Explicit Modeling of RNA Stability Improves Large-Scale Inference of Transcription Regulation

Konstantine Tchourine[1,2], Christine Vogel[*1,2], and Richard Bonneau[†1,2,3,4]

[1]Center for Genomics and Systems Biology, New York University, New York, NY 10003, USA
[2]Biology Department, New York University, New York, NY 10003, USA
[3]Courant Institute of Mathematical Sciences, Computer Science Department, New York University
[4]Flatiron Institute, Center for Computational Biology, Simons Foundation, New York, NY 10010, USA

January 31, 2017

Abstract

Inference of eukaryotic transcription regulatory networks remains challenging due to the large number of regulators, combinatorial interactions, and redundant pathways. Even in the model system *Saccharomyces cerevisiae*, inference has performed poorly. Most existing inference algorithms ignore crucial regulatory components, like RNA stability and post-transcriptional modulation of regulators. Here we demonstrate that explicitly modeling transcription factor activity and RNA half-lives during inference of a genome-wide transcription regulatory network in yeast not only advances prediction performance, but also produces new insights into gene- and condition-specific variation of RNA stability. We curated a high quality gold standard reference network that we use for priors on network structure and model validation. We incorporate variation of RNA half-lives into the *Inferelator* inference framework, and show improved performance over previously described algorithms and over implementations of the algorithm that do not model RNA degradation. We recapitulate known condition- and gene-specific trends in RNA half-lives, and make new predictions about RNA half-lives that are confirmed by experimental data.

1 Introduction

Inference of large-scale transcriptional regulatory networks (TRN) is an active research area, with broad applications in basic science and biomedical research. The key underlying assumption of TRN inference is that changes in RNA expression levels are informative of regulatory relationships between transcription factors (TFs) and their target genes. Information from expression data is often complemented with orthogonal data on protein-protein and protein-DNA interactions, such as from protein binding assays (Valouev *et al.*, 2008; Mundade *et al.*, 2014), DNA accessibility assays like FAIRE-Seq and ATAC-seq (Davie *et al.*, 2015), and motif enrichment analysis (Setty *et al.*, 2015; Guo *et al.*, 2012). Various machine learning approaches are used to infer regulatory networks. They have multiple levels of model complexity, ranging from the earliest Boolean network and network module approaches (Shmulevich *et al.*, 2002; Lähdesmäki *et al.*, 2003; Segal *et al.*, 2003; Pe'er *et al.*, 2001) to approaches that explicitly model dynamics, TF interactions and transcription factor post-transcriptional activity (Honkela *et al.*, 2010; Äijö *et al.*, 2013; Intosalmi *et al.*, 2016; Studham *et al.*, 2014). Advancements in genomics and transcriptomics technologies spurred the development of more complex methods, involving Mutual Information (Margolin *et al.*, 2006a,b; Faith *et al.*, 2007; Butte & Kohane, 2000), correlation (Butte & Kohane, 2000), ANOVA (Küffner *et al.*, 2012), conditional entropy (Karlebach & Shamir, 2012), Random Forest (Huynh-Thu *et al.*, 2010; Petralia *et al.*, 2015), Bayesian causality (Mani & Cooper, 2004; Mani *et al.*, 2012; Friedman *et al.*, 2000), expression module clustering (Reiss *et al.*, 2006, 2015), and constrained regression of biophysical models (Bonneau *et al.*, 2006; Greenfield *et al.*, 2013; Arrieta-Ortiz *et al.*, 2015).

[*]cvogel@nyu.edu
[†]rbonneau@simonsfoundation.org

Importantly, recent comprehensive blind assessments of these approaches, called DREAM4 and DREAM5, concluded that there is no single machine learning category of methods outperforming all others (Marbach *et al.*, 2012), and that inference in eukaryotes is systematically much more challenging than in prokaryotes. For example, despite the abundance of data, prediction in yeast *Saccharomyces cerevisiae* only reaches an Area Under Precision-Recall curve (AUPR) of 0.025 and fluctuates around random performance, compared to 0.05 to 0.1 in the bacterium *Escherichia coli*, in which it exceeded random performance several-fold. While DREAM challenges are limited because they did not allow methods to include prior interaction data, results from recent studies that incorporated prior interaction data also dramatically differed between prokaryotes and eukaryotes (Greenfield *et al.*, 2013; Wilkins *et al.*, 2016; Siahpirani & Roy, 2016; Bonneau & Aijo, 2016).

The challenges in inference of eukaryotic networks lie in extensive post-transcriptional regulation. For example, the eukaryotic genome encodes many transcription factors, complex promoter regions, and redundant regulatory pathways. Eukaryotes post-process their RNA and have extensive RNA degradation regulation. However, most inference methods, such as random forest (Huynh-Thu *et al.*, 2010; Petralia *et al.*, 2015), mutual information and related transfer entropy (Margolin *et al.*, 2006a,b), and correlation (Butte & Kohane, 2000), do not explicitly model biophysical parameters involved in expression regulation, and thereby cannot address the increased complexity of eukaryotic expression regulation in a straightforward manner.

The *Inferelator* is a method based on constrained regression (Bonneau *et al.*, 2006; Greenfield *et al.*, 2013). It is distinct from other large-scale inference methods as it allows explicit modeling of biophysical parameters (Equation 1). However, we and others have recently shown that inference of transcription- and translation-related biophysical parameters via ordinary differential equations produces robust genome-wide models of expression changes in response to various stresses in various organisms (Tchourine *et al.*, 2014; Schwanhäusser *et al.*, 2013; Peshkin *et al.*, 2015). Recent modifications to the Inferelator algorithm, such as robust incorporation of priors (Greenfield *et al.*, 2013) and condition-specific TF activity (TFA) estimation (Liao *et al.*, 2003; Arrieta-Ortiz *et al.*, 2015) dramatically improved the *Inferelator* performance in prokaryotes, boosting the Inferelator performance in *B. subtilis* from 0.1 to 0.48 in terms of AUPR. The effect of these modifications in eukaryotic data sets has only been tested in rice (*Oriza sativa*), with modest results (Wilkins *et al.*, 2016) - highlighting the need for new developments that include parameters that account for additional regulatory pathways.

One such crucial regulatory component that is typically ignored (or convolved with other parameters) is RNA degradation. While traditional RNA half-life measurements involved transcription inhibition (Wang *et al.*, 2002; Grigull *et al.*, 2004; Shalem *et al.*, 2008), newer approaches use non-invasive metabolic labeling (Miller *et al.*, 2011; Schwalb *et al.*, 2012; Neymotin *et al.*, 2014). For yeast, several experimental datasets have emerged recently, highlighting the large range in RNA half-lives and their extensive regulation under different conditions and in different genetic backgrounds (Miller *et al.*, 2011; Schwalb *et al.*, 2012; Sun *et al.*, 2012; Neymotin *et al.*, 2014; Munchel *et al.*, 2011). For example, typical RNA half-lives in yeast range between 10 and 18 minutes, while ribosomal genes are twice as long-lived (Neymotin *et al.*, 2014; Munchel *et al.*, 2011; Miller *et al.*, 2011). However, under glucose starvation or rapamycin stress, this relationship becomes inverted, and RNA half-lives systematically stabilize, except for those of ribosomal genes, which become less stable (Munchel *et al.*, 2011). In addition, evidence suggests extensive feedback between transcription and degradation regulation: RNA degradation rates across 46 mutant yeast strains had a strong positive correlation with transcription rates (Sun *et al.*, 2013).

Here, we present the first TRN inference approach that explicitly models RNA half-lives and demonstrate its ability to significantly improve prediction performance. To achieve this result, we first constructed a new, high-quality gold standard of signed transcription regulatory interactions. Then we clustered the 2,577 expression data sets into 20 bi-clusters and showed that combining individually inferred networks outperforms global modeling. We optimized the RNA degradation term for each bi-cluster and showed that the incorporation of this step into modeling further improved performance. Finally, we showed that optimizing network inference for each bi-cluster also results in accurate condition- and gene-specific RNA decay rate predictions. Our final prediction has an AUPR of 0.328, far larger than in other existing work in yeast.

2 Materials and Methods

2.1 Data Acquisition and Normalization

We acquired four prior known regulatory interaction data sets from various sources, as listed in Table S1, originating predominantly from ChIP-chip, ChIP-seq, knock-out, and overexpression assays (Teixeira *et al.*, 2006; Monteiro *et al.*, 2008; Abdulrehman *et al.*, 2011; Teixeira *et al.*, 2014; Cherry *et al.*, 2012; Costanzo *et al.*, 2014; Kemmeren *et al.*, 2014). The list of 563 TFs was assembled by including all gene names that were marked as either "DNA-binding"

or "Regulation of transcription, DNA-templated" in Saccharomyces Genomes Database (SGD) (Cherry *et al.*, 2012; Costanzo *et al.*, 2014), as well as all regulators in the YEASTRACT database of regulatory interactions (Teixeira *et al.*, 2006; Monteiro *et al.*, 2008; Abdulrehman *et al.*, 2011; Teixeira *et al.*, 2014). We downloaded 179 RNA expression data sets from 119 different labs, from Gene Expression Omnibus (GEO) (Edgar *et al.*, 2002; Barrett *et al.*, 2013) using the R Bioconductor package `GEOquery` (Huber *et al.*, 2015; Davis & Meltzer, 2007). To obtain a high-quality, consistent data set, and to avoid platform-specific batch effects, we exclusively used the Yeast Affymetrix 2.0 platform (GPL2529) for the analysis, as it contained the largest number of unique samples in the GEO database. Raw CEL files for every sample (GSM) measured on this platform were downloaded on March 23, 2015, along with the meta-data for each GSM. We processed and normalized the raw CEL files using the R packages `affy` (Gautier *et al.*, 2004) and `gcrma` (Wu *et al.*, 2004), adjusting for background intensities, optical noise, and non-specific binding in a probe sequence specific fashion. The meta-data was processed manually to identify samples that belonged to time series experiments. The full meta-data, as downloaded from the GEO website, is included in the Supplementary File SuppData1.zip. The final RNA expression data set included 2,577 samples, each containing the expression data for 5,716 genes.

2.2 Expression Data Clustering

The expression data was further scaled such that every row (gene) had mean 0 and variance 1. The 2,577 expression samples were then clustered using the Euclidean distance metric in R programming language. We first performed PCA on the entire RNA expression matrix, and removed all but the first 16 dimensions to facilitate condition-wise clustering and remove the cumulative effect of noisy low-variance components. We then performed k-means clustering with $k = 4$. All downstream analysis was performed on the resulting clusters using the original (normalized but unscaled) expression data. The number of clusters was optimized as described in Section 6.1.2.

To annotate the four condition clusters, we used meta-data that we downloaded from GEO for each sample to determine common biological themes, employing the R packages `tm` (Feinerer & Hornik, 2015; Feinerer *et al.*, 2008) and `SnowballC` (Bouchet-Valat, 2014). First, we used the binomial test to determine which words are enriched in a given condition cluster as compared to the remaining clusters. To avoid words with a p-value of 0 and minimize lab-specific effects, we then excluded words that had zero counts in all but one cluster. This resulted in a list of words sorted by p-value enrichment in each cluster. p-values were then corrected for multiple hypothesis testing using the Bonferroni correction. Word clouds were created from terms with p-values smaller than 10^{-20}, using the `wordcloud` package in R (Fellows, 2012). The final label assignments were confirmed by manual inspection. Section 6.1.2 provides more details. The list of terms with corrected p-values for each condition cluster and the results of the manual inspection can be found in Supplementary File SuppData2.zip.

In addition to condition-wise clustering, we also performed row (gene-wise) clustering. We first hierarchically clustered the 997 genes in the Gold Standard, and then generalized these clusters to the 5,716 genes present in the entire expression dataset. This resulted in five clusters, for which we performed gene ontology enrichment analysis, as described in Section 2.7. See Section 6.1.3 for more detail.

2.3 Curation of the Gold Standard of Regulatory Interactions

A key aspect of the work was the construction of a high-quality gold standard of regulatory interactions in yeast, which was used as prior interactions data for TRN training (GS-train in Figure 1), for fitting RNA half-lives (GS-fit in Figure 1), and for validating the predicted interactions (GS-fit or GS-validate in Figure 1). The gold standard was derived by combining binding and expression information from three major sources (Table S1). The core data was taken from the YEASTRACT repository (Teixeira *et al.*, 2006; Monteiro *et al.*, 2008; Abdulrehman *et al.*, 2011; Teixeira *et al.*, 2014), which is a curated repository of $> 200,000$ regulatory interactions in yeast, obtained from $> 1,300$ bibliographic references. The repository contains two types of evidence for each potential regulatory interaction: direct and indirect. *Direct* evidence denotes an interaction coming from an assay that directly established a physical binding event, such as ChIP-seq or one-hybrid assay. *Indirect* evidence comes from differential expression analysis after a TF knock-out or overexpression assay indirectly suggests a regulatory relationship. We first filtered these data to obtain a conservative list of 2,577 regulatory interactions that have one source of direct evidence and two sources of indirect evidence. At this stage, these interactions were unsigned, i.e. they did not include information about whether the regulatory interaction is positive or negative.

As TFA estimation performs best when all prior known interactions are signed (see Section 6.1.6.1), we processed the list further to maximize the number of signed interactions. YEASTRACT provides information on the signs for some interactions, e.g. those derived from expression analysis of knock-out mutants. To add signs from the YEASTRACT database, we used the following rule: a regulatory interaction was deemed "positive" if the target

gene was down-regulated upon TF knock-out, and "negative" if the opposite were the case. As some interactions were detected in multiple experiments with opposite sign annotations, we only considered the signs that were measured in assays conducted under normal conditions, labeled as "YPD medium; mid-log phase" in the YEASTRACT database. In case there was still conflict, we employed the majority rule, and in case of a tie, the interaction was discarded (set to 0). This procedure resulted in 1,155 signed interactions in total.

To expand this dataset, we obtained additional regulatory interactions from the Saccharomyces Genome Database (SGD), curated by biocurators (Cherry *et al.*, 2012; Costanzo *et al.*, 2014) and from the Kemmeren *et al.* (2014) dataset of 1,484 knock-out experiments (Kemmeren *et al.*, 2014). These interactions were only used to assign signs to interactions that were still unsigned in the list of 2,577 interactions that had one direct and two indirect evidence types in YEASTRACT. These additions expanded our list of signed interactions to 1,403.

These 1,403 interactions constitute the set of signed prior known interactions used throughout this paper, which we denote as the *Gold Standard* (GS). Section 6.1.4, Figures S2 and S3, and Supplementary File SuppDoc1.pdf describe more details about the creation of GS and test it against other collections of interactions.

2.4 Inferelator Model

We used and modified code for the Inferelator version 2015.03.03 (Bonneau *et al.*, 2006; Greenfield *et al.*, 2013; Arrieta-Ortiz *et al.*, 2015). We describe the Inferelator core model in this section, and more details can be found in Section 6.1.6. The Inferelator algorithm calculates the optimal model of regulation for each target gene independently of other genes. The model for each gene i is based on the assumption that the dynamics of transcription regulation are governed by the following relation:

$$\frac{dX_i}{dt} = -\alpha_i X_i + \sum_{j \in P_i} \tilde{\beta}_{i,j} A_j, \tag{1}$$

where X_i is the RNA expression level of gene i, P_i is the set of potential regulators of gene i, A_j is the activity of TF j, $\tilde{\beta}_{i,j}$ is the coefficient of regulatory interaction between TF j and gene i, and α_i is the RNA degradation rate of gene i.

