Winter 1-15-2021

Mapping Transcription Factor Networks and Elucidating Their Biological Determinants

Yiming Kang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science
Department of Computer Science and Engineering

Dissertation Examination Committee:
Michael Brent, Chair
Jeremy Buhler
Roman Garnett
Scott McIsaac
Robi Mitra

Mapping Transcription Factor Networks and Elucidating Their Biological Determinants

by
Yiming Kang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

January 2021
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Supplemental Figures

List of Supplemental Tables

Acknowledgments

I would like to thank the following people who have helped and supported me during my doctoral studies. First of all, I would like to thank my advisor, Dr. Michael Brent, whose commitment to science and dedication to research motivate me to dig into exciting research topics in systems biology. His encouragement and support allow me to explore new ideas and to overcome roadblocks. His mentorship of conveying complex science ideas to any audience is a life-long gift for me. I would also like to thank other committee members including Dr. Jeremy Buhler, Dr. Roman Garnett, Dr. Scott McIsaac, and Dr. Robi Mitra for their insight and knowledge to help improve my thesis work.

Next, I would like to thank current and former members of Brent Lab including Dhoha Abid, Sandeep Acharya, Dr. Daniel Agustinho, Sanji Bhavsar, Holly Brown, Brian Chen, Ryan Friedman, Jeffery Jung, Eduard Kotysh, Maryl Lambros, Hien-haw Liow, Cynthia Ma, Zeke Maier, Chase Mateusiak, Drew Michael, Nikhil Patel, Jessica Plaggenberg, and Mike Toomey, who share wise research ideas and thoughtful feedbacks that help advance my science journey. Furthermore, I would like to thank my collaborators including Dr. Andrew Chang, Dr. Tamara Doering, Dr. Bryan Leland, Dr. Scott McIsaac, Dr. Rob Mitra, and Dr. Amy Schmid, who generously share their resources that help enrich my work.

Finally, I would like to thank my parents Xiaorong and Shuguang for their continuous love, care, and encouragement all throughout my life from a young artist to a scientist. I would also thank my grandfather Zhenhuang, whose biomedical research sparked my curiosity of

science. Last but not least, I would like to thank my wife Yingying for her unconditional love and unreserved support.

Yiming Kang

Washington University in St. Louis

January 2021

ABSTRACT OF THE DISSERTATION

Mapping Transcription Factor Networks and Elucidating Their Biological Determinants

by

Yiming Kang

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2020

Professor Michael Brent, Chair

A central goal in systems biology is to accurately map the transcription factor (TF) network of a cell. Such a network map is a key component for many downstream applications, from developmental biology to transcriptome engineering, and from disease modeling to drug discovery. Building a reliable network map requires a wide range of data sources including TF binding locations and gene expression data after direct TF perturbations. However, we are facing two roadblocks. First, rich resources are available only for a few well-studied systems and cannot be easily replicated for new organisms or cell types. Second, when TF binding and TF-perturbation response data are available, they rarely converge on a common set of direct and functional targets for a TF. This dissertation explores and validates the best combination of experimental and analytic techniques to map TF networks. First, we introduce an unsupervised inference algorithm that maps TF networks by exploiting only gene expression and genome sequence data. We show that our "data light" method is more accurate at identifying direct targets of TFs than other similar methods. Second, we develop an optimization method to search for a convergent set of target genes that are independently identified by binding locations and perturbation responses of each TF. Combining this method with network inference greatly

expanded the high-confidence network maps, especially when applied on datasets obtained by using recently developed experimental methods. Third, we describe a framework for predicting each gene's responsiveness to a TF perturbation from genomic features. Using this framework, we identified properties of each gene that are independent of the perturbed TF as the major determinants of TF-perturbation responsiveness. This may lead to improvements in network mapping algorithms that exploit TF perturbation responses. Overall, this dissertation provides a scalable framework for mapping high-quality TF networks for a variety of organisms and cell types.

Chapter 1: Introduction

1.1 Motivation

Cells respond to environmental cues or stress by modulating their transcriptional states. Transcription factors (TFs) are the key regulators that directly bind to their target gene's regulatory DNA and thereby control the genes' transcription rates. Such interactions between TFs and targets form a directed graph, called a TF network, in which nodes represent TFs or genes, and edges represent the *direct* and *functional* regulatory interactions. Such network maps have essential roles in many areas of research and development including TF activity inference (Tran et al. 2005; Boorsma et al. 2008; Alvarez et al. 2016; Ma and Brent 2020), transcriptome engineering (Heinäniemi et al. 2013; D'Alessio et al. 2015; Rackham et al. 2016; Cahan et al. 2014; Michael et al. 2016), cancer systems biology (Carro et al. 2010; Aytes et al. 2014; Bhagwat and Vakoc 2015; Da Silveira et al. 2017), and drug discovery (Bansal et al. 2014; Gayvert et al. 2016; Garcia-Alonso et al. 2018).

To map the TF network of a particular organism, we need to leverage experimental data generated systematically to identify the genes that respond to the perturbation of each TF or those whose *cis*-regulatory DNA are bound by each TF. Model organisms such as *Saccharomyces cerevisiae* (yeast) and *Homo sapiens* (human) have been the systems for development of methods that seek to effectively exploit such datasets. Nevertheless, mapping TF networks for less well-studied organisms is difficult due to the lack of comprehensive resources. The data for TF binding is especially scarce, as technologies for its measurement are considerably more challenging than those for gene response quantification. Therefore, it is

essential to develop an unsupervised, “data light” algorithm to accurately reconstruct TF networks using only scalable resources such as gene expression data.

For well-studied model organisms, a wide range of data sources including TF binding locations for many TFs are available, in addition to gene expression profiles. The data for TF binding and those for TF-perturbation responses provide independent, orthogonal information for understanding TF regulation. Using these two types of data generated for the same TF, we would expect to find a large fraction of TF-bound genes to be responsive, and vice versa. However, little convergent evidence was found between these independent sources (Gitter et al. 2009; Lenstra and Holstege 2012; Cusanovich et al. 2014). This motivated us to develop techniques for improving the convergence and reconstructing high-confidence TF networks for model organisms. Furthermore, there still remains a challenge -- if the binding locations of a TF in a gene’s regulatory DNA are insufficient to explain why the gene would respond to the perturbation of this TF, then what are the other factors that constitute the discrepancy? To address this issue, we describe a systematic approach using machine learning to predict whether a gene will respond to the perturbation of a particular TF. Factors presenting a gene’s epigenomic context and its expression properties, in addition to the binding signals of the perturbed TF, are considered to be plausible predictors. Explaining how the models learn to tackle the prediction task expands our understanding of the determinants of transcriptional responses to TF perturbations.

1.2 Background

1.2.1 Experimental methods to generate data for TF network mapping

There are two major experimental approaches methods to generate data for mapping TF networks: (1) the measure of transcriptional responses after TF perturbation, and (2) the measure

of TF binding locations. The perturbation response assay measures how much the expression levels of the targets change upon perturbing the activity of a TF. The perturbations include knockout, knockdown, and induction. If a TF is a direct and functional regulator of a target, then the target is expected to show strong transcriptional response when the TF is perturbed. Most large-scale experiments have measured the steady-state response of genes after perturbing the TF (Hu et al. 2007; Kemmeren et al. 2014; Davis et al. 2018; Schmitges et al. 2016). A recent large-scale, time-series response dataset has been generated using the ZEV system for inducing yeast TFs to over express (Hackett et al. 2019). TF binding assays record the genomic coordinates where a TF binds. The TF binding sites (TFBS) appear frequently in the *cis*-regulatory regions of the targets. Direct evidence of binding can be obtained from experiments such as chromatin immunoprecipitation (ChIP) followed by microarray or sequencing, while indirect evidence can be attained by searching the genes' *cis*-regulatory regions for occurrences of short DNA motif that a TF can potentially bind. More recently developed technologies such as transposon calling cards (Wang et al. 2007, 2011a, 2012), ChIP-exo (Rhee and Pugh 2011; Rossi et al. 2018b, 2018a) and CUT&RUN (Skene and Henikoff 2017; Hainer and Fazzio 2019; Meers et al. 2019a) have improved over ChIP-chip/seq. Data are currently available for a small fraction of TFs. Auxiliary data types such as chromatin accessibility, histone modifications, and chromatin conformation (e.g., Hi-C and ChIA-PET) have also shown usefulness in mapping TF binding signals to each gene's regulatory DNA.

1.2.2 Computational methods to map TF networks

DREAM challenges (Madar et al. 2010; Greenfield et al. 2010) initiated the first large cohorts of inference algorithms for mapping TF networks from gene expression data. The fundamental idea is that regulatory interactions can be inferred from the correlated gene

expression levels of each TF and its target, measured across various conditions such as environmental stimulus, TF perturbation, and cell cycle. Among the best performing methods, Inferelator (Greenfield et al. 2010) and TIGRESS (Haury et al. 2012) use sparse linear regression to model the expression level of each gene as a function of TFs' expression levels. GENIE3 (Huynh-Thu et al. 2010) replaces linear models with random forest. CLR (Faith et al. 2007) and ARACNE (Margolin et al. 2006) estimate mutual information of the TF-target pairs. In the post-DREAM era, MERLIN (Roy et al. 2013) was developed to apply probabilistic graphical models to select the TFs that regulate each gene, where both regulators and targets are constrained to be co-expressed within a module of genes. NetProphet (Haynes et al. 2013) is a method developed in our lab to improve TF network mapping using gene expression data measured after genetic perturbations. It reduced overfitting of linear models and optimized the integration of two analyses -- co-expression and differential expression.

For model organisms such as yeast and human, the binding signals for a number of TFs have been either directly measured or inferred using other experimental data. Thereby, the recent generation of TF network mapping methods have been focused on incorporating the binding information into previously established expression-based inference. MERLIN-P (Siahpirani and Roy 2017) integrates the estimated prior probabilities into the existing graphical models. Inferelator was updated (Greenfield et al. 2013) to select the co-expression model that best utilizes the priors using Bayesian best subset regression. Another extension to Inferelator (Castro et al. 2018) infers TF activities from the priors, and subsequently uses them as predictors of gene expression in linear models.

1.2.3 Predicting gene expression levels

To understand gene regulation, many studies have focused on predicting gene expression levels from various combinations of genomic features. These features include TF data (e.g., TF binding signals and expression levels of TF-encoding genes), epigenetic factors (e.g., histone modifications and chromatin accessibility), and DNA sequence. Given the right combination of TFs for a particular gene, TF binding signals near the gene's transcription start site (TSS) have been shown to be predictive of the gene's expression level (Middendorf et al. 2004; Ouyang et al. 2009; Schmidt et al. 2017). Another class of methods demonstrated the value of using the localized signals of histone modifications in genes' *cis*-regulatory region as predictors (Karlić et al. 2010; Cheng et al. 2011; Dong et al. 2012; McLeay et al. 2012; Singh et al. 2016; Read et al. 2019). More recently, exploiting the DNA sequence flanking a gene by training deep neural networks has gained traction in the field (Kelley et al. 2018; Zhou et al. 2018; Washburn et al. 2019; Agarwal and Shendure 2020). Furthermore, models have been trained for an alternative task -- predicting the variability of gene expression within or across cell types (Ouyang et al. 2009; Zhou et al. 2014; González et al. 2015; Crow et al. 2019; Sigalova et al. 2020). In addition to the above genomic features, combining the binding signals of RNA-binding proteins and microRNA at gene bodies with TFBS at promoters has also shown predictive value (Tasaki et al. 2020).

It is worth noting that these methods aimed to predict gene expression levels in conditions that do not involve TF perturbations; therefore, they cannot be used to explain the functional associations between TFs and genes. It is rather important to directly predict whether a gene will change its transcription level when the activity of a TF is perturbed. This is because it serves as a benchmark of how well we understand the TF network.

1.3 Contributions

- **Development of an unsupervised algorithm for mapping TF networks using only scalable and low-cost resources.** We present NetProphet 2.0, a novel, unsupervised learning method that combines three key principles -- ensemble learning, TF-TF binding similarity, and motif inference. By exploiting only gene expression and genome sequence data, it improves over our lab's original method, NetProphet 1.0, and other expression-based methods. This contribution is presented in Chapter 2.
- **Development of a computational method to optimize the convergence from TF binding locations and TF perturbation responses.** We describe dual threshold optimization for setting significance thresholds on binding and perturbation-response data, which improves their convergence; processing response data through network inference further improves the outcome. This contribution is presented in Chapter 3.
- **Reconstruction of high-confidence TF networks for yeast and human cells.** We present high-confidence TF networks for yeast cells, human K562 and HEK293 cells by applying the best combination of experimental and analytic techniques. This contribution is presented in Chapter 3.
- **Development of a framework to predict TF-perturbation responses.** We describe a machine learning framework to train and test models for predicting each gene's responsiveness upon a TF perturbation by using genomic features including TF binding locations. This contribution is presented in Chapter 4.
- **Identification of TF-independent factors as major determinants of perturbation responses.** We identify gene expression properties and histone modifications measured in unperturbed conditions as the top determinants of responsiveness to TF perturbations.

Currently available data on TF binding locations is predictive of perturbation response in yeast but not in human. This contribution is presented in Chapter 4.

Chapter 2:

Mapping TF networks by exploiting scalable data resources

2.1 Introduction

A transcription factor (TF) network map is a directed graph comprising nodes that represent genes and the proteins they encode and edges that link the TFs to their direct, functional targets. Developing effective methods for mapping TF networks genome-wide is a long-standing goal in genomics (Harbison et al. 2004; Hu et al. 2007) and computational biology (Faith et al. 2007; Margolin et al. 2006); see (Brent 2016) for a recent review. TF network maps encode basic knowledge about the biochemical functions of molecules, much like metabolic network maps. They are thus a key part the encyclopedic knowledge that enables research and development. In addition, a TF network map is an essential input to at least two downstream applications. The first is TF activity inference. A TF network map links TFs with the target genes that they have the potential to bind and regulate, given the right circumstances, external signals, or developmental context. TF activity inference uses such a map to quantitatively model how much influence each TF is exerting on each target in a given context (Boorsma et al. 2008; Boulesteix and Strimmer 2005; Kao et al. 2004; Tran et al. 2005). Unlike ordinary regression of target RNA levels against TF RNA levels, this approach treats TF activity levels as latent variables that are not necessarily proportional to TF RNA levels. A second application is transcriptome engineering, in which the goal is to modify the transcriptional regulatory network of a cell in a way that drives it into an expression state associated with some desirable behavior

(Michael et al. 2016). The most common application of transcriptome engineering to date has been aimed at driving mammalian cells of one type (e.g. stem cells) into the transcriptional state associated with another cell type (e.g. liver cells) (Cahan et al. 2014; D'Alessio et al. 2015; Heinäniemi et al. 2013; Rackham et al. 2016).

Previous approaches to TF network mapping can be loosely categorized into those that rely exclusively on gene expression data ("expression only") and those that integrate a wide range of data types, including chromatin immunoprecipitation sequencing (ChIP-seq) of many TFs, genome-wide chromatin marks, and binding specificities for many TFs determined in vitro ("integrative"). The data required for integrative approaches are available only for major model systems, principally Saccharomyces cerevisiae (yeast), Drosophila melanogaster (fly) (Marbach et al. 2012b), and the mammalian cell lines that have been the focus of the ENCODE project (Brent 2016). Such resources are unlikely to become available soon for most other organisms and cell types. Even for fly and mammalian cell lines only a small fraction of the TFs encoded in the genome have been successfully subjected to ChIP-seq. Furthermore, most of the genes whose regulatory regions are bound by a TF show no evidence of being functionally regulated by that TF (Gitter et al. 2009; Cusanovich et al. 2018).

Gene expression data, by contrast, can be obtained from low cost, reliable, and easily scalable experiments. Expression-only approaches to network inference have had notable successes on bacterial networks (Faith et al. 2007; Ghanbari et al. 2015; Greenfield et al. 2010; Haury et al. 2012; Huynh-Thu et al. 2010; Lam et al. 2016). More recently the NetProphet algorithm, which directly compares expression profiles from TF-knockout strains and wild-type strains, has been shown to give good results on single-cell eukaryotes (Haynes et al. 2013; Brent 2016). There is evidence to suggest that when NetProphet is applied to yeast (Saccharomyces

cerevisiae) it identifies bound genes more accurately than existing yeast ChIP-chip data (Haynes et al. 2013). Unlike ChIP-chip, however, all of the targets NetProphet identifies for a TF are functionally regulated by that TF. However, the accuracy of this approach on animal networks, which are much more complex than that of yeast, has never been demonstrated.

Here, we report on a second-generation "data light" TF-network mapping algorithm called NetProphet 2.0. Our approach requires only data that can be generated from low-cost, reliable, and easily scalable experimental methods. NetProphet 2.0 relies on three fundamental ideas. First, combining several expression-based network algorithms that use different types of models can yield better results than using either one alone – the "wisdom of the crowds" idea (Marbach et al. 2012a). Second, TFs with similar DNA binding domains (in terms of amino acid sequence) tend to bind similar sets of target genes. Third, even an imperfect network map can be used to infer models of each TF's DNA binding preferences from the promoter sequences of its putative targets and these models can be used to further refine the network. We describe the modules of NetProphet 2.0, show that each module contributes to its overall accuracy on both yeast and fly, and show that its overall accuracy improves on that of earlier data light methods, which rely only on gene expression data.

2.2 Results

2.2.1 Overview of analysis steps in NetProphet 2.0

NetProphet 2.0 comprises six computational modules (Fig. 2.1), five of which take advantage of information obtained from gene expression profiling or genome sequencing. The output of each module is a map, represented as a score matrix with rows corresponding to TFs and columns corresponding to all genes, each of which is a potential target. The score vector (row) for a TF represents the strength of evidence that the TF regulates each potential target

gene. A discrete graph structure can always be constructed by including only edges whose scores exceed a chosen threshold.

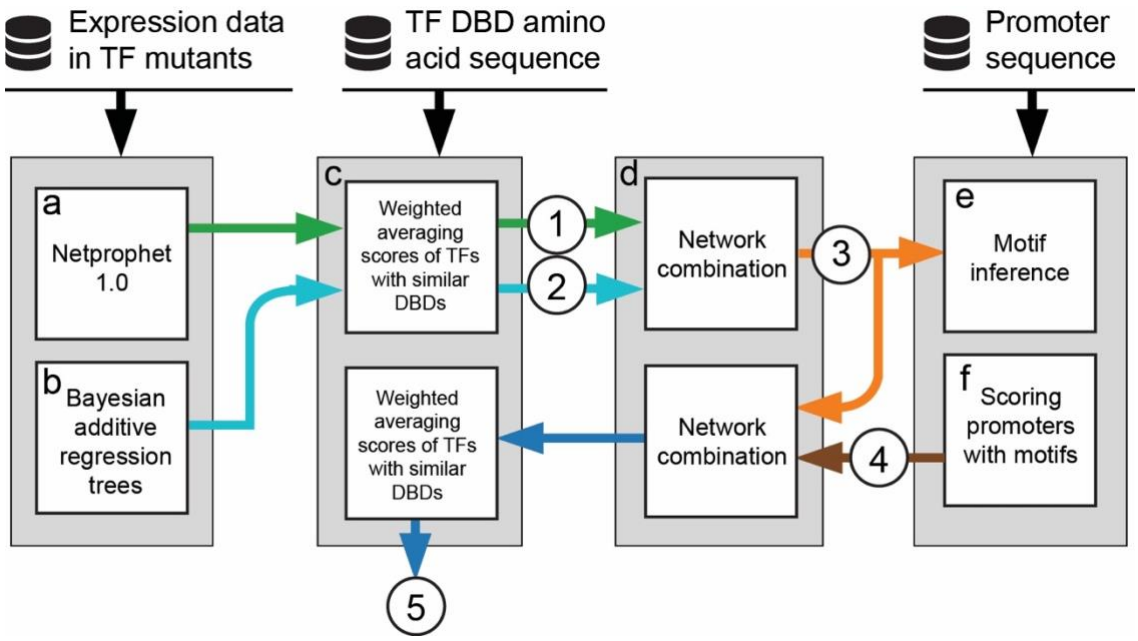


Figure 2. 1: Overview of NetProphet 2.0 pipeline. Database icons: input data sources. Rectangles: computational modules. Circles: network maps.

Module A (Fig. 2.1a) is NetProphet 1.0, as previously described (Haynes et al. 2013). It constructs a map from gene expression profiles and performs best when the data include expression profiles of single TF perturbation strains. Module B (Fig. 2.1b) constructs an independent network map from the same gene expression data by using a machine learning algorithm called Bayesian Additive Regression Trees (BART) (Chipman et al. 2012). For each gene, Module B trains a separate BART model to predict the RNA level of that gene as a function of the RNA levels of all TFs. It then simulates the effect of varying each TF's RNA level on the predicted RNA level of the target, holding the levels of all other TFs constant. Each TF's level is varied between its minimum and maximum observed levels. The difference between the two predicted target gene expression levels is used as the score of the TF-target pair.

Intuitively, the more a gene is predicted to change as a result of changing the level of a TF, the more likely it is to be a direct target of that TF.

Although Module B (BART) and Module A (NetProphet 1.0) use the same gene expression data, they do so in very different ways. NetProphet 1.0 relies primarily on the direct comparison of a gene's expression after genetic perturbation of a TF to its expression in unperturbed, wild type cells. Secondarily, it uses sparse linear regression of each gene's RNA level against the RNA levels of the TFs. BART does not explicitly compare expression of a gene before and after an experimental TF perturbation. Instead, it uses a non-linear, non-parametric regression model based on random forests to predict the effects of a TF perturbation on the expression of a gene.

Module C (Fig. 2.1c) capitalizes on the fact that TFs with similar DNA binding domains (DBDs) tend to bind similar sets of target genes (Weirauch et al. 2014). It replaces the score matrix row for each TF by a weighted average of rows for other TFs with similar DBDs. Each row is weighted according to how similar the DBD of its TF is to the DBD of the row being replaced (see Methods & Supplemental Fig. S2.1). The predicted amino acid sequence of the DBD can be obtained from automated annotation of the genome sequence. The outputs of modules A and B are independently passed through Module C. They are then combined into a single score matrix by Module D (Fig. 2.1d), which uses quantile normalization to make the score distributions of the two networks comparable (see Methods).

Modules E and F (Fig. 2.1e,f) make use of the target genes' promoter sequences to further refine the network map. Module E infers the DNA-binding specificity (motif) of each TF by identifying motifs whose presence in a promoter best distinguishes high scoring (likely) target genes from low scoring (unlikely) target genes (see Methods). Module F scans the inferred motif

for each TF over the promoters of all genes and computes a score reflecting the strength of evidence that the TF binds the promoter. If no significant motif is found for a TF, then its score vector remains unchanged after Module F. The resulting score matrix is then combined with the input score matrix by using module D again. In a final step, the combined matrix is passed through module C again.

In the following sections, we evaluate the contribution of each successive module to the overall accuracy of NetProphet 2.0. Finally, we compare the accuracy of the complete system to that of some previous systems for mapping TF networks from gene expression data.

2.2.2 Input data and benchmarking standards

We collected input data and benchmarking data for both yeast and fly. The gene expression data we used as inputs came from two sources. The first is a recently published yeast data set, which contains 1,487 samples including 265 TF knockout strains and 1,219 knockouts of non-TF-encoding genes (Kemmeren et al. 2014). The second is a fly data set, which contains 200 samples including 23 TF knockdown lines and 84 knockdowns of non-TF-encoding genes (Bonke et al. 2013). To evaluate the accuracy of the inferred network maps, we compared them to both ChIP-based binding data and motif-based binding potential. However, we do not assume that either of these networks is the correct network we are aiming to learn. Indeed, we know that most genes whose promoters are bound by a TF according to ChIP data show no evidence of being functionally regulated by that TF (Gitter et al. 2009; Cusanovich et al. 2014). However, a TF's direct, functional targets are likely to be a subset of the genes whose promoters are bound by that TF. In other words, binding is necessary, but not sufficient, for direct regulation. Because our predicted targets are based on evidence of functional regulation from gene expression data, those predicted targets that are also bound by the TF are likely to be its direct, functional targets.

For each species, we constructed two benchmark networks whose edges connect TFs to the genes whose promoters they bind (but do not necessarily regulate). The first is based on ChIP-chip/seq data, which assesses the physical binding locations of the TFs. For yeast, we compiled ChIP data from TNET (Babu et al. 2004) and YEASTRACT (Abdulrehman et al. 2011), which contains ~30,000 interactions for 184 TFs. For fly, we compiled ChIP data from FlyNet (Marbach et al. 2012b) and seven other ChIP-chip/seq studies, which together contain ~180,000 interactions for 82 TFs. The second benchmark is a motif network constructed by scoring the promoters using position weighted matrix (PWM) models of the DNA binding specificity of each TF. These models are derived from protein binding microarray (PBM) data (collected in UNIPROBE database (Gordân et al. 2011; Robasky and Bulyk 2011)), which are completely independent of both gene expression and ChIP experiments. PWM models are available for 150 yeast TFs and 98 fruit fly TFs.

2.2.3 Exploiting similarity between DNA binding domains improves accuracy

Previously we showed that NetProphet 1.0 (Module A) performed well on yeast by using an older gene expression data set (Hu et al. 2007). Here, our first step is to determine its accuracy on a new yeast data set and on the fruit fly. We evaluated the percentage of the top ranked edges that were supported by the ChIP network (Fig. 2.2A,C) or by the known-PWM network (Fig. 2.2B,D). In all cases, the predicted networks scored much better than randomly generated networks (gray shading), except for the PWM evaluation of the fly network when the number of predicted targets exceeded ~25 per TF encoded in the genome (24,225 total). The Kemmeren yeast data set yielded better results than those previously obtained using the smaller Hu data set (Supplemental Fig. S2.2). Module C (weighted averaging) improved the evaluations against

ChIP data except that it was neutral for large fly networks (Fig. 2.2). It slightly hurt the PWM evaluation of the smaller yeast networks, but it was otherwise neutral.

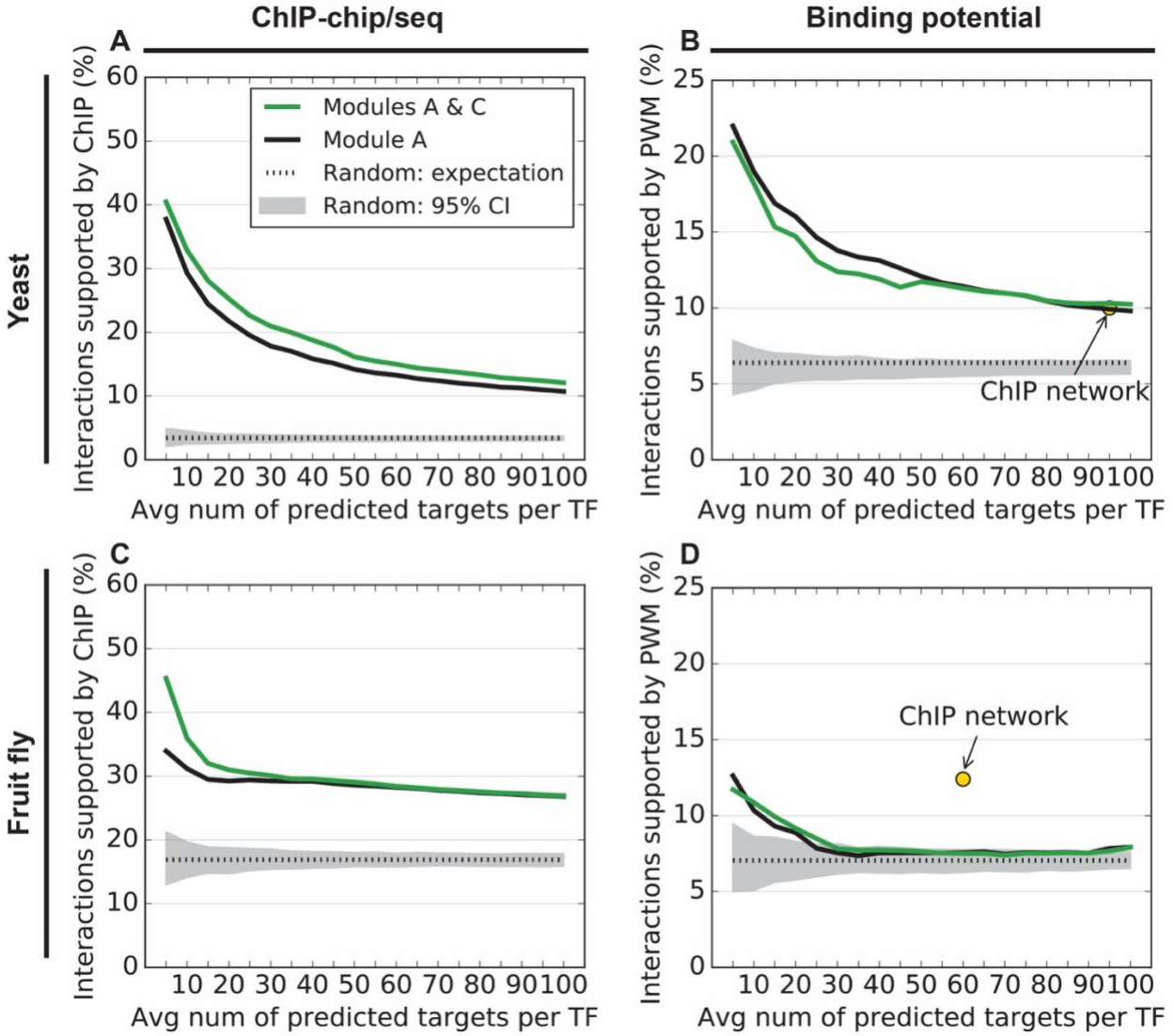


Figure 2. 2: Effect of weighted averaging on the edges of similar TFs. (A) Accuracy of NetProphet 1.0 on yeast before weighted averaging (black line) or after weighted averaging (green line). Horizontal axis: number of top ranked edges included in the network per TF encoded in the genome. E.g., since there are 320 TFs in the yeast genome, "10" on the horizontal axis corresponds to a network with 3,200 edges. Vertical axis: Percentage of edges supported by ChIP data. Dotted line: Expected accuracy of random networks. Gray area: 95% confidence interval for randomly selected networks. (B) Same as A for PWM support. The point labeled "ChIP network" indicates the number of ChIP-supported edges and the fraction of those edges that also have PWM support. (C) Same as A for the fly data. (D) Same as B for the fly data, except that the vertical axis shows support by conserved PWM hits only.

2.2.4 NetProphet 1.0 works on the fly network

Comparing the results for yeast and fly, it is apparent that the fly networks received slightly more support than the yeast network from ChIP data but less support from PWM data. In fact, the PWM support for fly networks with more than 25 edges per TF encoded in the genome does not significantly exceed the support for random networks. That is probably because the number of fly expression profiles in which a single TF has been knocked down represents 10-fold fewer TFs than for yeast (23 vs. 265) and the number of expression profiles from non-TF knockdowns is also much smaller (84 vs. 1,219). The number of known fly PWMs against which to evaluate is also smaller (98 vs. 150). Another difference is that the yeast ChIP network was supported by PWM evidence at the same rate as the similar sized networks predicted by NetProphet 1.0. The fly ChIP network, by contrast, was supported at a much higher rate than similar sized networks predicted by NetProphet 1.0. That may be the result of the smaller expression data set for fly and because the fly ChIP data are more recent than the yeast data, so the ChIP methodology may have matured in the interim.

2.2.5 Combining with Bayesian Additive Regression Trees improves accuracy

Module B uses Bayesian Additive Regression Trees (BART), which provides an alternative approach to making use of the gene expression data. As weighted averaging (Module C) improved the accuracy of the NetProphet 1.0 output, we applied it to the BART output (Fig 2.1, network 2), which it also improved (Supplemental Fig. S2.3). Finally, we tried combining the two resulting networks (Fig. 2.1, network 3). The effects of processing through these modules on accuracy are shown in Figure 2.3. NetProphet 1.0 with weighted averaging (Modules A & C, green) generally performed better than BART with weighted averaging (Modules B & C, blue), except that BART significantly outperformed in PWM support on yeast (Fig. 2.3B).

Remarkably, combining the two networks (Modules A-D) performed as well as the better of the two on the yeast ChIP and PWM metrics (Fig. 2.3A, B) and significantly better than either network on fly (Fig. 2.3C, D). This is consistent with the previously reported "Wisdom of Crowds" effect in TF network mapping (Marbach et al. 2012a).

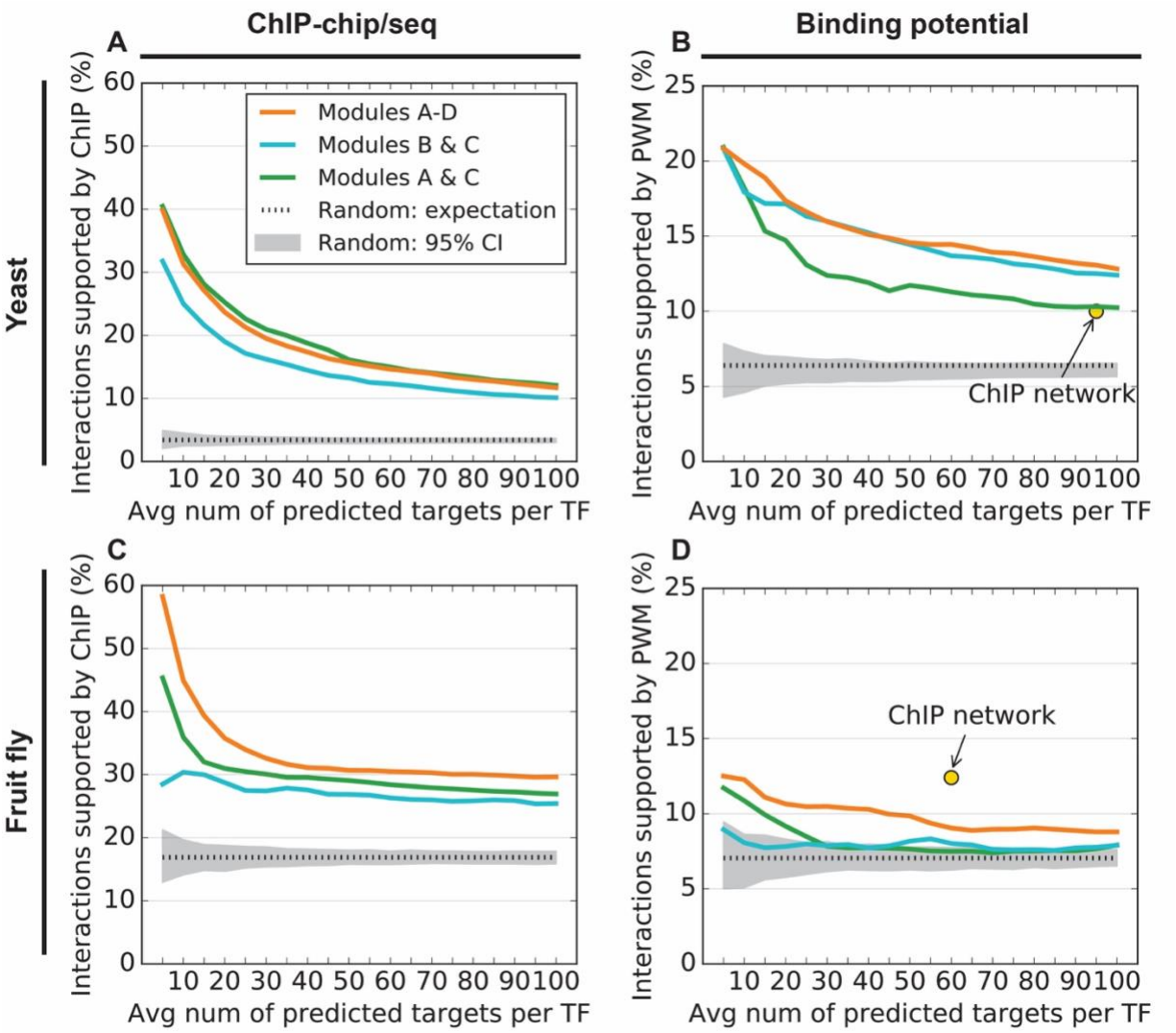


Figure 2. 3: Effect of combining intermediate networks. (A) Accuracy of NetProphet 1.0 on yeast after weighted averaging (Modules A & C, green line), BART after weighted averaging (Modules B & C, blue line), and the combination of the two (Modules A-D, orange line). Horizontal axis: number of top ranked edges included in the network per TF encoded in the genome. Vertical axis: Percentage of edges of included edges that are sup-ported by ChIP data. Dotted line: Expected accuracy of randomly networks. Gray area: 95% confidence interval for random networks. (B) Same as A for PWM support. The point labeled "ChIP network" indicates the number of ChIP-

supported edges and the fraction of those edges that also have PWM support. (C) Same as A for the fly data. (D) Same as B for the fly data, except that the vertical axis shows support by conserved PWM hits only.

2.2.6 Inferring TF binding preferences from promoter sequence improves accuracy

We hypothesized that knowing the DNA binding specificities of the TFs would enable us to improve on the accuracy of the maps output by Modules A-D. To test that hypothesis, we scanned the known yeast and fly PWMs across the promoter sequences of all genes in the genome, producing a binding potential score for each TF at each promoter (see Methods). This score matrix was then combined with the score matrix output by Modules A-D (Fig. 2.1, Network 3) by using Module D again. The resulting maps were evaluated as before (Fig. 2.4, purple dashed lines). For the evaluation by PWM support, using known PWMs constitutes "peeking" at the evaluation standard, so it would have been worrisome if performance had not improved. For the evaluation by ChIP support, using known PWMs provided a small but consistent accuracy improvement, except for mid-sized fly networks, where it had no effect. The fact that this helped the yeast results more than the fly results is not surprising, since the promoter regions in yeast are much smaller and a much higher fraction of yeast TFs have known PWMs (46.9% vs. 10.1%).

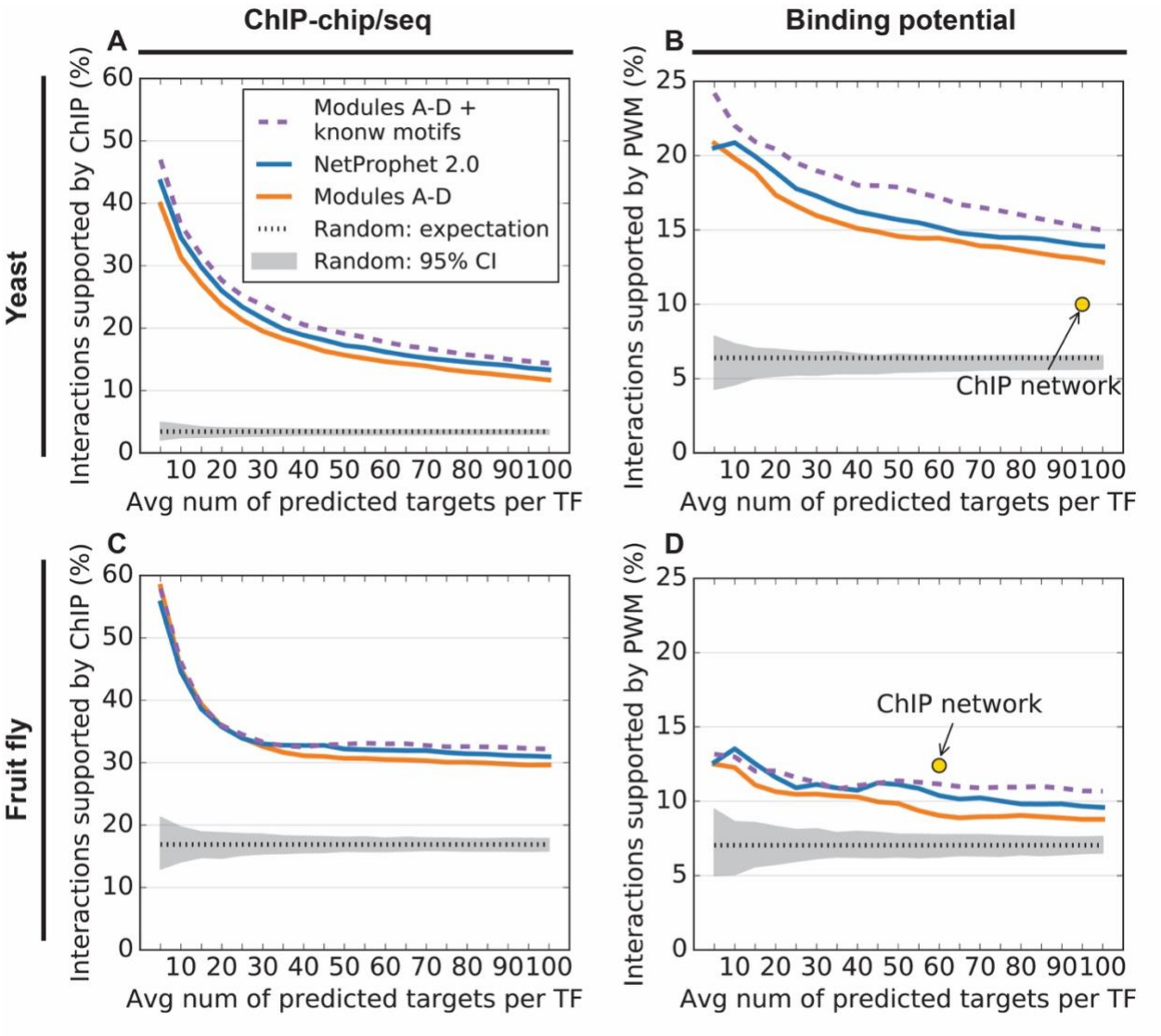


Figure 2. 4: Effect of adding a motif network. (A) Accuracy of Modules A-D (Combination of NetProphet 1.0 and BART after weighted averaging (orange line), Modules A-D with known yeast PWM motifs (dashed purple line), and Modules A-F (NetProphet 2.0). Horizontal axis: number of top ranked edges included in the network per TF encoded in the genome. Vertical axis: Percentage of edges of included edges that are supported by ChIP data. Dotted line: Expected accuracy of randomly networks. Gray area: 95% confidence interval for random networks. (B) Same as A for PWM support. The point labeled "ChIP network" indicates the number of ChIP-supported edges and the fraction of those edges that also have PWM support. (C) Same as A for the fly data. (D) Same as B for the fly data, except that the vertical axis shows support by conserved PWM hits only.

