8-2013

Mapping functional transcription factor networks from gene expression data

Brian C. Haynes
Washington University School of Medicine in St. Louis

Ezekiel J. Maier
Washington University School of Medicine in St. Louis

Michael H. Kramer
Washington University School of Medicine in St. Louis

Patricia I. Wang
Washington University School of Medicine in St. Louis

Holly Brown
Washington University School of Medicine in St. Louis

See next page for additional authors

Mapping functional transcription factor networks from gene expression data

Brian C. Haynes, Ezekiel J. Maier, Michael H. Kramer, et al.

Genome Res. 2013 23: 1319-1328 originally published online May 1, 2013
Access the most recent version at doi:10.1101/gr.150904.112

Mapping functional transcription factor networks from gene expression data

Brian C. Haynes,[1,2] Ezekiel J. Maier,[1,2] Michael H. Kramer,[1] Patricia I. Wang,[1,2] Holly Brown,[1] and Michael R. Brent[1,2,3,4]

[1]*Center for Genome Sciences and Systems Biology, Washington University, Saint Louis, Missouri 63108, USA;* [2]*Department of Computer Science and Engineering, Washington University, Saint Louis, Missouri 63130, USA;* [3]*Department of Genetics, Washington University School of Medicine, Saint Louis, Missouri 63110, USA*

A critical step in understanding how a genome functions is determining which transcription factors (TFs) regulate each gene. Accordingly, extensive effort has been devoted to mapping TF networks. In *Saccharomyces cerevisiae*, protein–DNA interactions have been identified for most TFs by ChIP-chip, and expression profiling has been done on strains deleted for most TFs. These studies revealed that there is little overlap between the genes whose promoters are bound by a TF and those whose expression changes when the TF is deleted, leaving us without a definitive TF network for any eukaryote and without an efficient method for mapping functional TF networks. This paper describes NetProphet, a novel algorithm that improves the efficiency of network mapping from gene expression data. NetProphet exploits a fundamental observation about the nature of TF networks: The response to disrupting or overexpressing a TF is strongest on its direct targets and dissipates rapidly as it propagates through the network. Using *S. cerevisiae* data, we show that NetProphet can predict thousands of direct, functional regulatory interactions, using only gene expression data. The targets that NetProphet predicts for a TF are at least as likely to have sites matching the TF's binding specificity as the targets implicated by ChIP. Unlike most ChIP targets, the NetProphet targets also show evidence of functional regulation. This suggests a surprising conclusion: The best way to begin mapping direct, functional TF-promoter interactions may not be by measuring binding. We also show that NetProphet yields new insights into the functions of several yeast TFs, including a well-studied TF, Cbf1, and a completely unstudied TF, Edsl.

[Supplemental material is available for this article.]

Genome sequencing is now a routine matter, and the vast majority of protein coding genes in human and major model organisms have been identified. A natural next step is to identify the transcription factors (TFs) that regulate the expression of each gene. Accordingly, a great deal of effort has been devoted to mapping TF networks. The best mapped eukaryotic TF network is that of the budding yeast *Saccharomyces cerevisiae*, where protein–DNA interactions have been identified for most yeast transcription factors by ChIP-chip (Lee et al. 2002; Harbison et al. 2004) and expression profiling has been done on strains deleted for most TFs (Hu et al. 2007; Reimand et al. 2010). Nonetheless, the TF network of yeast is still far from completely mapped. We calculated that at least 97.5% of yeast genes are regulated, in the sense that their transcript levels respond to deletion or overexpression of some TF, but at most 45% respond to perturbation of any TF that is known to bind their promoter regions (see Supplemental Methods). In general, the genes whose promoters are bound by a TF according to ChIP-chip experiments and those whose expression level responds to perturbation of the same TF show little overlap—typically 3%–5% (Gitter et al. 2009). At least 55% of yeast genes do not fall into that overlap for any TF and thus have no functional, direct regulator implicated by available genome-wide data sets.

Algorithms for mapping TF networks from gene expression data have been a focus of intensive research. An annual community effort to evaluate such algorithms, known as DREAM (Dialogue for Reverse Engineering Assessments and Methods), has been held six times (Stolovitzky et al. 2009; Prill et al. 2010). DREAM evaluations have included both simulated data and experimental data on small networks. These evaluations have fostered a motivated computational community, but the results do not appear to have inspired widespread adoption of network mapping algorithms for biological inquiry.

In order to improve the efficiency of TF network mapping from gene expression data, we developed a novel algorithm called NetProphet. In this paper, we describe NetProphet and demonstrate its value for biological inquiry. We compare the networks it infers to existing ChIP-chip networks and to networks inferred by two algorithms that have won DREAM competitions (Inferelator and Genie3) by using the genome-wide data sets available for yeast: expression data on strains in which nearly every TF has been deleted and/or overexpressed, curated databases of TF targets implicated by available ChIP data sets, and databases of position weight matrix (PWM) models for TF sequence specificity. To our knowledge, this is the first evaluation of expression-based network mapping against an objective, comprehensive set of physical TF-DNA interactions. We also take the unusual step of evaluating NetProphet by its ability to discover novel biology.

Algorithms for mapping TF networks from gene expression data can be broadly classified into those that exploit coexpression and those that exploit differential expression (DE). Coexpression approaches can be as simple as measuring the correlation or mutual information between the expression profiles of TFs and potential target genes (Butte and Kohane 2000; Faith et al. 2007). Alternatively, they can use multivariate predictor functions (Bonneau et al. 2006; Marbach et al. 2009; Huynh-Thu et al. 2010). In either

[4]Corresponding author
E-mail brent@wustl.edu
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.150904.112.

case, coexpression analysis can only make predictions for TFs whose expression varies sufficiently in the available expression profiles. Algorithms that exploit DE typically construct networks with regulatory edges from each TF to all the genes that are differentially expressed when the TF is deleted (Hu et al. 2007; Pinna et al. 2010), overexpressed, or otherwise perturbed. Some edges in this network reflect direct binding of the TF to the target (the intended result), while others reflect indirect effects of perturbing the TF (which are considered false positives). In order to reduce the number of edges that reflect indirect effects, edges are often removed until there is only one path from each TF to each target that is affected by perturbing the TF.

Coexpression and DE have complementary strengths and weaknesses. Coexpression can identify targets of TFs that have not been individually perturbed. This greatly increases the range of expression profiling data that can be used and opens up the possibility of analyzing all TFs, regardless of the data set. DE creates variation in the expression of a TF by direct experimental intervention, rather than relying on chance, and it does not risk confusing correlation with causation.

NetProphet capitalizes on the complementarity of the coexpression and DE strategies by combining them. The coexpression component enables NetProphet to exploit any expression data source, including environmental perturbations that affect many TFs simultaneously, and to predict the targets of TFs that have not been individually perturbed in the available expression profiles. The DE component enables NetProphet to exploit the power of targeted TF perturbations (see Results). For every TF and every possible target gene, NetProphet computes a confidence score that combines a number representing its coexpression analysis with a number representing its DE analysis. The coexpression number is a LASSO regression coefficient reflecting the degree to which coexpression patterns allow the level of the putative target to be predicted from that of the TF. This is similar to the calculation done by Inferelator (Bonneau et al. 2006), a method that has performed very well in DREAM assessments (Greenfield et al. 2010; Madar et al. 2010) and that we compare to NetProphet below (see Methods for differences). The DE number is the log odds that the putative target is differentially expressed when the TF is perturbed, given the available replicate expression profiles (see Methods). NetProphet then ranks all possible TF-target interactions by a confidence score that is, roughly speaking, a weighted sum of the coexpression number, the DE number, and their product. This method of integrating coexpression and DE is both simple and, empirically, very effective.

