Summer 8-15-2022

TFA inference: Using mathematical modeling of gene expression data to infer the activity of transcription factors

Cynthia Ma
Washington University in St. Louis

Recommended Citation

Ma, Cynthia, "TFA inference: Using mathematical modeling of gene expression data to infer the activity of transcription factors" (2022). *McKelvey School of Engineering Theses & Dissertations*. 794.
https://openscholarship.wustl.edu/eng_etds/794

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Dissertation Examination Committee:
Michael Brent, Chair
Richard Bonneau
Jeremy Buhler
Barak Cohen
Roman Garnett

TFA Inference: Using Mathematical Modeling of Gene Expression Data to Infer the Activity of Transcription Factors

by
Cynthia Zhou Ma

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

August 2022
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor, Professor Michael Brent, for his guidance and patience through my PhD study. This would not have been possible without his consistent enthusiasm for learning and discovery. My thanks also to my committee members, Professor Richard Bonneau, Professor Jeremy Buhler, Professor Barak Cohen, and Professor Roman Garnett for taking the time to evaluate my work and provide advice from their unique perspectives.

To my lab mates, thank you for your support, and special thanks to Dr. Yiming Kang, Dhoha Abid, Jeff Jung, Lisa Liao, Chase Mateusiak, Sandeep Acharya, and Holly Brown for their help with producing, analyzing, and understanding all the data I've had the privilege to work on.

Thank you to friends, including Lynn Ma and Jessie Wu, who have been part of my life for nearly two decades, and Melody Li, who I met here in St. Louis. You've given me the opportunity to make memories outside of research, and I'm grateful.

Lastly, I would like to thank my mom, Qin Zhou, my dad, Haiching Ma, and my brother, Kevin Ma, who have been patient and supportive through my many years of doctoral study. Finally, we made it!

Cynthia Zhou Ma

Washington University in St. Louis

August 2022

Dedicated to my parents.

ABSTRACT OF THE DISSERTATION

TFA Inference: Using Mathematical Modeling of Gene Expression Data
to Infer the Activity of Transcription Factors

by

Cynthia Zhou Ma

Doctor of Philosophy in Computer Science

Washington University in St. Louis, 2022

Professor Michael Brent, Chair

Transcription factors (TFs) are a set of proteins that play a key role in the information processing system that enables a cell to respond to changes in internal and external state. By binding near a gene in a cell's DNA, a TF can influence that gene's expression level, triggering the appropriate increase or decrease in production levels of proteins that are needed to handle stressors like a change in nutrient availability or damage to the cell's internal structures. Transcription factor activity (TFA) is a measure of how much effect a TF has on its target genes in a given sample of cells. TFA depends on several factors including expression of the gene that encodes the TF, the TF's access to genes, and how much of the TF protein has the modifications needed to activate it. Because there are so many molecular factors influencing TF activity, there is no one assay that can measure TFA directly.

In this dissertation, we build on previous work in TFA inference that uses the measurable output of cell signaling pathways – gene expression levels – to infer TFA values and to utilize these inferred values to better understand the roles of individual TFs within gene regulatory systems. First, we applied TFA inference to microarray data on the well-studied *Saccharomyces* *cerevisiae* (baker's yeast) in order to define systematic, objective accuracy metrics. With these metrics, we explore the robustness of TFA inference to changes in the studied organism, the type of data input, and the optimization approach. Finally, we optimize the TFA inference algorithm to study RNA-seq data from a pathogenic yeast, *Cryptococcus neoformans,* to analyze the signaling pathway involved in its capsule formation response to environmental stress, a major factor of its virulence in humans.

Chapter 1: Introduction

1.1 What is TFA

As an analogy, one can think of a single cell as an automaton, and in its core, the nucleus, a cell holds a master file of blueprints, the DNA. Each blueprint, or gene, contains instructions for building a protein, and the more blueprint copies being transcribed, i.e. the higher genes are being expressed, the more copies of that protein will be built. Though many proteins are best understood as having a simple function like "transports salt out of the nucleus" or "transports calcium into the cell", some proteins function as part of a cell's information processing system, which enables the cell to respond to changes in external and internal state. Transcription factors (TFs) are proteins that play a key role in these signaling pathways. By binding near a gene in a cell's DNA, a TF can influence that gene's expression level, triggering the appropriate increase or decrease in production levels of proteins that are needed to handle stressors like a change in nutrient availability or damage to the cell's internal structures.

Transcription factor activity (TFA) is a measure of how much effect a TF has on its target genes in a given sample of cells. As external signals are propagated through the information processing networks inside of cells, knowing how TFA levels change during that process would greatly clarify the role that specific TFs play in specific signaling pathways. However, TFA depends on several factors including expression of the gene that encodes the TF, the TF's localization to the cell's nucleus where it has access to genes, and how much of the TF protein has the post-translational modifications needed to activate it (Fig.1.1). Because there are so many molecular factors influencing TF activity, there is no one assay that can measure TFA directly, and the expression levels of genes that encode TFs is a poor substitute. The research in this dissertation builds on previous work in TFA inference that uses the measurable output of cell signaling pathways - gene expression levels of TF target genes - to infer TFA values and to utilize these inferred values to better understand the roles of individual TFs within gene regulatory systems.

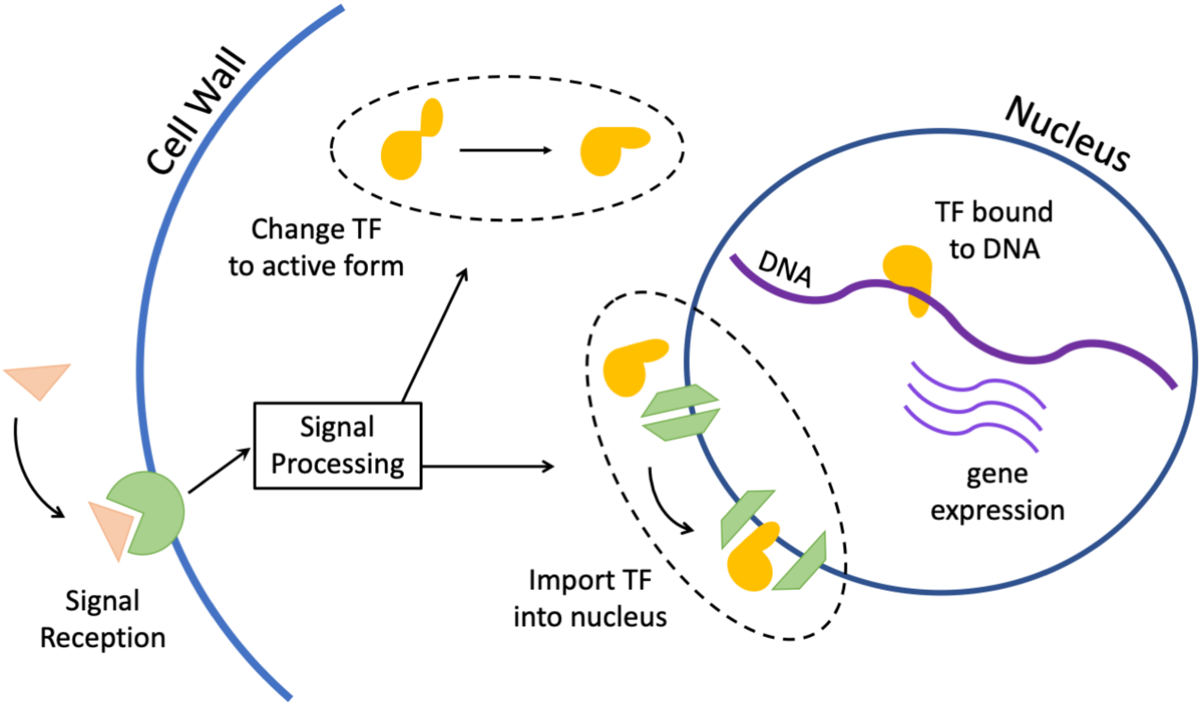


Figure 1.1 Example of signaling in cells. A signal molecule is recognized by a receptor on the cell's wall. As part of how that signal is processed, the activities of TFs are increased through form change and transportation into the nucleus. Active TF proteins bind upstream of a target gene in the DNA, triggering a change in that gene's expression.

1.2 Prior work in TFA inference

There are many algorithms that infer TFA from target gene expression data (Alvarez, *et al.*, 2016; Balwierz, *et al.*, 2014; Barenco, *et al.*, 2006; Boorsma, *et al.*, 2008; Boulesteix and

Strimmer, 2005; Chen, *et al*., 2013; Chen, *et al*., 2017; Cheng, *et al*., 2007; Cokus, *et al*., 2006; Fröhlich, 2015; Fu, *et al*., 2011; Gao, *et al*., 2004; Garcia-Alonso, *et al*., 2018; Gitter, *et al*., 2013; Jiang, *et al*., 2015; Khanin, *et al*., 2007; Li, *et al*., 2014; Liao, *et al*., 2003; Nachman, *et al*., 2004; Ocone and Sanguinetti, 2011; Sanguinetti, *et al*., 2006; Schacht, *et al*., 2014; Shi, *et al*., 2009; Tchourine, *et al*., 2018; Yu and Li, 2005; Zhu, *et al*., 2013). When successful, TFA inference provides insights into which TFs are involved in the transcriptional response to a given stimulus, such as a drug, extracellular signal, or nutrient influx. In principle, TF activity inference could also be used to predict the transcriptomic effects of direct perturbations of TF activity levels, as well as to improve TF network mapping (Arrieta-Ortiz, *et al*., 2015; Barenco, *et al*., 2009; Bussemaker, *et al*., 2017; Bussemaker, *et al*., 2001; Cokus*, et al.*, 2006; Fu*, et al.*, 2011; Gao*, et al.*, 2004; Gitter*, et al.*, 2013; Lam, *et al*., 2016; Lee and Bussemaker, 2010; Nachman*, et al.*, 2004; Shi*, et al.*, 2009; Tchourine*, et al.*, 2018; Wang, *et al*., 2008; Yang, *et al*., 2005; Yu and Li, 2005). However, these prior works have not established large-scale, systematic accuracy metrics that allow for objective evaluation of TFA inference results. This holds back computational researchers from improving their algorithms, and biology researchers from drawing new conclusions about gene regulatory pathways with confidence.

Most algorithms for inferring TF activity from gene expression data fit the parameters of a mathematical model to the expression data (Balwierz*, et al.*, 2014; Barenco*, et al.*, 2009; Boulesteix and Strimmer, 2005; Bussemaker*, et al.*, 2001; Chen*, et al.*, 2017; Cokus*, et al.*, 2006; Fröhlich, 2015; Fu*, et al.*, 2011; Jiang*, et al.*, 2015; Khanin*, et al.*, 2007; Li*, et al.*, 2014; Liao*, et al.*, 2003; Nachman*, et al.*, 2004; Sanguinetti*, et al.*, 2006; Schacht*, et al.*, 2014; Tchourine*, et al.*, 2018; Yu and Li, 2005). These models include parameters representing TFA levels, the values of which vary from one biological sample to another. Some models also include parameters that are constant across samples but vary as a function of the TF and target gene. These parameters reflect factors such as the affinity of the TF for sites in the promoter of each gene. In some approaches, these parameters, called *control strengths* (CSs), are obtained directly from TF binding data (Cokus*, et al.*, 2006; Gao*, et al.*, 2004; Jiang*, et al.*, 2015; Li*, et al.*, 2014) or by scanning models of TF binding specificity across promoters (Balwierz*, et al.*, 2014; Bussemaker*, et al.*, 2001; Conlon, *et al*., 2003). In other approaches, they are treated as unknowns and obtained by fitting the model to gene expression data (Boulesteix and Strimmer, 2005; Fu*, et al.*, 2011; Khanin*, et al.*, 2007; Liao*, et al.*, 2003; Nachman*, et al.*, 2004; Sanguinetti*, et al.*, 2006; Yu and Li, 2005). In this case, gene expression is typically modeled as depending linearly on the TFA levels and on the CSs (Boulesteix and Strimmer, 2005; Fu*, et al.*, 2011; Liao*, et al.*, 2003; Yu and Li, 2005), so the model as a whole is bilinear. More highly parameterized, non-linear models that more closely reflect the underlying biochemistry have also been tried when modeling a small number of TFs (Khanin*, et al.*, 2007; Nachman*, et al.*, 2004; Sanguinetti*, et al.*, 2006). Because it has relatively few parameters and can be fit by a simple algorithm, the work in this dissertation uses the bilinear framework.

To infer the activity of a set of TFs from the expression of their target genes, an inference algorithm must know at least some of the targets of each TF. We refer to this input as a *TF network map* (Brent, 2016). TF network maps link each TF to the targets it has the potential to regulate directly, given the right conditions. These maps are qualitative, so they can be represented by binary adjacency matrices. Fitting the bilinear model yields a control strength for each edge of the input map. Multiplying these CSs by the TFA levels inferred for a sample of cells yields a sample-specific network map showing how strongly each TF is influencing the expression of each of its targets in that sample. In previous work featuring inferred CS values, qualitative network maps have mostly been constructed from binding location data obtained by chromatin immunoprecipitation (ChIP) (Arrieta-Ortiz*, et al.*, 2015; Boscolo, *et al*., 2005; Boulesteix and Strimmer, 2005; Chen*, et al*., 2017; Liao*, et al*., 2003; Nachman*, et al*., 2004; Ocone and Sanguinetti, 2011; Rogers, *et al*., 2007; Sanguinetti*, et al*., 2006; Schacht*, et al*., 2014; Shi*, et al*., 2009; Tchourine*, et al*., 2018; Wang*, et al*., 2008; Yang*, et al*., 2005; Yu and Li, 2005). Garcia-Alonso *et al*. reported that a manually curated network map performed best for TFA inference in human (Garcia-Alonso, *et al*., 2019), but curated networks include very few TFs and are not available for most organisms. Details on how we analyzed the effects on TFA inference accuracy of using networks constructed from various high-throughput data sources can be found in later chapters.

In most previous studies, inferred TFA values were allowed to be positive or negative and their absolute value was interpreted as the magnitude of activity change relative to some reference sample. Thus, a smaller absolute TFA did not indicate less activity, but rather less change relative to the reference. As a result, the TFA levels did not distinguish between increasing and decreasing activity. Furthermore, the signs of the CS values had no meaning. Here, we propose, evaluate, and optimize a version of the bilinear approach in which TFA values are constrained to be non-negative, so that zero represents no activity, equivalent to deletion of the gene encoding the TF. We also include parameters representing the expression of each gene when all its regulators have activity zero (*baselines*). The combination of baseline expression levels with the non-negativity constraint on TFA values differentiates our model from previously proposed models. Together, they make the parameters interpretable. Positive control strength indicates that the TF activates the target and negative control strength indicates that it represses the target. If a TF's activity is larger in one sample than in another, then the TF is more active in the former sample than in the latter. We make extensive use of gene expression data after direct perturbations of TF activities (Arrieta-Ortiz*, et al.*, 2015; Tchourine*, et al.*, 2018; Tran, *et al.*, 2005), constraining each control strength parameter to be positive (activating) or negative (repressing) based on the direction in which the target gene's mRNA level changes when the TF is perturbed. If the gene encoding a TF is deleted in a sample, the TF's activity is fixed at zero; if it is overexpressed, the TF's activity is constrained to be greater than its activity in unperturbed samples

Evaluating the effects of various mathematical models, network mapping procedures, and perturbation-derived constraints, requires accuracy metrics that are objective, quantitative, and available for large numbers of TFs. This poses a challenge because TFA cannot be directly measured. As a result, most attempts to validate TFA inference algorithms have been small-scale and often qualitative, highlighting successes with just a few TFs. Some validation efforts have been based on inferring significant differential activity in a handful of samples subjected to stressors (Boorsma*, et al.*, 2008) or small molecules known to affect a particular TF's activity (Azofeifa, *et al*., 2018; Barenco*, et al.*, 2006; Ocone and Sanguinetti, 2011). Others have been based on inferring activity patterns that appear to match the periodicity of cell cycles (Liao*, et al.*, 2003; Nachman*, et al.*, 2004; Sanguinetti*, et al.*, 2006) or using changes in the nuclear localization of a GFP-tagged TF as a proxy for TFA (Boorsma*, et al.*, 2008). Other evaluation efforts have been based on internal consistency measures (Berchtold, *et al*., 2016), TF activity perturbations (Boorsma*, et al.*, 2008; Garcia-Alonso*, et al.*, 2019; Trescher and Leser, 2019), or identification of TFs important for proliferation of cancer cells (Alvarez*, et al.*, 2016; Azofeifa*, et al.*, 2018; Balwierz*, et al.*, 2014; Barenco*, et al.*, 2006; Chen*, et al.*, 2013; Chen*, et al.*, 2017; Cheng*, et al.*, 2007; Fröhlich, 2015; Garcia-Alonso*, et al.*, 2018; Jiang*, et al.*, 2015; Li*, et al.*, 2014; Ocone and Sanguinetti, 2011; Trescher and Leser, 2019; Tripodi, *et al*., 2018; Zhu*, et al.*, 2013). In this work, we take advantage of high-quality TF perturbation datasets in *Saccharomyces cerevisiae* and *Cryptococcuss neoformans* to present multiple quantitative, large-scale validation metrics. These validation metrics allowed us to reveal which high-throughput data types are most valuable for TFA inference in the bilinear framework, identify best practices for achieving high accuracy, and show that given the right input data, TFA inference works reasonably well.

1.3 Dissertation contributions

In part two of this dissertation, we detail our own model for TFA inference, and apply it to data from *Saccharomyces cerevisiae* (baker's yeast). One of the central novel contributions of this work is systematic, objective accuracy metrics for scoring TFA inference results, which leverages the existence of two large TF perturbation datasets collected using microarray technology. By assuming direct deletion or induction of a TF gene will result in true decrease or increase in the activity of that TF, we can define a set of ground truths to compare inferred TFA values against.

When performing TFA inference on both perturbation datasets, we analyze the impact that prior knowledge like TF network maps can have on TFA inference accuracy. The benchmarks allow us the quantify the difference between different sources of network constraints. It also allows us to estimate the effect of including more edges, at the risk that some of them may not be correct, versus including only high confidence edges. After optimizing our TFA inference approach by using the metrics as a guide, we demonstrate how these optimized TFA values can be used to draw conclusions about cellular information processing in *S. cerevisiae*, both novel and supported by the biological literature, by analyzing several gene expression datasets from experiments that altered nutrient availability. The work presented in this part comes from a prior publication:

> Cynthia Z Ma, Michael R Brent, Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data, *Bioinformatics*, Volume 37, Issue 9, 1 May 2021, Pages 1234–1245, https://doi.org/10.1093/bioinformatics/btaa947

In part three, we transition to work with a new organism, *Cryptococcus neoformans*, using gene expression data measured with a new technology. All gene expression data in part two was measured using microarrays, a technique first published forty years ago (Taub, 1983), that results in data points which are log2 fold changes in gene expression detected between an experimental sample and a reference sample. However, the newer technology of RNA-sequencing, which was developed in the mid-2000s and is increasingly more popular than microarrays, was used to measure the *C. neoformans* gene expression data to be analyzed for part three. RNA-seq data points are counts, measured without a reference sample, and have a much wider dynamic range for gene expression levels compared to microarray. We validate TFA inference on a subset of replicate TF perturbation samples of this dataset, simultaneously proving that TFA inference can be transferred to the new species as well as to the new data collection technology.