The known PWMs used above were obtained from protein binding microarray experiments. However, we hypothesized that we could infer PWMs using only gene expression

and genome sequence data, thereby avoiding the need for additional experiments. Thus, we applied the FIRE motif inference algorithm (Elemento et al. 2007) to the score vector of each TF after Modules A-D. FIRE attempts to find a motif whose presence in a promoter best discriminates between high and low scoring target genes. We then used the inferred motifs to score the promoter of each gene just as we had with the known PWMs. These scores were combined with the output of Modules A-D, except that the scores for TFs for which FIRE could not identify a high confidence motif were left unchanged. The resulting accuracy improvement (Fig. 2.4, blue line) was approximately half of that obtained from the known motifs. Importantly, this approach does not require any additional experiments, making it suitable for application to non-model systems.

Next, we directly compared the motifs inferred by Module E to the known motifs. For each TF with a known PWM, we calculated the Spearman correlation between the scores assigned to each promoter by the inferred and known PWMs (Fig. 2.5, blue bars). As a randomized baseline distribution, we calculated the median of the correlations between each inferred PWM and all other known PWMs (Fig. 2.5, orange bars). For yeast, 37.9% of the inferred PWMs correlated with the corresponding known PWMs at levels significantly above the baseline distribution. In the fly data, the baseline distribution showed much higher correlations than in the yeast data. This is probably because 42% of all known fly PWMs belong to the Homeodomain family, whose members share a preference for binding motifs containing ATTA (Hughes 2011). Additionally, there were only 2 fly TFs, whose inferred PWM scores had a correlation of > 0.5 with their known PWM score (as compared to 25 in yeast). This may reflect the larger size of the fly promoters, the smaller amount of expression data available for fly, and/or a greater tendency for gene regulation in the fly to be determined by combinatorial logic,

rather than by independently active binding sites. Although few of the inferred fly PWMs showed a statistically significant degree of similarity with their known counterparts, the use of the PWM inference module results in a small but noticeable increase in overall accuracy.

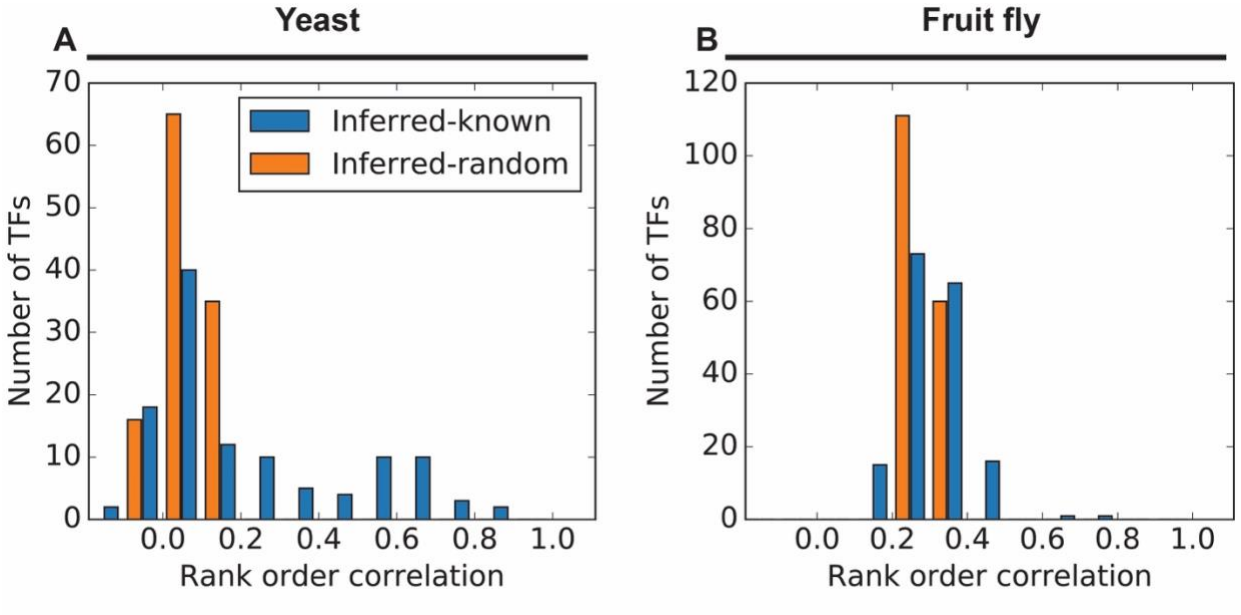


Figure 2. 5: Relationships between inferred and known PWMs. Blue bars: distribution of rank order correlations between binding potential scores assigned to each promoter by inferred PWMs and known PWMs. Orange bars: distribution of the medians of Spearman correlations between each inferred PWM and the known PWMs for all other TFs. (A) Yeast. (B) Fruit fly.

2.2.7 NetProphet 2.0 improves on previous network mapping methods

For several years, algorithms for mapping TF networks from gene expression data were compared in a series of community evaluation projects known as DREAM (Dialog on Reverse-Engineering Assessment and Methods; (Marbach et al. 2012a)). In a previous publication, we compared NetProphet 1.0 to several of the best performing algorithms from DREAM on a set of yeast expression profiles (Haynes et al. 2013). The comparison algorithms were Inferelator (Greenfield et al. 2010) and GENIE3 (Huynh-Thu et al. 2010). Here, we compare NetProphet 2.0 to those same algorithms plus two others: CLR (Faith et al. 2007) and TIGRESS (Haury et al. 2012), on a new set of yeast expression profiles and a set of fly expression profiles. We also

compare to using the squared Spearman correlation coefficient between the expression of each TF and each target gene as the TF-target score, the method used in the FlyNet paper (Marbach et al. 2012b).

To evaluate NetProphet and the four other algorithms, we ran them all on the same sets of expression profiles used throughout this study and selected the top scoring interactions from the output of each algorithm. The number of top scoring interactions selected was ten per TF encoded in the genome – i.e. 3,200 for yeast and 9,690 for fly. It is important to note that NetProphet 2.0 requires an annotated genome sequence as input, whereas the other algorithms use only the gene expression data. Therefore, we are not evaluating algorithms designed for exactly the same tasks. However, they can all be viewed as special cases of algorithms designed to infer direct, functional TF networks from data that can be produced by low-cost, reliable, scalable methods.

The results of the comparison showed that NetProphet 2.0 was more accurate than the other algorithms as evaluated by the yeast ChIP benchmark and by the fly ChIP and PWM benchmarks (Fig. 2.6). GENIE3 was slightly more accurate than NetProphet 2.0 on the yeast PWM benchmark. When comparing predictions to known interactions that are supported by both ChIP and PWM data, NetProphet 2.0 was substantially more accurate than all of the comparison algorithms. This is significant because ChIP hits that coincide with PWM hits are more likely to be functional than those that do not (Cusanovich et al. 2014; Van Nostrand and Kim 2013).

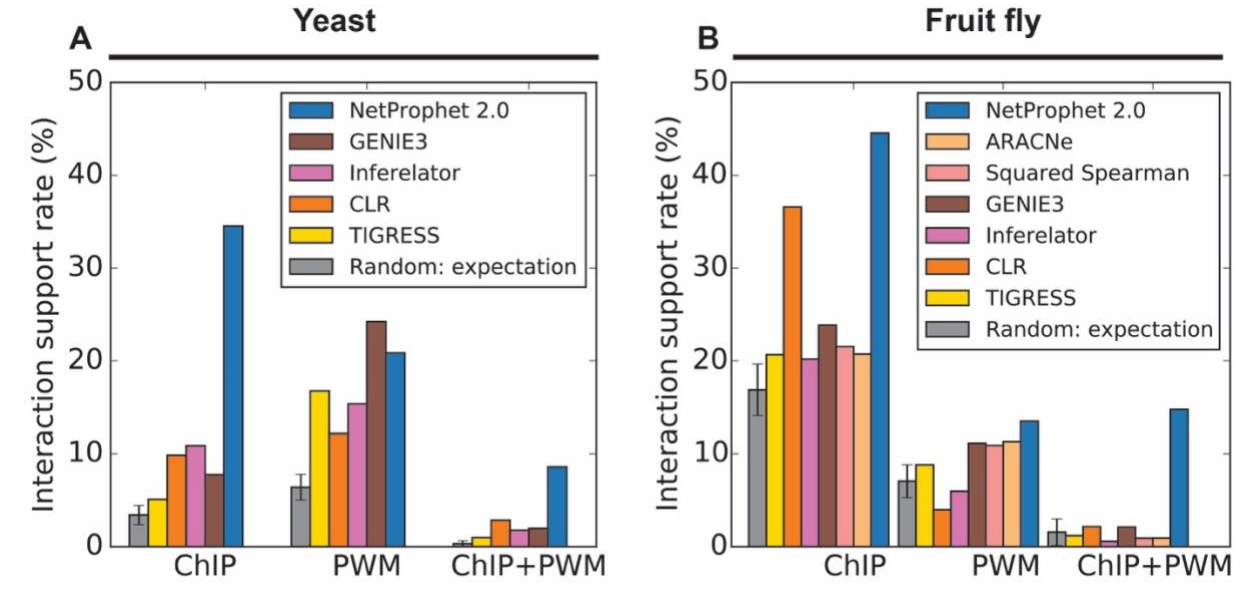


Figure 2. 6: Comparison between NetProphet 2.0 and other leading expression-based mapping algorithms. (A) Yeast. (B) Fruit fly.

2.3 Discussion

NetProphet 2.0 is designed around the principle of using only data that can be obtained with robust, predictable, and scalable experimental methods. Specifically, it requires only gene expression data after TF perturbation and genome sequence with automated annotation. It makes use of three fundamental ideas. First, combining the results of distinct approaches to mapping networks from gene expression data can significantly improve accuracy (Marbach et al. 2012a). Second, similar DNA binding domains bind similar sets of promoters (Weirauch et al. 2014). Third, even a noisy, imperfect network can be used to infer useful binding motifs from promoter sequences. By combining these three ideas, NetProphet 2.0 significantly outperforms NetProphet 1.0 and a range of other expression-based algorithms, as assessed by measured binding locations and by binding potentials. The fraction of predicted interactions that are supported by both ChIP and PWM substantially exceeds that of the other algorithms tested (Fig. 2.6).

There are many possible ways to implement the ideas behind NetProphet 2.0. For example, there are other non-parametric regression algorithms that could substitute for or supplement BART. Fused regression (Lam et al. 2016) is a possible alternative to our weighted averaging approach for exploiting the similarities between DNA binding domains. There are also many software packages for inferring TF binding motifs, which could be substituted for FIRE. Implementations using these alternative components, which are beyond the scope of this study, have the potential to improve accuracy in the future.

NetProphet's "data light" approach stands in contrast to the "integrative" approach, which has also been applied to mapping the fly TF network (Marbach et al. 2012b). In that study, a network was constructed by using all available data sources, including the same TF-ChIP and PWM data sets that we used only for validation. Because these two data sources were used as inputs, they could not also be used for validation of the integrative network. As a result, it is not possible to directly compare the accuracy of the two approaches on genome-scale networks. The integrative model also used ChIP of a wide range of chromatin marks as input. Thus, applying it to a new organism or cell type would require a data generation effort far beyond what can currently be done in a single lab. Integrative network construction is feasible for a few model systems that have been targeted for exhaustive data generation by large consortia. When the integrative approach is feasible, NetProphet 2.0 can be used to process the available gene expression data in place of methods such as the Spearman correlation of expression profiles (Marbach et al. 2012a). In addition, NetProphet 2.0 can integrate binding specificity models determined by methods such as yeast one hybrid (Fuxman Bass et al. 2016), high throughput selex (Jolma et al. 2013), and protein binding microarrays (Weirauch et al. 2014), for any TFs for which they are available. An interesting intermediate between data-light and integrative

approaches would be to combine NetProphet 2.0 with TF binding locations that are predicted from TF binding specificity, conservation, and cell-type specific DNA accessibility data, but not requiring ChiP-seq of individual TFs (Cuellar-Partida et al. 2012; Zhong et al. 2013). There has also been recent progress in formal frameworks for integration of prior knowledge into expression-based network mapping (Ghanbari et al. 2015; Lam et al. 2016).

TF-target interactions predicted by NetProphet 2.0 are supported by binding potential (PWMs derived from protein-binding microarray experiments) at a significantly higher rate than the interactions predicted by existing yeast ChIP-chip data (Fig. 2.4B). The ChIP-seq data on the fly genome are much more recent than the yeast data (Clough et al. 2014; Georlette et al. 2007; Hadzić et al. 2015; Ikmi et al. 2014; Liu et al. 2009; Marbach et al. 2012b; Page et al. 2005; Teleman et al. 2008). When networks of similar size (number of targets per TF) are compared, the fly ChIP edges are supported by PWMs at a slightly higher rate than the NetProphet 2.0 predictions (Fig. 2.4D). However, the NetProphet 2.0 edges that score among the top 14,535 (~15 targets per TF) are supported by strong binding potential at a rate comparable to those of the larger ChIP network. For practical purposes, it is also important to keep in mind that the ChIP data come at a much higher cost than the NetProphet 2.0 predictions, take much longer to generate, and are plagued by the uncertain success of individual ChIP-seq experiments. Furthermore, existing evidence suggests that only a very small fraction of ChIP-supported interactions are functional, in the sense that the expression of the gene whose promoter is bound changes when the TF is perturbed (typically < 10%; (Gitter et al. 2009; Cusanovich et al. 2014); reviewed in (Brent 2016)). Since NetProphet 2.0 is primarily an expression-based method, all its predictions are supported by expression data and hence are likely to be functional. Thus, NetProphet 2.0 provides an attractive alternative to TF ChIP, especially for experimental systems

that are unlikely to benefit from an ENCODE-style undertaking to systematically ChIP a large number of TFs.

NetProphet 2.0 is the first algorithm that has been shown to be effective on an animal genome without requiring any data beyond gene expression after TF perturbations and genome sequence. While the steps from bacteria to yeast and yeast to fly were significant (Haynes et al. 2013; Marbach et al. 2012b), the step from a compact invertebrate genome such as that of the fly to mammalian genomes will also be challenging. The primary challenges include limited data availability, large, poorly defined promoters, and long-range enhancers. The data limitation will probably be removed over the next few years, now that CAS9 has made deleting TFs in mammalian systems much easier. The problem of defining enhancers and identifying their target genes may also be alleviated before long. One source of data that will likely prove useful is the expression of enhancer RNAs, which can highlight active enhancers and the genes whose expression correlates with enhancer activity (Andersson et al. 2014; Core et al. 2008; Danko et al. 2015). Data on three-dimensional chromosome conformation from rapidly improving, sequencing based methods will also prove useful. We expect that these new data sources will make it possible to test, validate, and apply NetProphet 2.0 to mammalian systems in the near future.

2.4 Methods

2.4.1 Download and preparation of data sets

Yeast data

The microarray data with gene deletions used for mapping TF network in yeast was published in (Kemmeren et al. 2014) (accession GSE42217, downloaded from http://deleteome.holstegelab.nl/). This set of expression profiles contains 265 strains of single TF

knockouts and 1,219 strains of other gene knockouts. To assess the mapping accuracy, we constructed two benchmarks derived from ChIP-chip and protein binding microarray (PBM) experiments. The ChIP benchmark was compiled from TNET (Babu et al. 2004) and YEASTRACT (Abdulrehman et al. 2011). It contains 29,945 interactions among 184 TFs and 5,790 genes. The PWM benchmark was constructed using the PBM-derived PWMs of 150 TFs collected in UNIPROBE database (Gordân et al. 2011; Robasky and Bulyk 2011), as described in (Haynes et al. 2013).

Fruit fly data

The microarray data after TF knockdowns used for mapping the fly TF network was published in (Bonke et al. 2013) (accession E-MTAB-453). It consists of samples of 23 single TF knockdowns and 84 other gene knockdowns. We processed the raw CEL files using RMA normalization (affy package, R/Bioconductor (Gautier et al. 2004)). To build benchmarks for fly, ChIP data was combined from multiple ChIP-chip/seq studies. We combined the curated data in (Marbach et al. 2012b) with data described in several other publications (Clough et al. 2014; Georlette et al. 2007; Hadzić et al. 2015; Ikmi et al. 2014; Liu et al. 2009; Page et al. 2005; Teleman et al. 2008). The resulting network map contains 184,053 interactions between 82 TFs and 14,165 genes. The known PWM benchmark was as described in (Marbach et al. 2012b). Each PWM was scanned across the conserved regions within the promoter of each gene (1k bp regions centered on TSS) and the highest score within the promoter was used. Conservation was based on the analysis of the genomes of 12 Drosophila species. The resulting network map (binary score matrix) contains 71,090 interactions between 98 TFs and 10,299 genes. This relatively small number of edges per TF helps explain why the percentage of predicted edges that

are supported by the fly PWM network is much lower than the percentage supported by the fly ChIP network.

Promoter sequence

We collected the promoter sequences of yeast and fruit fly from Regulatory Sequence Analysis Tools (RSAT) database (Medina-Rivera et al. 2015) (http://pedagogix-tagc.univ-mrs.fr/rsat/). The promoter of each yeast gene is the region 600 bp upstream of the transcription start site (TSS). The promoter of each fly gene is the region between 2,000 bp upstream from the TSS and 200 bp downstream from the TSS. Any regions of the promoters that overlap with the coding sequences of a neighboring gene were excluded.

DNA binding domain sequence

The amino acid sequences of the yeast TFs was obtained from Saccharomyces Genome Database (Cherry et al. 2012). The amino acid sequences of the fly TFs were obtained from FlyBase (dmel v6.04, (Dos Santos et al. 2015)). The NCBI Web CD-Search Tool (Marchler-Bauer et al. 2015) (https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi, with default settings) was then used to search for the DNA binding domains (DBDs) of the TFs. Subsequently BEDTools (v2.25.0) (Quinlan and Hall 2010) was used to parse the search results and output DBD sequences in fasta format.

Data access

The resource data files, output networks, inferred motifs and benchmark networks are available for download at http://mblab.wustl.edu/software.html under NetProphet 2.0 software package.

2.4.2 Evaluation

ChIP support

We used the ChIP benchmarks to assess the mapping accuracy of our algorithm. These network maps are binary matrices in which ones represent positive ChIP interactions. Based on a certain stringency level, the top L interactions predicted by NetProphet 2.0 modules were evaluated against ChIP interactions. The mapping accuracy, termed as ChIP support rate, is the fraction of these predicted top interactions supported by ChIP evidence. The network size is based on all predicted edges above a given stringency, while the ChIP support rate is based on the edges whose TFs have ChIP data. We evaluated the accuracies of the mapped networks of different sizes as we varied the stringency levels.

PWM support

The PWM score matrices for yeast were binarized using a threshold for each TF. The threshold was the greatest binding potential score that was exceeded by at least 10% of the ChIP-supported interactions of that TF. We calculated the PWM-support rates using this binary matrix, just as we did for the ChIP binary matrix.

2.4.3 Weighted Averaging

Calculation of weighting function & threshold

We used a four-step process to characterize the relationship between the similarities of DBDs and the similarities of known PWMs. First, for each yeast TF, we obtained the sequences of any DBDs found within it as well as the PWM associated with it from the CIS-BP data base (Weirauch et al. 2014). We then aligned the DBDs of each TF to the DBDs of each other TF by using Clustal Omega (v1.2.1) (Sievers et al. 2011) and used the percent identity (PID) to quantify the similarity between the two of DBDs. If there were multiple DBDs within a TF all pairs of DBDs were aligned and the largest percent identity was used. Second, we aligned the PWM of each TF to that of each other by using Tomtom (v4.9.1) (Gupta et al. 2007) and used the E value

output from Tomtom as a measure of the similarity between the two PWMs. Third, for the TF pairs whose DBD similarity scores fall in a certain PID range, we calculated the fraction of the corresponding PWM pairs that are similar (Tomtom E value < 1). Finally, we fit a logistic function to model the relationship between the percent identities of DBDs and the fraction of significantly similar PWMs:

$$w(d) = \frac{0.9}{1 + exp\big(-0.1(d - 40)\big)} \tag{2.1}$$

where d is the percent identity of a pair of DBDs (Supplemental Fig. S2.1). The fraction of similar PWMs can also be seen as the probability of a pair of TFs at a given DBD-similarity level binding to similar DNA sequences.

Use of weighting function

To implement Module C, we calculated the PID between each pair of DBDs to predict the probability that the DBDs bind significantly similar sequences, according to the logistic model. For each TF *i*, this probability was used as a weighting factor for each other TF with PID >= 50%; for TFs with PID < 50%, the weighting was 0. Row *i* was then replaced by the weighted sum of all rows:

$$S'_i = \sum_k w\big(d_{k,i}\big) S_k \tag{2.2}$$

where S'_i is the updated row of edge scores of TF *i* to all genes, $d_{k,i}$ is the percent identity score between DBD's of TF *k* and TF *i*, and $w(\cdot)$ is the weighting factor calculated using the logistic function.

2.4.4 Bayesian Additive Regression Trees

We used the BART model trained for each target gene to predict the effects of varying each TF's level on the level of the target gene. Specifically, we varied the RNA level of each TF

between its minimum and maximum observed levels while keeping the levels of other TFs constant (fold change 1). The edge score of TF *i* to target *j* in the BART network map is the difference between the predicted level of target *j* in the two simulations, one with TF *i* at its maximum observed level and the other with TF *i* at its minimum observed value. BART package implemented in R was used (v0.3-1.3, https://cran.r-project.org/package=BayesTree; (Chipman et al. 2012)).

2.4.5 Quantile combination of network maps

Since the network maps output by various modules have different score distributions, we used quantile normalization (Module D) to combine score matrices. One matrix is designated as the reference and the other as the auxiliary. The scores in the auxiliary matrix are modified to have the same distribution as the reference matrix before averaging with the corresponding entries of the reference matrix. Formally, if $S_{i,j}^{ref}$ is the score for TF *i* as a regulator of gene *j* in the reference matrix and $S_{i,j}^{aux}$ is the score for TF *i* as a regulator of gene *j* in the auxiliary matrix:

$$\mathrm{S_{i,j}} = \frac{1}{2}\left(\mathrm{S_{i,j}^{ref}} + \mathrm{F_{ref}^{-1}}\left(\mathrm{F_{aux}}\left(\mathrm{S_{i,j}^{aux}}\right)\right)\right) \tag{2.3}$$

where F_{ref} and F_{aux} are the empirical cumulative distribution functions of the reference and auxiliary matrices, respectively. For combining the NetProphet-derived (Fig. 2.1, network 1) and BART-derived (Fig. 2.1, network 2) matrices, the former is designated as the reference. This approach was chosen over other quantile normalization methods empirically, because it gave better results.

2.4.6 PWM inference and promoter scoring

PWM inference

Module E uses an algorithm called FIRE (Elemento et al. 2007) to infer a motif for each TF based on its score vector and the promoter sequences of all genes. For each TF, Module E divides the range of target scores into 20 bins, each spanning 1/20th of the range. For each gene, the bin number corresponding to its score as a target of the TF is input to FIRE, along with the sequence of its promoter region. We used 7 as the k-mer seed size, 20/20 as the robustness threshold, and default parameters for other criteria. If more than one motif passed the criteria for a TF, we only considered the best one, according to FIRE.

Promoter scoring

Semantically, the motifs output by FIRE are patterns specifying which nucleotides are possible at each position of a binding site. However, these can be converted to PWMs by assigning each of the possible nucleotides at each position the same probability and assigning each impossible nucleotide probability zero. For example, if the motif specified is {A,T}{G}{G,C,T}, A or T in the first position would have probability 1/2, G in the second position would have probability 1, and G, C or T in the third position would have probability 1/3. With this interpretation, Module F uses the FIMO program (Grant et al. 2011) to score the binding potentials by scanning the inferred motifs over the promoters. The TF-promoter binding potential was calculated as the maximum of two scores:(1) the log odds of the most significant binding site, (2) the sum of log odds of all significant ($p < 0.05$) binding sites. Subsequently, we used Module D again to combine this binding potential matrix (the auxiliary matrix; Fig. 2.1, Network 4) with the input to Module E (the reference matrix; Fig. 2.1, Network 3). The rows of TFs for which we could not infer a motif were left unchanged from the input (Fig. 2.1, Network 3).

2.4.7 Other algorithms to which NetProphet 2.0 is compared

TIGRESS

Trustful Inference of Gene REgulation using Stability Selection (TIGRESS) uses stability selection to sample the expression data and scores the TF-target interaction as the frequency of each TF being chosen in LARS for each target gene (Haury et al. 2012). We used its MATLAB implementation (v2.1) downloaded from http://cbio.mines-paristech.fr/~ahaury/svn/dream5/html/index.html. We modified the code so that the TFs could be indexed at any position in the comprehensive gene list.

CLR

Context likelihood of relatedness (CLR) estimates the likelihood of the mutual information (MI) by contrasting the MI calculated using the RNA levels of each TF-target pair across all samples with the null model, given the local network context (Faith et al. 2007). We used minet (v3.30.0, R/Bioconductor package) downloaded from https://www.bioconductor.org/packages/release/bioc/html/minet.html to build MI matrix and infer CLR network.

Inferelator pipeline

The Inferelator pipeline in DREAM4 (Greenfield et al. 2010) is a mixture model that consists of median corrected Z-score, mutual information (CLR) and LASSO regression coefficient (Inferelator 1.0). The source code was downloaded from https://github.com/smidget/Network-Inference-Workspace/tree/master/algorithms/inferelator-pipeline. We wrote a script to pipeline Inferelator modules according to the provided pseudo-code.

GENIE3

GEne Network Inference with Ensemble of trees (GENIE3) uses random forests that estimate how much the expression level of each TF contributes to explaining the level of each target gene (Huynh-Thu et al. 2010). We used the Python implementation downloaded from http://www.montefiore.ulg.ac.be/~huynh-thu/software.html.

Chapter 3:

Expanding high-confidence maps using convergent evidence from TF binding locations and TF perturbation responses

3.1 Introduction

Mapping the circuitry by which cells regulate gene expression is a fundamental goal of systems biology. Such maps would facilitate a broad spectrum of research programs, much as maps of intermediary metabolism and genome sequences have. Transcriptional regulation has multiple layers and component types, including sensors and signal transduction cascades. The bottom layer of transcriptional regulation, which acts directly at the genome, features sequence-specific DNA binding proteins known as transcription factors (TFs). Signaling cascades often change the activity levels of specific TFs -- the extent to which they exert their regulatory potential on their target genes -- via mechanisms that affect TFs' abundance, localization, non-covalent interactions, or covalent modifications. To map and model transcriptional regulation as a whole, we must know which genes each TF regulates, or has the potential to regulate when activated.

A map of an organism's TF network would have powerful applications. It could be used to infer the effects of specific signals, drugs, or environments on the activity levels of TFs by analyzing their effects on gene expression (Balwierz et al. 2014; Boorsma et al. 2008; Liao et al. 2003; Tran et al. 2005). It could be used to predict the significance of naturally occurring genome variants in TFs or TF binding sites (TFBS). It could also be used to design genome edits

in TFs or TFBS to achieve a desired transcriptional state or behavior (Cahan et al. 2014; Michael et al. 2016; Rackham et al. 2016). Crucial to all of these applications is the distinction between the direct functional targets of a TF -- the genes it regulates because it binds to their cis-regulatory DNA -- and its indirect targets, which are regulated via intermediary proteins. For example, a mutation inactivating a binding site for a TF in the cis-regulatory DNA of one of its direct targets will affect the relationship between the TF and its direct target. However, a mutation in a non-functional binding site which happens to lie in the cis-regulatory DNA of an indirect target will not affect the relationship between the TF and its indirect target.

In this study, we analyze previously published and newly described genome-wide data sets (Table 3.1) with both standard and novel analytic techniques, to reveal the current state of the art in identifying the direct, functional targets of a TF. The data sets we focus on are those that aim to determine the binding locations of TFs and those that attempt to measure the transcriptional response to perturbations of TF activity, such as over expressing the TF or deleting the gene that encodes it. The binding location data are derived from either chromatin immunoprecipitation (ChIP) or transposon calling cards (Mayhew and Mitra 2016; Ryan et al. 2012; Shively et al. 2019; Wang et al. 2008).

Table 3. 1: Data resources

Data type	Technology	Species	Proteins targeted	Targeted TFs analyzed	Genome assembly	Strain/ Cell line	Publications
Binding location	ChIP-chip	*S. cerevisiae*	203	155	N/A	W303	Harbison, 2004
	ChIP-chip	*S. cerevisiae*	200	36	N/A	S288C	Venters, 2011
	ChIP-exo	*S. cerevisiae*	26	26	R55, R64 (SGD)	S288C	Rhee, 2011; Rossi, 2018a; Rossi, 2018b; Bergenholm, 2018; Holland, 2019
	Transposon calling cards	*S. cerevisiae*	15	15	R61 (SGD)	S288C	Wang, 2012; Shively, 2019; Kang, 2020
	ChIP-seq	*H. sapiens*	261	261	GRCh38	K562	Davis, 2018 (ENCODE)
	ChIP-seq	*H. sapiens*	131	131	GRCh37	HEK293	Schmitges, 2016
	ChIP-exo	*H. sapiens*	236	236	GRCh37	HEK293T	Imbeault, 2017
Perturbation response	TFKO	*S. cerevisiae*	1,484	164	N/A	S288C	Kemmeren, 2014
	ZEV TF induction	*S. cerevisiae*	201	139	N/A	S288C	Hackett, 2019
	TFKD (shRNA, siRNA)	*H. sapiens*	261	261	GRCh38	K562	Davis, 2018 (ENCODE)
	CRISPR + CRISPRi	*H. sapiens*	96	96	GRCh38	K562	Davis, 2018 (ENCODE)
	TF induction	*H. sapiens*	80	80	GRCh37	HEK293	Schmitges, 2016

Yeast data sets on TF binding locations and TF perturbation-responses are more complete than those of any other eukaryote and yeast has a simpler genome with more localized regulatory DNA. For those reasons, we start by focusing on yeast. In addition to evaluating data sets and

experimental and analytic methods, we construct a preliminary map of the yeast TF network by integrating the best available binding and perturbation response data sets. For model invertebrates, there are large data sets on TF binding location (Brown and Celniker 2015; Kudron et al. 2018), but there are currently no comparable data sets on the responses to TF perturbations. Such data is available, however, for human cell lines. We analyze large data sets on human K562 cells (Sloan et al. 2016; Dunham et al. 2012) and HEK293 cells (Schmitges et al. 2016), producing TF networks for each cell type.

3.2 Results

3.2.1 Simple comparison of yeast ChIP-chip to expression profiles of TF deletion strains yields few high-confidence regulatory relationships

Comprehensive binding and perturbation response data sets are available for yeast TFs

In 2004, Harbison et al. assayed the binding locations of all yeast TFs by using ChIP-chip (Harbison et al. 2004). In 2007, Hu et al. published gene expression data on yeast strains in which each non-essential yeast TFs was deleted (Hu et al. 2007). This made it possible to estimate the fraction of binding events that are functional, and Hu et al. remarked on how small that fraction is -- about 3-5% in their data. In 2014, Kemmeren et al. published a second such data set, which benefited from newer technology and the hindsight afforded by the earlier study (Kemmeren et al. 2014). In this section, we focus on the Kemmeren TF knockout (TFKO) data because it demonstrates better agreement with the Harbison ChIP data, on average.

Most bound genes in the Harbison ChIP data are not responsive in the TFKO data

We began by calculating the response rate of bound genes for each TF -- the fraction of bound genes that are differentially expressed in the TFKO strain, relative to the wild type (WT).

The microarrays used by Harbison et al. in their ChIP-chip study contained one probe for each promoter, so their analysis yielded a simple P-value for whether each promoter is bound. We eliminated from further consideration the 16 TFs that were not called as bound to any promoter. For the TFKO data, we used the authors' statistical analysis and considered a gene differentially if its p-value (adjusted for multiple comparisons) was < 0.05. We eliminated from further consideration any TF whose knockout resulted in no significant changes as well as the 32 TFs whose reported expression level in the strain lacking the TF was more than one half its reported level in the WT. This can happen when the wild-type expression level of the TF is near or below the detection limit of the microarray.

Figure 3.1A shows a histogram of the results. The median response rate for bound genes was 18%. The mode was 0% -- 25 of the 97 TFs (26%) had both bound targets and responsive targets, but none of the bound targets were responsive. Only 17 TFs (18%) had a response rate above 50%. Tightening the statistical significance threshold for responsiveness lowers the response rate further, while tightening the threshold for binding causes very few genes to be classified as bound and responsive (Supplemental Fig. S3.1A-C). Thus, these data do not support the notion that most binding is functional. The low response rate of bound genes cannot be explained by saying that the TFs are inactive in the conditions tested, since the median number of genes that respond with $p<0.05$ is 318. A lot of genes respond, but they are not the bound genes.

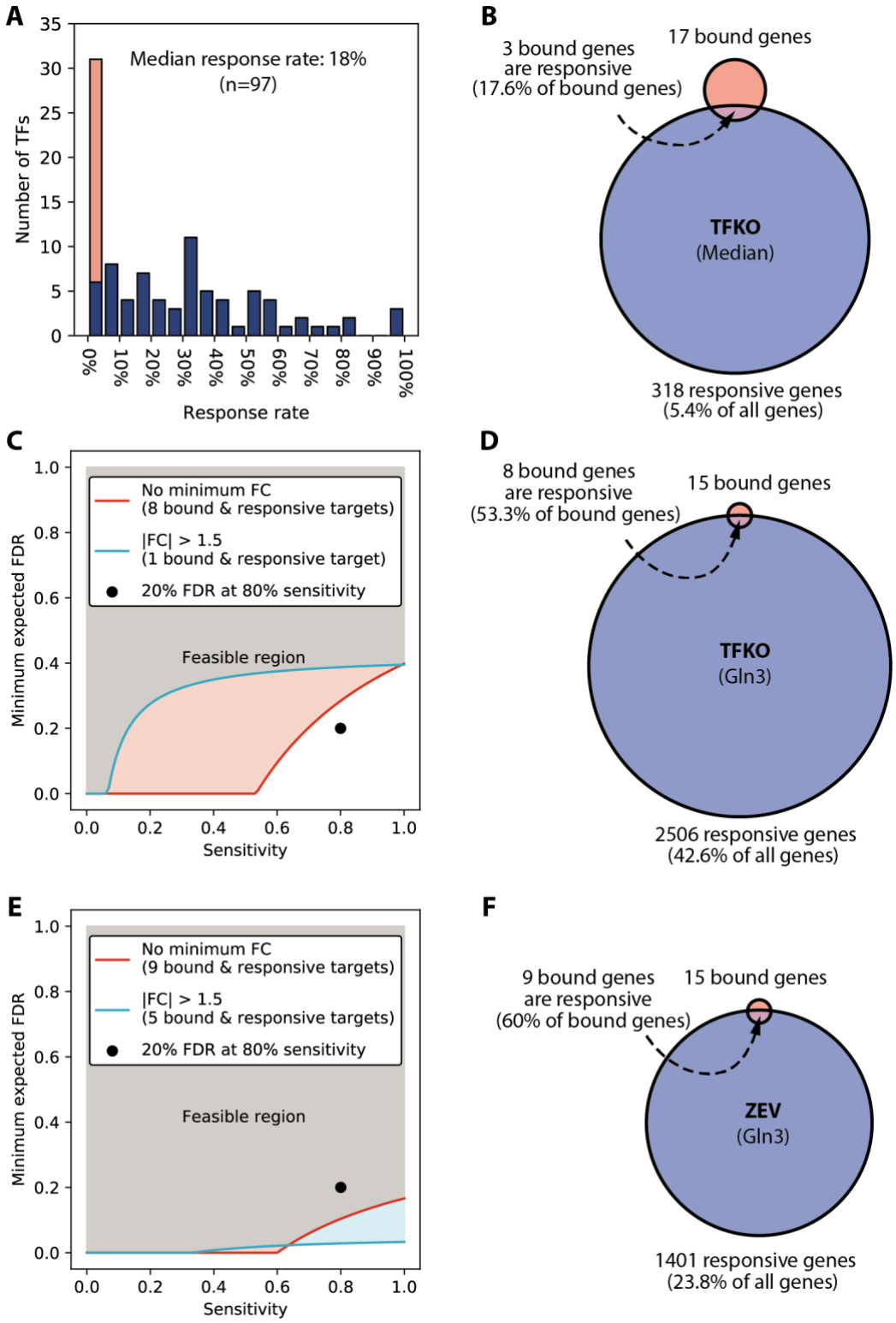


Figure 3. 1: Overlap between the bound and responsive gene sets. (A) Distribution of the response rates of TFs (fraction of bound genes that respond to TF perturbation) in the Harbison binding and Kemmeren TFKO data sets.

Stacked orange bar indicates the number of TFs with response rates of exactly 0. Binding threshold is $p<0.001$ and response threshold is $p<0.05$, as recommended in the original publications, with no minimum fold change. (B) Median numbers of bound genes (17), perturbation-responsive genes (318), and intersection size (3), when comparing the ChIP-chip data to the TFKO perturbation-response data. Thresholds are as in panel A. (C) Minimum expected FDR as a function of sensitivity for TF Gln3, when comparing ChIP to TFKO. Genes are counted as responsive if they have adjusted $P<0.05$ (blue line) or adjusted $P<0.05$ and fold-change > 1.5 (salmon line). 80% sensitivity with 20% FDR is not attainable at either threshold, when comparing ChIP to TFKO. (D) The bound set, responsive set, and intersection for Gln3, when comparing ChIP to TFKO. (E) Minimum expected FDR, as a function of sensitivity, with moderate and tight thresholds for responsiveness, when comparing ChIP to ZEV15. 80% sensitivity with 20% FDR is attainable at either threshold. (F) The bound set, responsive set, and intersection for Gln3, when comparing ChIP to ZEV15.

Many genes that are both bound and responsive in previously published data are probably not direct functional targets

Given that available data suggest most binding sites are non-functional, a logical procedure for finding the direct functional (DF) targets is to take the intersection of the genes bound by each TF with the genes that respond to perturbation of that TF, a procedure we refer to as the intersection algorithm. It is important to keep in mind, however, that most responsive genes are not bound. Comparing the ChIP data with the TFKO data, the median fraction of responsive genes that are bound is 1% (Fig. 3.1B). Thus, most of the responsive genes are indirect targets. Furthermore, it is reasonable to assume that the distribution of indirect targets among all genes is independent of the distribution of non-functional binding sites, or at least that non-functional binding sites do not systematically avoid the promoters of indirect targets. This suggests that some of the indirect targets also have non-functional binding sites. These genes would be false positives of the intersection algorithm -- genes that are bound and responsive, but are not responsive because they are bound.

In Methods (3.5.2 Expected false discovery rate of intersection algorithms), we derive a new lower bound on the expected false discovery rate (FDR) of the intersection algorithm, as a

function of its sensitivity (the fraction of direct functional targets that are in the intersection) and four other variables: number of bound genes, $|B|$, the number of responsive genes, $|R|$, the number of bound and responsive genes, $|R \cap B|$, and the total number of genes assayed, $|G|$.

$$\mathrm{E}[FDR] \geq \frac{\max(0,\ |B| - |R \cap B|/\mathrm{Sn}) \max(0,\ |R| - |R \cap B|/\mathrm{Sn})}{|G||R \cap B|} \qquad (3.1)$$

The formula shows that, if a large fraction of bound genes is not responsive and a large fraction of responsive genes is not bound, the intersection procedure cannot have both high sensitivity and low false-discovery rate. For example, Figure 3.1C shows the relationship between sensitivity and expected FDR for a fairly typical TF, Gln3, based on the Harbison ChIP data and the TFKO response data. The blue and red lines form the boundaries between the feasible and infeasible regions for two different response thresholds. They are calculated by varying the sensitivity and using the formula shown above to calculate the corresponding lower bound on the expected FDR. A reasonable minimum accuracy criterion for a procedure aimed at finding the DF targets of a TF is that it has sensitivity >= 80% (it detects at least 80% of the DF targets) and an FDR <= 20%. However, that is not possible for Gln3, using these two data sets (Fig. 3.1C, black dot). Intuitively, this is because the fraction of Gln3-bound genes that are responsive to the Gln3 perturbation (53%) is only a little more than the fraction of all genes that are responsive to the Gln3 perturbation (43%; Fig. 3.1D). The 80-20 criterion is achievable for only 43 TFs. Supplemental Figure S3.2 shows the cumulative fraction of TFs that have an FDR bound below a given level, assuming 80% sensitivity, at various significance thresholds for binding and response.

The FDR lower bound does not guarantee any maximum FDR for the intersection algorithm. In fact, of the 43 TFs that could possibly achieve the 80-20 criterion in the ChIP-TFKO comparison, only 27 have an intersection that is significantly larger than would be

expected by chance (hypergeometric P<0.01, not adjusted for multiple testing). Conversely, three TFs that passed the P<0.01 criterion failed the 80-20 criterion. If we define "TF with acceptable convergence" to be one that could pass the 80-20 criterion and has a larger overlap between bound and responsive targets than would be expected for randomly selected gene sets, then there are 27 acceptable TFs with 448 interactions regulating 366 target genes. If we take this to be our network map, ~85% of TFs do not have acceptable convergence, so they have no high confidence targets, while 94% of genes have no identifiable regulator. In summary, using the simple intersection algorithm with just these two data sets does not produce anything like a complete TF network map.

3.2.2 Comparing yeast ChIP-chip data to expression profiles measured shortly after TF induction enlarges the network map

Recently, some of us released a data set in which the expression of nearly every yeast TF was induced from a very low level to a high level (http://idea.research.calicolabs.com; (Hackett et al. 2020)). This was accomplished by expressing ZEV, an estradiol-activated artificial TF, and replacing the promoter of the gene to be induced with a ZEV-responsive promoter (McIsaac et al. 2014, 2013). (Some of the TFs were induced using an earlier iteration of the artificial TF called GEV (McIsaac et al. 2011), but we refer to the data set as ZEV for convenience.) Gene expression profiles were measured before induction and at 5, 10, 15, 20, 30, 45, and 90 minutes after inducing the expression of a natural yeast TF with estradiol. We reasoned that genes that respond rapidly might be enriched for direct targets of the induced TF, since there would be limited time for intermediary proteins to be transcribed and translated. If the responders were enriched for direct targets, the number of TFs showing acceptable convergence might increase, expanding the network map. In general, the expression profiles taken 15 minutes after TF

induction (ZEV15) were most enriched for bound genes, so we focus on the 15-minute time point for the remainder of the analyses (Supplemental Fig. S3.3). For a detailed description of the strains, experiments, and analysis, see (Hackett et al. 2020).