A surprising insight that we gained from these studies is that the most efficient way to start mapping a network of functional TF-DNA interactions may not be by measuring binding directly with methods like ChIP. Our results suggest that an expression-based approach like NetProphet can identify thousands of direct TF-DNA interactions with accuracy at least as good as that of existing ChIP-chip data. Whereas ChIP-chip reveals nonfunctional as well as functional binding, the NetProphet approach reveals only functional binding. Furthermore, the NetProphet approach is easier to scale up than in vivo binding experiments, and it requires less specialized expertise. Binding studies are a valuable complement that make the map more complete, but the most efficient path to a network map may start with the NetProphet approach.

Another surprising finding from this study is that the strength of evidence supporting a gene's response to perturbation of a TF is a good predictor of whether the gene is a direct target of the TF (see Methods). This reveals something about the fundamental nature of signal propagation in TF networks: The response to a TF perturbation tends to dissipate rapidly. This rapid dissipation of the effects of variations in TF level may help explain how cells tolerate noise in the expression of TFs.

Results

We ran NetProphet using a comprehensive collection of microarray expression profiling data on yeast TF deletion mutants (Hu et al. 2007) and analyzed the results as described below (the complete NetProphet output can be found in Supplemental Table S1).

Top NetProphet predictions identify direct binding potential better than ChIP data does

The top 4000 regulatory links predicted by NetProphet comprise 219 distinct TFs regulating 1744 distinct targets. We evaluated these links in a two-stage process. First, we evaluated the potential of each TF to bind each of its predicted targets by using a position weight matrix (PWM) model of its binding specificity. Second, we made the same calculation for all regulatory links implicated by ChIP experiments, regardless of NetProphet score. We then set high-, medium-, and low-stringency thresholds for support by each PWM. At each stringency level, the percentage of ChIP-implicated targets supported by PWMs serves to calibrate the NetProphet results, allowing us to determine what should count as a good outcome for NetProphet. Of the top 4000 regulatory links predicted by NetProphet, 1408 involved a regulator that had been studied by ChIP-chip or ChIP-seq (Lee et al. 2002; Harbison et al. 2004; Balaji et al. 2006; Abdulrehman et al. 2011) and for which a PWM was available in the UNIPROBE database (Gordan et al. 2011; Robasky and Bulyk 2011). These PWMs are based on in vitro data obtained using protein-binding microarrays (PBMs), so they are not influenced by either gene expression data or ChIP data. We calculated the percentage of these predictions that are supported by high binding potential using each of the three stringency levels for PWM support (Fig. 1). A TF-target relationship was counted as

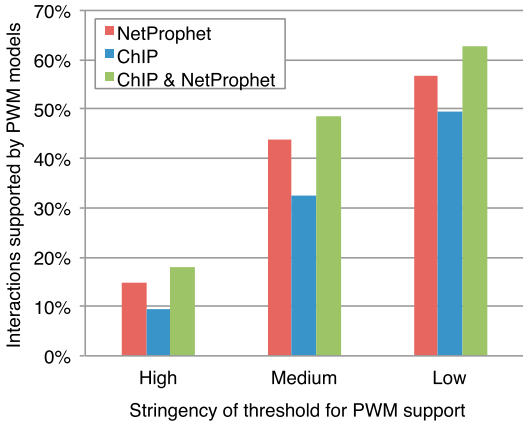

Figure 1. TF-promoter binding potential for the top 4000 NetProphet predictions (red), all direct targets implicated by ChIP hits in the Yeastract or Tnet data bases (blue), and targets implicated by ChIP hits that are also predicted by NetProphet (green). The high stringency threshold for each PWM was set such that ~10% of ChIP-implicated targets have PWM scores exceeding the threshold and hence count as "PWM supported." The medium and low stringency thresholds were set such that ~33% and ~50% of ChIP-implicated targets for each TF have PWM scores exceeding the threshold, respectively. For the high, medium, and low stringency PWM cutoffs, chance inclusion was 6.4%, 22.1%, and 36.8%, respectively.

supported when the promoter of the target contained a single high-scoring PWM hit or a number of individually significant hits whose total score was exceptionally high (see Methods). Similar results were obtained using PWMs based on in vivo data and manual curation (see Supplemental Fig. S1; Spivak and Stormo 2012).

We also created 1000 random networks of the same topology by randomly permuting node labels in the adjacency matrix. Only two of these had as many PWM-supported interactions as NetProphet, indicating that NetProphet's strong performance is based on choosing specific interactions with above-chance accuracy ($P = 0.002$), in addition to choosing a good network topology.

We were surprised to find that the top NetProphet predictions are supported by binding potential at a higher rate than the ChIP-implicated interactions (Fisher's exact $P < 10^{-9}$ at the most stringent level of PWM support). NetProphet predictions are also supported by conserved binding sites at a higher rate than ChIP hits (Fisher's exact $P < 10^{-8}$) (Supplemental Fig. S2). This is remarkable, given that NetProphet uses only gene expression data and knows nothing about binding, PWMs, or promoter sequences, while ChIP experiments measure binding directly. ChIP hits that are also NetProphet predictions are much more likely to be supported by high binding potential than ChIP hits in general (Fig. 1), indicating that NetProphet analysis adds value for predicting direct targets of TFs that have already been subjected to ChIP.

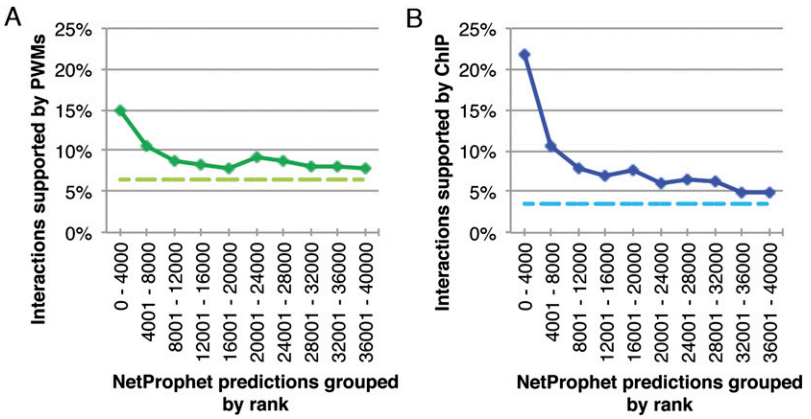

Figure 2. Evidence supporting NetProphet predictions as a function of NetProphet rank. (*A*) Percentage of predictions supported by binding potential at the high stringency threshold (green, solid) and expectation for randomly selected targets (green, dashed). (*B*) Percentage of NetProphet predictions supported by ChIP hits (blue, solid) and expectation for randomly selected targets (blue, dashed). All points represent groups of NetProphet predictions that are significantly enriched for predictions with PWM support (panel *A*) or ChIP support (panel *B*), $P < 0.05$.