To estimate the parameters $\tilde{\beta}_{i,j}$, we can approximate Equation 1 using finite differences, and divide both sides by α_i:

$$\tau_i \frac{X_i(t_{k+1}) - X_i(t_k)}{t_{k+1} - t_k} + X_i = \sum_{j \in P_i} \beta_{i,j} A_j(t_k), \tag{2}$$

where the time axis t has been broken up into discrete time points at which the data was collected, indexed by k. The left hand side of Equation 2 is the *response variable*, whereas the the right hand side is the *design variable*. Note that $\tau_i = \alpha_i^{-1}$, and is related to the RNA transcript half-life HL_i via $\text{HL}_i = \tau_i \log(2)$, and $\beta_{i,j} = \tau_i \tilde{\beta}_{i,j}$. Also note that throughout our analysis, no corrections for cell division times were made, as it was impossible to determine them for each of the 2,577 experiments coming from 119 labs. Given that median half-lives are much shorter than doubling times, the effect is tolerable.

The response variable is first used together with prior known interactions to calculate TF Activities (TFA) for every TF (Section 6.1.6.1). TFA is derived from expression changes of the prior known targets of a TF, and has been shown to improve TRN inference dramatically in prokaryotes (Arrieta-Ortiz *et al.*, 2015). The same prior known interactions are then used in a constrained regression step that selects the most likely model of regulation for every gene using a data-driven approach called *Bayesian Best Subset Regression* (BBSR). For calculating TFA and BBSR, we used our Gold Standard or a subset thereof as prior known interactions. Figure 1 and Section 2.5 outline the workflows employed in this paper, specifying how the Gold Standard was used in each of them. The final output of the Inferelator is a list of confidence scores for all possible regulatory interactions, determined using a computational knock-out assay. Each Inferelator run was performed on 50 bootstraps of the RNA expression data, and the final confidence scores for all interactions were computed by rank-combining the confidence scores across bootstraps. For more detail, see Section 6.1.6.2.

2.5 Subsampling the Gold Standard for RNA-half Life Fitting and Error Estimation

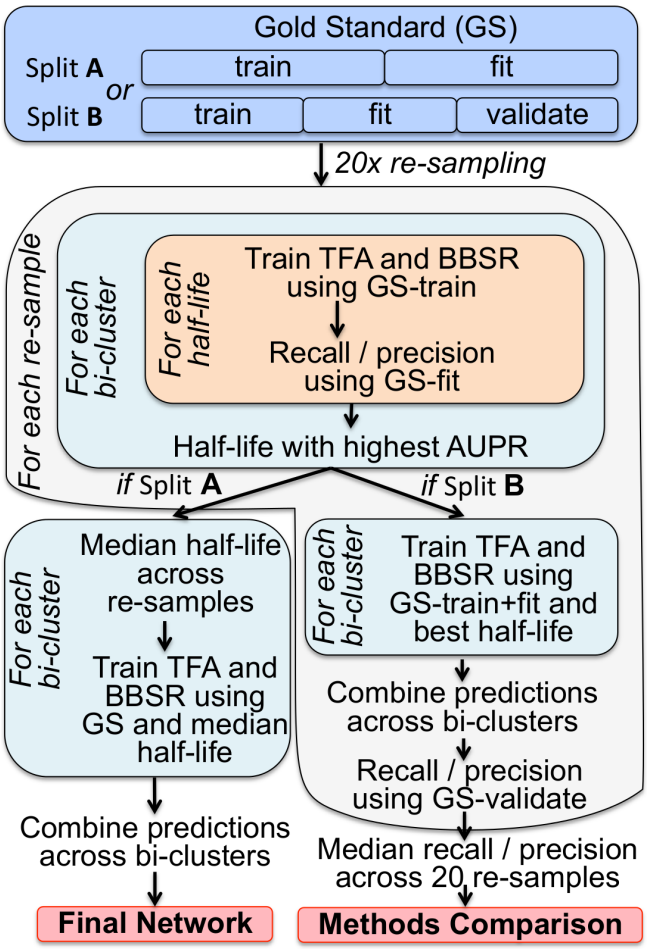

Figure 1: Outline of the workflows employed in this paper. We use two strategies for splitting the Gold Standard (GS) of interactions. Split **A** approach is used to make optimal predictions of condition- and gene- specific RNA half-lives. These half-lives are furthermore used to create the final predicted network. Split **B** approach is used to evaluate the improvement in network inference accuracy conferred by bi-clustering and estimating RNA half-lives. In both cases, GS was randomly re-sampled into two or three equal sets of interactions, with GS-train used to estimate Transcription Factor Activity (TFA) and run Bayesian Best Subset Regrssion (BBSR), and GS-fit to find the optimal condition- and gene- specific prediction of RNA half-lives, based on maximum area under precision-recall curve (AUPR).

To use our Gold Standard for both parameter fitting and method evaluation without overfitting, we designed two strategies for re-sampling the Gold Standard (Figure 1). For assessing the dependency of inference accuracy on RNA half-life (Figures 4 and 5A) and obtaining optimal gene- and condition-specific RNA half-lives (Figures 5B, 6, S4 and S5), we used Split **A**. This method involved randomly selecting a pre-specified fraction of Gold Standard interactions to be in the training set (GS-train), with the rest of the interactions to be used for fitting half-lives (GS-fit). We set the fraction of data used in the 'training' set to 0.5, although our results also hold for other values (Figure S9). This procedure was repeated 20 times, and for each re-sample, RNA half-lives were fit as described in Section 2.6.

To assess whether fitting condition- and gene-specific RNA half-lives in this manner improves performance (Figures 7A and 7C-D, Table 1), we used Split **B**, where a third set of GS interactions (GS-valid) was held out and used only for estimating accuracy of our algorithm's predictions. We created GS-valid to avoid over-fitting, and it was exclusively used to estimate the accuracy of the network computed using prior known interactions in GS-train and half-lives obtained using GS-fit. Each interaction was assigned one of the three categories randomly (GS-train, GS-fit, or GS-validate), with probabilities 0.34, 0.33, and 0.33, respectively. This way of splitting the Gold Standard was also applied to the other methods (Genie3 and iRafNet) for benchmarking purposes, keeping the random assignments of interactions into GS-train, GS-fit, and GS-valid identical across the methods for each GS re-sample.

We used two measures of network prediction accuracy to assess the quality of our predictions: Area Under Precision-Recall curve (AUPR) and Area Under the Receiver Operating Characteristic curve (AUROC). The two measures were calculated in the standard way, as described in Section 6.1.6.3. We focus here on AUPR, as it is more sensitive for high-scoring interactions compared to AUROC, which distributes the weights more equally across the entire list of predictions. Therefore, a model with maximal AUPR is desirable for small-scale, targeted validation experiments. AUROC is also inferior to AUPR in the class-imbalanced (skewed) regime, in which the sizes of true positives and false positives differ substantially (Davis & Goadrich, 2006), which is the case for our data.

2.6 RNA Half-Life Estimation

The primary advance described here is the explicit modeling and incorporation of RNA degradation rates into the large-scale inference of TRN. To do so, we first developed a procedure for comparing network prediction between models that assume different RNA half-lives (RNA half-lives are inversely proportional to RNA degradation rates, see Section 2.4).

As shown in Figure 1, Split **A** involves assembling two sets of interactions from the GS: one for training TFA and BBSR (GS-train), and one for calculating AUPR (GS-fit). We pre-specified values of τ that we used is $\tau = 0, 5, 10,$

20, 30, 40, 50, 60, 70, 80, 90, 100, 120, 140, 160, 200, and 250 minutes, designed to span the range of known RNA half-lives (Neymotin *et al.*, 2014; Munchel *et al.*, 2011; Sun *et al.*, 2013; Schwalb *et al.*, 2012; Miller *et al.*, 2011).

The Inferelator was run while setting τ to a given value on the list for every gene and every condition either in the given bi-cluster or in the entire data set, using GS-train as prior known interactions. Precision and recall curves were computed for each Inferelator run corresponding to a re-sample and a value of τ, treating GS-fit as the set of true interactions. Comparisons of precision-recall curves across RNA half-lives were made by taking the element-wise median of the precision and recall vectors across GS re-samples for a given value of RNA half-life (Figure 4A). Comparisons between AUPRs measured for different τ's were made while keeping GS-train and GS-fit constant for each re-sample, as represented by isochromatic curves in Figures 4B, 5A, and Figure S4, and an optimal τ was chosen by maximizing AUPR. We also compared performances between models with different RNA half-lives using AUROC instead of AUPR, yielding similar optimal half-lives (Figures S6, S7, S8).

Finally, our RNA half-life predictions for each condition and gene bi-cluster were made by considering the distribution of optimal half-lives across the 20 re-samples for a given gene and condition bi-cluster, using the Split **A** procedure in Figure 1. These distributions are shown in Figure 6 for various gene and condition clusters, and their medians are shown for every bi-cluster in Figure 5B. These median values of AUPR constitute our RNA half-life predictions for each gene and condition bi-cluster. To predict RNA half-lives of translation genes (Figure 6C and S5), we separately applied the same AUPR maximization procedure to each condition cluster, using only cytoplasmic translation genes and their known regulators for AUPR calculations, because the entire gene cluster that was enriched in translation genes had too many genes that were not related to translation. This was not an issue for nucleotide metabolism genes, so the entire gene cluster was used to predict their optimal half-lives. Supplementary File SuppData3.zip contains the final RNA half-life predictions for each gene and condition cluster. For more detail, see Section 6.1.5.

To estimate the improvement in TRN inference accuracy from fitting and incorporating bi-cluster-specific RNA degradation rates, we first split the Gold Standard according to Split **B**, into GS-fit, GS-train and GS-validate. For a given re-sample, the predicted RNA half-lives were determined by maximizing AUPR on each bi-cluster. Using those values of RNA half-lives and the same re-sample of the Gold Standard, the Inferelator was trained again, but now using a union of GS-train and GS-fit (GS-train+fit) for TFA and BBSR computation. The final precision-recall curve as reported in Figure 7A was calculated by adding confidence scores across condition clusters for each re-sample, calculating precision and recall on the GS-valid set corresponding to that re-sample, and then taking the element-wise median of the precision and the recall vectors across the 20 re-samples. Figures 7C and D were calculated the same way, but with true positives and false positives instead of precision and recall in the last (validation) step. Note that the magnitude of the increase in inference accuracy due to half-life fitting that we calculate using the Split **B** approach (Table 1) is an underestimate of the actual increase in accuracy of our final predicted network, which is produced using the Split **A** approach (see Section 6.1.7 for more detail).

2.7 Function Enrichment Analysis

For determining Gene Ontology (GO) enrichments, we employed the hypergeometric test using the `hyperGTest` function from the R Bioconductor package `'GOstats'`, and the R Bioconductor package `'org.Sc.sgd.db'` (Carlson *et al.*, 2014). We used the 997 genes that have at least one interaction in the GS as the background for the hypergeometric test. The p-values obtained from the hypergeometric test were corrected for multiple hypothesis testing using the Bonferroni correction. For each gene cluster, we reported the top GO enrichment terms, all of which were statistically significant ($p < 0.05$). For the full list of gene ontology terms enriched in each cluster, see Supplementary File SuppData1.zip.

2.8 Combining Condition-Specific Networks to Create the Final Yeast Network

The final predicted network of regulatory interactions was created in several steps (Figure 1). First, bi-cluster specific RNA half-lives were determined by re-sampling the Gold standard into GS-train and GS-fit 20 times (i.e. employing the Split **A** approach). The predicted half-life for each bi-cluster was determined by taking the median across the half-lives that optimized AUPR in each of the 20 GS re-samples. Then the Inferelator was trained on each bi-cluster using the full Gold Standard and the predicted value of the corresponding RNA half-life. The final combined confidence scores were obtained by adding the confidence scores across the condition clusters for every gene. To convert the final (combined) ranked list of interactions into a set of predicted interactions, we set a cutoff at precision=0.5, as calculated on the full GS. This resulted in 1,462 new (i.e. not present in the GS) interactions, which can be found in Supplementary File SuppNetwork1.tsv. Supplementary File SuppNetwork2.tsv contains all predictions and their precision values.

3 Results

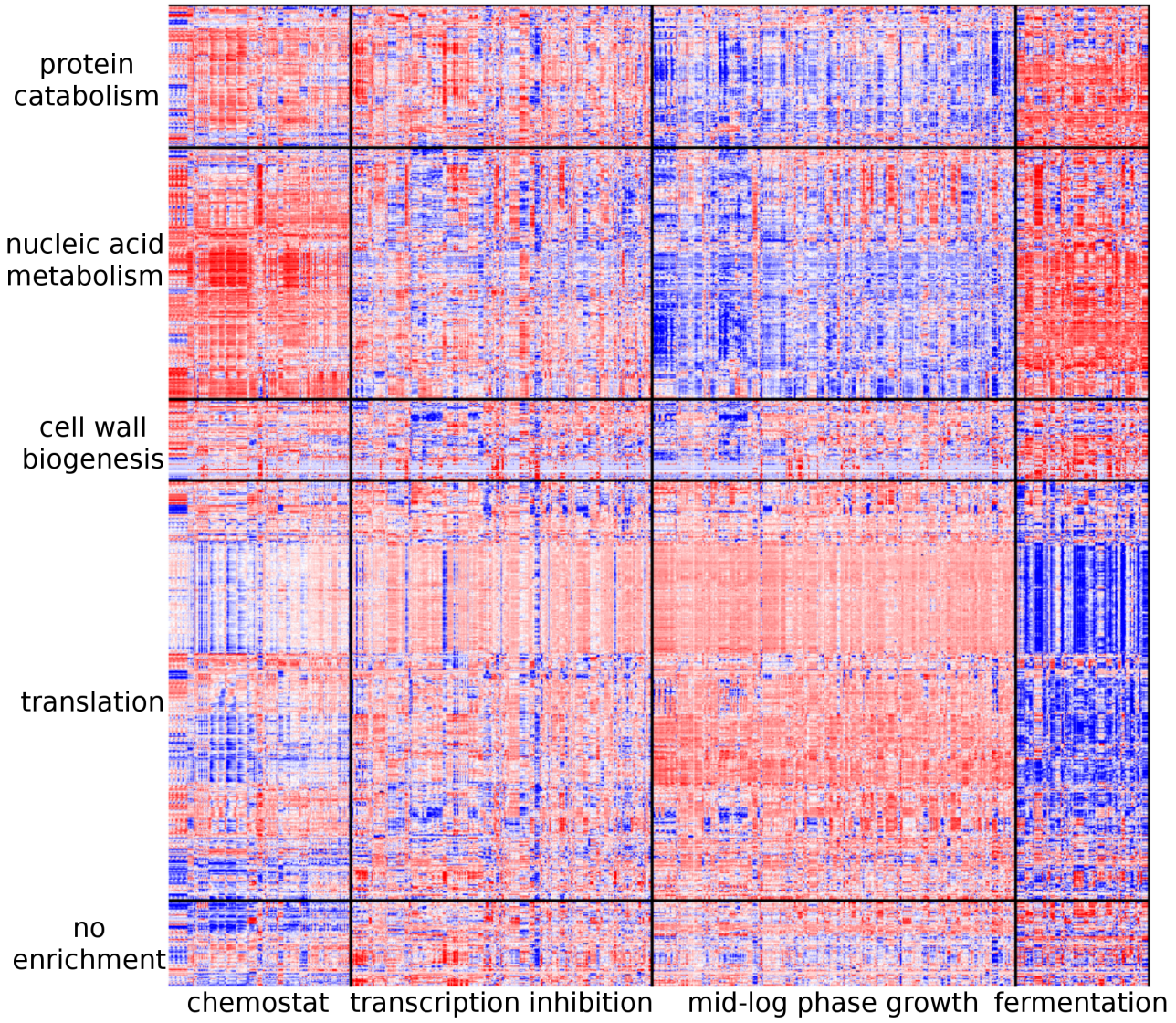

Figure 2: A heatmap of RNA expression data used for in this study. Genes are positioned on the vertical axis, whereas conditions are positioned on the horizontal axis. Bright red color denotes higher expression levels, whereas dark blue denotes lower expression. The four conditions clusters are shown at the bottom, whereas the five gene clusters are labeled on the left. The names of gene clusters correspond to a GO category that was most highly enriched in each cluster. Only the 997 genes that are also present in the GS are shown, although our final network was derived from expression data of 5,716 genes.