The TF Gln3, which could not achieve 80% sensitivity with 20% expected FDR in the ChIP-TFKO comparison (Fig. 3.1C), can in the ChIP-ZEV15 comparison (Fig. 3.1E). The reason is that the number of responsive genes has decreased from 43% of all genes to 24%, at the same time that the response rate of bound genes increased from 53% to 60% (Fig. 3.1D,F). Across all TFs, the ChIP-ZEV15 comparison identified 37 acceptable TFs, 23 of which had not been identified in the ChIP-TFKO comparison (Fig. 3.2A). The ChIP-ZEV15 comparison significantly expanded the network map. Still, >72% of TFs do not show acceptable convergence in either data set and hence have no identifiable targets, while >87% of genes have no identifiable regulators.

3.2.3 Dual threshold optimization expands the TF network map

A possible limitation of the previous analyses is its sensitivity to the statistical significance thresholds used to determine which genes are bound and which are responsive. The statistics are calculated separately for the binding and response data sets and statistical significance thresholds are, by their nature, arbitrary. Furthermore, statistically significant levels of binding or perturbation response might not be biologically significant. For example, a TF may bind a site consistently in the ChIP data even though the fractional occupancy of the site is too low to detectably affect transcription.

To address these problems, we developed dual threshold optimization (DTO), a method that sets the binding and response thresholds by considering both data sets together. DTO chooses, for each TF, the pair of (binding, response) thresholds that minimizes the probability

that the overlap between the bound and responsive sets results from random gene selection (Fig. 3.2B). For this analysis, we ranked all genes by their absolute log fold change in the ZEV15 data and, separately, by their negative log p-value in the ChIP-chip data. We then chose the pair of (binding, response) rank thresholds that minimized the nominal hypergeometric P-value of the overlap between bound and responsive gene sets. The only constraint on the thresholds chosen was that the p-value for the ChIP data could not exceed 0.1. To test the significance of the overlap at the chosen thresholds, we randomly permuted the assignment of binding and response signals to genes 1000 times and ran DTO on each random permutation (see Methods for details).

After DTO, we applied the same acceptable convergence criteria as before -- the bound-responsive overlap must be significant ($P<0.01$, permutation-based) and 20% FDR at 80% sensitivity must be theoretically achievable. DTO expanded the network map again (Fig. 3.2C). Combining the results from TFKO and ZEV15, 60 TFs showed acceptable convergence. For these 60, the bound-responsive overlap contained 2,074 regulatory interactions involving 1,430 unique target genes. The number of TFs that are acceptable in both response da ta sets, 29, now exceeds the number that are acceptable in either of the data sets alone (TFKO:14, ZEV15:17). In this map, ~33% of TFs have at least one target and ~24% of genes have at least one regulator.

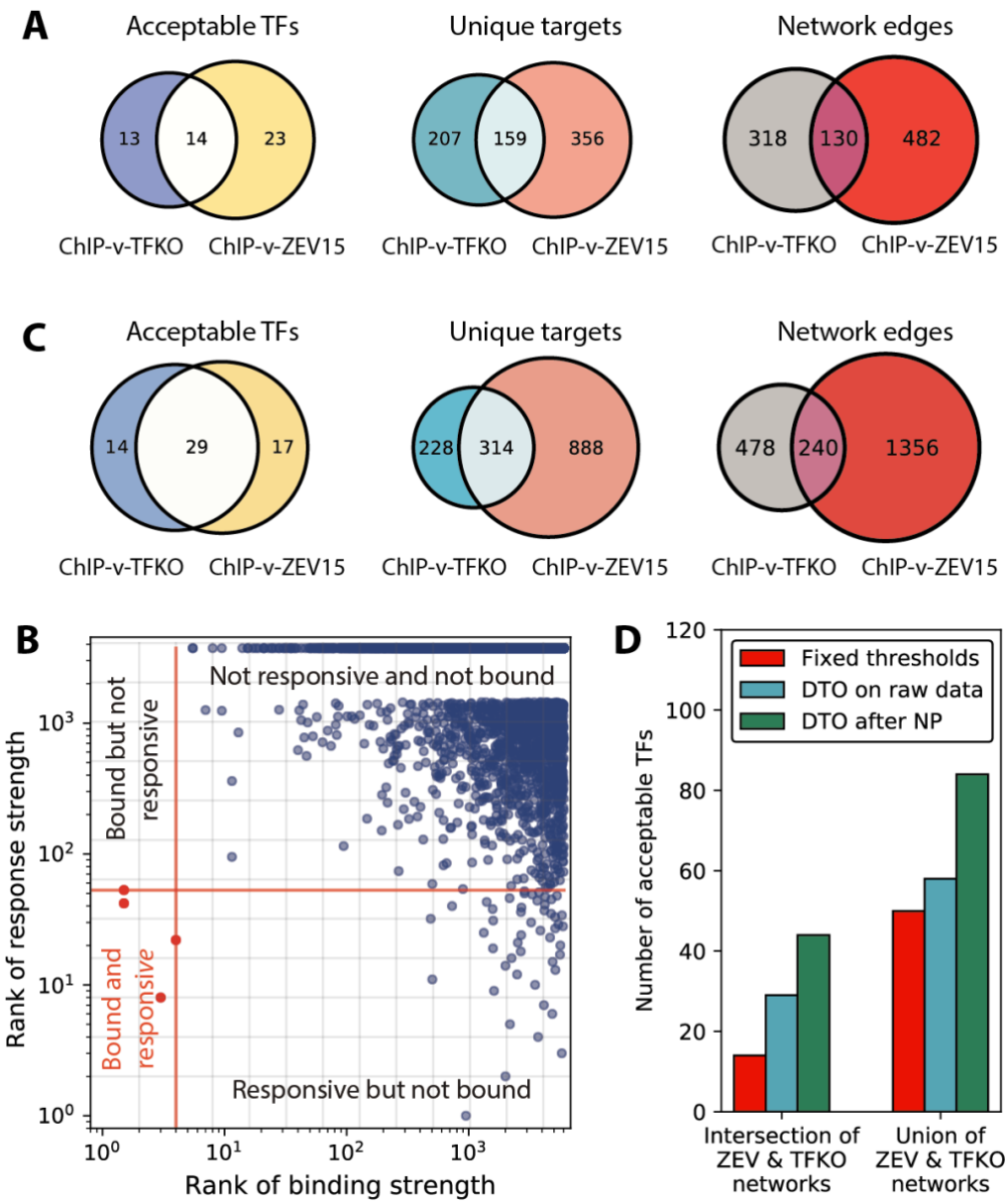


Figure 3. 2: Dual threshold optimization and network inference in yeast. (A) Numbers of acceptable TFs, unique target genes, and network edges, when comparing Harbison ChIP data to TFKO or ZEV15 response data. "Unique Targets" are genes that are in the bound-responsive intersection of at least one acceptable TF and thus are plausible direct functional targets. Edges connect acceptable TFs to the genes in their bound-responsive intersection. The ZEV15 response data yields more acceptable TFs, more unique targets, and more regulatory edges. (B) Illustration of DTO algorithm. Each dot represents one gene. Red lines indicate the chosen (optimal) thresholds for binding (vertical red line) and regulation (horizontal red line). The lower left quadrant, relative to the red lines, contains the bound and responsive genes, which are presumed to be direct functional targets (red dots). Gray lines indicate some of the other possible thresholds on binding or response and locations where the gray lines cross are possible combinations of binding and response thresholds, each of which is evaluated by the DTO algorithm. (C)

Numbers of acceptable TFs and unique target genes for comparison of Harbison ChIP binding data to TFKO or ZEV15 response data, after dual threshold optimization (DTO). The requirement that the overlap between the bound and responsive targets be greater than chance at $p<0.01$ was checked by comparing the nominal hypergeometric p-value for the overlap to a null distribution obtained by running dual threshold optimization on 1,000 randomly permuted binding and response data sets. DTO increases the network size, relative to using fixed significance thresholds. ZEV15 still yields more acceptable TFs, regulated genes, and regulatory interactions than TFKO. (D) Comparison of TFKO and ZEV15 networks derived from fixed thresholds, DTO on raw gene expression, and DTO on gene expression data processed by NetProphet 2.0. The use of DTO on the raw expression data (blue bars) increases the size of both the intersection of the ZEV15 and TFKO (left bar grouping) and their union (right bar grouping). Post processing with NetProphet 2.0 (green bars) further increases the number of acceptable TFs.

3.2.4 Processing yeast gene expression data with a network inference algorithm further expands the network map

There are many algorithms that attempt to infer TF-target relationships by processing gene expression data but not binding location data (e.g., (Faith et al. 2007; Greenfield et al. 2013; Haury et al. 2012; Haynes et al. 2013; Huynh-Thu et al. 2010; Kang et al. 2018; Margolin et al. 2006; Roy et al. 2013)). Typically, they assign a confidence score to each possible TF-target interaction. If all possible targets of a TF are ranked according to their score, DTO can be applied to compare this ranking to binding location data. As long as the network inference algorithm does not use any binding data, DTO can provide independent, convergent evidence. There are also network inference algorithms that weigh and integrate data sources including gene expression and TF binding location data or curated sources influenced by binding data (e.g. (Miraldi et al. 2019; Siahpirani and Roy 2017; Wang et al. 2018)). These algorithms are not suitable for our current purpose, which is to assess the convergence of independent evidence from gene expression and binding location data.

To test this idea, we focused on our lab's network inference algorithm, NetProphet 2.0 (Kang et al. 2018). A major component of the NetProphet score is the degree to which the target

gene responds to direct perturbation of the TF. However, it also considers the degree to which the mRNA level of the TF is predictive of the mRNA level of the potential target, across many different perturbations. NetProphet also makes use of two other ideas: (1) that co-regulated genes tend to have similar sequence motifs in their promoters, and (2) that DNA binding domains with similar amino acid sequences tend to bind similar motifs. It does not use any data on TF binding location, either directly or indirectly.

We built separate NetProphet networks using the TFKO and ZEV data (see Methods). For TFKO, we input 3 wild-type expression profiles and the complete set of 1,484 expression profiles from strains lacking one gene -- some of the deleted genes encode TFs, but others encode other putative regulatory proteins, such as kinases and phosphatases. For ZEV, we used 590 expression profiles from 15 minutes, 45 minutes, or 90 minutes post-induction. We then ranked the potential targets of each TF by their NetProphet scores and ran dual threshold optimization, treating the NetProphet score as we did the perturbation response strength. Combining the results from NetProphet applied to TFKO and ZEV data, dual threshold optimization yielded 84 TFs (46%) with acceptable convergence (Fig. 3.2D). For these TFs, the bound-responsive intersection had 2,153 regulatory interactions involving 1,327 unique target genes (23%, Supplemental Fig. S3.4A,B). The number of TFs that are acceptable in both perturbation data sets, 44, is now much larger than the number that are acceptable in either data set alone (TFKO:22, ZEV:18). Results from comparing binding data to output from three other network inference algorithms, Inferelator (Greenfield et al. 2013), GENIE3 (Huynh-Thu et al. 2010), and MERLIN (Roy et al. 2013), can be found in Supplemental Figure S3.4C.

Running NetProphet on gene expression data and feeding the result into dual threshold optimization has enlarged the map, but it is still smaller than what is generally expected for the

complete yeast TF network. To improve it further, we need binding data that is more accurate or more specifically focused on functional binding.

3.2.5 Without network inference, data on human cell lines yields a few acceptable TFs

The ENCODE Project (Dunham et al. 2012) has produced a wealth of data on human cell lines, including 743 TF ChIP-seq experiments and 391 RNA-seq experiments following knockdown of a TF by siRNA or shRNA (TFKD), or by CRISPR interference (Gilbert et al. 2014) or CRISPR knockout (CRISPRi+CRISPR KO). In K562 cells, 42 TFs have both ChIP-seq and TFKD data while 45 TFs have both ChIP-seq and CRISPRi or CRISPR KO data. We focus on this K562 data, as it is by far the biggest relevant data set.

We considered two ways of assigning ChIP-seq peaks to the genes they potentially regulate. The first is the traditional approach of choosing a fixed interval around the transcription start site (TSS) -- we used 10 kb upstream to 2 kb downstream. The second is to take a small proximal promoter region (TSS -500 bp to +500 bp) along with enhancer regions that have been identified and assigned to the target gene in the GeneHancer database (Fishilevich et al. 2017). GeneHancer uses a variety of data types including predicted and ChIP-based TF binding sites, enhancer RNAs, histone marks, chromosome conformation, and cis-eQTLs. We used only the 'elite' enhancers and 'elite' associations, each of which are supported by at least two sources of evidence. 91% of the 'elite' enhancers were supported by evidence from ENCODE, much of which comes from K562 cells. The enhancer-based approach generally yielded 1 or 2 more TFs with acceptable convergence than the fixed interval approach, so we used the enhancers in subsequent analyses.

Unlike the yeast array data, the human sequencing data yielded many more bound than responsive genes (Fig. 3.3A,B). Among the TFs that had at least one bound and one responsive gene, 7 (TFKD) and 7 (CRISPRi+CRISPR KO) had no genes that were both bound and responsive. The median response rate for bound genes was < 0.5%. In a fixed-threshold intersection with K562 ChIP-seq data, TFKD and CRISPRi+CRISPR KO each yielded 5 TFs with acceptable convergence. We then ran dual threshold optimization limiting the bound and responsive gene sets to have p<=0.1; such limits are necessary because DTO occasionally chooses implausible thresholds, such as counting all genes as responsive. Among all TFs with both binding and response data, TFKD yielded 14% acceptable TFs (6/43) and CRISPRi+CRISPR KO yielded 13% (6/45), a slight improvement over fixed-threshold intersections (Fig. 3.3C, left and center).

We also analyzed a data set on 88 human GFP-tagged C2H2 Zinc finger TFs with matched ChIP-seq data and response-to-overexpression data in HEK293 cells (Schmitges et al. 2016). Using DTO on the ChIP-seq and differential expression data and limiting the total number of responsive genes to 300,000, three of 88 TFs showed acceptable convergence (Fig. 3.3C, right).

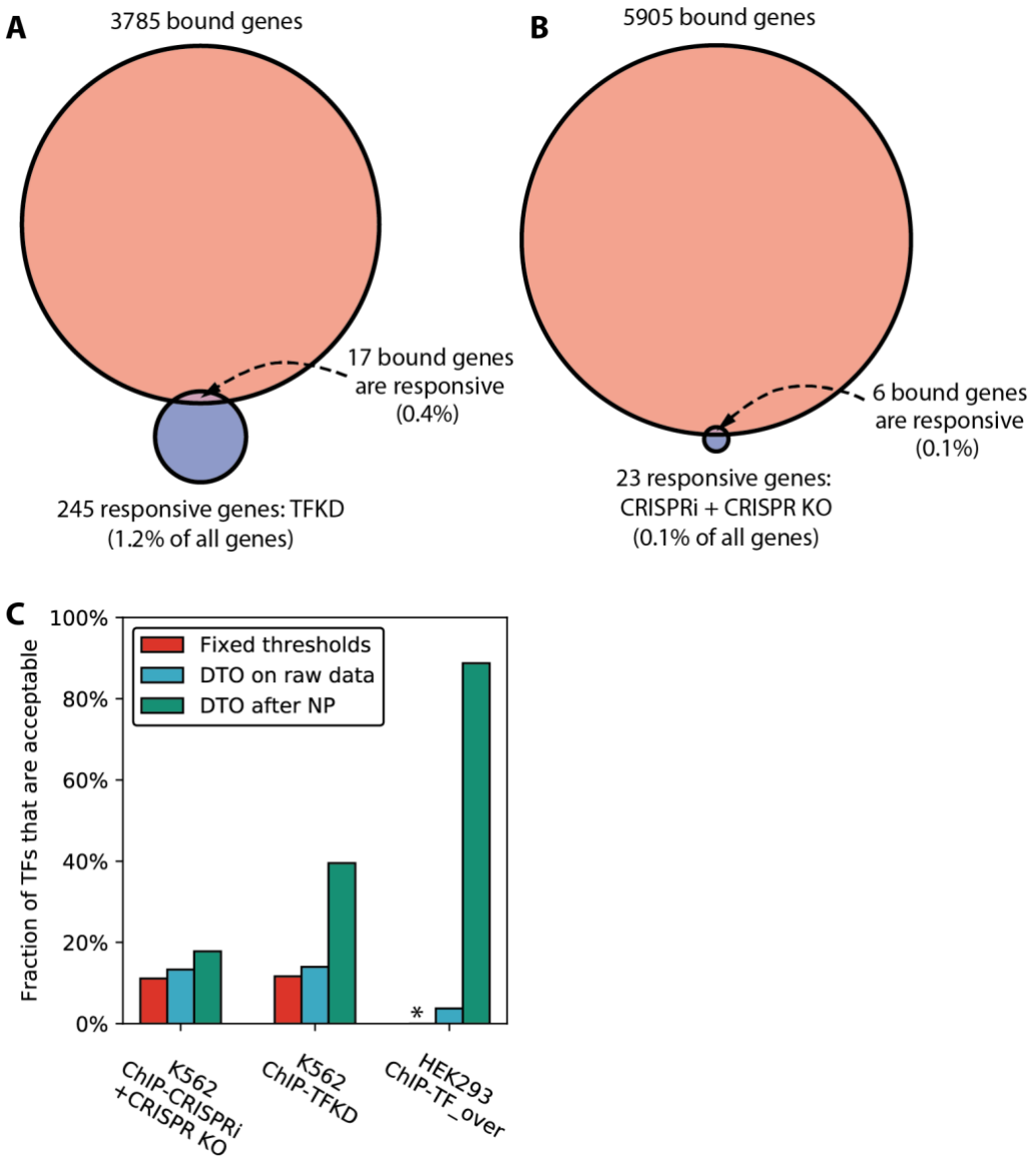


Figure 3. 3: Network inference with dual threshold optimization in human cell lines. (A) Medians of number of bound genes, number of perturbation-responsive genes, and number genes that are both bound and responsive, when comparing ENCODE K562 ChIP-seq data to ENCODE TFKD data. Excludes TFs with either no bound genes or no responsive genes. Binding threshold is $p<0.05$ and response threshold is $p<0.05$ with no minimum fold change. (B) Comparison of ENCODE K562 ChIP-seq data and ENCODE CRISPRi + CRISPR KO data, as in Panel A. (C) Comparison of human networks derived from fixed thresholds, dual threshold optimization (DTO) on raw perturbation-response data, and DTO on perturbation-response data processed by NetProphet 2.0. The vertical axis is the number of TFs showing acceptable convergence divided by the number that were both ChIPped and perturbed (K562: ChIP-CRISPRi + CRISPR KO = 45, K562: ChIP-TFKD = 43, HEK293: ChIP-TF_over = 80). Asterisk indicates that no fixed threshold analysis for HEK293 is available due to the lack of response p-values.

3.2.6 Processing human data through network inference algorithms greatly increases the number of acceptable TFs

We ran NetProphet 2.0 on both the K562 data (TFKD and CRISPRi+CRISPR KO) and the HEK293 data followed by DTO, limiting the total set of responsive genes to those with the top 500,000 (K562) or 300,000 (HEK293) NetProphet scores (see Methods for details). Among the TFs that were both perturbed and ChIPped, the number showing acceptable convergence increased from 6 to 8 (K562 CRISPRi+CRISPR KO), from 6 to 17 (K562 TFKD), and from 3 to 71 (HEK293 over expression; Fig 3.3C). Comparable results for other network inference algorithms are shown in Supplemental Fig. S3.4D. NetProphet and other inference algorithms can also infer targets for TFs that have not been directly perturbed, by exploiting correlation between the expression of the TF and its targets when other TFs are perturbed. Processing all the perturbation response data and evaluating only on the non-perturbed TFs, we found that the Inferelator scores yielded the largest number of TFs with acceptable convergence (Supplemental Fig. S3.4E). This is not surprising, since NetProphet weighs the response to direct perturbation heavily in its score. This suggests that, for TFs that have not been directly perturbed, Inferelator is the best choice of analysis tool.

We also compared the output of NetProphet 2.0 when run on HEK293 perturbation response data to a recently published ChIP-exo data set (Imbeault et al. 2017) focusing on KRAB Zinc finger TFs. ChIP-exo (Perreault and Venters 2016; Rhee and Pugh 2011, 2012; Rossi et al. 2018b) is a variant of ChIP-seq in which the affinity-purified chromatin is digested by an exonuclease, leaving much smaller pieces that are partially protected by protein. Of the 27 TFs that were in both perturbation and ChIP-exo data sets, 20 showed acceptable overlap with NetProphet scores. For the same 27 TFs, using the previously described ChIP-seq yielded 24 TFs

with acceptable convergence. This small difference may be due, in part, to the fact that the ChIP-exo experiments were done on a derivative cell line known as HEK293T.

3.2.7 In yeast, newer ChIP data do not necessarily yield better convergence with perturbation response

To assess whether the age of the Harbison ChIP-chip data was responsible for some of its limitations, we analyzed a 2011 ChIP data set from Venters et al. (Venters et al. 2011), which included 26 factors that were also chipped by Harbison and perturbed by TFKO and ZEV. The results did not improve on those of Harbison et al. (Fig. 3.4A).

3.2.8 In yeast, ChIP-exo yields better convergence than traditional ChIP

We also ran DTO on ChIP-exo data from yeast (Bergenholm et al. 2018; Holland et al. 2019; Rhee and Pugh 2011; Rossi et al. 2018a, 2018b). Twenty TFs had data in ChIP-exo, Harbison ChIP-chip, TFKO, and ZEV15, enabling all-way comparisons. Regardless of the perturbation-response data set, ChIP-exo showed acceptable convergence for more TFs than ChIP-chip did (Fig. 3.4B). (For the sixteen TFs with ChIP-exo data in four different growth conditions, we used the glucose-limited chemostat data as it gave the best results; dotted blue lines, Supplemental Fig. S3.5A, B). After processing the ZEV perturbation-response data through NetProphet 2.0, all 20 TFs showed acceptable convergence (Fig. 3.4B).

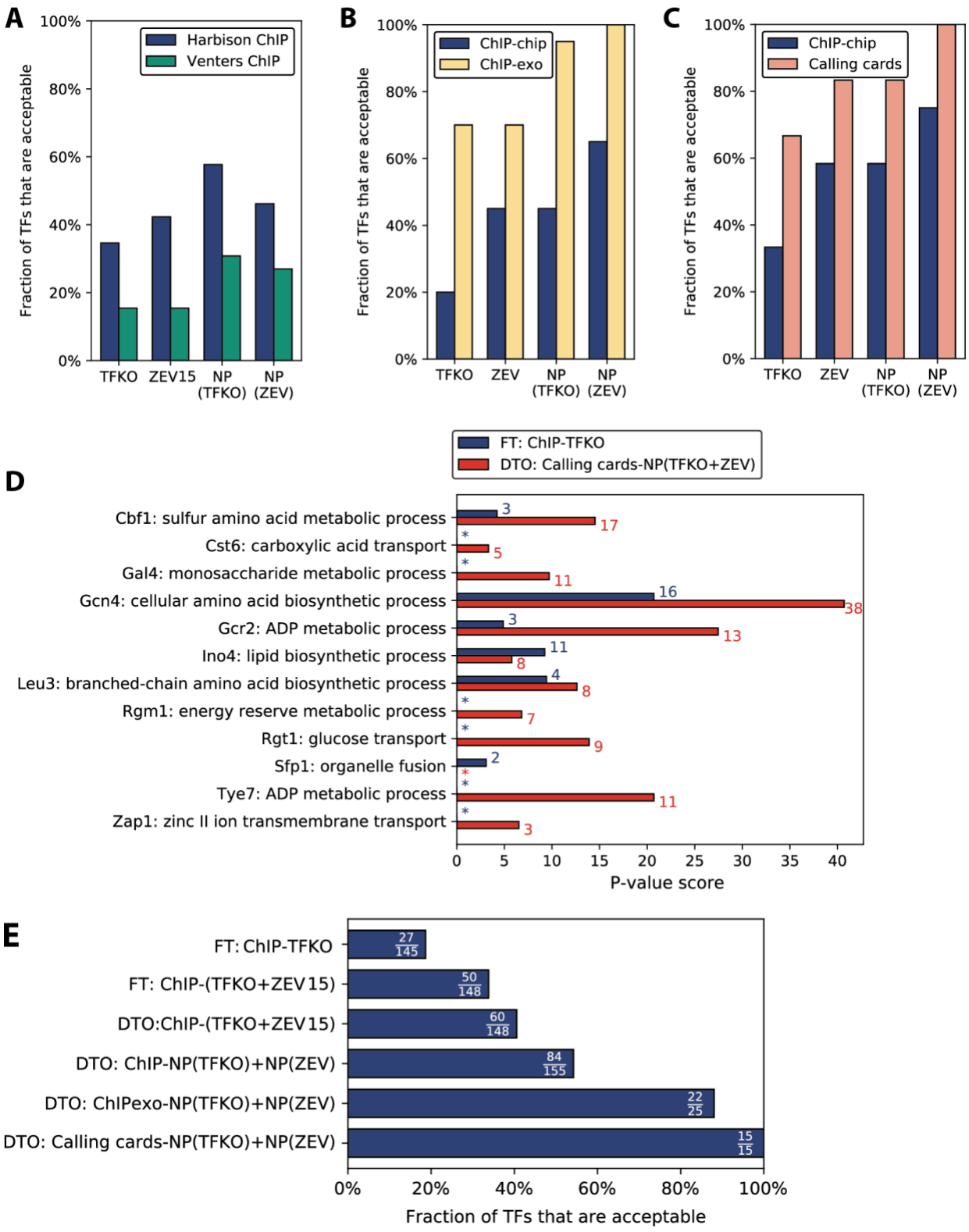


Figure 3. 4: Generating a high-confidence yeast TF network. (A) Percentage of TFs showing acceptable convergence, when comparing the Harbison ChIP and Venters ChIP data on the same 26 TFs. Regardless of the

perturbation data set or the processing by NetProphet 2.0, the Harbison ChIP data always yields more acceptable TFs. (B) Among the 20 TFs for which we have data in Harbison ChIP-chip, ChIP-exo, TFKO, and ZEV, the percentage that show acceptable convergence. Regardless of the perturbation data set or processing by NetProphet 2.0, ChIP-exo always yields more acceptable TFs. For both TFKO and ZEV, NetProphet postprocessing yields more acceptable TFs than raw differential expression. When NetProphet-processed ZEV data is compared to ChIP-exo data, all TFs show acceptable convergence. (C) Among the 12 TFs for which we have data in Harbison ChIP, calling cards, TFKO, and ZEV, the percentage that show acceptable convergence. When NetProphet-processed ZEV data is compared to calling cards, all TFs show acceptable convergence. (D) For each of the 12 TFs for which we have data in Harbison ChIP, calling cards, TFKO, and ZEV15, the Gene Ontology (GO) term that is most strongly enriched in the TF's targets. Targets are determined either by simple intersection of the bound and responsive genes in Harbison ChIP and TFKO data, using fixed thresholds (blue) or by dual threshold optimization on calling cards data and output from NetProphet 2.0 run on the TFKO and ZEV expression data (red). The colored numbers indicate the number of target genes annotated to the most significant GO term. Asterisk indicates no GO enrichment with $p<0.01$. (E) Among all TFs for which the indicated analyses can be carried out, the percentage that are acceptable in either TFKO or ZEV data or both. The fraction shows the number of acceptable TFs over the total number of TFs that could be analyzed. FT: Fixed threshold. DTO: Dual threshold optimization.

3.2.9 Transposon calling cards yields more acceptable TFs than traditional ChIP

Transposon calling cards is a method of determining TF binding locations by tethering a transposase to a TF, recovering the inserted transposons with their flanking sequences, and counting the insertions in a given genomic region. It does not require crosslinking, sonication, or affinity purification (Mayhew and Mitra 2016; Ryan et al. 2012; Wang et al. 2011b). Here, we analyze previously published calling cards data on seven TFs (Shively et al. 2019; Wang et al. 2011b) and new, never-before-analyzed data on eight TFs. Binding data from ChIP-chip and calling cards were compared to perturbation-response data from TFKO and ZEV15, using the 12 TFs present in all four data sets (Fig. 3.4C). In all comparisons, calling cards yielded substantially more acceptable TFs than ChIP-chip. This is particularly impressive given that the calling cards experiments were carried out in different growth conditions from the ZEV experiments -- synthetic complete medium with galactose on agarose plates at room temperature

versus minimal in phosphate-limited continuous-flow chemostats with glucose at 30°C (Hackett et al. 2020). Figure 3.4C also shows that, holding all other factors constant, ZEV was always better than TFKO and post-processing by NetProphet was always beneficial.

Figure 3.4D shows the -log p-value of the most significant Gene Ontology (GO) term for the predicted targets of each TF we have calling cards data on, excluding GO terms that describe more than 300 or fewer than 3 genes. To highlight the progress reported here, results are shown for the best combination of experimental and analytic methods (DTO on calling cards data and NetProphet output after processing TFKO and ZEV 15, 45, and 90-minute samples) compared to the simple intersection of bound and responsive genes using TFKO and ChIP-chip. For 10 of 12 TFs, the best combination of methods had a more significant GO term P-value, and the differences were large. For 2 of 12 (Ino4 and Sfp1), simple intersection had the more significant P-value, but the differences were smaller. The median -log10 P-value for the best combination of methods was 11.2, while that of simple intersection was 1.5. The best combination of methods assigned the top GO term to 117 target genes, whereas simple intersection assigned the top term to only 41 genes. For most TFs, the most significant GO term had a clear relationship to the known function of the TF. In some cases, the term selected is an immediate parent of the most familiar term associated with the TF. For example, Gcr2 (Glycolysis Regulation 2) is known as a regulator of genes encoding glycolytic enzymes. Its most significant GO term is “ADP metabolic process”, annotating 13 predicted Gcr2 targets, but 12 of those targets are also annotated with “Glycolytic process”, a subcategory of “ADP metabolic process”. This can be seen in Supplemental Figure S3.6, which shows the top 5 GO terms for each TF.

Another way to look at the contributions of various methods is to plot the fraction of available TFs that show acceptable convergence, combining TFKO and ZEV, using each

combination of methods described here (Fig. 3.4E). Only 15 TFs are currently available for calling cards and either ZEV15 or TFKO (12 for both), but analyzing these with DTO and NetProphet results in a much larger fraction of TFs being acceptable. This includes TFs that are not thought to be active in the ZEV or TFKO growth conditions, such as Gal4, presumably because ZEV overexpression of Gal4 significantly exceeds the number of Gal80 molecules available to bind and inactivate it. The second best percentage of TFs showing acceptable convergence was obtained by comparing NetProphet scores to ChIP-exo data (Fig. 3.4E).

3.2.10 The combination of ZEV and calling cards greatly increases response rates

We began this study by observing that, using fixed threshold analysis of the TFKO and ChIP data, most binding appears to be non-functional. To revisit the question of functionality using ZEV15 and calling cards data, we plotted the fraction of bound genes that are responsive, as a function of binding strength rank. Figure 3.5A shows that, for the TF Leu3, the combination of calling cards and ZEV15 gives much higher response rates than any of the other three combinations -- ChIP-ZEV15, calling cards-TFKO, or ChIP-TFKO -- regardless of binding strength. Nine out of the 10 mostly strongly bound and 48 out of 100 most strongly bound genes were responsive. To make the comparison between ZEV15 and TFKO fair, we fixed the number of Leu3-responsive genes in each perturbation data set to be the same. Thus, we labeled the 156 most strongly responsive genes in each data set as Leu3-responsive, because156 was the minimum of the numbers of genes that were significantly differentially expressed in the two data sets for Leu3. Although the number of responsive genes in each data set was the same, a larger fraction of the ZEV15-responsive genes was bound, as compared to the TFKO-responsive genes. Figure 3.5B shows a similar plot of the average response rates at each binding threshold, across

the 12 TFs for which we have all four combinations of data sets. Again, the combination of ZEV15 and calling cards gives higher response rates at all binding thresholds. On average, the response rate of the 10 most strongly bound genes is 61.7%. Individual rank response plots for the 11 other TFs present in all four data sets are shown in Supplemental Figure S3.7.

Figure 3.5C shows a direct comparison of binding strengths as assessed by calling cards, ChIP-exo, and ChIP-chip for the 8 TFs for which we had data from all methods. Each binding data set was compared to ZEV15 data on the same TF. At the highest binding strengths, calling cards appears to be a bit more discriminating, but ChIP-exo catches up when 20 or more top binding targets are considered. Both calling cards and ChIP-exo greatly outperform ChIP-chip.

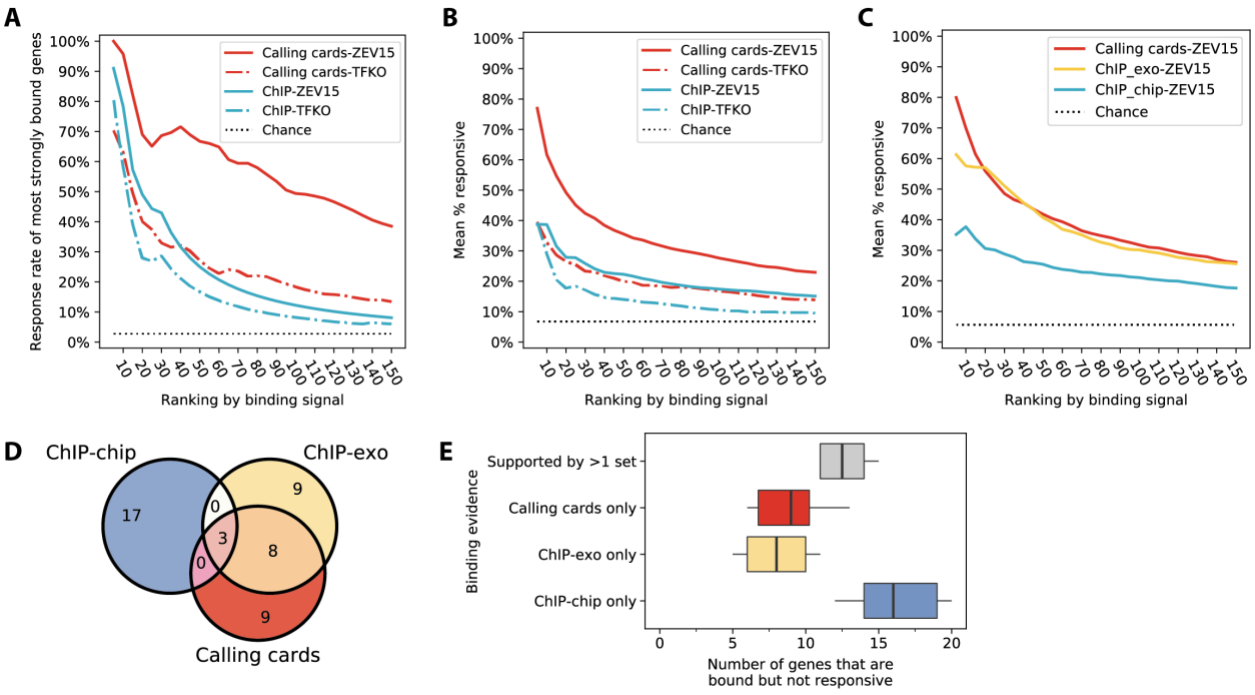


Figure 3. 5: Comparison of yeast perturbation-response and binding data sets. (A) The fraction of most strongly Leu3-bound genes that are responsive to Leu3 perturbation, as a function of the number of most-strongly bound genes considered. (B) Same as (A), with response rates averaged across the 12 TFs for which Harbison ChIP, calling cards, TFKO, and ZEV data were available. (C) Same as (B), with response rates averaged across the 8 TFs for which Harbison ChIP, calling cards, ChIP-exo, and ZEV15 data were available. (D) Venn diagram for the 20 genes that are most strongly bound by Leu3 in each assay but not responsive to Leu3 perturbation (ZEV15). Only

the top 20 non-responsive genes ranked by their binding strengths are shown. (E) The analysis for Leu3 shown in Panel A, applied to the eight TFs for which we have data in ChIP-chip, ChIP-exo, Calling cards, and ZEV. The three colored boxplots show the genes that are only bound in one of the three binding sets. The boxplot in grey shows the genes with evidence in at least two binding sets.

3.2.11 Comparison of non-responsive genes that are bound in each assay

Genes that appear to be bound by a TF but are not responsive to it could reflect false positives of the binding assay, non-functional binding sites, or genuinely bound genes that are not responsive because of network compensation, saturation, or other biological mechanisms (see Discussion). To estimate the contribution of false positives of the binding assays, we compared the bound but non-responsive targets according to each assay, for the 8 TFs for which we had all three assays (Fig. 3.5D). The non-responsive genes that are bound in only one assay are more likely to be false positives than those that are supported by multiple assays. Binding at these genes could be supported by another assay at a level below the threshold we used for this analysis, so we cannot conclude that they are definitely false positives. We found that ChIP-chip had more likely false positives than either calling cards or ChIP-exo, which were comparable to one another (Fig. 3.5E). The non-responsive genes that were supported by at least two assays are most likely true bound sites that are non-responsive for biological reasons. The relatively large size of this set suggests that there are a substantial number of truly bound, non-responsive genes.

3.2.12 Combining all available data sets yields the best result

ChIP-exo and calling cards data are not yet available for most yeast TFs. Furthermore, the data sets that are best overall may not be best on every TF. Therefore, we combined the data sets described above using NetProphet 2.0, DTO, and our FDR lower bound. We used the following procedure, which can be applied to any data sets available for any species:

Union the network edges produced by performing the following procedure on each perturbation-response data set:

1. Using the entire perturbation-response data set, run a suitable network inference algorithm that does not use binding location data either directly or indirectly. Rank all possible edges according to their score. If desired, multiple inference algorithms can be run (Marbach et al. 2012a).
2. For each TF:
    a. Compare the network inference scores of the TF's targets to each binding location data set using DTO to select thresholds. Among all binding data sets for the TF, choose the one that yields the best hypergeometric p-value.
    b. Using the chosen data set and DTO thresholds, check whether the TF is acceptable as defined above. If so, return edges from the TF to targets that are above the thresholds for both expression and binding.

We carried out this procedure on the TFKO and ZEV expression data with the Harbison ChIP, ChIP-exo, and calling cards binding data. For the TFs for which ChIP-exo or calling cards data were available, one of these data sets was chosen over Harbison ChIP 92% of the time (TFKO comparison) or 96% of the time (ZEV comparison). Considering all data sets, the resulting network comprises 96 acceptable TFs with 3,268 edges impinging on 1,686 unique target genes.

3.3 Discussion

The fundamental question behind this investigation is whether TF binding locations and TF perturbation responses could provide convergent evidence about the direct functional targets of each TF in an organism. Using standard methods to compare binding data from chromatin

immunoprecipitation (ChIP) to published perturbation-response data, we found that most of the genes whose cis-regulatory DNA is bound by a TF are not functionally regulated by that TF. We found this to be the case for two yeast ChIP datasets as well as ENCODE ChIP-seq experiments in human K562 cells and another 88 ChIP-seq experiments in human HEK293, consistent with previous reports based on different data sets (Cusanovich et al. 2014; Gitter et al. 2009; Hu et al. 2007; Lenstra and Holstege 2012).

If the problem is that most bound genes are not responsive, a natural solution would be to focus on those that are. That is, to take the intersection of the genes a TF binds and the genes that respond to perturbation of the TF as its direct functional targets. However, we proved that this procedure does not effectively identify the direct functional targets when the sets of bound and responsive genes are much larger than their intersection. The reason is that, when there are many genes with non-functional binding sites and many genes that respond to the perturbation because they are indirect targets, it is expected that some indirect targets will have non-functional binding sites in their cis-regulatory DNA. These are not direct functional targets, yet they inhabit and contaminate the intersection of bound and responsive genes.

We quantified this problem by setting minimal criteria for considering the genes that are bound and responsive to be likely direct functional targets. First, the intersection procedure must be able to achieve, in principle, 80% sensitivity with an expected false discovery rate of no more than 20%. Second, the intersection must be larger than would be expected by chance ($p<0.01$). We say that a TF shows acceptable convergence if it meets both those criteria. This designation does not guarantee that all or most of the TF's bound and responsive genes are responsive because they are bound. The 80-20 criterion is a lower bound on the expected FDR, not an upper bound. Furthermore, it does not guarantee a unique relationship between the bound and

responsive sets of an acceptable TF -- the bound set of one TF can show acceptable convergence when compared to the responsive set of a different TF. Acceptable simply means that there is no obvious red flag to prevent us from supposing that a good number of the TF's bound and responsive genes are direct functional targets. When combining ChIP data with steady-state perturbation-response data, the number of TFs showing acceptable convergence was no more than 15% of TFs assayed in both yeast and human data. For the remaining TFs, there is a clear red flag.

We identified four techniques that could substantially increase the number of TFs showing acceptable convergence.

1. Measuring the transcriptional response a short time after inducing overexpression of a TF by using a method such as ZEV.
2. Using dual threshold optimization (DTO) to set significance thresholds for binding and response data in a way that makes their intersection as significant as possible.
3. Processing all the perturbation-response data together through a network inference algorithm that does not use binding data, either directly or indirectly.
4. Measuring TF binding location by using transposon calling cards or (in yeast) ChIP-exo, rather than standard ChIP.

We combined all these methods to produce a high-quality yeast TF network, using the best binding data available for each TF. Currently, ~25% of the TFs in the network have binding data from calling cards or ChIP-exo; we expect the network to improve as these data are produced for more TFs. For mammalian cells, calling cards (Wang et al. 2012), dual threshold optimization, and network inference have all been shown to work to some degree. For TF activity perturbation, highly specific genome-targeting systems have been developed and tested with a variety of

activation and repression domains (Waryah et al. 2018) and linked to small-molecule inducers (Kundert et al. 2019; Oakes et al. 2016). However, the prospects for obtaining ZEV-like perturbation and calling cards binding data on large numbers of mammalian TFs remain uncertain.