C. neoformans is a pathogenic yeast that the Brent lab has studied in collaboration with the Tamara Doering's Lab for over a decade. Environmental signals can trigger a highly visible response in the form of a greatly enlarged capsule, which protects the cells from a host's immune response. Given that this capsule is required for virulence in humans, we're interested in understanding how different combinations of environmental signals are propagated through changes in TFA, leading to changes in gene expression and ultimately capsule growth. We know new transcription is required for capsule growth because blocking transcription blocks capsule growth (Haynes *et al*, 2011), and several TFs have been identified which, when deleted from the genome, cause capsule to be either larger or smaller (Jung *et al*, 2015; Maier *et al*, 2015). However, we don't know which TFs respond to which signals, and we suspect that there may be more TFs involved in signal transduction that have not yet been identified. To address these unknowns, we designed a massive experiment that ultimately yielded over 1500 samples of gene expression profiles by RNA-sequencing, including from TF deletion strains and from WT strains grown in different combinations of capsule-inducing conditions. This data set, which took over 10 years to develop, is presented here for the first time, and analyzed with TFA inference methods to link specific TFs to specific environmental signals. Such analyses ultimately allowed us to make predictions of capsule change in response to deletion of TF-encoding genes, which could be experimentally confirmed.

Lastly, we close by reviewing the limitations of TFA inference, and looking forward to the future for this field. While we have had success with the simple bi-linear model, we have also noticed that better fits of the model to the expression data do not lead to better performance on the TFA evaluation metrics. This is not an issue of overfitting, as using a cross-validation approach where a shrinkage constraint is chosen to optimize the model's fit on held-out gene expression values also does not lead to better performance on the metrics. This indicates that the model itself is insufficient to capture the relationship between TFA and gene expression, or that maximal performance on the metrics does not reflect true TFA accuracy, and we discuss both possibilities in part four. In addition, we speculate on the future of TFA inference, especially the impact of newer technologies like single-cell RNA-seq and nascent RNA-seq.

Chapter 2: TFA inference and evaluation in S. cerevisiae

2.1 Background

The activity level of a transcription factor (TF) in a given cell is the extent to which it is exerting its regulatory potential on its target genes. Cells process information, in part, by changing the activity levels of TFs, thereby changing the transcription rates of their target genes. Changes in TF activity can occur by several molecular mechanisms, making it difficult to measure directly, so several algorithms have been developed for inferring TF activity (TFA) from gene expression, usually by fitting the parameters of a mathematical model to the expression data (Balwierz*, et al.*, 2014; Barenco*, et al.*, 2009; Boulesteix and Strimmer, 2005; Bussemaker*, et al.*, 2001; Chen*, et al.*, 2017; Cokus*, et al.*, 2006; Fröhlich, 2015; Fu*, et al.*, 2011; Jiang*, et al.*, 2015; Khanin*, et al.*, 2007; Li*, et al.*, 2014; Liao*, et al.*, 2003; Nachman*, et al.*, 2004; Sanguinetti*, et al.*, 2006; Schacht*, et al.*, 2014; Tchourine*, et al.*, 2018; Yu and Li, 2005). We'll be following the bi-linear framework (Boulesteix and Strimmer, 2005; Fu*, et al.*, 2011; Liao*, et al.*, 2003; Yu and Li, 2005), where gene expression is modeled as depending linearly on a set of TFA parameters, the values of which vary from one biological sample to another, and a set of control strength (CS) parameters, the values of which are constant across samples but vary as a function of the TF and target gene.

In addition to the gene expression data, TFA inference models need to know at least some of the targets of each TF. Previous work constructed such input TF network maps from binding location data like chromatin immunoprecipitation (ChIP) or scoring the promoter regions of genes for the presence of TF binding motifs (Arrieta-Ortiz*, et al.*, 2015; Boscolo, *et al*., 2005; Boulesteix and Strimmer, 2005; Chen*, et al.*, 2017; Liao*, et al.*, 2003; Nachman*, et al.*, 2004; Ocone and Sanguinetti, 2011; Rogers, *et al*., 2007; Sanguinetti*, et al.*, 2006; Schacht*, et al.*, 2014; Shi*, et al.*, 2009; Tchourine*, et al.*, 2018; Wang*, et al.*, 2008; Yang*, et al.*, 2005; Yu and Li, 2005). Here, one of our goals is to analyze the effects on TFA inference accuracy when networks of comparable size are constructed from various high-throughput data sources, as well as the effects of constraining the model with other prior information, such as the direction of TF-target gene regulation. To do so, we take advantage of two independent, high-quality perturbation datasets in *Saccharomyces cerevisiae* to present multiple quantitative, large-scale validation metrics. By using one dataset for network construction and constraint generation, and the other for validation, we reveal which high-throughput data types are most valuable for TFA inference in the matrix factorization framework, identify best practices for achieving high accuracy, and show that TFA inference can be used to discover novel regulatory behavior.

2.2 The model and its optimization

We use a simple model in which the log expression level of a gene in a given sample is determined by its baseline expression level, when none of its regulators are active, plus the sum of the influences of all the TFs that regulate it. The influence of each TF is a product of the strength with which the TF regulates that gene (*control strength*) and the TF's activity in that sample:

$$e_{i,k} = baseline_i + \sum_{j \in TFs} (controlStrength_{i,j} \times activity_{j,k})$$

where $e_{i,k}$ is the log expression level of gene i in sample k, $baseline_i$ is the expression level of gene i absent any influence from TFs, $controlStrength_{i,j}$ is the condition-independent potential of TF j to activate or repress gene i, and $activity_{j,k}$ is the activity level of TF j in sample k. In matrix notation, **E** = **CS** ● **TFA**, where **E** is a gene expression matrix (genes by samples), **CS** is a matrix of control strengths (genes by TFs) augmented to incorporate baselines, **TFA** is a matrix of TF activity levels (TFs by samples), and ● indicates matrix multiplication (Fig. 2.1). Fitting the CS and TFA matrices to expression data is equivalent to factoring the expression matrix, under the constraints that CS signs are predetermined, TFA is non-negative, and the activities of perturbed TFs are constrained according to the perturbation.

In initial fits (Fig. 2.1A, top), models are fit to gene expression data by least-squares linear regression, alternating between TFA and CS matrices (which include baselines), starting from 20 random initializations of the CS matrix. If a TF does not regulate a gene in the TF network map, the corresponding CS is held at zero; otherwise, it is constrained to be either positive (activating) or negative (repressing). If a TF is deleted in a sample, its activity is held at zero; if it is overexpressed, its activity is constrained to be greater than its activity in unperturbed samples. Except for deletion samples, all activities are constrained to be >= 0.0001. When learning a CS matrix, the mean activity of each TF, across all samples, is constrained to be one, since scaling a TF's activities and control strengths by inverse factors does not affect the predictions. After each iteration, the non-baseline control strengths are held constant while the TFAs and baselines are fit to the second data set by alternating linear regression without constraining the mean activity of a TF (Fig. 2.1A, middle). Optimization against the first dataset is halted when R^2 in the second data set peaks or after 100 iterations, whichever comes first. This halting criterion does not use the perturbation key of the second dataset, which determines the gold standard for evaluation. It

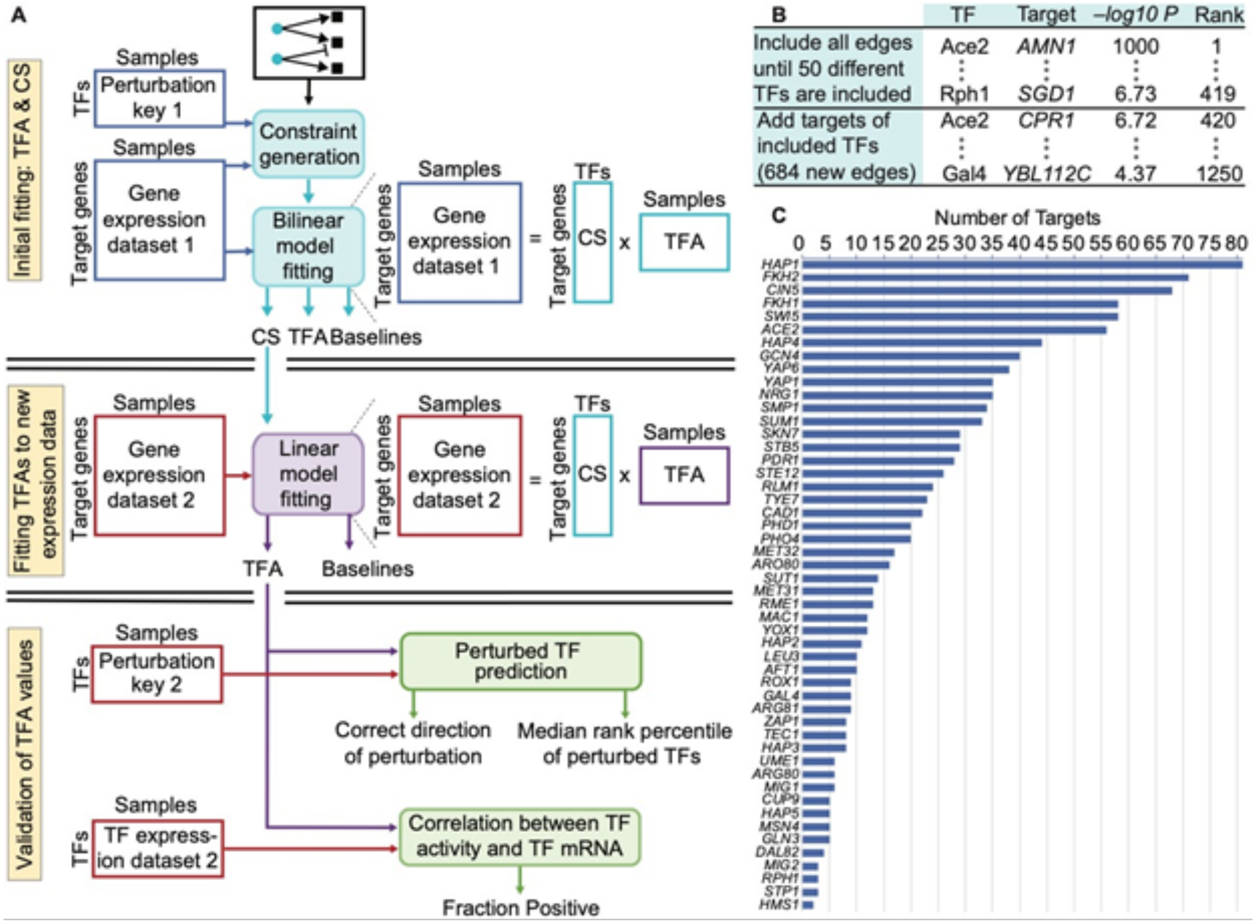


Figure 2.1. Evaluation framework and ChIP-based network construction. **A**. Overview of three-stage model fitting and TFA evaluation procedure. Gene expression levels and the perturbation key from dataset 1 are used only in the initial fitting. The CSs inferred in the initial fitting are fixed while the TFAs and baselines are refit to the target gene expression levels from dataset 2. The mRNA levels of the TFs and the perturbation key from dataset 2 are used only for evaluation. **B**. Illustration of how edges were selected for the ChIP-based network. All edges were ranked according to their -log *P*-value for the TF binding in the promoter of the target. Edges were selected in rank order until there was at least one edge from 50 different TFs. Lower-ranked edges were then selected for those TFs until rank 1250. After initial model construction, we removed any TFs with a single target and any set of TFs with identical target sets, along with their target genes. We then returned to the list and iteratively added edges that had previously been passed over until the network stabilized at 50 TFs. This yielded a network with 1,104 edges. **C**. The number of targets for each of the 50 different TFs in the ChIP network.

does not use the expression levels of the TFs, either, as they are removed from the input expression profiles Supplemental figures 2.S3 and 2.S4 replicate the results of Figure 2.2A without using the expression profiles from the second dataset to determine the stopping criteria. Comparable results were obtained by using Knitro (Byrd*, et al.*, 2006), a general non-linear solver, to optimize all parameters at once.

2.3 Evaluation metrics

Our approach to TFA evaluation relies on two independent expression data sets in which the activity of each TF has been individually perturbed. The first data set consists of expression profiles of strains in which a single TF was deleted from the genome (*TFKO* data) (Kemmeren, *et al*., 2014). The second consists of expression profiles collected 15 minutes after overexpression of a single TF was induced using the ZEV system (*ZEV* data) (Hackett, *et al*., 2020). These two data sets represent two very different growth conditions: TFKO profiles cells growing on synthetic complete medium in shake-flasks with no nutrient limitation, while ZEV profiles cells growing on minimal medium in phosphate-limited, continuous-flow chemostats. Two of our core evaluation metrics focus on whether TFA inference can determine which TF was perturbed in each sample and whether it was knocked out or induced. We refer to this information as the *perturbation key*. First, we use both the perturbation key and the expression profiles from one data set for network construction, constraint generation, and fitting (Fig. 2.1A, top). Next, we hold the control strengths from the initial fit fixed and refit the TFA and baseline parameters to the expression profiles in a second perturbation dataset (Fig. 2.1A, middle). Crucially, the perturbation key for the second dataset is not used for either fit, so it can be used as independent evaluation data (Fig. 2.1A, bottom).

We use three core evaluation metrics for TFA accuracy:

Direction of Perturbation To predict the direction of TF perturbation in a given sample, we compare the TF's activity in that sample to its activity in the WT sample. The percent of samples correctly predicted is calculated for where one of the two datasets serves training data and the other as the test data. We then swap the roles, with the dataset that was previously serving as training now serving as test, and vice-versa. The final score as an average between the two. Below, we refer to this averaging procedure as the average of the two train-test directions. To calculate the p-value, a binomial test is calculated for a 50% random chance of guessing the correct direction, where the number of trials is the total number of TFs evaluated in each dataset, averaged using Fisher's combined probability test.

Median Rank Percentile To predict the perturbed TF, we log and standardize all activities to Z-scores within each TF, then rank all standardized log activities in each sample. If a sample involves a TF overexpression, we rank from highest to lowest; if a sample involves a TF knockout or knockdown, we rank from lowest to highest. The rank 1 TF is given the rank percentile of 100%, while subsequent TFs are $\left(100 - \frac{(rank-1)}{numTFs}\right)\%$. We expect the perturbed TF to be highly ranked, so the rank percentile of the perturbed TF is used as an accuracy score for each sample, and the median rank percentile is used as an accuracy score for each dataset (TFKO and ZEV), with the final score being an average between the two train-test directions. To calculate the P-value, a binomial test is calculated where the number of trials is the number of samples evaluated, half are successes, and the probability of success is the probability of randomly achieving the final score or better as a rank percentile. This represents the desired null model of how likely it is that half the TFs achieve at least the median rank percentile, given the random probability of that rank percentile.

Positive correlation The fraction of TFs whose measured mRNA level and inferred activity are positively correlated is calculated for 1,000 bootstraps from the 180 samples used in the second fitting, and the median from the bootstrapping for each train-test direction is then averaged to return as the final score. (An analysis of correlation significance based on a model rather than bootstrapping yields the same conclusions; Supplemental Figure 2.S5.) To calculate the P-value, a binomial test is calculated for a 50% random chance of getting a positive correlation, where the number of trials is the total number of TFs evaluated in each dataset, averaged using Fisher's combined probability test. The statistical significance of the individual TFs' correlations is not considered, since the bootstrapping provides robustness against sampling error. We do not expect accurate TFA values to yield 100% for this metric, since post-transcriptional regulation is a key determinant of TFA levels, but we do expect the fraction of positive correlations to be substantially greater than half.

The metrics reported below are averages after training on each dataset, testing on the other, and then switch the roles of the two datasets. P-values from these two train-test directions are combined by using Fisher's combined probability test (Fisher, 1954). The first fitting only used samples in which one of the network TFs was directly perturbed, while the second fitting also used samples in which other TFs were perturbed. For all analyses, we considered only the 179 TFs that were perturbed in both ZEV and TFKO datasets.

2.4 Results

2.4.1 Constructing a ChIP-based network

Suppose binding location data are available for many of the TFs in a given organism, along with gene expression profiles from many samples. For now, assume that the expression dataset does not contain direct TF perturbations or, if it does, that the perturbation key is not used. To generate a network map as input for TFA inference, all possible TF-target *edges* can be ranked according to the strength of evidence that the TF binds in the promoter of the gene. We did this for yeast, ranking edges according to their negative log *P*-value in a comprehensive ChIP-chip dataset (Harbison, *et al*., 2004). This produced a single, global ranking of all edges involving all TFs (Fig. 2.1B). We then constructed a network map (Fig. 2.1B).

First, all possible TF-target interactions are first ranked according to the strength of evidence that the TF regulates the potential target gene. We did this for yeast (*S. cerevisiae*), ranking edges according to their negative log p-value in a comprehensive ChIP-chip dataset, their absolute differential expression in a TF perturbation sample, or their maximum negative log p-value in a comprehensive PWM dataset. To integrate data sources, the edges can be rank-averaged at this point, though the performance of such integrated networks is not shown in this paper. After ranking, we first dropped all but the top 1,250 edges. Then, starting from the top, edges were added until 50 TFs were included. If there were not enough TFs, the total number of edges considered was iteratively increased by 25 until at least 50 TFs could be recovered. Any remaining edges were only included if they emanated from TFs already in the map. This initial map was then checked for any TFs with a single target gene, and any set of TFs with identical target genes. These TFs and their target genes were removed from the map. Single-target TFs were removed to avoid having TFA values dependent on only one feature, which would be extremely vulnerable to noise or measurement error. TFs with identical targets were removed because it is impossible to separate the contributions to gene expression (i.e. the control strength matrix is not of full rank). If necessary, we returned to the list and added edges that were previously skipped over, repeating all steps until the network holds steady at 50 TFs.

2.4.2 Evaluating TFAs inferred from the ChIP network with correlation-based constraints

First, we evaluated the ChIP network without any constraints on the parameters (except non-negative TFA) and found that it did not perform better than chance on any metric (Fig. 2.S1). Next, we tried adding constraints on the signs of the control strength parameters, based on the intuition that a positive correlation between the mRNA levels of a TF and a target suggests activation while a negative correlation suggests repression. We constrained the sign of each control strength to match the sign of that correlation (even if the correlation was not significant) using the dataset reserved for the initial fitting, which improved performance. We refer to the ChIP-based network with correlation-based constraints as ChIP-CC (File S1-4). Using ChIP-CC, the direction of activity change between a TF's perturbation sample and the unperturbed sample was predicted correctly for 66% of TFs ($P < 0.01$, binomial test; Fig. 2.2A, left), the median rank percentile was 76.5% ($P < 0.0001$, binomial test; Fig. 2.2A, middle), and at least 65% of TFs' activity levels were positively correlated with their mRNA levels in half the sets of bootstrapped samples (Fig. 2.2A, right; $P < 0.01$).

