccessible in the lab[28]. Furthermore, for human and agriculture pathogens, lab experimental approaches are limited at least by regulations and at worst by dangers[29].

We frame inference of cis-regulatory networks from genome sequence as a representation learning problem[30]. The aim is to find a compact, quantitative and biologically interpretable representation for non-coding sequences[31] that can be used to understand their regulatory function. Using inference methods developed for deep learning, recent work aims to build mechanistic biophysical or bioinformatic models of transcriptional regulation by learning to predict large-scale experimental data [32–35]. Following this approach, we learn 256 TF binding motifs (position weight matrices, PWMs[36]) to represent orthologous sets of promoters.

Our key innovation is to learn a motif-based representation of regulatory regions from genome sequences alone. To do so, we infer PWMs and other parameters using reverse homology, a contrastive approach that leverages evolutionary conservation in the representation space as the learning signal[37,38]. Because our approach requires no direct experimental data, it can be applied to conditions and organisms for which large-scale functional genomics data is not available. In not relying on experimental data, our approach is similar to unsupervised training of large-language-models[39] where motifs and regulatory connections can in some cases be identified through post-hoc analysis [40]. Unlike a language model, however, our approach yields an interpretable motif-based representation of non-coding sequences from which regulatory connections can be read directly.

We find that patterns in the motif-based representations are strongly associated with gene expression and biological function, and that we can identify known and novel regulatory connections from visualizations of the representation space. Using this approach, we identify a putative recombination/parasex pathway in *C. albicans* and predict mating and virus-response pathways in *N. crassa*. Finally, we show that for the relatively unstudied genome of the pathogen *F. graminearum*, the promoter representation reveals a useful regulatory map. Our work demonstrates, in fungi, the feasibility of scaling sequence-based regulatory network inference beyond alignable genomes, and yields rich biology, comparable to functional genomics approaches.

Results

Learning motif-based representations of orthologous sets of promoter regions without sequence alignments

At evolutionary distances where we can still identify orthologs for most genes, fungal upstream non-coding regions (which we refer to subsequently as promoter regions, see Discussion) show conservation of regulatory function, but little sequence similarity. For example, the upstream

regions of *S. cerevisiae PHO84* and its orthologs contain many matches to the Pho4 consensus site (Figure 1a), but these are only alignable in very closely related species (Figure 1a, Saccharomyces *sensu stricto*, See Methods). The presence of motif matches in many orthologous promoters is unlikely to occur by chance, and therefore improves the signal-to-noise of motif-scanning and motif-discovery[10,19,22,31]. We sought to design an approach to exploit to exploit this improved signal-to-noise available in multiple orthologous promoter regions, without relying on sequence alignments.

To detect conservation of motifs without alignments, we first use a motif-based encoder to convert each set of orthologous promoter regions into a fixed-length representation vector (or "embedding"), where each dimension corresponds to a single motif[31]. We followed recent work[35] showing that motif match cutoffs, match counting and motif interactions can be modelled explicitly using trainable scaling, pooling and attention layers, respectively (Figure 1b, See Methods). We ensured that the model learns *bona fide* PWMs[36] by constraining the probabilities at each position to add up to 1 (see Methods). The resulting biologically interpretable neural network[35] embeds (unaligned) individual sequences in a space of 256 learned motifs. To infer the parameters of this model, we used reverse homology ([38], see Methods). This is a contrastive learning approach that uses orthologous sequences alone to develop a representation that maximizes mutual information between homologs[37]. Reverse homology trains the model to classify orthologs vs. non-homologous sequences by comparing individual sequence representations to an estimate of the phylogenetic average representation of its orthologs (see Methods). This procedure aims to learn similar representations of orthologs and different representations of unrelated sequences. In the case of our motif-based encoder, it encourages learning of motifs that can be identified in individual sequences and match the conserved motifs in the orthologs. Once the parameters of the encoder have been learned, to measure evolutionary conservation in the motif representation space, we simply average each of the motifs' scores over the orthologs (Figure 1b). We refer to this as phylogenetic averaging[31]. Thus, using reverse homology (RH) we train an interpretable encoder (IE) and obtain a representation for a set of orthologous upstream regions by phylogenetic averaging (PA). We abbreviate this as RHIEPA.

To first test the power of motif-based embedding learned through reverse homology (RHIEPA) on a well-studied regulatory network, we obtained orthologous upstream sequences for *S. cerevisiae* promoters from 89 species from the y1000+ project[24] (see Methods) yielding nearly 380,000 non-coding sequences. This represents an order of magnitude more data than classical comparative motif-finding approaches[9], but an order of magnitude less than used to train large-language-models[40] (See Discussion). We found that the PWMs learned by the model qualitatively agree with known motifs for TFs from the JASPAR fungi database [41] (Figure 2a), including both long 'easy' motifs (Figure 2a, matrix 81/Sfp1, matrix 66/Mcm1), but also short motifs with little information content (Figure 2a, matrix 29/Crz1). Like known transcription factor binding motifs, the learned PWMs showed overlapping sequence preferences (Figure 2a, matrix 121/Swi4, matrix 83/Mbp1::Swi6). Overall, 63 of the 256 learned PWMs match known motifs when we compare to two current databases ([41,42], Figure 2b) using the MEME Suite's TomTom motif database search tool ([43], see Methods). 63 known motifs is nearly half of the number that match when we search the databases using a expert curated set of PWMs for yeast [44] (based on 100s of *in vitro* and *in vivo* experiments related to transcription factor specificity) (Figure 2b). This indicates that the motifs learned from sequences alone contain a

reasonable fraction of the knowledge about the cis-regulatory vocabulary that has been learned from decades of direct experiments.

Location in the learned motif space is associated with gene expression patterns and gene function.

We next sought to assess the biological information within the motif-based phylogenetic average representations of promoter regions (RHIEPA, Figure 1c). To do so, we first compared the promoter representations to gene expression patterns over more than 1400 yeast gene deletion experiments[45]. We identified nearest neighbours (using cosine distance, see Methods) for each promoter region in the RHIEPA representation space and used them to predict relative expression of the downstream gene for individual experiments (using 1-NN regression, see Methods) and to predict co-expressed pairs (classification of downstream gene pairs with correlation >0.5 over the 1487 experiments). These represent measures of prediction of between-gene and between-condition variation, respectively[46], based on the promoter sequence alone. We emphasize we did not train any parameters using the gene expression data or include any prior knowledge to make predictions. This is in contrast to current regulatory network inference approaches for well-studied cells that leverage prior knowledge about transcription factors, their motifs, as well as large-scale data for gene expression, chromatin and/or transcription factor binding [47–49]. We compared the predictions of gene expression and co-expression to predictions using nearest neighbours obtained from the regulatory network inferred by NetProphet3[49] (see Methods). Averaging over all the experiments, RHIEPA representations of promoters predict gene expression with R=0.22 (s.e. 0.0026), greater than R=0.12 (s.e. 0.0012) for NetProphet, and far greater than R=0.01 obtained using random neighbours. Similarly, over all genes, the RHIEPA nearest neighbour is co-expressed in the experiments for 768 genes (14.3% or 0.14 precision, n=5353 genes total, Figure 2c), more than 236 genes (4.4% or 0.04 precision, n=5364 genes total, Figure 2c) for NetProphet, and far more than the 72 genes (1.35%) observed for random neighbours. Because the classification of co-expressed gene pairs is outside of the training data of the sequence-based RHIEPA, this is a measure of zero-shot generalization.

Furthermore, simply using the phylogenetic average of motif scores [31] of expert curated PWMs[44] also yields higher correlations with expression than both NetProphet3 neighbours and experimental binding data (Figure 2c). However, we find that matches to the expert curated PWMs in *S. cerevisiae* alone have little power, consistent the idea that matches to motifs in single genomes are mostly non-functional noise (the so-called futility theorem[4]). Since single nearest neighbours explores only highly local relationships in the high-dimensional space, we repeated these experiments for 50 nearest neighbours and found similar results (figure 2c), indicating that the similarity in the motif-based space extends beyond pairs of promoters, as expected if the representations are grouping together co-regulated pathways. Taken together these results suggest that much of the power to predict gene expression (relative to transcription factor binding data and methods based upon it) is due to the phylogenetic averaging, which we believe increases the signal-to-noise ratio.

We next visualized the RHIEPA representation in 2 dimensions using UMAP (see Methods), where each point corresponds to the 256-dimensional vector for a promoter (and its orthologs, Figure 1c). We coloured the points representing promoter regions matching English text in the associated gene names or descriptions (see Methods) and found that promoter regions flanking genes with the same matches to their names and descriptions were nearby each other (Figure

2d). We emphasize that only genomic sequence information is used in training the model or making the UMAP, and the colouring is not based on clustering. The adjacency of promoters with the same text in their gene names and descriptions strongly supports the idea the motif-based representation contains functional representations of the promoters.

Taken together these results show that training a motif-based representation on genome sequences alone identifies known motifs, predicts gene expression patterns and groups promoters according to their biological functions.

Reading the cis-regulatory code by visualizing the motif-based representation

Given the associations of the RHIEPA representation with gene function and gene expression, we next sought to explore whether we could understand how the presence of motifs determines promoter function (the cis-regulatory code). We clustered and visualized[50] the *S. cerevisiae* promoter regions based on their RHIEPA representations and named motifs using matches to the JASPAR fungi database (See Methods, Figure 2a). We then identified patterns of motifs that were strongly associated with biological functions. For example, consistent with known regulation [51], a cluster of 236 promoters with strong signals for Dot6, Stb3, Sfp1 and Abf1 is associated with genes is strongly enriched for ribosome biogenesis (126/236, 9.6x enrichment, P= 5.45E-133, Figure 3a) and is distinct from a cluster of 199 promoters that does not show signals for Stb3, but does show signals for the other factors, and is associated with genes enriched for translation machinery (61/119, 6.8X enrichment, P=4.09E-35, Figure 3a). This cluster includes promoter regions for t-RNA synthetases, which additionally show a signal for a motif matching Gcn4 (known to regulate amino acid biosynthesis[52]) consistent with a recent report of activation of Gcn4 when tRNA synthetase activity is low[53] and more generally, a connection between regulation of amino acid biosynthesis and translation. Also consistent with known regulation[51], a cluster of 124 promoters associated with ribosomal proteins (82/124, 24x enrichment, P=3.10E-104, Figure 3a) are distinguished by signals for a learned motif matching Rap1. Finally, a cluster of 260 promoters associated with mitochondrial translation (36/260, 6.7x enrichment, P=7.56E-18, Figure 3a) also shows signals for Sfp1 and Abf1, but also Nsi1 (which is similar to Reb1). Thus, the combinatorial control of translation and associated cellular machinery can be read from this visualization of the motif-based promoter representation (see Discussion).

To further explore whether the signals in the RHIEPA representation show evidence of combinatorial regulation, we next examined a large cluster of *S. cerevisiae* promoters showing signals for many of the known cell cycle regulatory transcription factors and enriched for promoters regions upstream of cell cycle genes (262/623 promoters, 3.2x enrichment, P=1.93E-79). Comparing the cluster of promoter regions to gene expression measurements for the associated genes (Figure 3b) revealed that the groups of promoters with different patterns of transcription factors correspond to groups of genes with coordinated gene expression patterns in classic cell cycle synchronization experiments[54] (Figure 3b, full comparison in Supplementary Figure 1a) and correlated cell heterogeneity in sc-RNA-seq[55] (Supplementary Figure 1b). These groups of genes showed different functional enrichment (Figure 3b) consistent with regulation by different combinations of cell cycle transcription factors that control different stages of the cell cycle. For example, a cluster of promoters with signals for learned motifs matching Mbp1::Swi6 shows enrichment for genes involved in DNA replication (Figure 3b), while a cluster with signals for motifs matching Hcm1 and Fkh1 shows enrichment for

genes involved in chromosome segregation (Figure 3b). These observations are consistent with the distinct cell cycle functions of these transcription factors[56,57].

We also observed that TF-TF regulatory connections could be inferred while analyzing this cluster of promoters associated with the cell cycle: of the 12 learned motifs that matched JASPAR (Figure 3b, red text), the promoters of genes encoding 8 (66%) of the corresponding TFs (Figure 3b, purple text) were found in the cluster. Since less than 15% of the genes in the genome are in the cluster, this represents a strong statistical enrichment (P<0.001, Fisher's exact test) and is consistent with the known regulation of the cell cycle TFs by each other[58].