Other new technologies for measuring TF binding locations have shown great promise (Policastro and Zentner 2018), but have not yet yielded a sufficiently large, systematic data set, with matched perturbation-response data, for comparison to ChIP and calling cards. One such technology is DamID, in which a DNA-methyltransferase is tethered to a DNA-binding protein and changes in DNA methylation relative to a control are assayed to determine binding location (Hass et al. 2015; Tosti et al. 2018; Van Steensel and Henikoff 2000). Another is CUT&RUN, in which an endonuclease tethered to an antibody against a TF enters permeabilized nuclei and releases the DNA bound by the TF, which diffuses out of the cell and is recovered for sequencing (Hainer and Fazzio 2019; Meers et al. 2019b; Skene et al. 2018; Skene and Henikoff 2017). A promising approach for measuring perturbation-response in mammalian cells is to transfect cells with a library of constructs encoding guide-RNAs that target a variety of TFs and then use single-cell RNA-seq to identify the TF perturbed and measure the response. Variants of this general approach include Perturb-seq (Adamson et al. 2016; Dixit et al. 2016; Replogle et al. 2018), CROP-seq (Datlinger et al. 2017), and CRISP-seq (Jaitin et al. 2016). As these technologies mature, they will likely be used to produce large, systematic data sets that can be analyzed using the methods described here.

Even when we apply the best combination of analytic and experimental methods, a large fraction of the genes whose regulatory DNA is significantly bound by a TF does not respond to a perturbation of that TF. Such non-responsiveness could be caused by several mechanisms.

- Insufficient occupancy -- rank response plots (Fig. 3.5A-C) indicate that the most strongly bound sites are much more likely to be functional than sites that are bound less strongly, even when the weaker sites are statistically significant.
- Saturation -- if a gene is already expressed at its maximum possible level and an activator of that gene is induced, no response will be seen. However, if other TFs were removed, lowering the expression level of the gene, it would respond to the induction. The same situation arises when a repressor of an unexpressed gene is induced or an activator of it is depleted.
- Inactivity -- the TF may bind DNA even when the TF is in an inactive or partially active state. However, ZEV induction of Gal4 activates galactose genes even in the absence of galactose and presence of glucose, showing that overexpression can elicit a response in conditions where a TF is normally inactive.
- Compensation -- the regulatory network as a whole may compensate for the change in TF activity in a way that damps the effect of the initial perturbation. Measuring responses shortly after the perturbation should reduce the prevalence of such compensation, but some mechanisms can compensate quickly. A simple example would be two essentially equivalent TFs that can bind to the same sites, so that the effects of perturbing one TF are buffered by the other. This was shown to be a contributing factor in a comparison of the Harbison ChIP data to the TFKO data from Hu et al (Gitter et al. 2009; Hu et al. 2007).
- Override -- some regions of a genome may be shut down in a way that overrides the effects of TFs, even when the TFs can bind to the cis-regulatory DNA. For example, the transcribed region of a gene might be in inaccessible, tightly compacted DNA even though the cis-regulatory region remains somewhat accessible to TFs.

- Synergistic regulation -- some TFs that are bound to cis-regulatory DNA may be active only where there is a binding site for a cofactor nearby.

Regardless of the mechanism that renders a bound gene non-responsive, it remains the case that many binding sites are non-functional under the conditions tested, in the sense that the transcription rate of the associated gene is unaffected by the presence or absence of the TF. Currently, we do not know how much each of the factors listed above contributes to explaining why so many genes that are bound by a TF do not respond to a perturbation of that TF. For now, technical limitations of the available data sets may be a significant contributing factor. Once those have been mitigated by newer methods like transposon calling cards, we will be in a strong position to investigate the biological factors that explain the non-responsiveness of genes whose cis-regulatory DNA is bound by a TF. Determining the prevalence of each factor will bring the landscape of transcriptional regulation into much clearer focus.

3.4 Methods

3.4.1 Data preparation

Yeast gene and TF definitions

For all yeast analyses, we considered the 5,887 genes labeled as "ORF verified" or "uncharacterized" in the Saccharomyces Genome Database (SGD), discarding the 1,127 labeled as "dubious", "ncRNA", "rRNA", "snoRNA", "snRNA", or "tRNA". We only considered TFs with evidence of direct DNA binding via a DNA binding domain. To identify these, we compared multiple lists, including those that had been ChIPped by Harbison et al., those that were over-expressed in the ZEV data, and those that had DNA binding specificity models in the CIS-BP database (Weirauch et al. 2014). In cases of disagreement, we curated the list manually by consulting data in SGD, focusing primarily on domain analysis of the protein and on gene

ontology categories assigned via high-throughput experiments such as protein-binding microarrays. In most cases the judgment is clear but there are some borderline cases that require a best guess.

Yeast ChIP-chip data sets

The Harbison ChIP-chip binding location data was published in (Harbison et al. 2004). We downloaded the p-values that represent the significance of TF binding within the intergenic regions from http://younglab.wi.mit.edu/regulatory_code/GWLD.html. Following the authors' recommendation, targets were considered significantly bound if their p-value was less than or equal to 0.001. TFs with no significantly bound targets were eliminated from further analysis. The Venters ChIP-chip data were published in (Venters et al. 2011). We downloaded the occupancy-level profiles for 200 transcription-related proteins from Table S4a in (Venters et al. 2011). The log2 fold change of experimental signal over background signal within each promoter was used as the binding signal strength. The probes covered a distal region (260-320 bp upstream of ATG) and a proximal region (30-90 bp upstream of ATG). The downloaded occupancy level took the maximal level from either regulatory region. The authors used an FDR threshold of 5%, but we used 1% in order to make the data more comparable to those from the Harbison data set. For each TF, an FDR cutoff was calculated by searching for an occupancy level such that the ratio of number of targets in the mock IP control over the number in the experimental sample reaches the desired FDR. The "25&37C merged MockIP controls" file was obtained directly from the authors as it was unpublished. TFs with no significantly bound target were eliminated from further analysis.

Yeast ChIP-exo data sets

The ChIP-exo data for 26 TFs were compiled from four resources (Bergenholm et al. 2018; Holland et al. 2019; Rhee and Pugh 2011; Rossi et al. 2018a, 2018b). We downloaded genomic coordinates of the ChIP peaks for Reb1, Gal4, Phd1 and Rap1 published in ref. (Rhee and Pugh 2011) from https://ars.els-cdn.com/content/image/1-s2.0-S0092867411013511-mmc2.xls. We mapped the peaks to genes using the coordinates of each gene's promoter region (700 bp upstream to ATG) in reference genome S288C-R55, which was the last release prior to the date cited in the paper, "(build: 19-Jan-2007)" (ref. (Engel et al. 2014) lists all releases). We then calculated each TF's binding strength at each promoter as the sum of all the TF's in that promoter. We downloaded peaks for Abf1 and Ume6 generated using a newer protocol, ChIP-exo 5.0, described in ref. (Rossi et al. 2018b), from GEO Series GSE110681. We also downloaded peaks for Cbf1 from GEO Series GSE93662 (see ref. (Rossi et al. 2018a)). The assignment of peak-promoter and binding strength at promoter were calculated as for the data from ref. (Rhee and Pugh 2011), except that both peak and promoter coordinates were based on gene annotation from reference genome S288C-R64. Lastly, we obtained ChIP-exo data for 20 TFs (Cat8, Cbf1, Ert1, Gcn4, Gcr1, Gcr2, Hap1, Ino2, Ino4, Leu3, Oaf1, Pip2, Rds2, Rgt1, Rtg1, Rtg3, Sip4, Stb5, Sut1, Tye7) directly from the authors of (Holland et al. 2019) and (Bergenholm et al. 2018). Each TF was assayed in four different environmental conditions, but we focused on the glucose limited chemostat data as that gave the best agreement with both TFKO and ZEV15 response data. This data set contains scores for each promoter that had at least one peak assigned to that promoter. We directly used the highest score at each promoter as the binding strength, after removing any peak that was > 700 bp upstream from ATG. Any promoter without a score in the file was assigned a score of zero.

Yeast transposon calling cards data

We combined calling cards data from (Wang et al. 2011b) and (Shively et al. 2019) on Cbf1, Cst6, Gal4, Gcr1, Gcr2, Rgm1, and Tye7 with new data on Eds1, Gcn4, Ino4, Leu3, Lys14, Rgt1, Sfp1, and Zap1. For each TF, all data from all replicates were combined. Transpositions within the promoters of yeast genes (700 bp upstream to ATG, reference genome S288C-R61) were used for calculating the significance of TF binding. For each promoter, a Poisson p-value was calculated by comparing the experiment sample with a no-TF control sample as described (Wang et al. 2012). To obtain a ranking by calling cards signal strength for dual threshold optimization, we used the normalized transposition count of the experimental samples minus that of the control samples to break ties when promoters had identical p-values.

Yeast TFKO data

The microarray expression data on gene knockout strains was published in ref (Kemmeren et al. 2014). The gene expression profiles of 1,484 single gene deletion strains and 3 wild type replicates were downloaded from http://deleteome.holstegelab.nl/data/downloads/deleteome_all_mutants_controls.txt. In addition, we downloaded the gene expression profiles after removal of the slow growth signature removed http://deleteome.holstegelab.nl/data/downloads/deleteome_all_mutants_svd_transformed.txt. This transformed data set did not contain new p-values and analyzing it did not produce better results than the untransformed data, so we focused on the untransformed data.

Yeast ZEV induction data

The ZEV induction system was described in ref. (Hackett et al. 2020). The shrunken expression profiles were used as the quantitative responsiveness of target genes after TF induction. Specifically, the file “Raw & processed gene expression data” was downloaded from https://idea.research.calicolabs.com/data and the column labeled “log2_shrunken_timecourses”

was used. The responsive set contains all targets with non-zero expression levels. We systematically analyzed ZEV expression profiles measured at all time points (5, 10, 15, 20, 30, 45, and 90 minutes) after TF induction. To make different time points comparable, we only focused on 103 TFs that were available in the Harbison ChIP-chip data and each time point of the ZEV data. The maximal number of acceptable TFs was obtained at 15 min, so we chose to move forward with this time point for all subsequent analyses except those that involve network inference. For network inference, we used the 15, 45, and 90 minute samples.

Human ChIP-seq data

Two human ChIP-seq data sets were analyzed in this work: ChIP-seq in K562 cell line published by ENCODE, and ChIP-seq in HEK293 cell line published in ref. (Schmitges et al. 2016). All ENCODE data were downloaded from www.encodeproject.org as of January 21st, 2019. We focused on the data on K562 cells because it had by far the most TFs with both ChIP-seq and perturbation response data. We downloaded the "conservative" ChIP-seq peaks mapped to GRCh38 as called by the ENCODE pipeline, which uses the Irreproducible Discovery Rate (IDR) analysis of biological replicates with 2% IDR cutoff. Using the ENCODE definition of transcription factor, there was ChIP-seq data for 261 TFs in K562. To quantify the significance of each TF-target binding interaction, we summed the log10 q-values of significant peaks that were within the regulatory regions of each gene (defined below). For the HEK293 cell line, ChIP-seq was carried out using an antibody against GFP. We downloaded the combined summits for 131 ChIPped zinc finger proteins from GEO Series GSE76494 (see ref. (Schmitges et al. 2016) for details). The binding strength within the regulatory regions of a gene was the summed scores of all summits assigned to those regions. We tried two definitions of regulatory region: (1) a single long promoter extending from 10 Kb upstream of the 5'-most transcription start site

(TSS) to 2Kb downstream (Ensembl Release 92), or (2) a core promoter extending from 500 bp upstream of the TSS to 500 bp downstream combined with the gene's enhancers from the GeneHancer database V4.8 (Fishilevich et al. 2017). We used only the "double elite" enhancers, for which both the existence of the enhancer and the gene-enhancer association are supported by at least two evidence sources. This double-elite list was obtained by emailing the authors of the paper. In order to properly use the ChIP summits in HEK293 whose coordinates were based on GRCh37, we used the LiftOver tool in UCSC genome browser to lift over the coordinates of regulatory regions from GRCh38 to GRCh37.

Human ChIP-exo data

A collection of ChIP-exo data of 221 KRAB zinc-finger proteins in HEK293T cell lines was downloaded from GSE78099 (Imbeault et al. 2017). We mapped the MACS peaks obtained from the supplemental files to the regulatory regions defined above (GRCh37). We then summed the scores of peaks for each gene to represent the binding strength between each protein and its target.

Human TFKD, CRISPRi and TF-induction data

We considered three human perturbation response data sets: TF knockdown (TFKD) in K562, CRISPRi in K562, and TF-induction in HEK293 (Schmitges et al. 2016). The RNA-seq expression profiles of wild-type controls, TFKDs and CRISPRi were downloaded from the ENCODE web site. Knockdowns using small-interfering RNA (siRNA) or small-hairpin RNA (shRNA) were combined in the data set we referred to as TFKD while the CRISPRi and CRISPR TF-disablement data were combined in the data set we referred to as CRISPR. For K562 cells, there were TFKD experiments targeting 261 different proteins and CRISPRi experiments targeting 96 different proteins. The expected counts were reported by the RSEM program in the

ENCODE RNA-seq processing pipeline using gene annotation from GENCODE V24 (GRCh38). Differentially expressed genes in each perturbed TF strain were processed by comparing the experimental replicate set to the control set using DESeq2 (V1.10.1). On the TF-induction for HEK293, RNA-seq was carried out 24 hours after overexpressing the TF from a tetracycline-inducible plasmid. For the majority of TFs there was only a single replicate of the RNA-seq experiment, which prevents the calculation of statistical significance by traditional methods. The processed RNA-seq expression profiles (after lowly-expressed gene removal and batch normalization) for 80 induced zinc finger proteins were downloaded from GEO Series GSE76495. Since there were no control replicates, we used the expression levels in each profile, normalized to the medians of the respective batches, as the response strength (Schmitges et al. 2016).

3.4.2 Expected false discovery rate of intersection algorithms

Intersection algorithms identify the direct functional targets of a TF as those whose promoters are bound by the TF in an assay such as ChIP-seq and are responsive when the same TF is perturbed. A true direction function (DF) target is responsive when the TF is perturbed *because it is bound by the TF*. One possible alternative is that a gene is in the intersection because it is an indirect target of a TF and happens, by chance, to have a non-functional binding site for the same TF in its promoter. Although we use the terms non-functional binding site and indirect target, the analysis is unaffected if there are simply false positives of the binding or response assays.

We started by defining the following notation for any given TF:

B the set of genes whose promoters appear to be bound by the TF in an experiment

R the set of genes that appear to be responsive when the TF is perturbed in an experiment

G the set of all genes assayed in both the binding and response experiments

DF the true set of direct functional targets of the TF

DF is unknow, but B, R and G are all observed outcomes of the experiments. B ∩ $\overline{\mathrm{DF}}$ is the set of genes with only non-functional binding of the TF (the overbar indicates set complement). Genes with functional binding are in $B \cap DF$. Likewise, $R \cap \overline{DF}$ is the set of indirect targets – genes that are responsive but not direct functional targets.

This analysis is based on the idea that the promoters with only non-functional binding for a TF can be modeled as though they were scatter randomly across genes, without regard to whether the gene are indirect targets of the same TF. (In fact, we only need to assume that promoters with non-functional binding don't systematically avoid indirect target genes.) We believe this is a good assumption because we cannot think of any molecular or evolutionary mechanism by which non-functional binding sites could be enriched or depleted in the promoters of indirect target genes. Since they are non-functional, they are not under any evolutionary selection. The same applies to false positives of the binding and/or responsive assays – there is no reason to believe that false positives of the binding assay would be enriched or depleted among the false positive of the response assay.

According to this model, the genes with only non-functional binding are selected at random from $\overline{DF}$, so the expected fraction that are also in $R \cap \overline{DF}$ is simply $|\mathrm{R} \cap \overline{DF}|/|\overline{DF}|$, where the vertical bars indicate set size. Thus

$$E[|R \cap B \cap \overline{DF}|] = |B \cap \overline{DF}| \frac{|R \cap \overline{DF}|}{|\overline{DF}|} \tag{3.2}$$

By way of analogy, it is as though $|B \cap \overline{DF}|$ balls were selected at random from a jar containing $|\overline{DF}|$ balls, of which $|R \cap \overline{DF}|$ are red and the remainder are white. The expected number of red balls is given by the right-hand side of formula (3.2).

By definition, the false discovery rate (FDR) of the intersection algorithm is:

$$\text{FDR} = \frac{|R \cap B \cap \overline{DF}|}{|R \cap B|} \tag{3.3}$$

The denominator is directly observable from the assays. We do not know the true set of DF targets, but have just derived the expectation of the numerator with respect to the random process that distributes promoters with non-functional binding sites to genes.

We also observed that the sensitivity of the intersection algorithm is, by definition:

$$Sn = \frac{|DF \cap R \cap B|}{|DF|} \leq \frac{|R \cap B|}{|DF|}$$

So

$$|DF| \leq \frac{|R \cap B|}{Sn} \tag{3.4}$$

Putting these equations togethers,

$$\begin{aligned}
\text{E}[FDR] &= \frac{E[|R \cap B \cap \overline{DF}|]}{|R \cap B|} \\
&= \frac{|B \cap \overline{DF}||R \cap \overline{DF}|}{|\overline{DF}||R \cap B|} \\
&\geq \frac{max(0, |B| - |DF|)max(0, |R| - |DF|)}{|\overline{DF}||R \cap B|} \\
&\geq \frac{max(0, |B| - |R \cap B|/Sn)max(0, |R| - |R \cap B|/Sn)}{|\overline{DF}||R \cap B|} \\
&\geq \frac{max(0, |B| - |R \cap B|/Sn)max(0, |R| - |R \cap B|/Sn)}{|\text{G}||R \cap B|}
\end{aligned} \tag{3.5}$$

3.4.3 NetProphet analysis

NetProphet 2.0 is a TF network inference algorithm that exploits gene expression data under genetic or environmental perturbation and genome sequences with annotations. The algorithm is described in detail in ref. (Kang et al. 2018). Here, two yeast TF networks were mapped, one using the Kemmeren TFKO gene expression data and one using the ZEV data. Three human TF networks were mapped using the three perturbation response data sets.

Yeast NP networks

NetProphet 2.0 requires gene expression data in the form of a gene expression matrix and differential expression matrix. The Kemmeren gene expression matrix was represented as the log2 fold-change (logFC) values of strains with gene deletions over wild-type strains. The ZEV gene expression matrix was represented as the logFC values of the levels measured at a certain time point after the TF induction relative to time 0. For the differential expression (DE) module of the algorithm, we used Kemmeren samples in which a TF-encoding gene was knocked out, not those in which some other type of gene was knocked out, and ZEV samples from 15 min after TF induction. For the co-expression module of the algorithm, we used the complete set of 1,484 Kemmeren expression profiles from strains lacking one gene (not necessarily encoding a TF) or 590 ZEV expression profiles from 15 minutes, 45 minutes, or 90 minutes post-induction. The other two inputs were DNA sequences of yeast promoters and amino acid sequences of TFs' DNA binding domains (DBDs), as described in ref. (Kang et al. 2018). PWM models of TFs' binding specificity were not used. Each output network is an adjacency matrix, where the rows represent TFs, the columns represent genes, and the entries are NetProphet scores representing the aggregate strength of evidence that the gene is a direct functional target of the TF.

Human NP networks

For K562 data, we calculated differential expression (DE) p-values for TFKD and CRISPRi independently using DESeq2. The DE matrix input to NetProphet 2.0 contained the -log p-values, with a negative sign for apparent repression (the knockdown of the TF makes the target gene go up). For HEK293 data, we directly used the logFC values because there were no replicates, hence no p-values, for most TFs. The co-expression matrix contained the logFC of individual mutant strain replicates over the median expression level of control replicates (K562) or the median expression level of the gene across all perturbations (HEK293). There were 765 expression profiles (including replicates) for K562 TFKD, 252 for K562 CRISPRi, and 107 for HEK293 TF inductions. We obtained DNA sequences of the regulatory regions based on their coordinates in GRCh38 using our definition (2) of regulatory regions. We concatenated the enhancers and promoter of each gene into a single sequence for the purpose of motif inference in NetProphet 2.0. Each pair of concatenated regions was separated by 50 N's to ensure that no inferred motif instances crossed between one enhancer and another. We also queried the CIS-BP database (Weirauch et al. 2014) for the amino acid sequences of human TFs' DBDs. The details of DBD preprocessing are described in (Kang et al. 2018). The TFKD and CRISPRi networks had 392 TFs each, which were ENCODE TFs being ChIPped in either of the major cell lines K562 or HepG2. The TF-induction network had 103 TFs, which were the zinc finger proteins being ChIPped in HEK293.

3.4.4 Analysis using other network inference algorithms

GENIE3 (Huynh-Thu et al. 2010) (v1.16.5, Python implementation) was downloaded from http://www.montefiore.ulg.ac.be/?huynh-thu/software.html. Default parameters in GENIE3 were used. The version of Inferelator that incorporates Bayesian Best Subset Regression

(Greenfield et al. 2013) was downloaded from https://github.com/ChristophH/Inferelator. No prior network was used because our intention is to infer networks from perturbation response data without any influence from data on TF binding locations. Otherwise, default parameters were used. MERLIN (Roy et al. 2013) was downloaded from https://github.com/marbach/gpdream and used with default parameters. We input the same gene expression matrix, TF list and gene list to NetProphet 2.0, GENIE3, Inferelator, and MERLIN.

3.4.5 Acceptable TFs

For each pair of binding and expression data on a given TF, the positive gene sets (bound genes or responsive genes) were compared. The TF was deemed acceptable if they met two criteria. (1) The lower bound on the expected FDR had to be less than or equal to 20% when the sensitivity was fixed at 80%, as calculated by using the formula (3.1) derived in Methods. (2) The p-value for the significance of the overlap between the bound and responsive gene sets had to be <= 0.01. For fixed threshold analysis, the p-value was calculated using the hypergeometric null distribution. For dual threshold analysis, it was calculated using the randomization-based null distribution, not the nominal p-value (see description of dual threshold optimization).

3.4.6 Dual threshold optimization

Software availability

Software implementing dual threshold optimization and instructions can be found at https://github.com/BrentLab/Dual_Threshold_Optimization.

DTO algorithm

For each TF, dual threshold optimization (DTO) uses one binding location data set and one gene expression data set. The genes in each data set are ranked by the strength of their binding or expression signal. By default, the signal strength is the negative log p-value, but

different experiments and methods may require different calculations of signal strength (see sections below). For each data set, DTO chooses a threshold on the ranks such that genes ranking above the threshold are considered positives for binding or response (see Fig. 3.2C). A series of rank-threshold combinations (places where the gray lines cross in Fig. 3.2C) are used to generate positive subsets of genes in each dataset. The series of thresholds for each data set, T_1, T_2, …, were generated using the recurrence:

$$T_1 = 1$$

$$T_n = Floor(T_{n-1} * 1.01 + 1)$$

This formula produces a fine spacing among smaller subsets that becomes coarser as the subsets grow. If a threshold would split a group of genes that all have the same score that threshold is skipped. For each pair of subsets, a hypergeometric p-value was computed using a hypergeometric survival function (Scipy's hypergeom.sf) with the following parameters:

k = # of genes in the intersection of the subsets - 1

M = # of genes in the universe of assayed genes

n = # of genes in the expression subset

N = # of genes in the bound subset

This hypergeometric p-value is the probability of an intersection as large as, or larger than, the observed intersection, when choosing random subsets of genes, with the number of genes in each random subset equal to the number of genes in the positives defined by the rank threshold pair. We refer to this as the nominal p-value because it is only used for selecting the best pair of thresholds, not for determining whether the resulting overlap is significantly larger than would be expected by running DTO on random rankings. DTO returns the threshold combination that minimizes the nominal p-value.

Randomization-based p-values for overlaps identified by DTO

To produce a null distribution for testing the significance of the overlap chosen by DTO, a randomization procedure was used. A new set of data was generated using random assignment of signal strength scores to genes and then DTO was run. The best nominal p-value for each randomized data set was used to calculate a null distribution of nominal p-values. This was done by running the randomized DTO procedure 1000 times for each TF in each analysis and the distribution of nominal (hypergeometric) p-values enabled us to determine a $P<0.01$ significance threshold on the nominal p-value that is specific to that TF. When the nominal p-value of the rank pair chosen by DTO using the true data was below the threshold defined by the randomizations, the overlap was considered significant.

Application of DTO to yeast data

In the analysis of yeast data, the universe was defined as the set of all genes assayed in either of the two datasets being compared. For data sets that do not have p-values, the signal strength is the log fold change (ZEV) or score (NetProphet 2.0, GENIE3, Inferelator or MERLIN). For Calling cards, where many p-values were identical, ties were broken by the difference between the number of insertions in the experimental sample and the number of insertions in the control sample. The number of insertions was normalized to the total insertion count in each sample.

Occasionally, DTO can produce implausible results, such as concluding that all genes are responsive to a perturbation. To prevent this, we set very relaxed limits on the bound or responsive genes in certain data sets. For Harbison ChIP, we required $P<0.1$. For each inferred network output we required that the score of the TF-target relationship be among the top 150,000

scores. These were sufficient to eliminate any anomalous results; no constraints on the TFKO or ZEV data were required.

Application of DTO to human ENCODE data

In the analysis of human data on K562 cells, the universe was defined as the set of all genes detected in the gene expression dataset. Response signal strength for DTO was the absolute value of the log fold change, relative to non-perturbed control samples. DTO was limited to choosing bound or responsive genes with P<=0.1. For each inferred network output, the score of the TF-target relationship was required to be among the top 500,000 scores.

Application of DTO to human HEK293 data

In the analysis of human data on HEK293 cells, the universe was defined as the set of all genes detected in the gene expression dataset. Response signal strength for DTO was the absolute value of the log fold change, relative to non-perturbed control samples. No replicates or p-values were available for most TFs. For both raw perturbation-response data and inferred network scores, the score of the TF-target relationship was required to be among the top 300,000 scores.

3.4.7 Comparisons among binding data sets

Harbison ChIP-chip compared to Venters ChIP-chip

The TFs used in the comparison shown in Figure 3.4A are: Ash1, Cha4, Cin5, Fkh1, Fkh2, Gal4, Gcn4, Gln3, Ino4, Leu3, Msn2, Pho2, Rfx1, Rph1, Sfp1, Skn7, Stp1, Swi5, Uga3, Wtm1, Wtm2, Xbp1, Yap5, Yap6, Zap1, Zms1

Harbison ChIP-chip compared to ChIP-exo

The TFs used in the comparison shown in Figure 3.4B are: Cat8, Cbf1, Ert1, Gal4, Gcn4, Gcr2, Hap1, Ino4, Leu3, Oaf1, Phd1, Pip2, Rds2, Rgt1, Rtg1, Rtg3, Sip4, Stb5, Sut1, Tye7

Harbison ChIP-chip compared to transposon calling cards

The TFs used in the comparison shown in Figure 3.4C are: Cbf1, Cst6, Gal4, Gcn4, Gcr2, Ino4, Leu3, Rgm1, Rgt1, Sfp1, Tye7, Zap1

3.4.8 Rank response plots

To create the lines in the rank response plots such as Figure 3.5A, we first determined the minimum of the number of responsive genes in the TFKO and the ZEV15 data -- call it n. A gene was considered responsive in the TFKO data if it had adjusted $P<0.05$ and in the ZEV15 data if it had shrunken absolute log fold change > 0. We then labeled the top n most strongly responsive genes in the TFKO and ZEV15 data as responsive for purposes of this plot (see "DTO algorithm" above for definitions of signal strength). This equalized the number of ZEV15-responsive and TFKO-responsive genes for each TF. We then sorted genes by the strength of their binding signal for the TF in question. Next, we considered the top 1, 2, 3, 4, etc. most strongly bound genes. For each such group, we calculated and plotted the fraction of genes that were responsive. For the mean rank-response plot (Fig. 3.5B) we simply averaged the response rates across the 12 TFs. In comparison of ChIP-chip, ChIP-exo and calling cards (Fig. 3.5C), we averaged the response rates across the 8 TFs that are present in all 3 binding data sets.

3.4.9 GO enrichment analysis

Gene ontology (GO) enrichments for each TF were analyzed using two networks mapped using different methods: (1) fixed threshold on Harbison ChIP data and TFKO data; (2) DTO on calling cards data and output from NetProphet 2.0 run on the TFKO and ZEV response data. For each TF, its target set in network 2 was the union of the output of DTO applied to NetProphet scores from ZEV expression data and DTO applied to NetProphet scores for TFKO expression data. The mapping of GO term to gene for Saccharomyces cerevisiae was queried using R

Bioconductor library org.Sc.sgd.db (V3.5.0). Any GO terms annotated with less than 3 or greater than 300 genes were eliminated. Using the target genes of a TF identified from network (1) or (2), the GO enrichment in biological process was analyzed using the hypergeometric test implemented in R Bioconductor library GOstats (V2.44.0). The output p-values were used for ranking the enriched terms, from most significant to the least significant. When plotting the top GO term shown in Figure 3.4D, for each TF, we combined all terms from both networks into a single rank list. If multiple terms were enriched by the same set of targets, only the most specific term was retained, based on the GO hierarchical structure, i.e. redundant ancestral terms were removed from the rank list. Subsequently, the top GO term was chosen for the corresponding TF. When plotting the top GO terms shown in Supplemental Figure S3.6, we used the GO terms enriched in one network and chose the top five (if available) as described above.

Chapter 4:

Elucidating the biological determinants of transcriptional responses to TF perturbations

4.1 Introduction

Understanding the function of a genome requires knowing which transcription factors (TFs) directly regulate each gene. A systems-level understanding should also enable us to predict which genes will change in expression level in response to direct perturbations of TFs. It was hoped that determining where in the genome each TF binds by chromatin-immunoprecipitation (ChIP) would go a long way toward solving these problems, but several studies have shown that the set of genes whose promoters are bound by a TF and the set of genes that respond when that TF is perturbed do not overlap much (Gitter et al. 2009; Lenstra and Holstege 2012; Cusanovich et al. 2014; Kang et al. 2020). Genes that are responsive but not bound may be indirect targets of the TF. The genes that are not responsive despite the fact that their regulatory DNA is bound by the perturbed TF constitute a greater mystery. Currently, we cannot predict which bound genes will respond to a perturbation and which will not. In this study, we take on the challenge of predicting whether a gene will respond to perturbation of a TF by using data on where the TF binds along with a variety of TF-independent features of each gene, including histone marks (HMs), chromatin accessibility, dinucleotide frequencies, and the gene's pre-perturbation expression level and expression variation.

A number of studies have shown success in predicting the expression *levels* of genes by using TF binding signals (Middendorf et al. 2004; Ouyang et al. 2009; Schmidt et al. 2017), or

HMs in each gene's regulatory region (Karlić et al. 2010; Cheng et al. 2011; Dong et al. 2012; McLeay et al. 2012; Singh et al. 2016; Read et al. 2019). Recently, deep neural networks have been used to predict the expression level of a gene from the DNA sequence flanking it (Kelley et al. 2018; Zhou et al. 2018; Washburn et al. 2019; Agarwal and Shendure 2020). All these models predict expression level in a given sample by using data from the same cell type and similar growth conditions. As result, the features used for prediction could be causes, consequences, or merely correlates of gene expression level (Henikoff and Shilatifard 2011). Models have also been trained to predict the variability of gene expression within or across cell types (Ouyang et al. 2009; Zhou et al. 2014; González et al. 2015; Crow et al. 2019; Sigalova et al. 2020). In addition to the above genomic features, combining the binding signals of RNA-binding proteins and microRNA at gene bodies with TFBS at promoters was also determined to be predictive (Tasaki et al. 2020).

We have taken on a different challenge – training machine learning models to predict which genes will respond to perturbation of a TF without using any data from perturbed cells. Because the predictive features are measured in unperturbed cells, they cannot be consequences of the perturbation or the response. The overall accuracy of the models serves as a benchmark for our understanding of global regulatory networks. Perhaps more important, analysis of the trained models can provide insight into the factors that determine which genes respond to a TF perturbation. Many methods have been developed to explain how specific features and feature values influence a complex model's predictions (Molnar 2019). One class of methods computes feature importance as the drop in prediction accuracy when the assignment of a feature's values to training examples is randomly permuted (Breiman 2001; Zeiler and Fergus 2012; Zhou and Troyanskaya 2015; Fisher et al. 2019). Another class focuses on explaining why the predictions

for individual examples differ from the mean prediction. In this study we rely on one such method – SHAP values (Lundberg and Lee 2017). SHAP values are based on how the prediction for a particular example is affected when the value of a feature is replaced by the value from another randomly selected example. A positive SHAP value for a particular feature of a particular example indicates that that feature pushes the model to predict a higher response for that example. Conversely, a negative SHAP value indicates that the feature value pushes the model to predict a lower response for that example. The magnitude of the SHAP value for a particular feature of a particular example indicates how influential the feature value is.

SHAP values are specific to one example because, in a non-linear model, the effects of a feature depend not only on its value but on the values of other features of the same example. However, several summary calculations make it possible to draw conclusions that apply to all examples or a specific subset of examples. Separately summing the positive and negative SHAP values for a feature over a set of examples reveals the relative strength of the positive and negative influences of the feature. This can be especially useful when looking at just the positive examples, just the negative examples, examples which are predicted accurately or inaccurately, etc. Summing the positive and negative SHAP values of a feature together provides a sense of the feature's overall direction of influence, which we refer to as the *net influence* of the feature. Summing the absolute values of the SHAP values for a set of examples shows how important the feature is in determining the model's predictions, regardless of direction. We refer to the sum of absolute SHAP values over all examples as *global feature importance*.

SHAP analysis, complemented by analyses of model accuracy, provides several surprising biological and methodological insights.

1. Existing genome-scale data on TF binding locations, including ENCODE data on human K562 cells, are not useful for predicting which genes will respond to perturbation of a TF. However, yeast data obtained by newer methods (transposon calling cards or ChIP-exo) are.
2. A few HMs have value for predicting perturbation responses, primarily when they occur in the gene body downstream of the transcription start site (TSS).
3. For both yeast and human, the preperturbation gene expression level and gene expression (GEX features) were surprisingly useful for predicting whether a gene would respond to perturbation of any TF or other regulatory protein; for human cells, they were far and away the most useful features. When these features are available, HMs provide no additional information that is useful for predicting perturbation responses in human K562 cells.
4. In summary, properties of the gene itself have a major influence on its tendency to respond to regulatory perturbations. The extent to which this tendency is determined by the gene's epigenetic state or its inherent properties remains to be seen.

4.2 Results

4.2.1 Modeling frameworks, features, and datasets

We took a two-step approach to understanding the determinants of transcriptional responses to TF perturbations: (1) train machine learning models to predict whether each gene will respond to a perturbation of a particular TF and (2) analyze the trained models to identify which genomic features they used to make their predictions. We provided the models with three types of genomic features (Fig. 4.1A). First, data on the binding locations of the perturbed TF (location features). Second, data on the median and variance of each gene's expression levels in

unperturbed samples (GEX features). Third, data on each gene's epigenomic context, including DNA accessibility, selected histone modifications, and dinucleotide frequencies (epigenetic features). We focused on eight histone marks that were previously shown to be most useful for predicting gene expression level (Karlić et al. 2010; Zhou et al. 2014; González et al. 2015; Singh et al. 2016; Roadmap Epigenomics Consortium et al. 2015) (Supplemental Table S1). Neither GEX features nor epigenetic features are tied to any specific TF – if they predict a gene's responsiveness to perturbation of one TF, they should also predict its responsiveness to perturbation of other TFs.

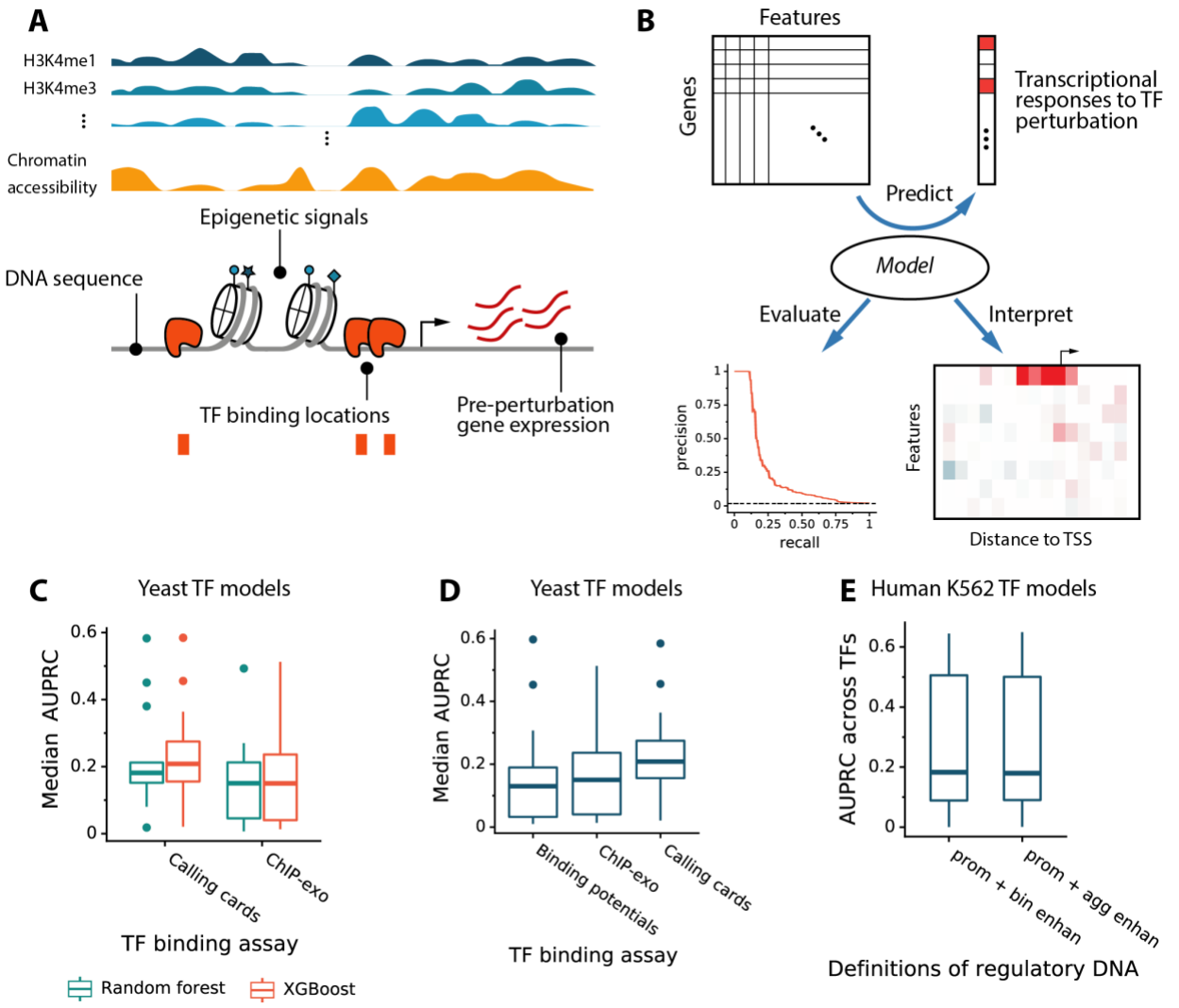


Figure 4. 1: Model and performance. (A) Features for predicting transcriptional responses to TF perturbation. (B) Framework for predicting responses, evaluating model performance, and estimating local feature influences. (C) Model accuracy on yeast TFs using all binding location data. (D) Model accuracy on eight yeast TFs with binding location data from both ChIP-exo and calling cards assays. (E) Model accuracy on human K562 cells using two methods of aggregating data from enhancers associated with each gene.

To generate a feature matrix, we defined cis-regulatory regions for each gene and mapped genomic data to them. For yeast genes, we assumed a regulatory region ranging from 1000 bp upstream of the transcription start site (TSS) to 500 bp downstream. Although most studies assume the yeast promoter is smaller than this, we expected that the models would learn which parts of this region are most predictive. For human genes, we included both proximal promoters

(4 kb centered on the 5'-end TSS) and distal enhancers (taken from (Fishilevich et al. 2017); see Methods). 47% of alternative TSS's fell within 4 kb region around the 5' TSS. The 5' TSS and others within 2 kb of it account for the vast majority of transcription (Supplemental Fig. S4.1). Alternative promoters outside of this region were treated as enhancers (Andersson and Sandelin 2020).To test whether certain locations within a regulatory region are more important than others, we divided the promoter regions into 100 bp subregions, each with its own features. We tried two methods of subdividing enhancer regions, as described below. Within each 100 bp subregion, signals from assays for TF binding location, DNA accessibility, or histone marks were aggregated and discretized. For yeast, we used TF binding location data generated by two *in vivo* assays: transposon calling cards (Wang et al. 2011a; Shively et al. 2019; Kang et al. 2020) and ChIP-exo (Bergenholm et al. 2018; Rossi et al. 2018a). We showed previously that these datasets predict perturbation responses much better than older ChIP-chip data (Kang et al. 2020). We used data on yeast histone marks from ref. (Weiner et al. 2015) and chromatin accessibility from ref. (Schep et al. 2015), both assayed in steady-state growth conditions. For human models, we used data from the K562 cell line because the it has the most TFs that were ChIPped and perturbed in the ENCODE Project (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020). Histone marks, DNA accessibility, and perturbation-response data were also from ENCODE (see Methods). For both yeast and human, preperturbation expression variance was adjusted to make it independent of expression level (Methods; Supplemental Fig. S4.2).

We trained the models to predict whether a gene will respond to a TF perturbation (Fig. 4.1B). For yeast, responsiveness was determined by using data from Hackett et al. (2020), who measured transcriptional responses shortly after chemically inducing overexpression of each TF. We focused on the responses at 15 minutes after the induction (see Methods). For human,

responsiveness was determined by using ENCODE RNA-Seq data measured after TF knockdown or knockout. Our datasets included 25 yeast TFs and 56 human TFs with both binding and perturbation-response data. The average number of genes that responded to each perturbation was ~6.7% in yeast (median: 2.7%, sd: 9.4%) and ~6.4% in human (median: 2.5%, sd: 8.3%).

We trained and tested two ensemble classifiers for each perturbed TF—random forests and a gradient boosting implementation called XGBoost (Chen and Guestrin 2016) – by using ten-fold cross-validation on genes. Below, we analyze how the features influence the prediction for each gene using the model that was not trained on that gene. We used precision recall curves for accuracy evaluation and the area under the curve (AUPRC) as a summary statistic. This approach is appropriate because only a small fraction of genes is responsive to each perturbation, creating large class imbalances.