2.4.3 Adding constraints obtained from TF perturbation data

When expression data after direct TF perturbation are available, the perturbation key can be used to constrain both the CS signs and the activity of the TF perturbed in each sample, as described above. We evaluated the effect of using the perturbation key from the first data set to constrain the control strengths and activity levels during the first fit (Fig. 2.1A, top). The perturbation key for the second dataset is only used for evaluation (Fig. 2.1A, bottom), not during the second fit (Fig. 2.1A, middle). We call the ChIP network with perturbation-constraints ChIP-PC (File S7-10). Performance on all three metrics increased, relative to using correlation-based constraints (Fig. 2.2A, blue and orange bars).

2.4.4 Generating network and constraints from TF perturbation data, without binding data

If expression data from direct TF perturbations is available, it is possible to build a network from the perturbation data rather than ChIP data. The same network building procedure is used, but instead of TF-target interactions being ranked by the strength of ChIP evidence, they are ranked by the absolute value of the log fold change of the target when the TF is perturbed (DE-PC, Files S13-16). The top 1250 edges were not sufficient to build a network of 50 TFs when using the TFKO dataset, so additional edges were considered in increments of 25 until the top 1400 edges was found to be sufficient. The performance of this network with perturbation-based constraints was similar to that of ChIP-PC, with the biggest change being an increase in median rank percentile from 92% to 96% (Fig. 2.2A, red and orange bars). We conclude that differential expression data from direct TF perturbations are necessary and sufficient for accurate TFA inference performance -- binding location data are not necessary.

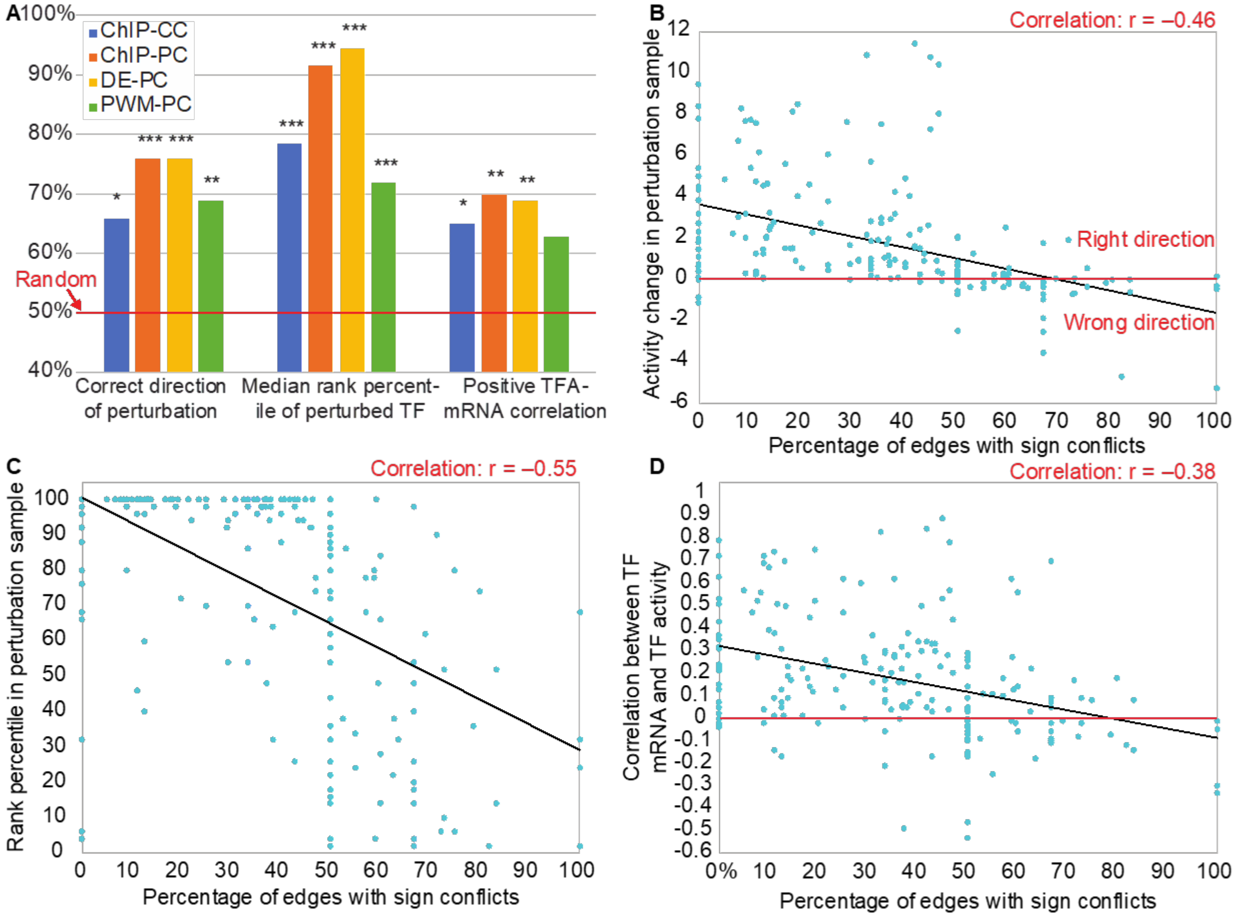


Figure 2.2. Determinants of TFA accuracy. A. Effects of network construction and constraint generation on TFA accuracy. Blue: ChIP network with correlation-based constraints. Orange: ChIP network with perturbation-based constraints. Yellow: Differential expression network with perturbation-based constraints. Green: Binding-specificity (PWM) network with perturbation-based constraints. Asterisks above the bars indicate magnitude of significance compared to a random model, with 1, 2, or 3 asterisks representing p-value thresholds of 0.01, 0.001, or 0.0001 B. Vertical axis: The activity of each TF in the sample in which it was perturbed minus its activity in the unperturbed sample, oriented so that higher is better. TFs plotted below the horizontal axis have been inferred to change activity in the wrong direction. Horizontal axis: The fraction of each TF's targets for which the TFKO and ZEV data sets suggest conflicting CS signs. TFs with < 50% conflict edges are almost all predicted in the correct direction, while most TFs with > 50% conflict edges are not. C: Vertical axis: Rank percentile of the perturbed TF's activity change in each perturbation sample (higher is better). Horizontal axis: Same as B. TFs with a higher percentage of conflict edges tend to be ranked lower. D: Vertical axis: median fraction of bootstrap samples in which a TF's mRNA level and its inferred activity level are positively correlated (see main text). TFs with a higher percentage of conflicting edges tend to have low or negative correlation. B-D: Results from the 50-TF ChIP-PC and DE-PC networks, trained on each of the datasets and tested on the other, have been combined, but each individual set of 50 points showed similar, highly significant correlations.

2.4.5 Using binding specificity models in network generation

A popular source of data for building gene regulatory networks is models of TF binding specificity, typically represented as position weight matrices (PWMs) (Boorsma*, et al.*, 2008; Boscolo*, et al.*, 2005; Bussemaker*, et al.*, 2001; Cheng*, et al.*, 2007; Garcia-Alonso*, et al.*, 2018; Lee and Bussemaker, 2010). To test this approach, we ranked all possible TF-target interactions by the maximum, across all positions in the target gene's promoter, of the negative-log P-value for presence of the TF's motif, as defined in the ScerTF database (Spivak and Stormo, 2012) and scored by FIMO (Grant, *et al*., 2011). We then built a network from this ranking just as we did with ChIP-chip data and the differential expression data (Fig. 2.1B). Using this network, we optimized with perturbation-based constraints (PWM-PC, Files S19-22). This network performed significantly worse than both ChIP-PC and DE-PC networks (Fig. 2.2A, green bars).

2.4.6 Increasing the number of TFs

To infer activity for more than 50 TFs, the input network maps can be extended by considering more than just the 1250 top ranked edges. To quantify the loss in performance from using lower ranked edges, we built independent networks of 50 TFs where the support for all edges decreased, without overlap. In anticipation of situations like the DE network, where the total number of edges had to be increased beyond 1250 in order to obtain a network of 50 TFs, a generous 2,000 edges were selected to comprise a "block," even though the networks themselves never needed all 2,000 edges. Block 1 networks are the same networks described above, while the Block 2 networks were created after reassigning ranks when the top 2,000 edges were zero-ed out, the Block 4 networks were created after reassigning ranks when the top 6,000 edges were zero-ed out, etc... The number of edges considered to build each network was kept to 1,250 whenever possible, with the only exceptions being two DE networks based on TFKO, where Block 1 used 1,400 as described above, and Block 2 used 1,450 edges.

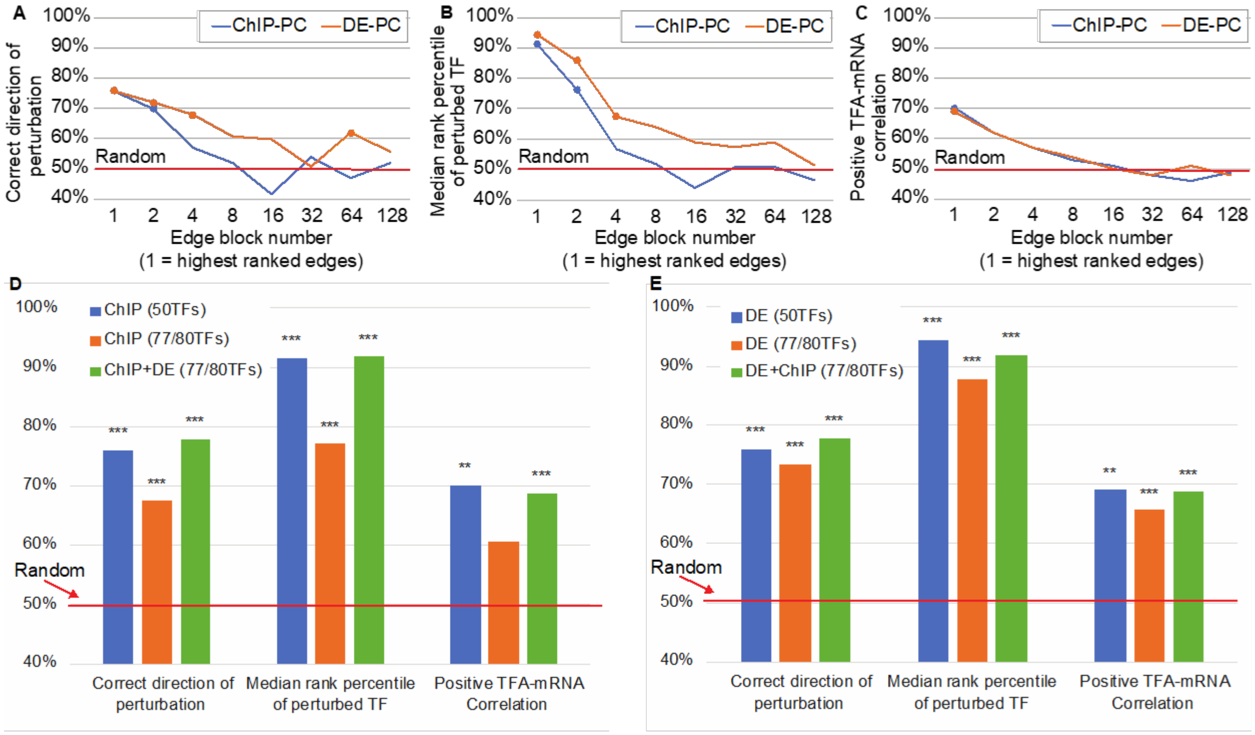


Figure 2.3. Effects of increasing the number of network TFs on accuracy. A-C: Accuracy metrics for networks constructed from the ChIP or DE edge lists by taking successively lower ranked edges. Edges were divided into blocks of 2,000 and blocks are plotted in an exponential series. For example, Block 1 is edges ranked 1-2,000 and Block 4 is edges ranked 6,001-8,000. Points are plotted for results that are significantly better than random ($P < 0.001$) A: Percent of TFs whose direction of perturbation is predicted correctly. B: Median rank percentile of the perturbed TF. C: Percent of TFs with a positive TF-mRNA correlation. In A and B, the ChIP-PC performance starts out similar to DE-PC, but it drops faster to no better than random in any measure by Block 4. D. Comparison of two ways of increasing the number of TFs in the network -- going further down the list of ChIP edges or using 50-TF ChIP and DE networks and averaging standardized TFAs of TFs that are in both networks. Consistent with panels A-C, performance degrades when lower ranked edges are included in the ChIP network. Inferring TFAs separately and averaging them, by contrast, yields performance on a larger network that is as good as performance on the smaller, 50-TF networks. E. Same as D, but blue and orange bars are for DE networks.

Both ChIP-PC and DE-PC lost accuracy steadily as lower ranked edges were used. DE-PC lost accuracy more slowly in the first two metrics (Fig. 2.3A, B) while both networks lost ground at the same rate for TFA-mRNA correlation (Fig. 2.3C). Thus, DE-PC appears to be the better choice for building larger networks with more TFs.

If both ChIP and TF perturbation data are available, another option is to infer activities separately for 50-TF ChIP-PC and DE-PC networks and combine the results, averaging the standardized activities for TFs that are in both networks. This provides activities for a larger number of TFs with accuracy that is better than including lower-ranked edges from a single data source (Fig. 2.3D, E, orange vs. green bars).

2.4.7 Effect of optimizing control strengths on TFA accuracy

To determine how much optimizing control strengths contributes to the accuracy of inferred TFAs, we compared TFAs obtained by optimization of both TFA and CS matrices to TFAs obtained by using fixed control strengths of +1 for activation or -1 for repression (*signed binary* CSs). Signed binary CSs were also used in (Arrieta-Ortiz*, et al.*, 2015; Chen*, et al.*, 2017; Gitter*, et al.*, 2013; Tchourine*, et al.*, 2018). The optimized CSs performed slightly better on some metrics and some networks, but there was little difference overall (Fig. 2.4A, B). To investigate the potential value of CS optimization further, we considered two additional metrics.

Known regulators of TF activity The TFKO dataset contains 1,484 samples from strains in which a gene was deleted, including many known regulators of TF activity. Our next evaluation task was to determine the target TFs in samples where a known regulator of TF activity is perturbed. To evaluate performance on this task, we compiled a map of known TF activity regulators and their targets (File S31). For each TF, an activity regulator was assigned to it if there was published literature that proposed a direct interaction that affects TF activity by a specific mechanism, such as phosphorylation, nuclear localization, or complex formation. Networks, constraints, and initial fitting used the ZEV dataset. The resulting CS matrix was then used to fit TFAs and baselines to the 194 TFKO samples in which a known TFA regulator was perturbed. In each perturbation sample, all TFs were ranked by the absolute difference between their standardized log activities in the perturbed and unperturbed samples. The absolute value was used because the literature is not always clear on the direction of regulation. We then plotted the fraction of literature-supported targets that were ranked above a given percentile in the sample where their TFA regulator was perturbed (Fig. 2.4C). The optimized CS matrices (solid lines) identify more known TFA regulators than the signed binary matrices (dashed lines), especially at rank percentiles above 85%. The ChIP-PC network (blue) also outperforms the DE-PC (orange) by this metric.

ZEV time course data Although we have focused on the ZEV data from 15 minutes after TF induction, they are part of time courses with samples taken 2.5, 5, 10, 15, 20, 30, 45, 60, and 90 minutes after induction (some timepoints are not available for some TFs). The expression data show that each TF is rapidly induced to a very high level and remains highly expressed throughout the 90 minutes. Therefore, we decided to infer TFAs for the entire time course, using a CS matrix derived from the TFKO data. For each time point, we computed a log fold change of the induced TF's inferred activity, relative to its activity at time 0. We then fit a 4-parameter, sigmoidal, saturating curve to the time series: $h_0 + \left(\frac{h_1 - h_0}{1+e^{\beta(x-t)}}\right)$. To eliminate curves that fit poorly, we tried several different thresholds on the variance explained by the sigmoidal fit. For each threshold, we calculated the fraction of fits that showed increasing activity throughout the time series (expected behavior) rather than decreasing activity. Overall, the TFAs with optimized CS matrices (solid lines) performed better than those with signed binary CS matrices (dashed lines) and this effect got stronger for curves with better fits (Fig. 4E). Furthermore, the DE-PC network (orange lines) performed substantially better than the ChIP-PC network (blue lines).

2.4.8 Condition-independence of control strengths

Control strengths are intended to be condition-independent, quantitative measures of each TF's potential to regulate each target gene. We have seen that good performance on TFA inference tasks can be obtained by using control strengths optimized on a different data set (Fig. 2.1A, 2.2A, 2.3, 2.4). Another indication that inferred CSs are transferable between data sets is that using the CS values inferred from one data set for TFA inference in a second data set increases the variance explained by 5%, relative to using the signed binary CSs (Fig. 2.S2).