Further supporting the idea that TF regulation is visible in the sequence representation, we found we could infer examples of autoregulation in the *S. cerevisiae* promoter map. For example, the promoter region of the gene encoding the iron-responsive transcription factor Aft1[59] was found among promoter regions of transition metal transporters (31/47, 60x enrichment, P=1.19E-49) in a cluster with strong signals for learned motifs that matched the Aft1 and Aft2 motifs in JASPAR (Figure 3c shows a subset of this cluster). Although to our knowledge auto-regulation of *AFT1* has not been reported, the promoters of *AFT1* orthologs contain matches to the Aft1 consensus (Figure 3d), consistent with binding of Aft1 to its own promoter. This prediction is also consistent with a peak of Aft1 binding observed at the AFT1 promoter (Supplementary Figure 1c) under standard growth conditions [60]. More generally, we found at least 9 (Figure 3e) examples of TF promoters showing strong signals for the learned motifs that matched their own motifs in JASPAR and were clustered among their known target promoters in the motif-based sequence representation. Of these, at least 5 are known (Gcr1[61], Upc2[62], Ndt80[63], Zap1[64] and Yrr1[65]), while 4 (Aft1, Lys14, Crz1 and Hap1) appear to be new predictions of TF autoregulation.

Taken together, the cluster analysis of the phylogenetic averages of learned motif representations for the well-characterized regulatory network of budding yeast supports the idea that we can infer regulatory networks from genome sequence alone.

Prediction of a meiosis/parasex pathway in C. albicans

Because the regulatory inference based on genome-sequences is unbiased by experimental conditions or lab constraints, we next tried to infer a regulatory network for *C. albicans*, a distantly related pathogenic fungus with many shared genomic features to *S. cerevisiae*. We obtained orthologous promoter sequences for *C. albicans* from 69 species from the Y1000+ project (see Methods) again yielding more than 300,000 non-coding sequences. We inferred the parameters of interpretable encoder using reverse homology with no changes to hyperparameters and created representations for the promoters by phylogenetic averaging (RHIEPA). Consistent with the relative paucity of experimental measurements in *C. albicans*, many fewer of the learned motifs matched known motifs in the databases relative to *S. cerevisiae* (Figure 1b). Nevertheless, we observed many of the same patterns in the representation space, where clusters of promoters showed strong functional enrichments (Figure 4a) and some of the known differences in regulation, such as signals for Tye7 in a cluster of promoter regions of glycolysis genes[66] (9/9, >300x enrichment, P=1.15E-21) and absence of Rap1 motifs in promoter regions associated with the ribosome genes [67] (59/71, 34x enrichment, P= 1.40E-85). Although it has no significant match in the database, a learned motif (matrix 133) associated with the promoters of the ribosome cluster matches the reported

consensus for Tbf1[68] (Figure 4a), which is expected to regulate ribosome gene expression in *C. albicans* (along with Sfp1 and Stb3) [67].

Unexpectedly, we identified a group of promoter regions strongly associated with genes annotated to have a function in the meiotic cell cycle (Figure 4b). Meiosis has never been observed in *C. albicans*, although *C. albicans* is known to have several of the homologs of *S. cerevisiae* meiosis genes and undergoes a recombination process known as parasex[69]. In this cluster, we also found upstream regions of homologs of yeast meiosis regulators (*UME7*, *CSK1* and *IME2*). The motifs with strong signals in this cluster (Figure 4b) did not match any known motifs in JASPAR. We suggest that they represent the regulators of parasex in *C. albicans*, and the (at least) 14 genes of unknown function in this cluster are candidates for genes involved in meiosis/parasex.

Inferring regulatory networks for multicellular fungi

Encouraged by the analysis of unicellular fungi, we next attempted to apply this approach to more complex organisms. *Neurospora crassa* is a genetic model system for filamentous fungi. Its genome is 40MB (compared with 12MB for budding yeast) and contains approximately 10,000 genes (compared with approximately 5,000-6,000 for budding yeast) nearly half of which are annotated as encoding unknown proteins[70]. We obtained orthologs from the 52 closest genomes (see Methods), again yielding more than 300,000 non-coding sequences and with no changes to hyperparameters created a motif-based RHIEPA representation. The motifs learned matched more known motifs in the Neurospora-specific cis-bp database than JASPAR fungi (37 vs. 27) consistent with many of these specificities being predicted from *in vitro* TF binding experiments[42].

As with yeasts, we found that clusters of promoter regions with similar signals in the RHIEPA representation showed strong association with essential cellular functions (Figure 5a). For example, we observed a signal for a matrix that matches JASPAR fungi Gat1 in a cluster of promoters associated with nitrogen transport (30/189, 3.2x enrichment, P= 7.99E-06), consistent with previously observed conserved regulation[71]. As in yeasts, we find separate clusters of upstream regions for ribosome biogenesis genes (109/233, 18x enrichment, P=8.66E-113), translation factors (19/68, 39x enrichment, P=7.86E-24) and ribosome components (53/81, 38x enrichment, P=2.98E-74). Ribosome biogenesis promoters include signals for Sfp1 (as in *S. cerevsisiae*) but also include signals for a GATA factor motif that matches Gat1, suggesting different mechanisms may regulate growth in Neurospora.

We find a cluster of ~500 promoters strongly associated with the cell cycle (163/505, 6.1x enrichment, P= 1.16-86) that share some motifs with yeast (Mbp1[71] and Fkh2) alongside signals for different motifs (Tye7 and Rfx1), suggesting different regulation. In contrast to the well-studied budding yeast cell cycle, 40.6% of the genes in the cell cycle cluster in *Neurospora* are annotated as encoding “hypothetical protein” and little is known about the transcription factors that control the *Neurospora* cell cycle. Thus, our analysis provides hypotheses about the identity of cell cycle regulators in this species.

Other clear signals include a cluster of promoters strongly associated with peroxisome genes (21/96, 32x enrichment, P=9.57E-25) shows a strong signal for matrix 234 that has no matches in the databases, but matches the CCGAGG consensus of FarA/B, transcription factors that regulate the peroxisome genes in Aspergillus[72]. Furthermore, the promoters of proteasome subunits contain signals for unknown motifs (matrix 145, Figure 5a) consistent with previous

reports of evolution of proteasome regulation in fungi[71]. Similarly, upstream regions associated with Golgi transport, cell-wall components and genes involved in catabolism also show unknown motifs (matrix 88, matrix 56, matrix 133, Figure 5a). These represent new hypotheses for basic regulatory mechanisms.

Beyond basic cellular mechanisms, we identified a group of 165 promoter regions with signals for a matrix that matched ASL-1, a regulator of osmotic stress and circadian rhythm gene expression [73]. The genes associated showed no functional enrichment (though 55% are annotated as hypothetical protein) but do include 3 of the 7 known ASL-1 targets[74] (cat-1, bli3, and ccg-1, 27x enrichment, P = 0.00012, Fisher's exact test). To test for circadian rhythm gene expression, we counted genes with both significant RNA and RNAPII-S2P rhythms [75]. We found 16 of 67 (15x enrichment, P = 2.85e-15, Fisher's exact test) associated with the promoter regions in this cluster, indicating that many of the promoters in this group may be uncharacterized direct targets of ASL-1.

To confirm that similar promoters in the RHIEPA representation space showed similar gene expression patterns, we next compared the sequence-based clustering of Neurospora promoters with large-scale gene expression measurements for growth on varying carbon sources[76] (Figure 5a). Indeed, groups of promoters with motif signals matching known carbohydrate regulators (vib-1, xlr-1 and NCU04848, whose motif is similar to CRE-1[76]) correspond to groups of genes with strong induction of gene expression and the genes associated with these promoters are also strongly enriched for genes involved in carbohydrate metabolism and transmembrane transport (10x and 7.4x enrichment, P=6.74E-51 and P=1.82E-28, respectively). We also noted strong expression changes for the ribosome and ribosomal biogenesis genes, perhaps consistent with differences in growth on the different carbon sources (Figure 5a). Interestingly, the cluster of promoters associated with rhythmic expression shows somewhat reciprocal gene expression changes (Figure 5a). Although it is tempting to speculate that variation in carbon source may influence circadian transcription in Neurospora, we cannot rule out simple time-of-day effects.

Taken together these show that we can identify known and predict new regulation for known genes in Neurospora, thus inferring a regulatory network in a multi-cellular organism from genome sequences alone.

Prediction of mating and RNA-virus response pathways in Neurospora

We next searched for possible new pathways among the clusters of upstream regions associated with large numbers of uncharacterized genes in Neurospora and found several compelling examples. We found a cluster of promoter regions associated with signals for a motif matching fmf-1 (Female and male fertility protein 1) and an unknown motif (matrix 191) (Figure 5b). Although >80% of the genes in this cluster have unknown function, it includes MatA-1, MatA-2/3 (Mating-type proteins), both Male-barren genes and the pheromone receptor genes, Pre-1 and Pre-2 (Figure 5b). We therefore speculate that this represents the (largely uncharacterized) mating pathway in Neurospora (7/214 are annotated as response to pheromone, 15x enrichment, P=9.92E-05). Consistent with this, the motif for fmf-1 also matches the CTTTG consensus for MatA-1[77] and the group of genes includes Ccg-4, the only confirmed target of MatA-1[74]. We also noted a group of ~500 upstream regions with signals from unknown motifs associated with mostly uncharacterized genes (87% unclassified). This group includes several of the genes believed to be involved in recombination in Neurospora

(e.g., sad-1[78], sad-2[79], tsn, sms-2 [80], mlh-2 and msh4), so we speculate that this represents the meiosis pathway. Consistent with this hypothesis, gene expression of both the putative mating and meiosis clusters is strongly upregulated during sexual development (last columns of life cycle RNA-seq[81], Figure 5a).

Perhaps the most interesting uncharacterized pathway is a cluster of promoter regions with signals for matrix 110 (Figure 5c). Although 87% of these genes are unclassified, they include dicer-like 2 and rrp3 (an RNA-dependent RNA polymerase) that were both induced in response to viral infection[82]. We propose that this pathway represents an RNAi-based immune response. Like the putative parasex pathway in *C. albicans*, these results show how clusters in the motif-based promoter representation space lead to hypotheses about new pathways in difficult to study experimental conditions (see Discussion).

A partial regulatory map for Fusarium

To explore how far these results can be extended, we attempted to infer regulation from a RHIEPA representation for a much less characterized organism, *Fusarium graminearum*, an important agricultural pathogen. We obtained orthologs for 12,710 genes from 68 species (see Methods), but noted that a large number of orthologous upstream sequences were very similar at the DNA level, possibly reflecting the large number of closely-related Fusarium strains that have been sequenced[83]. We therefore filtered the orthologs to remove very similar sequences (see Methods) and were left with nearly 280,000 upstream sequences.

Given the much sparser functional annotation of the *F. graminearum* genome, we found fewer clusters of promoter regions that were associated with known functions. Nevertheless, we observed functional enrichments similar to Neurospora for some aspects of basic cellular machinery (Figure 6a) such as Sfp1 signals in promoters of associated with the ribosome genes (40/80, 51x enrichment, P= 8.87E-59), FarA/B-like signals in promoters associated with the peroxisome (15/145 19x enrichment, P= 1.33E-19) and promoters with signals for several motifs, including one similar to CRE-1 associated with carbohydrate metabolism and transport (4.8x and 3.6x enriched, P= 8.11E-26 and 1.83E-32, respectively). As with Neurospora, we could predict unknown pathways: a large cluster of promoters, some of which show signals for a learned motif that matches fmf-1, includes the *F. graminearum* mating type proteins, as well as orthologs of several Neurospora genes involved in mating and meiosis (Figure 6a “mating/meiosis”).

We next tested whether the RHIEPA promoter representation contained information about the pathways controlling virulence in *F. graminearum*. To do so we reduced the dimensionality of the promoter representations to 2D (see Methods) and indicated promoters of 165 genes annotated as showing “reduced virulence” phenotypes ([84] see Methods). These promoters were strongly enriched in a region of the RHIEPA representation space distinct from the promoters associated with basic cellular processes (Figure 6b, black symbols vs. colours). This region of the representation space corresponds to a large number of promoters with relatively weak signals for learned motifs i (Figure 6a), but a 3.8-fold enrichment for genes annotated as signaling (78/1650 of the 1688 where we could map genes, P=4.62E-25). We found that these promoters were upstream of 71 (43%) of the genes annotated as reduced virulence. Since this group represents 13% of the upstream regions in the genome, this is a highly significant enrichment (Figure 6c, Fisher’s exact test, P < 2.2e-16). Although it is tempting to speculate that promoter regions can predict virulence, given the enrichment of annotated signaling proteins in

this group, we cannot rule out that the enrichment is due to bias in researchers to study the effects of mutations in signaling proteins on virulence. Given the relatively weak signals in the promoters of these genes, we did not identify an obvious transcriptional regulatory pathway that controls signaling proteins or virulence.