NetProphet rank is predictive of support by binding potential and by ChIP results

We wanted to investigate the degree to which the rank assigned to a potential TF-target link by NetProphet corresponds to its likelihood of being supported by independent evidence. Thus, we calculated PWM support and ChIP support as a function of NetProphet rank. Using the high-stringency threshold for PWM support, we calculated the fraction of predictions supported for the top 4000 NetProphet predictions, the next 4000, and so on (Fig. 2A, solid green line). These are incremental evaluations of predictions in a given range of NetProphet ranks, not cumulative evaluations of all predictions above a given rank. There is a clear trend in Figure 2A: potential links that are ranked more highly by NetProphet are more likely to be supported by PWM evidence. The analysis of ChIP support as a function of NetProphet rank shows the same trend (Fig. 2B). Thus, NetProphet rank is a good predictor of PWM and ChIP support.

The top 40,000 NetProphet predictions are enriched for PWM and ChIP support

To determine the lowest rank at which NetProphet predictions are better than chance, we tested each block of 4000 predictions for significant enrichment in either PWM-supported interactions or ChIP-supported interactions. The results showed significant enrichment for both PWM support and ChIP support in all blocks of edges down to rank 40,000, below which we did not test (Fig. 2). For TFs that have been subjected to both ChIP and PBM studies, 9% of the targets in the top 40,000 NetProphet predictions are supported by PBM-derived PWMs. For the same TFs, 9% of the targets in the top 29,946 ChIP-implicated interactions are supported by PBM-derived PWMs. Thus, the complete NetProphet network may contain as many direct interactions as the complete ChIP-implicated network.

Additional data improves lower ranked predictions

Previous studies have shown that TF overexpression data can be complementary to deletion data and can help to identify direct targets (Chua et al. 2006; Sopko et al. 2006), so we added profiling data from cells grown in YPGal and overexpressing each of 55 TFs (Chua et al. 2006). Since many genes are only expressed in stress conditions, we also added profiling data from time courses of responses to multiple stressors (Gasch et al. 2000). Combined, these additional data sets resulted in a marginal improvement in the fraction of the top 4000 predictions supported by PWMs and ChIP (Supplemental Fig. S3). Predictions ranked 24,000–40,000 showed the greatest improvement.

To investigate the effects of having smaller data sets, we ran NetProphet on subsamples of the data from Hu et al. (2007), comprising expression profiles from deletions of one-fourth, one-half, or three-fourths of the TFs. For each input data set, we considered TFs whose deletion profiles were or were not included in the data separately (Supplemental Methods). For TFs whose deletion profiles were included in the input data, adding more data comprising deletion profiles for other TFs did not improve their predicted targets (not shown). For TFs whose deletion profiles were not included in the input data, target prediction was less accurate overall but adding more data improved prediction accuracy (as assessed by PWM support) at all NetProphet ranks (Supplemental Fig. S4); ChIP support did not change (data not shown).

To gain further insight into the interplay between data-set size and NetProphet rank, we evaluated NetProphet's LASSO regression and DE components separately. The results showed that DE performs substantially better than LASSO in terms of its top 4000 picks, slightly better on ranks 4001–20,000 (at least by PWM support), and about the same on 20,001–40,000 (Supplemental Fig. S5). By

combining DE and LASSO, NetProphet performs somewhat better than DE on the top 4000 and on 20,001–40,000, and about the same as DE on 4001–20,000. Thus, LASSO regression enhances NetProphet accuracy in some rank ranges more than others, and the extent of its contribution depends on how much data is available.

NetProphet rank predicts PWM and ChIP support better than other methods tested

We compared the top 4000 NetProphet predictions to the top 4000 predictions from two other publicly available methods that have won DREAM competitions, Inferelator (Bonneau et al. 2006; Greenfield et al. 2010; Madar et al. 2010) and Genie 3 (Huynh-Thu et al. 2010). The metrics used were percentage of interactions supported by high-stringency PWM evidence, ChIP evidence, or both (Fig. 3A). The NetProphet predictions received substantially more support by all three metrics. To determine whether this good performance was broadly distributed across TFs, we considered the five most highly ranked targets of each TF and asked whether at least one of the five was supported by PWM evidence, ChIP evidence, or both. NetProphet was able to identify at least one correct target for many more TFs than either of the other methods tested (Fig. 3B). We found similar results when comparing NetProphet to Inferelator and Genie3 on simulated data from the DREAM4 evaluation (Supplemental Fig. S6).

NetProphet predictions identify novel TF functions that cannot be found by ChIP

After we validated NetProphet by both PWM and ChIP support, we used it to predict TF functions that could not be identified using existing ChIP data. For each TF, we performed Gene Ontology (GO) enrichment analysis on all its targets in the top 10,000 interactions ranked by NetProphet (see Methods and Supplemental Table S2). We then removed any GO terms that were enriched among the targets implicated by existing ChIP data. Since there can be many overlapping GO terms for a gene set, some of which are extremely broad, we chose a single, specific GO term to represent sets of redundant terms for the same genes. The results include 44 functional annotations on 42 TFs.

Of the 44 functional annotations that could not be derived from existing ChIP data, 29 (66%) were supported in the literature by non-ChIP evidence, such as mutant phenotypes and protein-protein interactions (Supplemental Table S3). Examples include involvement of *GCR1* and *GCR2* in regulating hexose catabolism (Clifton et al. 1978), *DIG1* in pheromone response (Tedford et al. 1997), *PHO2* in polyphosphate metabolism and dephosphorylation (Ogawa et al. 2000), and *SFP1* in glycolysis (Cipollina et al. 2008). In two cases, we could find evidence supporting the function in other species but not in *S. cerevisiae* (Table 1). For example, NetProphet predicts that *RIM101* regulates iron metabolism in *cerevisiae*. The orthologous TF in *Cryptococcus neoformans*, a pathogenic fungus of phylum Basidyomicota, is known to regulate iron metabolism (O'Meara et al. 2010). Thirteen of the functional predictions are completely novel (Table 1), including regulation of cytokinesis by *HAP2, HAP4*, and *HSF1*, response to nitrogen starvation by *SNF6*, and lysine biosynthesis by *EDS1* (discussed below).

NetProphet identifies a novel biological function for Cbf1

NetProphet predicts that centromere binding factor 1 (Cbf1) activates genes involved in phosphate acquisition, including all three repressible acid phosphatases (*PHO5*, *PHO11*, *PHO12*) and a repressible alkaline phosphatase (*PHO8*). Cbf1 has long been known to regulate sulfate metabolism (Kuras et al. 1996) but not phosphate metabolism, and the Cbf1 ChIP studies in the curated databases used for our computational analyses detected no significant binding of Cbf1 to these phosphatase promoters. Recently, it was found that Cbf1 binds the promoters of genes in the Pho4 regulon and regulates their expression (Zhou and O'Shea 2011). However, Zhou and O'Shea found that Cbf1 *represses* acid phosphates while NetProphet predicts that it *activates* them. Following up on this apparent discrepancy, we noticed that Zhou and O'Shea's experiments involved cultures grown in synthetic complete medium with 10 mM inorganic phosphate (SC+10 mM Pi), whereas the NetProphet predictions were based on microarray data from cultures grown in YPD (Hu et al. 2007). YPD has ample organic phosphate that can be liberated by phosphatases but relatively little free (i.e., inorganic) phosphate. To validate the predicted novel function of Cbf1 as an *activator* of phosphatases, we assayed the expression of acid phosphatases from cultures of the same strains used by Zhou and O'Shea, grown in either SC+10 mM Pi or YPD (see Methods). The results showed that Cbf1 is a repressor of acid phosphatases in SC+10 mM Pi, as reported by Zhou and O'Shea (2011), but an activator of acid phosphatases in YPD, as predicted by NetProphet (Fig. 4). Thus, rather than simply serving as a repressor, Cbf1 serves to amplify the effect of medium on expression of phosphatases. The activating role of Cbf1 explains the fact that the Cbf1 deletion mutant grows slowly in conditions that derepress phosphatases but not in conditions that repress