Another way to test the condition-independence of control strengths is to calculate the correlation between CSs inferred from two data sets collected in different growth conditions. To maximize the number of TFs and their target genes we could use, we first created a large network from the union of edges from ChIP-PC and the two DE-PC networks, one derived from the TFKO data and the other from the ZEV. After filtering out edges with conflicting sign constraints and dropping two TFs that were left with only a single target, this new network (Union-PC, File S25) contains 94 TFs, 1,416 target genes, and 2,731 edges. We optimized both TFA and CS matrices using both the TFKO and the ZEV datasets together. Due to the datasets coming from different conditions, labs, and microarray technology, different baselines were optimized for the different datasets to compensate for any constant shifts in gene expression

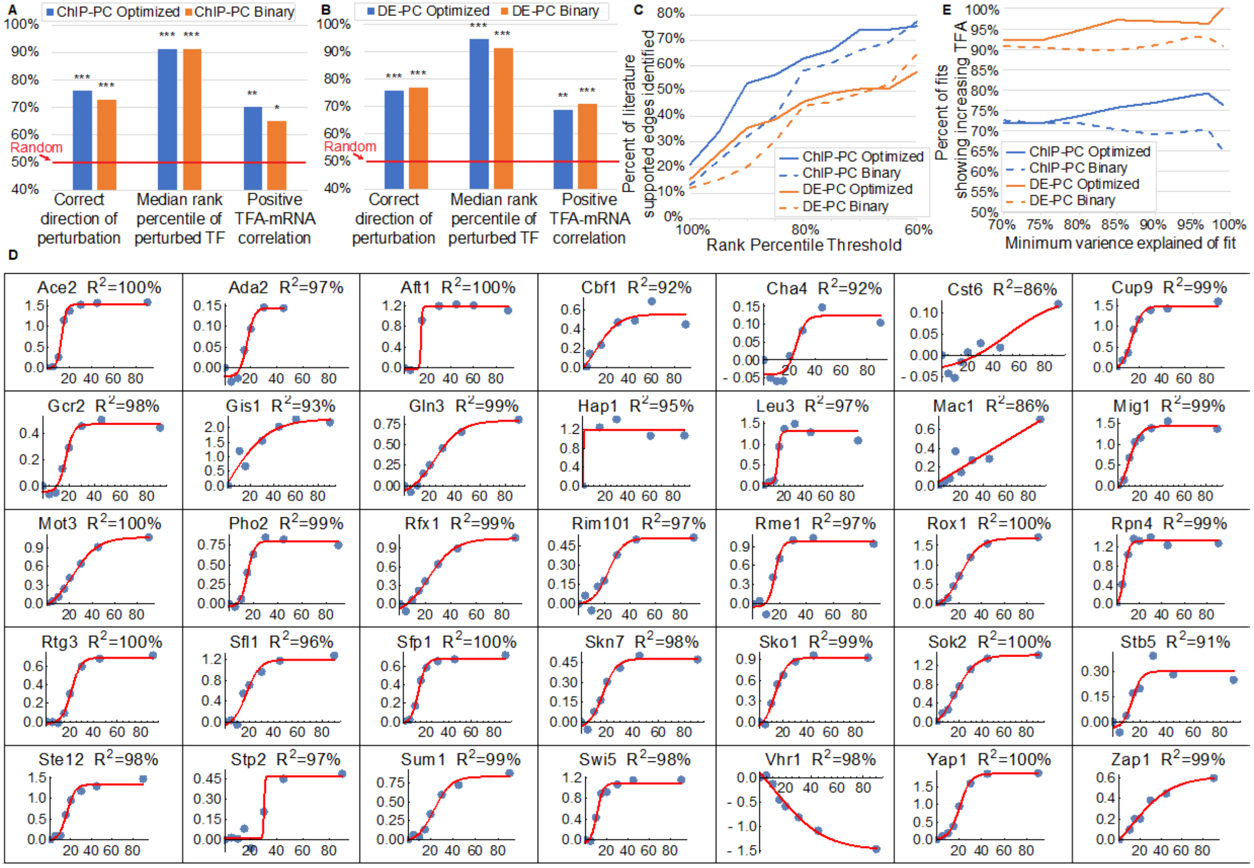


Figure 2.4. Impact of using a CS matrix optimized on a different data set versus using a signed binary CS matrix. A. ChIP-PC network. B. DE-PC network. C. Percent identification of literature-supported edges between TFA regulators and TFs, as a function of minimum rank percentile for identification. Solid lines: CS matrix optimized on the ZEV dataset and used to infer TFA's in the samples in which a TF regulator was deleted. Dashed lines: Signed binary CS matrix. For TFs whose change in standardized log activity from WT ranks above 85th percentile, more literature supported edges are identified by using optimized CS matrices than by using signed binary matrices. D. Sigmoidal fits to log2 fold change of TFAs inferred for the ZEV time course data, using the DE-PC network and a CS matrix optimized on the TFKO dataset, relative to the 0min timepoint. Only fits with variance explained above 85% are shown. In all but one of the 35 fits, TF activity is correctly inferred to be increasing (97%). Only Vhr1 activity is inferred to change in the wrong direction, probably because 9 of its 11 targets have sign conflict (80%, see Fig. 2.2B-D). E. After fitting sigmoidal curves as in D and imposing various thresholds on the variance explained by the fit, the percentage of fits that correctly show increasing activity. The DE-PC network (orange lines) performs better than the ChIP-PC network (blue lines). For each network, using a CS matrix optimized on the TFKO data (solid lines) generally shows better performance than using a signed binary CS matrix (dashed lines), and this effect increases as the variance explained by the sigmoidal fits increases.

measurements. One perturbation sample for each TF in the network from each of the two datasets was included in the gene expression set, plus a WT sample for each of the two datasets.

As an initial validation of the resulting CS matrix, we used it to infer TFAs in a new data set consisting of 69 double-deletion strains (Sameith, *et al*., 2015). For this set, we want to score our ability to predict the perturbed TFs in any sample where a network TF is deleted, so we first check each sample if the knocked-out TFs are in the network. If none are, we skip the sample, and if one is, we rank and score in the same way we would rank and score a sample with a single TF knock-out. If both TFs are in the network, the standardized log2 activity values are ranked twice, once without including the first TF, and once without including the second TF. The rank percentile of the perturbed TF is used as an accuracy score for each ranking, and the median rank percentile is used as the summary accuracy score. The results for predicting the direction of perturbation (86.4% correct) and identifying the perturbed TF (rank percentile 95.7%) were even better than the results for smaller networks (compare to Fig. 2.2A). The percentage of TFs that showed positive correlation between their mRNA levels and their inferred TFAs, 66%, is slightly below the results for the smaller networks, but still much better than chance. Readers can use the optimized CS matrix (File S26) for TFA inference in other datasets.

To calculate the correlation of CS values inferred in two growth conditions, we re-optimized Union-PC using only the TFKO data (synthetic complete medium with 2% glucose and no nutrient limitation) and then only the ZEV data (minimal medium with 2% glucose in phosphate-limited chemostats). Focusing on the 76 TFs with at least 5 targets, we constructed 1000 bootstrap samples of target genes for each TF and for each sample, calculated the correlation between target genes' CS inferred from one dataset and their CS inferred from the other dataset. We then scored each TF by its median correlation across bootstrap samples. The median score (across TFs) was +0.32 and 61 of 76 TFs had positive scores (80%). (Figure 2.S6 shows these correlations and their significance thresholds calculated with a traditional model-based approach.) This shows that, while there may be some over-fitting of CSs to a particular growth condition, there is also a substantial amount of condition independence.

2.4.9 Evaluating control strengths directly

As another evaluation, we asked whether the control strengths inferred for the targets of a TF would correspond in any way to the strength with which the TF binds to the promoters of those targets in genome-wide binding location data. To do this, we turned to binding data obtained by the transposon calling cards method (Mayhew and Mitra, 2016; Wang, *et al*., 2011). In this method, a TF is linked to a transposase, which deposits a transposon in the genome near where the TF is bound. The number of transposons in a gene's promoter is an approximate measure of the amount of time the TF spends bound to that promoter. We predicted that promoter occupancy would be positively correlated with the inferred control strength, in most cases. Importantly, we considered only the genes that were targets of a TF in the input network and therefore had inferred control strengths. The input network itself does not contain quantitative binding strength information. We optimized both TFA and CS using the Union-PC network and both the TFKO and ZEV data together. Using 1,000 bootstrap samplings of target genes for inferring TFAs, the median correlation between inferred CS and measured transposons was calculated for each TF. Of the 11 TFs with Calling Cards data and at least 5 target genes, 9 (82%) have positive median correlation between measured binding events and inferred CS value, with a median correlation of 0.31. A histogram of correlations and their model-based P-values are shown in Supplemental Figure 2.S7.

2.4.10 Analyzing time courses after glucose influx

Next, we used the CS matrix for Union-PC (File S26) to infer baselines and TFAs for three expression time courses after yeast cells were provided with glucose. In the first dataset, yeast cells growing in galactose-limited chemostats were provided glucose to a final concentration of 0.02% (w/v) or 0.2% (Ronen and Botstein, 2006). In the second, batch cultures depleted glucose over a 24hr growth period before being transferred to fresh media with 2% glucose (Apweiler, *et al*., 2012). We plotted the inferred activity of each TF as a function of time, fitting both a 4-parameter sigmoid curve (as in Fig. 2.4) and a 6-parameter impulse curve (Chechik and Koller, 2009). The sigmoid curve is the same as used for ZEV time-course response pattern analysis, while the impulse curve allows for a return to a new baseline level: $\frac{1}{h_1}\left(h_0 + \frac{h_1-h_0}{1+e^{\beta(x-t_1)}}\right)\left(h_2 + \frac{h_1-h_2}{1+e^{-\beta(x-t_2)}}\right)$. This results in five general categories of behavior: an upward spike, a downward spike, monotonically increasing, monotonically decreasing, and poor fits (R^2<80%). We chose between the models using the Bayes Information Criterion (Schwarz, 1978) and filtered out the poor fits (R^2<80%).

For four well-studied TFs, Gcr2, Gln3, Gcn4, and Msn2, the inferred TFA levels and fits for all three time courses are shown as points and lines inside turquoise circles in Figure 2.5. The shapes of the activity curves in response to glucose made sense: Gcr2, an activator of glycolytic genes, increased in activity upon glucose addition; Gln3, Gcn4, and Msn2, activators of genes needed during nutrient deprivation, generally decreased activity upon glucose addition (Gln3 showed a very small increase in one time course). Gcr2 activity also makes sense in terms of the glucose concentrations added, returning to baseline quickly at the lowest concentration, more slowly at the intermediate concentration (note the last gold data point, which is not reflected in the fitted curve), and not at all at the highest concentration.

To gain a deeper understanding of the network structure, the control strengths, and how they led to the TFA patterns shown in Figure 2.5, we carried out enrichment analysis on the network targets of the four TFs using Gene Ontology biological process annotations and Kyoto Encyclopedia of Genes and Genomes metabolic pathway annotations (Liao, *et al*., 2019). We first discarded targets that, after optimization, had absolute CS <= 10^{-4}, then analyzed the remaining activated and repressed targets of each TF separately, and removed redundant terms (Bodenhofer, *et al*., 2011). All significantly enriched annotations that apply to four or more target genes are listed above the black squares in Figure 2.5. Each TF is connected to a target square by an arrow if the corresponding annotation was enriched among its activated targets and a T-head if the annotation was enriched among its repressed targets. Gcn4 and Msn2 had no repressed targets and the few repressed targets of Gcr2 had no enriched annotations that applied to four or more targets. The annotations and directions of regulation of target sets made sense in terms of the known function of each TF (Broach, 2012; Conrad, *et al*., 2014; De Virgilio, 2012; Ljungdahl and Daignan-Fornier, 2012; Rodkaer and Faergeman, 2014).

For each gene set in Figure 2.5, we calculated the median log fold change in each time course (not shown) and fit sigmoidal or impulse curves to them (shown in black boxes). In general, the shapes of the inferred activity curves for each TF made sense in light of the expression patterns of their target groups. For example, the inferred activity patterns of Gcr2 mirror those of the glycolysis genes it activates. Note that curves are expected to be inverted for repressed targets and that some of the lines occlude others at early time points, hiding early dips. Although the mRNA levels of target genes show a clear relationship to inferred activity levels, they do not always mirror each other perfectly. For example, the amino acid synthesis genes spike upward after addition of 0.2% glucose, whereas the inferred activity of Gcn4 dips and returns, presumably responding to influential target genes whose individual expression levels do not follow the pattern of the median levels shown.

Finally, we attempted to identify key regulators driving the TFA activity patterns observed in Figure 2.5, by inferring TFAs from two more data sets, using the same CS matrix. The first consists of expression profiles of strains deleted for 567 different regulatory factors, including many known TF activity regulators (Kemmeren*, et al.*, 2014). The second includes time courses after addition of rapamycin (an inhibitor of TORC1) or inhibitors of Snf1, Tpk1-3, or Sch9 (Zaman, *et al*., 2009). The inferred TFAs for these datasets are provided as Files S29 and S30. By examining the effects of these perturbations on the inferred activities of Gcr2, Gln3, Gcn4, and Msn2 and comparing them to published literature, we were able to confirm many previously described regulatory interactions (Fig. 2.5A, solid maroon lines) and to identify a few potentially new ones (Fig. 2.5A, dashed blue lines). The TF regulatory edges in Figure 2.5A were derived by manual curation of the changes in inferred TF activity when TF activity regulators are perturbed in these datasets. The evidence supporting most of the hypothesized novel TF regulatory interactions is shown in Figure 2.5B, C, but space limitations precluded showing the evidence for the SWI/SNF or Ure2 activating Gcr2. A description of the TFA regulatory interactions shown in Fig. 2.5 and of the literature supporting those that were previously known can be found in the online supplement. To summarize key highlights, we have identified several likely regulators of Gcr2 activity, about which little was previously known, and discovered that Grr1, previously known for its role in glucose repression, is probably a positive regulator of glycolysis (via Gcr2) and a negative regulator of stress-induced TFs. These novel observations,

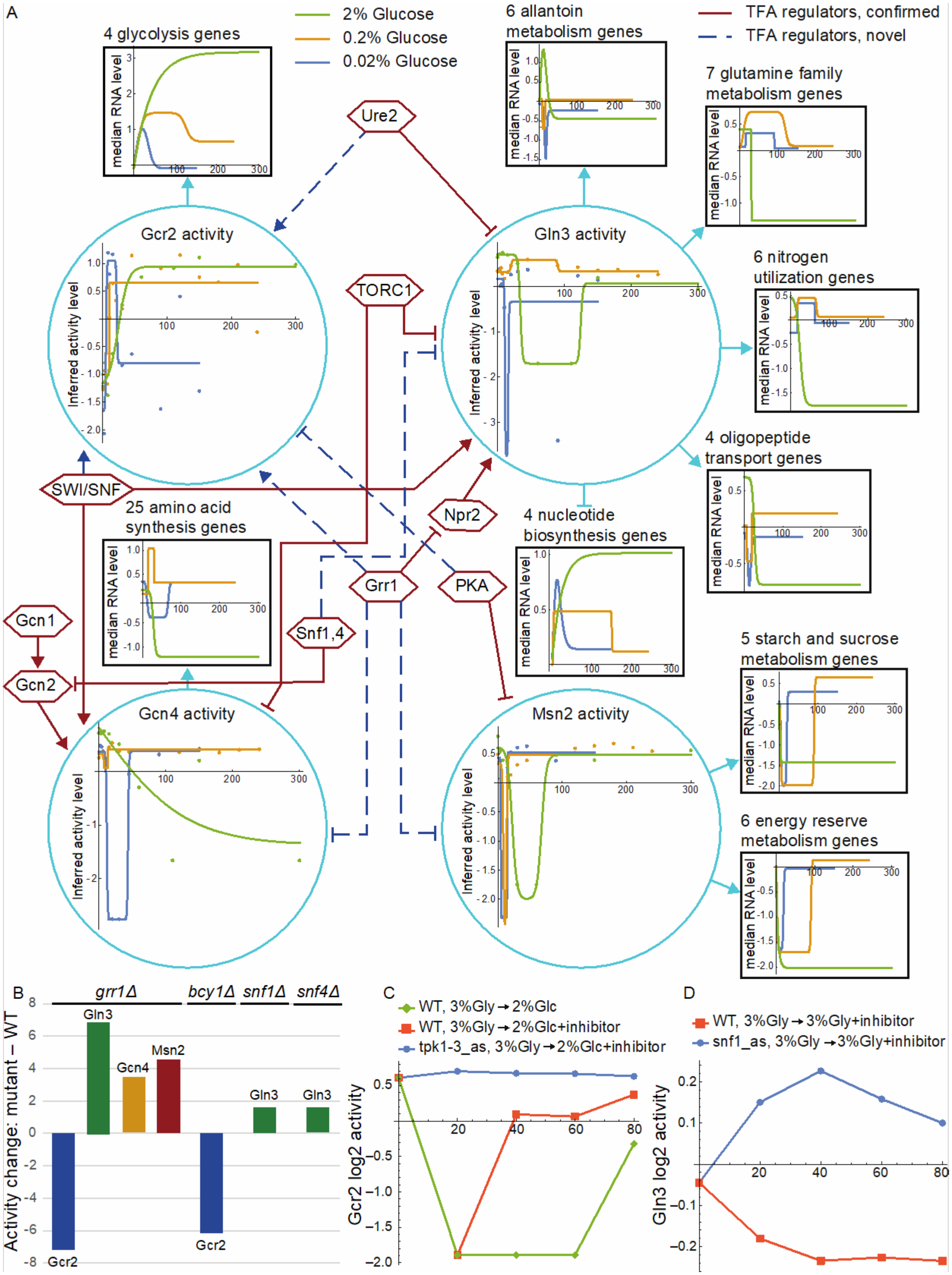

A
2% Glucose
0.2% Glucose
0.02% Glucose
TFA regulators, confirmed
TFA regulators, novel
4 glycolysis genes
6 allantoin metabolism genes
7 glutamine family metabolism genes
6 nitrogen utilization genes
4 oligopeptide transport genes
4 nucleotide biosynthesis genes
25 amino acid synthesis genes
5 starch and sucrose metabolism genes
6 energy reserve metabolism genes
median RNA level
Gcr2 activity
Gln3 activity
Gcn4 activity
Msn2 activity
Inferred activity level
Ure2
TORC1
SWI/SNF
Npr2
Grr1
PKA
Gcn1
Gcn2
Snf1,4
B
grr1Δ
bcy1Δ
snf1Δ
snf4Δ
Activity change: mutant – WT
Gcr2
Gln3
Gcn4
Msn2
Gcr2
Gln3
Gln3
C
WT, 3%Gly→2%Glc
WT, 3%Gly→2%Glc+inhibitor
tpk1-3_as, 3%Gly→2%Glc+inhibitor
Gcr2 log2 activity
D
WT, 3%Gly→3%Gly+inhibitor
snf1_as, 3%Gly→3%Gly+inhibitor
Gln3 log2 activity

Figure 2.5. Gcr2, Gln3, Gcn4, and Msn2, their activity regulators, activity changes in response to different glucose concentrations, target gene sets, and target set expression patterns. **A**. Turquoise circles: Changes in inferred TF activity after addition of 2% glucose to post-diauxic-shift shake-flasks with synthetic complete medium (green) or addition of 0.2% (gold) or 0.02% (blue) glucose to cultures grown in galactose-limited chemostats with minimal medium. Points are Log2 of inferred activity level and lines are impulse or sigmoidal fits to the points, chosen by the Bayes Information Criterion. Black boxes: Sets of target genes that are regulated in the same direction and are annotated to a Gene Ontology or KEGG term enriched among targets of the TF that regulates them. Arrowheads indicate activation and T-heads repression. Colored lines are impulse or sigmoidal fits to the median Log2 fold-change of the annotated genes at each time point, relative to time 0. Hexagons: TF activity regulators inferred from analysis of two data sets as described in the text. Solid maroon lines indicate clear literature support while dashed blue lines indicate hypothesized novel edges. **B**. Change in activity of transcription factors in response to deletion of *GRR1*, *BCY1* (inhibitory subunit of PKA), *SNF1*, or *SNF4* (activating subunit of Snf1 complex). **C**. Gcr2 activity after addition of 2% glucose to cells growing on 3% glycerol. In wild-type cells, glucose initially reduces Gcr2 activity (green, orange). (This response is different from Gcr2's response to glucose under the conditions of Fig. 2.5A.) Addition of the Tpk1-3 inhibitor with glucose to analog-sensitive cells (blue) eliminates that response, suggesting that PKA represses Gcr2 activity. This is consistent with the observation that deletion of *BCY1* reduces Gcr2 activity in 2% glucose (panel B). **D**. Gln3 activity is slightly elevated when inhibitor is added to cells growing in 3% glycerol and expressing an analog-sensitive Snf1 (blue), relative to WT cells (orange), suggesting that the Snf1,4 complex represses Gln3 activity. This is consistent with the observation that deletion of either Snf1 or Snf4 increases Gln3 activity in 2% glucose (panel B).

which were mined from gene expression data via TF activity inference, constitute a rich trove of hypotheses for future experimental investigation. The repression of Gcn4 and Msn2 activity by Grr1 is lent plausibility by the fact that the mRNA levels of *GCN4 and MSN2* increase significantly when Grr1 is deleted (the *GCR2* mRNA level does not change). One possible mechanism for regulation of Gcn4, Gcr2, and Msn2 by Grr1, a change in the nuclear concentration of the TF, could be tested by perturbing Grr1 in strains in which one of these TFs is linked to a fluorescent protein and comparing images of the perturbed and unperturbed samples. A complementary approach would be to carry out calling cards experiments on each TF in unperturbed cells and in cells in which Grr1 has been perturbed. This could detect changes promoter occupancy by the putatively regulated TFs.