3.1 Assembly of Comprehensive Data Sets for High-Quality Inference

One of the main advantages of using baker's yeast to study eukaryotic TRNs is the rich availability of existing literature that characterizes yeast regulatory phenomena in a broad sample of experimental conditions. To produce a comprehensive and accurate yeast regulatory network, we assembled all components, such as a list of 563 transcription factors (TF), the Gold Standard of interactions, and the RNA expression data set. The expression data set originated from 119 labs and spanned a wide range of experimental conditions, but used the same transcriptomics platform throughout. Comprised of 5,716 genes and 2,577 samples, as shown in Figure 2, it is one of the largest expression data sets used in network inference in yeast.

The Gold Standard (GS) of regulatory interactions combines multiple types of regulatory evidence from multiple databases (Table S1) and includes 1,403 signed interactions (i.e. distinguishing between activation and repression).

This number represents only a small fraction of all potential regulatory interactions in yeast, but each interaction is confirmed by at least three sources of evidence: one of which is direct (e.g. ChIP-chip) and two are indirect (e.g. based on TF knock-out expression changes). The GS is therefore highly enriched in true positives. Figure S2 shows that this strategy – maximizing confidence rather than number of interactions in the GS – results in more consistent regulatory network models than other approaches.

Another development addresses the notion that several discrete modes of gene expression regulation are employed in responding to different categories of stress (Lehtinen *et al.*, 2013; Hart *et al.*, 2015; Yang & Leskovec, 2014). Indeed, when clustering the entire expression data set, we identified four clusters with distinct expression characteristics (Figure 2). Figure 3 shows the characteristics of each condition cluster, as extracted from the associated meta-data. Condition cluster 1 is related to chemostat growth environment under which nutrients are continuously supplied. Experiments include nutrient limitation and long time-series. Condition cluster 2 is related to transcriptional inhibition and oxidative stress experiments. Many of the transcriptional inhibition studies were early attempts to measure RNA degradation rates by inhibiting Pol II (Shalem *et al.*, 2008, 2011). Cluster 3 is enriched in mutant-wildtype comparisons for cells grown in a rich medium. Finally, condition cluster 4 includes alcoholic fermentation experiments involving industrial strains and processes of producing wine and beer. The time series experiments in this condition cluster often last several days.

Finally, we also clustered the data with respect to similarities between gene expression patterns, and analyzed the clusters for function enrichments. The most significant enrichments included cytoplasmic translation, cell wall biogenesis, nucleic acid metabolism, and protein catabolism across four of the five clusters, respectively (p-value < 0.05, Figure 2). Combining condition and gene clustering resulted in 20 bi-clusters of the RNA expression data matrix.

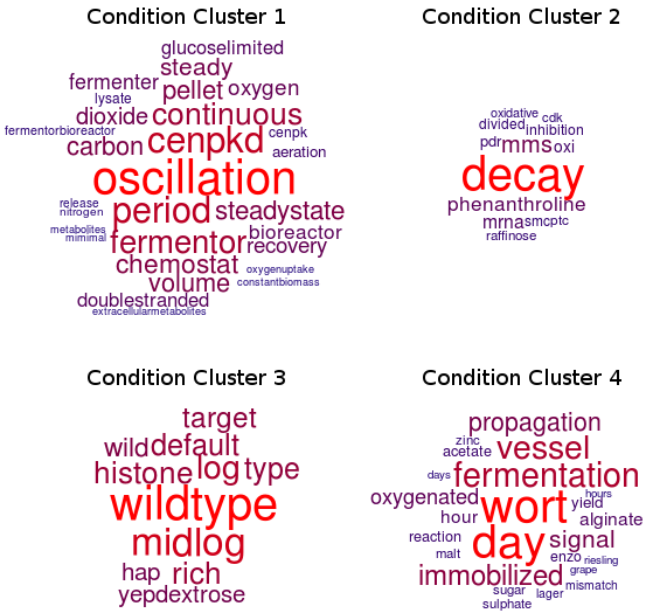

Figure 3: Terms enriched in each condition cluster. Words that are most highly enriched, as determined by the procedure described in the main text and in Methods, are displayed for each cluster. The higher the enrichment, the larger the font. The same cutoff of $p < 10^{-20}$ (after the Bonferroni correction) was used for all four clusters.

3.2 Global and Local Network Prediction is Sensitive to RNA Stability and Recapitulates Known Trends

To assess whether the accuracy of the predicted network is sensitive to RNA degradation rates, we first ran the Inferelator on the entire unclustered expression data set for a range of preset RNA half-lives and plotted the different precision-recall curves as shown in Figure 4. We found that an RNA half-life between 20 and 50 minutes maximizes the area under the precision-recall curve (AUPR). Intriguingly, the optimal half-life range is consistent with experimental measurements, which estimate median RNA half-lives in unperturbed conditions between 10 and 15 minutes (Miller *et al.*, 2011; Neymotin *et al.*, 2014; Munchel *et al.*, 2011).

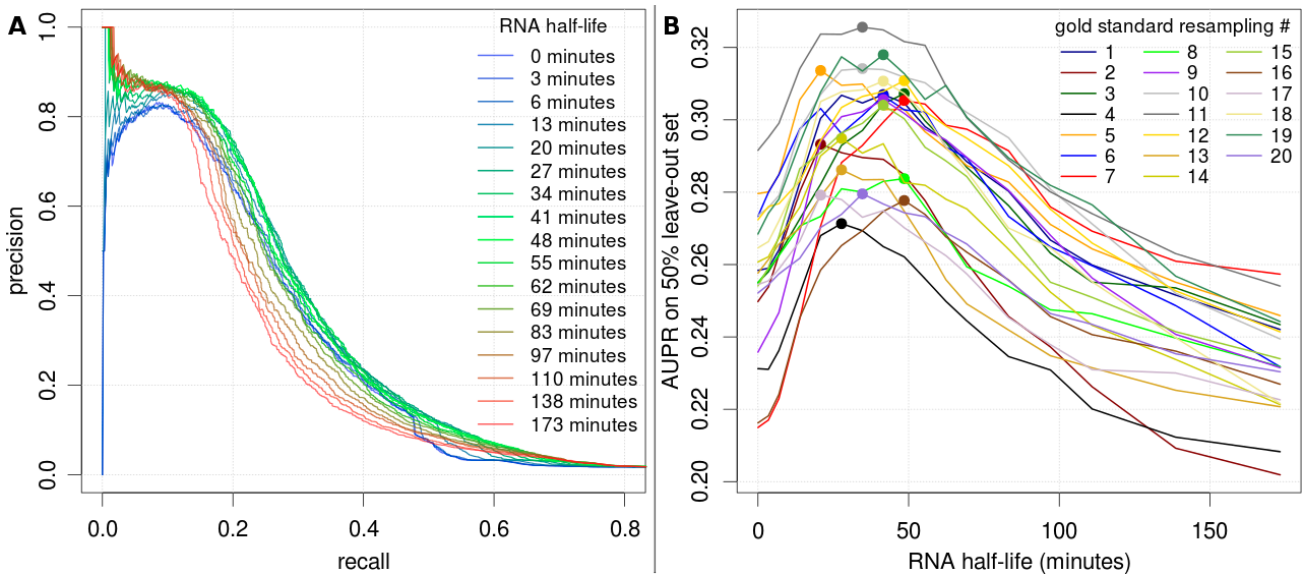

Figure 4: Network inference is sensitive to RNA half-lives. A) Precision-recall curves on the Inferelator output, with each line corresponding to a different pre-set value of RNA half-life. Each line displays the median precision and recall across 20 Gold Standard (GS) re-samples. B) Area under the precision-recall curve (AUPR) as a function of pre-set RNA half-life. Different lines denote 20 independent GS re-samples, and colored dots represent the maximum AUPR for a given GS re-sample.

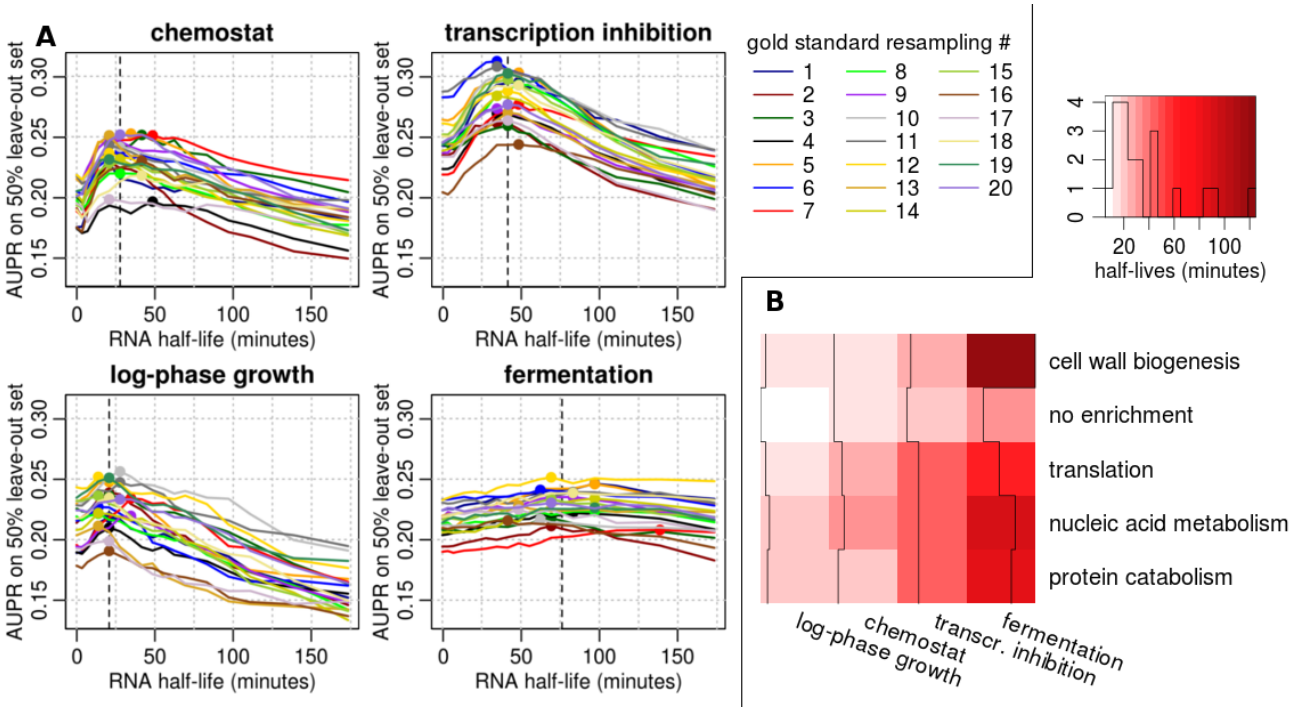

Figure 5: Network inference accuracy is sensitive to RNA half-lives in a condition- and gene-specific manner. A) AUPR as a function of pre-set gene-independent RNA half-life, where different colored lines correspond to 20 independent GS re-samples. The four panels correspond to condition clusters. Colored dots denote optimal half-lives, and horizontal dotted lines correspond to the median optimal half-life across re-samples. B) Optimal half-life for each condition and gene bi-cluster. Vertical black lines trace the magnitude of half-lives across gene clusters for every condition cluster. For the full plot of AUPR trajectories for every bi-cluster, see Figure S4.

As the data used in our work spans a variety of conditions very different from unperturbed cells growing in

rich medium, we proceeded to optimize RNA half-lives for the condition clusters separately. Figure 5A shows that the optimal RNA half-life differs between condition clusters. The distribution of optimal half-lives varies in width depending on the condition cluster, suggesting a variation in half-life specificity in different conditions. The lower optimal RNA half-lives for the "chemostat" and "log-phase growth" condition clusters - which represent conditions that are perturbed less severely than the "transcription inhibition" cluster, and represent typical laboratory strains (unlike the "fermentation" cluster) - are consistent with the observed median RNA half-lives of 10 to 15 minutes in unperturbed laboratory strains.

To assess the dependence of inference accuracy on half-life for each gene cluster, we extended this analysis to the 20 bi-clusters. Figure 5B summarizes the bi-clustering results by showing a heatmap of median values of half-life that optimized the AUPR on that bi-cluster. Figure S4 shows that AUPR trajectories for each bi-cluster peak at distinct values of RNA half-life.

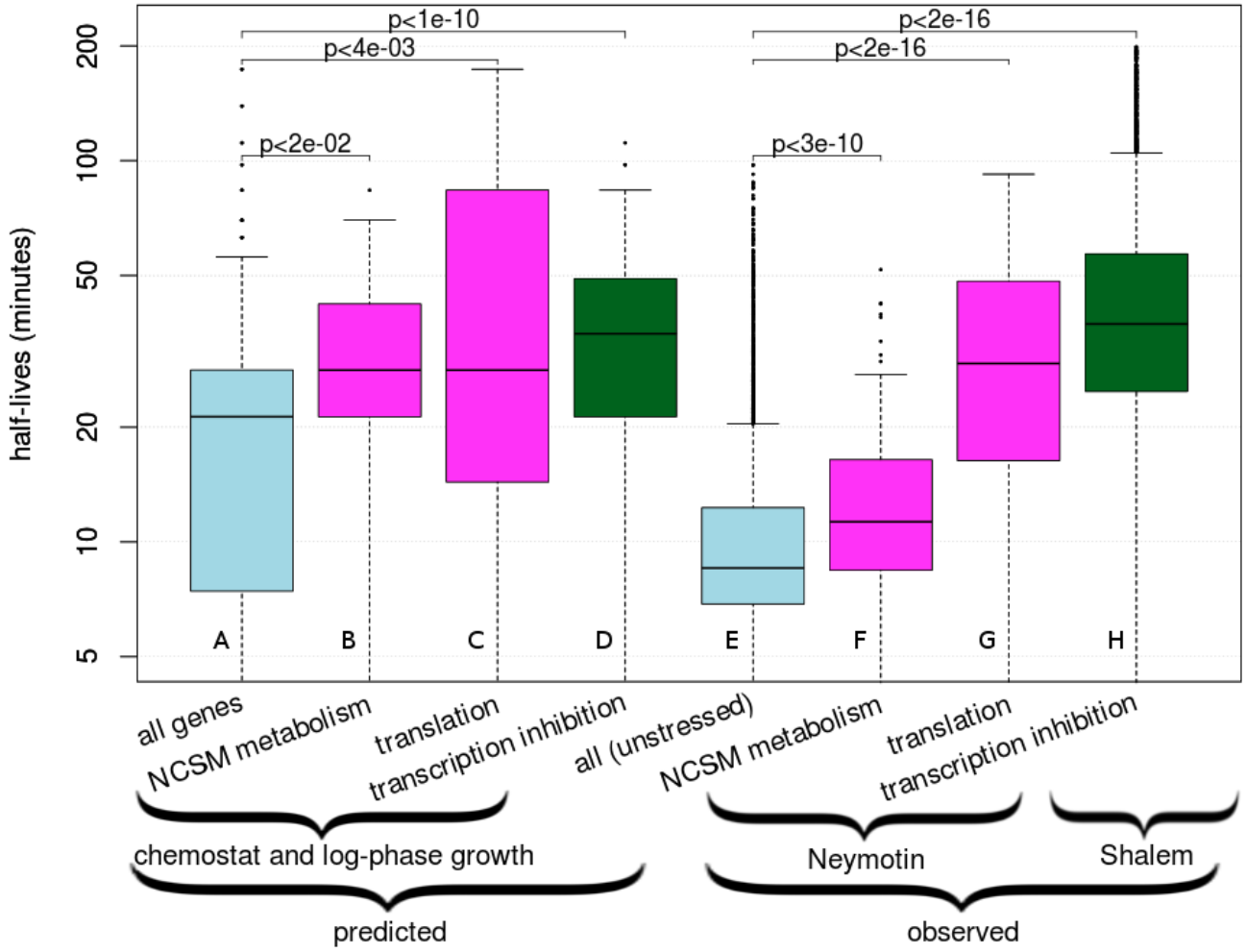

Figure 6: Predicted and observed differences in distributions of RNA half-lives between different groups of genes and conditions. A-D show predicted values, and E-H show experimentally measured values. Predicted values reflect distributions of RNA half-lives across the 20 GS re-samples. Magenta color denotes subsets of genes whose half-lives were predicted in non-extreme condition clusters ("chemostat" and "log-phase growth") or measured in normal external conditions (Neymotin *et al.*, 2014). These gene clusters, nucleotide metabolism (NCSM) and translation, are defined in the main text. Light blue denotes all genes predicted or measured in non-extreme or normal conditions, respectively. Green denotes half-lives of the entire transcriptome, predicted for the "transcription inhibition" condition cluster or measured under conditions that inhibited transcription (Shalem *et al.*, 2008).