First, we tested the two classifiers using yeast TF binding-location data from either transposon calling cards or ChIP-exo, keeping all other features constant. The best combination of classifier and binding data was XGBoost on calling cards data (Fig. 4.1C). However, calling cards and ChIP-exo data assayed different sets of TFs. To make a direct comparison, we trained and tested XGBoost models for the eight TFs assayed by both methods. The calling cards data again yielded greater prediction accuracy (Fig. 4.1D). We also tried replacing binding location data with TF binding potentials obtained by scanning a binding specificity model for the perturbed TF (Spivak and Stormo 2012; Grant et al. 2011) over promoter sequences. Binding potential was least useful, even when data on chromatin accessibility was also included in the model (Fig. 4.1D). Going forward, we use the XGBoost models. For yeast, we focused on TFs

that had either calling cards or ChIP-exo data; for those that had both, we used the data that yielded the best prediction accuracy.

Using XGBoost on the K562 ENCODE data, we investigated two ways of incorporating binding and epigenetic features from enhancers. The two methods divide the region around the promoter into subregions in different ways and sum the signals from enhancers within each subregion to form a single feature value. The first method (*bin enhan*) sums signals over enhancers within subregions whose widths increase exponentially with their distance from the TSS (Supplemental Fig. S4.3). The second approach (*agg enhan*) sums signals from all enhancers upstream of the TSS to create one feature and all enhancers downstream of the TSS to create another. Models trained using the two strategies of enhancer-feature mapping show no significant difference in accuracy ($P = 0.63$, paired t-test; Fig. 4.1E), so we used less numerous aggregated enhancer features in the remainder of the study.

The prediction accuracy varied quite a bit from one TF to another (Supplemental Fig. S4.4). In general, accuracy was lower for TFs that had few responsive targets than for those that had many (Supplemental Fig. S4.5A,B). This is likely the result of extreme class imbalance, which is known to hinder classification algorithms (Japkowicz and Stephen 2002) and the lack of enough positive examples to learn from. In the ENCODE data, another major factor was the effectiveness of the TF perturbation. The larger the absolute log fold change of the TF in the perturbed sample relative to the unperturbed, the better the TF model performed (Supplemental Fig. S4.5D). There was no such trend in the yeast data because the induced TFs were highly over-expressed (typically at least 16-fold, Supplemental Fig. S4.5C).

4.2.2 SHAP analysis shows that the TF binding signal is useful for prediction in yeast

The next step in our analysis was to determine what XGBoost learned about genomic features and how it used them. The model interpretation approach we used is based on SHAP values (Lundberg and Lee 2017; Lundberg et al. 2018). SHAP values explain why the prediction for one particular test example – one TF-gene pair – differs from the average prediction for all genes in response to perturbation of that TF. Of course, explaining what an algorithm learned is only interesting if it learned something significant, as indicated by its prediction accuracy. Thus, we analyzed only models with a median cross-validation AUPRC greater than 0.1, yielding models for 17 yeast TFs (Supplemental Table S2) and 30 human TFs (Supplemental Table S3).

Taking the XGBoost model for yeast TF Lys14 as an example, Figure 4.2A illustrated the SHAP values calculated for each feature of *LYS9*, a responsive gene, and *ECM23*, a non-responsive gene. The primary factors that caused the model to predict that *LYS9* would be responsive are (1) the Lys14 binding signal in the 500 bp upstream of the TSS and (2) *LYS9*'s pre-perturbation expression level (Fig. 4.2A, left, red). For *ECM23*, these positive influences were absent (Fig. 4.2A, right). Furthermore, *ECM23*'s pre-perturbation expression level and, to a lesser extent, its pre-perturbation expression variation, pushed the model to predict that it would not respond to Lys14 perturbation (blue). To aggregate these influences across promoter regions, we separately summed all positive SHAP values for each feature, which are plotted in red to the right of the heatmaps, and all negative SHAP values for each feature, which are plotted in blue. If a feature has a positive influence in some regions of the promoter but negative in others, it is shown with both a red bar (to the right of the centerline) and a blue bar (to the left of the

centerline). For *LYS14* and *ECM23*, most features have only positive or only negative SHAP values, so only one bar is visible.

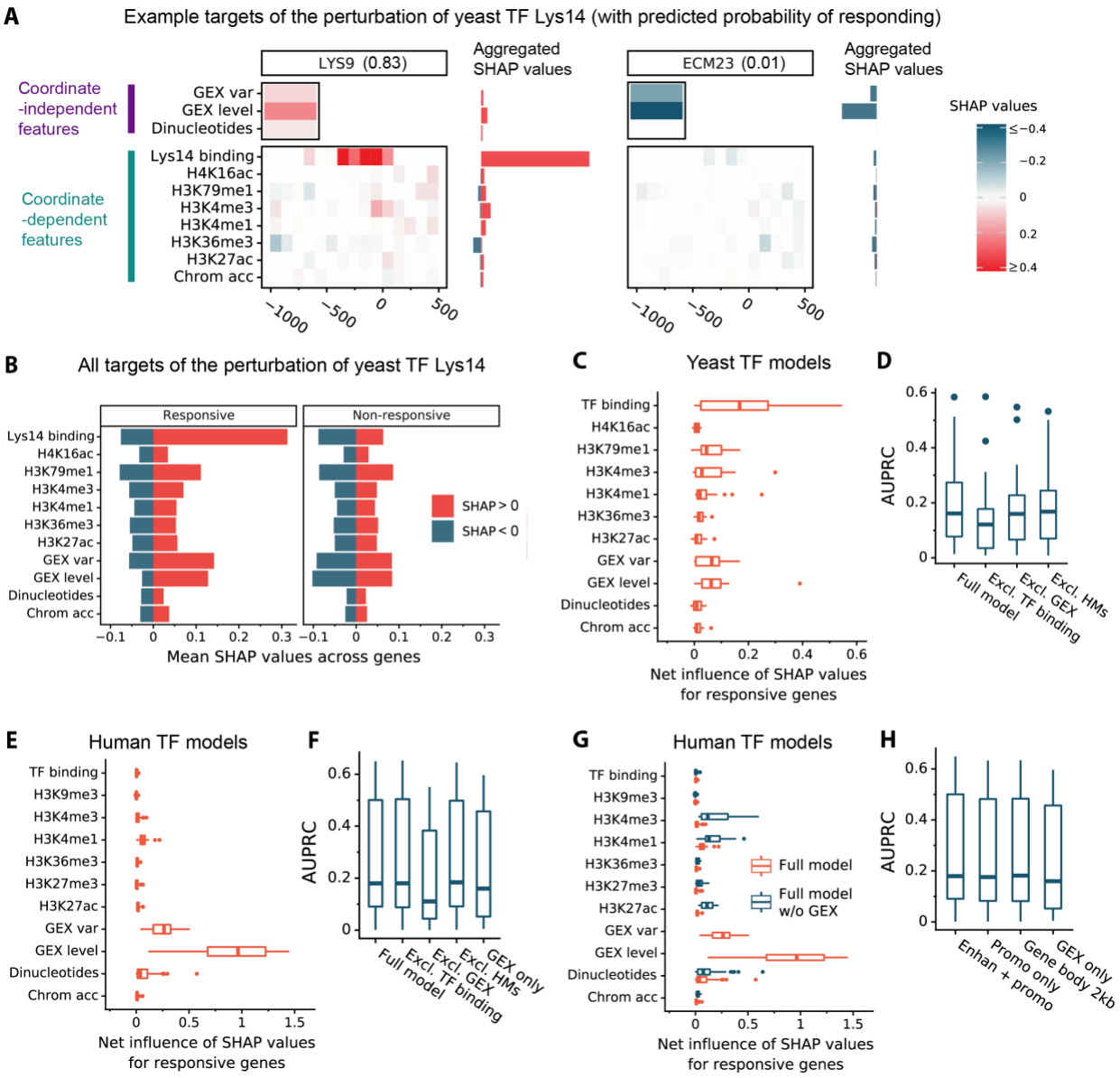


Figure 4. 2: Quantification of feature influences. (A) An example of decomposing the predicted score using SHAP values. Lys9 is a responsive target of yeast TF Lys14 with predicted response probability 0.83 and Ecm23 is an unresponsive gene with predicted response probability 0.01. The top panel shows the features that are independent of genomic coordinates; the bottom panel shows the features that depend on genomic coordinates. The right horizontal bars show the respective sums of SHAP values that are positive (red) and negative (blue), regardless of their genomic coordinates. (B) Left: For yeast TF Lys14, the positive (red) or negative (blue) SHAP values for each feature, summed over genomic positions relative to each gene and averaged over genes that respond to Lys14

perturbation. Right: The same analysis for genes that do not respond to Lys14 perturbation. (C) Distribution across TFs of the "net influence" of each feature on predictions, averaged over responsive targets. Net influence is the sum of all SHAP values for a feature, regardless of sign or genomic position. (D) Comparison of yeast model accuracy using four types of input features: the model described previously (*Full model*), the model trained without TF binding features, the model without gene expression features, and the model without histone marks (HMs). (E) Same as (C) except for human K562 TF perturbations. (F) Human model accuracy as in (D), with the addition of a model trained only on gene expression features (*GEX only*). (G) Comparison of net influences of features on predictions for responsive human genes. Models were trained with gene expression features (*Full model*) or without. (H) Comparison of model accuracy using four types of input features: the model described previously (*Full model*), the model excluding enhancer features (*Promo only*), the model excluding enhancer features and features mapped upstream of the TSS (*Gene body 2Kb*), and the model using only pre-perturbation gene expression features (*GEX only*).

To get a sense of how feature values affected the model's predictions for all genes, we first divided genes into responsive and non-responsive. Within each group, for each feature, we separately summed its positive SHAP values from all promoter regions of all genes and its negative SHAP values (Fig. 4.2B). For Lys14-responsive genes (Fig. 4.2B, left), Lys14 binding data in the gene's promoter tends to have a much bigger effect on predictions when it pushes the predicted probability of response up (red bar) than when it pushes the predicted probability of response down (blue bar). Comparing the red and blue bars for other features reveals net positive effects from pre-perturbation gene expression level and variation. Histone marks H3K79me1 and H3K4me3 have smaller positive influences and they have negative influences that are almost as large as the positive ones, on average. Thus, depending on the promoter position, the value of the histone mark feature, and the gene, these features can either increase the predicted probability of a gene's responding or decrease it. For genes that do not respond to the Lys14 perturbation, the net influences of all features are close to zero (Fig. 4.2B, right), indicating that they do not push predictions for non-responsive genes very far from the average prediction for all genes. This

average is low – 6.8% probability of being responsive – since the vast majority of genes do not respond to perturbation of Lys14.

To generalized from Lys14 to all TFs, we calculated the net influences of features on predictions for genes that respond to perturbation of each TF and plotted the distributions (Fig. 4.2C). This showed that the findings for Lys14 generalize well to the other TFs. The biggest net influence was the binding signal from the perturbed TF, followed by gene expression level, gene expression variation, and histone marks H3K79me1 and H3K4me3. Supplemental Figure S4.6 shows the positive and negative influences of each feature on both responsive and non-responsive genes. Complementary analysis of the effects of dropping feature classes from the model confirmed that TF binding features contribute most to the accuracy of the full model, followed by gene expression features (Fig. 4.2D). Dropping histone marks had a marginally significant but very small effect (the mean AUC dropped 0.01, $P<0.04$). This was due to an effect on a minority of TFs, since the median AUC actually *increased* by 0.006.

4.2.3 In human cells, ChIP-seq peaks and epigenetic marks have relatively little value for response prediction

Next, we summarized SHAP values for each human TF model, focusing first on genes that respond to perturbation of the TF. Strikingly, ChIP-seq peaks for a TF, which reflect its binding location, had essentially no net influence on predictions for genes that are in fact responsive (Fig. 4.2E). This is consistent with earlier studies (Gitter et al. 2009; Lenstra and Holstege 2012; Cusanovich et al. 2014; Kang et al. 2020). Gene expression level in unperturbed control samples was the most influential factor, followed by expression variation in the control samples. H3K4me1 and dinucleotide frequencies in the cis-regulatory DNA had very small influences on the predictions for some TF models, but the effects of the other histone marks and

of chromatin accessibility were negligeable. Analysis of non-responsive genes yielded similar conclusions (Supplemental Fig. S4.6B). The impacts of these features on predictive accuracy supported these conclusions: Dropping the ChIP-Seq or the HM features had negligible impact on prediction accuracy, whereas dropping the gene expression features greatly reduced accuracy (Fig. 4.2F). In fact, a model using only the gene expression features was almost as accurate as the full model (median AUPRC dropped by 0.001, $P<0.001$).

We hypothesized that the lack influence of histone marks in the model might be due to the fact that gene expression features summarize any useful information provided by histone marks as well as other aspects of a gene's epigenetic state, rendering the information from histone marks redundant with and less useful than gene expression information. To test this, we trained a model without the gene expression features and analyzed the influence of the remaining features on predictions for genes that are in fact responsive (Fig. 4.2G). Removing the gene expression features from the model did increase the influence of H3K4me3 and H3K4me1, supporting our hypothesis. However, the model without gene expression features has low accuracy (median AUPRC 0.11), so the predictive value of HMs is very small.

These findings drove us to investigate the utility of features mapped to various regions of the cis-regulatory DNA associated with each gene. When we dropped the TF binding signal, histone modifications, dinucleotide frequencies, and chromatin accessibility from the enhancer regions associated with each gene, the effect on prediction accuracy was negligible (median AUPRC decreased by 0.004, $P<0.001$ Fig. 4.2E). We then tried dropping all features from both the enhancers and the promoter regions upstream of the TSS, leaving only the first 2 Kb of the gene body. Again, the effect on accuracy was negligible (median *AUPRC increased* by .002 relative to the full model while the mean decreased by 0.007). Finally, we tried dropping these

features altogether, leaving only the gene expression level and expression variation in control samples. The effect of dropping these features entirely was small, compared to the full model (median AUPRC decreased by 0.035, or 14% of the full model's AUPRC). While the TF binding signal and epigenetic features significantly enhanced prediction accuracy in yeast, they had little predictive value in the human data. What predictive value they did have was entirely due to features mapped to the 2 Kb downstream of the TSS.

4.2.4 In yeast cells, TF binding locations and strengths discriminate between bound genes that are responsive and those that are not

The most common use of *in vivo* binding location data is to classify genes into those whose regulatory DNA is or is not bound by the TF. However, this typically yields a large set of genes that are bound by the TF at a statistically significant level but are not responsive to perturbation of that TF (Gitter et al. 2009; Lenstra and Holstege 2012; Cusanovich et al. 2014; Kang et al. 2020). Thus, we investigated whether the model could use the strength and location of the binding signal to better predict which bound genes would be responsive. In Figure 4.3A, each row shows SHAP values of the TF binding signal in each promoter bin, averaged across the genes that were significantly bound by the perturbed TF. For all but three TFs, the binding signal in the 600 bp upstream of a gene's TSS influenced the model toward predicting (correctly) that the gene would respond to the perturbation. The Calling Cards and ChIP-exo technologies showed a general concordance on the relative utilities of various positions, but the influence of ChIP-exo was even more localized to 300 bp upstream (Supplemental Fig. S4.7). Using Leu3 calling cards data as a typical example, stronger binding signals were more influential than weaker ones and signals of the same strength were more influential in the region 100-200 bp upstream of the TSS than in the region 400-500 bp upstream (Fig. 4.3B). For Leu3, SHAP values

in five promoter bins were significantly higher among the bound and responsive group than in the bound but unresponsive group (Fig. 4.3C). For most TFs, a similar pattern was found in one or more promoter bins (Fig. 4.3D, Supplemental Fig. S4.8). Thus, the strength and location of the binding signal are meaningful predictors of whether significantly bound genes will respond to the perturbation, consistent with our earlier findings (Kang et al. 2020).

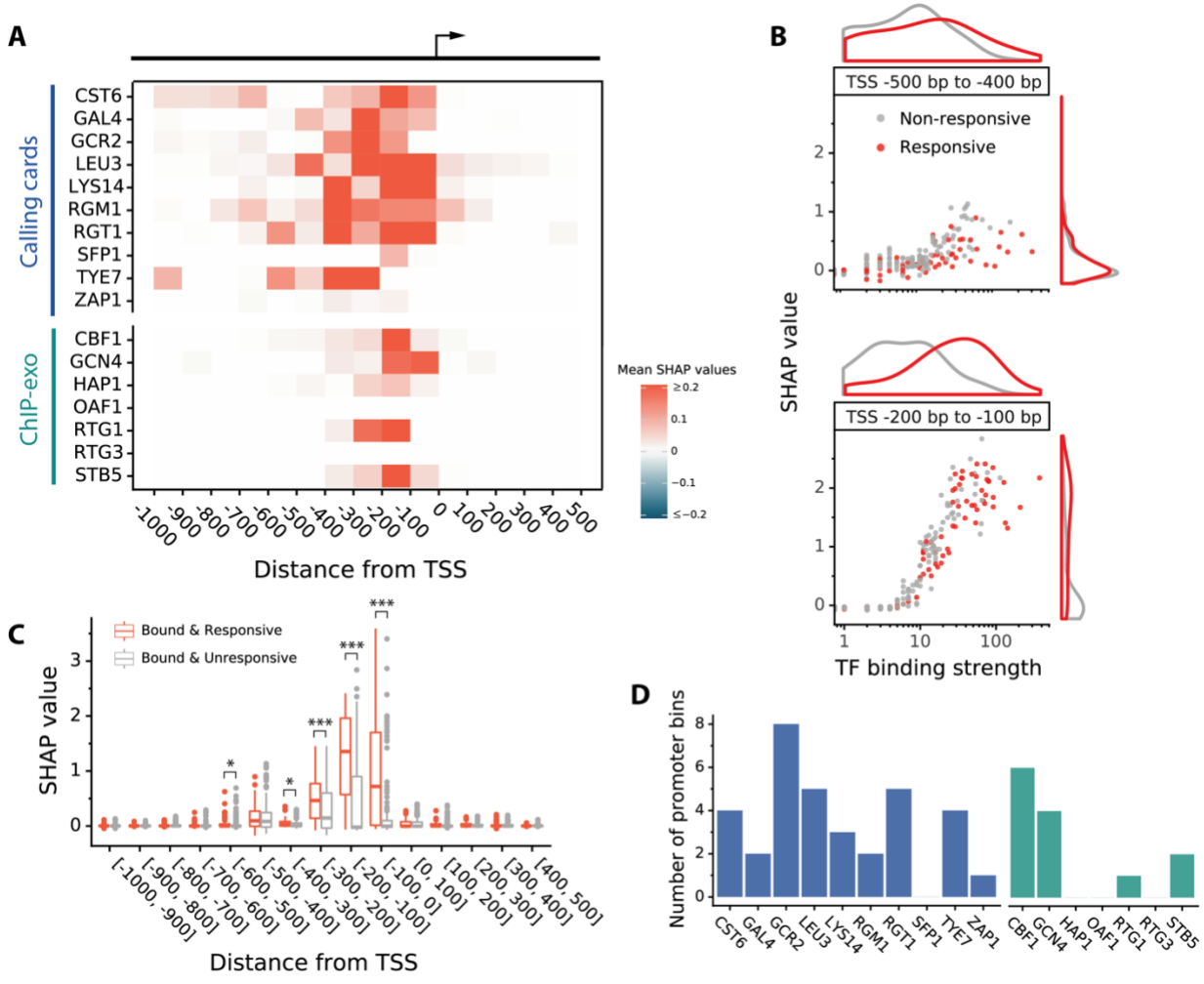


Figure 4. 3: TF binding features in yeast models. (A) Heatmap of the influence of yeast TF binding signals along regulatory DNA. Each pixel is the mean SHAP value over all target genes that were bound by the perturbed TFs. (B) Comparison of two upstream bins ([-500, -400] and [-200, -100]) of yeast TF Leu3. Among the genes that are bound by Leu3, the responsive genes are more clearly distinguished from the unresponsive one in the [-200, -100] bin. This shows that even within 500 bp of the TSS, Leu3 binding near the TSS is more likely to be functional than Leu3 binding further away. (C) Comparison of feature influences on responsive and unresponsive targets that were bound

by Leu3. Statistical significance used Wilcoxon rank-sum test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$(***). The significant differences all show that responsive genes are bound more strongly than unresponsive genes. Furthermore, all significant effects of binding strength are within 600 bp upstream of the TSS. (D) Number of promoter bins in which the bound and responsive genes have significantly higher SHAP values for TF binding ($p < 0.05$) than the bound but non-responsive genes. Blue bars: calling cards; Green bars: ChIP-exo data.

4.2.5 Highly expressed genes and genes with high expression variation are more likely to be responsive

Given the predictive power of gene expression level and variation, we investigated how the model used these features. Starting with yeast TF Lys14, we noted a monotonic relationship in which the more highly a gene was expressed before the perturbation, the more the model expected it to respond (Fig. 4.4A). We also noted that, the more a gene's expression varied from one pre-perturbation sample to another, the more the model expected it to respond (Fig. 4.4B). This this was not due to the relationship between expression level and expression variation, which we removed by fitting a model that predicts expression variation from expression level and using the residuals from that model as our variation feature (Supplemental Fig. S4.1). For most yeast TFs, both expression level and expression variation are positively correlated with SHAP value – higher expression level and expression variation push the model to predict a higher probability of response (Fig. 4.4C). The same pattern holds for human TFs (Fig. 4.4D). The two yeast TFs (Rgt1 and Zap1) and one human TF (ZC3H8) that have negative correlations were those for with the lowest model accuracies, which suggests that either (1) the data on these TFs are not particularly accurate or (2) the model's feature utilization may not reflect the true patterns in the data.

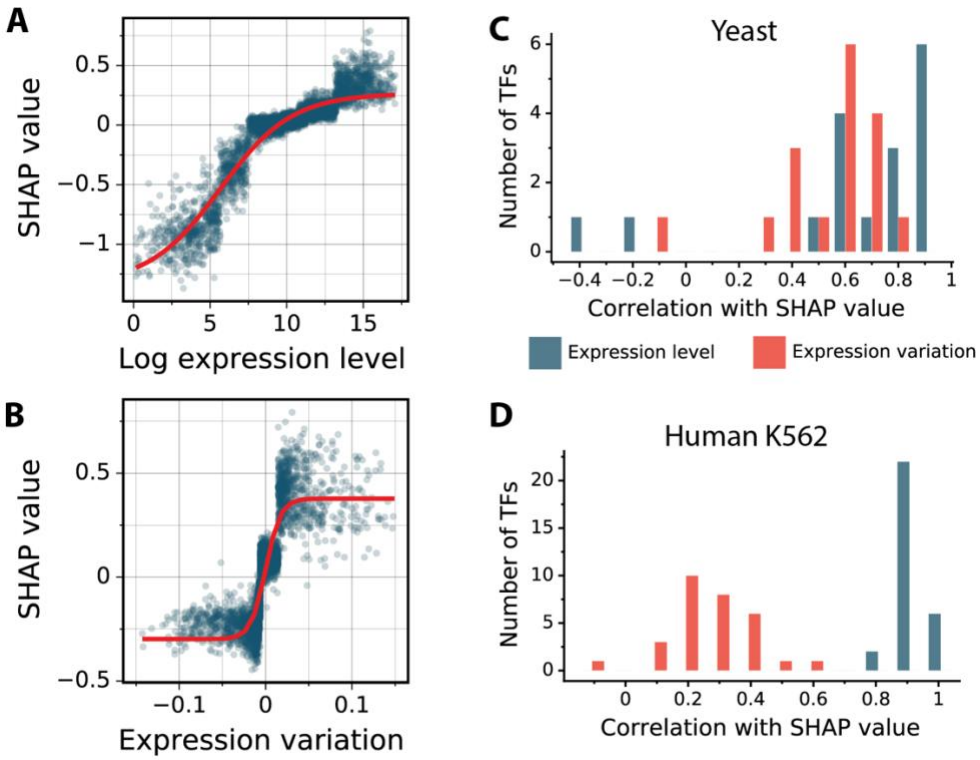


Figure 4. 4: Gene-specific features. (A) Relationship between feature input and SHAP values of gene expression level for Lys14 model. Red curve is a fitted sigmoid function. The model predicts that more highly expressed genes are more likely to be responsive to Lys14 perturbation. (B) Relationship between feature input and SHAP values of gene expression variation for the same model. The model predicts that genes whose expression levels are more variable after correction for their expression level are more likely to be responsive to Lys14 perturbation. (C) The distribution of the correlations of input and SHAP values for the two expression-related features in yeast cells. For most TFs, both expression level and expression variation are positively correlated with response to a perturbation. On average, expression level is more positively correlated than expression variation. All correlations are statistically significant with the largest P-value < E-18. (D) Same as (C) except for human K562 cells. All correlations are statistically significant.

4.2.6 Histone marks downstream of the TSS are more predictive of responsiveness than upstream histone marks

We showed above that for human TFs, models trained using coordinate-dependent features in enhancers, the 2Kb upstream of the 5'-end TSS, and the 2Kb downstream of the 5' TSS were no more accurate than those that used only the downstream features (Fig. 4.2H). For

both yeast and human, the downstream histone marks had a much greater influence on the predictions than the upstream marks (Fig. 4.5A). This was quantified for each TF by the mean absolute SHAP values across all genes. Among the six histone marks we analyzed in yeast, downstream H3K79me1 had the biggest influence on predictions, followed by downstream H3K4me3 and downstream H3K4me1. For these three marks, the differences between their influence when they occur downstream of the TSS compared to upstream of the TSS are statistically significant (Fig. 4.5A). This echoes the previous report that H3K79me1 and H3K4me3 are predictive of gene expression for genes whose promoters have low CpG content (Karlić et al. 2010). In human cells we did not have data on H3K79me1, but downstream H3K4me3 and H3K4me1 are the two most influential marks, followed by downstream H3K27ac (Fig. 4.5A, right). The differences between the influences of these marks when they occur downstream of the TSS compared to upstream are statistically significant.

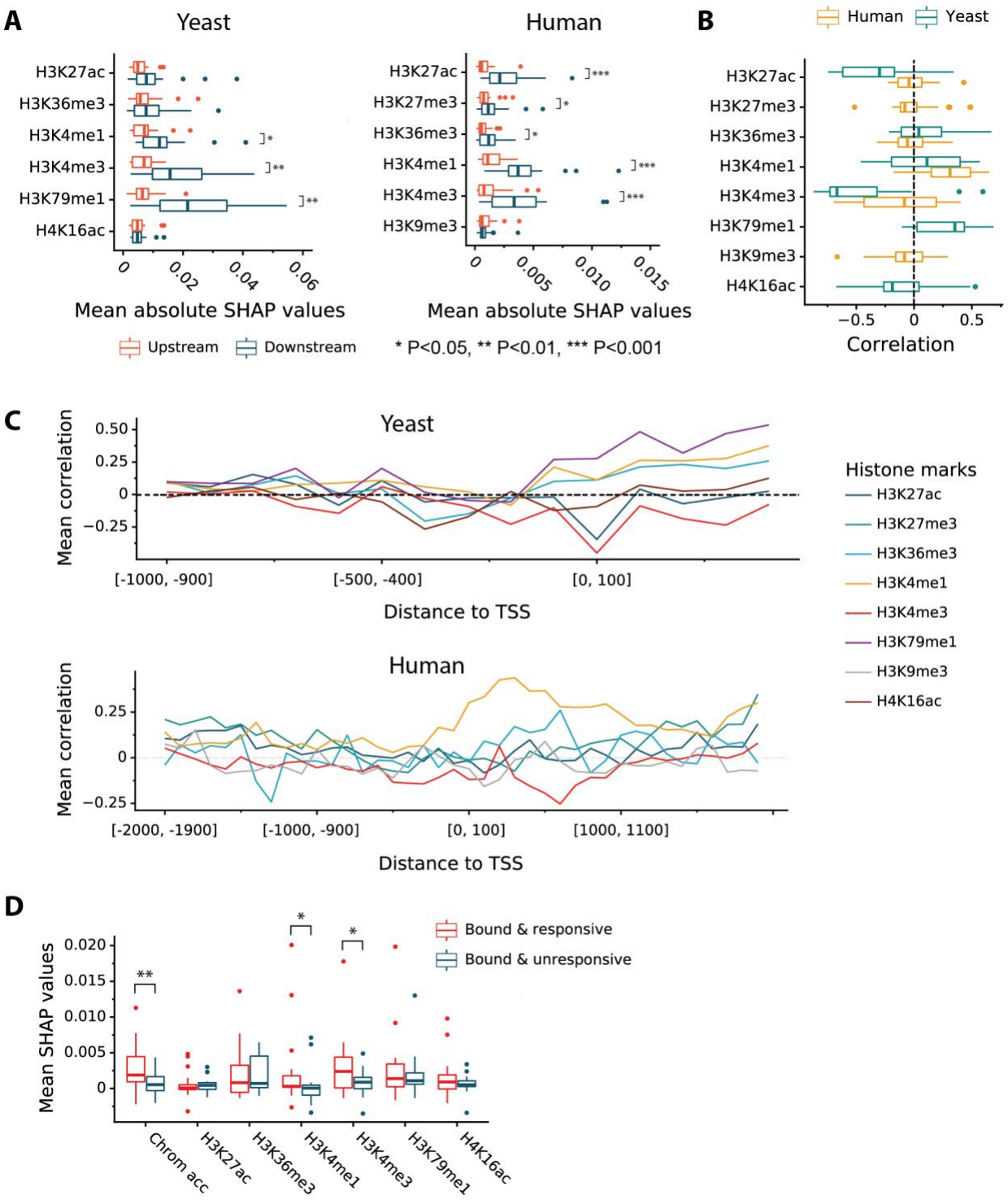


Figure 4. 5: Epigenetic features. (A) Comparison of the global importance of histone mark features in the regions upstream or downstream of the TSS. The distribution across TFs is shown. For each TF, the global importance of each feature is the absolute SHAP value for that feature averaged across all genes and all bins upstream or downstream of the TSS. Marks showing a significant difference are more influential when they occur downstream of

the TSS than when they occur upstream (* $P<0.05$, ** $P<0.01$, *** $P<0.001$, Wilcoxon signed-rank test). (B) Correlation of histone modification signals and their corresponding SHAP values at bin [0, 100] (first downstream bin from TSS). For example, both the yeast and human models predict that genes with stronger H3K4me1 signal immediately downstream of the TSS are more likely to be responsive to TF perturbations. Genes with stronger H3K4me3 signal, by contrast, are less likely to be responsive. (C) Average input-SHAP correlations along the genomic coordinates in yeast and human cells. In both organisms, H3K4me1 was the strongest and most consistent predictor of responsiveness while H3K4me3 was the strongest predictor of unresponsiveness. (D) Distributions across yeast TFs of the average signed influences of features located upstream of the TSS, averaged across TF-bound genes that were either responsive or non-responsive. Asterisks indicate significant difference between bound targets that are perturbation responsive and those that are non-responsive.

Focusing in on the direction of influence in the 100 bp downstream of the TSS, H3K4me3 signal was negatively correlated with SHAP value for most yeast TFs, indicating that the presence of this mark pushes the model to predict that a lower probability of response (Fig. 4.5B). H3K27ac was also negatively correlated with SHAP value for most yeast TFs, while H3K79me1 was positively correlated. For H3K4me3 and H3K4me1, the sign of correlation was generally consistent between yeast and human, but some TFs are exceptions to this generalization. Looking at different positions relative to the TSS in yeast (Fig. 4.5C, top), we see that the influences of H3K4me3 and H3K27ac presence are most consistently negative when they occur immediately downstream of the TSS, whereas the influences of H3K79me1, H3K4me1, and H3K36me3 presence downstream of the TSS are consistently positive. In human, the influences of H3K4m1, H3K4me3, and H3K36m3 peak slightly downstream of the TSS (Fig. 4.5C, bottom). However, one should not read too much into this, as the inclusion of these features in the model has a very small effect on prediction accuracy.

Next, we focused on yeast histone marks upstream of the TSS and asked whether the model picked up differences between the TF-bound genes that respond to perturbation of the TF and the TF-bound genes that do not respond (Fig. 4.5D). The presence of either H3K4me1 or H3K4me3 upstream of the TSS influences the model to correctly predict a higher probability of

response among genes that are responsive than among those that are not. Thus, the model is able to take advantage of these features in discriminating between bound-responsive genes and bound-unresponsive genes. The same is true of chromatin accessibility upstream of the TSS. However, these influences are orders of magnitude lower than the influences of the TF binding signal (Fig. 4.3C) or the gene expression features (Fig. 4.4A,C).

4.2.7 Responses to any genetic perturbation are partially explained by TF-independent factors

Above, we reported that TF binding signals at cis-regulatory regions have little value for modeling responses to TF perturbations in ENCODE data on human K562 cells. This drove us to investigate whether the features that are independent of any particular TF can predict a gene's predisposition to respond to perturbations of TFs or other regulators. We therefore calculated the response frequency of each gene – the number of perturbations to which each gene responds divided by the total number of perturbations. Next, we trained an XGBoost regression model to predict each gene's responsive frequency using only the TF-independent features and it by 10-fold cross-validation on genes. The median variance explained is 45% for yeast and 37% for human (Fig. 4.6A). Since training on this task does not require DNA binding location data, we also tried it on a larger set of regulator perturbations in K562 cells, including TFs for which no binding data is available and regulators that are not DNA-binding proteins. This model was even better, explaining 56% of variance in the frequency of response to regulator perturbations. These results clearly demonstrate that some genes are poised to respond to perturbations and others are resistant. In this task, H3K79me1 and H3K4me3 are the most influential features in yeast cells – more influential even than gene expression level and variation (Fig. 4.6B, left). In human cells, gene expression level and variation are by far the most influential features (Fig. 4.6B, middle,

right). These features are likely read-outs of some as-yet-unidentified molecular feature that makes genes sensitive to perturbations.

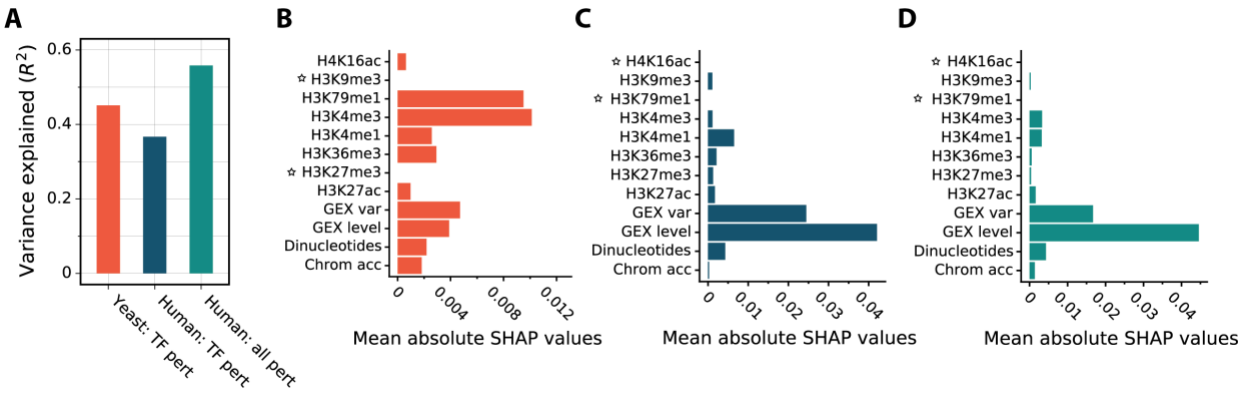


Figure 4. 6: TF-independent prediction of each gene's tendency to respond to genetic perturbations. (A) The variance explained for predicting DE frequency in yeast TF perturbations (n=194), human TF perturbations for which binding data are available (n=56), and all ENCODE genetic perturbations of K562 cells (n=355). Bar height indicates the median, across genes, of the variance explained (R^2) based on cross validation using held-out genes. (B) The mean absolute influence of each TF-independent feature across genes. Asterisks indicate missing data.

4.3 Discussion

Determining which genes are regulated by each TF in an organism is a fundamental goal of regulatory systems biology. Furthermore, the ability to predict which genes will respond to perturbation of a TF serves as a benchmark of how well we understand the TF network. Note that there is a body of work focused on predicting gene expression level (Middendorf et al. 2004; Ouyang et al. 2009; Schmidt et al. 2017; Tasaki et al. 2020; Karlić et al. 2010; Cheng et al. 2011; Dong et al. 2012; McLeay et al. 2012; Singh et al. 2016; Read et al. 2019; Kelley et al. 2018; Zhou et al. 2018; Washburn et al. 2019; Agarwal and Shendure 2020; Zhou et al. 2014; González et al. 2015; Crow et al. 2019; Sigalova et al. 2020), but this is a very different task from predicting the response of expression level to perturbations by using only data from unperturbed cells. Data on where in the genome each TF binds was expected to be of great value in

determining its targets, but multiple studies have shown that, in the available large ChIP-chip and ChIP-seq datasets, the genes in whose regulatory DNA a TF binds do not correspond well to those that respond to perturbation of the TF (Gitter et al. 2009; Lenstra and Holstege 2012; Cusanovich et al. 2014; Kang et al. 2020). We followed up on these observations by training machine learning models to predict which genes would respond to perturbation of a TF, given both data on a TF's binding locations and several features reflecting the gene's epigenetic context. We found that data on yeast TF binding locations obtained by the calling cards (Wang et al. 2011a; Shively et al. 2019; Kang et al. 2020) method and, to a lesser extent, the ChIP-exo method (Bergenholm et al. 2018; Holland et al. 2019), are useful for predicting which genes will respond to a perturbation of the TF. In fact, binding location was the most influential and valuable among the features we provided (Fig. 4.2A-D). Since earlier ChIP-chip data on yeast are known to correspond poorly to perturbation response, we conclude that the newer technologies are yielding better results. Binding signals influenced predictions mainly in the 500 bp upstream of the TSS, suggesting that this is the extent of functional yeast promoter regions (Fig. 4.3A-B). Even among genes with significant binding signal for a TF in their promoter, the strength and location of the signal helped to differentiate between functional and non-functional binding (Fig. 4.3C-D). In ENCODE data on human K562 cells, however, the situation was strikingly different. The models did not identify any patterns in ChIP-seq data that were useful for predicting which genes would respond to perturbation of the TF (Fig. 4.2E-F).

We also investigated the predictive value of a selection of histone marks and chromatin accessibility features. In yeast, these features had predictive value that was less than that of TF binding locations, but it was not negligible (Fig. 4.2A-D). In human, however, these features were not useful in models that included GEX features; HM features had a larger (though still

small) impact in models that did not include GEX features, which are discussed below. Among HMs for which we had data in both yeast and human, H3K4me1 and H3K4me3 were most influential. H3K4me1 increased the predicted probability of a gene's response to perturbations while H3K4me3 decreased it (Fig. 4.5). Both marks were most influential when they occurred downstream of the TSS, in the gene body. In fact, dropping all features that mapped to the enhancers and the promoter region upstream of the TSS had only a small impact on predictive accuracy in K562 cells. This surprising observation likely reflects both the low utility of the existing ChIP-seq data and incomplete knowledge of the enhancer locations and enhancer-gene associations. Future datasets on TF binding locations and enhancer-gene associations will likely reveal at least some predictive power for enhancer features.

The preperturbation gene expression and gene expression variation (GEX features) were surprisingly good predictors of which genes would respond to a perturbation. In fact, a model using only these two features predicted responses in K562 cells almost as well as the full model, which includes ChIP-seq, histone marks, chromatin accessibility, and dinucleotide frequencies (Fig. 4.2F). Genes that were expressed at higher level and genes that showed more variability in their expression level were more likely to respond to perturbations (Fig. 4.4). We hypothesize that these features are readouts of many molecular features of the genes sequence context and / or epigenetic state which have limited predictive power individually but much greater predictive power when aggregated by their effects on gene expression level and variation. This hypothesis is supported by the observation that the influence of several histone marks increases when GEX features are omitted from the model, though these influences are still small compared to the impact of GEX features when they are provided (Fig. 4.2G). However, the epigenetic state that is reflected in the GEX features is not as simple open chromatin versus closed chromatin, since

chromatin accessibility features have little influence even in the absence of GEX features. Identifying the molecular states and sequence elements that are reflected by in GEX features and showing that they can predict perturbation response is an important direction for future research.

The predictive power of features that are independent of the TF perturbed -- GEX features, HMs, chromatin accessibility, and dinucleotide sequence -- shows that some genes are poised to respond to perturbations while others are not. To confirm this, we trained models to predict how many TF perturbations a gene would respond to using only these features of the target gene (not TF binding locations). This model proved quite accurate. The yeast model relied most on histone marks, followed by GEX features, whereas the human model relied almost exclusively on GEX features. This was a significant finding -- if you want to predict which genes will respond to perturbation of a TF, it is not enough to know where the TF binds -- you must also know whether the gene is predisposed to respond to perturbations. An important direction for future research is to discover the extent to which a gene's predisposition to respond to perturbations is an epigenetic feature that depends on cell type and conditions or an inherent feature of the gene.

Our findings raise many questions. Neither the yeast nor the human model is able to predict which genes will respond to a TF perturbation reliably, even when provided with the binding locations of the TF and a host of epigenetic features. This is not simply a consequence of discretizing perturbation responsiveness, as models trained to predict the quantitative response yielded similar results (not shown). For human cells, the inability to predict responsiveness may reflect limitations of the technologies used for measuring TF binding locations and perturbing TFs as well as limited knowledge of enhancer locations and enhancer-gene associations. It may also be possible to improve on the way enhancer-associated features were coded for the model,

which could lead to better utilization of histone marks and chromatin accessibility for determining enhancer activity in a given sample of cells. Other types of data, such as levels of enhancer-associated transcription may also help. For yeast, however, these explanations are less applicable. Obtaining the right genomic data and developing the right models for predicting which yeast genes will respond to perturbation of a TF is a major challenge. Progress in overcoming this challenge will serve as a benchmark of our understanding of regulatory systems biology.

4.4 Methods

4.4.1 Data preparation

TF-perturbation response data

We used two large collections of TF perturbations followed with gene expression profiling in yeast and human K562 cells respectively. For yeast data, we downloaded the microarray data for transcriptional responses to each of the 194 induced yeast TFs using the ZEV TF-induction system (Hackett et al. 2020). Specially, column *log2_shrunken_timecourses* from the file "Raw & processed gene expression data" at https://idea.research.calicolabs.com/data was used as the levels of responses. A gene associated with a non-zero value was defined as responsive. The response profiles measured at 15 minutes after TF inductions were used to create labels for the corresponding TF models.