2.5 Discussion

The ability to accurately infer changes in TF activity from changes in gene expression profiles provides a vital tool in the systems-biology toolbox. It enables us to look inside the cell, seeing not only the output of the circuits that control gene expression, but also their internal state. Observing the internal states of regulatory circuits is the key to understanding how these circuits control the cell's transcriptional program in response to internal and external signals. We demonstrated how this tool can be used to gain new insights into the regulation of TF activity, such as the probable activation of Gcr2 and repression of Gcn4, Gln3, and Msn2 by the Grr1 ubiquitin ligase in the presence of glucose (Fig. 2.5). In future work, we hope to automate this process of identifying TFA regulators and develop genome-scale benchmarks with which to evaluate it.

Previous studies introduced various versions of the matrix factorization approach (Boulesteix and Strimmer, 2005; Liao*, et al.*, 2003; Sanguinetti*, et al.*, 2006; Yu and Li, 2005). Here, we presented an objective, genome-scale evaluation of this approach, using multiple measures of accuracy and multiple independent data sets. We found that, when inferring TFA and CS matrices, it is essential to have data in which TFs are directly perturbed and constraints derived from such data. We also showed that a CS matrix derived from the TFKO and ZEV data can be used to successfully analyze other expression data sets that do not contain direct perturbations of TF activity. In two of our core metrics, matrix factorization with constraints yielded TFA values that cover more than half the distance from random (50%) to ceiling (100%; Fig. 2.2A). However, this probably underestimates the true accuracy of the method, since even perfect TFA inference would not necessarily yield 100% on these metrics. For example, deletion of an inactive TF will not result in decreased activity, and perturbation of a TF's activity could lead to an equal or bigger change in the activity of a downstream transcription factor, pushing the true rank percentile of the perturbed TF below 100%. Similarly, the true percentage of TFs whose activity is positively correlated with their mRNA level is almost certainly less than 100%, due to post-transcriptional regulation.

Building the input network from binding specificity models (e.g. PWMs) is a popular approach that is condition-independent in principle, but it did not do well in our evaluation (Fig. 2.2A). Network structures derived from either perturbation-response data (DE) or currently available binding location data (ChIP) perform about equally when using the highest-scoring edges, but as more edges are added, the DE network is more robust (Fig. 2.3). For the ChIP-PC network, we observed the best performance when using only the top 1250 edges, even though the recommended P-value threshold of 0.001 includes at least three times more. Using edges ranked 2001 through 3250 (Fig. 2.3A-C, Block 2) resulted in decreased performance on all metrics. The extended ChIP-PC networks of 77-80 TFs, which considered the top 2000 edges rather than the top 1250, also resulted in decreased performance on all the metrics (Fig. 2.3D). Since perturbation-response data are needed for sign constraints in any case, generating binding location data in addition may not be worth the effort and expense required. In other words, perturbation-response data is both necessary and sufficient for good performance.

Another recent paper (Trescher and Leser, 2019) evaluated TFA inference algorithms by using expression data after TF knockdowns in human cell lines and *E. coli*, similar to our second metric. It reported that the perturbed TF was rarely among those with the greatest inferred activity changes and that there was very little agreement among the algorithms tested. One factor that likely contributes to the difference in findings is that the algorithms they tested did not use sign constraints on control strengths and could not distinguish between increasing and decreasing activity. In the absence of sign constraints, we also saw poor performance. Another difference is that the input networks they used were largely based on manual curation rather than automated processing of high-throughput data, so they lacked confidence scores, making it impossible to select only the most confident edges. We saw performance degrade when less confident edges were included in the network. Finally, their evaluation was carried out using perturbations of only a handful of TFs for evaluation, making the findings vulnerable to sampling error.

We were surprised to find that, by our three basic metrics, optimizing control strengths is not necessary for achieving good performance -- signed binary control strengths, taken directly from the input network and sign constraints, do almost as well (Fig. 2.4A, B). Optimized CS matrices result in somewhat better performance on two other metrics -- detecting literature-supported TFA regulators (Fig. 2.4C) and detecting the trend in TF activity from a time course (Fig. 2.4D, E). Furthermore, we found evidence that the optimized control strengths correlated positively with those learned by optimizing on data from different growth conditions and with the strength of binding to target promoters in independent binding data. Nonetheless, these correlations were far from perfect, so the limited impact of optimizing CS matrices on TFA accuracy may reflect the fact that the control strengths in our testing framework are optimized on one data set while TFAs are optimized and evaluated on another (Fig. 2.1). In many real applications, control strengths and activities would be optimized on the same data set. Improvements to the mathematical model could also increase the importance of CS optimization (see below). For now, however, using signed binary control strengths may be a reasonable choice for some organisms, especially when a limited amount of gene expression data is available. When control strengths are not optimized, the overall optimization changes from non-linear to linear, making it much faster and simpler.

Any approach to TFA inference relies on having a reasonably large and accurate set of targets for each TF. This can be challenging for TFs that are not very active in any of the conditions in which the data used to build the network were obtained. For example, Gal4 had only 3 targets in the Union-PC network. One of those, *GAL10*, is an established target with a known role in galactose metabolism, but CS optimization reduces the link between Gal4 and *GAL10*, emphasizing instead the link between Gal4 and dubious ORF YDR544C. This may be due to the fact that none of the samples to which we fit the activity levels were grown with galactose, so the expression of *GAL10* does not vary much. As a result, there is little need to explain *GAL10* expression as resulting from changes in Gal4 activity. A possible approach to this problem would be to discard inferences about TFs that have two or fewer significant targets after CS optimization.

Analysis of the error patterns on our three core benchmarks showed that poor performance was highly correlated with the fraction of a TF's targets that exhibited opposite signs in the TFKO and ZEV data sets (Fig. 2.2B-D). Apparent sign conflict can occur when the true sign of regulation is consistent, but the effect of the perturbation on the target gene is so weak that random measurement noise leads to a sign error. This can be remedied by using multiple perturbation data sets to determine sign and discarding edges with sign conflicts, as we did when constructing the Union-PC network. However, conflicts can also occur because some TFs are repressors in some conditions and activators in others. For example, Rgt1 represses *HXT1* in low glucose but activates it in high glucose (Polish, *et al*., 2005). This points to a limitation of any model that constrains TFAs to be non-negative and CSs to be one sign or the other. A possible solution would be to release the non-negativity constraint on TFs when there is sufficient evidence of a true sign change that applies to most of its targets.

The matrix factorization approach has several limitations. First, predicted gene expression does not saturate as TF activity gets large, whereas in reality each gene has maximum and minimum expression levels and the binding sites for each TF eventually become fully occupied. Second, the model assumes that each TF-target relationship is either activating or repressing in all conditions. Third, TF-TF interactions, such as competitive or cooperative binding, are not accounted for. Fourth, it does not model condition-dependent epigenetic effects on the susceptibility of each gene to regulation, for example by making the promoter more or less accessible to TFs. Fifth, as parameters are optimized during the fitting process, the variance explained is not a good predictor of a model's value for accurate TFA inference. Finally, when control strengths learned on one data set are used to model a different data set, the variance explained in the second data set is much lower than in the first. This suggests a degree of over fitting that might be remedied by parameter shrinkage. However, improving the variance explained in cross-validation is not guaranteed to improve the accuracy of the TFA parameters learned.

We found that TFA inference in yeast works reasonably well when best practices are followed, but there is still room for improvement. We anticipate improvements coming from better network maps. One likely source of better maps is new, more accurate methods for measuring TF binding locations (Bergenholm, *et al*., 2018; Holland, *et al*., 2019; Kang, *et al*., 2020; Mayhew and Mitra, 2016; Shively, *et al*., 2019). The input network could also be improved by obtaining TF perturbation data from cells grown in new conditions. More improvement in TFA inference could come with better sign constraints and the possibility of allowing negative TF activity when the data strongly justify it. Mathematical models that more closely reflect the underlying biochemistry could also lead to better results, although such models come with new challenges. As new approaches are tried, the benchmarks presented here can be used to determine whether they robustly improve the accuracy of TFA inference, across multiple data sets and network maps.

2.6 Supplementary materials

2.6.1 Supplementary files

S1-S35 can be found at *Bioinformatics* online

https://academic.oup.com/bioinformatics/article/37/9/1234/5949002

Code for TFA inference and evaluation is available here:

https://doi.org/10.5281/zenodo.4050573

2.6.2 Evidence supporting inferred TFA regulators

The inferred TFAs for potential TF activity regulators, obtained by analyzing expression data from refs. are provided as File S29 and S30. We manually curated the results to obtain the TFA regulatory interactions shown in Figure 2.5. The evidence supporting most of the hypothesized novel TF regulatory interactions is shown in Figure 2.5B, C.

Some of the findings shown in Figure 2.5A are as follows. We confirmed that Gcn2 activates Gcn4 (Yang *et al*., 2000; Broach, 2012; Conrad *et al*,. 2014; Ljungdahl and Daignan-Fornier, 2008), Gcn1 activates Gcn4 (probably via its effects on Gcn2 (Zaman *et al*,. 2008)), and Ure2 represses Gln3 (probably by anchoring it to the plasma membrane, (Broach, 2012; Conrad *et al*,.

2014)). We also saw evidence that Ure2 may activate Gcr2 directly or indirectly. We confirmed the well-known role of TORC1 as a repressor of Gln3 and Gcn4 activity (Staschke *et al*., 2010; Rodkaer *et al*., 2014; Virgilio, 2012) but did not see unequivocal evidence that it represses Msn2 as previously reported (Rodkaer *et al*., 2014; Virgilio, 2012). Our analyses showed that Grr1 represses Gln3, probably via Npr2 (Avendano *et al*., 2005), which our analysis confirms as an activator of Gln3 (Broach, 2012)), and activates Gcr2 (probably indirectly; Fig. 2.5B). Although it has not been previously reported that Grr1 activates Gcr2, Grr1 is known to be required for glucose suppression, so activation of glycolytic genes via Gcr2 would be a consistent role. We also found that Grr1 represses Gcn4 and Msn2 (Fig. 2.5B), consistent with its being active in nutrient-replete conditions. Since Grr1 activates Gcr2 and represses the other three, a transient spike in its activity upon glucose influx could explain the upward and downward spikes we see in the activities of the four TFs. We confirmed that the SWI/SNF chromatin remodeling complex contributes to the activities of Gln3 (Avendano *et al*., 2005; Riego *et al*., 2002) and Gcn4 (Ljungdahl and Daignan-Fornier, 2008) and discovered that it also works with Gcr2. We confirmed that PKA represses Msn2 (Broach, 2012; Conrad *et al*,. 2014; Rodkaer *et al*., 2014; Virgilio, 2012; Wever *et al*., 2005) and saw evidence that it also represses Gcr2 (Fig. 2.5C), which has not been previously reported. We confirmed that Snf1 represses Gcn4 in the absence of glucose and amino acids and saw evidence that it weakly represses Gln3 (Fig. 2.5B, D), contrary to previous claims that Snf1 activates Gln3 (Virgilio, 2012; Bertram *et al*., 2002). In summary, we have identified several likely regulators of Gcr2 activity, about which little was previously known, and discovered that Grr1, previously known for its role in glucose repression, is probably a positive regulator of glycolysis (via Gcr2) and a negative regulator of stress-induced TFs. These novel observations, which were mined from gene expression data via TF activity inference, constitute a rich trove of hypotheses for future experimental investigation.

2.6.3 Datasets used

Yeast TFKO data The microarray expression data of 1,484 single gene knockout strains (Kemmeren *et al*., 2014) was downloaded from http://deleteome.holstegelab.nl/data/downloads/deleteome_all_mutants_controls.txt A sample using expression level of 0 for all genes was assumed in order to stand in for WT.

Yeast ZEV induction data The microarray expression data of 199 single gene ZEV induction strains (Hackett *et al*., 2020) was downloaded from https://storage.googleapis.com/calico-website-pin-public-bucket/datasets/pin_tall_expression_data.zip Only the column labeled log2_cleaned_ratio was considered for this work. A sample using expression level of 0 for all genes was assumed in order to stand in for the 0min timepoint, when induction of over-expression had not yet started.

Yeast ChIP-chip data P-values that represent TF binding significance from ChIP-chip experiments (Harbison *et al*., 2004) were downloaded from http://younglab.wi.mit.edu/regulatory_code/GWLD.html Values were transformed to negative log10 p-values for the purpose of treating them as confidence scores, where greater values indicate greater support.

Yeast PWM data Position weight matrices for S. cerevisiae motifs in the ScerTF database (Spivak and Stormo, 2012) were downloaded from http://stormo.wustl.edu/ScerTF/ Values from FIMO scanning (Grant *et al*., 2011) for motif hits were transformed to negative log10 p-values for the purpose of treating them as confidence scores, where greater values indicate greater support. If multiple hits were found between linking the same TF and target gene, the maximum score was used.

Yeast double-deletions data The microarray expression dataset of 69 double-deletion strains (Sameith *et al*., 2015) was downloaded from http://www.holstegelab.nl/publications/GSTF_geneticinteractions/ A sample using expression level of 0 for all genes was assumed in order to stand in for WT.

Yeast time course data The microarray expression data of 9 time-points (0min, 3, 7.5, 15, 30, 60, 110, 150, 300min) after 2% glucose influx for WT strains (Apweiler *et al*., 2012) was downloaded from http://www.holstegelab.nl/publications/glucose_regulatory_system/ A sample using expression level of 0 for all genes was assumed in order to stand in for 0min.

The microarray expression data of 13 (0min, 2, 4, 6, 8, 10, 15, 20, 30, 45, 90, 120, 150) and 15 (0min, 3, 5, 7, 10, 15, 20, 30, 45, 90, 120, 150, 180, 210, 240) time-points after 0.02% and 0.2% glucose influx for WT strains (Ronen and Botstein, 2006) was downloaded from GEO with accession ID GSE4158. A sample using expression level of 0 for all genes was assumed in order to stand in for 0min.

The microarray expression datasets of 5 time-points (0min, 20, 40, 60, 80) for multiple conditions and multiple strains (Zaman *et al*., 2009) were downloaded from https://puma.princeton.edu/cgi-bin/publication/viewPublication.pl?pub_no=524. Where possible, all samples were re-scaled to use the WT in 3% glycerol condition as the reference, and a sample using expression level of 0 for all genes was assumed in order to stand in for this reference.

2.6.4 Supplementary figures

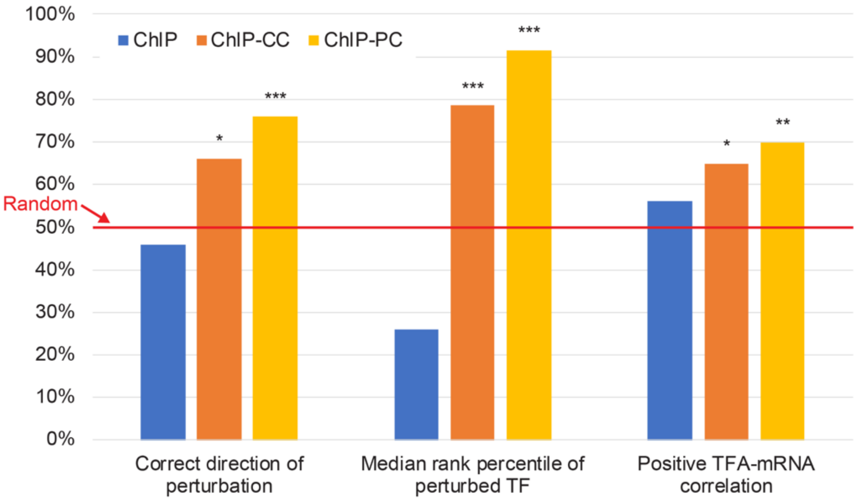


Figure 2.S1 Using ChIP to define the network without constraining the signs (blue) resulted in performance not significantly better than random. Performance of the ChIP network with correlation-based (orange) and perturbation-based (yellow) sign constraints are plotted for comparison.

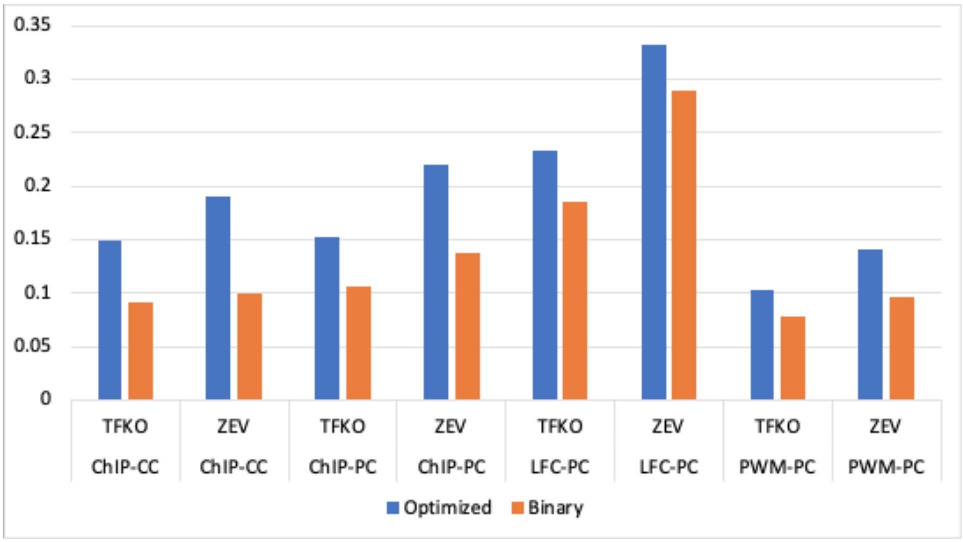
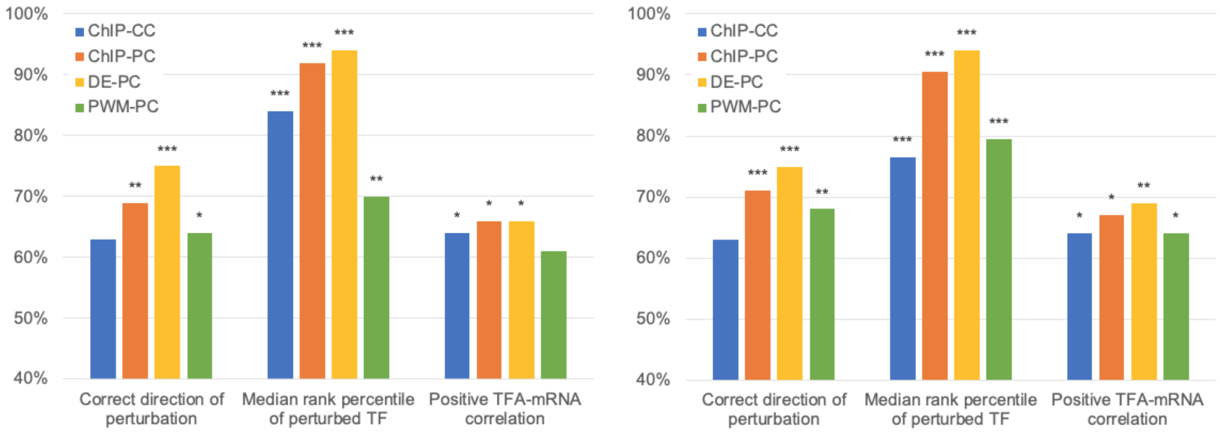


Figure 2.S2 Comparison of variance explained when using CS values optimized on a different data set (blue bars) or signed binary CS values (orange bars). Annotation below each pair of bars indicates the network and constraints used and the data set on which they were optimized. In all cases, both CS matrices are used to infer TFAs and baselines on the other data set and the variance explained is plotted. All pairs of bars show that CS matrices optimized on a different data set yield better fits than signed binary matrices, indicating that optimized CS matrices are, to some degree, transferrable from one growth condition to another.