The bi-cluster specific, optimal RNA half-lives represent gene- and condition-specific RNA half-life predictions and recapitulate known biology (Figure 6). First, the range of predicted half-lives under normal conditions is similar to the range of experimentally measured half-lives reported in recent experimental measurements under normal conditions (Figures 6A and E). The most prominent pattern in Figure 5B is the increased half-lives in the "fermentation"

condition cluster compared to the other condition clusters. We hypothesize that longevity (or slow growth rates) correlates with long RNA half-lives.

The next most prominent pattern in Figure 5B are the long half-lives in the "transcription inhibition" cluster compared to the "log-phase" and "chemostat" condition clusters (Figure 6D vs. Figure 6A, Wilcoxon $p < 10^{-10}$). This condition cluster is enriched in studies that measured RNA decay rates by inhibiting transcription. Indeed, methods that use transcription inhibition, e.g. Shalem *et al.* (2008), identify significantly longer RNA half-lives than those using less invasive metabolic labeling methods, e.g. Neymotin *et al.* (2014), as is shown in a comparison between Figure 6H and Figure 6E (Wilcoxon $p < 2 \times 10^{-16}$), (Neymotin *et al.*, 2014; Pelechano & Pérez-Ortín, 2008). This phenomenon is also closely related to buffering, wherein increased transcription rates in mutant strains correspond to decreased transcript stability, and vice versa (Sun *et al.*, 2012).

Another example illustrating successful prediction of RNA half-lives is shown for ribosomal genes, which are known to be more stable than other genes under normal conditions (Neymotin *et al.*, 2014; Munchel *et al.*, 2011). Again, the prediction of RNA half-lives in our framework – which is completely independent of experimental half-life measurements and is exclusively based on expression data and network priors – confirms this bias. The predicted half-life for the 115 ribosomal genes in the "translation" gene cluster is significantly higher than that of other genes (Figure 6, Wilcoxon $p < 4 \times 10^{-3}$, see Section 2.6).

Finally, the most prominent pattern in the "chemostat" and "log-phase" condition clusters is the "nucleic acid metabolism" gene cluster with very high half-lives (Figure 5, 6, Wilcoxon $p < 0.02$). Genes in this cluster are enriched in nucleobase-containing small molecule (NCSM) metabolism. Experimentally measured half-lives of these 207 genes confirmed this trend: under normal conditions, the genes exhibited higher RNA half-lives compared to all genes (Wilcoxon $p \leq 5 \times 10^{-10}$, Figure 6F). Therefore we demonstrated that optimizing TRN inference over the biophysical RNA half-life parameter accurately predicts both known and novel condition- and gene-specific trends, confirmed by direct experimental measurements.

3.3 Global Network Inference is Improved through the Use of Optimized RNA Stability

Having shown that the RNA half-lives optimized in our prediction are biologically relevant, we proceeded to use these half-lives to enhance prediction of regulatory interactions. To do so, we used separate, randomly chosen subsets of the Gold Standard to train the model, fit RNA half-lives for individual bi-clusters, and validate the predictions (Figure 1, Split **B**). A separate comparison was made for each of the 20 GS re-samples across each pair of competing methods (Figure 7, Table 1). These comparisons were made in terms of AUPR, as calculated on the respective GS-valid set of interactions, using the single value of half-life optimized on the respective GS-fit set of interactions. See Section 2.6 for more details on the methodology of methods comparisons.

Inference accuracy is improved drastically by incorporating bi-cluster-specific RNA half-life predictions, compared to inference without clustering of the expression data and use of optimized half-lives (Figure 7). A significant but smaller improvement is seen with each of the two modifications taken separately: either with bi-clustered expression data but without half-life optimization, or with half-life optimization but without expression clustering (Tables 1 and S2). A total of 113 of 120 pairwise comparisons provided larger AUPR for inference with these modifications (Table 1), and the median improvement of using both modifications is $\approx 14\%$. Using AUROC yielded a similar outcome (Table S2). The final AUPR value of 0.33 represents an almost eight-fold increase compared to Genie3 (Table 1), an inference method that performed best in a recent network inference competition (Marbach *et al.*, 2012). It represents a 10-fold increase compared to another method, iRafNet (Table 1) (Petralia *et al.*, 2015). See Sections 2.5, 2.6, and 6.1.7 for further details.

To maximize the size and accuracy of the final, integrated network, we repeated the whole procedure, but used the entire Gold Standard for training, setting the half-lives for each condition and gene cluster to the predictions made using the Split **A** approach from Figure 1. At 50% precision, this final transcription regulatory network contained 1,462 newly predicted interactions that were not present in the Gold Standard. Of these 1,462 interactions, 631 (43%) were validated by external data that was not included in this work, derived either from direct binding evidence (Yeastract-Direct), indirect evidence (Yeastract-Indirect, SGD, Kemmeren), or both (Figure 7B). This high fraction of independently confirmed interactions suggests that the remaining 831 new interactions are also strongly enriched in true positives.

Table 1: Both expression data bi-clustering and RNA half-life fitting independently improve Inferelator performance (second column). Furthermore, combining the two modifications improves performance as compared to using either one of them separately (third column). Columns 2 and 3 show the number of times the AUPR measured on GS-fit from the same re-sample was higher for one method than the other, given a pair of methods specified by the row and the column. Median AUPR for each approach is reported in the fourth column. See Sections 2.5, 2.6, and 6.1.7 for further details.

Method	Re-samples outperforming no clustering + fitting	Re-samples outperformed by clustering + fitting	Median AUPR
Clustering + Fitting	19/20	-	**0.328**
Clustering	20/20	17/20	0.319
Fitting	19/20	19/20	0.305
none	-	19/20	0.290
Genie3	0/20	20/20	0.042
iRafNet	0/20	20/20	0.031

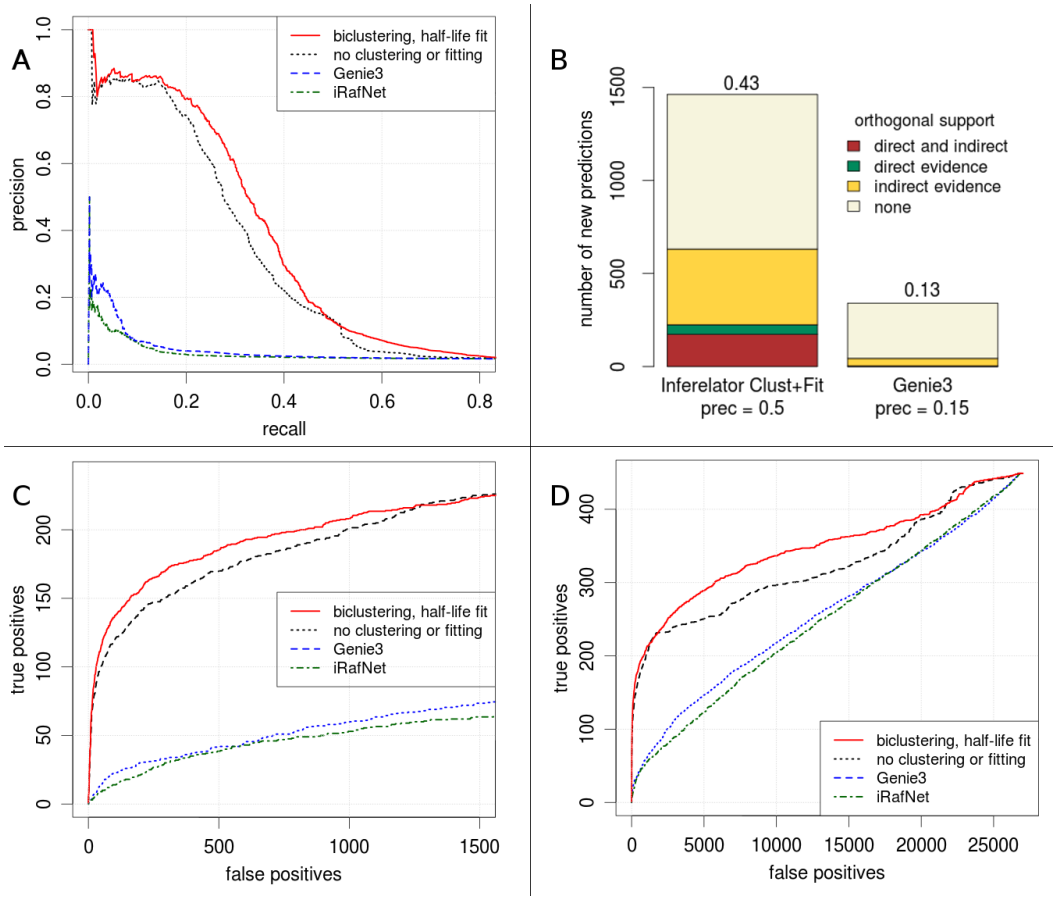

Figure 7: Inferring bi-cluster specific RNA half-lives improves network inference accuracy, resulting in better network predictions than other state-of-the-art methods. A) The estimated improvement in the precision-recall curve as a result of bi-clustering expression data and optimizing the half-life for each bi-cluster (red line), compared to using the Inferelator without bi-clustering or half-life optimization (black dotted line), or compared to Genie3 and iRafNet (blue line and green line, respectively). B) Number of new predicted interactions (i.e. interactions not in the Gold Standard), obtained using the optimized bi-cluster-specific half-lives and the full GS for training, compared to new predictions from Genie3. The vertical length of each color section within each bar corresponds to the number of new interactions that were confirmed by the corresponding type of evidence in an orthogonal data source. Direct evidence refers to physical binding, and indirect refers to knock-out and overexpression assays. The number above each bar denotes the fraction of new interactions supported by at least one orthogonal source. C and D show the results of the same comparison as in A by plotting false positive vs true positive rates instead of precision and recall. C) is a zoomed-in version of D.

3.4 The High-Quality Final Network Produces Functionally Relevant Predictions

To illustrate the value of newly predicted regulatory interactions, Table S3 lists the top ten most confidently predicted targets of the ten most connected and medium connected TFs, respectively. Many of these interactions are validated by literature. Table 2 and Table S3 display interactions that are not already in our Gold Standard, but many of these interactions are present in orthogonal collections of interactions (non-bolded targets in Table 2, superscripts in Table S3). For example, SFP1 is a well-known regulator of ribosome biogenesis (Marion *et al.*, 2004; Cipollina *et al.*, 2008; Reja *et al.*, 2015), and indeed, all of its top predicted targets are ribosomal subunits. Similarly, RPN4 regulates proteasome expression (Karpov *et al.*, 2008a,b), and proteasomal genes form the majority of its most confidently predicted targets.

Table 2: New final predictions and their precision values. A) New targets (i.e. interactions not in the Gold Standard) of the 10 TFs that have the highest number of known targets in the Gold Standard, and the precision values of these new interactions. TFs are listed in the top row, and their new targets with the corresponding precision values are listed below the TF. B) New targets of the median 10 TFs in terms of the number of known targets, out of the 97 TFs with at least one target in the GS. Precision values are calculated using the entire matrix of prediction confidence scores, containing 5,716 genes and 557 TFs. The list of true positives was defined as the entire Gold Standard. Shade of green denotes the preicision value. Bold targets correspond to interactions that were not found in any of the four major sources of interactions used in this study listed in Table S2. See Table S3 and Supplementary File SuppNetwork3.tsv for more details.

A

RAP1		GCN4		SFP1		MSN2		SOK2		YAP1		HSF1		RPN4		ABF1		TEC1	
SEC14	0.67	LYS1	0.85	RPS24A	0.98	YDR391C	0.71	YRO2	0.67	**GAC1**	0.48	OPI10	0.81	RPN12	0.93	**YER010C**	0.31	**TRX2**	0.26
YEL1	0.62	HOM2	0.84	RPL18B	0.92	CMK2	0.64	UIP4	0.67	TAH18	0.46	GGA1	0.66	RPT2	0.92	COA3	0.28	CRH1	0.25
PMT4	0.54	BAT1	0.82	RPS16A	0.92	YLR257W	0.43	YNL194C	0.61	ISF1	0.44	RTC3	0.64	PRE10	0.92	GET2	0.24	YOL019W	0.23
FEN1	0.52	HIS2	0.82	RPL8A	0.90	STF2	0.38	GAD1	0.60	YNR034W-A	0.25	YRO2	0.62	RPT5	0.92	BSD2	0.23	RAX2	0.20
ALG7	0.50	STR2	0.80	RPS18B	0.91	RCN2	0.32	JIP4	0.59	MBR1	0.22	APJ1	0.61	RPN7	0.88	**YAH1**	0.22	YOL014W	0.18
RBD2	0.44	SDT1	0.76	RPS18A	0.90	HAL5	0.25	MSC1	0.59	**MRK1**	0.20	**LSB1**	0.57	RPN1	0.85	**TDA5**	0.22	TGL2	0.16
RPL18B	0.43	RIB3	0.75	RPL40B	0.90	OXR1	0.24	TFS1	0.56	TCB2	0.18	**YGR127W**	0.53	RPN10	0.83	MIM1	0.22	YPS3	0.15
RPL16B	0.41	PSF2	0.75	RPL42B	0.88	PNC1	0.24	YNL195C	0.56	GAT2	0.18	YGR250C	0.51	PRE7	0.82	YDR541C	0.18	PHM8	0.14
YLR412C-A	0.40	POS5	0.74	RPS22A	0.88	DOA1	0.23	YJR096W	0.54	YLR460C	0.16	CUR1	0.51	YBR062C	0.81	**MGR2**	0.18	**YBL111C**	0.12
HXK2	0.40	SRY1	0.74	RPL9A	0.88	MRP8	0.22	OM45	0.49	ATR1	0.16	**SAF1**	0.45	PUP1	0.81	SLC1	0.17	**RIB4**	0.11