Discussion

Our approach is able to generate regulatory maps for a variety of ascomycete fungi from genome sequences alone, without experimental data or whole genome alignments. For budding yeast, this allowed us to scale inferences of regulatory connections well-beyond the closely related *Saccharomyces sensu stricto* where genome alignments are available [85]. For the relatively well-studied species, we could take advantage of the large collections of known TF binding motifs to validate the learned motifs to assign TFs to them. In some cases, we could understand the combinations of factors that control basic cellular processes, (such as the cell cycle and translation, Figure 3a) directly from the clustering of signals for motifs in the promoter regions. We obtain new insights into poorly characterized species, including a plant pathogen difficult to study in the lab. We predict both novel regulatory sequences and also groups of potentially co-regulated genes, including uncharacterized genes that may function in novel pathways.

Our unsupervised approach allows us to study regulatory function without relying on training data relating to biological function. Even with annotations or functional data, for pathways with small numbers of genes, such as glycolysis (24 genes annotated in *S. cerevisiae*) it would not be feasible to reliably learn the parameters of a motif of width 15 [5] or train a model to predict expression patterns from sequences in the genome[86]. By using orthologous non-coding sequences from more than 50 genomes to learn the motif-based representation, we augment the number of glycolysis promoters to more than 1000 [37]. By clustering and visualizing the representations, we can focus on strong patterns. Indeed, we identified a cluster of 16 promoters that contains 12 of the annotated glycolysis genes (Precision = 12/16 = 75% and Recall = 12/24 = 50%). Furthermore, these upstream regions contain signals for Gcr1, the known regulator of glycolysis as well as the upstream region for the *GCR1* gene, consistent with the known autoregulation[61].

An alternative approach to unsupervised learning from genomes is using large-language-models[39], one of which identified 51 motifs in *S. cerevisiae* 5' regions through post-hoc motif-finding analysis, of which 18 were matched to known motifs [40]. Although 18 is far fewer than the 63 learned PWMs matched to known motifs here, it is not clear whether the difference is an advantage of a motif-based encoder relative to post-hoc motif-finding, or some other limitation of the large-language-model. More clear, however, is that the motif-based models here are trained on less than 300 million bases and have 80 thousand parameters, while the 5' region large-language-model was trained on 13 billion bases and has 90 million parameters[40]. Thus, the computational expense of the motif-based approach is at least an order of magnitude less and the model complexity is 3 orders of magnitude less.

For many of the promoter clusters in the RHIEPA representation, we find statistical enrichment of GO annotations comparable to classic gene expression analysis (Figures 3, 4 and 5) without any lab experiments. Because the analysis is based on sequence alone, we can identify pathways in relatively unstudied experimental conditions. This allowed us to identify putative pathways for parasex/meiosis in *C. albicans* (Figure 4), and mating and RNA-virus response in

N. crassa (Figure 5). Thus, we believe that with sufficient evolutionary resolving power, unsupervised sequence-based analysis can represent a powerful alternative to direct functional genomics experiments for discovery in challenging organisms. Of course, where such data exists, future efforts can aim to integrate evolutionary conservation with experimental data[47,87,88].

Despite the successful examples, matching of the learned motifs to the database has limitations. For example, the learned motif in the upstream regions of heat shock promoters (matrix 36 for *S. cerevisiae*, matrix 241 for *C. albicans*), did not match any known motifs, even though it shows the well-known GAA repeats recognized by Hsf1[89]. Moreover, methionine pathway promoters in *S. cerevisiae* show a signal for matrix 233, whose best matches in JASPAR and cis-bp are Tye7 and Rtg3, respectively (E-value 4.83e-04 and 4.11e-03, respectively), TFs with similar specificity to one of the known regulators, Cbf1[90] (which is the second best match for matrix 233 in both databases, E-value 6.55e-03 and 1.96e-02, respectively). Thus, even interpretable motif-based approaches combined with extensive databases for well-studied organisms cannot definitively identify which transcription factor controls the pathways. Nevertheless, like high-throughput functional genomics experiments, the motif-based approach directly provides hypotheses that could be confirmed in simple experiments.

An oversimplification in our work (so far) is that we treat promoter and upstream regions interchangeably. Even in fungal genomes, 5' UTRs contain important post-transcriptional and translational regulatory elements[91] that are likely to be conserved and therefore provide signal for learning. However, because of the strong inductive bias of our encoder to learn transcription factor binding motifs (Figure 1b), we are not currently able to interpret these signals very well, although we do see PUF3 and PUF4 motifs[92] being learned in *S. cerevisiae*. RNA regulation is a possible explanation for some of the large clusters of upstream regions that show functional enrichments, but no strong signals for transcription factor motifs (e.g., Fusarium virulence genes). Future applications of reverse homology could use more sophisticated encoders to model RNA structure and binding motifs[93]. For genomes with large non-coding regions, to apply a motif-based approach, the encoder will need to either operate over a large DNA region, or identify the important regions via segmentation.

Another current limitation of our approach is that it relies on ortholog assignments for the species of interest. Here we chose the species and number of orthologs arbitrarily based on the number of orthologs for each reference genome (see Methods), and we assume that all orthologs are the same. We find that the approach is relatively robust to the number and distance of species chosen[31,38]. However, even within the groups of fungi considered here, current ortholog assignment software is difficult to scale to 100s or 1000s of genomes[94]. In our study, the ortholog finding was more computationally demanding than creating the RHIEPA representations (see Methods). It is already infeasible to apply high-confidence ortholog finding pipelines to the 100,000s of bacterial genomes currently available. This is an advantage of unsupervised training of large-language-models[40] that do not rely on ortholog assignments. Future motif-based representation-learning approaches that can leverage evolutionary conservation without requiring sequence alignment ***or*** ortholog assignments might be even more powerful tools for regulatory network inference.

Methods

Ortholog finding for fungal promoters

Y1000p [24] genomes were obtained from http://y1000plus.org. For filamentous fungi, 440 Sordariomycetes genomes were downloaded from NCBI using datasets tools (https://www.ncbi.nlm.nih.gov/datasets/). We ran OrthoFinder[94] using diamond_unltra_sens, --fewer-files and -p. Example codes for orthofinder runs are available at https://github.com/stajichlab/Y1000_orthologs.

We used NC12 as reference for *Neurospora* and ASM24013v3 as reference for *F. graminearum*. To choose the species to provide orthologs for each reference, we ranked species by how many reference genes had orthologs and then arbitrarily chose a number of species based on the coverage obtained. After ortholog finding, we noted redundant genome sequences and assemblies among the y1000+ genomes and manually removed them. The numbers were *S. cerevisiae*: 90 (1 redundant was filtered) species with at least 4,400 orthologs, *C. albicans*: 69 (7 redundant were filtered) species with at least 5,000 orthologs, *N. crassa*: 52 species with at least 6,500 orthologs, *F. graminearum*: 68 species with at least 8,000 orthologs (before filtering, see below). Ortholog coverage distribution is shown in Supplementary Figure 2. The complete lists of species used are available as supplementary data.

We extracted upstream sequence for each ortholog up to 800 bp or the next annotated coding sequence. For *F. graminearum*, we aligned the orthologs using mafft[95] we filtered out any ortholog sequences that were less than 0.01 subs. per site away from the reference, and using heuristics as for filtering protein orthologs[38].

A motif-based encoder

Inspired by recent advances in fully interpretable encoders based on standard biophysical and bioinformatics models[32–35], we implemented standard bioinformatics motif-scanning steps in an end-to-end, trainable encoder using keras/tensorflow. Given an input of N orthologous sequences of varying lengths, we first pad shorter sequences to a fixed sequence length and then take the reverse complement. We scan both strands with the learned PWMs, reverse the order of scores of the reverse complement, and then take the better score of the two strands at each position. Although not indicated for simplicity in Figure 1b, in these steps we only consider matches in L – W positions, where W=15 is the motif width and, to avoid counting overlapping matches, we take the best match in a 10 nt window. We next scale the scores for each motif using a trainable scaling layer[35] and summarize the number and strength of matches using the trainable pooling layer[35]. This allows the model to choose the cutoff for each motif and decide if single strong matches or multiple weak matches should be considered for each motif. The outputs of these layers are weighted by an interaction matrix, scaled between 0 and 1[35]. We then applied batch normalization[96] which we found stabilized the training of the encoder.

To create a bona fide motif layer based on the standard 1d convolution layer, we implemented a constraint to ensure that a conv1d layer could be interpreted as a PWM. The kernel of a 1d convolution is simply a matrix of size W x C, where C are the channels (in our case C=4 for the 4 nucleotides of DNA represented as a matrix of indicator variables) and W is the width of the kernel (in our case the width of the PWM). Although a PWM can match the shape of a 1d convolution kernel (W x 4), the entries of a true PWM matrix, M, are defined as $M_{ib} = \log \frac{f_{ib}}{g_b}$,

where *i* and *b* index the position and nucleotide, respectively, and *f* and *g* are probability distributions over nucleotides under the motif model and a background model, respectively[36]. Throughout we use *g* = (0.25, 0.25, 0.25, 0.25). To ensure that the probabilities under the motif model sum to 1 (and therefore each position has only 3 degrees of freedom, not 4), each position *i* of M should follow

$$1 = \sum_b f_{ib} = \sum_b e^{\log f_{ib}} = \sum_b g_b e^{\log \frac{f_{ib}}{g_b}} = g_b \sum_b e^{M_{ib}}$$

Where in the last step we took advantage of the assumption of equal background probabilities. Taking logs implies that for valid PWMs, $\log g_b + \log \sum_b e^{M_{ib}} = 0$. We therefore implemented a PWM constraint in Keras, resetting the weights of the one-dimensional convolutional layer as follows:

$$M_{ib} \leftarrow M_{ib} - \log g_b - \log \sum_b e^{M_{ib}}$$

For example, consider the matrix of W=2, $M = \begin{bmatrix} 3 & 0 \\ 3 & 1 \\ 3 & 2 \\ 3 & 0 \end{bmatrix}$. Neither of the columns is a valid position in a PWM (they contain log probability ratios greater than the background, but none smaller). Using base 2 logarithms for numerical convenience we have $\log_2 g_b = -2$, $\log_2 \sum_b 2^{M_{1b}} = 5$ and $\log_2 \sum_b 2^{M_{2b}} = 3$ for positions 1 and 2, respectively. Applying the PWM constraint, we have

$$M = \begin{bmatrix} 3+2-5 & 0+2-3 \\ 3+2-5 & 1+2-3 \\ 3+2-5 & 2+2-3 \\ 3+2-5 & 0+2-3 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 0 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$$

Converting these to motif probabilities using $f_{ib} = g_b 2^{M_{ib}}$ gives $f = \begin{bmatrix} 0.25 & 0.125 \\ 0.25 & 0.25 \\ 0.25 & 0.5 \\ 0.25 & 0.125 \end{bmatrix}$ which are now valid probabilities. We added this PWM constraint to the convolutional layer in our model and confirmed empirically that our learned motifs are PWMs with *g* as defined.

Learning non-coding sequence representations using reverse homology

Reverse homology is contrastive learning using evolution as the training signal[37,38]. It trains an encoder to tell apart members of a family of homologous (in this case orthologous, see above) sequences from randomly selected sequences. We largely followed the procedure suggested in previous work[38], with some minor simplifications. A single sequence is chosen as the 'anchor' (or positive) and it is encoded. n_f of the anchor's homologs are sampled to represent the family, and their embeddings are simply per motif averaged to create the 'family embedding'. Unlike previous work[38], we used family size n_f =4 and the same encoder for the family and the single sequences. We used each training sequence as the 'anchor' once in each epoch. During training, we filtered any sequences that were shorter than 10% of the max sequence length (L=800 bp) and removed any sequences that contain Ns (missing nucleotides). As in previous work[37], we use the infoNCE loss[97], which is the categorical cross entropy, $H(y,p) = -\sum_j^B y_j \log p_j$ where y is an indicator variable defined so that the positive anchor is 1

and all other single sequences are 0, and $p = softmax(\frac{z_F \cdot z_A}{t})$ where · is the dot product of the family embedding z_F and the single sequence embedding z_A. In reverse homology, the infoNCE is using the dot product with the family embedding to classify the anchor as a homolog relative to all other anchors in the batch (which are negatives for that family). Throughout, we used temperature t=0.1 and batch size B= 256. We also applied reverse complement augmentation with probability 0.5, even though our encoder does consider both strands. With 256 motifs, the encoder includes 82688 parameters. We trained each model using Adam with learning rate 0.001 for a maximum of 100 epochs, and used a held-out validation set for early stopping (patience of 10 epochs). All models were trained on a workstation equipped with an Nvidia Quadro P6000. Example of a loss curve showing convergence can be found as Supplementary Figure 3.