For human data, we used all RNA-seq expression profiles measured after gene knockout (KO) or knockdown (KD) in K562 cells from the ENCODE Project database (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020). These TFKO and TFKD mechanisms include CRISPR TF-disablement, CRISPR inference (CRISPRi), small-interfering RNA (siRNA), and small-hairpin RNA (shRNA). We downloaded the expected counts of experimental and control

profiles that were estimated using RSEM in the ENCODE RNA-seq pipeline and genome assembly GENCODE V24 (GRCh38). For each of the 355 experiments, we ran DESeq2 (V1.10.1) (Love et al. 2014) to identify differentially expressed genes by comparing the experimental replicates to the corresponding control replicates. A gene that has a Benjamini-Hochberg adjusted P-value < 0.05 and log2 fold-change > 0.5 was defined as responsive.

Pre-perturbation gene expression data

The gene expression features—pre-perturbation expression level and expression variations—were derived from the above gene expression datasets. One feature is the median of gene expression level across all samples measured prior to the TF perturbation. The other feature is the coefficient of variation (COV) of gene expression levels in these pre-perturbation samples. However, there exists dependency between expression level and expression variation, where COV decreases as expression level increases (Supplemental Fig. S4.2: left panels). To make the two features independent, we performed a correction procedure described in ref. (Sigalova et al. 2020). First, a smooth curve for COV was fitted as a function of the median expression level, using locally estimated scatterplot smoothing (LOESS) regression (Python scikit-misc V0.1.3). Second, for each gene, the residual of LOESS based on the gene's median expression level was calculated to represent the corrected COV, namely pre-perturbation expression variation (Supplemental Fig. S4.2: right panels). Regarding the specific input for calculating the two expression features, for each yeast gene, we took its log fluorescence levels of red (experimental) channel measured at 0 minute (before each of the TF inductions) as the pre-perturbation gene expression levels. For each human gene, we took its log TPM levels among all replicates of control samples as the pre-perturbation gene expression levels.

TF binding location data

The genome-wide binding location measures of yeast TFs were obtained using transposon calling cards (Wang et al. 2011a; Shively et al. 2019; Kang et al. 2020) and ChIP-exo (Bergenholm et al. 2018; Holland et al. 2019). To consistently map coordinate-dependent features, we used genome assembly sacCer3 for yeast and GRCh38 for human throughout this study. For the calling cards data that are available for 16 yeast TFs, we lifted over the transposon insertion coordinates, which were originally mapped based on sacCer2, to sacCer3 using the LiftOver tool in UCSC genome browser. No peak calling was used to further process the binding signals.

As regards ChIP-exo data, we obtained the peaks for TF binding sites (TFBSs) of 20 yeast TFs from the authors of Bergenholm et al. (2018) and Holland et al. (2019). Kang et al. (2020) reported that among the four environmental conditions, the bound targets of these TFs in glucose limited chemostat condition have the best agreement with the responsive targets at 15 minutes after TF inductions. We therefore only focused on the binding data in glucose limited condition. Furthermore, as the binding locations were reported in an alternative strain CEN.PK, we lifted over the TFBSs to assembly sacCer3 for strain S288C as follows. First, since the loci of TFBSs were reported as relative distances to CEN.PK TSSs, these loci were converted to the relative distances to the ORFs using the CEN.PK TSS annotation (https://github.com/SysBioChalmers/ChIPexo_Pipeline/blob/master/Data/TSSData.tsv). Due to high similarity of the two yeast strains, we assumed that the relative distance of each TFBS to CEN.PK ORF are the same for the matching S288C ORF. Next, the relative distances were converted to absolute genomic coordinates in sacCer3 using S288C gene annotation from the *Saccharomyces* Genome Database (SGD).

Turning now to human TFBS data, we downloaded the ChIP-seq peaks for 54 TFs in human K562 cells from the ENCODE Project (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020). K562 was by far the cell type that have the largest number of TFs that were both ChIPped and perturbed. These TFs were also restricted to be the well-defined DNA-binding factors from ref. (Lambert et al. 2018). Furthermore, we specifically used the "conversative" peaks, which underwent the Irreproducible Discovery Rate (IDR) correction using biological replicates at 2% IDR threshold. The log10 q-value reported for each peak was considered the binding strength of the TFBS.

Histone modifications and chromatin accessibility data

We used the coverage data for histone modifications (Weiner et al. 2015) and chromatin accessibility (Schep et al. 2015) in yeast cells. Specifically, the histone modifications data were measured in timepoint 0 minute before a diamide stress response using MNase-ChIP-Seq. Our choice of yeast histone marks includes H3K27ac, H3K36me3, H3K4me1, H3K4me3, H3K79me1, and H4K16ac. The chromatin accessibility data were measured at 0 minute before an osmotic response using ATAC-seq. We downloaded the coverage data in bigWig files under accession GSE61888 for histone marks and GSE66386 for chromatin accessibility from NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/). The coverage of each feature was mapped to yeast genome assembly sacCer3.

For human K562 cells, we downloaded the coverage data (fold change over control) in bigWig format for histone modifications and chromatin accessibility from ENCODE (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020). Our choice of human K562 histone marks includes H3K27ac, H2K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3. The coverage of each feature was mapped to human genome assembly GRCh38.

Mapping genome-wide features to cis-regulatory regions

For the yeast genome, the promoter of a gene is a fixed interval ranging from 1,000 bp upstream to 500 bp downstream from the transcription start site (TSS). The genomic coordinate of the TSS for each yeast gene were obtained from de Boer et al. (2020). The genome assembly is R64 (sacCer3). Next, each promoter region was split into 15 equal-sized bins (100 bp in width). The inputs of each genome-wide feature that were mapped to these promoter bins based on the relative position to TSS were then summed into a single value to represent a quantized input level for a certain feature in a certain bin.

For the human genome, we define three types of cis-regulatory regions for each gene. (1) *5' promoter*: a 4 Kb region centered at the 5'-end TSS of each gene (2 Kb on either side). The TSS coordinates were downloaded from Ensembl Release 92 (Cunningham et al. 2019). (2) *Alternative promoter*: a 4 Kb region centered at each TSS that is more than 2 Kb from the 5'-end TSS (if there exits alternative TSS for a particular gene). (3) *Enhancer*: a distal locus that is linked to target gene(s), which were compiled in the GeneHancer V4.8 database (Fishilevich et al. 2017). We require a legible enhancer-target link to have the "double elite" status, meaning that the identification of each enhancer must be supported by at least two distinct types of evidence, and the link to target must also be supported by more than one evidence. Enhancers that are more than 500 Kb away from the 5'-end TSSs of the linked genes were removed.

Here, we considered several strategies to create genomic bins that cover a gene's regulatory regions. Equal number of bins across genes is required to guarantee a rectangular feature matrix without missing values. To begin with, we primarily focused on the promoter centered at the 5'-end TSS of each gene for two reasons. First, approximately half of all TSSs (excluding the 5'-end TSSs) fall within the 2 Kb region downstream from the 5'-end TSS

(Supplemental Fig. S4.1A). Those TSSs that are more than 2 Kb away have median distance of 26.3 Kb. Second, the usage of the TSSs within the 2 Kb region (including the 5'-end TSS) is approximately three times of the TSS usage outside the region according to Fantom5 CAGE data (Forrest et al. 2014; Lizio et al. 2019) (Supplemental Fig. S4.1B). Therefore, 2 Kb is a reasonable range because of the coverage and usage of TSSs. Meanwhile, alternative promoters outside of this region were treated as enhancers, since enhancers and promoters share properties and functions as reviewed in ref. (Andersson and Sandelin 2020). If overlap exists, the regions would be merged to prevent double mapping of the localized features. Consequently, to combine the 5' promoter and distal regulatory elements, we devised two approaches as illustrated in Supplemental Figure S4.3. (1) *Prom + bin enhan* (blue) includes 40 equal-width bins of the promoter centered around the 5'-end TSS, 45 bins within the upstream region (-500 Kb to -2 Kb) from 5'-end TSS, and another 45 bins within the downstream region (2 Kb to 500 Kb). The bin widths for the distal regions are the multiple of 500 bp, e.g. the width of the first three bins closest to the TSS are 500, 1000, and 1500 bp respectively. (2) *Prom + agg enhan* (green) includes 40 equal-width bins of the promoter centered around the 5'-end TSS, one single upstream bin covering the entire region between -500 Kb and -2 Kb, and one single downstream bin covering the region between 2 Kb and 500 Kb. Next, the signals of each genome-wide feature that fall within the defined cis-regulatory regions were quantized within the corresponding bins based on their relative genomic distance.

Response frequency in perturbations

The response frequency of each gene to any perturbation is the number of perturbations to which it is responsive divided by the total number of perturbations. We inherently used the criteria for responsiveness described above to binarize target genes. The inductions of 194 yeast

TFs available in (Hackett et al. 2020) were incorporated in this calculation. And two types of human K562 perturbation samples in ENCODE (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020) were incorporated: (1) the perturbations of 56 TFs used for TF specific perturbation-response prediction, and (2) the perturbations of 355 genes.

4.4.2 Predicting TF-perturbation responses using cross-validation

For each TF perturbation, we simultaneously trained and tested the model for predicting whether a gene will respond using cross-validation on all genes. Specifically, every gene can be thought to be an instance for training or testing. As the genes were divided into ten folds at random, genes in one of the folds was reserved for testing while those in the other folds combined were used to train a model. Class stratification on gene split was applied to assure that all folds have equal proportion of responsive and unresponsive genes.

We trained and tested two ensemble classification algorithms—random forest implemented in scikit-learn library (V0.22.1) (Pedregosa et al. 2011), and gradient boosted trees implemented in XGBoost library (V 0.90) (Chen and Guestrin 2016). Both algorithms capitalize on the idea of "the wisdom of crowds", i.e. the ensemble of weak learners. For random forest, the number of trees was set to 500 while other hyperparameters were kept as default. For XGBoost, 500 shallow gradient-boosted trees were estimated, learning rate of 0.01 was set for the "gbtree" booster, and other hyperparameters were set as default. The complexity of the boosted trees, by default, are penalized by applying L2 regularization on feature weights. In each cross-validation run, the training data was standardized using Z-score transformation such that every feature is zero-centered with a unit standard deviation. Consequently, the testing data was standardized using the scaler learned from training.

The trained model was evaluated using precision recall curve, where the probabilities predicted for the held-out genes were compared against the measured binary responsiveness. The area under this precision recall curve (AUPRC) was used as the summary statistic for the corresponding cross-validation fold.

4.4.3 Using SHAP to quantify the predictive values of features

We employed SHAP (SHapley Additive exPlanations) framework (V0.35.0) (Lundberg and Lee 2017) to quantify the extent to which each feature contributes to the predicted probability of responsiveness for a gene. Briefly, SHAP exploits on the idea of using a linear model that are explainable to approximate the predicted probability of each example in a black-box model agnostically. Instead of directly interrogating a complex, nonlinear model, SHAP explains what the linear model learns. To train the linear model, SHAP samples new data points from a particular example (gene), each of which has a weight that is derived from Shapley value estimate, which quantifies the effect of removing a particular feature out of all possible combinations of other features. The use of Shapely values guarantee key desirable mathematical properties (Lundberg and Lee 2017). For the tree-based models, we utilized TreeExplainer function, which makes use of the node dependency in a tree structure to effectively reduce the approximation to polynomial complexity (Lundberg et al. 2018).

We calculated the SHAP values only for genes unseen by the model in each testing set; accordingly, we obtained each feature's SHAP value for every gene in the cross-validation framework. In this work, SHAP values explain why the prediction for one particular test example – one TF-gene pair – differs from the average prediction for all genes in response to perturbation of that one TF. Positive values indicate how strongly a particular feature value pushes the model toward assigning the gene a higher probability of responding, while negative values represent

how strongly the value pushes the model toward assigning the gene a lower probability of responding.

4.4.4 Aggregating SHAP values across genes of interest

To characterize the effect of a particular feature for a group of genes, we separately averaged all its positive SHAP values and its negative SHAP values. For each coordinate-dependent feature (e.g. localized TF binding), we independently summed its positive and negative SHAP values over genomic bins for each gene before averaging within the gene group. Concretely, we calculated mean positive SHAP value S_k^+ and mean negative SHAP value S_k^- as:

$$\mathrm{S}_{\mathrm{k}}^{+} = \sum_{i}^{G'} \sum_{j}^{B} \phi_{ijk}\left[\phi_{ijk} > 0\right] / |G'| \tag{4.1}$$

$$S_k^- = \sum_{i}^{G'} \sum_{j}^{B} \phi_{ijk}\left[\phi_{ijk} < 0\right] / |G'| \tag{4.2}$$

where ϕ_{ijk} is the SHAP value for gene i in bin j for feature k, G' is the set of gene indices, ($G' \subseteq$ G, where G is for all genes), and B is the set of bin indices. For coordinate-independent features (e.g., pre-perturbation gene expression), B has size of one.

We defined two terms for quantifying how each feature influences model prediction. *Net influence* is the sum of the positive and negative SHAP values of a feature together for a set of genes. It provides a sense of the feature's overall direction of influence. *Global feature importance* is the sum of absolute values of the SHAP values of a feature for all genes. It shows how important the feature is in determining the model's prediction, regardless of direction.

4.4.5 Modeling and interpreting generic responses in any genetic perturbation

The above classification framework was modified to predict how frequently that a gene would respond in any genetic perturbation in absence of TF information. Specifically, we trained and tested a XGBoost regressor rather than a classifier for predicting each of the response-frequency vectors. Each entry of the vector represents the fraction of conditions that a particular gene responds across all genetic perturbations. We created three vectors respectively for all induction conditions in yeast (n=194), TFKO/TFKD conditions in human K562 (n=56), and all genetic perturbations in K562 (n=355). The corresponding feature matrix for each label vector includes all features except for the localized TF binding data. The regression model for each label was cross validated.

Subsequently, we calculated the SHAP values for the testing genes. In a regression model, SHAP values explain why the prediction for one particular gene differs from the average frequency for all genes in response to any genetic perturbation. Positive values indicate how strongly a particular feature value increases the likelihood of generic response, while negative values represent how strongly the value decreases the likelihood of generic response. In addition, to summarize the overall importance of a feature, we calculated the mean absolute SHAP value across all genes. The higher the absolute value, the stronger overall influence the feature has.

Chapter 5: Discussion

5.1 Conclusion

Mapping high quality, genome-wide TF networks has been a long-standing goal in regulatory systems biology. Being able to map such networks from scalable and low-cost data resources (i.e., gene expression and annotated genome) provides researchers with powerful tools. In Chapter 2, we described a new network mapping framework -- NetProphet 2.0 -- based on the success of an earlier method developed in our lab. Our new method improved over the original and many other expression-based inference methods by capitalizing on three principles. First, assembly of multiple intermediary networks, or "wisdom of the crowds", outperforms any one of them. Second, TFs that have similar protein domains are likely to regulate similar sets of target genes. Third, an incomplete network is valuable for inferring TFs' motifs, which in turn help improve the network mapping quality.

Even when rich data resources (e.g., ChIP-seq) are available for well-studied yeast and human cells, the genes bound by a TF and those that respond upon the perturbation of the same TF were found to have little convergence. To address this mystery, we asked two question -- Can we find better convergence? And what factors in addition to TF binding determine the response?

In Chapter 3, we the developed dual threshold optimization method for setting significance thresholds on binding and perturbation-response data to improve the convergence. Integrating NetProphet 2.0 was found to further improve the results. Moreover, we found that data from new technologies for measuring TF perturbation responses (i.e., ZEV induction system) and TF binding locations (i.e., transposon calling cards and ChIP-exo) give further

advantages. Collectively, we progressed towards high-confidence network maps for yeast and human by applying the best combination of analytical and experimental methods.

In Chapter 4, we described a two-step process to elucidate the determinants of a gene's responsiveness under a TF perturbation. First, we trained machine learning models using TF binding, epigenetic, and gene expression features in unperturbed samples. Second, we applied SHAP values to explain the feature influences on each prediction. The binding signals of the perturbed TF were found only predictive in yeast promoters, while those mapped to enhancers and promoters in the human genome were of no use. Interestingly, inherent properties of each gene showed substantial predictive values. For instance, a gene whose expression level or expression variation is regularly high is poised to respond to any genetic perturbation regardless of which TF is perturbed.

Throughout this thesis work, we made steps towards mapping high-quality TF networks for several organisms, from simple eukaryotes to compact invertebrates, and from invertebrates to complex mammalian systems. Looking ahead, we anticipate that we will be better equipped to map reliable TF networks by taking several immediate steps.

5.2 Future directions

5.2.1 Improving the quality of expression-based network mapping using more precise input and expanded data resources

In Chapter 2, the premise behind regression modules is that a certain combination of TFs' activity levels is predictive of the expression level of a gene. Because we currently lack the experimental approach to directly measure TF activities, we made an assumption that the gene expression level is a reasonable approximation of the TF activity (TFA) level. However, factors

such as post-transcriptional and post-translational modifications are known to alter TFA levels. To address this issue, computational methods have been developed to infer TFA from multiple resources (reviewed in ref. (Ma and Brent 2020)). A recent method developed in our lab accurately inferred TFA from large-scaled expression datasets and a prior network (Ma and Brent 2020). We anticipate that TF network mapping will be greatly benefited from a more precise representation of TFA. One way to achieve the goal is to iterate between TF network mapping and TFA inference, where the output network map is the input to TFA inference and the output TFA levels are the input to network mapping, until the scores of network edges and TFA levels are stabilized.

From Chapter 4 and several other studies (Crow et al. 2019; Sigalova et al. 2020), we have learned that some genes are naturally poised to respond to a TF perturbation or any regulator perturbation. This implies that we need to be careful about calling a responsive gene the target of a particular perturbation. Computational tools have become more and more sophisticated in performing statistical tests to quantify the extent to which each gene changes its expression when comparing a perturbed sample to a control. However, a gene with a significant change in expression (fold change over control or P-value) should not always be interpreted as the perturbation target, just because it responds strongly. This means we cannot so safely say that such a gene is the functional target of the perturbed TF. So far, we know that some genes with certain properties are more likely to vary their expression levels regardless of the circumstances. Therefore, incorporating the prior knowledge (such as gene expression level and expression variation in unperturbed conditions) of genes can help generate adjusted transcriptomic response profiles that properly reflect the consequences of perturbations. This will bring substantial value to expression-based network mapping.

The rise of single cell technology, especially the wide adoption of transcriptomic profiling at single cell level (scRNA-seq) has brought a new horizon to study systems biology. The advantages that single cell data brings to network inference are (1) the large sample size, as a single sequencing run can simultaneously measure expression profiles of tens of thousands of cells; (2) the variance of expression levels across cells, which include underrepresented cell types that are not typically detectable by bulk sequencing. Moreover, the integration of CRISPR technology and single cell sequencing makes it possible to knock out multiple TF-encoding genes in a single tube of cells, where one of the targeted TFs in each cell is randomly perturbed (e.g. CRISP-seq (Jaitin et al. 2016), Perturb-seq (Dixit et al. 2016; Adamson et al. 2016; Replogle et al. 2020), CROP-seq (Datlinger et al. 2017), Mosaic-seq (Xie et al. 2017)), or activate the transcription of many genes (e.g. CRISPRa followed by scRNA-seq (Alda-Catalinas et al. 2020)). However, one caveat to keep in mind is that the network mapping algorithms designed to work with bulk gene expression data are not guaranteed to succeed when substituting the input with single cell data. Because of the low sequencing depth in each cell, profiles for scRNA-seq samples are discretized and zero-inflated, which produces different distributions compared to those for bulk RNA-seq samples. To make use of single cell data in NetProphet without algorithmic change, we may consider preprocessing scRNA-seq data using imputation algorithms. In a recent review (Hou et al. 2020), methods that smooth expression levels using cell-to-cell similarities (Wagner et al. 2017; Dijk et al. 2018) and capitalize on gene-to-gene relationships within each cell (Huang et al. 2018) were considered top contenders that accurately recover bulk expression data. Conversely, imputation methods were found to have no clear advantage of improving differential expression analysis over no imputation. Another study suggested that mapping TF networks without preprocessing scRNA-seq data using complex

normalization or imputation worked reasonably well (Jackson et al. 2020). Furthermore, other newly developed imputation and augmentation methods using deep generative models (Xu et al. 2020; Marouf et al. 2020) could be potential alternatives for this task. While the rapid growth of single cell data can provide us with rich transcriptomic information for mapping TF networks, these data should be used with caution as different modules in NetProphet may require different input preprocessing, the effect of which remains to be seen. Alternatively, devising novel algorithms that directly make use of single cell profiles is a completely feasible avenue.

5.2.2 Improving the accuracy and efficiency of expression-based network inference using better implementation

The overarching goal of regression modules in NetProphet is to assign importance scores to TFs for each gene. For each gene, a BART model is trained to predict its expression level from that of all TFs. Then, the influence of each TF on the gene is calculated as the predicted gene responses under a simulated change of the TF's expression level while other TFs remain at their respective mean levels. This is a strong assumption that may create counterfactual conditions -- because the expression levels of the TFs can be highly correlated, changing one between extreme values while maintaining others at their mean values may be unrealistic. To address the issue, we can directly explain the BART model by applying SHAP values, which is one of many interpretable machine learning methods, without simulating new unrealistic conditions. Precisely, for the trained BART model for a particular gene, we would first calculate the SHAP values for each TF (predictor), which represents how much influence this TF exerts on the predicted level of the gene in a particular expression profile. Next, we would try scoring the TF-gene edge by summarizing SHAP values. (1) A straightforward way is to calculate the global importance of the TF by averaging the absolute SHAP values over samples. However, this

approach loses information about the sign of the TF's influence on the gene's predicted expression level. (2) To retain the sign, we need to put the magnitude of each feature's influence into the equation. The idea is to take the signed SHAP value only if it makes up a significant fraction of the target's response level and their signs are in the same direction. A more sophisticated implementation may better capitalize on this idea.

Many transcriptomic profiling experiments have been conducted in time-series, including ZEV TF-induction (Hackett et al. 2020). NetProphet, as currently implemented, treats gene expression profiles in a time-series as independent samples. As each input profile is assumed to be measured at equilibrium, the inference is unable to fully exploit the temporal information. Therefore, at minimal, our regression modules will need to incorporate the classic kinetic model for TF regulation (Bonneau et al. 2006). We will implement an additional layer of processing to generate the pseudo-time transcriptomic profiles, which requires a vector of RNA half-life (or RNA degradation rate) data for all genes as input parameters. In a recent collaboration of mapping NetProphet networks for archaea cells, we incorporated the use of kinetic model on time-series data (data not shown). Furthermore, we may consider updating the differential expression analysis for time-series profiles (Spies et al. 2019).

We developed NetProphet in a high throughput computing environment, where CPU time and memory are not limiting. For ease of use in a desktop environment, there is the need to optimize individual modules for computing efficiency such as runtime and memory. (1) The estimation of global shrinkage in LASSO module is computationally intensive. Its novelty lies in the search for a single shrinkage parameter that is universal for all regression models, each of which corresponds to a gene. This is computationally expensive, because all potential shrinkage parameter values obtained by separate LARS runs on each gene are currently checked for cross-

validation error on all genes. To speed up the search for shrinkage parameter, we can apply grid search or Bayesian Optimization on a pre-defined range of parameters. (2) The implementation used for BART regression has a long runtime. Placing with a more recent BART implementation (XBART) or other regression software such as XGBoost are viable solutions for speedup (He et al. 2018). However, we should not trade accuracy for speed, which would defeat the purpose of mapping high-confidence TF networks.

5.2.3 Mapping a high-confidence global human network

One ambition in the field of regulatory systems biology is to accurately map the global human TF network, which is a key component for many downstream applications, from developmental biology to transcriptome engineering, and from disease modeling to drug discovery. Such a network can be defined in many ways. In our view of the global network, each edge represents the maximal potential a TF exerts on each of its direct and functional targets. If one's interest is in a cell-type specific network, we will first infer TFA levels based on the global network along with a particular expression profile or a subset of profiles representing the cell type. And subsequently update the edge weights by integrating the inferred TFA levels; specifically, up-weight edges of the global network outgoing from active TFs and down-weight the ones outgoing from inactive TFs. On the other hand, we can try a different approach that directly maps cell-type specific network without having to first map a global network; specifically, the regression modules will use expression data for the cell type of interest, while the differential expression module will use perturbed samples for all cell types (since TF-perturbation samples for human cells are relatively limited).

In Chapter 3, we made our effort to map human networks for independent cell types using the combination of DTO and NetProphet. A natural next step is to run the same procedure

using an expanded collection of datasets that include more cell and tissue types. The following are several large-scaled sets we identified so far: ENCODE database (Dunham et al. 2012; Davis et al. 2018; Abascal et al. 2020) is a reliable source that has been frequently updated over the past decade. ReMap (Chèneby et al. 2020) also provides large collections of human TF binding data from ChIP-seq, ChIP-exo, and DAP-seq (DNA affinity purification sequencing). Alternatively, computational approaches can help recovering the *in vivo* TF binding sites (TFBSs) that have not been assayed, due to high cost and intensive labor of experimental approaches. The idea is to filter potential binding sites whose DNA sequences match a TF's motif by using chromatin accessibility (Pique-Regi et al. 2011; Li et al. 2018). Other epigenomic features including DNA sequence, DNA shape, and histone marks near each motif hit were also predictive of *in vivo* binding sites. (Xin and Rohs 2018). For gene expression, in addition to ENCODE database, several large-scaled TF perturbation datasets are available (Hurley et al. 2012; Nakatake et al. 2020) as well as GTEx database for tissue expression (with no perturbation) (Lonsdale et al. 2013; Aguet et al. 2017). As each dataset may have batch effect that is inherent in data generation and processing, it is not yet clear whether running network mapping programs using input combined all together or combining networks mapped separately using individual datasets brings more benefit.

Our initial motivation for developing NetProphet was to map a TF network for yeast, which has a compact genome. A network map containing two sets of nodes -- TFs and genes -- seemed adequate, as TFs bind to the proximal region of a gene to effectively modulate the gene's expression level. Human cells, by contrast, contain multiple *cis*-regulatory elements (CREs), such as enhancers that are located distally (in a genome coordinates) from the associated genes and promoters that are located nearby the genes. Our current network architecture may be

oversimplified. Therefore, we anticipate a new design of directed acyclic graph that includes a new layer of nodes for CREs between TF and gene nodes: a TF-CRE edge represents a TFBS in a CRE, while a CRE-gene edge represents the association between a CRE and its target (Brent 2016). The effect of direct and functional regulation between a TF and its target gene can quantified as the combination of non-zero scores for the two edges and the activities levels of the corresponding TF and CRE. To build such a network, we will need three arms: (1) the *in vivo* binding evidence or binding potential of TFBSs in CREs; (2) the CRE-gene association (whether physical chromatin contact or functional linkage); (3) the transcriptional association between TF and gene. The first arm can be obtained from established experimental and computational methods. The second arm is an area of active research, which will be discussed in detail in the following section. The last arm can inherit co-expression and differential expression analyses in NetProphet. Subsequently the inferred score of each TF-gene pair will be integrated with all potential TF-CRE-gene paths. In addition, we will also need to modify TFA inference to accommodate the new architecture and collect evidence or integrate inference for the CRE activity. Since cooperativity within small groups of TFs occur frequently in complex organisms, it may be worthwhile constraining TFs that are likely to interact to have similar activity levels in the TFA inference module, where TF interactions that have been systematically evaluated and compiled from multiple sources are available in STRING (Szklarczyk et al. 2019) and BioGRID (Oughtred et al. 2019). Overall, the TF-CRE-gene architecture is expected to offer the versatility of more accurately characterizing the direct, functional TF-gene interactions mediated by CREs.

5.2.4 Improving the identification of activities and gene associations of *cis*-regulatory elements

A key component for mapping accurate TF networks, especially for complex mammalian genomes, is the association between each active CRE and its target genes (Gasperini et al. 2020; Brent 2016). Such characteristics are also expected to improve our ability to reliably predict TF-perturbation responses (Chapter 4); thereby, the influences of TF binding signals and epigenetic features can be better explained. Currently, technologies have been developed to identify (1) the locations of active CREs and (2) the associations between CREs and genes.

Nascent RNAs transcribed from CREs appear to be a good indicator of the CREs' activity. Technologies such as GRO-seq (Core et al. 2014), PRO-seq (Mahat et al. 2016), and CoPro (Tome et al. 2018) were developed for measuring such transcriptions *in vivo*. Data are currently limited to a few cell types. Therefore, we expect data for systematically evaluating a range of cell types to become available in the near future.

The goals for identifying the CRE-gene associations can be divided into two classes -- direct interactions and functional interactions. To measure the physical contact in 3D genome, experimental methods such as Hi-C (Lieberman-Aiden et al. 2009), ChIA-PET (Fullwood and Ruan 2009), and HiChIP (Mumbach et al. 2016, 2017) were developed. While 3D loops identified by Hi-C requires feature processing to filter for the loop ends that contain activate CREs, those identified by ChIPA-PET or HiChIP contain active CREs marked by H3K27ac (a well-known indicator of active enhancer) in at least one loop end. On the other hand, to measure the functional CRE-gene association, experimental methods were developed to determine the causal effects of interfering or activating CREs on the nearby genes (Fulco et al. 2016; Klann et al. 2017; Simeonov et al. 2017). As *in vivo* evidence is currently limited to a few well-studies

human cell types, computational methods can offer wider coverage. GTEx *cis*-eQTL data (Lonsdale et al. 2013; Aguet et al. 2017) have been used to identify CRE-gene associations, if a SNP in the CRE is statistically associated with the expression variation of a gene. Moving from bulk to single cell sequencing, technologies such as Sci-CAR (Cao et al. 2018) and 10x Chromium Single Cell Multiome ATAC + Gene Expression provide the possibility to jointly measure chromatin accessibility and gene expression in the same cell. A functional association can be established if the expression level of a gene increases as a nearby CRE becomes accessible, and vice versa. For a single cell sequencing run on cell mixtures, the CRE-gene associations for multiple cell types can be simultaneously inferred. Nevertheless, an accurate algorithm that is thoroughly validated by experimental evidence remains to be seen.

5.2.5 Improving response prediction using network maps for TF-TF interactions and TF-gene regulations

Among the features used to predict which gene will respond to a regulatory perturbation (Chapter 4), we only used the TF binding information of the perturbed TF, which might hinder our ability to reliably predict transcriptional responses. From the perspective of systems biology, the limitations are likely caused by two reasons. First, TF interactions such as cooperation and competition can affect how a gene is regulated through TF binding. Second, regulatory dependencies between TFs in a TF network can affect how a gene is indirectly regulated.

To address the first issue, for each TF perturbation, we can add the binding signals of other TFs that interact cooperatively or competitively with the perturbed TF as new features. Some TFs form a protein complex and bind their target as a single entity. Thus, we expect that knowing the sites in a gene's regulatory region that are adjacently bound by the complex members will improve our ability to predict the gene's likelihood of response. On the other hand,

since some TFs compete for the binding sites by recognizing similar sequence motifs, we expect that the unresponsive targets with the same sites bound by competing TFs can be more accurately predicted. For example, yeast TF Tye7 is known to cooperate with three other factors while competing with Cbf1 (Shively et al. 2019). In the perturbation of Tye7, targets bound by Tye7 and its cooperative factors are expected to be responsive; but those bound by Cbf1 are expected to be unresponsive. The cooperative factors can be identified using the evidence of direct physical contacts curated in protein-protein interaction (PPI) network (e.g., STRING (Szklarczyk et al. 2019) or BioGRID (Oughtred et al. 2019)). The competing factors can be identified as those that have no connection in the PPI network but have similar motifs or DBDs (if the motif is unavailable) (Chapter 2). We then need to define features to encode the two types of TF interaction and use them to annotate the binding features of each of the interacting TFs.

To address the second issue where a gene can be indirectly regulated by the perturbed TF, we need to propagate the effect of perturbation to indirect targets through the responses of intermediary TFs. Specifically, we will incorporate the transcriptional responses of all genes that encode TFs as an additional set of features. Moreover, we need to inform the model about the regulatory relationship from these intermediary TFs to their respective targets. To do so, we will need a prior -- the TF network map based on binding data for ChIP or PWMs processed with epigenetic data (discussed in 5.2.3), or network inference using NetProphet 2.0 with differential expression module disabled (to avoid data peeking). Next, for each gene, we will reweigh the response of each intermediary TF using the edge score of a network map. Finally, we will characterize how each of the new features contributes to the predicted probability. We hypothesize that if a gene is the indirect target of the perturbed TF, then a fraction of the new features representing intermediate regulators of the gene would show some certain degrees of

influence on the prediction. This would indicate that the effect of TF perturbation has been propagated through the TF network. Conversely, if the gene is the direct target of the perturbed TF, then none of these features would be expected to have influence on the outcome. Taken together, we anticipate that informing models with features for the synergistic and competitive interactions among TFs and those for network structure will provide a comprehensive picture of transcriptional responses.

References

Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, Akiyama JA, Jammal O Al, Amrhein H, Anderson SM, et al. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**: 699–710.

Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço 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*.

Adamson B, Norman TM, Jost M, Cho MY, Nuñez JK, Chen Y, Villalta JE, Gilbert LA, Horlbeck MA, Hein MY, et al. 2016. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*.

Agarwal V, Shendure J. 2020. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. *Cell Rep* **31**: 107663. https://doi.org/10.1016/j.celrep.2020.107663.

Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, Mohammadi P, Park YS, Parsana P, Segrè A V., et al. 2017. Genetic effects on gene expression across human tissues. *Nature*.

Alda-Catalinas C, Bredikhin D, Hernando-Herraez I, Santos F, Kubinyecz O, Eckersley-Maslin MA, Stegle O, Reik W. 2020. A Single-Cell Transcriptomics CRISPR-Activation Screen Identifies Epigenetic Regulators of the Zygotic Genome Activation Program. *Cell Syst*.

Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Hilda Ye B, Califano A. 2016. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet* **48**: 838–847.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–461.

Andersson R, Sandelin A. 2020. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* **21**: 71–87. http://dx.doi.org/10.1038/s41576-019-0173-8.

Aytes A, Mitrofanova A, Lefebvre C, Alvarez MJ, Castillo-Martin M, Zheng T, Eastham JA, Gopalan A, Pienta KJ, Shen MM, et al. 2014. Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy. *Cancer Cell*.

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA. 2004. Structure and

evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*.

Balwierz PJ, Pachkov M, Arnold P, Gruber AJ, Zavolan M, Van Nimwegen E. 2014. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res*.

Bansal M, Yang J, Karan C, Menden MP, Costello JC, Tang H, Xiao G, Li Y, Allen J, Zhong R, et al. 2014. A community computational challenge to predict the activity of pairs of compounds. *Nat Biotechnol*.

Bergenholm D, Liu G, Holland P, Nielsen J. 2018. Reconstruction of a Global Transcriptional Regulatory Network for Control of Lipid Metabolism in Yeast by Using Chromatin Immunoprecipitation with Lambda Exonuclease Digestion. *mSystems*.

Bhagwat AS, Vakoc CR. 2015. Targeting Transcription Factors in Cancer. *Trends in Cancer*.

Bonke M, Turunen M, Sokolova M, Vähärautio A, Kivioja T, Taipale M, Björklund M, Taipale J. 2013. Transcriptional networks controlling the cell cycle. *G3 Genes, Genomes, Genet*.

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**: 1. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-5-r36.

Boorsma A, Lu XJ, Zakrzewska A, Klis FM, Bussemaker HJ. 2008. Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One*.

Boulesteix AL, Strimmer K. 2005. Predicting transcription factor activities from combined analysis of microarray and ChIP data: A partial least squares approach. *Theor Biol Med Model*.

Breiman L. 2001. Random forests. *Mach Learn*.

Brent MR. 2016. Past Roadblocks and New Opportunities in Transcription Factor Network Mapping. *Trends Genet* **32**: 736–750. http://dx.doi.org/10.1016/j.tig.2016.08.009.

Brown JB, Celniker SE. 2015. Lessons from modENCODE. *Annu Rev Genomics Hum Genet*.

Cahan P, Li H, Morris SA, Lummertz Da Rocha E, Daley GQ, Collins JJ. 2014. CellNet: Network biology applied to stem cell engineering. *Cell* **158**: 903–915.

Cao J, Cusanovich DA, Ramani V, Aghamirzaie D, Pliner HA, Hill AJ, Daza RM, McFaline-Figueroa JL, Packer JS, Christiansen L, et al. 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science (80-)* **361**: 1380–1385.

Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, Sulman EP, Anne SL, Doetsch F, Colman H, et al. 2010. The transcriptional network for mesenchymal transformation of brain tumours. *Nature* **463**: 318–325. http://dx.doi.org/10.1038/nature08712.

Castro DM, Veaux N de, Miraldi ER, Bonneau R. 2018. Multi-study inference of regulatory networks for more accurate models of gene regulation. *bioRxiv* 279224. https://www.biorxiv.org/content/early/2018/03/08/279224.

Chen T, Guestrin C. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chèneby J, Ménétrier Z, Mestdagh M, Rosnet T, Douida A, Rhalloussi W, Bergon A, Lopez F, Ballester B. 2020. ReMap 2020: A database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res*.

Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M. 2011. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol*.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, et al. 2012. Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Res*.

Chipman HA, George EI, McCulloch RE. 2012. BART: Bayesian additive regression trees. *Ann Appl Stat* **6**: 266–298.

Clough E, Tedeschi T, Hazelrigg T. 2014. Epigenetic regulation of oogenesis and germ stem cell maintenance by the Drosophila histone methyltransferase Eggless/dSetDB1. *Dev Biol*.

Core LJ, Martins AL, Danko CG, Waters CT, Siepel A, Lis JT. 2014. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet*.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (80-)*.

Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. 2019. Predictability of human differential gene expression. *Proc Natl Acad Sci U S A* **116**: 6491–6500.

Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. 2012. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S, et al. 2019. Ensembl 2019. *Nucleic Acids Res*.

Cusanovich DA, Hill AJ, Aghamirzaie D, Daza RM, Pliner HA, Berletch JB, Filippova GN, Huang X, Christiansen L, DeWitt WS, et al. 2018. A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**: 1309-1324.e18. https://doi.org/10.1016/j.cell.2018.06.052.

Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. 2014. The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet* **10**.

D'Alessio AC, Fan ZP, Wert KJ, Baranov P, Cohen MA, Saini JS, Cohick E, Charniga C, Dadon D, Hannett NM, et al. 2015. A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports*.

Da Silveira WA, Palma PVB, Sicchieri RD, Villacis RAR, Mandarano LRM, Oliveira TMG, Antonio HMR, Andrade JM, Muglia VF, Rogatto SR, et al. 2017. Transcription factor networks derived from breast cancer stem cells control the immune response in the basal subtype. *Sci Rep*.

Danko CG, Hyland SL, Core LJ, Martins AL, Waters CT, Lee HW, Cheung VG, Kraus WL, Lis JT, Siepel A. 2015. Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12**: 433–438.

Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. 2017. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods*.

Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al. 2018. The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res*.

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

Dijk D Van, Sharma R, Nainys J, Wolf G, Krishnaswamy S, Pe D, Dijk D Van, Sharma R, Nainys J, Yim K, et al. 2018. Recovering Gene Interactions from Single-Cell Data Resource Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell* **174**: 716-729.e27. https://doi.org/10.1016/j.cell.2018.05.061.

Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, Marjanovic ND, Dionne D, Burks T, Raychowdhury R, et al. 2016. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**: 1853-1866.e17. http://dx.doi.org/10.1016/j.cell.2016.11.038.

Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M,

Guigó R, Birney E, et al. 2012. Modeling gene expression using chromatin features in various cellular contexts. *Genome Biol*.

Dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, Brown NH, Kaufman T, et al. 2015. FlyBase: Introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res*.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Elemento O, Slonim N, Tavazoie S. 2007. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Mol Cell* **28**: 337–350.

Engel SR, Dietrich FS, Fisk DG, Binkley G, Balakrishnan R, Costanzo MC, Dwight SS, Hitz BC, Karra K, Nash RS, et al. 2014. The Reference Genome Sequence of Saccharomyces cerevisiae: Then and Now. *G3 Genes, Genomes, Genet*.

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**: 0054–0066.

Fisher A, Rudin C, Dominici F. 2019. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J Mach Learn Res*.

Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al. 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**: 1–17. https://academic.oup.com/database/article-lookup/doi/10.1093/database/bax028.

Forrest ARR, Kawaji H, Rehli M, Baillie JK, De Hoon MJL, Haberle V, Lassmann T, Kulakovskiy I V., Lizio M, Itoh M, et al. 2014. A promoter-level mammalian expression atlas. *Nature* **507**: 462–470.

Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, Kane M, Cleary B, Lander ES, Engreitz JM. 2016. Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science (80-)*.

Fullwood MJ, Ruan Y. 2009. ChIP-based methods for the identification of long-range chromatin interactions. *J Cell Biochem*.

Fuxman Bass JI, Pons C, Kozlowski L, Reece-Hoyes JS, Shrestha S, Holdorf AD, Mori A,

Myers CL, Walhout AJ. 2016. A gene-centered C. elegans protein– DNA interaction network provides a framework for functional predictions . *Mol Syst Biol*.

Garcia-Alonso L, Iorio F, Matchan A, Fonseca N, Jaaks P, Peat G, Pignatelli M, Falcone F, Benes CH, Dunham I, et al. 2018. Transcription factor activities enhance markers of drug sensitivity in cancer. *Cancer Res*.

Gasperini M, Tome JM, Shendure J. 2020. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet* **40**. http://dx.doi.org/10.1038/s41576-019-0209-0.

Gautier L, Cope L, Bolstad BM, Irizarry RA. 2004. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*.

Gayvert KM, Dardenne E, Cheung C, Boland MR, Lorberbaum T, Wanjala J, Chen Y, Rubin MA, Tatonetti NP, Rickman DS, et al. 2016. A Computational Drug Repositioning Approach for Targeting Oncogenic Transcription Factors. *Cell Rep*.

Georlette D, Ahn S, MacAlpine DM, Cheung E, Lewis PW, Beall EL, Bell SP, Speed T, Manak JR, Botchan MR. 2007. Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb-MuvB/dREAM complex in proliferating cells. *Genes Dev*.

Ghanbari M, Lasserre J, Vingron M. 2015. Reconstruction of gene networks using prior knowledge. *BMC Syst Biol*.

Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*.