Figures 2.S3 and 2.S4 In comparison with Figure 2.2A, Fig. 2.S3 and Fig. 2.S4 show the results from using stopping criteria alternative to the peak in variance explained for the second dataset. Fig. 2.S3 shows the results from stopping each random start when the improvement in variance explained from the last iteration drops below 0.1%. This approach shows similar trends between the networks and has the added benefit of not requiring additional data to implement. Fig. 2.S4 shows the results from using the peak in variance explained for held-out samples of the first dataset. Again, we see similar trends in performance. Asterisks above the bars indicate magnitude of significance compared to a random model, with 1, 2, or 3 asterisks representing p-value thresholds of 0.01, 0.001, or 0.0001.

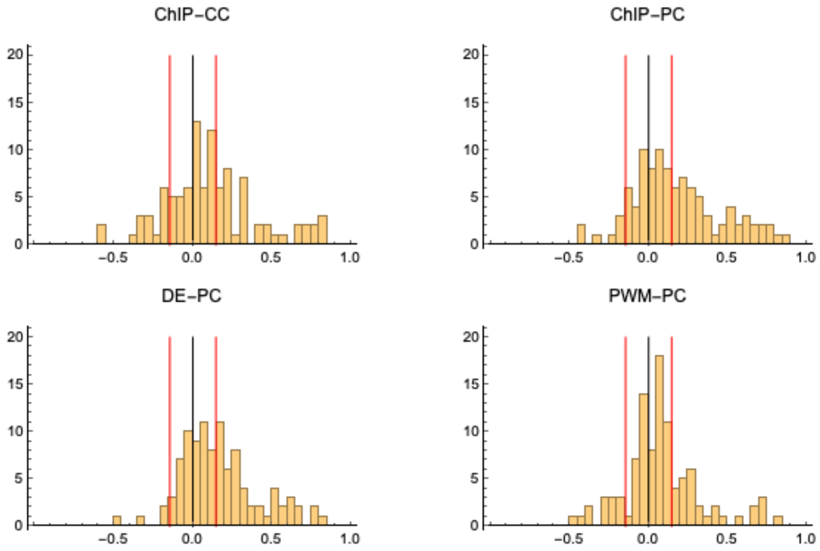


Figure 2.S5 For the *positive correlation* metric, we used bootstrapping to robustly estimate the percentage of TFs with positively correlated TFA and gene expression. This method keeps all 50 TFs in the evaluation. As a more traditional alternative, this figure shows histograms of the correlations calculated between TF activity and gene expression without bootstrapping. Correlations from evaluating both datasets are included, and the red bars at +/- 0.146 indicate where a correlation value for 180 samples passes P<=0.05. ChIP-PC and DE-PC clearly outperform ChIP-CC and PWM-PC, with more significant positive correlations, and fewer significant negative correlations.

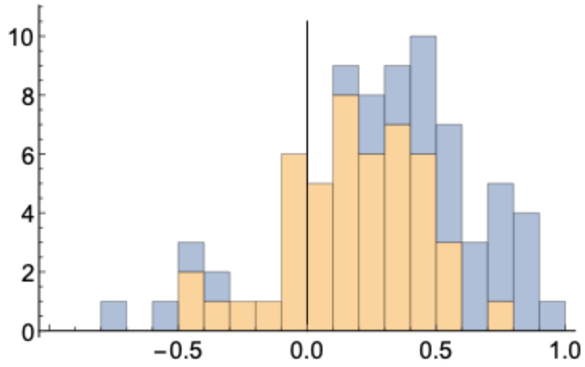


Figure 2.S6 A stacked histogram of correlations between 76 TF's CS values inferred from TFKO data and their CS values inferred from ZEV data, without bootstrapping. Blue indicates counts of TFs with significant correlation at P < 0.05. Most correlations are positive and most significant correlations are also positive.

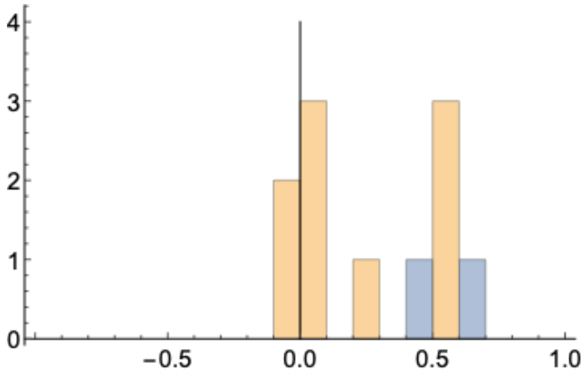


Figure 2.S7 A stacked histogram of correlations between 11 TFs' inferred CS values and Calling Cards binding signal, without bootstrapping. Blue indicates counts of TFs with significant correlation at P < 0.05. Calling Cards data is only available for 11 network TFs and correlations are not significant when considered individually because they have only a few targets over which to calculate the correlation. Nonetheless, most correlations are positive, positive correlations tend to have much greater magnitude than negative ones, and the two correlations that are significant when considered individually are strongly positive.

Chapter 3: Analysis of C. neoformans capsule production using TFA inference

3.1 Background

Cryptococcus neoformans is a virulent yeast organism that can cause cryptococcosis in the immunocompromised when its spores are inhaled (Coelho *et al*, 2014). A unique feature of the genus Cryptococcus compared to other pathogenic fungi is a capsule made of complex sugars that surrounds the cell wall. The formation of this capsule is induced by entering a host organism, where the environmental conditions differ in nutrient availability, pH, CO2 availability, temperature, and threats from the host's immune system (Caza and Kronstad, 2019). No individual signal is sufficient to trigger capsule production, although a subset can. Capsules are generally induced in labs using a set of signals such as 37C, 5% CO2, and DMEM (Delbruck's Modified Eagles' Medium), which is a chemically defined tissue culture medium. This capsule is arguably the most important virulence factor for its role in protecting cells from the host's immune response (Casadevall *et al*, 2019). It greatly enlarges during the course of infection, and mutant strains that are unable to form capsule are avirulent. As such, understanding the many overlapping signaling pathways that regulate capsule formation could be key to treatment discovery.

Transcription factor (TF) proteins play a key role in how cells respond to changes in external and internal state, and certain TFs in *C. neoformans* are known to affect capsule production. Some TF knock-out (TFKO) mutants show differences in capsule phenotype compared to wild-type strains, and regulatory network mapping has linked certain TFs to subsets of genes with known virulence functions, including capsule production (Maier *et al*, 2015). However, these analyses are not able to clarify the activity of TFs in response to environmental signals, nor are there efficient experimental methods to measure TF activity (TFA) directly. As such, the roles of TFs in the signaling pathways that trigger capsule growth in response to host-like conditions are still mostly unknown.

In the previous section of this dissertation, we defined systematic and objective evaluation metrics for validating the inferred activity values of this model using TFKO and over-expression (OE) mutants (Ma and Brent, 2020). In this work, we analyze two large gene expression datasets in *C. neoformans* by inferring TF activity (TFA) levels from measured target gene expression (Fig 3.1A). The first dataset consists of 105 TFKO strains, with an average of four replicates each, that were RNA-sequenced at 90min after introduction to capsule inducing conditions of 37C, 5% CO2, and DMEM tissue culture medium. This dataset was compiled over the past decade in collaboration with the Doering lab and this thesis is the first public analysis of the data, which we use to validate that TFA inference can identify meaningful activity changes (Fig. 3.1A), despite the change in both species and gene expression measurement technology from previous work.

After validation, we then infer TFA in a second dataset of environmental perturbations. These include samples subjected to many combinations of host-like environmental signals, such as changes in temperature, CO2 concentration, and nutrient availability (Fig. 3.1B). Each condition set was done in replicate sets of at least three, and RNA-sequenced at 30min, 90min, 180min, and 24hrs, with an additional India ink imaging step at the 24hr timepoint to measure capsule size. The inferred activity values, as well as measured capsule phenotypes for both TF KO strains and environmental conditions, are used to model the relationship between external signals, TFA, and capsule production.

In summary, we show how certain host-like conditions trigger changes in the activity of certain TFs, allowing us to hypothesize more specific links between signals and TFs, as well as make capsule phenotype predictions for knock-out mutants that had yet to be imaged. Finally, we report experimental data showing that some of those TF knockouts did in fact alter capsule size as predicted.

3.2 Results

3.2.1 Validation

Previous work has shown that TFA inference is able to accurately detect increased and decreased activity of TFs in their over-expression and knock-out mutants respectively (Ma and Brent, 2020). However, that analysis was done for the model organism of *S. Cerevisiae* (baker's yeast), using gene expression data that was measured with microarray technology. To demonstrate the applicability of TFA inference for RNA-seq measured gene expression measurements in *C. neoformans*, we first analyze a large dataset of *C. neoformans* TFKO replicates, RNA-sequenced at 90min after introduction to capsule inducing conditions of 37C, 5% CO2, and DMEM tissue culture medium, compiled over the past decade in collaboration with the Doering lab.
Leaving one replicate out per set, RNA-seq results are processed with NetProphet3.0 (Abid and Brent, unpublished) to produce a network edge scores matrix. A signed binary network of 71 TFs was created (Fig. 3.1C) with a minimum of 5 target genes per TF, and TFA inference was then

run on the entire gene expression matrix (normalized using DESeq2) to infer the activity patterns of those 71 TFs (See Methods for details). Only replicates that were left out of the network creation were included in the evaluation.

To validate the results of TFA inference, we used two of the three core evaluation metrics defined in Ma and Brent, 2020. The first is the *direction of perturbation* metric, which is the fraction of samples in which the activity of a deleted TF is inferred correctly as lower than in WT samples. For this dataset, we compared the activity of TFs in their deletion samples to the average activity in WT samples and found that 57 of the network TFs did in fact have lower activity (Fig 1D, purple points), giving us a score of 80% ($p < 0.001$). For the second metric, we first normalized all TFA values as a percentile score across samples, as a better performing alternative to the standardization of the log2 activity values that was used for *S. cerevisiae*. All TFs in a deletion sample are ranked, starting from the one whose inferred activity was most decreased relative to other samples. We expect that the deleted TFs will be highly ranked near 100%. The median rank percentile, across all perturbation samples, is the *median rank percentile* metric, which was 83% ($p < 0.001$) for this dataset. These results are significantly better than random and are comparable to the performance achieved in *S. cerevisiae* for similarly sized networks.

3.2.2 Visualizing TFA patterns in changing environmental conditions

Once validation has shown TFA inference is able to detect meaningful changes of TF activity in RNA-seq measurements of C. neoformans, we use it to analyze a dataset of environmental perturbations (Fig 1B). Replicate sets of cells were transitioned from standard laboratory growth conditions (YPD media, 30C, room air) in which capsule production is not induced, into several

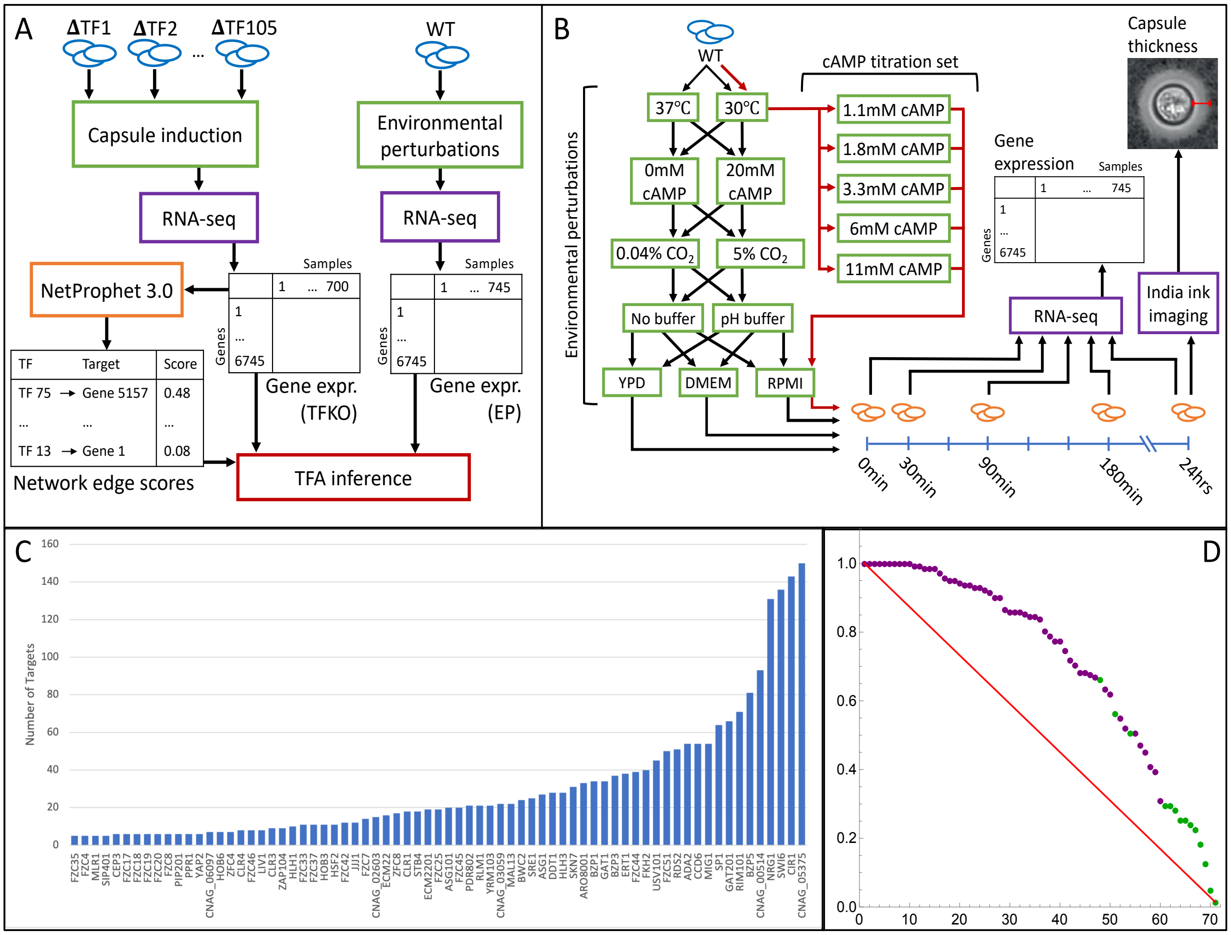


Figure 3.1. Experimental framework and validation of TFA inference. **A**: Overview of data used as inputs for TFA inference. A series of TFKO strains grown in capsule inducing conditions (left) and WT strains subject to environmental perturbations (right) were RNA-sequenced to compile large matrices of gene expression measurements. The TFKO set was analyzed with NetProphet 3.0 to create a list of network edge scores, which is used as input to TFA inference, in addition to the gene expression data, to infer the activity patterns of TFs in the KO strains and in response to a range of environmental signals. **B**: The environmental perturbations experimental design. WT samples were transitioned from YPD to a combination of other conditions as listed in the chart, before being RNA-sequenced at 5 timepoints, and phenotyped at 24hrs for capsule size phenotype using India ink imaging. **C**: The number of target genes for each of the 71 TFs in the network used for TFA inference. **D**: Validation of the TFA inference using TFKO data. Each KO sample is plotted using the percentile rank of the true perturbed TF, where 1.0 indicates it had the most significantly decreased activity compared to all other network TFs, as expected. Samples plotted in purple had lower activity in its KO strain than the median inferred activity across WT samples, while samples in green did not. The red line indicates random expectation.

combinations of environmental signals known to encourage capsule induction (Zaragoza *et al*, 2003), and samples were RNA-sequenced at 30min, 90min, 180min, and 24hr after environmental conditions changed. These RNA-seq results were processed with DESeq2 to produce a log2 normalized gene expression matrix, and TFA inference was run on this new matrix using the same 71 TF network that was validated in the previous section (See network construction under section 2.4 for details).

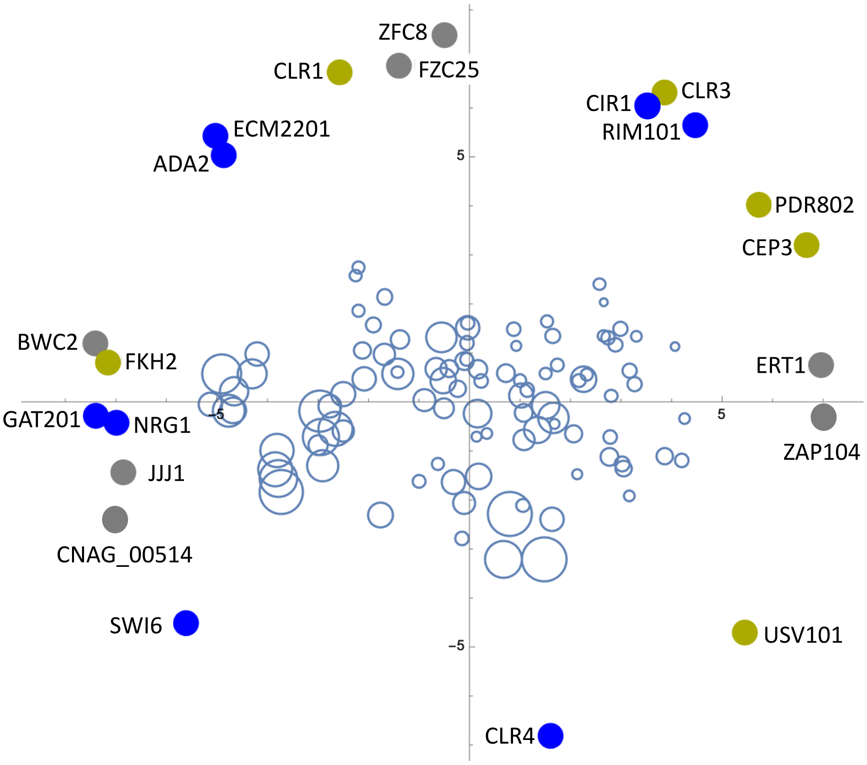


Figure 3.2. PCA plot of the activity values inferred for the environmental perturbations set, using the second (x-axis) and fourth (y-axis) principal components. The open circles in the center plot each condition replicate along PC2 and PC4, with the capsule size measured at 24hrs reflected by the size of the plotted circle. The solid circles around the border of the plot indicate the directions of the projections of selected TFs' activities at 24hrs onto PCs 2 and 4. They are colored by the capsule phenotype of their KO mutants in capsule inducing conditions (blue = capsule decrease; yellow = capsule increase; gray = unknown).