B

MGA2		MOT3		INO2		MSN4		FKH2		FLO8		RTG3		STP1		LEU3		NRG1	
PLB2	0.83	FIG2	0.49	HNM1	0.84	YHR022C	0.47	AIM20	0.83	FLO9	0.33	IDH2	0.71	VHR2	0.33	OAC1	0.74	**PDE2**	0.60
TDA4	0.71	PRM1	0.48	SAH1	0.81	CRS5	0.37	HOF1	0.84	**TIR4**	0.24	PDH1	0.30	**MET17**	0.33	BAT1	0.72	**PRM7**	0.40
YLR413W	0.66	SAG1	0.46	FAS1	0.81	**JLP1**	0.22	ALK1	0.82	**DED81**	0.17	AAT2	0.27	MET3	0.32	FRS2	0.66	**VTS1**	0.28
ALE1	0.65	TIR1	0.46	SAM2	0.78	SNO4	0.20	CLB1	0.81	**PAU7**	0.15	**CPR4**	0.22	**SUL1**	0.30	ILV5	0.52	YLR407W	0.27
FAS1	0.50	PAU24	0.39	CHO1	0.75	FRE7	0.16	**YMR030W-A**	0.79	**MSC7**	0.15	**ARP3**	0.14	MET10	0.28	LEU1	0.49	**GFD2**	0.26
HEM13	0.46	AAC3	0.39	EPT1	0.67	PUG1	0.13	BUD4	0.78	**PAU5**	0.14	WTM1	0.14	MET5	0.27	**SSB1**	0.43	**YGR079W**	0.24
SUR2	0.36	TIR4	0.35	ADO1	0.63	YEL073C	0.13	CDC5	0.77	**ERG24**	0.14	IDP1	0.13	MEP2	0.27	MAE1	0.33	**YNL095C**	0.21
NEM1	0.35	EUG1	0.31	OPI3	0.63	PDC6	0.12	**YMR001C-A**	0.76	**PAU24**	0.14	**URA8**	0.12	**PET8**	0.21	PAB1	0.26	**TRM5**	0.20
PEX31	0.35	TIR2	0.27	EHT1	0.59	GCY1	0.11	KIN3	0.75	**NCP1**	0.13	**ADE3**	0.10	GNP1	0.21	**YPL260W**	0.25	**PHO90**	0.19
FHN1	0.35	FIG1	0.26	YIP3	0.55	FMP48	0.11	HST3	0.67	**YGL108C**	0.13	**SLP1**	0.10	DAL80	0.18	**ILV1**	0.22	MCH5	0.16

Of the 100 new interactions predicted as the top ten most likely targets of the top ten most highly-connected genes, only 13 (13%) have not been validated by one of the four regulatory interactions data bases that we employed. However, existing literature suggests that many of these completely new interactions are likely to be correct. For example, HSF1 is a key regulator of diverse stresses and monitors translation status through interaction with the Ribosome Quality Control complex (RQC) (Brandman *et al.*, 2012). We make a completely new prediction that HSF1 regulates LSB1 and SAF1. SAF1 has four other transcription regulators (BUR6, MED6, SPT10, SUA7) all detected under various stresses, especially heat shock (Mendiratta *et al.*, 2006; Venters *et al.*, 2011) – suggesting that HSF1 may be a true regulator. LSB1 regulates actin assembly and prion modulation in yeast (Ali *et al.*, 2014), but has not been previously linked to HSF1. However, several recent studies have linked HSF1 to actin assembly: one study identified altered actin cytoskeletal structures in yeast deficient in the RQC-Hsf1 regulatory system (Yang *et al.*, 2016); another study showed that overexpressing HSF1 in worms increases actin cytoskeleton integrity and lifespan (Baird *et al.*, 2014); a third study in mammalian cells confirmed that active HSF1 affects the actin cytoskeleton (Toma-Jonik *et al.*, 2015). Therefore, it is tempting to hypothesize that actin assembly regulator Lsb1 is the missing link by which HSF1 affects the actin skeleton.

Another validation for newly predicted regulatory interactions arises from gene knockout phenotypes. For example, the TEC1 transcription factor regulates filamentation genes, but also positively affects lifespan (Mösch & Fink, 1997; Garay *et al.*, 2014). Its newly predicted target TRX2 is a thioredoxin isoenzyme involved in the oxidative stress response (Garrido & Grant, 2002; Greetham *et al.*, 2010). A screen by Postma *et al.* (2009) demonstrated increase lifespan in a TRX2 null mutant strain Postma *et al.* (2009), again providing supporting evidence for the regulatory interaction between TEC1 and TRX2 that was newly predicted in our work.

Finally, similar reasoning supports another example: MRK1 is a newly predicted target for the YAP1 transcription factor, which is a basic leucine zipper required for tolerance to oxidative stress and cadmium exposure (Kuge & Jones, 1994; Wemmie *et al.*, 1994; Lee *et al.*, 1999). MRK1 is a Glycogen synthase kinase 3 (GSK-3) homolog, which activates Msn2p dependent transcription of stress response genes (Hardy *et al.*, 1995; Hirata *et al.*, 2003). Interestingly, a genome-wide screen showed increased cadmium levels for the MRK1 knockout, indicating that the gene is indeed linked to a YAP1 function (Yu *et al.*, 2012). Cadmium exposure is an example of a rare environmental condition for which binding or perturbation experiments would likely be missing, highlighting the advantage of network inference from large-scale expression data of diverse experimental design.

4 Discussion

Genome-wide inference of transcription regulatory networks in eukaryotes is challenging. Here, we present conceptual advances over existing work (Greenfield *et al.*, 2013; Arrieta-Ortiz *et al.*, 2015), demonstrating, for the first time, that using biophysically relevant models that incorporate, for example, RNA degradation, improves automatic large-scale network prediction. In addition, our approach includes other substantial improvements, such as the use of a high-quality Gold Standard of regulatory interactions, and the construction of different network models across subsets of genes and conditions. Using these advances, we present a genome-wide regulatory network, which at 50% precision predicts $> 1,400$ new interactions, 43% of which are validated by independent data sets, and 57% are entirely new (Figure 7B).

The high-quality Gold Standard dataset of regulatory interactions is built on several benchmark datasets of regulator–target interactions (Teixeira *et al.*, 2006; Monteiro *et al.*, 2008; Abdulrehman *et al.*, 2011; Teixeira *et al.*, 2014; Cherry *et al.*, 2012; Costanzo *et al.*, 2014; Kemmeren *et al.*, 2014), but improves on these sets by adding signs and accounting for confidence measures. In constructing this Gold Standard, we found that the quality, but not necessarily the size of a benchmark set of interactions improves predictions (Figure S2).

With this signed Gold Standard, we combined network predictions from several disjoint expression sub-spaces. Combining networks that were modeled assuming distinct regulatory regimes resulted in higher recovery of known interactions than when assuming the same regulatory regime across all conditions (Figure 7, Table 1). This result is consistent with the findings that both RNA degradation rates and regulatory networks experience fine-tuned and global changes in response to changing environmental conditions (Munchel *et al.*, 2011; Miller *et al.*, 2011; Lehtinen *et al.*, 2013; Hart *et al.*, 2015; Yang & Leskovec, 2014).

Most importantly, we showed that including RNA degradation in the mathematical modeling of RNA expression changes substantially boosted inference of the transcriptional regulatory network. To the best of our knowledge, the inference framework that we used, the Inferelator, is the only approach capable of doing so on a genome-wide scale. Notably, we learned RNA degradation rates directly from time-series and steady-state expression data, without the use of values known *a priori*. The resulting optimal rates are surprisingly similar to experimentally measured rates and accurately reflect known trends. For example, ribosomal RNAs are more stable than other gene transcripts under normal conditions, and transcription inhibition appears to correlate with degradation globally, as has been shown experimentally (Neymotin *et al.*, 2014; Sun *et al.*, 2012; Pelechano & Pérez-Ortín, 2008) (Figure 6).

Given that it is still highly challenging to measure RNA degradation in living cells and relevant conditions, only a few such datasets exist (Miller *et al.*, 2011; Schwalb *et al.*, 2012; Sun *et al.*, 2012; Neymotin *et al.*, 2014; Munchel *et al.*, 2011), and our framework can be used to predict missing rates. It can be used to reveal trends that have so far gone unnoticed, such as the long half-lives of nucleic acid metabolism genes (Figure 6). While the majority of predicted interactions were validated by external large-scale regulatory interaction data sets, we also illustrated that even the interactions not seen in any previous literature are biologically meaningful and supported by orthogonal evidence – e.g. YAP1 regulating MRK1, TEC1 regulating TRX2, and HSF1 regulating LRB1.

In a broader context, this work illustrates the promising prospect of more biophysically motivated modeling approaches to network inference. Other popular methods currently use techniques such as Random Forest (Huynh-Thu *et al.*, 2010; Petralia *et al.*, 2015), Mutual Information and Related Transfer Entropy (Margolin *et al.*, 2006a,b), correlation (Butte & Kohane, 2000), and others. These methods do not provide a clear way to incorporate or recapitulate measurements of biophysical dynamics parameters that govern transcription regulation. The *Inferelator's* explicit modeling of the RNA synthesis and degradation processes via a differential equation highlights the advantage of using this framework for TRN inference.

The results of this study encourage further developments of the *Inferelator* algorithm that would allow for an efficient incorporation and recovery of biophysical parameters, such as RNA decay rates and interaction terms between co-regulating TFs, a more careful separation of the transcription term into transcriptional activation and repression, which has only been done on the small scale (Noman & Iba, 2005; Liu & Wang, 2008; Bonneau & Aijo, 2016;

Intosalmi *et al.*, 2016), and modeling functional modifications of TFs that can affect their transcriptional activity. Given the growing body of literature on RNA-binding proteins (RBPs) (Hogan *et al.*, 2008; Mittal *et al.*, 2009; Janga & Mittal, 2011), our results also inspire potential approaches to model the RNA decay term explicitly as a sum of contributions from RNA degradation factors. Therefore, it is time to move inference of transcription regulatory networks to biophysically relevant models, and the work presented here provides an important step towards this goal.

5 References

References

Abdulrehman D, Monteiro PT, Teixeira MC, Mira NP, Lourenço AB, dos Santos SC, Cabrito TR, Francisco AP, Madeira SC, Aires RS, Oliveira AL, Sá-Correia I, Freitas AT (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic acids research* **39**: D136–40

Äijö T, Granberg K, Lähdesmäki H (2013) Sorad: a systems biology approach to predict and modulate dynamic signaling pathway response from phosphoproteome time-course measurements. *Bioinformatics (Oxford, England)* **29**: 1283–91

Ali M, Chernova TA, Newnam GP, Yin L, Shanks J, Karpova TS, Lee A, Laur O, Subramanian S, Kim D, McNally JG, Seyfried NT, Chernoff YO, Wilkinson KD (2014) Stress-dependent proteolytic processing of the actin assembly protein Lsb1 modulates a yeast prion. *The Journal of biological chemistry* **289**: 27625–39

Arrieta-Ortiz ML, Hafemeister C, Bate AR, Chu T, Greenfield A, Shuster B, Barry SN, Gallitto M, Liu B, Kacmarczyk T, Santoriello F, Chen J, Rodrigues CDA, Sato T, Rudner DZ, Driks A, Bonneau R, Eichenberger P (2015) An experimentally supported model of the Bacillus subtilis global transcriptional regulatory network. *Molecular systems biology* **11**: 839

Baird NA, Douglas PM, Simic MS, Grant AR, Moresco JJ, Wolff SC, Yates JR, Manning G, Dillin A, Dillin A (2014) HSF-1-mediated cytoskeletal integrity determines thermotolerance and life span. *Science (New York, NY)* **346**: 360–3

Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, Sullivan J, Micklem G, Cherry JM (2012) YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. *Database* **2012**: bar062

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research* **41**: D991–5

Bonneau R, Aijo T (2016) Biophysically motivated regulatory network inference: progress and prospects. *bioRxiv*

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

Bouchet-Valat M (2014) *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library.* R package version 0.5.1

Brandman O, Stewart-Ornstein J, Wong D, Larson A, Williams CC, Li GW, Zhou S, King D, Shen PS, Weibezahn J, Dunn JG, Rouskin S, Inada T, Frost A, Weissman JS (2012) A ribosome-bound quality control complex triggers degradation of nascent peptides and signals translation stress. *Cell* **151**: 1042–54

Butte AJ, Kohane IS (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing* : 418–29

Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, AmiGO Hub tA, Web Presence Working Group tWPW (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics (Oxford, England)* **25**: 288–9

Carlson M, Falcon S, Pages H, Li N (2014) Org. sc. sgd. db: Genome wide annotation for yeast. *R package version* **2**

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, *et al.* (2012) Saccharomyces Genome Database: The genomics resource of budding yeast. *Nucleic Acids Research* **40**

Cipollina C, van den Brink J, Daran-Lapujade P, Pronk JT, Porro D, de Winde JH (2008) Saccharomyces cerevisiae SFP1: at the crossroads of central metabolism and ribosome biogenesis. *Microbiology (Reading, England)* **154**: 1686–99

Costanzo MC, Engel SR, Wong ED, Lloyd P, Karra K, Chan ET, Weng S, Paskov KM, Roe GR, Binkley G, Hitz BC, Cherry JM (2014) Saccharomyces genome database provides new regulation data. *Nucleic acids research* **42**: D717–25

Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, Aerts S, Thurman R, Rynes E, Humbert R, Vierstra J, Maurano M, Dunham I, Kundaje A, Aldred S, Collins P, Davis C, Jones P, Andersson R, Gebhard C, *et al.* (2015) Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling. *PLOS Genetics* **11**: e1004994

Davis J, Goadrich M (2006) The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, pp. 233–240

Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)* **23**: 1846–7

Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30**: 207–210

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. *PLoS Biology* **5**: e8

Feinerer I, Hornik K (2015) *tm: Text Mining Package*. R package version 0.6-2

Feinerer I, Hornik K, Meyer D (2008) Text Mining Infrastructure in R. *Journal of Statistical Software* **25**: 1–54

Fellows I (2012) wordcloud: Word clouds. *R package version* **2**: 109

Friedman N, Linial M, Nachman I, Pe'er D (2000) Using Bayesian Networks to Analyze Expression Data. *Journal of Computational Biology* **7**: 601–620

Garay E, Campos SE, González de la Cruz J, Gaspar AP, Jinich A, Deluna A (2014) High-resolution profiling of stationary-phase survival reveals yeast longevity factors and their genetic interactions. *PLoS genetics* **10**: e1004168

Garrido EO, Grant CM (2002) Role of thioredoxins in the response of Saccharomyces cerevisiae to oxidative stress induced by hydroperoxides. *Molecular Microbiology* **43**: 993–1003

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics (Oxford, England)* **20**: 307–15

Greenfield A, Hafemeister C, Bonneau R (2013) Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. *Bioinformatics (Oxford, England)* **29**: 1060–7

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

Greetham D, Vickerstaff J, Shenton D, Perrone GG, Dawes IW, Grant CM, Dalle-Donne I, Rossi R, Giustarini D, Colombo R, Milzani A, Klatt P, Lamas S, Shenton D, Grant C, Rouhier N, Lemaire S, Jacquot J, Davis D, Newcomb F, *et al.* (2010) Thioredoxins function as deglutathionylase enzymes in the yeast Saccharomyces cerevisiae. *BMC Biochemistry* **11**: 3

Grigull J, Mnaimneh S, Pootoolal J, Robinson MD, Hughes TR (2004) Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molecular and cellular biology* **24**: 5534–47

Guo Y, Mahony S, Gifford DK (2012) High Resolution Genome Wide Binding Event Finding and Motif Discovery Reveals Transcription Factor Spatial Binding Constraints. *PLoS Computational Biology* **8**: e1002638