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**: 1–7. http://dx.doi.org/10.1038/msb.2009.33.

González AJ, Setty M, Leslie CS. 2015. Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation. *Nat Genet*.

Gordân R, Murphy KF, 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*.

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

Greenfield A, Hafemeister C, Bonneau R. 2013. Robust data-driven incorporation of prior

knowledge into the inference of dynamic regulatory networks. *Bioinformatics* **29**: 1060–1067.

Greenfield A, Madar A, Ostrer H, Bonneau R. 2010. DREAM4: Combining genetic and dynamic information to identify biological networks and Dynamical Models. *PLoS One*.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol*.

Hackett S, Baltz E, Coram M, Wranik B, Kim G, Baker A, Fan M, Hendrickson D, Berndl M, McIsaac RS. 2019. Time-resolved genome-scale profiling reveals a causal expression network. *Mol Syst Biol*.

Hackett SR, Baltz EA, Coram M, Wranik BJ, Kim G, Baker A, Fan M, Hendrickson DG, Berndl M, Mcisaac RS. 2020. Learning causal networks using inducible transcription factors and transcriptome-wide time series. 1–15.

Hadzić T, Park D, Abruzzi KC, Yang L, Trigg JS, Rohs R, Rosbash M, Taghert PH. 2015. Genome-wide features of neuroendocrine regulation in Drosophila by the basic helix-loop-helix transcription factor DIMMED. *Nucleic Acids Res*.

Hainer SJ, Fazzio TG. 2019. High-Resolution Chromatin Profiling Using CUT&RUN. *Curr Protoc Mol Biol*.

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*.

Hass MR, Liow H haw, Chen X, Sharma A, Inoue YU, Inoue T, Reeb A, Martens A, Fulbright M, Raju S, et al. 2015. SpDamID: Marking DNA Bound by Protein Complexes Identifies Notch-Dimer Responsive Enhancers. *Mol Cell*.

Haury A-C, Mordelet F, Vera-Licona P, Vert J-P. 2012. TIGRESS: Trustful Inference of Gene REgulation using Stability Selection. *BMC Syst Biol* **6**: 145. http://www.ncbi.nlm.nih.gov/pubmed/23173819.

Haynes BC, Maier EJ, Kramer MH, Wang PI, Brown H, Brent MR. 2013. Mapping functional transcription factor networks from gene expression data. *Genome Res* **23**: 1319–1328.

He J, Yalov S, Hahn PR. 2018. XBART: Accelerated Bayesian additive regression trees. *arXiv* **89**.

Heinäniemi M, Nykter M, Kramer R, Wienecke-Baldacchino A, Sinkkonen L, Zhou JX, Kreisberg R, Kauffman SA, Huang S, Shmulevich I. 2013. Gene-pair expression signatures reveal lineage control. *Nat Methods*.

Henikoff S, Shilatifard A. 2011. Histone modification: Cause or cog? *Trends Genet* **27**: 389–396. http://dx.doi.org/10.1016/j.tig.2011.06.006.

Holland P, Bergenholm D, Börlin CS, Liu G, Nielsen J. 2019. Predictive models of eukaryotic transcriptional regulation reveals changes in transcription factor roles and promoter usage between metabolic conditions. *Nucleic Acids Res*.

Hou W, Ji Z, Ji H, Hicks SC. 2020. A Systematic Evaluation of Single-cell RNA-sequencing Imputation Methods. *bioRxiv* 1–30.

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

Huang M, Wang J, Torre E, Dueck H, Shaffer S, Bonasio R, Murray JI, Raj A, Li M, Zhang NR. 2018. SAVER: Gene expression recovery for single-cell RNA sequencing. *Nat Methods*.

Hughes TR. 2011. Introduction to "a handbook of transcription factors." *Subcell Biochem*.

Hurley D, Araki H, Tamada Y, Dunmore B, Sanders D, Humphreys S, Affara M, Imoto S, Yasuda K, Tomiyasu Y, et al. 2012. Gene network inference and visualization tools for biologists: Application to new human transcriptome datasets. *Nucleic Acids Res*.

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

Ikmi A, Gaertner B, Seidel C, Srivastava M, Zeitlinger J, Gibson MC. 2014. Molecular evolution of the Yap/Yorkie proto-oncogene and elucidation of its core transcriptional program. *Mol Biol Evol*.

Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*.

Jackson CA, Castro DM, Saldi GA, Bonneau R, Gresham D. 2020. Gene regulatory network reconstruction using single-cell rna sequencing of barcoded genotypes in diverse environments. *Elife* **9**: 1–34.

Jaitin DA, Weiner A, Yofe I, Lara-Astiaso D, Keren-Shaul H, David E, Salame TM, Tanay A, van Oudenaarden A, Amit I. 2016. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*.

Japkowicz N, Stephen S. 2002. The class imbalance problem: A systematic study. *Intell Data Anal*.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013. DNA-binding specificities of human transcription factors. *Cell* **152**:

327–339. http://dx.doi.org/10.1016/j.cell.2012.12.009.

Kang Y, Liow HH, Maier EJ, Brent MR. 2018. NetProphet 2.0: Mapping transcription factor networks by exploiting scalable data resources. *Bioinformatics* **34**: 249–257.

Kang Y, Patel NR, Shively C, Recio PS, Chen X, Wranik BJ, Kim G, McIsaac RS, Mitra R, Brent MR. 2020. Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. *Genome Res* gr.259655.119.

Kao KC, Yang YL, Boscolo R, Sabatti C, Roychowdhury V, Liao JC. 2004. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *Proc Natl Acad Sci U S A*.

Karlić R, Chung HR, Lasserre J, Vlahoviček K, Vingron M. 2010. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci U S A* **107**: 2926–2931.

Kelley DR, Reshef YA, Bileschi M, Belanger D, Mclean CY, Snoek J. 2018. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. 1–12.

Kemmeren P, Sameith K, Van De Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, Van Wageningen S, Ko CW, et al. 2014. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**: 740–752. http://dx.doi.org/10.1016/j.cell.2014.02.054.

Klann TS, Black JB, Chellappan M, Safi A, Song L, Hilton IB, Crawford GE, Reddy TE, Gersbach CA. 2017. CRISPR-Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nat Biotechnol*.

Kudron MM, Victorsen A, Gevirtzman L, Hillier LW, Fisher WW, Vafeados D, Kirkey M, Hammonds AS, Gersch J, Ammouri H, et al. 2018. The modern resource: genome-wide binding profiles for hundreds of Drosophila and Caenorhabditis elegans transcription factors. *Genetics*.

Kundert K, Lucas JE, Watters KE, Fellmann C, Ng AH, Heineike BM, Fitzsimmons CM, Oakes BL, Qu J, Prasad N, et al. 2019. Controlling CRISPR-Cas9 with ligand-activated and ligand-deactivated sgRNAs. *Nat Commun*.

Lam KY, Westrick ZM, Müller CL, Christiaen L, Bonneau R. 2016. Fused Regression for Multi-source Gene Regulatory Network Inference. *PLoS Comput Biol*.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell* **172**: 650–665.

https://doi.org/10.1016/j.cell.2018.01.029.

Lenstra TL, Holstege FCP. 2012. The discrepancy between chromatin factor location and effect. *Nucl (United States)*.

Li H, Quang D, Guan Y. 2018. Anchor: Trans-cell Type Prediction of Transcription Factor Binding Sites. *Genome Res*.

Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP. 2003. Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*.

Lieberman-Aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-)*.

Liu J, Ghanim M, Xue L, Brown CD, Iossifov I, Angeletti C, Hua S, Nègre N, Ludwig M, Stricker T, et al. 2009. Analysis of Drosophila segmentation network identifies a JNK pathway factor overexpressed in kidney cancer. *Science (80-)*.

Lizio M, Abugessaisa I, Noguchi S, Kondo A, Hasegawa A, Hon CC, De Hoon M, Severin J, Oki S, Hayashizaki Y, et al. 2019. Update of the FANTOM web resource: Expansion to provide additional transcriptome atlases. *Nucleic Acids Res*.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 1–21.

Lundberg S, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. *NIPS* **16**: 426–430.

Lundberg SM, Erion GG, Lee S. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. http://arxiv.org/abs/1802.03888.

Ma C, Brent M. 2020. Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data. 1–28.

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R. 2010. DREAM3: Network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS One*.

Mahat DB, Kwak H, Booth GT, Jonkers IH, Danko CG, Patel RK, Waters CT, Munson K, Core LJ, Lis JT. 2016. Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat Protoc*.

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Aderhold A, et al. 2012a. Wisdom of crowds for robust gene network inference. *Nat Methods*.

Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M. 2012b. Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. *Genome Res* **22**: 1334–1349.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res*.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, Califano A. 2006. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*.

Marouf M, Machart P, Bansal V, Kilian C, Magruder DS, Krebs CF, Bonn S. 2020. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nat Commun*.

Mayhew D, Mitra RD. 2016. Transposon calling cards. *Cold Spring Harb Protoc*.

McIsaac RS, Gibney PA, Chandran SS, Benjamin KR, Botstein D. 2014. Synthetic biology tools for programming gene expression without nutritional perturbations in Saccharomyces cerevisiae. *Nucleic Acids Res*.

McIsaac RS, Oakes BL, Botstein D, Noyes MB. 2013. Rapid synthesis and screening of chemically activated transcription factors with GFP-based reporters. *J Vis Exp*.

McIsaac RS, Silverman SJ, McClean MN, Gibney PA, Macinskas J, Hickman MJ, Petti AA, Botstein D. 2011. Fast-acting and nearly gratuitous induction of gene expression and protein depletion in Saccharomyces cerevisiae. *Mol Biol Cell*.

McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. 2012. Genome-wide in silico prediction of gene expression. *Bioinformatics*.

Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, et al. 2015. RSAT 2015: Regulatory sequence analysis tools. *Nucleic Acids Res*.

Meers MP, Janssens DH, Henikoff S. 2019a. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell*.

Meers MP, Janssens DH, Henikoff S. 2019b. Pioneer Factor-Nucleosome Binding Events during Differentiation Are Motif Encoded. *Mol Cell*.

Michael DG, Maier EJ, Brown H, Gish SR, Fiore C, Brown RH, Brent MR. 2016. Model-based transcriptome engineering promotes a fermentative transcriptional state in yeast. *Proc Natl Acad Sci* **113**: E7428–E7437. http://www.pnas.org/lookup/doi/10.1073/pnas.1603577113.

Middendorf M, Kundaje A, Wiggins C, Freund Y, Leslie C. 2004. Predicting genetic regulatory response using classification. In *Bioinformatics*.

Miraldi ER, Pokrovskii M, Watters A, Castro DM, De Veaux N, Hall JA, Lee JY, Ciofani M, Madar A, Carriero N, et al. 2019. Leveraging chromatin accessibility for transcriptional regulatory network inference in T Helper 17 Cells. *Genome Res* **29**: 449–463.

Molnar C. 2019. Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*.

Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. 2016. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**: 919–922. http://www.ncbi.nlm.nih.gov/pubmed/27643841%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5501173.

Mumbach MR, Satpathy AT, Boyle EA, Dai C, Gowen BG, Cho SW, Nguyen ML, Rubin AJ, Granja JM, Kazane KR, et al. 2017. Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat Genet* **49**: 1602–1612. http://www.ncbi.nlm.nih.gov/pubmed/28945252%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5805393.

Nakatake Y, Ko SBH, Sharov AA, Wakabayashi S, Murakami M, Sakota M, Chikazawa N, Ookura C, Sato S, Ito N, et al. 2020. Generation and Profiling of 2,135 Human ESC Lines for the Systematic Analyses of Cell States Perturbed by Inducing Single Transcription Factors. *Cell Rep* **31**: 107655. https://doi.org/10.1016/j.celrep.2020.107655.

Oakes BL, Nadler DC, Flamholz A, Fellmann C, Staahl BT, Doudna JA, Savage DF. 2016. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat Biotechnol*.

Oughtred R, Stark C, Breitkreutz BJ, Rust J, Boucher L, Chang C, Kolas N, O'Donnell L, Leung G, McAdam R, et al. 2019. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*.

Ouyang Z, Zhou Q, Wong WH. 2009. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci U S A*.

Page AR, Kovacs A, Deak P, Torok T, Kiss I, Dario P, Bastos C, Batista P, Gomes R, Ohkura H, et al. 2005. Spotted-dick, a zinc-finger protein of Drosophila required for expression of

Orc4 and S phase. *EMBO J*.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine learning in Python. *J Mach Learn Res*.

Perreault AA, Venters BJ. 2016. The chIP-exo method: Identifying protein-DNA interactions with near base pair precision. *J Vis Exp*.

Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res*.

Policastro RA, Zentner GE. 2018. Enzymatic methods for genome-wide profiling of protein binding sites. *Brief Funct Genomics*.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*.

Rackham OJL, Firas J, Fang H, Oates ME, Holmes ML, Knaupp AS, Suzuki H, Nefzger CM, Daub CO, Shin JW, et al. 2016. A predictive computational framework for direct reprogramming between human cell types. *Nat Genet*.

Read DF, Cook K, Lu YY, Le Roch KG, Noble WS. 2019. Predicting gene expression in the human malaria parasite Plasmodium falciparum using histone modification, nucleosome positioning, and 3D localization features. *PLoS Comput Biol* **15**: 1–23. http://dx.doi.org/10.1371/journal.pcbi.1007329.

Replogle J, Xu A, Norman T, Meer E, Terry J, Riordan D, Srinivas N, Mikkelsen T, Weissman J, Adamson B. 2018. Direct capture of CRISPR guides enables scalable, multiplexed, and multi-omic Perturb-seq. *Nat Biotechnol*.

Replogle JM, Norman TM, Xu A, Hussmann JA, Chen J, Cogan JZ, Meer EJ, Terry JM, Riordan DP, Srinivas N, et al. 2020. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat Biotechnol*.

Rhee HS, Pugh BF. 2012. ChiP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol*.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*.

Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature*.

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*.

Rossi MJ, Lai WKM, Pugh BF. 2018a. Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res*.

Rossi MJ, Lai WKM, Pugh BF. 2018b. Simplified ChIP-exo assays. *Nat Commun*.

Roy S, Lagree S, Hou Z, Thomson JA, Stewart R, Gasch AP. 2013. Integrated Module and Gene-Specific Regulatory Inference Implicates Upstream Signaling Networks. *PLoS Comput Biol* **9**.

Ryan O, Shapiro RS, Kurat CF, Mayhew D, Baryshnikova A, Chin B, Lin ZY, Cox MJ, Vizeacoumar F, Cheung D, et al. 2012. Global gene deletion analysis exploring yeast filamentous growth. *Science (80-)*.

Schep AN, Buenrostro JD, Denny SK, Schwartz K, Sherlock G, Greenleaf WJ. 2015. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res*.

Schmidt F, Gasparoni N, Gasparoni G, Gianmoena K, Cadenas C, Polansky JK, Ebert P, Nordstrom K, Barann M, Sinha A, et al. 2017. Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res*.

Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, Jolma A, Zhong G, Guo H, Kanagalingam T, et al. 2016. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome Res*.

Shively CA, Liu J, Chen X, Loell K, Mitra RD. 2019. Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc Natl Acad Sci U S A*.

Siahpirani AF, Roy S. 2017. A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic Acids Res* **45**: 1–22.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**: 539. http://msb.embopress.org/content/7/1/539.abstract.

Sigalova O, Shaeiri A, Forneris M, Furlong E, Zaugg J. 2020. Predictive features of gene expression variation reveal a mechanistic link between expression variation and differential expression. 1–24.

Simeonov DR, Gowen BG, Boontanrart M, Roth TL, Gagnon JD, Mumbach MR, Satpathy AT, Lee Y, Bray NL, Chan AY, et al. 2017. Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*.

Singh R, Lanchantin J, Robins G, Qi Y. 2016. DeepChrome: Deep-learning for predicting gene expression from histone modifications. In *Bioinformatics*.

Skene PJ, Henikoff JG, Henikoff S. 2018. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. *Nat Protoc*.

Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife*.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res*.

Spies D, Renz PF, Beyer TA, Ciaudo C. 2019. Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Brief Bioinform*.

Spivak AT, Stormo GD. 2012. ScerTF: A comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res*.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, et al. 2019. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*.

Tasaki S, Gaiteri C, Mostafavi S, Wang Y. 2020. Deep learning decodes the principles of differential gene expression. *Nat Mach Intell*.

Teleman AA, Hietakangas V, Sayadian AC, Cohen SM. 2008. Nutritional Control of Protein Biosynthetic Capacity by Insulin via Myc in Drosophila. *Cell Metab*.

Tome JM, Tippens ND, Lis JT. 2018. Single-molecule nascent RNA sequencing identifies regulatory domain architecture at promoters and enhancers. *Nat Genet*.

Tosti L, Ashmore J, Tan BSN, Carbone B, Mistri TK, Wilson V, Tomlinson SR, Kaji K. 2018. Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo. *Genome Res*.

Tran LM, Brynildsen MP, Kao KC, Suen JK, Liao JC. 2005. gNCA: A framework for determining transcription factor activity based on transcriptome: Identifiability and numerical implementation. *Metab Eng*.

Van Nostrand EL, Kim SK. 2013. Integrative analysis of C. elegans modENCODE ChIP-seq

data sets to infer gene regulatory interactions. *Genome Res*.

Van Steensel B, Henikoff S. 2000. Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. *Nat Biotechnol*.

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, et al. 2011. A Comprehensive Genomic Binding Map of Gene and Chromatin Regulatory Proteins in Saccharomyces. *Mol Cell* **41**: 480–492. http://dx.doi.org/10.1016/j.molcel.2011.01.015.

Wagner F, Yan Y, Yanai I. 2017. K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data. *bioRxiv*.

Wang H, Heinz ME, Crosby SD, Johnston M, Mitra RD. 2008. "Calling Cards" method for high-throughput identification of targets of yeast DNA-binding proteins. *Nat Protoc*.

Wang H, Johnston M, Mitra RD. 2007. Calling cards for DNA-binding proteins. *Genome Res*.

Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. 2011a. Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res*.

Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. 2011b. Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res* **21**: 748–755.

Wang H, Mayhew D, Chen X, Johnston M, Mitra RD. 2012. "Calling Cards" for DNA-binding Proteins in mammalian Cells. *Genetics*.

Wang Y, Cho DY, Lee H, Fear J, Oliver B, Przytycka TM. 2018. Reprogramming of regulatory network using expression uncovers sex-specific gene regulation in Drosophila. *Nat Commun*.

Waryah CB, Moses C, Arooj M, Blancafort P. 2018. Zinc fingers, TALEs, and CRISPR systems: A comparison of tools for epigenome editing. In *Methods in Molecular Biology*.

Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, Wang H. 2019. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc Natl Acad Sci U S A* **116**: 5542–5549.

Weiner A, Hsieh TS, Rando OJ, Friedman N, Weiner A, Hsieh TS, Appleboim A, Chen H V, Rahat A, Amit I. 2015. High-Resolution Chromatin Dynamics during a Yeast Resource High-Resolution Chromatin Dynamics during a Yeast Stress Response. *Mol Cell* **58**: 371–386. http://dx.doi.org/10.1016/j.molcel.2015.02.002.

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. 2014. Determination and Inference of Eukaryotic

Transcription Factor Sequence Specificity. *Cell* **158**: 1431–1443. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4163041&tool=pmcentrez&rendertype=abstract.

Xie S, Duan J, Li B, Zhou P, Hon GC. 2017. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol Cell*.

Xin B, Rohs R. 2018. Relationship between histone modifications and transcription factor. 321–333.

Xu Y, Zhang Z, You L, Liu J, Fan Z, Zhou X. 2020. scIGANs: single-cell RNA-seq imputation using generative adversarial networks. *Nucleic Acids Res*.

Zeiler MD, Fergus R. 2012. Visualizing and Understanding Convolutional Networks.

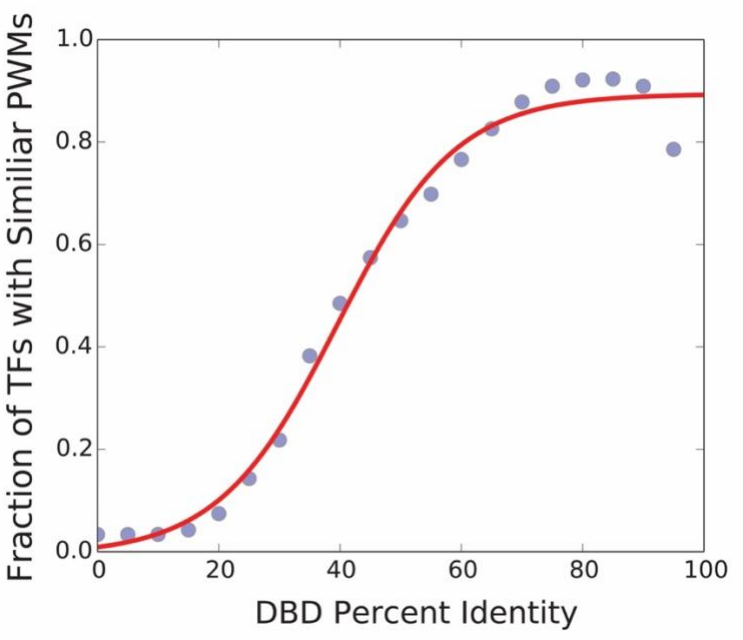
Zhong S, He X, Bar-Joseph Z. 2013. Predicting tissue specific transcription factor binding sites. *BMC Genomics*.

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet*.

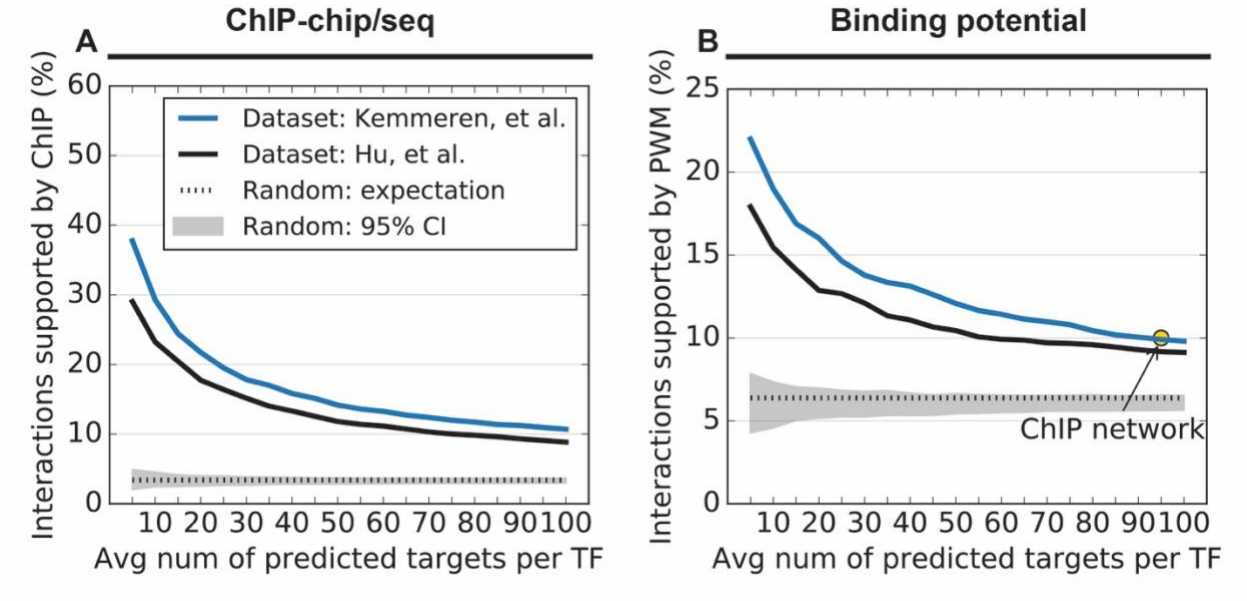
Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods*.

Zhou X, Cain CE, Myrthil M, Lewellen N, Michelini K, Davenport ER, Stephens M, Pritchard JK, Gilad Y. 2014. Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol*.

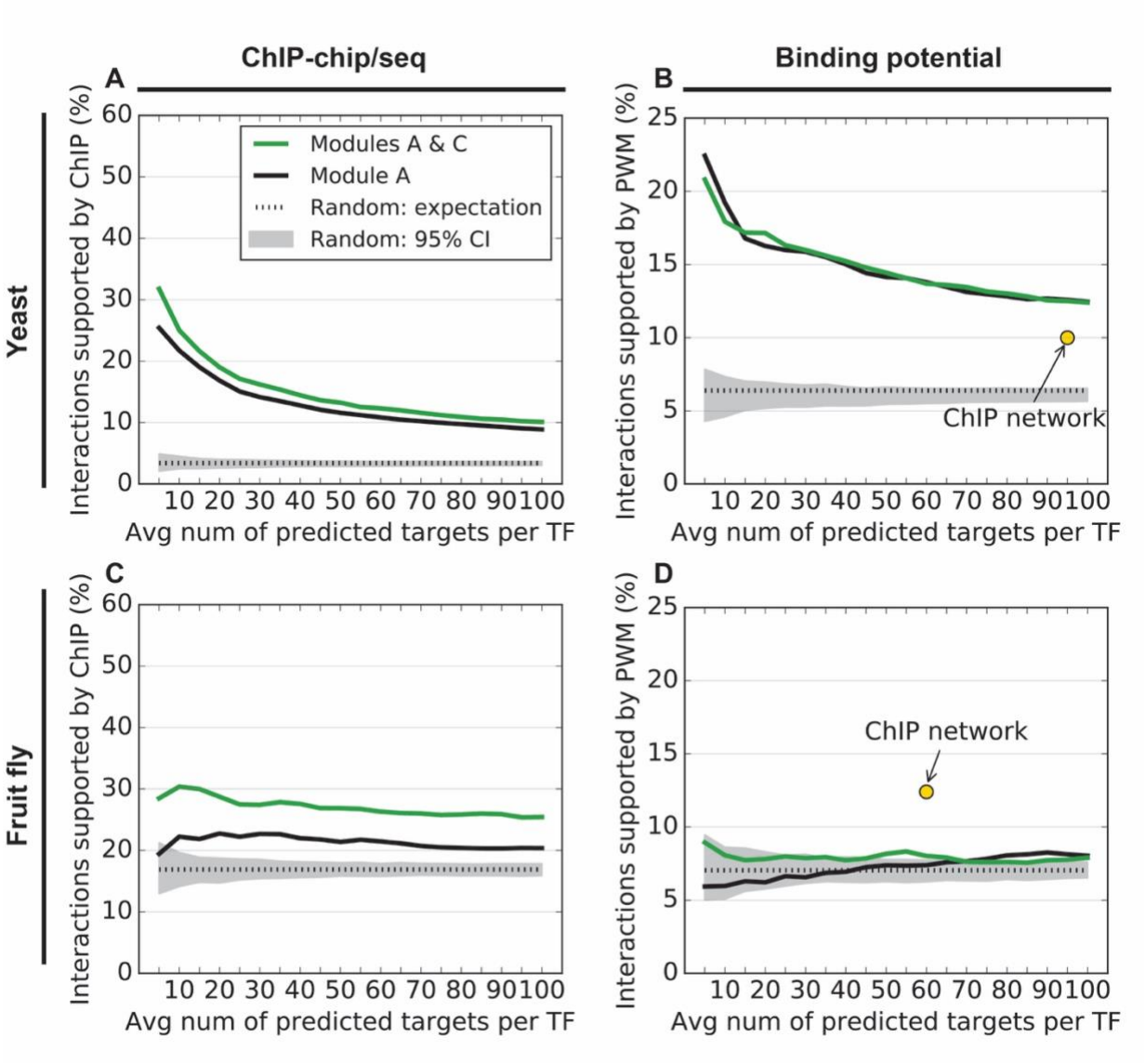
Appendix



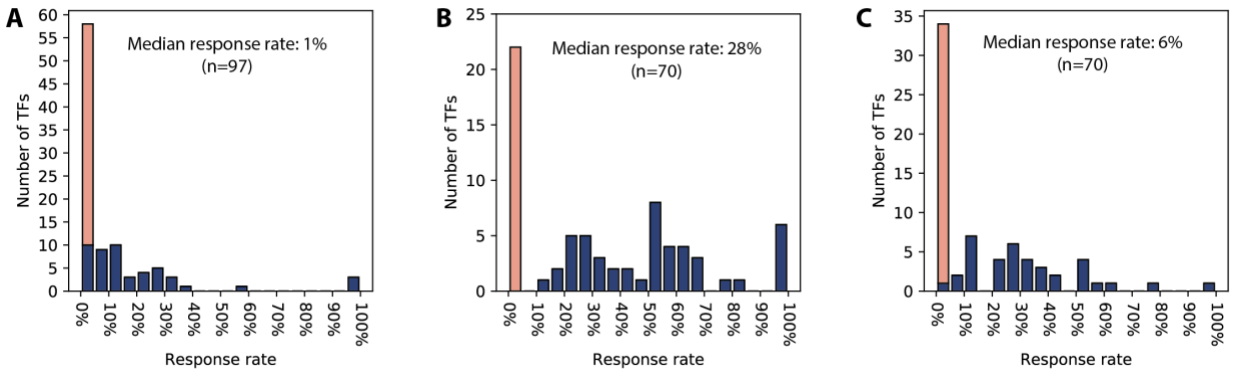
Supplemental Figure S2. 1: Relationship between the percent identities of DBD pairs and the similarities of their PWMs. Horizontal axis: the bins of percent identities (PIDs) of all pairs of DBDs. Each value on the horizontal axis represents the lower bond of the corresponding bin, e.g. if a pair of DBD are 82% identical, they are in the PID bin of 80%-85%. Vertical axis: the fraction of TFs in a certain PID bin that have similar known PWMs (Tomtom E-value < 1). Red line: a logistic fit.



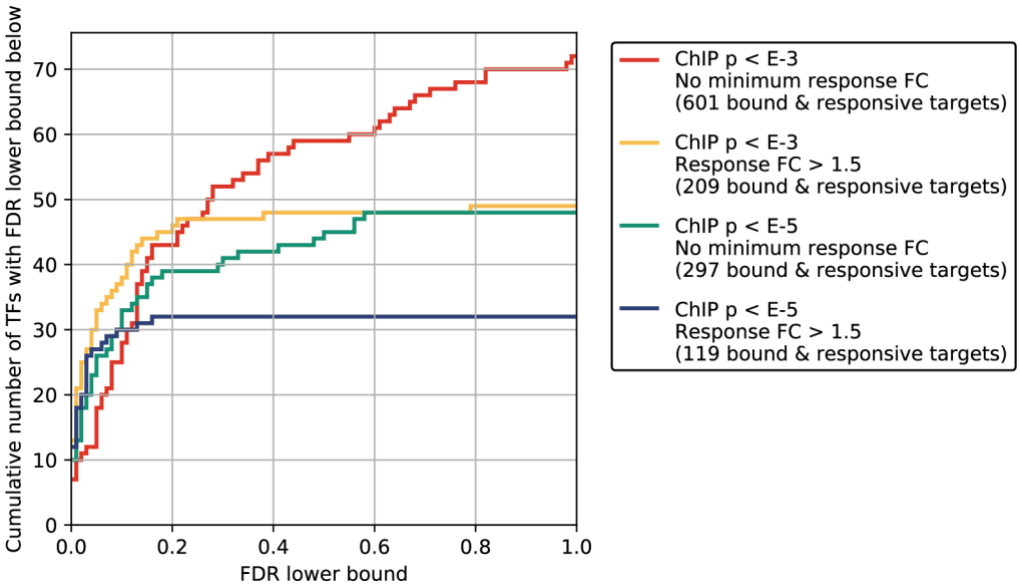
Supplemental Figure S2. 2: Accuracy of NetProphet 1.0 using two yeast expression data sets. The smaller dataset (Hu et al. 2007) contains 269 TF knockout strains (black line); the larger dataset (Kemmeren et al. 2014) contains 265 strains of single TF knockouts and 1,219 strains of other gene knockouts (blue line). (A) Vertical axis: Percentage of edges supported by ChIP data. Horizontal axis: number of top ranked edges included in the network per TF encoded in the genome. E.g., since there are 320 TFs in the yeast genome, "10" on the horizontal axis corresponds to a network with 3,200 edges. Dotted line: Expected accuracy of random networks. Gray area: 95% confidence interval for randomly selected networks. (B) Same as A for PWM support. The point labeled "ChIP network" indicates the number of ChIP-supported edges and the fraction of those edges that also have PWM support.



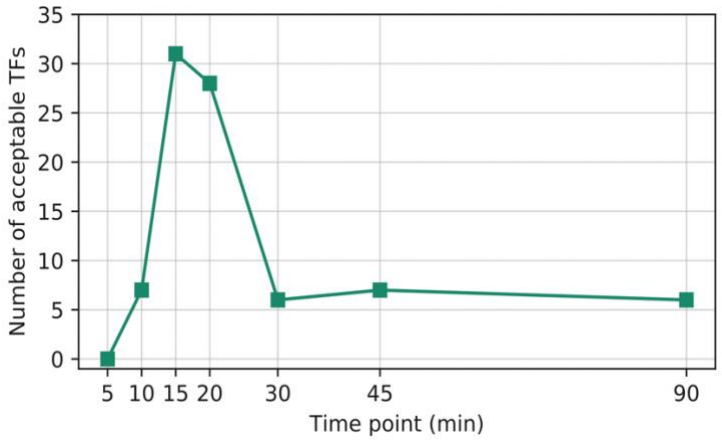
Supplemental Figure S2. 3: Effect of weighted averaging applied to BART network. Accuracy of BART network on yeast before weighted averaging (black line) or after weighted averaging (green line). Horizontal axis: number of top ranked edges included in the network per TF encoded in the genome. E.g., since there are 320 TFs in the yeast genome, "10" on the horizontal axis corresponds to a network with 3,200 edges. Vertical axis: Percentage of edges supported by ChIP data. Dotted line: Expected accuracy of random networks. Gray area: 95% confidence interval for randomly selected networks. (B) Same as A for PWM support. The point labeled "ChIP network" indicates the number of ChIP-supported edges and the fraction of those edges that also have PWM support. (C) Same as A for the fly data. (D) Same as B for the fly data, except that the vertical axis shows support by conserved PWM hits only.



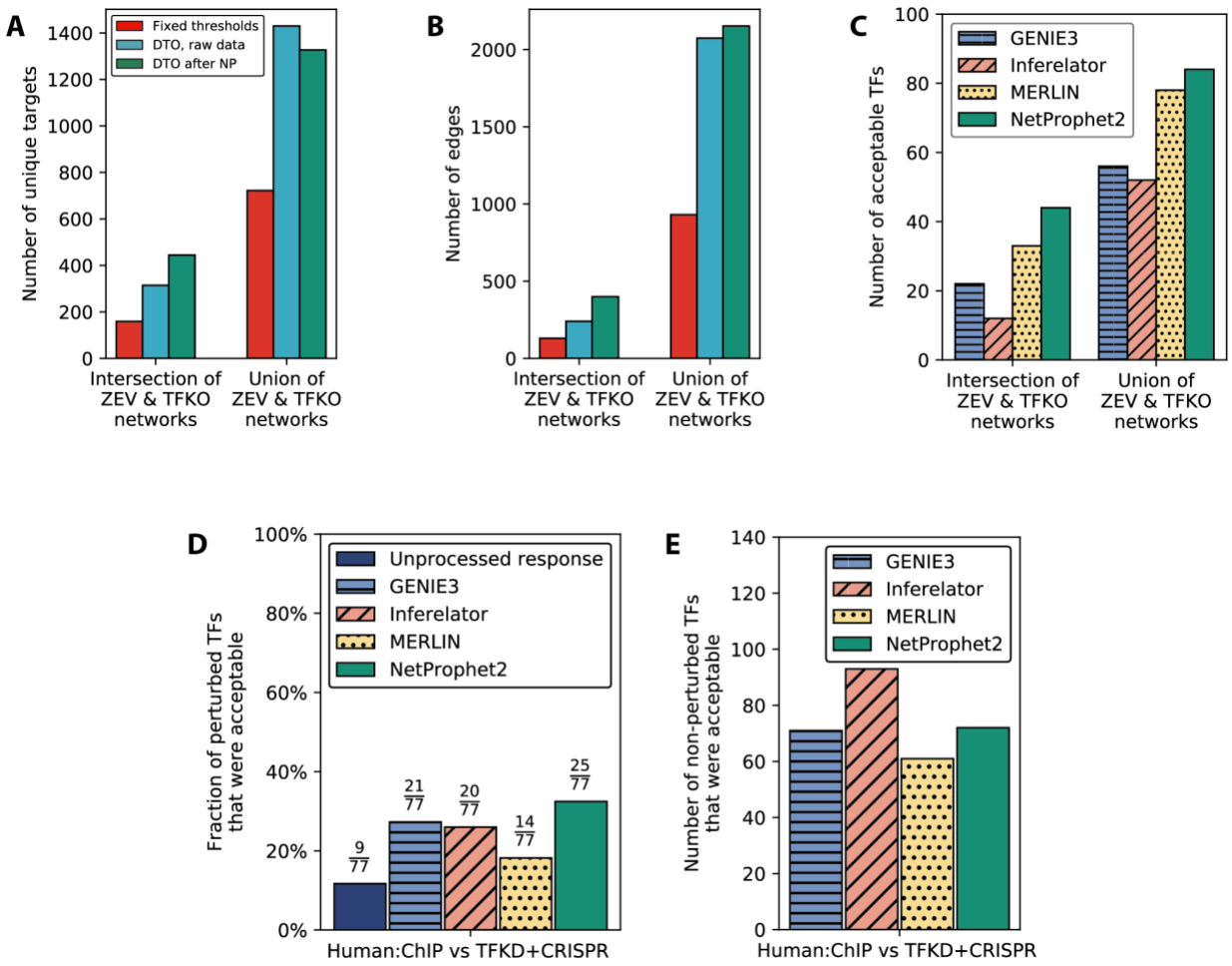
Supplemental Figure S3. 1: Overlap between bound and responsive gene sets at different thresholds. Same as Figure 3.1A except: (A) Binding threshold is $p<0.001$ and response threshold is $p<0.05$ with fold change > 1.5. Total bound and responsive genes is 209. (B) Binding threshold is $p<0.00001$ and response threshold is $p<0.05$ with no minimum fold change. Total bound and responsive genes is 297. (C) Binding threshold is $p<0.00001$ and response threshold is $p<0.05$ with fold change > 1.5. Total number of bound and responsive genes is 119.



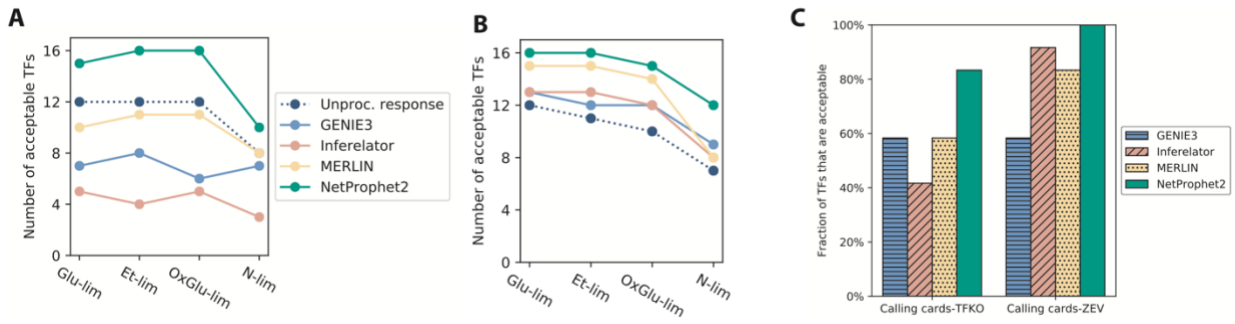
Supplemental Figure S3. 2: Cumulative number of TFs with expected FDR lower bound less than the number on the horizontal axis, assuming sensitivity of 80%. Red line: moderate binding and response thresholds; orange line: moderate binding threshold and tight response threshold; green line: tight binding threshold and moderate response threshold; blue line: tight binding and response thresholds.



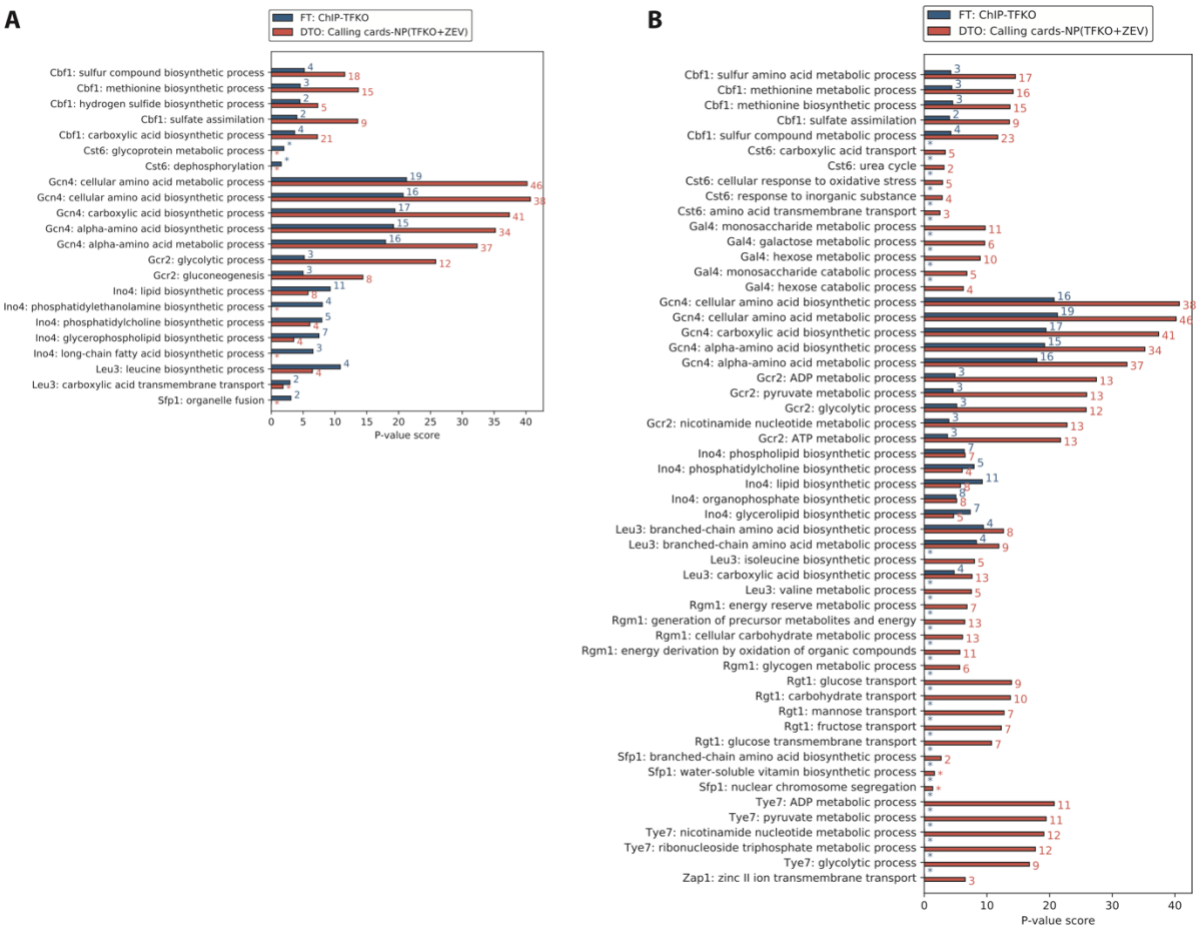
Supplemental Figure S3. 3: Numbers of acceptable TFs when comparing Harbison ChIP data to ZEV response data at various time points.