To visualize the results and how activity might relate to capsule induction, the inferred TFA matrix is re-arranged so that each (TF, timepoint) pair is a feature, and each replicate is a sample, before being transformed with Principal Components Analysis (PCA). Rather than blindly pick the top two principal components (PCs), we check each PC for correlation with mean capsule size across cells from the same conditions. This reveals PC2 and PC4 as most relevant, with PC1 coming in third, which seems to suggest that capsule production accounts for a large amount of the variance in the inferred activity values. Figure 3.2 plots each sample along PCs 2 and 4, with capsule size indicated by the size of the plotted circle. In general, capsule sizes trend larger for samples plotted to the left and to the bottom.

After choosing the PCs, we then provided orientation by projecting a subset of (TF, 24hr) features as solid circles around the plot's border, where colored TFs were selected for having known null mutant capsule phenotypes. These feature projections in combination with the sample projections can suggest directional relationships between capsule size and a TF's activity at 24hrs. PDR802's presence in the upper right quadrant where samples have small capsules indicates a negative correlation between PDR802 activity and capsule size, which is consistent with its KO mutant having an increased capsule phenotype. SWI6's presence in the bottom left quadrant indicates a positive correlation, which is consistent with its KO mutant having a decreased capsule phenotype. Among the remaining TFs with known KO mutant phenotypes (Maier *et al* 2015) that are also in the network used for TFA inference, most are similarly consistent with the capsule size trend, with the most striking exceptions being CIR1, RIM101, and FKH2.

3.2.3 Environmental signals modeling of TFA patterns

To better understand how the different signals that make up host-like conditions influence TFs in the signaling pathways of capsule induction, a multivariate linear regression model was fit to predict inferred TFA values from the environmental signals (Fig 3.1B) of the samples. This model uses YPD, 30C, room air, no cAMP, and no pH buffer as the baseline condition, and was fit to learn coefficient values for each environmental signal to predict the inferred activity values. Media (YPD, DMEM, and RPMI), temperature (30C and 37C), atmosphere (room air and 5% CO2), cAMP treatment (none or 20mM), and pH buffer (none or present) were all treated as nominal features and were included in the model as interaction terms with the timepoint feature. The pH buffer signal was found to be insignificant across almost all TFs and had little effect on capsule induction, so we decided to ignore it in this model and all following analyses. To measure the importance of each remaining signal for predicting each TF's activity pattern, we calculate the drop in variance explained when the signal is excluded. The overwhelmingly important signal was the tissue culture media, with 50 out of the 71 network TFs losing at least 10% variance explained when the signal was dropped. CO2 came second with 13 TFs, then cAMP with 9 TFs, and finally 37C with 2 TFs. Further exploration was done by adding interaction terms between tissue culture media and the other signals. Media-CO2 and Media-cAMP each explained an additional 5% or more of the variance in activity for a handful of TFs (Fig 3.3A).

Among the TFs that were included in the signal network plot of Figure 3.3A for being most predictable by environmental signals, 8 are known to have significant KO mutant capsule phenotype. Given that the signals were chosen for being capsule inducing, it's expected that positive (blue) edges, where a signal increases a TF's activity level, would predict TFs who play

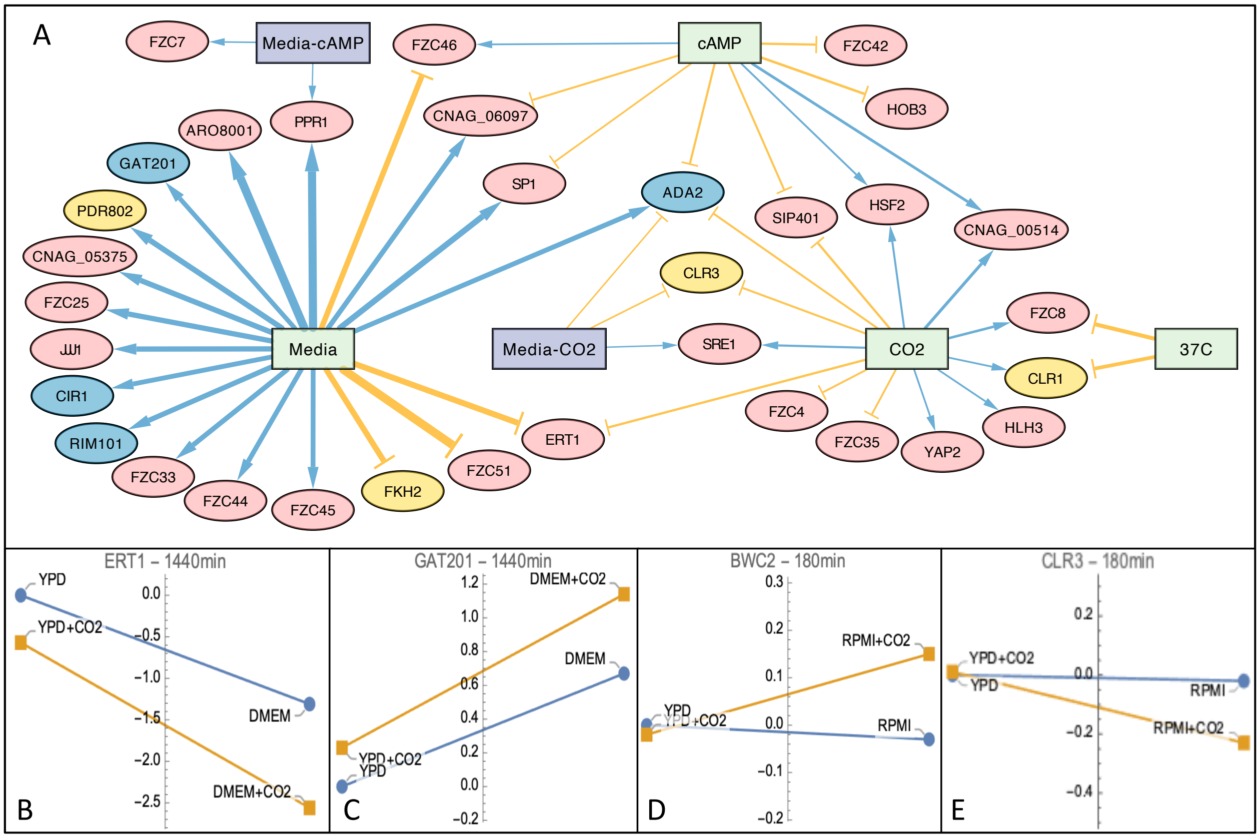


Figure 3.3. Explaining TFA values with environmental signals. **A**: A network map of the top TFs whose activity pattern is explainable through the perturbation of environmental conditions. Edges are included with thresholds chosen for easier visualization at Media > 30%, all other main effects (green rectangles) >=10%, and interaction effects (purple rectangles) > 5%. Edge thickness is weighted by the gain in variance explained when the signal information is included in the model of the TF's activity values, with blue arrowhead edges for signals that increase the TF's activity, and yellow T-head edges for signals that decrease the TF's activity. TF nodes are colored by the capsule phenotype of their KO mutants (blue = capsule decrease; yellow = capsule increase; pink = unknown). **B-D**: The effects of media, CO2, and their interaction on the activity of certain TFs. The blue vs. yellow lines indicates the presence of the 5% CO2 signal. The left side of the plot is in the baseline YPD media, while the right side is in tissue culture media (DMEM or RPMI). When CO2 and tissue culture media have independent effects on TFA (no interaction), these lines are parallel.

a role in inducing capsule production, while yellow edges would predict TFs who play a role in repressing capsule. We cannot directly verify which TFs are inducing or repressing of capsule production, but we can check for consistency with their KO phenotype. Indeed, all except for

PDR802 show a smaller capsule after deleting a TF that increases in activity in response to capsule inducing conditions, and larger capsules otherwise.

In addition to these results, ongoing capsule phenotype experiments seem to indicate that tissue culture media is necessary for capsule, and CO2 is the strongest among the other signals. As such, the TFs were analyzed for interesting patterns of interaction between these two signals, and 2 TFs with known KO mutant phenotypes as well as 5 novel TFs (the gray dots on Fig. 3.2) were selected for having one of two patterns.

The first pattern is consistency in response direction to a tissue-culture media (DMEM or RPMI), CO2, and the interaction term. ERT1, FZC25, GAT201, JJJ1, and ZFC8 follow this pattern, with two examples illustrated in Fig 3B-C. ERT1 activity at the 24hr timepoint is repressed by the CO2 signal, the DMEM tissue culture media signal, and further repressed by the interaction of both. As such, we would predict that ERT1 activity is negatively correlated with capsule production, which is consistent with its projected direction in the PCA plot. In contrast, GAT201 activity at the 24hr timepoint in induced by these signals. This pattern suggests a positive correlation between GAT201 activity and capsule production, which is consistent with its projected direction in the PCA plot, as well as its known KO mutant phenotype of decreased capsule size.

The second pattern applies to BWC2 and CLR3 (Fig. 3.3D-E), where a main effect like RPMI or CO2 is less significant than the effect of the interaction term. At 180min, both TFs have minimal response to RPMI or CO2 alone, but a significant response when both are present. BWC2 increases in activity when both signals are present, and CLR3 decreases, which is both consistent with their projected directions on the PCA plot, with CLR3 also having a known KO mutant phenotype of increased capsule size.

TF common name	TF systematic name	Variance explained		TF common name	TF systematic name	Coeff %
ERT1	CNAG_04588	67%		CLR3	CNAG_00871	6.19%
ARO8001	CNAG_04345	61%		BWC2	CNAG_02435	3.64%
CLR3	CNAG_00871	59%		CNAG_03059	CNAG_03059	3.31%
MLR1	CNAG_00031	52%		FZC51	CNAG_02877	3.20%
FZC51	CNAG_02877	49%		ASG101	CNAG_03018	3.11%
ECM22	CNAG_03710	48%		ADA2	CNAG_01626	2.80%
CNAG_00514	CNAG_00514	46%		NRG1	CNAG_05222	2.75%
GAT201	CNAG_01551	45%		FZC19	CNAG_02364	2.68%
JJJ1	CNAG_05538	43%		FZC8	CNAG_04807	2.61%
ZAP104	CNAG_05392	43%		ZAP104	CNAG_05392	2.60%

Table 3.1. Modeling capsule size with inferred TFA values. Cells are colored when KO mutant capsule phenotype is known (blue = capsule decrease; yellow = capsule increase; gray = no significant change). **(Left):** Using the inferred activity of a single TF at the four timepoints to predict the capsule size measured at 24hr, the top ten most predictive TFs by variance explained. **(Right):** Using the inferred activity of all TFs and timepoints to predict the capsule size with a ridge regression model, the top ten most predictive TFs based on the percentage of the model's absolute coefficient values attributed to each TF, summed over timepoints.

3.2.4 Modeling capsule size with TFA values

So far, we have shown that the activity of TFs can be explained by the presence of capsule inducing signals, but to complete the pathway analysis, we must also show that the activity of TFs can be used to explain capsule size. To do this, we tested two different approaches for predicting capsule size from TFA. The first was to fit linear models for predicting capsule size using the activity levels at all timepoints of only one TF at a time. Table 3.1 (Left) shows the top ten TFs, based on their model fits. The second was to use all the TFs and timepoints in a ridge regression model. Table 3.1 (Right) shows the top ten TFs, based on the percentage of absolute coefficient values attributed to them, summed over timepoints.

For both analyses, many TFs learned both positive and negative coefficients at different timepoints, so we chose not to use the directional relationships between inferred TFA values and capsule size. However, we do expect that TFs whose activity patterns are strongly predictive of capsule size are more likely to be part of the capsule production signaling pathway, and therefore have a capsule phenotype in its KO mutant. Of the 71 TFs in the network, 14 (20%) are known to have a significant capsule phenotype, 7 (10%) are known to have no significant capsule phenotype, and the remaining 50 are unknown. Given the distribution of these phenotypes among the TFs of Table 1, there is some evidence that the second approach is better at finding candidates for capsule phenotype, though the small sample size limits our ability to draw conclusions.

3.2.5 Validation through capsule phenotype prediction

To validate the compilation of all the analyses done, a set of TFs was chosen as candidates for having a capsule phenotype in their KO mutants. Ideal candidates have activity patterns that are predictive of measured capsule size and are consistently affected by capsule-inducing signals. The former suggests that deleting just that TF is highly likely to induce a change in capsule production, while the latter allows us to make a directional prediction of capsule phenotype. Table 2 lists the chosen TFs and their predicted directional role in capsule production based on all the above analyses.

To test the prediction of decreased capsule size, the candidates were grown in standard host-like conditions of DMEM tissue culture media, 37C, and 5% CO2. In this condition, capsules are induced in WT cells, as shown in the second box-and-whisker of Fig 4A. Among the candidates tested (ZAP104 results not yet available), the JJJ1 KO mutant has a capsule phenotype in the predicted direction.

TF name	CNAG ID	PCA trend b/w TFA and capsule sizes	Greatest signal effect on TFA	CO2+Media interaction effect on TFA	Capsule predictive (perTF / ridge)	KO mutant predicted capsule size
ZFC8	02700	Unclear	Positive	Positive	Rank 24 / 56	Decrease
BWC2	02435	Positive	Positive	Positive	Rank 17 / 2	Decrease
JJJ1	05538	Positive	Positive	Positive	Rank 9 / 12	Decrease
FZC25	06483	Unclear	Positive	Positive	Rank 26 / 57	Decrease
CNAG_ 00514	00514	Positive	Positive	Insignificant	Rank 7 / 17	Decrease
ERT1	04588	Negative	Negative	Negative	Rank 1 / 33	Increase
ZAP104	05392	Negative	Negative	Insignificant	Rank 10 / 10	Increase

Table 3.2. Predicting capsule phenotypes of KO mutant strains. For each TF candidate, column 3 indicates whether the projected direction of that TF's 24hr timepoint on PCs 2 and 4 trends positively or negatively with the pattern of capsule sizes (Fig 2), and columns 4 and 5 indicates whether the TF's activity is positively or negatively predicted using capsule inducing signals (Fig 3). Column 6 lists the ranks of how well the TF predicts measured capsule sizes in the two approaches outlined in the previous section. The last column is a prediction of the capsule size phenotype for experimental testing.

To test the prediction of increased capsule size, the last two candidates were grown in "almost-inducing" conditions of RPMI tissue culture media, 37C, and room air, along with two other strains that have known capsule phenotypes in standard inducing conditions. In this condition, capsules are mostly not induced in WT cells, as shown in the first box-and-whisker of Fig. 3.4B. However, there is evidence of capsule induction in the ZAP104 candidate. This is particularly evident when compared to capsule induction in the CLR3 mutant, which is known to have a larger capsule size than WT cells in the standard inducing condition (Maier *et al*, 2015).

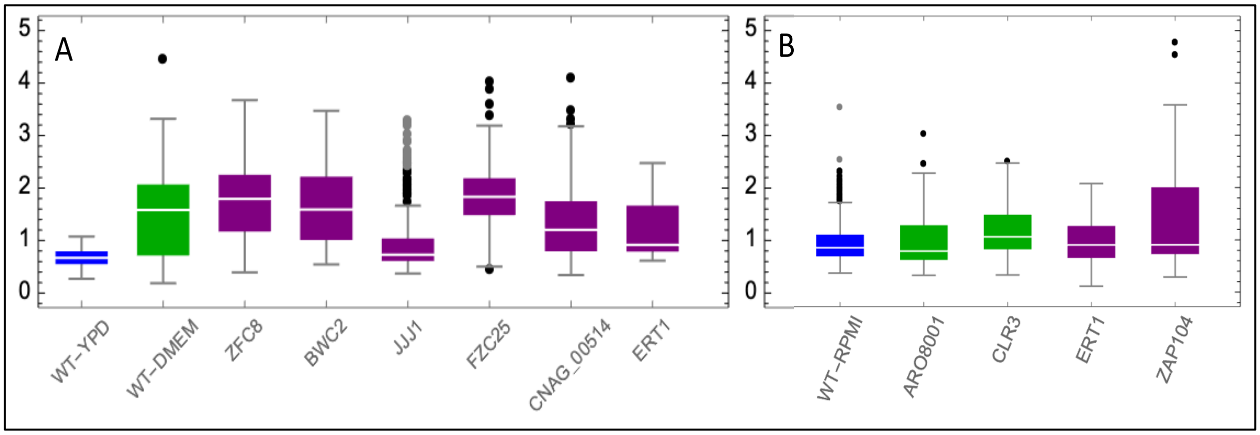


Figure 3.4. Capsule phenotype measurements of KO mutant strains. **A**: The first two box-and-whiskers are of WT cells in conditions that don't (YPD, 30C, room air) and do (DMEM, 37C, 5% CO2) induce capsule. The remaining box-and-whiskers are cells from KO mutants that were grown in the same capsule inducing condition as WT-DMEM. **B**: The first box-and-whiskers is of WT cells in the almost-inducing condition (RPMI, 37C, room air). The remaining box-and-whiskers are cells from KO mutants that were grown in the same almost-inducing condition as WT-RPMI.

3.3 Discussion

TFA inference enables systematic and efficient analysis of how TFs operate in the signaling pathways that control response to host-like conditions, far beyond the insights that experiments like phenotyping can give us. In this paper, we prove that TFA inference can recover meaningful values in *C. neoformans*, and then use TFA values inferred from a large dataset of environmental perturbations to link environmental signals to TFA, and TFA to capsule size. This allows us to speculate about how each TF is potentially involved in capsule production, as well as make phenotype predictions for un-tested KO mutants, two of which were experimentally confirmed in the correct direction.

One challenge of this work was transitioning TFA inference from S. cerevisiae microarray data to *C. neoformans* RNA-sequencing data. With minor tweaks to how TFA values are normalized between inference and analysis, adjustments to how the input network map was defined, and batch-normalizing the RNA-seq data with DESeq2 (see Methods for details), it was possible to get comparably good performance for two evaluation metrics, *direction of perturbation* and *median rank percentiles*, as explained in the first subsection of Results. However, there was a third metric which checks for more positive than negative correlations between inferred activity values and measured TF gene expression values. This metric could not be optimized to be consistently better than random without hurting the performance of the other metrics or shrinking the network size to 50 TFs. In the end, we chose to drop this metric, but we hope to explore this discrepancy further to find an explanation.

By inferred TFA values across the environmental perturbation samples, we could define how each TF was reacting to each condition set at the varying time points. This opened the possibility of directly analyzing these values as if we had knowledge of the intermediaries between the experimentally controlled environmental signals, and the experimentally measured capsule phenotypes. Our PCA visualization could link known capsule phenotypes to capsule size trends using just two principal components of TFA, and our models of TFA response to environmental signals could link known capsule phenotypes to known capsule inducing signals. With just these two methods, we could hypothesize which TFs are likely involved in capsule production and which TFs carry specific environmental signals. In future work, these results can be further analyzed as time-courses to fill in the dynamics of how TFs propagate through signal pathways to finally induce capsule production, though it's unclear how this could be validated.

For this paper, we chose to make capsule phenotype predictions from our analyses to enable easier validation. Most TFs are not expected to have a capsule phenotype, and several have already been identified. However, by filtering through our results for TFs whose activity are consistently influenced by environmental signals and highly ranked in being able to predict capsule size, we identified six candidates for capsule phenotype. In the end, we were able to confirm novel capsule phenotypes, one in each direction. This success is indicative of the value of TFA inference to elucidate the inner workings of cell response.