One the parameters have been trained, we encode promoter sequences into the 256-dimensional RHIEPA space using up to L=500bp upstream for each ortholog, as we found that this reduced the frequency of clustering together promoters of an adjacent genes (compared to using the full 800bp used in training the model.) We averaged over all available orthologs and filtered any upstream regions with less than 2 orthologs. Codes are available at https://github.com/alanmoses/promoter_reverse_homology

Comparison of learned motifs to databases using TomTom

We converted the motifs to MEME format and uploaded them to the online MEME suite[43]. We searched JASPAR fungi [41]and cis-bp[42] for each species of interest. For the learned databases of motifs obtained using reverse homology, we used E-value<1/256 to control for false matches. For the expert curated yetfasco PWMs [44], we used only the 244 high-confidence motifs, and E-value <1/244. TomTom outputs showing logos and E-values for all the motifs in each map are available as supplementary data.

Comparison of representations to gene expression data

We converted binding data (entire matrix of binding evidence downloaded from yeastract[98]) and netprophet3 regulatory network[49] to binary representation vectors where for each promoter where we have a 1 if a transcription factor binds it and 0 otherwise. We also tried using the confidence weights from netprophet3 as a high-dimensional representation, but obtained worse results. We used the sci-kit learn[99] neighbours module to identify nearest neighbours using cosine distance for each representation. We median centered the columns of each representation and the columns of the gene expression data, and removed any expression experiments with >50% missing data. We set the remaining missing values to 0 (expression data is in log space) and removed any experiments with 0 standard deviation.

For nearest neighbour expression experiment prediction, we replaced the observed gene expression values for a given experiment with the values of the nearest neighbour (or average of 50 nearest neighbours) in the representation space. We then computed the Pearson correlation between the nearest neighbour expression values and the observed gene expression values. This represents 1-nn regression with leave-one-out cross-validation.

For nearest neighbour prediction of co-expression, we computed the genes by genes correlation matrix and considered any gene pairs above correlation 0.5 to be co-expressed. Approximately 1% of gene pairs are considered co-expressed by this criterion. For each gene, we then checked whether a co-expressed gene was the nearest neighbour in the representation space

(top 1 precision) or whether a co-expressed gene was among the 50 nearest neighbours (top 50 precision).

Visualization and clustering of motif-based representations

We obtained gene descriptions from SGD[100] (budding yeast, *S. cerevisiae*) CGD[101] (*C. albicans*) Ensembl[70] (*N. crassa*, NC12, and *F. graminearum*, PH-1). We supplemented the *F. graminearum* annotations using annotations of our orthologs in neurospora.

We made UMAPs using the uwot package [102] in R and coloured points using grep of the gene names and descriptions. For example, for 'proteasome' in Figure 2s, we searched for any genes that had 20S proteasome or 26S proteasome in their gene description, while for 'ribosome', we searched for gene names that matched RPL or RPS. For PHO, ERG and MET pathways, we coloured genes with names matching PHO, ERG or MET, respectively.

For clustered heatmaps, we median centered and clustered using cluster3.0 (average linkage hierarchical clustering). Clustered representations were explored interactively in java treeview[103]. The datafiles for treeview for each of the 4 regulatory maps are provided as supplementary data.

Enrichment analyses

GO analysis was done using Gprofiler[104] (using all known genes in the species as the background) and other data was tested for association using Fisher's test in R. Known targets of Neurospora transcription factors were obtained [74] and we used the low-throughput evidence as evidence of bona fide transcription factor-target interactions. Rhythmic promoters were identified using RNA pol binding and gene expression[75]. We parsed PHI-base[84] for *F. graminearum* genes that were annotated as 'reduced virulence': these are genes that result in reduced virulence when mutated. Gene lists for clusters discussed in the text are available as Supplementary File 1.

Data and code availability

Supplementary data can be found at https://zenodo.org/records/14920044

Codes can be found at https://github.com/stajichlab/Y1000_orthologs and https://github.com/alanmoses/promoter_reverse_homology

Acknowledgements

Research was performed using a GPU generously donated by Nvidia to AMM and hardware obtained through support to AMM from the Canada Foundation for Innovation. We thank members of the Moses lab, Knowles Lab, Philip Fradkin, Sanjana Bhatnagar and Dr. John Calarco for Discussions and Cameron Dufault for comments on the manuscript. We acknowledge Aqsa Alam for sharing data that was used for earlier versions of the analysis.

References

1. Aerts S. Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol. 2012;98: 121–145. doi:10.1016/B978-0-12-386499-4.00005-7

2. Kim S, Wysocka J. Deciphering the multi-scale, quantitative cis-regulatory code. Mol Cell. 2023; S1097-2765(22)01215–1. doi:10.1016/j.molcel.2022.12.032

3. Zeitlinger J. Seven myths of how transcription factors read the cis-regulatory code. Curr Opin Syst Biol. 2020;23: 22–31. doi:10.1016/j.coisb.2020.08.002

4. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet. 2004;5: 276–287. doi:10.1038/nrg1315

5. Zia A, Moses AM. Towards a theoretical understanding of false positives in DNA motif finding. BMC Bioinformatics. 2012;13: 151. doi:10.1186/1471-2105-13-151

6. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. 2004;5: R98. doi:10.1186/gb-2004-5-12-r98

7. Aerts S, Quan X-J, Claeys A, Sanchez MN, Tate P, Yan J, et al. Robust Target Gene Discovery through Transcriptome Perturbations and Genome-Wide Enhancer Predictions in Drosophila Uncovers a Regulatory Basis for Sensory Specification. PLOS Biol. 2010;8: e1000435. doi:10.1371/journal.pbio.1000435

8. Eddy SR. A Model of the Statistical Power of Comparative Genome Sequence Analysis. PLOS Biol. 2005;3: e10. doi:10.1371/journal.pbio.0030010

9. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature. 2003;423: 241–254. doi:10.1038/nature01644

10. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005;434: 338–345. doi:10.1038/nature03441

11. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45: 891–898. doi:10.1038/ng.2684

12. Kheradpour P, Stark A, Roy S, Kellis M. Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res. 2007;17: 1919–1931. doi:10.1101/gr.7090407

13. Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 2000;403: 564–567. doi:10.1038/35000615

14. Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 2008;4: e1000106. doi:10.1371/journal.pgen.1000106

15. Doniger SW, Fay JC. Frequent gain and loss of functional transcription factor binding sites. PLoS Comput Biol. 2007;3: e99. doi:10.1371/journal.pcbi.0030099

16. Weirauch MT, Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet TIG. 2010;26: 66–74. doi:10.1016/j.tig.2009.12.002

17. He X, Ling X, Sinha S. Alignment and Prediction of cis-Regulatory Modules Based on a Probabilistic Model of Evolution. PLOS Comput Biol. 2009;5: e1000299. doi:10.1371/journal.pcbi.1000299

18. Lunter G, Rocco A, Mimouni N, Heger A, Caldeira A, Hein J. Uncertainty in homology inferences: assessing and improving genomic sequence alignment. Genome Res. 2008;18: 298–309. doi:10.1101/gr.6725608

19. Ray P, Shringarpure S, Kolar M, Xing EP. CSMET: Comparative Genomic Motif Detection via Multi-Resolution Phylogenetic Shadowing. PLOS Comput Biol. 2008;4: e1000090. doi:10.1371/journal.pcbi.1000090

20. Arnold P, Erb I, Pachkov M, Molina N, van Nimwegen E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinforma Oxf Engl. 2012;28: 487–494. doi:10.1093/bioinformatics/btr695

21. Glenwinkel L, Wu D, Minevich G, Hobert O. TargetOrtho: a phylogenetic footprinting tool to identify transcription factor targets. Genetics. 2014;197: 61–76. doi:10.1534/genetics.113.160721

22. Gordân R, Narlikar L, Hartemink AJ. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Res. 2010;38: e90. doi:10.1093/nar/gkp1166

23. Christmas MJ, Kaplow IM, Genereux DP, Dong MX, Hughes GM, Li X, et al. Evolutionary constraint and innovation across hundreds of placental mammals. Science. 2023;380: eabn3943. doi:10.1126/science.abn3943

24. Opulente DA, LaBella AL, Harrison M-C, Wolters JF, Liu C, Li Y, et al. Genomic factors shape carbon and nitrogen metabolic niche breadth across Saccharomycotina yeasts. Science. 2024;384: eadj4503. doi:10.1126/science.adj4503

25. Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, et al. Highly contiguous assemblies of 101 drosophilid genomes. Coop G, Wittkopp PJ, Sackton TB, editors. eLife. 2021;10: e66405. doi:10.7554/eLife.66405

26. Hahn S, Young ET. Transcriptional Regulation in Saccharomyces cerevisiae: Transcription Factor Regulation and Function, Mechanisms of Initiation, and Roles of Activators and Coactivators. Genetics. 2011;189: 705–736. doi:10.1534/genetics.111.127019

27. Rodriguez DL, Quail MM, Hernday AD, Nobile CJ. Transcriptional Circuits Regulating Developmental Processes in Candida albicans. Front Cell Infect Microbiol. 2020;10: 605711. doi:10.3389/fcimb.2020.605711

28. Gostinčar C, Stajich JE, Gunde-Cimerman N. Extremophilic and extremotolerant fungi. Curr Biol. 2023;33: R752–R756. doi:10.1016/j.cub.2023.06.011

29. Case NT, Berman J, Blehert DS, Cramer RA, Cuomo C, Currie CR, et al. The future of fungi: threats and opportunities. G3 Bethesda Md. 2022;12: jkac224. doi:10.1093/g3journal/jkac224

30. Koo PK, Eddy SR. Representation learning of genomic sequence motifs with convolutional neural networks. PLOS Comput Biol. 2019;15: e1007560. doi:10.1371/journal.pcbi.1007560

31. Alam A, Duncan AG, Mitchell JA, Moses AM. Functional similarity of non-coding regions is revealed in phylogenetic average motif score representations. bioRxiv; 2023. p. 2023.04.09.536185. doi:10.1101/2023.04.09.536185

32. Tareen A, Kinney JB. Biophysical models of cis-regulation as interpretable neural networks. arXiv; 2020. doi:10.48550/arXiv.2001.03560

33. de Boer CG, Vaishnav ED, Sadeh R, Abeyta EL, Friedman N, Regev A. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat Biotechnol. 2020;38: 56–65. doi:10.1038/s41587-019-0315-8

34. Novakovsky G, Fornes O, Saraswat M, Mostafavi S, Wasserman WW. ExplaiNN: interpretable and transparent neural networks for genomics. bioRxiv; 2022. p. 2022.05.20.492818. doi:10.1101/2022.05.20.492818

35. Balcı AT, Ebeid MM, Benos PV, Kostka D, Chikina M. An intrinsically interpretable neural network architecture for sequence to function learning. bioRxiv; 2023. p. 2023.01.25.525572. doi:10.1101/2023.01.25.525572

36. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16: 16–23. doi:10.1093/bioinformatics/16.1.16

37. Lu AX, Lu AX, Moses A. Evolution Is All You Need: Phylogenetic Augmentation for Contrastive Learning. arXiv; 2020. doi:10.48550/arXiv.2012.13475

38. Lu AX, Lu AX, Pritišanac I, Zarin T, Forman-Kay JD, Moses AM. Discovering molecular features of intrinsically disordered regions by using evolution for contrastive learning. PLOS Comput Biol. 2022;18: e1010238. doi:10.1371/journal.pcbi.1010238

39. Consens ME, Dufault C, Wainberg M, Forster D, Karimzadeh M, Goodarzi H, et al. To Transformers and Beyond: Large Language Models for the Genome. arXiv; 2023. doi:10.48550/arXiv.2311.07621

40. Karollus A, Hingerl J, Gankin D, Grosshauser M, Klemon K, Gagneur J. Species-aware DNA language models capture regulatory elements and their evolution. Genome Biol. 2024;25: 83. doi:10.1186/s13059-024-03221-x

41. Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Lemma RB, Turchi L, Blanc-Mathieu R, et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2022;50: D165–D173. doi:10.1093/nar/gkab1113

42. Lambert SA, Yang AWH, Sasse A, Cowley G, Albu M, Caddick MX, et al. Similarity regression predicts evolution of transcription factor sequence specificity. Nat Genet. 2019;51: 981–989. doi:10.1038/s41588-019-0411-1

43. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009;37: W202-208. doi:10.1093/nar/gkp335

44. de Boer CG, Hughes TR. YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. Nucleic Acids Res. 2012;40: D169-179. doi:10.1093/nar/gkr993

45. Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. Cell. 2014;157: 740–752. doi:10.1016/j.cell.2014.02.054

46. Karollus A, Mauermeier T, Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biol. 2023;24: 56. doi:10.1186/s13059-023-02899-9

47. Zhang S, Pyne S, Pietrzak S, Halberg S, McCalla SG, Siahpirani AF, et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets. Nat Commun. 2023;14: 3064. doi:10.1038/s41467-023-38637-9