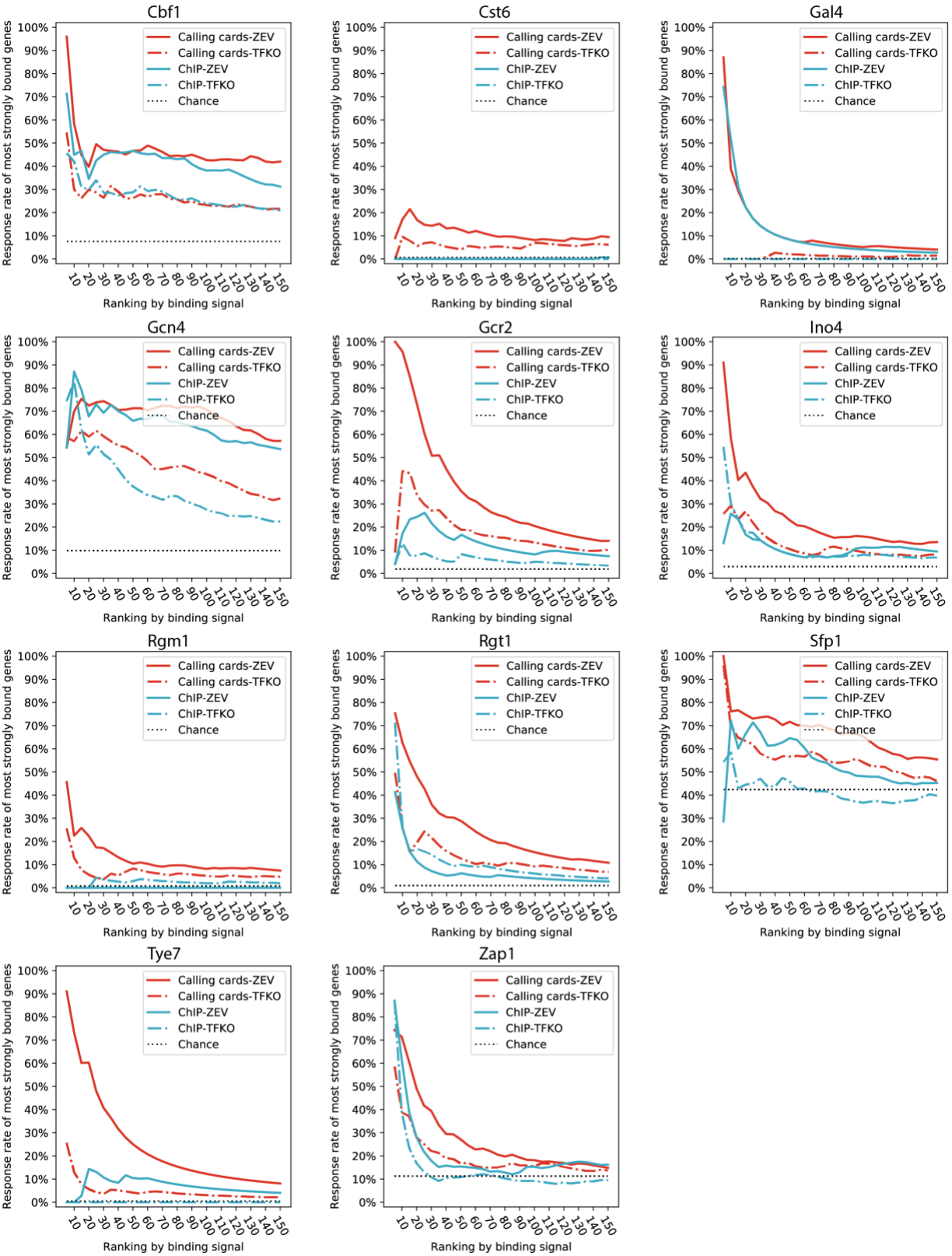
Supplemental Figure S3. 4: (A) and (B) Comparison of TFKO and ZEV15 networks derived from fixed thresholds, dual threshold optimization (DTO) on raw gene expression, and DTO on gene expression data processed by NetProphet 2.0. DTO on the raw expression data (blue bars) increases the size of the networks over fixed thresholds (red bars). This is true for both the intersection of the TFKO and ZEV networks (left bar grouping) and their union (right bar grouping). Post processing expression data with NetProphet 2.0 (green bars) further increases the size of the networks. See Figure 3.2 for the numbers of TFs showing acceptable convergence in these analyses. (C) Comparison of yeast networks derived from DTO on ChIP data and response data processed using several network inference algorithms (D) Comparison of human K562 networks derived from DTO on ChIP-seq and raw response data or network inference processed response data. Bar height is the fraction of TFs showing acceptable convergence divided by the number of TFs that were ChIPped and perturbed by either TFKD or CRISPR. Network inference postprocessing improves convergence over unprocessed response data. (E) Similar to (D), but considering only TFs that were not directly perturbed. These TFs cannot be analyzed for convergence without network inference.



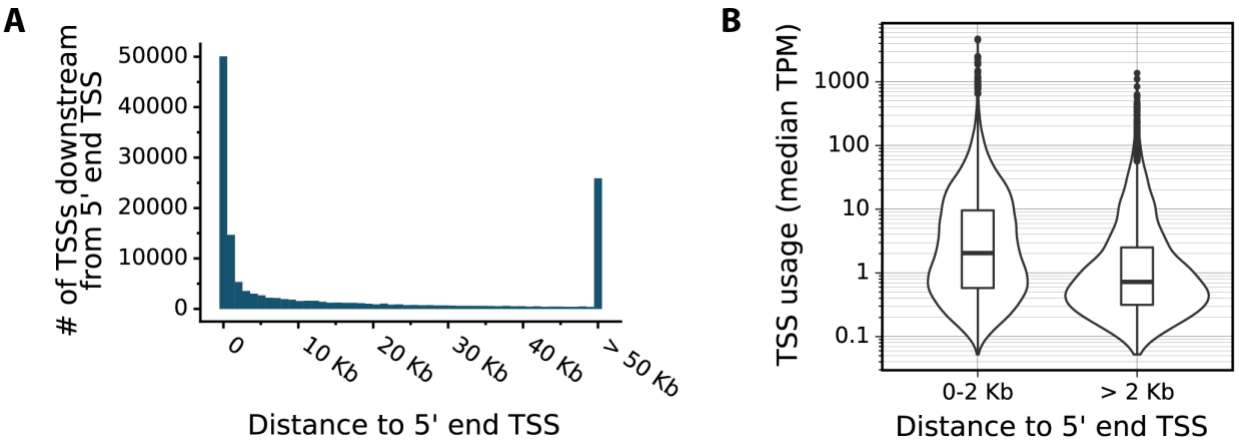
Supplemental Figure S3. 5: (A) Among 16 TFs for which we have data in TFKO and ChIP-exo in four nutrient limited conditions, the number of TFs that show convergence between ChIP-exo data and TFKO response data (dotted blue) or network inference processed response data (other colored lines). (B) Same as (A) except that ZEV15 data replaces TFKO data. (C) Comparison of networks derived from DTO on calling cards data and response data processed through several network inference algorithms.



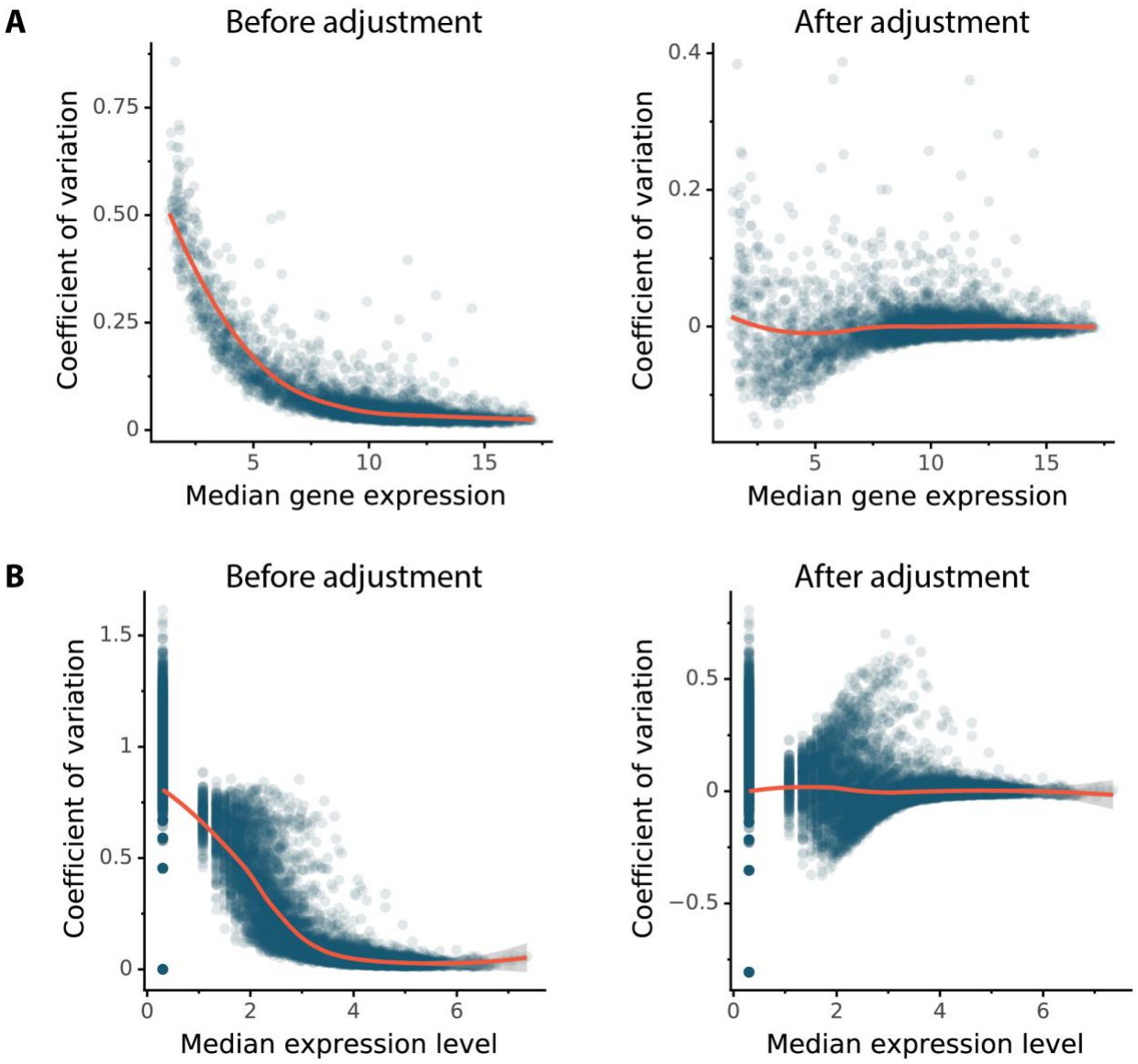
Supplemental Figure S3. 6: Same as Figure 3.4D except: (A) For each TF that has any enriched GO term, the five most enriched terms when the targets of each TF are chosen by using fixed thresholds on Harbison ChIP and TFKO data. In most cases these same terms are even more significantly enriched when the targets are chosen by using a different method -- dual threshold optimization comparing calling cards data to output from NetProphet 2.0 run on the TFKO and ZEV expression data (red bars). The numbers to the right of the bars indicate the number of genes with a given GO term among the targets of the TF. (B) For each TF that has any enriched GO term, the five most significantly enriched terms chosen by using dual threshold optimization comparing calling cards data to output from NetProphet 2.0 run on the TFKO and ZEV expression data. In many cases, terms with one or two fewer genes are more familiar than the terms with the most genes. For example, Gcr2 has 12 target genes annotated with "glycolytic process", corresponding to its accepted function, but the most significant term is "ADP metabolic process", which contains those 12 glycolytic genes plus one additional gene.


Cbf1
Cst6
Gal4
Gcn4
Gcr2
Ino4
Rgm1
Rgt1
Sfp1
Tye7
Zap1
Calling cards-ZEV
Calling cards-TFKO
ChIP-ZEV
ChIP-TFKO
Chance
Response rate of most strongly bound genes
Ranking by binding signal

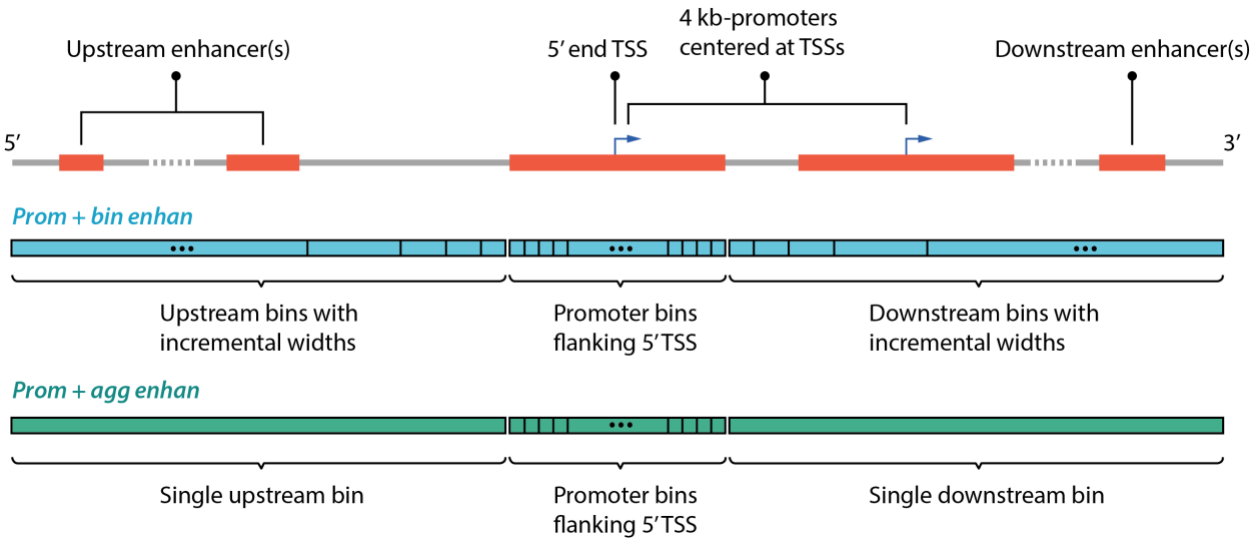
Supplemental Figure S3. 7: Same as Figure 3.5A. The fraction of most strongly TF-bound genes that are responsive to the perturbation of that TF, as a function of the number of most-strongly bound genes considered. Shown are the other 11 TFs, in addition to Leu3, that had data available in Harbison ChIP, transposon calling cards, TFKO, and ZEV15.



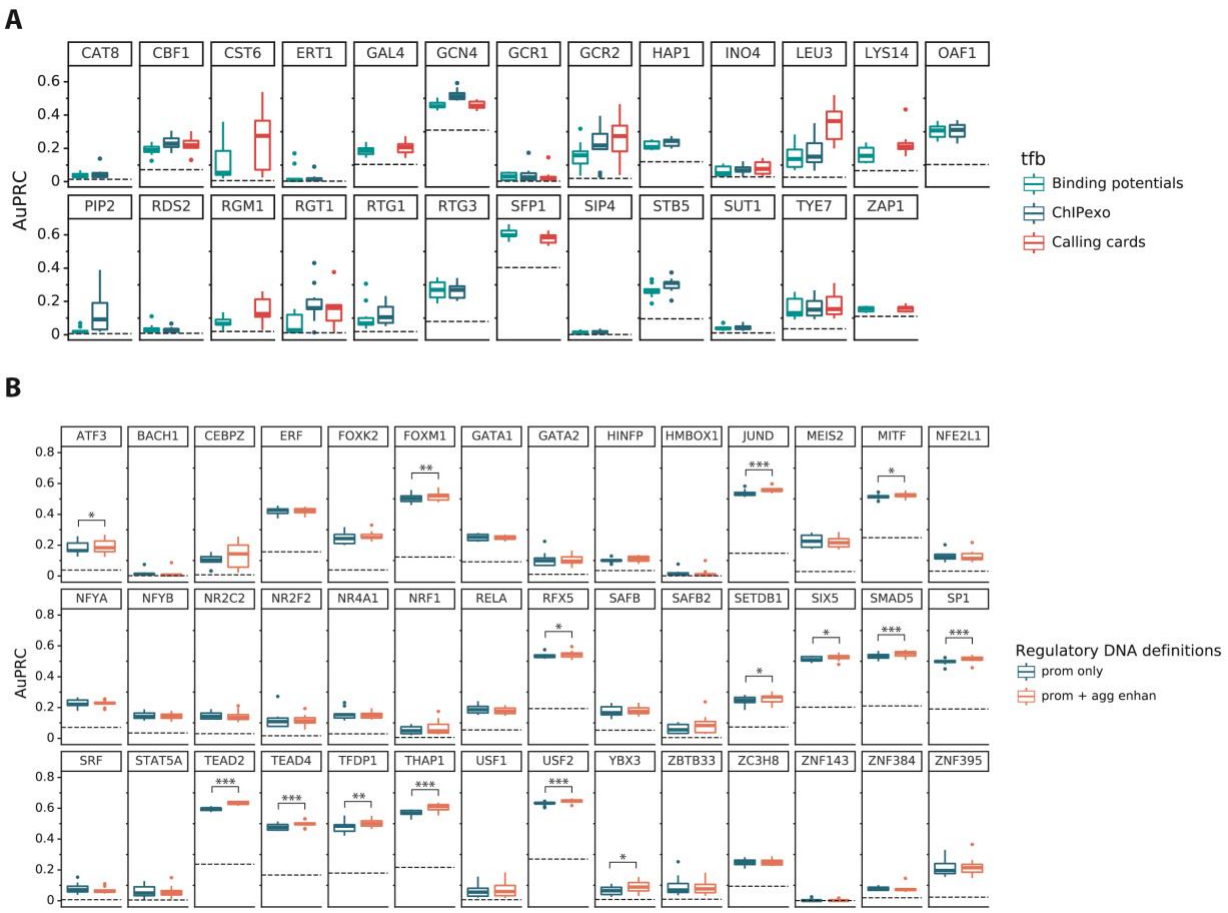
Supplemental Figure S4. 1: Statistics on human TSS. (A) Distances between each 5’ end TSS and other downstream TSSs of the corresponding gene. Among the downstream TSSs for all genes, ~47% of them are within the 2 Kb range of their paired 5’ end TSS. The median distance for the TSSs within 2 Kb range is 163 bp, while the median distance for those that are more than 2 Kb away is 26.3 Kb. (B) Relationship between TSS usage and distance. TSS usage for each TSS is represented as the median Tags Per Million (TPM) level across all samples in Fantom5 CAGE expression data (Forrest et al. 2014; Lizio et al. 2019). The median TPM for the TSSs within 2 Kb range to their corresponding 5’ end TSS (including these 5’ end TSS themselves) is approximately three times of the median TPM for the TSSs that are more than 2 Kb away from their corresponding 5’ end TSS.



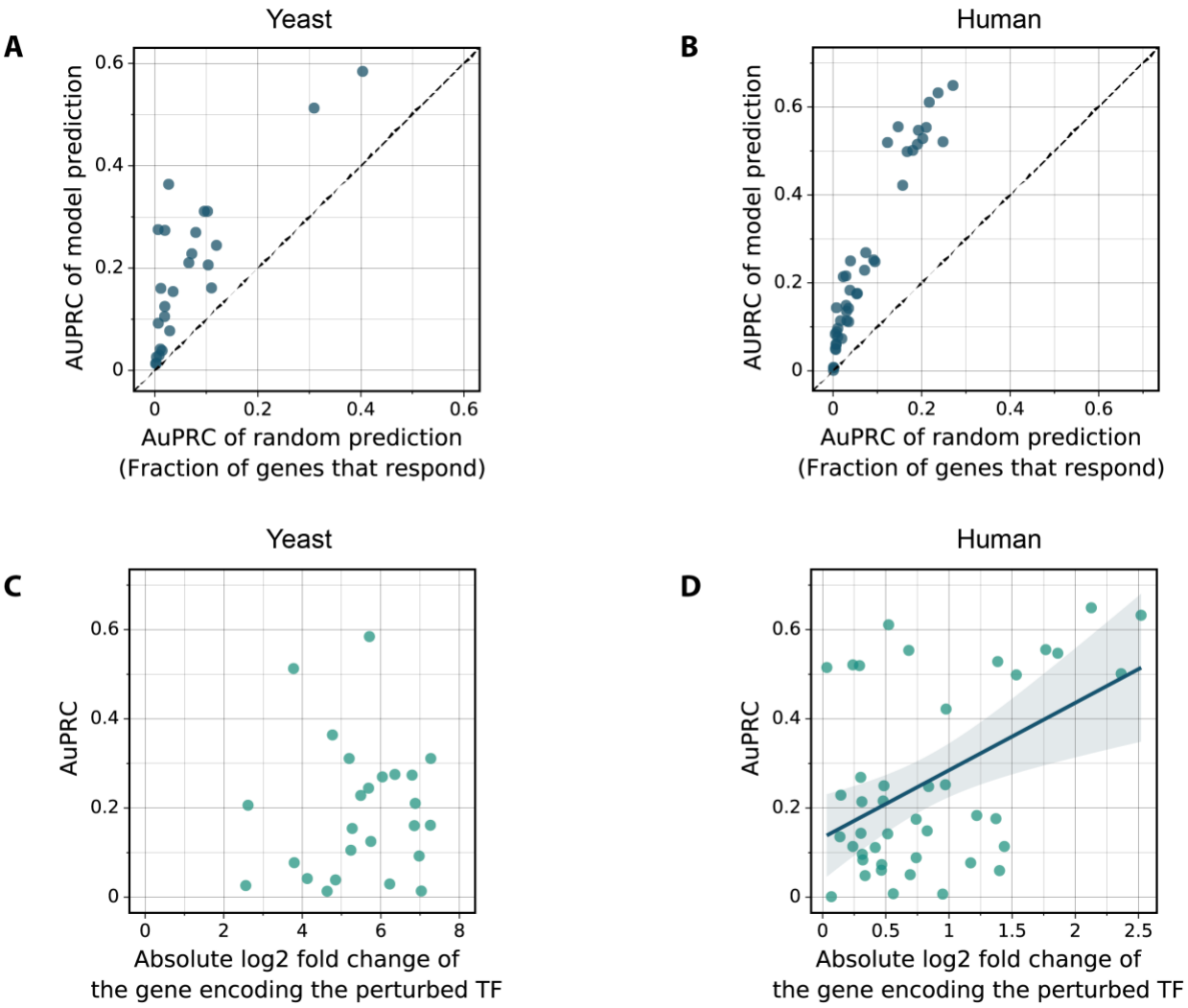
Supplemental Figure S4. 2: (A) Expression variation for yeast cells. Left: Relationship between the median expression level of each gene across pre-perturbation (or control) conditions and its expression variation measured by the coefficient of variation. Orange curve was fitted using locally estimated scatterplot smoothing (LOESS) regression. Right: Expression variation adjusted for the median expression level by taking the residual of LOESS regression (the orange curve from left). (B) Expression variation for human cells. Same analysis as in (A).



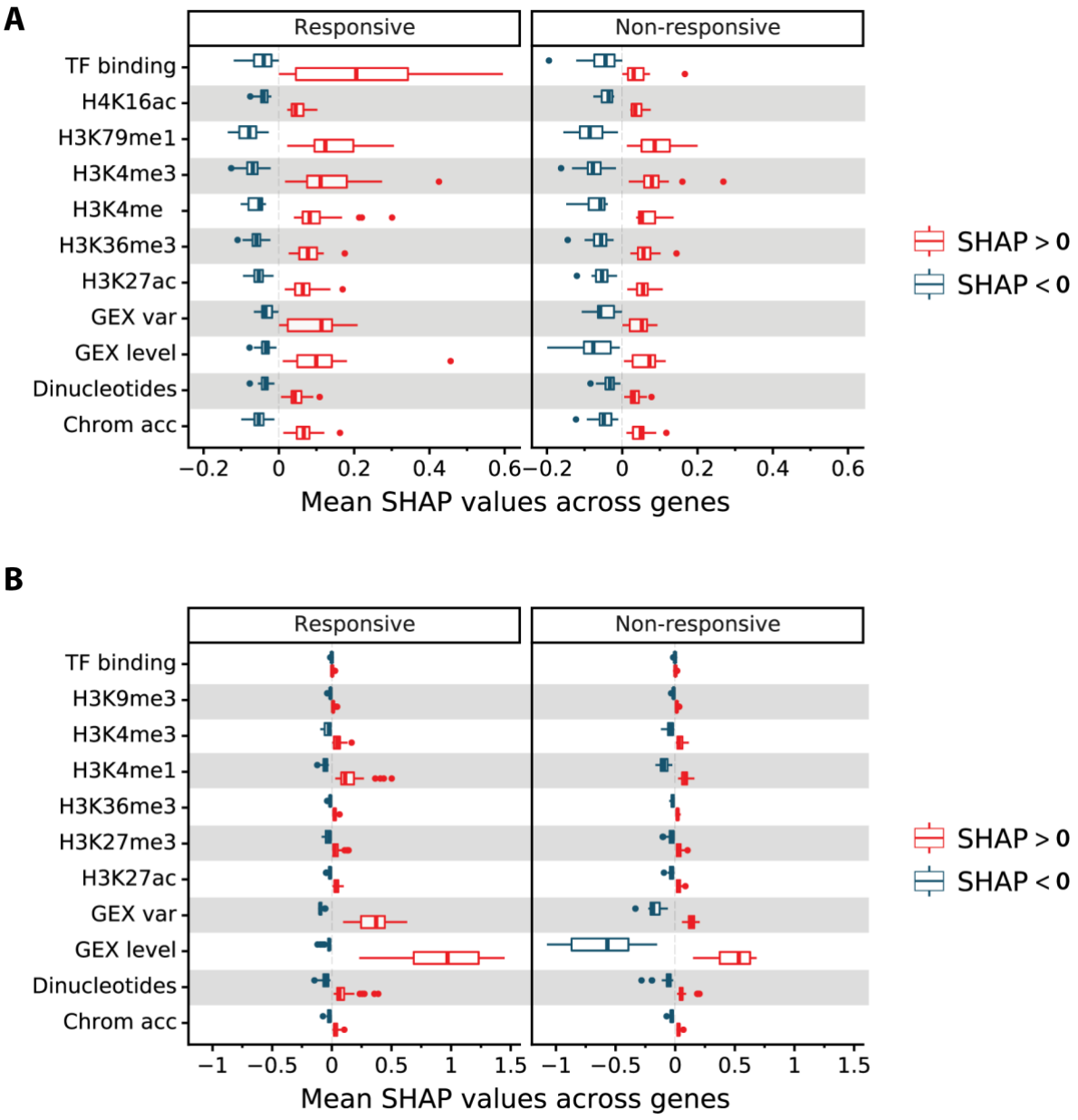
Supplemental Figure S4. 3: Definition of human cis-regulatory regions. The top panel illustrates a 1 Mb region centered at the 5' end TSS of a gene. Orange boxes indicate the enhancers linked to the gene, and the 4 Kb promoter(s) centered around the gene's TSS(s). The bottom two panels illustrate the approaches for binning the 1Mb cis-regulatory region. *Prom + bin enhan* (blue) includes 40 equal-sized bins of the promoter centered around the 5' TSS, and 45 bins with incremental widths for the upstream regions between -500 Kb and -2 Kb and another 45 bins for the downstream regions between 2 Kb and 500 Kb respectively. As the distance between the distal bin and TSS increases, the width of the bin increases exponentially. *Prom + agg enhan* (green) includes 40 equal-sized bins of the promoter centered around the 5' TSS, one single upstream bin covering the entire region between -500 Kb and -2 Kb, and one single downstream bin covering the region between 2 Kb and 500 Kb. The signals of each coordinate-dependent feature that are mapped to the defined cis-regulatory regions (orange, top panel) linked to the corresponding bins according to the genomic coordinates. Within each bin, the signals are summed into a single aggregated input value.



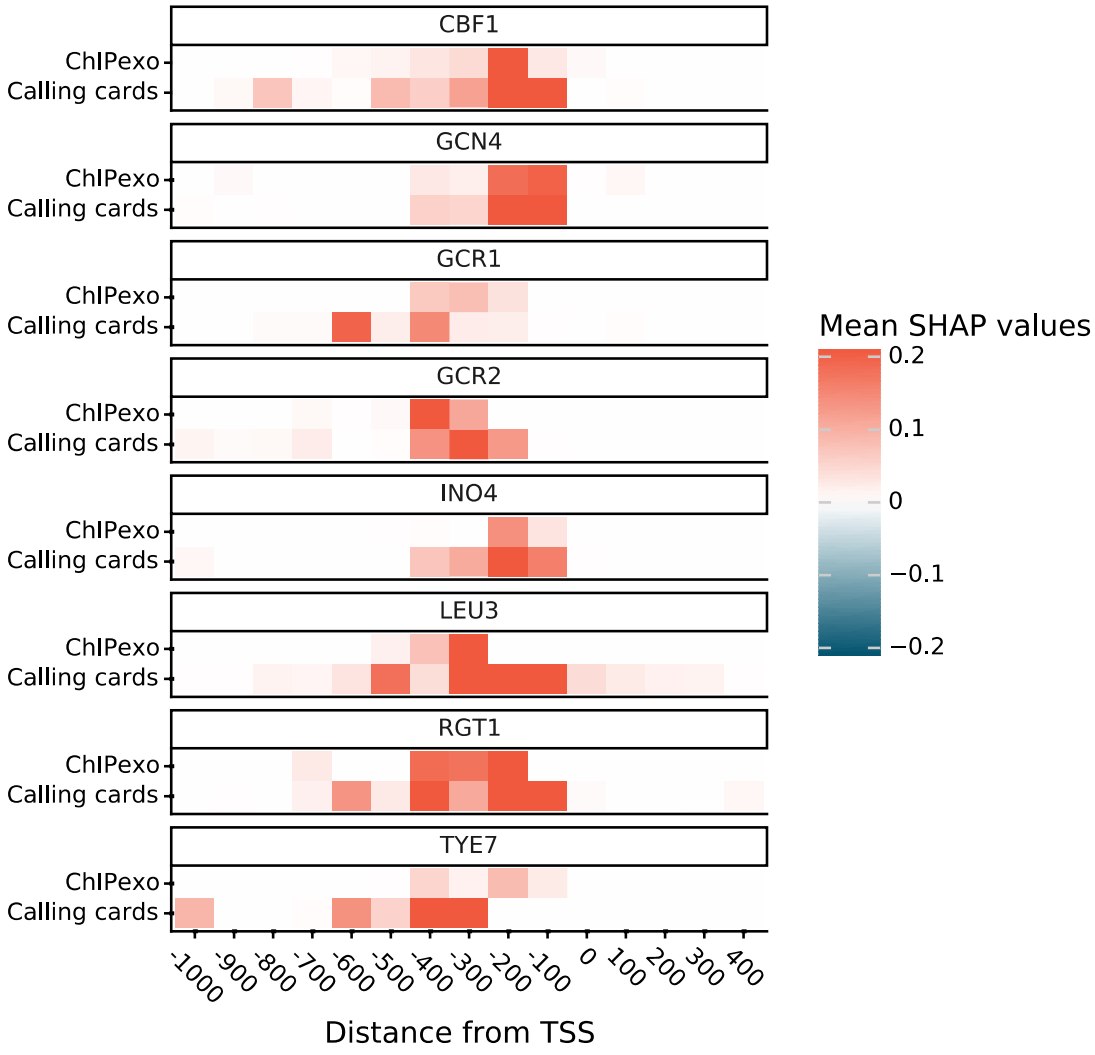
Supplemental Figure S4. 4: (A) Performance of individual yeast TF models that were trained on different types of TF binding data. No boxplot is shown if the TF binding data from the corresponding assay was unavailable. Each boxplot shows the results of ten-fold cross-validation on all genes. (B) Performance of individual human TF models that were trained using various definitions of regulatory DNA. Statistical significance used paired t-test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$(***).



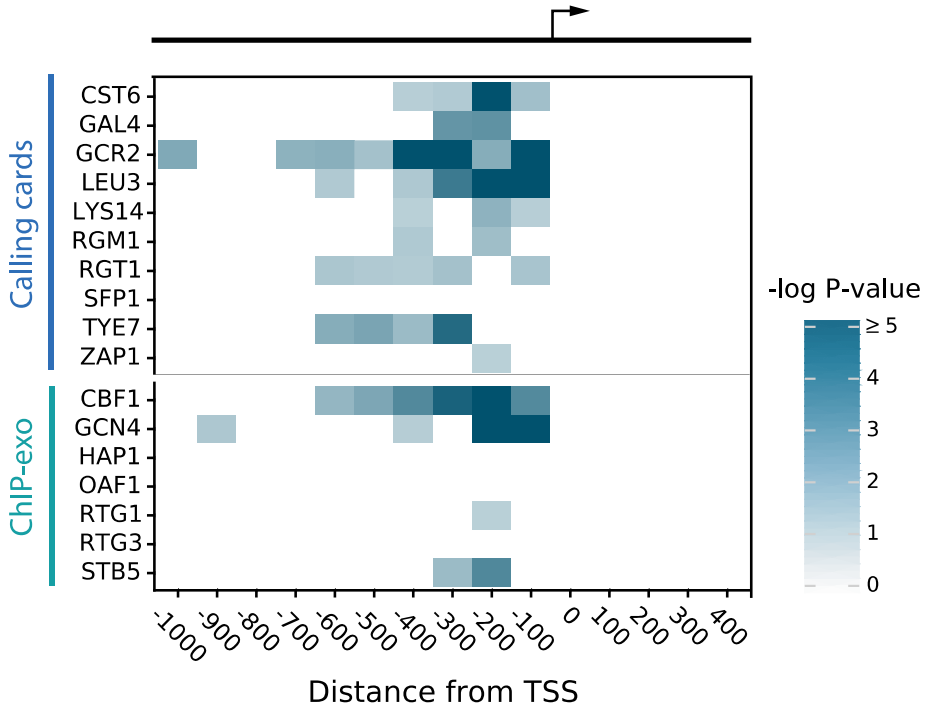
Supplemental Figure S4. 5: (A) Relationship between AUPRC of random prediction and AUPRC of model prediction. Dashed diagonal line has slope of 1. (B) Same as (A) but for human K562 TFs. (C) Relationship between the log fold change of the mRNA for the perturbed TF and model accuracy, for yeast. Pearson correlation = 0.06, *P* = 0.76. (D) Same as (B) but for human TFs. Pearson correlation = 0.47, *P* = 0.002.



Supplemental Figure S4. 6: (A) Mean SHAP values for all responsive and unresponsive targets of each yeast TF perturbation. (B) Mean SHAP values for all responsive and unresponsive targets of human TF perturbations.



Supplemental Figure S4. 7: Comparison of the influence of yeast TF binding data generated from two types of assays: transponson calling cards and ChIP-exo. Each pixel is the mean SHAP values of all target genes that were bound by the perturbed TFs.



Supplemental Figure S4. 8: Heatmap of the statistics that indicate the degree to which the bound but unresponsive genes have insufficient TF occupancy. The P-values for all bins in each row (TF) were estimated using the method in Figure 3C. Bins with P-values no less than 0.05 are in blank.

Supplemental Table S4. 1: Selection of histone modifications

	Literatures			Data Availability	
	Karlić, 2010	Zhou, 2014; González, 2015	Kundaje, 2015; Singh, 2016	Yeast (Weiner, 2015)	Human K562 (ENCODE, 2020)
H3K27ac	1	1		1	1
H3K27me3		1	1		1
H3K36me3	1		1	1	1
H3K4me1		1	1	1	1
H3K4me3	1	1	1	1	1
H3K79me1	1			1	
H3K9me3			1		1
H4K16ac	1			1	

Supplemental Table S4. 2: Yeast TF models.

TF	Binding dataset	AUPRC	For SHAP analysis
YLR403W (SFP1)	Calling cards	0.58446718	TRUE
YEL009C (GCN4)	ChIP-exo	0.51278158	TRUE
YLR451W (LEU3)	Calling cards	0.36390632	TRUE
YHR178W (STB5)	ChIP-exo	0.3110723	TRUE
YAL051W (OAF1)	ChIP-exo	0.31106409	TRUE
YIL036W (CST6)	Calling cards	0.27532851	TRUE
YNL199C (GCR2)	Calling cards	0.27381191	TRUE
YBL103C (RTG3)	ChIP-exo	0.26980529	TRUE
YLR256W (HAP1)	ChIP-exo	0.24455603	TRUE
YJR060W (CBF1)	ChIP-exo	0.22821206	TRUE
YDR034C (LYS14)	Calling cards	0.21046638	TRUE
YPL248C (GAL4)	Calling cards	0.20618276	TRUE
YJL056C (ZAP1)	Calling cards	0.16135172	TRUE
YKL038W (RGT1)	Calling cards	0.16046672	TRUE
YOR344C (TYE7)	Calling cards	0.15421627	TRUE
YMR182C (RGM1)	Calling cards	0.12508439	TRUE
YOL067C (RTG1)	ChIP-exo	0.10535961	TRUE
YEL009C (GCN4)	Calling cards	0.45564519	FALSE
YJR060W (CBF1)	Calling cards	0.21908991	FALSE
YNL199C (GCR2)	ChIP-exo	0.21725465	FALSE
YKL038W (RGT1)	ChIP-exo	0.16007984	FALSE
YLR451W (LEU3)	ChIP-exo	0.1505335	FALSE
YOR344C (TYE7)	ChIP-exo	0.1503378	FALSE
YOR363C (PIP2)	ChIP-exo	0.09239194	FALSE
YOL108C (INO4)	Calling cards	0.07745282	FALSE
YOL108C (INO4)	ChIP-exo	0.06640978	FALSE
YGL162W (SUT1)	ChIP-exo	0.04185123	FALSE
YMR280C (CAT8)	ChIP-exo	0.03881355	FALSE
YPL133C (RDS2)	ChIP-exo	0.029646	FALSE
YPL075W (GCR1)	ChIP-exo	0.02614397	FALSE
YPL075W (GCR1)	Calling cards	0.02100747	FALSE
YBR239C (ERT1)	ChIP-exo	0.01386913	FALSE
YJL089W (SIP4)	ChIP-exo	0.01344086	FALSE

Supplemental Table S4. 3: Human TF models.

TF	AUPRC	For SHAP analysis
ENSG00000105698 (USF2)	0.64906387	TRUE
ENSG00000074219 (TEAD2)	0.63229787	TRUE
ENSG00000131931 (THAP1)	0.61089246	TRUE
ENSG00000130522 (JUND)	0.55512811	TRUE
ENSG00000113658 (SMAD5)	0.55354296	TRUE
ENSG00000143390 (RFX5)	0.54710206	TRUE
ENSG00000177045 (SIX5)	0.52840098	TRUE
ENSG00000187098 (MITF)	0.52106216	TRUE
ENSG00000111206 (FOXM1)	0.51931609	TRUE
ENSG00000185591 (SP1)	0.51506909	TRUE
ENSG00000198176 (TFDP1)	0.50122105	TRUE
ENSG00000197905 (TEAD4)	0.49858104	TRUE
ENSG00000105722 (ERF)	0.42194973	TRUE
ENSG00000143379 (SETDB1)	0.26877678	TRUE
ENSG00000102145 (GATA1)	0.25225519	TRUE
ENSG00000141568 (FOXK2)	0.24994403	TRUE
ENSG00000144161 (ZC3H8)	0.24814691	TRUE
ENSG00000001167 (NFYA)	0.22888833	TRUE
ENSG00000134138 (MEIS2)	0.21580418	TRUE
ENSG00000186918 (ZNF395)	0.2140645	TRUE
ENSG00000162772 (ATF3)	0.18328793	TRUE
ENSG00000173039 (RELA)	0.17614689	TRUE
ENSG00000160633 (SAFB)	0.17497082	TRUE
ENSG00000123358 (NR4A1)	0.14874691	TRUE
ENSG00000115816 (CEBPZ)	0.14337738	TRUE
ENSG00000120837 (NFYB)	0.14218726	TRUE
ENSG00000177463 (NR2C2)	0.1354988	TRUE
ENSG00000185551 (NR2F2)	0.11380771	TRUE
ENSG00000082641 (NFE2L1)	0.11378921	TRUE
ENSG00000172273 (HINFP)	0.11147308	TRUE
ENSG00000179348 (GATA2)	0.09625769	FALSE
ENSG00000060138 (YBX3)	0.08859001	FALSE
ENSG00000130254 (SAFB2)	0.08360437	FALSE
ENSG00000177485 (ZBTB33)	0.07704929	FALSE
ENSG00000126746 (ZNF384)	0.07324795	FALSE
ENSG00000112658 (SRF)	0.06051329	FALSE

ENSG00000158773 (USF1)	0.05962755	FALSE
ENSG00000126561 (STAT5A)	0.05056387	FALSE
ENSG00000106459 (NRF1)	0.0484205	FALSE
ENSG00000147421 (HMBOX1)	0.00781738	FALSE
ENSG00000156273 (BACH1)	0.00700893	FALSE
ENSG00000166478 (ZNF143)	0.00116493	FALSE