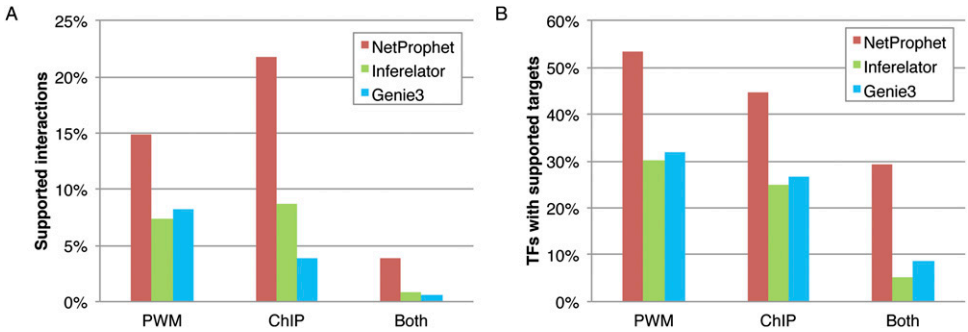

Figure 3. Evaluation of top 4000 predictions from NetProphet (red), Inferelator (green), and GENIE3 (blue). (*A*) For each method, percentage of predictions supported by binding potential at the high stringency PWM threshold, ChIP data, or both. (*B*) Percentage of TFs for which at least one of the top five targets predicted by each method is supported by either binding potential or ChIP hits or both. (See Supplemental Methods for additional details.)

Table 1. Novel TF functions predicted by NetProphet but not by existing ChIP data

Name	NP #	GO #	Overlap	*P*-value	Description
EDS1	18	10	7	1×10^{-16}	Lysine biosynthesis
CST6	595	167	49	1×10^{-11}	Cytoplasmic translation
STB1	18	9	4	1×10^{-8}	Siderophore transport
SWI4	40	250	11	9×10^{-7}	Cell wall organization/ biogenesis
RIM101	62	80	8	2×10^{-6}	Metal ion transport
SNF6	351	6	5	6×10^{-6}	Cellular response to nitrogen starvation
HSF1	66	11	4	6×10^{-6}	Cytokinesis, completion of separation
STP1	27	80	5	3×10^{-5}	Metal ion transport
SNF2	245	6	4	5×10^{-5}	Cellular response to nitrogen starvation
RPN4	60	98	7	8×10^{-5}	Response to pheromone
SIN4	505	34	11	1×10^{-4}	Glycolysis
HAP2	67	114	7	4×10^{-4}	Cytokinesis
ASH1	9	98	3	4×10^{-4}	Response to pheromone
AFT1	112	8	3	4×10^{-4}	Lysine biosynthesis
HAP4	18	11	2	5×10^{-4}	Cytokinesis, completion of separation

(NP #) Number of targets predicted by NetProphet; (GO #) number of yeast genes in the indicated Gene Ontology (GO) Biological Process category; (Overlap) number of genes in both the NetProphet target set and the GO category; (*P*-value) probability of such a large overlap occurring by chance; (Description) description of the GO category.

phosphatases—if Cbf1 merely repressed phosphatases, then it should not be needed in derepressing conditions.

NetProphet identifies a novel biological function for Eds1

Eds1 ("Expression dependent on Slt2") is a putative zinc finger TF homologous to Rgt1, a glucose-inactivated repressor of glucose transporter genes. Eds1 currently has no known function and, to the best of our knowledge, is not the subject of any papers. Existing ChIP data on Eds1 identifies only three significant targets that do not appear to have any common function. However, NetProphet predicts that Eds1 is a highly specific repressor of lysine biosynthesis. The top seven predicted targets of Eds1 are all in the pathway for conversion of cytosolic citrate to lysine (Fig. 5A). Furthermore, these seven targets encode proteins that catalyze seven of the eight steps required for conversion of citrate to lysine (GO enrichment $P < 10^{-16}$). One of these steps is also part of the TCA cycle. To follow up on this prediction, we computed the total score of all significant matches to the Eds1 PWM in the promoters of all yeast genes. Among all genes, the percentiles of the seven targets predicted by NetProphet were 97% (*LYS4*), 97% (*LYS9*), 97% (*CTP1*), 87% (*ACO2*), 77% (*LYS12*), 27% (*LYS21*), and 26% (*LYS1*). The mean percentile was 72.5% and the probability of picking seven genes at random with such a high mean percentile is <0.02, suggesting that Eds1 binds at least some of these promoters directly.

In follow-up experiments, we used QuantiGene to assay the expression of *EDS1*, *LYS4*, *LYS9*, *CTP1*, *ACO2*, and *LYS12* in a wild-type strain (BY4741), the *eds1Δ* mutant from the yeast deletion collection (YBR033W), and *eds1Δ* complemented with *EDS1* under the control of the $tet0_2$ promoter (Gari et al. 1997). The results verified that *EDS1* is deleted in YBR033W and that each of the five lysine-associated genes is induced three- to fourfold relative to WT (Fig. 5B). Furthermore, *eds1:Ptet02-EDS1* complemented the *eds1Δ* mutant, restoring WT expression levels to the five lysine genes and demonstrating that their induction in *eds1Δ* was caused by the loss of *EDS1* rather than a collateral lesion.

Discussion

Mapping TF networks is a longstanding goal in genomics and computational biology. Currently, the two chief sources of experimental data for systematic network mapping are gene expression data and in vivo TF-binding data. *S. cerevisiae* is the only eukaryote for which we have nearly complete data sets of both types. However, these data sets do not agree very well, with typical overlaps of 3%–5% between the genes that respond to deletion of a TF and those that the TF binds in ChIP assays. For the majority of TFs (52%), the binding motif indicated by protein binding arrays is not significantly enriched in the targets indicated by ChIP-chip (Gordan et al. 2009). The failure of these methods to yield a consistent map of functional, direct regulatory interactions leaves us without an efficient, systematic means of producing such maps.

We set out to produce a practical, efficient, and systematic method of TF network mapping that would be attractive for global mapping (identifying direct targets of all TFs encoded in a genome) or for mapping subnetworks that regulate specific physiologic responses. We assumed that any such mapping effort would require generating expression and/or binding data. We also assumed it would have a limited budget and might be carried out by scientists whose main focus is not genomic technologies.