48. Skok Gibbs C, Jackson CA, Saldi G-A, Tjärnberg A, Shah A, Watters A, et al. High-performance single-cell gene regulatory network inference at scale: the Inferelator 3.0. Bioinformatics. 2022;38: 2519–2528. doi:10.1093/bioinformatics/btac117

49. Abid D, Brent MR. NetProphet 3: a machine learning framework for transcription factor network mapping and multi-omics integration. Bioinformatics. 2023;39: btad038. doi:10.1093/bioinformatics/btad038

50. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A. 1998;95: 14863–14868. doi:10.1073/pnas.95.25.14863

51. Shore D, Zencir S, Albert B. Transcriptional control of ribosome biogenesis in yeast: links to growth and stress signals. Biochem Soc Trans. 2021;49: 1589–1599. doi:10.1042/BST20201136

52. Hinnebusch AG, Natarajan K. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress. Eukaryot Cell. 2002;1: 22–32. doi:10.1128/EC.01.1.22-32.2002

53. McFarland MR, Keller CD, Childers BM, Adeniyi SA, Corrigall H, Raguin A, et al. The molecular aetiology of tRNA synthetase depletion: induction of a GCN4 amino acid starvation response despite homeostatic maintenance of charged tRNA levels. Nucleic Acids Res. 2020;48: 3071–3088. doi:10.1093/nar/gkaa055

54. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9: 3273–3297. doi:10.1091/mbc.9.12.3273

55. Gasch AP, Yu FB, Hose J, Escalante LE, Place M, Bacher R, et al. Single-cell RNA sequencing reveals intrinsic and extrinsic regulatory heterogeneity in yeast responding to stress. PLoS Biol. 2017;15: e2004050. doi:10.1371/journal.pbio.2004050

56. Koch C, Moll T, Neuberg M, Ahorn H, Nasmyth K. A role for the transcription factors Mbp1 and Swi4 in progression from G1 to S phase. Science. 1993;261: 1551–1557. doi:10.1126/science.8372350

57. Pramila T, Wu W, Miles S, Noble WS, Breeden LL. The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. Genes Dev. 2006;20: 2266–2278. doi:10.1101/gad.1450606

58. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, et al. Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. Genes Dev. 2002;16: 3017–3033. doi:10.1101/gad.1039602

59. Ramos-Alonso L, Romero AM, Martínez-Pastor MT, Puig S. Iron Regulatory Mechanisms in Saccharomyces cerevisiae. Front Microbiol. 2020;11: 582830. doi:10.3389/fmicb.2020.582830

60. Rossi MJ, Kuntala PK, Lai WKM, Yamada N, Badjatia N, Mittal C, et al. A high-resolution protein architecture of the budding yeast genome. Nature. 2021;592: 309–314. doi:10.1038/s41586-021-03314-8

61. Sasaki H, Kishimoto T, Mizuno T, Shinzato T, Uemura H. Expression of GCR1, the transcriptional activator of glycolytic enzyme genes in the yeast Saccharomyces cerevisiae, is positively autoregulated by Gcr1p. Yeast Chichester Engl. 2005;22: 305–319. doi:10.1002/yea.1212

62. Abramova NE, Cohen BD, Sertil O, Kapoor R, Davies KJ, Lowry CV. Regulatory mechanisms controlling expression of the DAN/TIR mannoprotein genes during anaerobic remodeling of the cell wall in Saccharomyces cerevisiae. Genetics. 2001;157: 1169–1177. doi:10.1093/genetics/157.3.1169

63. Prugar E, Burnett C, Chen X, Hollingsworth NM. Coordination of Double Strand Break Repair and Meiotic Progression in Yeast by a Mek1-Ndt80 Negative Feedback Loop. Genetics. 2017;206: 497–512. doi:10.1534/genetics.117.199703

64. Zhao H, Eide DJ. Zap1p, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in Saccharomyces cerevisiae. Mol Cell Biol. 1997;17: 5044–5052. doi:10.1128/MCB.17.9.5044

65. Le Crom S, Devaux F, Marc P, Zhang X, Moye-Rowley WS, Jacq C. New insights into the pleiotropic drug resistance network from genome-wide characterization of the YRR1 transcription factor regulation system. Mol Cell Biol. 2002;22: 2642–2649. doi:10.1128/MCB.22.8.2642-2649.2002

66. Askew C, Sellam A, Epp E, Hogues H, Mullick A, Nantel A, et al. Transcriptional regulation of carbohydrate metabolism in the human pathogen Candida albicans. PLoS Pathog. 2009;5: e1000612. doi:10.1371/journal.ppat.1000612

67. Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, Purisima E, et al. Transcription factor substitution during the evolution of fungal ribosome regulation. Mol Cell. 2008;29: 552–562. doi:10.1016/j.molcel.2008.02.006

68. Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, Whiteway M. Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol. 2010;8: e1000329. doi:10.1371/journal.pbio.1000329

69. Bennett RJ. The parasexual lifestyle of Candida albicans. Curr Opin Microbiol. 2015;28: 10–17. doi:10.1016/j.mib.2015.06.017

70. Ensembl Fungi. [cited 24 Feb 2025]. Available: https://fungi.ensembl.org/index.html

71. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol. 2004;2: e398. doi:10.1371/journal.pbio.0020398

72. Hynes MJ, Murray SL, Duncan A, Khew GS, Davis MA. Regulatory genes controlling fatty acid catabolism and peroxisomal functions in the filamentous fungus Aspergillus nidulans. Eukaryot Cell. 2006;5: 794–805. doi:10.1128/EC.5.5.794-805.2006

73. Lamb TM, Finch KE, Bell-Pedersen D. The *Neurospora crassa* OS MAPK pathway-activated transcription factor ASL-1 contributes to circadian rhythms in pathway responsive clock-controlled genes. Fungal Genet Biol. 2012;49: 180–188. doi:10.1016/j.fgb.2011.12.006

74. Hu Y, Qin Y, Liu G. Collection and Curation of Transcriptional Regulatory Interactions in Aspergillus nidulans and Neurospora crassa Reveal Structural and Evolutionary Features of the Regulatory Networks. Front Microbiol. 2018;9: 27. doi:10.3389/fmicb.2018.00027

75. Sancar C, Sancar G, Ha N, Cesbron F, Brunner M. Dawn- and dusk-phased circadian transcription rhythms coordinate anabolic and catabolic functions in Neurospora. BMC Biol. 2015;13: 17. doi:10.1186/s12915-015-0126-4

76. Wu VW, Thieme N, Huberman LB, Dietschmann A, Kowbel DJ, Lee J, et al. The regulatory and transcriptional landscape associated with carbon utilization in a filamentous fungus. Proc Natl Acad Sci U S A. 2020;117: 6003–6013. doi:10.1073/pnas.1915611117

77. Philley ML, Staben C. Functional analyses of the Neurospora crassa MT a-1 mating type polypeptide. Genetics. 1994;137: 715–722. doi:10.1093/genetics/137.3.715

78. Shiu PK, Raju NB, Zickler D, Metzenberg RL. Meiotic silencing by unpaired DNA. Cell. 2001;107: 905–916. doi:10.1016/s0092-8674(01)00609-2

79. Shiu PKT, Zickler D, Raju NB, Ruprich-Robert G, Metzenberg RL. SAD-2 is required for meiotic silencing by unpaired DNA and perinuclear localization of SAD-1 RNA-directed RNA polymerase. Proc Natl Acad Sci U S A. 2006;103: 2243–2248. doi:10.1073/pnas.0508896103

80. Lee DW, Pratt RJ, McLaughlin M, Aramayo R. An Argonaute-Like Protein Is Required for Meiotic Silencing. Genetics. 2003;164: 821–828. doi:10.1093/genetics/164.2.821

81. Wang Z, Lopez-Giraldez F, Slot J, Yarden O, Trail F, Townsend JP. Secondary Metabolism Gene Clusters Exhibit Increasingly Dynamic and Differential Expression during Asexual Growth, Conidiation, and Sexual Development in Neurospora crassa. mSystems. 2022;7: e0023222. doi:10.1128/msystems.00232-22

82. Honda S, Eusebio-Cope A, Miyashita S, Yokoyama A, Aulia A, Shahi S, et al. Establishment of Neurospora crassa as a model organism for fungal virology. Nat Commun. 2020;11: 5627. doi:10.1038/s41467-020-19355-y

83. Alouane T, Rimbert H, Bormann J, González-Montiel GA, Loesgen S, Schäfer W, et al. Comparative Genomics of Eight Fusarium graminearum Strains with Contrasting Aggressiveness Reveals an Expanded Open Pangenome and Extended Effector Content Signatures. Int J Mol Sci. 2021;22: 6257. doi:10.3390/ijms22126257

84. Urban M, Cuzick A, Seager J, Nonavinakere N, Sahoo J, Sahu P, et al. PHI-base - the multi-species pathogen-host interaction database in 2025. Nucleic Acids Res. 2025;53: D826–D838. doi:10.1093/nar/gkae1084

85. Scannell DR, Zill OA, Rokas A, Payen C, Dunham MJ, Eisen MB, et al. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the Saccharomyces sensu stricto Genus. G3 Bethesda Md. 2011;1: 11–25. doi:10.1534/g3.111.000273

86. de Boer CG, Taipale J. Hold out the genome: a roadmap to solving the cis-regulatory code. Nature. 2024;625: 41–50. doi:10.1038/s41586-023-06661-w

87. Carriel CC, Pyne S, Halberg-Spencer SA, Park SC, Seo H, Schmidt A, et al. A network-based model of Aspergillus fumigatus elucidates regulators of development and defensive natural products of an opportunistic pathogen. bioRxiv. 2023; 2023–05.

88. Zhang S, Knaack S, Roy S. Enabling Studies of Genome-Scale Regulatory Network Evolution in Large Phylogenies with MRTLE. Methods Mol Biol Clifton NJ. 2022;2477: 439–455. doi:10.1007/978-1-0716-2257-5_24

89. Sakurai H, Takemori Y. Interaction between Heat Shock Transcription Factors (HSFs) and Divergent Binding Sequences: BINDING SPECIFICITIES OF YEAST HSFs AND HUMAN HSF1*. J Biol Chem. 2007;282: 13334–13341. doi:10.1074/jbc.M611801200

90. Petti AA, McIsaac RS, Ho-Shing O, Bussemaker HJ, Botstein D. Combinatorial control of diverse metabolic and physiological functions by transcriptional regulators of the yeast sulfur assimilation pathway. Mol Biol Cell. 2012;23: 3008–3024. doi:10.1091/mbc.E12-03-0233

91. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, et al. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci U S A. 2013;110: E2792-2801. doi:10.1073/pnas.1222534110

92. Gerber AP, Herschlag D, Brown PO. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. PLoS Biol. 2004;2: E79. doi:10.1371/journal.pbio.0020079

93. Saman Booy M, Ilin A, Orponen P. RNA secondary structure prediction with convolutional neural networks. BMC Bioinformatics. 2022;23: 58. doi:10.1186/s12859-021-04540-7

94. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20: 238. doi:10.1186/s13059-019-1832-y

95. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

96. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv; 2015. doi:10.48550/arXiv.1502.03167

97. Oord A van den, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding. arXiv; 2019. doi:10.48550/arXiv.1807.03748

98. Teixeira MC, Viana R, Palma M, Oliveira J, Galocha M, Mota MN, et al. YEASTRACT+: a portal for the exploitation of global transcription regulation and metabolic model data in yeast biotechnology and pathogenesis. Nucleic Acids Res. 2023;51: D785–D791. doi:10.1093/nar/gkac1041

99. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830.

100. Engel SR, Aleksander S, Nash RS, Wong ED, Weng S, Miyasato SR, et al. Saccharomyces Genome Database: Advances in Genome Annotation, Expanded Biochemical Pathways, and Other Key Enhancements. Genetics. 2024; iyae185. doi:10.1093/genetics/iyae185

101. Skrzypek MS, Binkley J, Binkley G, Miyasato SR, Simison M, Sherlock G. The Candida Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high throughput sequencing data. Nucleic Acids Res. 2017;45: D592–D596. doi:10.1093/nar/gkw924

102. Melville J. jlmelville/uwot. 2025. Available: https://github.com/jlmelville/uwot

103. Saldanha AJ. Java Treeview—extensible visualization of microarray data. Bioinformatics. 2004;20: 3246–3248. doi:10.1093/bioinformatics/bth349

104. Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). Nucleic Acids Res. 2023;51: W207–W212. doi:10.1093/nar/gkad347

Figure Legends

Figure 1. Motif-based encoder to represent orthologous sets of upstream regions a) Shows the alignment of the upstream region for a single yeast promoter (*PHO84*) and orthologs. In yeast, *PHO84* is known to be a target of Pho4, and consensus sites for that transcription factor are found in the promoters from all species (black boxes), but are only aligned among the most closely related species (black box over text) b) a motif-based encoder changes the

sequence data into a 256 dimensional numerical representation. Dimensions of data inputs and outputs are shown as black text, transformations (implemented as layers in Keras) are showed in red text and the number of free parameters in each layer is indicated in purple text. c) the reverse homology, interpretable encoder, phylogenetic average (RHIEPA) representation for PHO84 has a strong activity for matrix 0, which is similar to the Pho4 consensus site. Each yellow bar represents the signal for each motif and brighter yellow indicates stronger signal, black indicates 0.