Chapter 4: Conclusions

4.1 Summary

Cells process and respond to information, in part, by changing the activity levels of TFs and the production rates of the genes they regulate. Since TFA levels cannot be easily measured experimentally, inferring them from gene expression measurements is a promising approach towards better understanding how much influence TFs have on genes of interest, how much the activity of TFs is influenced by other proteins in cellular systems, and the role TFs play in cellular signaling.

To make TFA inference a practical tool for studying gene regulation, my thesis work accomplished three main tasks. The first was defining accuracy metrics for TFA inference by implementing a proof of concept, using the large amount of data available for the well-studied model organism of S. cerevisiae. The second was to prove TFA inference as a method was robust to changing the model organism from S. cerevisiae to C. neoformans, changing the collection of input gene expression data from microarray to RNA-sequencing technology, as well as changing the optimizers used to fit the model. Finally, TFA inference was applied to the real-world problem of understanding C. neoformans capsule formation. With a database of gene expression levels and capsule size measurements collected over the last decade, TFA inference allowed us to further disentangle the signaling pathways of C. neoformans and its capsule formation response to host-like conditions. Using the insights from these analyses, we were able to successfully predict changes in capsule size for TFKO mutants, demonstrating the power of TFA inference to find causal relationships between TFs and capsule phenotype.

4.2 Limitations of TFA inference and evaluation

TFA inference works to find meaningful changes in activity from gene expression data, and by establishing evaluation metrics, it was possible to confirm that in multiple datasets, from different organisms and different data measurement technologies. However, our explorations did reveal some limitations.

While testing different optimization software packages, we found it was possible to improve the TFA inference model's fit to gene expression data when using a non-linear optimization approach over the alternating least squares approach. However, better fit did not correlate with better performance on the TFA inference metrics. To test if this was an overfitting problem, we also used a cross-validation approach to fit the model with L2 regularization. Cross-validation for a matrix factorization model like this is a little different from traditional cross-validation. Because holding out samples would mean we can't infer TFA parameter values for those samples and holding out genes would mean we can't infer CS parameter values for those genes, we have to hold out values that are distributed across the set such that all TFs still have some subset of their target gene measurements left in for each sample. Using this method, a shrinkage constraint can be selected to optimize model fit on held-out data, but this too did not correlate with better performance on the accuracy metrics. This seems to indicate that the model itself is too simple to explain the relationship between TFA and target gene expression, which isn't too surprising. Some concepts that the model cannot explain include changing regulatory influence between TFs and target genes when the environmental conditions change, non-linear regulatory interactions between TFs, and the dynamics of a sample from a time course compared to a sample in steady state. Increasing the complexity of the model could help address the gaps.

There are also limitations to the evaluation metrics themselves. While we designed the metrics to range between 50-100%, where 50% is random expectation and 100% is perfect performance, there is reason to believe the true perfect performance is much lower than 100%. For the *Direction of Perturbation* metric, we assume that direct perturbation of the TF encoding gene will lead to a measurable change in the activity of that TF in the direction of perturbation. However, it's entirely possible that the TF was already inactive before deletion, or that the TF was already maximally active before induction. These scenarios would result in no real change in TFA, so inferring any such change to get a higher score on this metric would actually be less accurate. The *Median Rank Percentile* metric is similarly limited, with the additional possibility that perturbing one TF could lead to a significant change in the activity of other TFs that are downstream in a signaling pathway. In a such a scenario, it may be true that the downstream TFs' activity change more than the directly perturbed TF, leading to lower metric scores that fail to reflect the accuracy of such a prediction. Lastly, we have the *Positive correlation*. As stated when we first introduced it in section 2.3, we do not expect accurate TFA values to yield 100% for this metric, since post-transcriptional regulation is a key determinant of TFA levels. We ultimately could not even find the fraction of positive correlations to be substantially greater than half for *C. neoformans*, but considering the success of other analyses, it does not seem like poor performance on this metric revealed a substantial failure of TFA inference.

For now, we conclude that the choice of optimizer does not substantially impact TFA inference as currently implemented. For next steps, the model itself should be updated to include more nuanced relationships between TFs and their target genes. As for the TFA evaluation metrics, there's reason to believe that their defined upper limit of 100% does not reflect the reality of TFA, and therefore should not be used as the goal of perfect TFA inference accuracy. Further research into the frequency of speculated exceptions to the metric assumptions would be useful for putting the metrics performance scores into perspective.

4.3 Future directions

Beyond the endless possibilities of tweaking the model, analyzing new datasets, and analyzing new organisms, the field of TFA inference will be greatly impacted by newer technologies being developed to measure biological output. One such technology is single-cell RNA-sequencing (scRNAseq), which can measure the expression within individual cells, instead of the expression from a bulk sample of cells. While the output tends to be sparser and noisier, scRNAseq gives us much higher resolution measurements of how cells are responding to changes in external and internal state. TFA inference in these datasets could capture changes in TFA that would otherwise be averaged out across cells, and preliminary tests have resulted in significantly better than random results on the *Median Rank Percentile* metric (results not shown).

Another new technology is nascent RNA-sequencing, including GRO-seq (Gardini, 2017) and PRO-seq (Mahat *et al.*, 2016). These methods can measure RNA that is currently being transcribed, allowing us to measure real-time transcription without the confounding factor of unequal RNA accumulation, as well as measure transcription of RNA that is not stable enough to be measured by other RNA-sequencing methods. TF influence on target gene expression is mediated through the binding of promoters, as well as through binding of more distal regions called enhancers, with the latter more common in more complex organisms. The measurement of enhancer RNA (eRNA) as it is transcribed can be understood as a measurement of enhancer activity, and there is an argument to be made that the activity of enhancers and promoters measured through nascent RNA-sequencing would be more tightly linked to TFA than the expression of TF target genes (Azofeifa *et al*., 2018; Wang *et al*., 2018). By updating the model to replace stable target gene expression levels in favor of directly using the activity of the enhancers and promoters, we could remove several confounding factors, such as the uncertain assignment of target genes to enhancers. However, it's still necessary to assign the enhancers and promoters to TFs through techniques like PWMs, and data from these new techniques is limited, so it's not likely that this approach can replace TFA inference yet. As these datasets grow, however, we look forward to being able to apply our metrics to objectively evaluate any benefits to this alternative data.

References

Alvarez, M.J. *et al.* (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet*., 48, 838–847.

Apweiler, E.*, et al.* (2012) Yeast glucose pathways converge on the transcriptional regulation of trehalose biosynthesis. *BMC Genomics*, **13**, 239.

Arrieta-Ortiz, M.L.*, et al.* (2015) An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Mol. Syst. Biol.*, **11**, 839.

Avendano, A. *et al.*,(2005) Swi/SNF-GCN5-dependent chromatin remodelling determines induced expression of GDH3, one of the paralogous genes responsible for ammonium assimilation and glutamate biosynthesis in Saccharomyces cerevisiae. *Mol. Microbiol.* **57**, 291-305

Azofeifa, J.G.*, et al.* (2018) Enhancer RNA profiling predicts transcription factor activity. *Genome Res.*

Balwierz, P.J.*, et al.* (2014) ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.*, **24**, 869-884.

Barenco, M.*, et al.* (2009) rHVDM: an R package to predict the activity and targets of a transcription factor. *Bioinformatics*, **25**, 419-420.

Barenco, M.*, et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.*, **7**, R25.

Berchtold, E.*, et al.* (2016) Evaluating Transcription Factor Activity Changes by Scoring Unexplained Target Genes in Expression Data. *PLoS One*, **11**, e0164513.

Bergenholm, D.*, et al.* (2018) Reconstruction of a Global Transcriptional Regulatory Network for Control of Lipid Metabolism in Yeast by Using Chromatin Immunoprecipitation with Lambda Exonuclease Digestion. *mSystems*, **3**.

Bertram, P.G. *et al.* (2002) Convergence of TOR-nitrogen and Snf1-glucose signaling pathways onto Gln3. *Mol. Cell. Biol.* **22**, 1246-1252

Bodenhofer, U.*, et al.* (2011) APCluster: an R package for affinity propagation clustering. *Bioinformatics*, **27**, 2463-2464.

Boorsma, A. et al. (2008) Inferring condition-specific modulation of transcription factor activity in yeast through regulon-based analysis of genomewide expression. *PLoS One*, 3, e3112.

Boscolo, R.*, et al.* (2005) A generalized framework for network component analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **2**, 289-301.

Boulesteix A.L. and Strimmer K. (2005) Predicting transcription factor activities from combined analysis of microarray and ChIP data: a partial least squares approach. *Theor. Biol. Med. Model.*, 2, 23.

Brent, M.R. (2016) Past Roadblocks and New Opportunities in Transcription Factor Network Mapping. *Trends Genet*, **32**, 736-750.

Broach, J.R. (2012) Nutritional control of growth and development in yeast. *Genetics*, **192**, 73-105.

Bussemaker, H.J*., et al.* (2017) Network-based approaches that exploit inferred transcription factor activity to analyze the impact of genetic variation on gene expression. *Curr Opin Syst Biol*, **2**, 98-102.

Bussemaker, H.J*., et al.* (2001) Regulatory element detection using correlation with expression. *Nat Genet*, **27**, 167-171.

Byrd, R*., et al.* Knitro : An Integrated Package for Nonlinear Optimization. In.; 2006. p. 35-59.

Casadevall, A. *et al.* (2019). The capsule of *Cryptococcus neoformans*. *Virulence*, *10*(1), 822–831.

Chechik, G. and Koller, D. (2009) Timing of gene expression responses to environmental changes. *J. Comput. Biol.*, **16**, 279-290.

Chen J. *et al*. (2013) Genome-wide signatures of transcription factor activity: connecting transcription factors, disease, and small molecules. *PLoS Comput. Biol*., 9, e1003198.

Chen, Y*., et al.* (2017) Systems-epigenomics inference of transcription factor activity implicates aryl-hydrocarbon-receptor inactivation as a key event in lung cancer development. *Genome Biol.*, **18**, 236.

Cheng, C*., et al.* (2007) Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinformatics*, **8**, 452.

Coelho, C. *et al.* (2014) The tools for virulence of Cryptococcus neoformans. *Adv Appl Microbiol* **87**, 1-41

Cokus, S*., et al.* (2006) Modelling the network of cell cycle transcription factors in the yeast Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 381.

Conlon, E.M*., et al.* (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 3339-3344.

Conrad, M*., et al.* (2014) Nutrient sensing and signaling in the yeast Saccharomyces cerevisiae. *FEMS Microbiol. Rev.*, **38**, 254-299.

De Virgilio, C. (2012) The essence of yeast quiescence. *FEMS Microbiol. Rev.*, **36**, 306-339.

De Wever, C. *et al.* (2005) A dual role for PP1 in shaping the Msn2-dependent transcriptional response to glucose starvation. *EMBO J.* **24**, 4115-4123

Fisher, R.A. (1954) Statistical methods for research workers. Edinburgh: Oliver and Boyd.

Fröhlich H. (2015) biRte: Bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics*, 31, 3290–3298.

Fu Y. *et al*. (2011) Reconstructing genome-wide regulatory network of E. coli using transcriptome data and predicted transcription factor activities. *BMC Bioinformatics*, 12, 233.

Gao, F*., et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Garcia-Alonso, L*., et al.* (2019) Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome Res.*, **29**, 1363-1375.

Garcia-Alonso, L*., et al.* (2018) Transcription Factor Activities Enhance Markers of Drug Sensitivity in Cancer. *Cancer Res.*, **78**, 769-780.

Gardini A. (2017) Global Run-On Sequencing (GRO-Seq). *Methods Mol Biol*. 1468:111-120. doi:10.1007/978-1-4939-4035-6_9

Gitter, A*., et al.* (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. *Genome Res.*, **23**, 365-376.

Grant, C.E*., et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017-1018.

Gurobi Optimization, Gurobi Optimizer Reference Manual. (2020)

Hackett, S.R*., et al.* (2020) Learning causal networks using inducible transcription factors and transcriptome-wide time series. *Mol. Syst. Biol.*, **16**, e9174.

Harbison, C.T*., et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99-104.

Haynes BC, Skowyra ML, Spencer SJ, Gish SR, Williams M, Held EP, Brent MR, Doering TL. 2011. Toward an integrated model of capsule regula- tion in Cryptococcus neoformans. PLoS Pathog 7: e1002411.

Holland, P*., et al.* (2019) Predictive models of eukaryotic transcriptional regulation reveals changes in transcription factor roles and promoter usage between metabolic conditions. *Nucleic Acids Res*, **47**, 4986-5000.

Jiang, P*., et al.* (2015) Inference of transcriptional regulation in cancers. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 7731-7736.

Jung, K.W. *et al*. (2015) Systematic functional profiling of transcription factor networks in Cryptococcus neoformans. *Nat Commun*.

Kang, Y.*, et al.* (2020) Dual threshold optimization and network inference reveal convergent evidence from TF binding locations and TF perturbation responses. *Genome Res.*, **30**, 459-471.

Kemmeren, P.*, et al.* (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**, 740-752.

Khanin, R.*, et al.* (2007) Statistical reconstruction of transcription factor activity using Michaelis-Menten kinetics. *Biometrics*, **63**, 816-823.

Lam, K.Y.*, et al.* (2016) Fused Regression for Multi-source Gene Regulatory Network Inference. *PLoS Comput. Biol.*, **12**, e1005157.

Lee, E. and Bussemaker, H.J. (2010) Identifying the genetic determinants of transcription factor activity. *Mol. Syst. Biol.*, **6**, 412.

Li, Y.*, et al.* (2014) Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput. Biol.*, **10**, e1003908.

Liao J.C. et al. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA*, 100, 15522–15527.

Liao, Y.*, et al.* (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.*, **47**, W199-W205.

Liu, O.W. *et al*. (2008) Systematic genetic analysis of virulence in the human fungal pathogen Cryptococcus neoformans. *Cell*. 135(1):174-88.

Ljungdahl, P.O. and Daignan-Fornier, B. (2012) Regulation of amino acid, nucleotide, and phosphate metabolism in Saccharomyces cerevisiae. *Genetics*, **190**, 885-929.

Ma, C.Z. and Brent, M. R. (2020) Inferring TF activities and activity regulators from gene expression data with constraints from TF perturbation data, *Bioinformatics*, Volume 37, Issue 9, Pages 1234-1245,

Mahat, DB *et al. (2016)* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). Nat Protoc. 2016

Maier, E.J. *et al.*, Model-driven mapping of transcriptional networks reveals the circuitry and dynamics of virulence regulation. *Genome Res* **25**, 690-700 (2015)

Mayhew, D. and Mitra, R.D. (2016) Transposon Calling Cards. *Cold Spring Harb Protoc*, **2016**, pdb top077776.

Nachman I. *et al*. (2004) Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20, i248–256.

Nielsen K, Cox *et al*. (2005). Cryptococcus neoformans α strains preferentially disseminate to the cen- tral nervous system during coinfection. Infect Immun 73: 4922–4933

Ocone A. , Sanguinetti G. (2011) Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, 27, 2873–2879.

Polish, J.A*., et al.* (2005) How the Rgt1 transcription factor of Saccharomyces cerevisiae is regulated by glucose. *Genetics*, **169**, 583-594.

Riego, L. *et al.* (2002) GDH1 expression is regulated by GLN3, GCN4, and HAP4 under respiratory growth. *Biochemical and Biophysical Research Communications* **293**, 79-85

Rodkaer, S.V. and Faergeman, N.J. (2014) Glucose- and nitrogen sensing and regulatory mechanisms in Saccharomyces cerevisiae. *FEMS Yeast Res*, **14**, 683-696.

Rogers, S*., et al.* (2007) Bayesian model-based inference of transcription factor activity. *BMC Bioinformatics*, **8 Suppl 2**, S2.

Ronen, M. and Botstein, D. (2006) Transcriptional response of steady-state yeast cultures to transient perturbations in carbon source. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 389-394.

Sameith, K*., et al.* (2015) A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions. *BMC Biol.*, **13**, 112.

Sanguinetti, G*., et al.* (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775-2781.

Schacht T. *et al*. (2014) Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*, 30, i401–407.

Schwarz, G. (1978) Estimating the Dimension of a Model. *Ann. Statist.*, **6**, 461-464.

Shi, Y*., et al.* (2009) A combined expression-interaction model for inferring the temporal activity of transcription factors. *J. Comput. Biol.*, **16**, 1035-1049.

Shively, C.A*., et al.* (2019) Homotypic cooperativity and collective binding are determinants of bHLH specificity and function. *Proc. Natl. Acad. Sci. U. S. A.*, **116**, 16143-16152.

Spielewoy, N. *et al.* (2010) Npr2, yeast homolog of the human tumor suppressor NPRL2, is a target of Grr1 required for adaptation to growth on diverse nitrogen sources. *Eukaryotic cell* **9**, 592-601

Spivak, A.T. and Stormo, G.D. (2012) ScerTF: a comprehensive database of benchmarked position weight matrices for Saccharomyces species. *Nucleic Acids Res*, **40**, D162-168.

Staschke, K. *et al.* (2010) Integration of general amino acid control and target of rapamycin (TOR) regulatory pathways in nitrogen assimilation in yeast. *J. Biol. Chem.* **285**, 16893-16911

Taub, F. (1983) "Laboratory methods: Sequential comparative hybridizations analyzed by computerized image processing can identify and quantitate regulated RNAs". *DNA*. **2** (4): 309–327

Tchourine, K*., et al.* (2018) Condition-Specific Modeling of Biophysical Parameters Advances Inference of Regulatory Networks. *Cell Rep*, **23**, 376-388.

Tran, L.M*., et al.* (2005) gNCA: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab Eng*, **7**, 128-141.

Trescher, S. and Leser, U. (2019) Estimation of Transcription Factor Activity in Knockdown Studies. *Sci. Rep.*, **9**, 9593.

Tripodi, I.J*., et al.* (2018) Detecting Differential Transcription Factor Activity from ATAC-Seq Data. *Molecules*, **23**.

Wächter, A. and Biegler, L.T. (2006) On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57

Wang, C*., et al.* (2008) Motif-directed network component analysis for regulatory network inference. *BMC Bioinformatics*, **9 Suppl 1**, S21.

Wang, H*., et al.* (2011) Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins. *Genome Res.*, **21**, 748-755.

Wang, J *et al.* (2018) Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. *BMC Genomics* **19,** 633

Yang, Y.L*., et al.* (2005) Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics*, **6**, 90.

Yu, T. and Li, K.C. (2005) Inference of transcriptional regulatory network by two-stage constrained space factor analysis. *Bioinformatics*, **21**, 4033-4038.

Zaman, S*., et al.*(2008) How Saccharomyces responds to nutrients. *Annu. Rev. Genet.* **42**, 27-81

Zaman, S*., et al.* (2009) Glucose regulates transcription in yeast through a network of signaling pathways. *Mol. Syst. Biol.*, **5**, 245.

Zaragoza, O. *et al*. (2003). Induction of capsule growth in Cryptococcus neoformans by mammalian serum and CO(2). *Infection and immunity*, *71*(11), 6155–6164. https://doi.org/10.1128/IAI.71.11.6155-6164.2003

Zhu M. *et al*. (2013) REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC Genomics*, 14, 504.