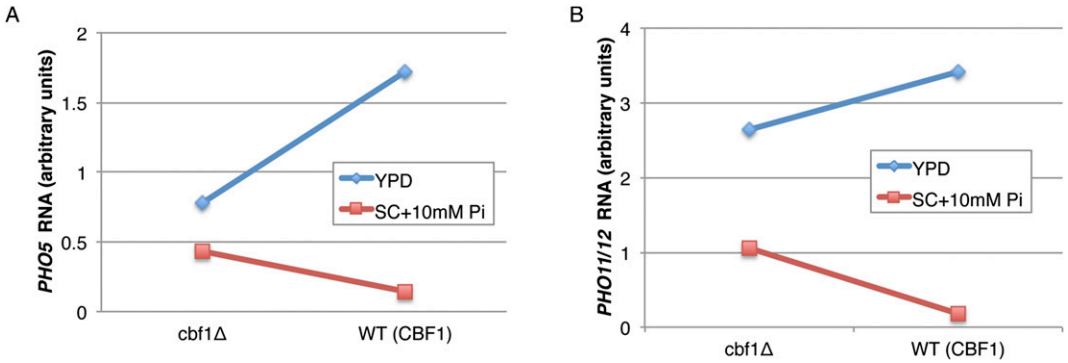

Figure 4. Effect of Cbf1 on expression of acid phosphatases in cells grown in synthetic complete medium with 10 mM inorganic phosphate (red) or on cells grown in YPD (blue). (*A*) Expression of *PHO5*. (*B*) Combined expression of *PHO11* and *PHO12*, which have so much sequence similarity that we were not able to distinguish their transcripts. Error bars representing one standard error of the mean of two technical replicates were too small to be seen in the figure.

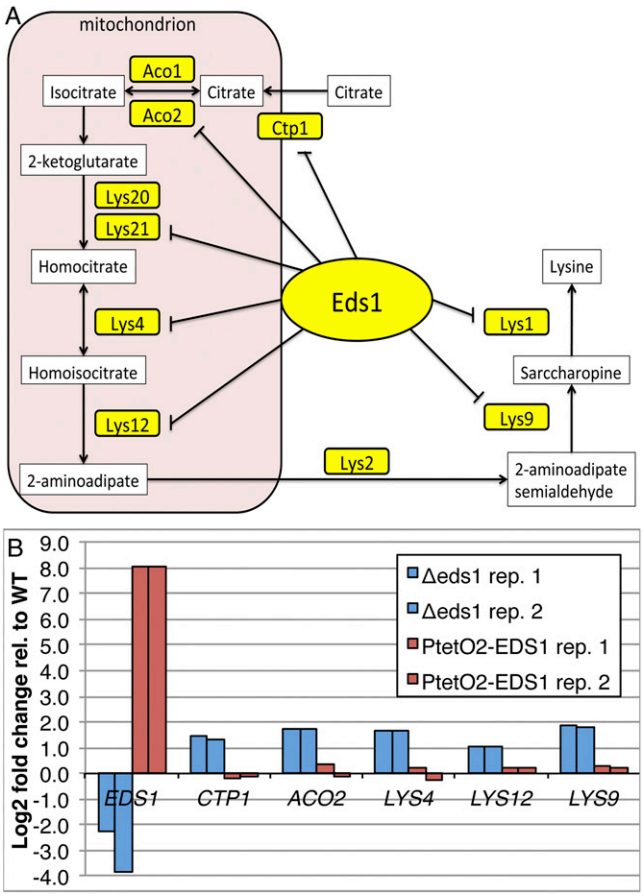

Figure 5. (*A*) The top seven NetProphet-predicted targets of Eds1 are all in the pathway for conversion of cytosolic citrate to lysine. (*B*) Log_2-fold change of *EDS1* and five lysine biosynthesis genes relative to wild-type cells in an eds1 deletion mutant (two replicate cultures) and the same deletion mutant complemented by overexpression of Eds1 under the control of the $tetO_2$ promoter (cultures of two independent transformants).

The studies described above suggest that measuring physical binding with methods such as ChIP may not be the most efficient way to begin mapping a network of direct, functional TF-target interactions. We described, validated, and applied a different approach based on perturbation of TF expression levels, followed by expression profiling and NetProphet analysis. This approach has several advantages over in vivo binding experiments:

1. For most organisms that can be grown in a laboratory, perturbation of TF expression (by disruption, RNA interference, or overexpression) and genome-wide expression profiling (by microarray or RNA-seq) are basic, commodity methods that can be scaled to dozens or hundreds of TFs without specialized expertise or outsized budgets. Although great strides have been made in in vivo binding analysis using methods such as ChIP-seq (Johnson et al. 2007; Landt et al. 2012), Bio-ChIP (van Werven and Timmers 2006), or Calling Cards (Wang et al. 2012), these are challenging experiments that are not easily scaled to many TFs outside of specialized genomics labs.
2. Most of the TF binding sites revealed by in vivo binding experiments show no evidence of affecting transcription rates. Even after binding experiments have been carried out, TF perturbation and expression profiling are required for determining which of the genes that are bound by a TF are also regulated by it.
3. Whereas binding data contain limited information about functional regulation, expression data carry substantial information about binding. For example, the top 4000 interactions predicted by NetProphet were supported by PWM models of binding specificity more often than the interactions indicated by significant ChIP hits were (Fig. 1). Even the top 40,000, taken as a group, were supported by PWM models at the same rate as the complete set of just under 30,000 ChIP-supported interactions in the TNET and YEASTRACT databases.

Building a more comprehensive network

Once an initial map has been built by the NetProphet approach, binding studies can be used to make it more comprehensive. Several algorithms have been developed for inferring networks by combining multiple data types. For example, Marbach et al. (2012) and Beyer et al. (2006) both combined coexpression data with in vivo binding data from ChIP, evolutionary conservation, and PWM models of binding specificity. The latter can be derived by several experimental methods, including protein binding arrays (Zhu et al. 2009), Selex (Jolma et al. 2010), and yeast one-hybrid assays (Reece-Hoyes et al. 2011). Interestingly, the analysis by Beyer et al. (2006), which included a comprehensive yeast ChIP data set (Harbison et al. 2004), identified only 5245 TF-target relationships that they considered "high confidence." Neither this analysis nor that of Marbach et al. (2012) used differential expression data, so adding NetProphet scores to the inputs they combine has the potential to enhance their results substantially.

Key insight behind NetProphet

To determine whether the method of analyzing expression profiles from TF perturbation studies matters, we applied NetProphet and two highly regarded methods (Inferelator and Genie3) to the same expression data set. The results showed a substantial accuracy advantage for NetProphet (Fig. 3). Much of this advantage is due to NetProphet's global ranking of interactions by the likelihood that the target is differentially expressed when the TF is perturbed. The key insight is that *the effect of deleting a TF is strongest on its direct targets and diminishes rapidly as it propagates through the network*. This does not mean that only direct targets are differentially expressed or even that the majority of the differentially expressed genes are direct targets. Indeed, the typical approach to DE analysis—setting a threshold for significance and treating all significant targets as equal—is not very useful for network mapping. However, the direct targets show a *stronger* differential expression effect, on average, than indirect targets. Because it exploits the *strength* of differential expression, NetProphet does not need an explicit mechanism for dealing with the fact that many indirect targets are differentially expressed at statistically significant levels. Algorithms based on a strength-of-DE ranking have been tried on simulated data (Greenfield et al. 2010; Pinna et al. 2010; Yip et al. 2010), but the studies presented here are the first to demonstrate that this approach is useful for inferring real TF networks from real gene expression data.

The effects of data-set size and the importance of combining DE with regression

We carried out a number of experiments to answer two intertwined questions: "How important is it to have expression profiles for a lot

of TF perturbations?" and "How important is it to use regression as well as global ranking by probability of DE?" The results indicate that NetProphet effectively predicts the targets of TFs whose deletions have been profiled even if the number of such TFs is small. When the number of TF-deletion profiles is small, as it may be in many focused studies of organisms other than yeast, NetProphet's effectiveness derives primarily from its DE-based ranking algorithm. Adding deletion profiles of additional TFs to a data set improves accuracy in two ways. First, it greatly improves the accuracy of target prediction for the TFs deleted in the new data. Second, it improves the accuracy of target prediction for all TFs for which a deletion profile is not available. The latter effect is due entirely to NetProphet's LASSO regression algorithm. In Supplemental Table S1, which shows NetProphet predictions based on all available deletion profiles, there are many ChIP-supported predictions for TFs whose deletion profiles are not available. Such predictions could not be made by DE alone. The regression component also makes it possible to improve accuracy by including expression profiles that do not involve single-TF perturbations, such as profiles of strains grown in stress conditions.