Figure 2. Unsupervised learning of the cis-regulatory code from sequence alone. a) Examples of learned yeast motifs (Matrix x) with matches to motifs the JASPAR database (MAx.x). E-values (from TomTom) indicate the probability of observing a motif in the database as similar by chance b) summary of learned motif matches against JASPAR and Cis-bp databases for 4 clades considered here (S. c.: *S. cerevisiae*, C.a: C. albicans, N.c: *N. crassa* F.g.: *F. graminearum*. High-confidence motifs from the Yetfasco database are included as a positive control for the database search. Numbers above the bars indicate the number of matches. c) comparison of predictive power of the reverse homology, interpretable encoder, phylogenetic average (RHIEPA) promoter sequence representations (blue bars) against various other representations (see text for details). Left panel represents nearest neighbour regression (nn) to predict relative gene expression levels. Right panel represents precision (positive predictive power) to identify a co-expressed gene as the 1 closet neighbour (Top 1) or among the 50 nearest neighbours (Top 50). Error bars represent 3 times standard error of the mean (n=1486 experiments, n>5300 genes) d) biological organization emerges in a 2D representation of the 256 dimensional RHIEPA space for *S. cerevisiae*. Each symbol represents a set or orthologous upstream regions. Colours indicate matches to the *S. cerevisiae* gene name or description. Inset is a zoom in showing that the PHO84 upstream region is found among upstream regions from other genes in the PHO pathway.

Figure 3. Reading the cis-regulatory code in yeast. Reading the cis-regulatory code in yeast. a) yeast promoters can be organized by sequence alone one they are transformed into the motif-based RHIEPA representation. Rows represent orthologous sets of upstream regions and columns represent learned motifs. Yellow colour indicates stronger signal for that motif in that promoter. Biological functions and known transcription factor (TF) motifs (best match of the learned motif JASPAR, TomTom E-value<0.05) are indicated to the right of clusters. Grey boxes in the heatmap indicate clusters in subsequent panels. b) the upstream regions of genes encoding cell cycle transcription factors (purple) are in a cluster where the learned motifs match their known motifs (red). Gene expression data is ordered according to the clustering of motif-based promoter representations and red indicates higher relative expression, while green indicates lower relative expression. GO and other functional enrichments for smaller clusters are indicated to the right of the heatmaps. c) a small section of the clustered promoter heatmap shows the promoter for AFT1 among known iron-responsive genes. d) Aft1 is predicted to be autoregulated because the upstream regions of AFT1 and its orthologs contain matches to the Aft1 consensus site, and one of these sites is conserved among closely related yeasts. e) transcription factor (TF) autoregulation is predicted based on the promoter representation cluster analysis in at least 9 cases, 5 of which are known (see text for details)

Figure 4. A regulatory map for C. albicans predicts a meiosis/parasex pathway a) clustered motif-based RHIEPA representations of *C. albicans* promoter regions presented as in Figure 3a. TF motifs with '?' did not have a significant match in JASPAR, but are named based

on similarity of the logo (displayed) to known motifs. b) cluster of upstream regions statistically associated with meiosis annotations. Upstream regions of homologs of meiosis genes are indicated in black, while upstream regions of homologs of meiosis regulators are indicated in purple. Two learned matrices with no matches to JASPAR (Matrix 114 and Matrix 246) show strong signals in this cluster.

Figure 5. Reading the cis-regulatory code in N. crassa. a) heatmap of motif-based RHIEPA promoter region representation and comparison with gene expression is shown as in Figure 3a. Matrices with no matches are indicated as matrix x. question marks indicate motifs that are similar to known motifs, but not statistically significant in database search. Predicted b) mating pathway and c) virus response pathway displayed as in Figure 4b. Ccg-4 is the only known target of MatA-1 (purple).

Figure 6. Association of promoter representation with virulence in Fusarium a) Map of Fusarium promoter regions visualized as in Figure 3a shows many clusters with unknown motifs and no functional enrichments. A large group of upstream regions of genes enriched for signaling, but relatively weak motif signals. b) genes annotated as reduced virulence are enriched in a region of the high-dimensional motif-based representation. This corresponds to the large groups of promoters with weak signals in panel a. Colours indicate terms found in the gene descriptions as in Figure 2d. c) these genes are enriched for reduced virulence annotations in PHI-base suggesting. Error bars represent 2 times the standard error of the fraction (using the normal approximation to the binomial). Numbers above the bars represent the sample sizes used to compute the fraction. P-value is based on a Fisher's exact test.

Supplementary Figures

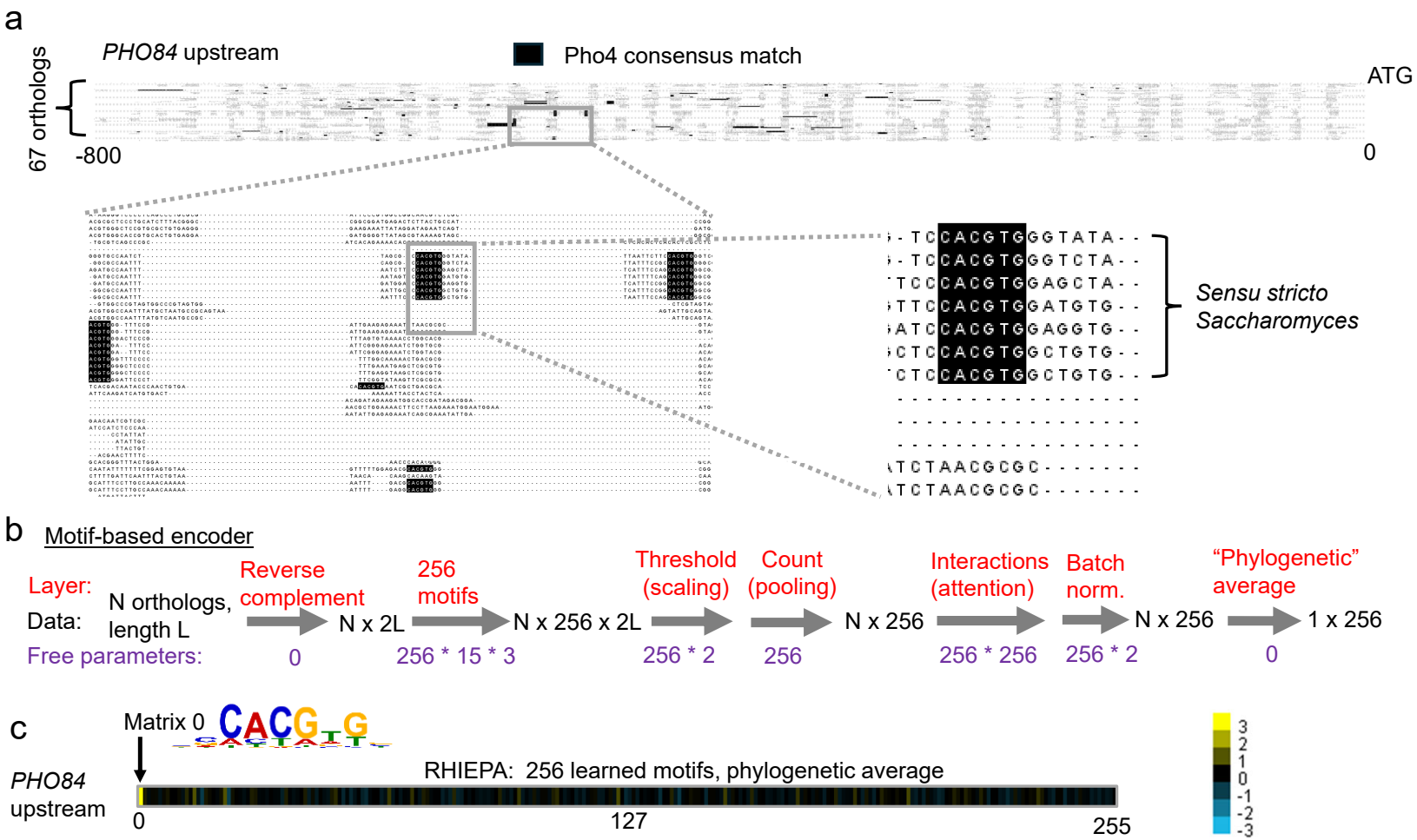
Supplementary Figure 1. Additional evidence for understanding of the yeast regulatory code. a) comparison of the RHIEPA promoter representations to the entire cell synchronized microarray data. Gene expression data has been ordered according to the clustered RHIEPA promoter representation. In addition to the cell cycle cluster, clusters of promoters associated with translation, heat shock and general stress response, show changes in gene expression. These represent the various stress and growth responses due to the experimental conditions (alpha factor, temperature sensitive and other mutants) needed to synchronize the cells. Genes with meiosis regulator motifs in their promoters change expression in the Cdc28 mutant strains. b) genes with promoters in cell cycle clusters (Figure 3) show coordinated expression in single cell RNA-seq experiments. Rows (genes) in the single cell RNA-seq data have been ordered by the RHIEPA clustering of genes, and columns have been clustered based on the entire single-cell RNA-seq dataset. c) Evidence of binding of Aft1 to the *AFT* promoter as displayed on the UCSC genome browser link from http://yeastepigenome.org/

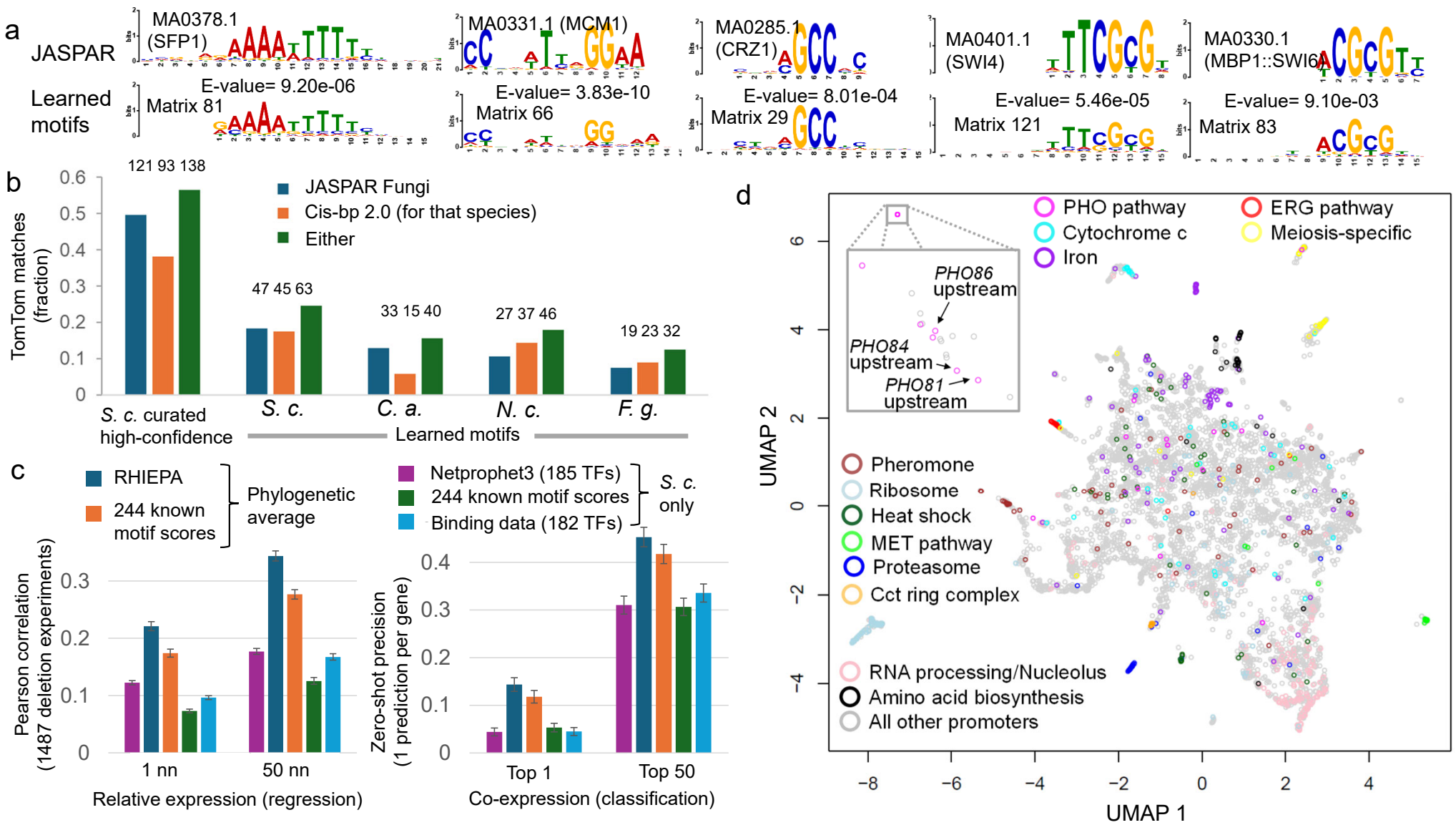
Supplementary Figure 2. Ortholog coverage distribution. Histograms showing the distribution of the number of orthologs for each gene in the reference genomes.