Hardy T, Wu D, Roach P (1995) Novel Saccharomyces cerevisiae Gene, MRK1, Encoding a Putative Protein Kinase with Similarity to Mammalian Glycogen Synthase Kinase-3 and Drosophila Zeste-White3/Shaggy. *Biochemical and Biophysical Research Communications* **208**: 728–734

Hart Y, Sheftel H, Hausser J, Szekely P, Ben-Moshe NB, Korem Y, Tendler A, Mayo AE, Alon U (2015) Inferring biological tasks using Pareto analysis of high-dimensional data. *Nature Methods* **12**: 233–235

Hirata Y, Andoh T, Asahara T, Kikuchi A (2003) Yeast glycogen synthase kinase-3 activates Msn2p-dependent transcription of stress responsive genes. *Molecular biology of the cell* **14**: 302–12

Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS biology* **6**: e255

Honkela A, Girardot C, Gustafson EH, Liu YH, Furlong EEM, Lawrence ND, Rattray M (2010) Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 7793–8

Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś AK, Pagès H, *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature Methods* **12**: 115–121

Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. *PLoS ONE* **5**: e12776

Intosalmi J, Nousiainen K, Ahlfors H, Lähdesmäki H (2016) Data-driven mechanistic analysis method to reveal dynamically evolving regulatory networks. *Bioinformatics* **32**: i288–i296

Janga SC, Mittal N (2011) Construction, structure and dynamics of post-transcriptional regulatory network directed by RNA-binding proteins. *Advances in experimental medicine and biology* **722**: 103–17

Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**: 118–127

Karlebach G, Shamir R (2012) Constructing Logical Models of Gene Regulatory Networks by Integrating Transcription Factor–DNA Interactions with Expression Data: An Entropy-Based Approach. *Journal of Computational Biology* **19**: 30–41

Karpov DS, Osipov SA, Preobrazhenskaia OV, Karpov VL (2008a) [Rpn4p is a positive and negative transcriptional regulator of the ubiquitin-proteasome system]. *Molekuliarnaia biologiia* **42**: 518–25

Karpov DS, Tiutiaeva VV, Beresten' SF, Karpov VL (2008b) [Mapping of Rpn4p regions responsible for transcriptional activation of proteasome genes]. *Molekuliarnaia biologiia* **42**: 526–32

Kemmeren P, Sameith K, van de Pasch LAL, Benschop JJ, Lenstra TL, Margaritis T, O'Duibhir E, Apweiler E, van Wageningen S, Ko CW, van Heesch S, Kashani MM, Ampatziadis-Michailidis G, Brok MO, Brabers NACH, Miles AJ, Bouwmeester D, van Hooff SR, van Bakel H, Sluiters E, *et al.* (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell* **157**: 740–52

Küffner R, Petri T, Tavakkolkhah P, Windhager L, Zimmer R (2012) Inferring gene regulatory networks by ANOVA. *Bioinformatics (Oxford, England)* **28**: 1376–82

Kuge S, Jones N (1994) YAP1 dependent activation of TRX2 is essential for the response of Saccharomyces cerevisiae to oxidative stress by hydroperoxides. *The EMBO journal* **13**: 655–64

Lähdesmäki H, Shmulevich I, Yli-Harja O (2003) On Learning Gene Regulatory Networks Under the Boolean Network Model. *Machine Learning* **52**: 147–167

Lee J, Godon C, Lagniel G, Spector D, Garin J, Labarre J, Toledano MB (1999) Yap1 and Skn7 control two specialized oxidative stress response regulons in yeast. *The Journal of biological chemistry* **274**: 16040–6

Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**: 882–883

Lehtinen S, Marsellach FX, Codlin S, Schmidt A, Clément-Ziza M, Beyer A, Bähler J, Orengo C, Pancaldi V (2013) Stress induces remodelling of yeast interaction and co-expression networks. *Molecular bioSystems* **9**: 1697–707

Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, Roychowdhury VP (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 15522–7

Liu PK, Wang FS (2008) Inference of biochemical network models in S-system using multiobjective optimization approach. *Bioinformatics (Oxford, England)* **24**: 1085–92

MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC bioinformatics* **7**: 113

Madar A, Greenfield A, Vanden-Eijnden E, Bonneau R (2010) DREAM3: Network inference using dynamic context likelihood of relatedness and the inferelator. *PLoS ONE* **5**

Mani S, Cooper GF (2004) Causal discovery using a Bayesian local causal discovery algorithm. *Studies in health technology and informatics* **107**: 731–5

Mani S, Spirtes PL, Cooper GF (2012) A theoretical study of Y structures for causal discovery. *arXiv preprint arXiv:12066853*

Marbach D, Costello JC, Küffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. *Nature methods* **9**: 796–804

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A (2006a) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics* **7**: S7

Margolin AA, Wang K, Lim WK, Kustagi M, Nemenman I, Califano A (2006b) Reverse engineering cellular networks. *Nature Protocols* **1**: 662–671

Marion RM, Regev A, Segal E, Barash Y, Koller D, Friedman N, O'Shea EK (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 14315–22

Mendiratta G, Eriksson PR, Shen CH, Clark DJ (2006) The DNA-binding domain of the yeast Spt10p activator includes a zinc finger that is homologous to foamy virus integrase. *The Journal of biological chemistry* **281**: 7040–8

Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dölken L, Martin DE, Tresch A, Cramer P (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular systems biology* **7**: 458

Mittal N, Roy N, Babu MM, Janga SC (2009) Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 20300–5

Monteiro PT, Mendes ND, Teixeira MC, D'Orey S, Tenreiro S, Mira NP, Pais H, Francisco AP, Carvalho AM, Lourenço AB, Sá-Correia I, Oliveira AL, Freitas AT (2008) YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. *Nucleic acids research* **36**: D132–6

Mösch HU, Fink GR (1997) Dissection of filamentous growth by transposon mutagenesis in Saccharomyces cerevisiae. *Genetics* **145**: 671–84

Munchel SE, Shultzaberger RK, Takizawa N, Weis K (2011) Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Molecular biology of the cell* **22**: 2787–95

Mundade R, Ozer HG, Wei H, Prabhu L, Lu T (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond

Neymotin B, Athanasiadou R, Gresham D (2014) Determination of in vivo RNA kinetics using RATE-seq. *RNA (New York, NY)* **20**: 1645–52

Noman N, Iba H (2005) Inference of gene regulatory networks using S-system and differential evolution

Pe'er D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics (Oxford, England)* **17 Suppl 1**: S215–24

Pelechano V, Pérez-Ortín JE (2008) The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast. *Yeast* **25**: 85–92

Peshkin L, Wühr M, Pearl E, Haas W, Freeman RM, Gerhart JC, Klein AM, Horb M, Gygi SP, Kirschner MW (2015) On the relationship of protein and mRNA dynamics in vertebrate embryonic development. *Developmental cell* **35**: 383–394

Petralia F, Wang P, Yang J, Tu Z (2015) Integrative random forest for gene regulatory network inference. *Bioinformatics (Oxford, England)* **31**: i197–205

Postma L, Lehrach H, Ralser M (2009) Surviving in the cold: yeast mutants with extended hibernating lifespan are oxidant sensitive. *Aging* **1**: 957–960

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria

Reiss DJ, Baliga NS, Bonneau R (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC bioinformatics* **7**: 280

Reiss DJ, Plaisier CL, Wu WJ, Baliga NS, DM P, E W, LH H, A T, J I, T M, E S, R D, AN B, E S, DJ R, K L, C H, Y H, K M, SA A, *et al.* (2015) cMonkey ₂ : Automated, systematic, integrated detection of co-regulated gene modules for any organism. *Nucleic Acids Research* **43**: e87–e87

Reja R, Vinayachandran V, Ghosh S, Pugh BF (2015) Molecular mechanisms of ribosomal protein gene coregulation. *Genes & development* **29**: 1942–54

Schwalb B, Schulz D, Sun M, Zacher B, Dümcke S, Martin DE, Cramer P, Tresch A (2012) Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics (Oxford, England)* **28**: 884–5

Schwanhäusser B, Wolf J, Selbach M, Busse D (2013) Synthesis and degradation jointly determine the responsiveness of the cellular proteome. *BioEssays : news and reviews in molecular, cellular and developmental biology* **35**: 597–601

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**: 166–176

Setty M, Leslie CS, Mathelier A, Newburger D, Bulyk M, Wingender E, Bailey T, Machanick P, Bailey T, Liu X, Brutlag D, Liu J, Wang J, Georgiev S, Pique-Regi R, Neph S, Arvey A, Agius P, Leslie C, Kuang R, *et al.* (2015) SeqGL Identifies Context-Dependent Binding Signals in Genome-Wide Regulatory Element Maps. *PLOS Computational Biology* **11**: e1004271

Shalem O, Dahan O, Levo M, Martinez MR, Furman I, Segal E, Pilpel Y (2008) Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular systems biology* **4**: 223

Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y (2011) Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. *PLoS genetics* **7**: e1002273

Shmulevich I, Dougherty ER, Kim S, Zhang W (2002) Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics (Oxford, England)* **18**: 261–74

Siahpirani AF, Roy S (2016) A prior-based integrative framework for functional transcriptional regulatory network inference. *Nucleic acids research* : gkw963

Studham ME, Tjärnberg A, Nordling TEM, Nelander S, Sonnhammer ELL (2014) Functional association networks as priors for gene regulatory network inference. *Bioinformatics (Oxford, England)* **30**: i130–8

Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, Tresch A, Cramer P (2013) Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Molecular cell* **52**: 52–62

Sun M, Schwalb B, Schulz D, Pirkl N, Etzold S, Larivière L, Maier KC, Seizl M, Tresch A, Cramer P (2012) Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome research* **22**: 1350–9

Tchourine K, Poultney CS, Wang L, Silva GM, Manohar S, Mueller CL, Bonneau R, Vogel C (2014) One third of dynamic protein expression profiles can be predicted by a simple rate equation. *Molecular BioSystems* **10**: 2850–2862

Teixeira MC, Monteiro P, Jain P, Tenreiro S, Fernandes AR, Mira NP, Alenquer M, Freitas AT, Oliveira AL, Sá-Correia I (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic acids research* **34**: D446–D451

Teixeira MC, Monteiro PT, Guerreiro JF, Gon??alves JP, Mira NP, Dos Santos SC, Cabrito TR, Palma M, Costa C, Francisco AP, Madeira SC, Oliveira AL, Freitas AT, S??-Correia I (2014) The YEASTRACT database: An upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. *Nucleic Acids Research* **42**

Toma-Jonik A, Widlak W, Korfanty J, Cichon T, Smolarczyk R, Gogler-Piglowska A, Widlak P, Vydra N (2015) Active heat shock transcription factor 1 supports migration of the melanoma cells via vinculin down-regulation. *Cellular Signalling* **27**: 394–401

Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, Batzoglou S, Myers RM, Sidow A (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods* **5**: 829–34

Venters BJ, Wachi S, Mavrich TN, Andersen BE, Jena P, Sinnamon AJ, Jain P, Rolleri NS, Jiang C, Hemeryck-Walsh C, Pugh BF (2011) A comprehensive genomic binding map of gene and chromatin regulatory proteins in Saccharomyces. *Molecular cell* **41**: 480–92

Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO (2002) Precision and functional specificity in mRNA decay. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 5860–5

Wemmie JA, Wu AL, Harshman KD, Parker CS, Moye-Rowley WS (1994) Transcriptional activation mediated by the yeast AP-1 protein is required for normal cadmium tolerance. *The Journal of biological chemistry* **269**: 14690–7

Wilkins O, Hafemeister C, Plessis A, Holloway-Phillips MM, Pham GM, Nicotra AB, Gregorio GB, Jagadish K, Septiningsih EM, Bonneau R, Purugganan MD (2016) EGRINs (Environmental Gene Regulatory Influence Networks) in Rice That Function in the Response to Water Deficit, High Temperature, and Agricultural Environments. *The Plant cell* : tpc.00158.2016

Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays

Yang J, Hao X, Cao X, Liu B, Nyström T (2016) Spatial sequestration and detoxification of Huntingtin by the ribosome quality control complex. *eLife* **5**

Yang J, Leskovec J (2014) Overlapping Communities Explain Core-Periphery Organization of Networks. *Proceedings of the IEEE* **102**: 1892–1902

Yu D, Danku JMC, Baxter I, Kim S, Vatamaniuk OK, Vitek O, Ouzzani M, Salt DE (2012) High-resolution genome-wide scan of genes, gene-networks and cellular systems impacting the yeast ionome. *BMC genomics* **13**: 623

6 Supplement

6.1 Supplementary Methods

6.1.1 Primary Data Processing

The meta data for every whole-genome expression sample (GSM) was downloaded using the R function `getGEO` from the `GEOquery` package (Edgar *et al.*, 2002; Barrett *et al.*, 2013). These GSMs were filtered such that the final list only contained samples measured in *Saccharomyces cerevisiae* using the Yeast Affymetrix 2.0 platform (GPL2529). For each expression series (GSE) that contained at least one of those GSMs, a TAR file was downloaded from the GEO website on March 23, 2015, using the R function `download.file`. The raw CEL files for every sample in that GSE (filtering for *S. cerevisiae* and GPL2529) were furthermore extracted from the associated TAR file. For processing and normalizing the raw CEL files, we used the R packages `affy` (Gautier *et al.*, 2004) and `gcrma` (Wu *et al.*, 2004), using the functions `ReadAffy` and `gcrma`. All samples were processed simultaneously as one batch. For time series meta data, all time differences were converted into minutes. To test for batch effects, we used ComBat (Johnson *et al.*, 2007; Leek *et al.*, 2012), treating either laboratory of origin or experiment series as batches, but did not observe an improvement in network prediction (data not shown). All computations were made in R programming language, version 3.1.2 (R Core Team, 2014).

6.1.2 Condition Clustering

Principal Component Analysis (PCA) of the entire expression matrix was performed using the R command `prcomp`, treating each expression sample as a $5,716$-dimensional vector. K-means clustering of the dimensionally-reduced (post-PCA) expression samples was performed using the R function `kmeans`, using the default parameters with `nstart=25`, and `iter.max=1000`, corresponding to 25 initial random configurations (among which the best one is automatically picked by the built-in algorithm in R) and a maximum of 1000 iterations. PCA and k-means clustering was performed on scaled expression data (mean shifted to 0 and variance scaled to 1), and all subsequent analysis was done on the resulting clusters using the original (processed and normalized, but not scaled) expression data. We also performed some of the downstream analysis using clusters that were obtained the same way but without scaling, which did not lead to any significant differences in terms of inference performance (not shown).