Identifying novel TF functions

One of the applications of genome scale network mapping is discovering the functions of TFs based on Gene Ontology (GO) terms that are significantly overrepresented in the annotations of their predicted targets. (In less-well-studied organisms, many target genes can be confidently assigned GO terms by orthology.) Applying this approach to NetProphet predictions produced significant enrichment of biological process terms for 239 of the 263 *S. cerevisiae* TFs (Supplemental Table S2). Even when we excluded GO terms that were significant in ChIP-implicated targets of each TF, we were left with 44 functions assigned to 42 TFs, of which 66% were supported by mutant phenotypes or protein-protein interactions and 34% were novel. This suggests that the majority of such inferences are correct and highlights the fact that NetProphet analysis provides valuable information even after most TFs have been ChIPed.

One novel prediction is that Cbf1 would activate expression of certain phosphatases when cells are grown in YPD. We verified this experimentally and showed that Cbf1 can best be understood as amplifying the effect of inorganic phosphate availability on the expression of phosphatases, rather than simply repressing or activating them in response to phosphate availability. Another novel prediction is that, Eds1, a TF with no known function, is a highly specific repressor of genes involved in lysine biosynthesis (including one that encodes an enzyme in the TCA cycle). We identified strong binding sites for Eds1 in the promoters of five of these genes, and verified experimentally that deletion of *EDS1* results in substantial up-regulation of these five and complementation of the *EDS1* deletion restores wild-type expression levels. Eds1 is the closest homolog of Rgt1, also a repressor, which regulates glucose transporters. The promoter of *EDS1* contains a significant binding site for Mig1/Mig2, major enforcers of glucose repression whose targets also include glucose transporters, and *EDS1* is differentially expressed in the *mig1*, *mig2* double mutant but not in either single mutant (Westholm et al. 2008). Thus, Eds1 may link lysine biosynthesis to glucose availability and the fermentative/oxidative balance.

Testing other applications of network models

By comparing NetProphet predictions to ChIP data and to PWM models of binding potential, we have demonstrated the application of network models to predicting TF binding. Other important applications for future studies are predicting gene expression and phenotype in response to new TF perturbations. A good benchmark for predictive power is the ability to predict gene expression and phenotype in double TF mutants using only data from single mutants. Ideally, a network map should predict epistatic interactions and thereby make it possible to predict expression in a double mutant more accurately than the average of the expression profiles of the two single mutants. Developing systematic network mapping methods that reliably achieve this will require innovations in both mapping and quantitative modeling. In the meantime, network maps produced by the current version of NetProphet correctly predict functional TF-promoter interactions and overall TF functions using only gene expression data that can be generated by scalable, commodity methods within modest budgets.

Methods

Materials

All chemicals were from Sigma-Aldrich. All kits were used according to manufacturer recommendations unless otherwise specified.

Strains and growth conditions

All strains used in the *CBF1* study, including *cbf1*Δ and the wild-type parent strain, have been described (Zhou and O'Shea 2011) and were kindly provided by the O'Shea lab.

Yeast microarray normalization and quantification

The microarray data used to map the yeast transcriptional network was originally published in Hu et al. (2007) and later reanalyzed in Reimand et al. (2010). We downloaded the raw GenePix files for each of the 588 microarrays from the Longhorn Microarray Database (Killion et al. 2003) and normalized this data following the scheme described by Reimand et al. with minor modifications (see Supplemental Methods).

ChIP data curation

A list of protein-DNA interactions in *S. cerevisiae* was compiled from two sources: TNET (Babu et al. 2004) and YEASTRACT (Abdulrehman et al. 2011). Taken together, TNET and YEASTRACT include the results of several major ChIP studies (Svetlov and Cooper 1995; Horak et al. 2002; Lee et al. 2002; Harbison et al. 2004; Luscombe et al. 2004; Teichmann and Babu 2004) as well as many smaller studies. YEASTRACT is regularly updated to include new ChIP data as it becomes available. TNET contains 12,873 interactions over 157 TFs and 4410 target genes. YEASTRACT contains 28,145 ChIP-supported interactions over 160 TFs and 5683 target genes. There was good agreement between these two sources, with 11,073 interactions in common. We used the union of TNET and YEASTRACT, which contained 29,945 interactions covering 184 TFs and 5790 target genes. This union is referred to below as the interactions in curated ChIP studies.

Estimating binding potential with position weight matrices

To establish binding site evidence for an interaction, the curated PWMs from the UNIPROBE database (Gordan et al. 2011; Robasky and Bulyk 2011) were scanned over the yeast promoters using FIMO (Grant et al. 2011). A gene's promoter region was defined as extending from 800 bases upstream of the transcription start site

(excluding sequence from neighboring open reading frames) to 200 bases downstream from the transcription start site. Binding sites that were identified by FIMO at a P-value < 0.005 were considered in subsequent analyses. Two models of binding were considered. For each TF, the strong site model ranks promoters containing one or more significant binding sites according to the negative log P-value of the most significant site. The weak site model ranks promoters containing one or more significant binding sites by the sum of the negative log P-values for all significant sites in a promoter. If a promoter is ranked in the top K promoters for one or both binding models, it is considered to be supported by binding potential evidence. Reasonable values of K are determined on a TF by TF basis. For each TF, K is set such that a certain fraction of ChIP interactions for that TF are recovered in the top K ranked promoters. Three different binding potential stringencies were considered using this formulation: high (~10% ChIP interactions recovered), medium (~33% ChIP interactions recovered), and low (~50% ChIP interactions recovered). At the high, medium, and low stringencies, the random baselines (percentage of random TF-gene pairs supported by binding potential) are 6.4%, 22.1%, and 36.8%, respectively. This process was repeated using the curated PWMs from the ScerTF database (Spivak and Stormo 2012), for which the random baselines at the high, medium, and low stringencies are 5.7%, 20.4%, and 35.9%, respectively.

NetProphet

The NetProphet method for inferring a transcriptional network from a gene expression data set combines the scores from two independent analyses. The first analysis is a sparse linear model that predicts the expression of each gene as a function of the expression of one or more transcriptional regulators. The second analysis assesses differential expression on each expression profile in the data set in which a specific regulatory gene has been perturbed (via knockout, knockdown, or overexpression) compared to wild-type control in the same growth condition. The two analyses are combined through a weighted model averaging scheme.