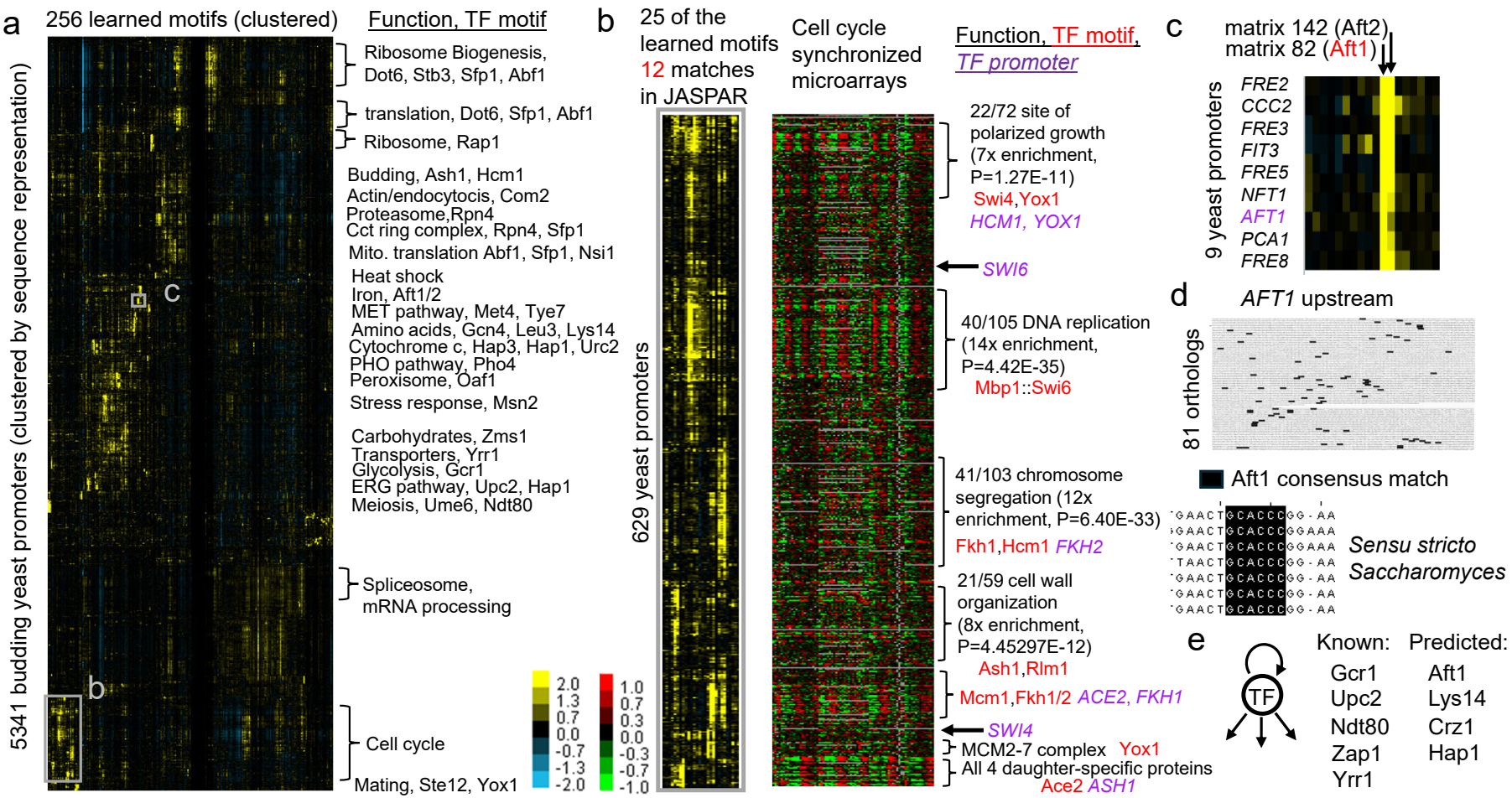
Supplementary Figure 3. Example training curves. Example training information for the S. cerevisiae model. All 375 promoters (and orthologs) on chromosome 2 were held out as the validation sample. In this example, the model was trained for the full 100 epochs, as validation loss continued to decrease.

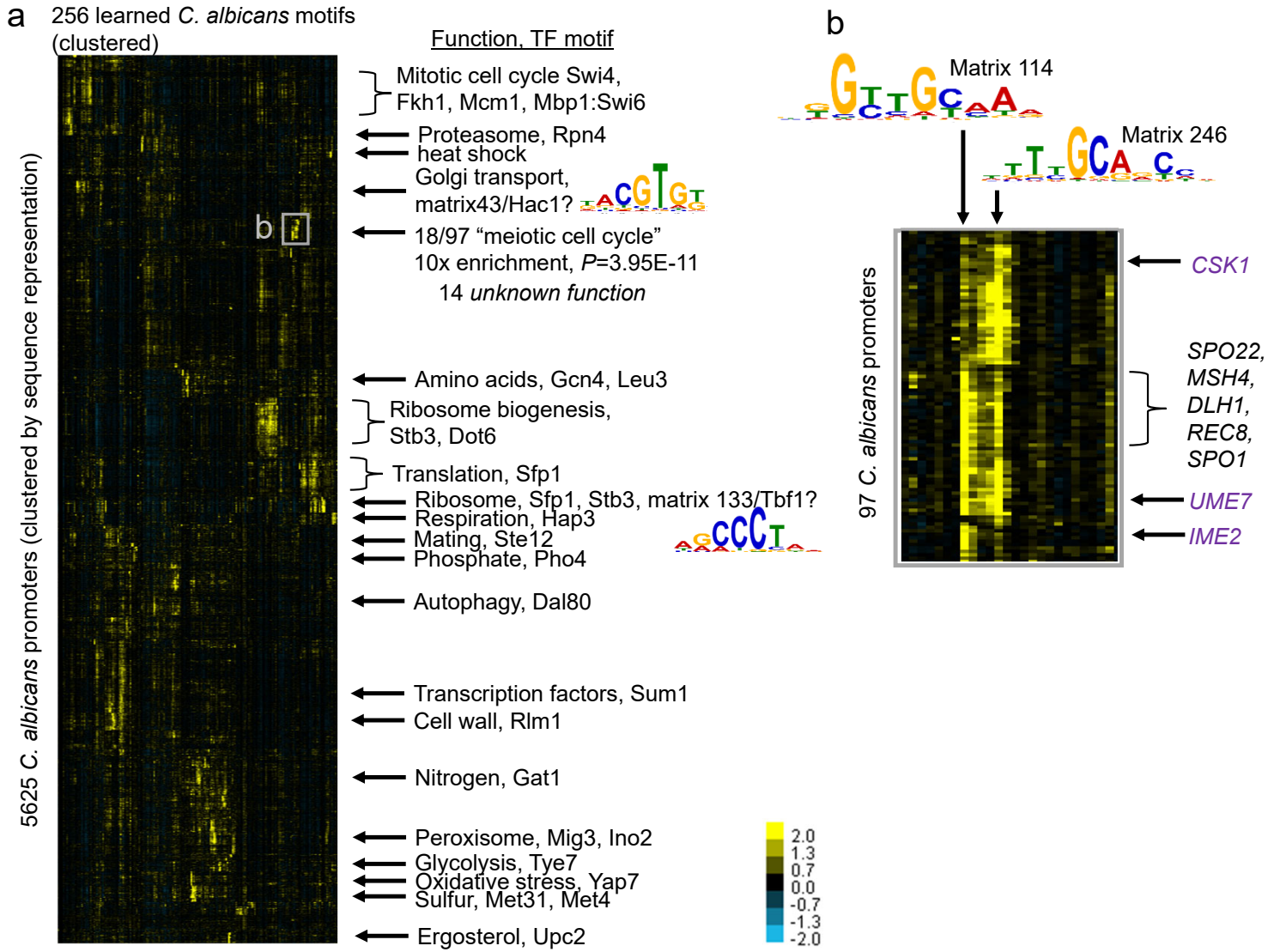
Supplementary Files

Supplementary File 1. Enrichment analysis. This file is an Excel spreadsheet of gene lists and enrichment analysis for clusters discussed in the text. The first 4 tabs are the clusters discussed for each species. Each cluster is represented as two columns, gene IDs and gene descriptions. Subsequent columns include the Gprofiler results including the enriched annotation category, GO id, P-value, log P-value, number of genes in the genome with that annotation, number of genes in the cluster, number of genes in the cluster with that annotation and total number of genes. Below the fraction of genes in the cluster and genome are indicated, as well as their ratio, which is the fold enrichment used in the text. The final two tabs include the data used to compute enrichments for *F. graminerum* and *N. crassa* described in the text.

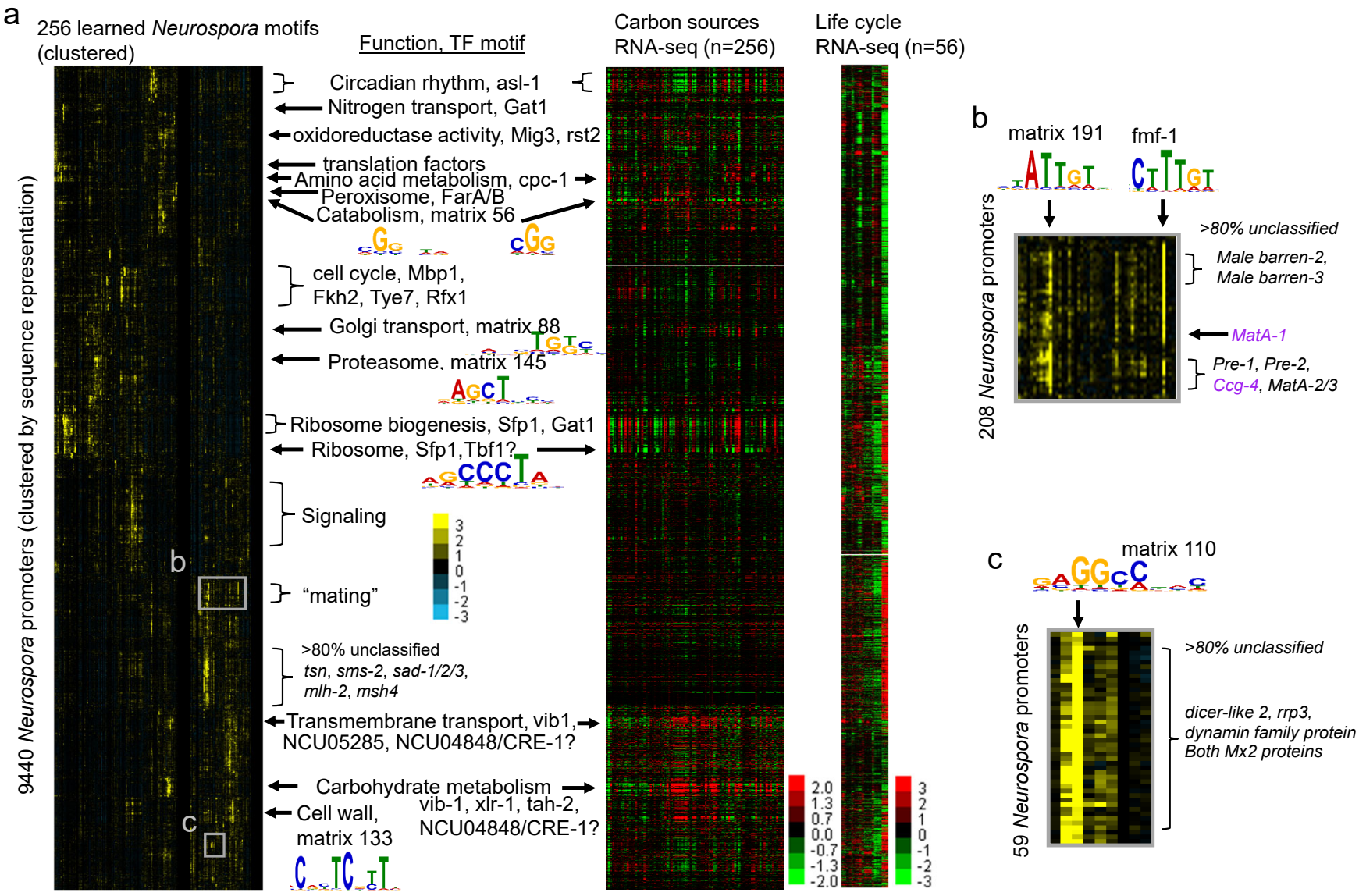

a
PHO84 upstream
Pho4 consensus match
67 orthologs
-800
ATG
0
Sensu stricto Saccharomyces
b
Motif-based encoder
Layer:
Data:
N orthologs, length L
Reverse complement
N x 2L
256 motifs
N x 256 x 2L
Threshold (scaling)
Count (pooling)
N x 256
Interactions (attention)
Batch norm.
N x 256
"Phylogenetic" average
1 x 256
Free parameters:
0
256 * 15 * 3
256 * 2
256
256 * 256
256 * 2
0
c
Matrix 0
PHO84 upstream
RHIEPA: 256 learned motifs, phylogenetic average
0
127
255
3
2
1
0
-1
-2
-3