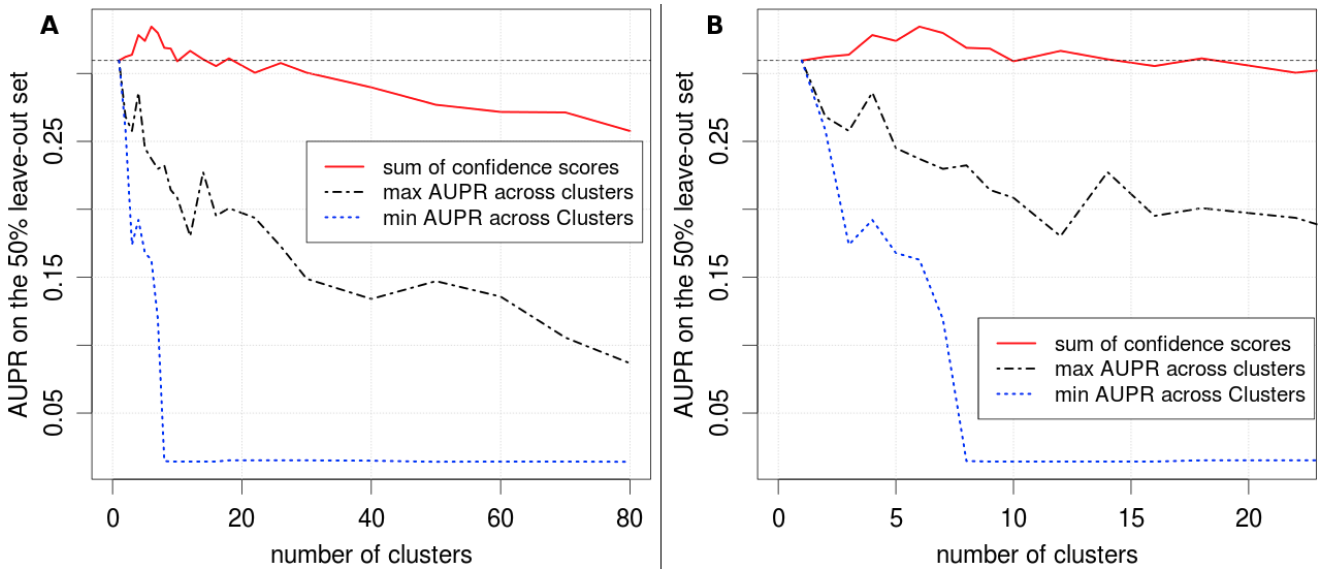

Figure S1: Network inference accuracy is improved when the expression data is clustered sample-wise and the predictions from the resulting clusters are combined. A) AUPR on the leave-out set is a function of the number of clusters, into which the expression data is split. B) same as A, but the x-axis is zoomed in around where the AUPR of the combined confidence scores is peaked. The red line shows the AUPR of the confidence scores combined across the clusters, the black dash-dotted line shows the highest AUPR across the clusters, and the dotted blue line shows the lowest AUPR across the clusters.

The number of condition clusters was determined by rank-combining the confidence score outputs of the Inferelator

across a pre-specified number of clusters, and maximizing the prediction accuracy of this procedure as a function of the number of condition clusters. Rank-combining across the predictions obtained from a set of disjoint expression clusters was done by taking the sum of the confidence score matrices (`combinedconf*.RData` files). We used the AUPR of the rank-combined prediction as a proxy for prediction accuracy. The pre-specified numbers of clusters that we tested were n = 2-10, 12, 14, 16, 18, 22, 26, 30, 40, 50, 60, 70, and 80. For a given n, the entire expression data set was first clustered into n clusters (as described in the paragraph above). For each n, we ran the Inferelator separately for each of n expression clusters, using 50% of the GS as the leave-in set for TFA and model selection steps, and the rest as the leave-out set for calculating AUPR. We did not resample the GS for this analysis, we used 10 response matrix bootstraps, and set $\tau = 60$ for this analysis. The same randomly sampled leave-in set was used across the Inferelator runs for all clusters.

Figure S1 shows AUPR as a function of the number of clusters n. Furthermore, it shows the AUPR of the worst-performing and the best-performing cluster among the n clusters. We notice that the AUPR is maximized around $n = 4$ and $n = 7$, where it is approximately equal to 0.35. We selected $n = 4$ as our optimal number of clusters because after $n = 4$, there is a sharp drop in the AUPR of the worst-performing cluster, rendering it uninformative. All downstream analysis was performed on these four clusters.

To annotate the condition clusters, all text that was provided in every available sample of the meta data table was sorted into four bins according to the sample's assigned cluster. The following processing was done for the list of words in every cluster. First, common English words, commas, and numbers were removed. Text was converted to lower case. These operations were done using the R package `tm` (Feinerer & Hornik, 2015; Feinerer *et al.*, 2008). Unnecessary whitespaces were removed, and term frequency statistics were calculated using the R package `SnowballC` (Bouchet-Valat, 2014). Multiple hypothesis testing for enriched terms in each cluster was performed using the Bonferroni correction. To foster visualization and interpretability, we removed spurious terms, which we did not consider biologically relevant, from these lists. The final word clouds were made using the R function `wordcloud` from the package `wordcloud` (Fellows, 2012).

In order to assign each cluster with a final label, we compared the occurrence of enriched terms in each cluster and assessed whether they come only from one lab or from multiple labs, in order to ensure that each topic is not enriched in its cluster as a result of lab bias in term usage (Supplementary File SuppData2.zip). Our condition cluster label assignments were also supported by GO enrichments in genes that had the highest relative variance compared to other genes in their cluster (not shown).

6.1.3 Gene Clustering

Gene clustering was performed as described in Section 2.2. The reason for scaling expression data before clustering is that this approach results in more evenly-sized gene clusters. We performed hierarchical clustering using the R function `hclust` with default arguments (e.g. euclidean distance metric). We determined the number of clusters by comparing several qualities of the resulting gene clusters as a function of the number of clusters. Given a pre-specified number of clusters k, we used the R function `cutree` to cut the similarity tree into k clusters. For each of the k resulting gene clusters, we performed Gene Ontology (GO) enrichment analysis (see Section 2.7). After performing this analysis for k ranging from 2 to 8, we determined that setting $k > 5$ does not result in any new clusters with meaningful new GO enrichments as compared to the enrichments obtained with $k = 5$ clusters. Rather, setting $k > 5$ creates new clusters with no GO enrichments.

Since our method for determining gene- and condition-cluster specific RNA half-lives relies on the availability of prior known interactions with the target genes, we first performed gene clustering for the 997 genes that were present in our gold standard of interactions. For this reason, setting the number of clusters higher than $k > 5$ resulted in clusters with too few genes to result in significant GO enrichments. For example, with $k = 5$, the smallest cluster consists of 79 genes and the largest consists of 409 genes. With $k = 6$, the former smallest cluster was further broken down into two clusters with 44 and 35 genes in each. Small gene clusters were also unfavorable because our RNA half-life inference for a given gene cluster relies on AUPR maximization calculated on a leave-out list of prior known interactions involving the genes in that cluster. AUPR calculation gets increasingly noisy as the number of known true positives is reduced. For $k = 5$, the smallest number of prior known interactions with the genes in one of the five clusters was 101, and the largest was 592. This range was sufficient for producing relatively smooth half-life vs. AUPR curves (Figure S4).

For the final yeast network, we included regulatory network predictions for all 5,716 genes in our expression data set. This required us to assign each gene with one of the clusters that was determined using the 997 genes that had prior known interactions. Denote the set of those 997 genes as gGS. For a given gene g, its cluster membership was assigned by finding the gene $g_{\text{prior}} \in gGS$ that minimizes the euclidean distance between g and g_{prior}, and assigning the cluster membership of gene g_{prior} to gene g. This approach was undertaken instead of clustering all 5,716 genes

Table S1: A survey of the databases of regulatory interactions used in this study. The first column denotes the name of the interactions database. The second column lists the number of interactions in each database. An interaction may be duplicated within a database if the database contains multiple entries for the same interaction (e.g. coming from different labs or different conditions), so in the parentheses we specify the number of unique interactions for each database. The third column lists the most common types of regulatory evidence in each database, including the number of interactions for each type of evidence. These top types of evidence account for over 98% of interactions in each database.

Database	Total number of interactions (unique)	Top types of evidence for interactions
Yeastract-Direct	55,653 (41,654)	ChIP (Chip: 47,535, Seq: 2,302, unspecified: 5382), EMSA: 267, DNA footprinting: 61
Yeastract-Indirect	194,083 (172,929)	Microarray expression level (WT vs. TF mutant: 167,288, WT vs. TF overexpression: 15,916, WT vs. TF loss of function mutant: 3,711, WT vs. TF chimera: 1,511), Northern Blot – WT vs. TF mutant: 1,406, Proteomics – WT vs. TF mutant: 993, lacZ – WT vs. TF mutant: 875
SGD	33,838 (32,338)	ChIP (Chip: 16,638, Seq: 154, unspecified: 11), Microarray expression level: 13,165, computational combinatorial: 3,782
Kemmeren	193,253 (193,253)	Microarray expression level – WT vs. TF mutant: 193,253

and then subsetting them to the 997 genes in *gGS* because the latter approach resulted in highly unevenly distributed cluster membership numbers among those 997 genes, which would result in noisy AUPR calculations on clustered data.

6.1.4 Gold Standard Curation

Gold Stanard was created as described in Section 2.3. Here we specify some further details. In the YEASTRACT dataset, in case the sign of an interaction was specified, it appeared in the "Association Type" column. When "Association Type" had a label "Negative", it denoted a decrease in abundance after a TF knock-out (and a positive sign according to our convention), and "Positive" denoted an increase (resulting in a negative sign according to our convention).

To obtain interactions from Saccharomyces Genome Database, we downloaded 32,338 unique regulatory interactions using Yeastmine (Balakrishnan *et al.*, 2012; Costanzo *et al.*, 2014). We refer to these 32,338 interactions as the *SGD* standard (such as in Table S3). The SGD website provided signs for 10,022 of those interactions. We observed 363 agreements and 12 disagreements in signs of the overlapping signed interactions between the list of 1,155 high-confidence signed interactions obtained from YEASTRACT (as described in Section 2.3) and the 10,022 from SGD. We used the signed portion of the SGD standard to expand our list of signed GS interactions in the following way. If the signed interaction from SGD was already in the list of 1,155 from YEASTRACT, we kept the sign that we obtained from YEASTRACT regardless of its sign in the SGD standard. If it was among the 2,577 high-confidence interactions obtained from YEASTRACT, but without a sign from YEASTRACT, we assigned it with the sign from SGD. This procedure provided an additional 117 signed interactions towards the final gold standard of interactions.

Furthermore, we used the data from Kemmeren *et al.* (2014), which contained the results of a study that measured genome-wide expression level changes for 1,484 knock-outs, including most TFs in yeast. First, we created a collection of interactions by considering all knock-out-induced expression changes that were reported in the original publication with a p-value of $p < 0.01$ after the Bonferroni correction, yielding $> 193,000$ interactions (which we furthermore denote as the *Kemmeren* standard). Of these, 131 were present but so far unsigned amongst the 2,577 high-confidence interactions described above, therefore expanding the set of high-confidence signed interactions to 1,403.

Figure S2 compares the performance of the Inferelator on the 50% leave-out set of seven different collections of interactions, when trained on the 50% leave-in set of the same seven collections (respectively). This provides a measure of consistency of a collection of interactions, estimating how effectively the Inferelator trained on one half of a collection of interactions can recover the other half of the same collection of interactions. According to both the precision-recall curve and the Receiver Operating Characteristic curve, the final Gold Standard of 1,403 signed interactions is by far the most consistent (has the highest area under the curve) as compared to any of the sources of interactions shown in Table S1. The biggest improvement in AUPR is achieved by restricting the collection of interactions from YEASTRACT to the 2,577 unsigned interactions that have 1 piece of evidence in Yeastract-Direct and 2 pieces of evidence in Yeastract-Indirect (blue line, AUPR=0.20, AUROC=0.70). Reducing those interactions to only those, for which we have evidence of the direction of the regulation in either of the three

sources of interactions (YEASTRACT, SGD, and Kemmeren), further improves performance (Gold Standard, red line, AUPR=0.30, AUROC=0.73). Our Gold Standard also outperformed the MacIsaac regulatory interactions database (MacIsaac *et al.*, 2006), which is commonly employed for evaluating network inference algorithms in yeast (Marbach *et al.*, 2012; Siahpirani & Roy, 2016). We extended this comparison to other possible combinations of these collections of interactions (Figure S3). Our results led us to conclude that the 1,403 interactions of our Gold Standard constitute the most consistent and reliable collection of interactions in yeast known to date.

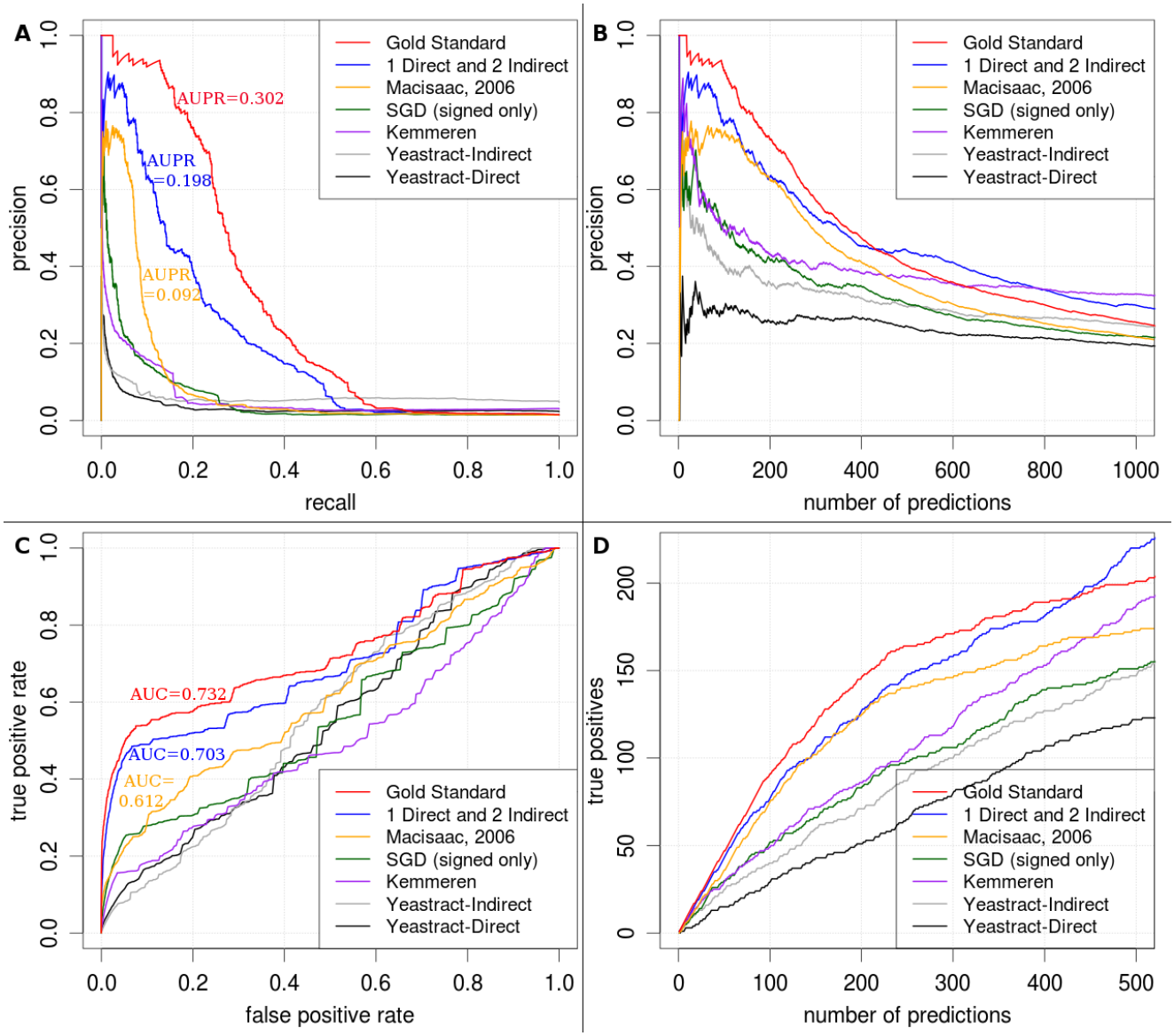

Figure S2: Combining multiple sources of evidence creates more consistent collections of regulatory interactions. Each of the collections of interactions listed in the legend was split into two equal parts, one for training the Inferelator and one for validation. Each line represents statistics calculated on the 50% leave-out set of the corresponding collection of interactions. Black, gray, purple, and green lines correspond to the collections of interactions described in Table S1. The green line only includes the signed portion of the SGD collection (where positive and negative regulation is pre-specified). Blue line corresponds to the interactions that appear at least once in the Yeastract-Direct collection and at least twice in the Yeastract-Indirect collection, without positive and negative regulation specified. The red line shows the final signed Gold Standard we use throughout this study. A) Precision vs. recall for the six collections of interactions. B) Precision as a function of number of predictions. C) Receiver Operating Characteristic curve for each collection of interactions. D) Number of true positives as a function of the number of predictions. Note that in B and D, performance on the Gold Standard decreases more quickly after several hundred interactions because the positives exhaust all true interactions more quickly due to the smaller size of the Gold Standard as compared to other collections shown.