LASSO regression is used to learn a sparse linear model that predicts the jth gene's expression level in measurement k, Y_{jk}, from a weighted combination of the regulators' expression levels in measurement k:

$$Y_{jk} = \sum_i B_{ij} \cdot X_{ik},$$

where the X_{ik} is the expression level of TF i in measurement k and B_{ij} is the coefficient that the LASSO procedure learns to describe the influence of regulator i on gene j. (In other words, the X_{ik} are the subset of the Y_{jk} for which gene i encodes a TF.) LASSO chooses the influence coefficients that minimize the sum of the squared prediction errors and a term called the L1 penalty:

$$\underset{B}{\operatorname{argmin}} \sum_{jk} (Y_{jk} - (\sum_i B_{ij} \cdot X_{ik}))^2 + t \sum_{ij} |B_{ij}|.$$

The L1 penalty is a "shrinkage term" that keeps the model sparse to avoid overfitting. The L1 penalty is scaled, relative to the sum of squared prediction errors, by a parameter t. When t is 0, the optimization of B is equivalent to ordinary least squares regression. As t grows, components of B are forced to zero, yielding a sparser solution. In this application, we disallow autoregulation by prohibiting B_{ij} from becoming nonzero when regulator i is encoded by gene j. Unlike the typical LASSO regression carried out by Inferelator, NetProphet uses a single, global weighting parameter t for all target genes. This reduces the number of learned parameters by more than 5000, thereby reducing the risk of overfitting the data. Parameter t is determined using 10-fold cross-validation, minimizing predictive error. We handle gene perturbations in the regression by omitting measurements in which gene j has been perturbed when fitting the coefficients B_{ij}.

DE analysis is used to rank potential regulator-target interactions based on the estimated probability that each gene changes expression in response to the regulator deletion according to a log-odds statistic:

$$L_{\Delta i}(j) = \log\left(\frac{\Pr(F_{\Delta i}(j) \neq 0)}{\Pr(F_{\Delta i}(j) = 0)}\right),$$

where $F_{\Delta i}(j)$ is the true $\log_2$-fold change of gene j in the wild type relative to a Δi background. We compute $L_{\Delta i}(j)$ using LIMMA, which estimates the posterior log odds score by way of a moderated t-statistic in which gene specific variances are shrunk toward a common value as described (Smyth 2004, 2005). A signed confidence score D_{ij} is assigned to each potential regulator-target interaction as follows:

$$D_{ij} = \begin{cases} L_{\Delta i}(j) \cdot \operatorname{sgn}(Y_{\Delta i}(j)) & L_{\Delta i}(j) > 0 \\ 0 & \text{otherwise} \end{cases}.$$

D_{ij} represents the signed confidence score that regulator i directly regulates gene j, and $Y_{\Delta i}(j)$ is the observed $\log_2$-fold change of gene j in the wild type relative to a Δi background. The sign of D_{ij} indicates whether regulator i is repressing or activating gene j. When it is more likely that gene j's expression is unchanged in the Δi background (i.e., $L_{\Delta i}(j) < 0$), the interaction is assigned a confidence score of 0. If the gene expression compendium contains no measurements for the Δi strain, D_{ij} is set to zero for all j.

LASSO regression and DE analyses are combined using a model averaging scheme. Before combining the score matrices $\boldsymbol{B}$ (from regression) and $\boldsymbol{D}$ (from differential expression), each matrix is normalized such that its values lie on the interval $[-1,1]$. After normalization the combined scores, M_{ij}, are computed as follows:

$$M_{ij} = \left(|B_{ij}| + c_b\right) \cdot \left(|D_{ij}| + c_d\right) \cdot \omega_{R(B_{ij},D_{ij})},$$

where c_b and c_d are constants that prevent M_{ij} from becoming zero when only one of the two individual scores is zero. The product of the two analyses is scaled by $\boldsymbol{\omega}$, which is a six-component vector that weights the average differently according to the signs of B_{ij} and D_{ij} which indicate the predicted regulatory influence (activating, repressing, or no influence). $\boldsymbol{\omega}$ is indexed as a function of B_{ij} and D_{ij} as follows:

$$R(B_{ij}, D_{ij}) = \begin{cases} I & B_{ij} > 0;\ D_{ij} > 0 \\ II & B_{ij} < 0;\ D_{ij} > 0 \\ III & B_{ij} < 0;\ D_{ij} < 0 \\ IV & B_{ij} > 0;\ D_{ij} < 0 \\ B & B_{ij} \neq 0; D_{ij} = 0 \\ D & B_{ij} = 0; D_{ij} \neq 0 \end{cases}.$$

In the absence of training data, c_b and c_d are set to 0.01 and $\boldsymbol{\omega}$ is set to 1. Otherwise, the offset coefficients c_b and c_d and the weight vector $\boldsymbol{\omega}$ are learned by cross-validation on the training data (labeled interactions), maximizing the area under the precision recall curve.

GO enrichment analysis to detect novel TF function

GO process enrichment was performed using R Bioconductor package GOstats (Falcon and Gentleman 2007). GO process enrichment was performed for each TF's targets as predicted by NetProphet, using the top 10,000 predicted targets. GO process enrichment was also performed over the set of targets identified for each TF by curated ChIP studies. Processes that received a *P*-value of 0.01 or less by NetProphet predictions and were not enriched (*P*-value > 0.05) by ChIP data were considered to be novel relative to ChIP. The novel processes identified for each TF were manually curated to remove redundant and generic processes.

Expression profiling of Δ*cbf1*

The wild-type and Δ*cbf1* strains (see Methods) were assessed for acid phosphatase expression using a QuantiGene assay (Affymetrix). Briefly, 4-mL YPD overnight cultures were inoculated from single colonies. The following day, cells were washed and transferred to 50 ml of media containing low or high levels of inorganic phosphate (YPD or SC +10 mM Pi, respectively). The high phosphate media, SC + 10 mM Pi, was made according to the recipe described in Zhou and O'Shea (2011). Cultures were inoculated at a density selected to achieve OD between 0.3 and 0.6 the following day, based on the doubling times determined for each strain/media pairing. Cells were harvested the following day, and selected RNAs were quantified using QuantiGene (Affymetrix) (see Supplemental Methods).

Raw probe counts were background-normalized according to the negative control probes (Affymetrix), and the background corrected counts of *PHO11+PHO12* (which could not be distinguished) and *PHO5* were normalized by the geometric mean of the background-corrected counts of three housekeeping genes: *TFC1*, *UBC6*, and *PDA1*. The results of technical duplicates were averaged.

Expression profiling of Δ*eds1*

Four-milliliter YPD overnight cultures of wild type (BY4741), *eds1Δ* (YBR033W), and *eds1Δ* complemented with *EDS1* under the control of the $tet0_2$ promoter were inoculated from single colonies. The following day, cells were washed and transferred to 30 ml SC + 2% glucose, grown overnight, synchronized to OD 0.5, and grown for 4 h to approximately OD 0.8. After harvesting by centrifugation, the cells were lysed. Crude lysate was quantified using QuantiGene (Affymetrix) (see Supplemental Methods).

Background correction and normalization of *EDS1*, *LYS4*, *LYS9*, *CTP1*, *ACO2*, and *LYS12* were as described above.

Data access

No data suitable for submission to databases were produced; however, the full QuantiGene data sets from which the data in Figures 4 and 5B were extracted are provided as supplemental spreadsheets.

Acknowledgments

We thank Tamara Doering for collaboration on network inference in *Cryptococcus neoformans*, a project that greatly informed our NetProphet work. We have had many fruitful conversations about network inference with Barak Cohen and members of his lab. Finally, we thank Erin O'Shea and Xu Zhou for generously sharing their strains and discussing them with us. This work was supported in part by National Institutes of Health (NIH) grants GM100452 and AI087794. B.H. was supported by NIH T32 grant HG000045.