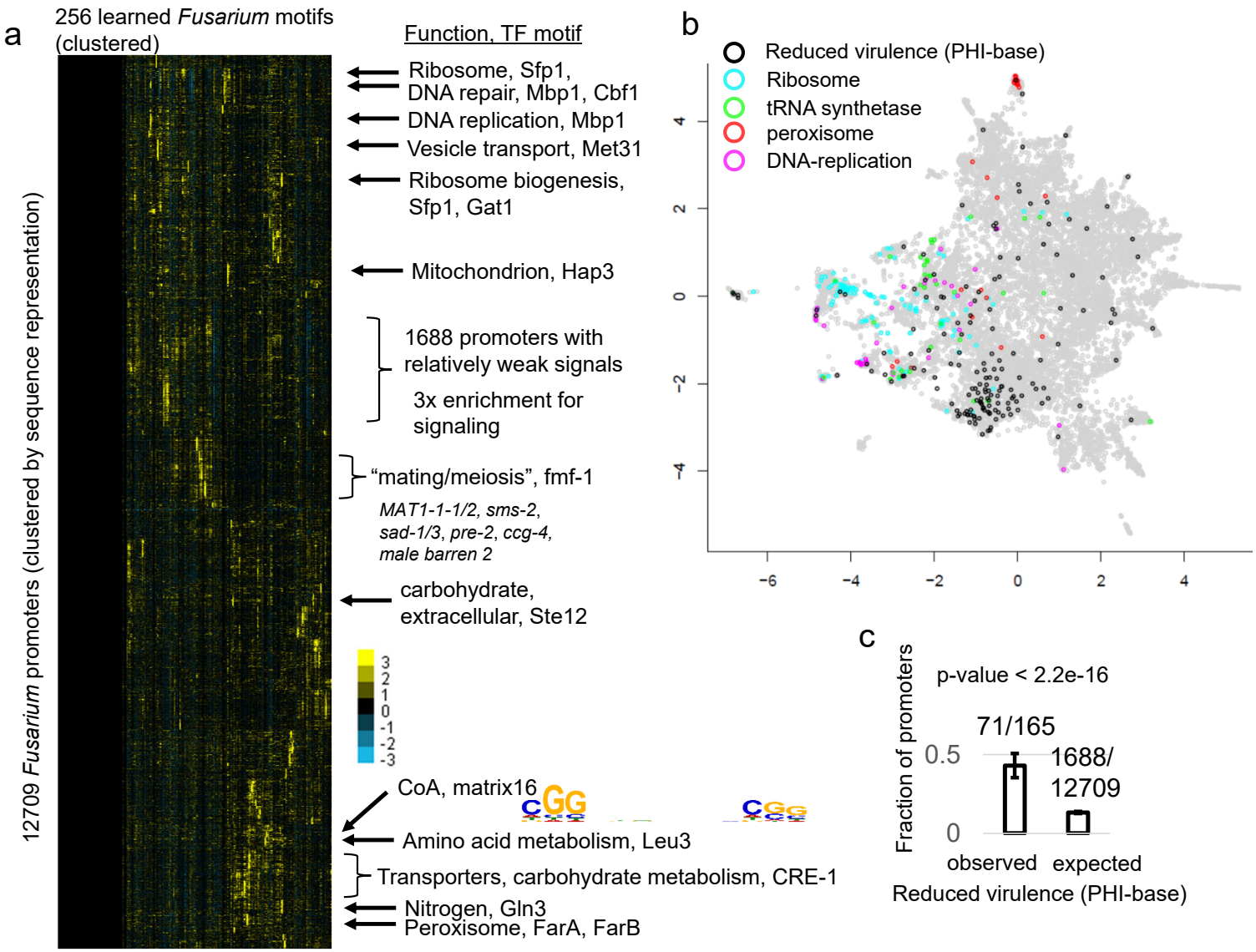

a
JASPAR
Learned motifs
MA0378.1 (SFP1)
E-value= 9.20e-06
Matrix 81
MA0331.1 (MCM1)
E-value= 3.83e-10
Matrix 66
MA0285.1 (CRZ1)
E-value= 8.01e-04
Matrix 29
MA0401.1 (SWI4)
E-value= 5.46e-05
Matrix 121
MA0330.1 (MBP1::SWI6)
E-value= 9.10e-03
Matrix 83
b
TomTom matches (fraction)
121 93 138
47 45 63
33 15 40
27 37 46
19 23 32
JASPAR Fungi
Cis-bp 2.0 (for that species)
Either
S. c. curated high-confidence
S. c.
C. a.
N. c.
F. g.
Learned motifs
c
RHIEPA
244 known motif scores
Phylogenetic average
Netprophet3 (185 TFs)
244 known motif scores
Binding data (182 TFs)
S. c. only
Pearson correlation (1487 deletion experiments)
1 nn
50 nn
Relative expression (regression)
Zero-shot precision (1 prediction per gene)
Top 1
Top 50
Co-expression (classification)
d
PHO pathway
Cytochrome c
Iron
ERG pathway
Meiosis-specific
PHO86 upstream
PHO84 upstream
PHO81 upstream
Pheromone
Ribosome
Heat shock
MET pathway
Proteasome
Cct ring complex
RNA processing/Nucleolus
Amino acid biosynthesis
All other promoters
UMAP 2
UMAP 1


a
256 learned motifs (clustered)
Function, TF motif
5341 budding yeast promoters (clustered by sequence representation)
Ribosome Biogenesis, Dot6, Stb3, Sfp1, Abf1
translation, Dot6, Sfp1, Abf1
Ribosome, Rap1
Budding, Ash1, Hcm1
Actin/endocytocis, Com2
Proteasome,Rpn4
Cct ring complex, Rpn4, Sfp1
Mito. translation Abf1, Sfp1, Nsi1
Heat shock
Iron, Aft1/2
MET pathway, Met4, Tye7
Amino acids, Gcn4, Leu3, Lys14
Cytochrome c, Hap3, Hap1, Urc2
PHO pathway, Pho4
Peroxisome, Oaf1
Stress response, Msn2
Carbohydrates, Zms1
Transporters, Yrr1
Glycolysis, Gcr1
ERG pathway, Upc2, Hap1
Meiosis, Ume6, Ndt80
Spliceosome, mRNA processing
Cell cycle
Mating, Ste12, Yox1
c
b
2.0
1.3
0.7
0.0
-0.7
-1.3
-2.0
1.0
0.7
0.3
0.0
-0.3
-0.7
-1.0
b
25 of the learned motifs 12 matches in JASPAR
629 yeast promoters
Cell cycle synchronized microarrays
Function, TF motif, TF promoter
22/72 site of polarized growth (7x enrichment, P=1.27E-11)
Swi4,Yox1
HCM1, YOX1
SWI6
40/105 DNA replication (14x enrichment, P=4.42E-35)
Mbp1::Swi6
41/103 chromosome segregation (12x enrichment, P=6.40E-33)
Fkh1,Hcm1 FKH2
21/59 cell wall organization (8x enrichment, P=4.45297E-12)
Ash1,Rlm1
Mcm1,Fkh1/2 ACE2, FKH1
SWI4
MCM2-7 complex Yox1
All 4 daughter-specific proteins
Ace2 ASH1
c
matrix 142 (Aft2)
matrix 82 (Aft1)
9 yeast promoters
FRE2
CCC2
FRE3
FIT3
FRE5
NFT1
AFT1
PCA1
FRE8
d
AFT1 upstream
81 orthologs
Aft1 consensus match
GAACTGCACCGGGAA
GAACTGCACCGGGAAA
GAACTGCACCGGGAAA
TAACTGCACCGGGAA
GAACTGCACCGGGAA
GAACTGCACCGGGAA
GAACTGCACCGGGAA
Sensu stricto Saccharomyces
e
TF
Known:
Gcr1
Upc2
Ndt80
Zap1
Yrr1
Predicted:
Aft1
Lys14
Crz1
Hap1


a
256 learned C. albicans motifs (clustered)
Function, TF motif
5625 C. albicans promoters (clustered by sequence representation)
b
Mitotic cell cycle Swi4, Fkh1, Mcm1, Mbp1:Swi6
Proteasome, Rpn4
heat shock
Golgi transport, matrix43/Hac1?
18/97 "meiotic cell cycle"
10x enrichment, P=3.95E-11
14 unknown function
Amino acids, Gcn4, Leu3
Ribosome biogenesis, Stb3, Dot6
Translation, Sfp1
Ribosome, Sfp1, Stb3, matrix 133/Tbf1?
Respiration, Hap3
Mating, Ste12
Phosphate, Pho4
Autophagy, Dal80
Transcription factors, Sum1
Cell wall, Rlm1
Nitrogen, Gat1
Peroxisome, Mig3, Ino2
Glycolysis, Tye7
Oxidative stress, Yap7
Sulfur, Met31, Met4
Ergosterol, Upc2
2.0
1.3
0.7
0.0
-0.7
-1.3
-2.0

b
Matrix 114
Matrix 246
97 C. albicans promoters
CSK1
SPO22, MSH4, DLH1, REC8, SPO1
UME7
IME2


a
256 learned Neurospora motifs (clustered)
Function, TF motif
Carbon sources RNA-seq (n=256)
Life cycle RNA-seq (n=56)
9440 Neurospora promoters (clustered by sequence representation)
Circadian rhythm, asl-1
Nitrogen transport, Gat1
oxidoreductase activity, Mig3, rst2
translation factors
Amino acid metabolism, cpc-1
Peroxisome, FarA/B
Catabolism, matrix 56
cell cycle, Mbp1, Fkh2, Tye7, Rfx1
Golgi transport, matrix 88
Proteasome, matrix 145
Ribosome biogenesis, Sfp1, Gat1
Ribosome, Sfp1,Tbf1?
Signaling
"mating"
>80% unclassified
tsn, sms-2, sad-1/2/3, mlh-2, msh4
Transmembrane transport, vib1, NCU05285, NCU04848/CRE-1?
Carbohydrate metabolism
vib-1, xlr-1, tah-2, NCU04848/CRE-1?
Cell wall, matrix 133
b
c
b
matrix 191
fmf-1
208 Neurospora promoters
>80% unclassified
Male barren-2, Male barren-3
MatA-1
Pre-1, Pre-2, Ccg-4, MatA-2/3
c
matrix 110
59 Neurospora promoters
>80% unclassified
dicer-like 2, rrp3, dynamin family protein Both Mx2 proteins


a
256 learned *Fusarium* motifs (clustered)
12709 *Fusarium* promoters (clustered by sequence representation)
Function, TF motif
Ribosome, Sfp1,
DNA repair, Mbp1, Cbf1
DNA replication, Mbp1
Vesicle transport, Met31
Ribosome biogenesis, Sfp1, Gat1
Mitochondrion, Hap3
1688 promoters with relatively weak signals
3x enrichment for signaling
"mating/meiosis", fmf-1
MAT1-1-1/2, sms-2, sad-1/3, pre-2, ccg-4, male barren 2
carbohydrate, extracellular, Ste12
3
2
1
0
-1
-2
-3
CoA, matrix16
Amino acid metabolism, Leu3
Transporters, carbohydrate metabolism, CRE-1
Nitrogen, Gln3
Peroxisome, FarA, FarB
b
Reduced virulence (PHI-base)
Ribosome
tRNA synthetase
peroxisome
DNA-replication
4
2
0
−2
−4
−6
−4
−2
0
2
4
c
p-value < 2.2e-16
71/165
1688/
12709
0.5
0
Fraction of promoters
observed
expected
Reduced virulence (PHI-base)