Author contributions: B.C.H. and M.R.B. conceived and designed the research, approach, and analysis. B.C.H. carried out most of the research. M.H.K., E.J.M., and P.I.W. assisted in selected analyses. H.B. carried out follow-up experiments involving Cbf1 and Eds1.

References

Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenco AB, dos Santos SC, Cabrito TR, Francisco AP, Madeira SC, Aires RS, et al. 2011. YEASTRACT: Providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface. *Nucleic Acids Res* **39:** D136–D140.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* **14:** 283–291.

Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L. 2006. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J Mol Biol* **360:** 213–227.

Beyer A, Workman C, Hollunder J, Radke D, Moller U, Wilhelm T, Ideker T. 2006. Integrated assessment and prediction of transcription factor binding. *PLoS Comput Biol* **2:** e70.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. 2006. The Inferelator: An algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol* **7:** R36.

Butte AJ, Kohane IS. 2000. Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pac Symp Biocomput* **2000:** 418–429.

Chua G, Morris QD, Sopko R, Robinson MD, Ryan O, Chan ET, Frey BJ, Andrews BJ, Boone C, Hughes TR. 2006. Identifying transcription factor functions and targets by phenotypic activation. *Proc Natl Acad Sci* **103:** 12045–12050.

Cipollina C, van den Brink J, Daran-Lapujade P, Pronk JT, Porro D, de Winde JH. 2008. *Saccharomyces cerevisiae* SFP1: At the crossroads of central metabolism and ribosome biogenesis. *Microbiology* **154:** 1686–1699.

Clifton D, Weinstock SB, Fraenkel DG. 1978. Glycolysis mutants in *Saccharomyces cerevisiae*. *Genetics* **88:** 1–11.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. 2007. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5:** e8.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. *Bioinformatics* **23:** 257–258.

Gari E, Piedrafita L, Aldea M, Herrero E. 1997. A set of vectors with a tetracycline-regulatable promoter system for modulated gene expression in *Saccharomyces cerevisiae*. *Yeast* **13:** 837–848.

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* **11:** 4241–4257.

Gitter A, Siegfried Z, Klutstein M, Fornes O, Oliva B, Simon I, Bar-Joseph Z. 2009. Backup in gene regulatory networks explains differences between binding and knockout results. *Mol Syst Biol* **5:** 276.

Gordan R, Hartemink AJ, Bulyk ML. 2009. Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res* **19:** 2090–2100.

Gordan R, Murphy K, McCord RP, Zhu C, Vedenko A, Bulyk ML. 2011. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biol* **12:** R125.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27:** 1017–1018.

Greenfield A, Madar A, Ostrer H, Bonneau R. 2010. DREAM4: Combining genetic and dynamic information to identify biological networks and dynamical models. *PLoS ONE* **5:** e13397.

Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, et al. 2004. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431:** 99–104.

Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M. 2002. Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev* **16:** 3017–3033.

Hu Z, Killion PJ, Iyer VR. 2007. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet* **39:** 683–687.

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. 2010. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **5:** e12776.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316:** 1497–1502.
Jolma A, Kivioja T, Toivonen J, Cheng L, Wei G, Enge M, Taipale M, Vaquerizas JM, Yan J, Sillanpaa MJ, et al. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **20:** 861–873.
Killion PJ, Sherlock G, Iyer VR. 2003. The Longhorn Array Database (LAD): An open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD). *BMC Bioinformatics* **4:** 32.
Kuras L, Cherest H, Surdin-Kerjan Y, Thomas D. 1996. A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism. *EMBO J* **15:** 2519–2529.
Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22:** 1813–1831.
Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, et al. 2002. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298:** 799–804.
Luscombe NM, Babu MM, Yu H, Snyder M, Teichmann SA, Gerstein M. 2004. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431:** 308–312.
Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. 2010. DREAM3: Network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE* **5:** e9803.
Marbach D, Mattiussi C, Floreano D. 2009. Replaying the evolutionary tape: Biomimetic reverse engineering of gene networks. *Ann NY Acad Sci* **1158:** 234–245.
Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. 2012. Wisdom of crowds for robust gene network inference. *Nat Methods* **9:** 796–804.
Ogawa N, DeRisi J, Brown PO. 2000. New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol Biol Cell* **11:** 4309–4321.
O'Meara TR, Norton D, Price MS, Hay C, Clements MF, Nichols CB, Alspaugh JA. 2010. Interaction of *Cryptococcus neoformans* Rim101 and protein kinase A regulates capsule. *PLoS Pathog* **6:** e1000776.
Pinna A, Soranzo N, de la Fuente A. 2010. From knockouts to networks: Establishing direct cause-effect relationships through graph analysis. *PLoS ONE* **5:** e12912.
Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. 2010. Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PLoS ONE* **5:** e9202.
Reece-Hoyes JS, Diallo A, Lajoie B, Kent A, Shrestha S, Kadreppa S, Pesyna C, Dekker J, Myers CL, Walhout AJ. 2011. Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. *Nat Methods* **8:** 1059–1064.
Reimand J, Vaquerizas JM, Todd AE, Vilo J, Luscombe NM. 2010. Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res* **38:** 4768–4777.
Robasky K, Bulyk ML. 2011. UniPROBE, update 2011: Expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **39:** D124–D128.
Smyth GK. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* **3:** Article 3.
Smyth G. 2005. Limma: Linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor* (ed. Gentleman R, et al.), pp. 397–420. Springer, New York.
Sopko R, Huang D, Preston N, Chua G, Papp B, Kafadar K, Snyder M, Oliver SG, Cyert M, Hughes TR, et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* **21:** 319–330.
Spivak AT, Stormo GD. 2012. ScerTF: A comprehensive database of benchmarked position weight matrices for *Saccharomyces* species. *Nucleic Acids Res* **40:** D162–D168.
Stolovitzky G, Prill RJ, Califano A. 2009. Lessons from the DREAM2 Challenges. *Ann NY Acad Sci* **1158:** 159–195.
Svetlov VV, Cooper TG. 1995. Review: Compilation and characteristics of dedicated transcription factors in *Saccharomyces cerevisiae*. *Yeast* **11:** 1439–1484.
Tedford K, Kim S, Sa D, Stevens K, Tyers M. 1997. Regulation of the mating pheromone and invasive growth responses in yeast by two MAP kinase substrates. *Curr Biol* **7:** 228–238.
Teichmann SA, Babu MM. 2004. Gene regulatory network growth by duplication. *Nat Genet* **36:** 492–496.
van Werven FJ, Timmers HT. 2006. The use of biotin tagging in *Saccharomyces cerevisiae* improves the sensitivity of chromatin immunoprecipitation. *Nucleic Acids Res* **34:** e33.
Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. 2012. "Calling cards" for DNA-binding proteins in mammalian cells. *Genetics* **190:** 941–949.
Westholm J, Nordberg N, Muren E, Ameur A, Komorowski J, Ronne H. 2008. Combinatorial control of gene expression by the three yeast repressors Mig1, Mig2 and Mig3. *BMC Genomics* **9:** 601.
Yip KY, Alexander RP, Yan KK, Gerstein M. 2010. Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data. *PLoS ONE* **5:** e8121.
Zhou X, O'Shea EK. 2011. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* **42:** 826–836.
Zhu C, Byers KJ, McCord RP, Shi Z, Berger MF, Newburger DE, Saulrieta K, Smith Z, Shah MV, Radhakrishnan M, et al. 2009. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **19:** 556–566.

Received October 16, 2012; accepted in revised form April 24, 2013.