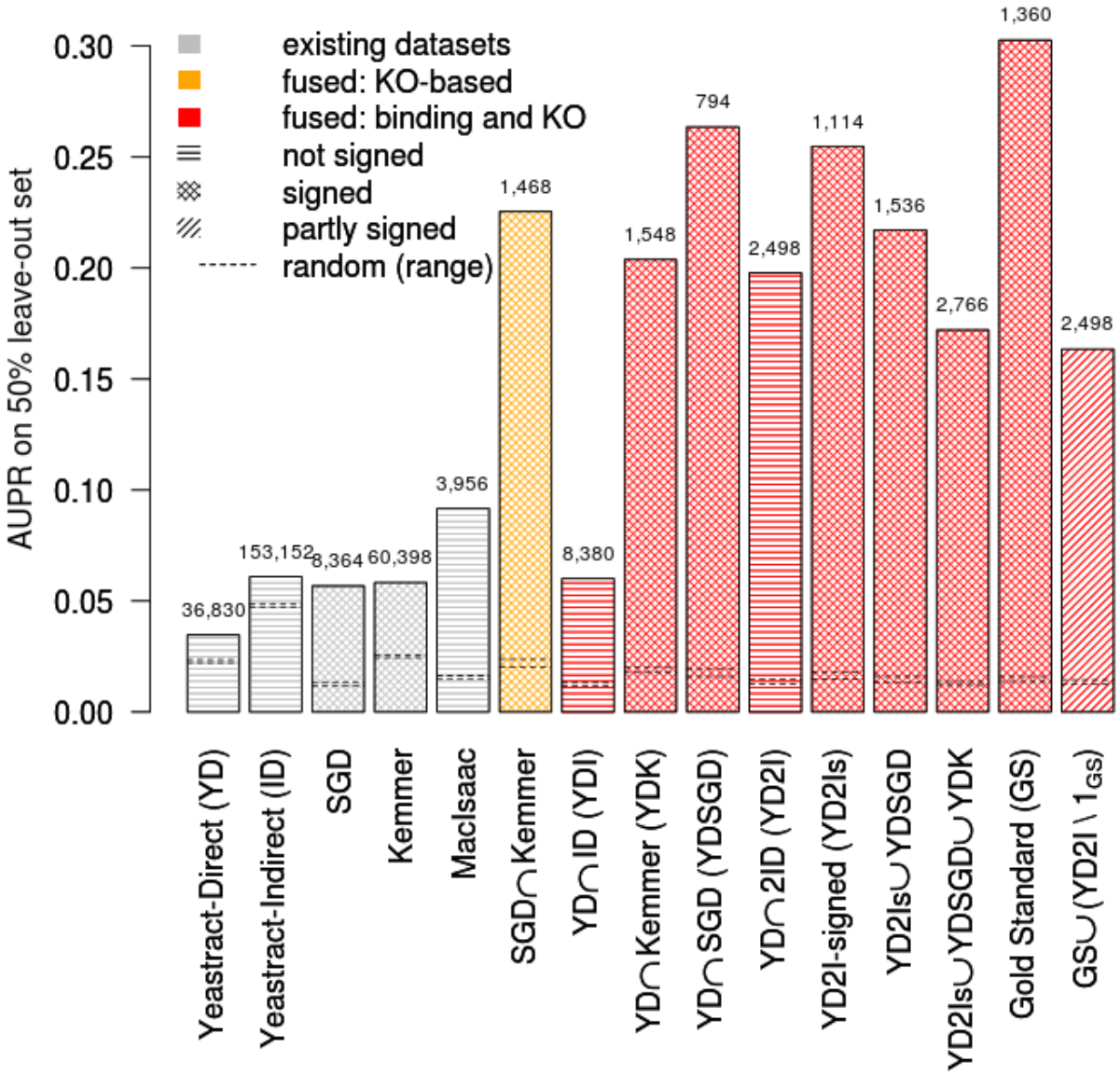

Figure S3: The Gold Standard outperforms other ways of combining the commonly used collections of interactions. Each collection of interactions was split into two equal parts, one for training the Inferelator and one for validation. The height of each bar represents the AUPR calculated on the 50% leave-out set of the corresponding collection of interactions. The number above each bar indicates the number of interactions in the collection, after removing genes and TFs that do not appear in our RNA expression dataset. The color of each bar represents whether it is an existing dataset or a combination of these datasets, either based solely on knock-out evidence (orange), or on an intersection of knock-out and a ChIP binding evidence (red). The shading pattern represents whether none, some, or all interactions are signed. Except for GS, signed interactions came from the signed database in the intersection (Kemmeren or SGD). For the partially signed collection (the rightmost bar), which is a superset of the GS, the same signs as in GS were used when available, and all other interactions were set to 1 (1_{GS} the indicator function of GS, i.e. the GS with all of its interactions set to 1). Standard set theory notation was used, with $\cup$ denoting union, $\cap$ denoting intersection, and $\setminus$ denoting set difference.

6.1.5 Gene Cluster Half-Life Prediction and Validation

The gene cluster that was most prominently enriched in cytoplasmic translation was also the largest cluster (409 genes), and also contained many genes unrelated to translation. Therefore, for predicting RNA half-lives for cytoplasmic translation genes, we used only the genes that were annotated as "cytoplasmic translation" genes in Saccharomyces Genome Database (GO:0002181). This consisted of 171 unique gene names, 127 of which were ribosomal protein coding genes (either RPLs or RPSs). These genes were used to calculate the predicted translation gene RNA half-lives in Figure 6 (see Section 2.6). The list of these genes can be found in Supplementary File SuppData1.zip.

The predicted RNA half-life values for nucleobase containing small molecule (NCSM) metabolism genes were obtained using the genes in the "nucleic acid metabolism" gene cluster in "log-phase growth" and "chemostat" condition clusters. We used the AmiGO 2 website (Carbon *et al.*, 2009) to obtain 207 genes contained in the nucleobase containing small molecule metabolic process category (GO:0055086). These were the genes for which the experimentally measured RNA half-lives were reported in Figure 6 (Neymotin *et al.*, 2014).

6.1.6 Inferelator Worfklow

The Inferelator is a multi-step framework that aims to predict a TRN from expression data (Bonneau *et al.*, 2006; Greenfield *et al.*, 2013; Arrieta-Ortiz *et al.*, 2015). As inputs, it takes RNA expression data as well as orthogonal sources of data. This orthogonal data can originate from Pol II binding, DNA accessibility, knock-out, overexpression assays, or motif data, and is normally compiled into a matrix of Gold Standard (GS) of interactions or prior known interactions, to be used for training and validation of the resulting network. The main idea behind the Inferelator is that the rate of change of RNA levels of a gene i can be expressed as a difference between its transcription and degradation of its transcript, where the transcription term is approximated as a linear combination of the activities of the gene's potential TFs. This section describes the details of how we implemented the Inferelator which were not described in Section 2.4.

6.1.6.1 Estimation of Transcription Factor Activities

We calculate TF activities (TFA) of a given TF from the expression levels of its prior known targets in the respective conditions, as described in (Liao *et al.*, 2003; Arrieta-Ortiz *et al.*, 2015). The primary motivation for calculating TFA, rather than using a TF's RNA expression levels as a proxy for its activity, is that the expression level of a TF is a poor proxy for its activity, and the activity is much better estimated by examining the expression level changes of a TF's known targets, as demonstrated in Arrieta-Ortiz *et al.* (2015). We calculate TFA by expressing the response variable, as defined in Equation 2, as a linear combination of TFA levels of each of its known regulators.

Mathematically, this relation is expressed as such:

$$\text{Resp} = PA, \tag{3}$$

where P is the connectivity matrix, A is the full TFA matrix (with rows corresponding to TFs and columns to samples), and Resp is the *response matrix* with each row corresponding to the response variable as defined in Equation 2. The elements of the connectivity matrix P come from the GS or its subset, such that a value of 0 corresponds to no known interaction, 1 to a known positive interaction (activation), and -1 to a known negative interaction (repression). We calculate the activity matrix A by multiplying both sides by the pseudoinverse of P.

6.1.6.2 Model Selection and Prediction Confidence Score

Selecting the set of TFs that regulate genes involves several steps and is the essence of TRN inference, because this is the process in which we select the most likely regulatory network for every gene. The first step involves narrowing down the list of all genes known to be acting as TFs in the organism (in this paper, we use all genes labeled as "transcription factors" or "DNA-binding genes" in SGD as well as all regulators in the YEASTRACT collection of interactions), to a small set of p TFs that are specific to a given gene. This step is completed using time-lagged Context Likelihood of Relatedness (tlCLR), a method for calculating context-dependent mutual information between gene i and its potential TF (Greenfield *et al.*, 2010; Madar *et al.*, 2010). Typically when using the Inferelator, this value is set to $p = 10$, which is what we use for this paper as well. In addition to these p regulators, TFs known to regulate gene i from prior known interactions are appended to its set of potential regulators. Denote this set as $\mathcal{P}_i$.

Selecting the most accurate model of regulation for a gene i is thereby equivalent to selecting the set of regulators $P_i \subset \mathcal{P}_i$ that optimizes an objective function. The choice of the objective function is a version of the Bayesian Information Criterion (BIC), modified with a Zellner's g-prior, which incorporates prior known interactions into the

model by reducing the sparsity penalty for prior known regulatory interactions. This approach, known as *Bayesian Best Subset Regression* (BBSR), is described in more detail in Greenfield *et al.* (2013). Note that for this step, we linearly scaled and shifted every gene in the response matrix and every TF in the design (TFA) matrix such that every row has mean 0 and variance 1 across samples, which was necessary in order to avoid having to estimate an extra y-intercept parameter corresponding to basal transcription rate.

Once the model P_i that minimizes the BIC is selected, we employ a computational knock-out approach to calculate the confidence we have in each predicted regulatory interaction. The confidence score of the predicted interaction between gene i and TF $j \in P_i$ is determined in the following manner. We first determine the parameters $\beta_{i,k}$ for all $k \in P_i$ by performing linear regression on Equation 2, and use them to calculate the right-hand side of Equation 2. Denote this as the *predicted profile*. Accordingly, the left-hand side of Equation 2 is denoted as the *observed profile*. Let $\sigma^2_{i,j}$ be the variance of residuals between the predicted profile and the observed profile. Let $P_i^{\neg j}$ be the same model of regulation of gene i, but without TF j. Then let $\sigma^2_{i,\neg j}$ be the variance of residuals between the observed profile and the predicted profile as calculated using $P_i^{\neg j}$. Then the confidence score of the interaction between TF j and gene i is defined as

$$\text{conf}_{i,j} = 1 - \frac{\sigma^2_{i,j}}{\sigma^2_{i,\neg j}} \tag{4}$$

To avoid overfitting and derive empirical confidence intervals for model parameters, tlCLR and BBSR are repeated on different but possibly overlapping subsets of the response matrix, denoted as *bootstraps* of the expression data. Every such bootstrap is a column-wise (i.e. sample-wise) subset of the response matrix. The columns are selected randomly using the R function `sample` with default settings. Using this command, we sample n columns with replacement, where n is the total number of columns in the full response matrix. For each bootstrap of the response matrix, the confidence score of every possible interaction is determined using Equation 4. The final *combined confidence score* for every interaction is determined by rank-combining (adding the ranks of) the confidence scores of that interaction across bootstraps. We use 50 bootstraps for all analyses in this paper, except when otherwise specified.

6.1.6.3 Estimating Network Prediction Accuracy

We define *validation set* as the set of known true interactions that we validate our prediction against. Because the output of the Inferelator is a list of predicted interactions, ranked by their cobmined confidence scores from highest to lowest, we may define precision and recall as functions of i in the following way: precision is the fraction of the interactions in the top i highest-ranked predicted interactions that are also in the validation set, whereas recall is the number of interactions in the top i highest-ranked predicted interactions divided by the total number of interactions in the validation set. We calculate precision and recall for every value of i in the full list of predicted interactions. Every PR curve is plotted by connecting the precision and recall values for consecutive values of i with straight lines, and the reported Area Under Precision-Recall curve (AUPR) is the area under this curve.

True Positives (TPs) is a function of i that maps i to the number of interactions among the top i highest-ranked predicted interactions that are also in the validation set, whereas False Positives (FPs) is the number of interactions among the top i highest-ranked predicted interactions that are not in the validation set. We plot all TP-FP curves by connecting the TP and FP values for consecutive values of i with straight lines. We report Area Under the ROC curve (AUROC) by dividing the area under the TP-FP curve by the product of the total number of TPs and the total number of FPs.

For calculating precision and recall, as well as TPs and FPs, on a *leave-out set* (i.e. when a certain *leave-in set* of interactions was used for training the network prediction algorithm, i.e. the TFA step and the BBSR step), we use the leave-out set as the validation set for calculations described above, and the confidence scores of predicted interactions that appear in the leave-in set are set to 0, effectively moving them to the end of the ranked list of predicted interactions. The sign of the interactions is not taken into account in any of our AUPR and AUROC calculations. Predictions involving TFs and genes with no prior known interactions in the leave-out set are excluded from AUPR and AUROC calculations.

6.1.7 Methods Comparisons

For implementing Genie3 (Huynh-Thu *et al.*, 2010), we used the R code downloaded from the Genie3 website (http://www.montefiore.ulg.ac.be/~huynh-thu/software.html) on September 22, 2016. We modified the original R code to allow for parallel processing, such that calculations on different genes could be performed on separate processors simultaneously. The implementation of the code was conducted as instructed in the `README` file downloaded

from the same website. The R code for iRafNet (Petralia *et al.*, 2015) was downloaded from the iRafNet website (http://research.mssm.edu/tulab/software/irafnet.html) on January 27, 2016. We implemented it using the parameter specifications recommended in the original iRafNet paper (Petralia *et al.*, 2015). In particular, we built the iRafNet weight matrix `W` using our prior known interactions (the leave-in portion of the Gold Standard) by setting all values of 0 in our leave-in to 0.1 in the weight matrix, and all values of -1 and 1 to 1.1. We set the number of trees `ntree` to 1000, and number of potential regulators to be sampled from every node to `mtry`=`round`$(\sqrt{n_g - 1})$, where n_g is the number of genes, and `round` is the rounding function.

For the "Clustering" (no half-life fitting) approach in Table 1 and Table S2, we set $\tau_i = 20$ for all genes i and all four condition clusters (see Equation 2 for a definition of τ_i), which roughly corresponds to an assumed half-life of 14 minutes. This number was based on the median experimentally measured RNA half-life from various recent studies (Neymotin *et al.*, 2014; Munchel *et al.*, 2011; Miller *et al.*, 2011). For the "Fitting" (but no clustering) approach, we used the optimal value of RNA half-life as determined by maximizing the AUPR on the GS-fit collection of leave-out interactions from Split **B** approach (Figure 1), using the entire RNA expression data set as the input, rather than doing this analysis separately for each condition cluster.

For each of these methods, training was done on the same set of 20 re-samples of the Gold Standard, as described in Sections 2.5 and 2.6, using the Split **B** aproach, and the AUPRs and AUROCs were calculated using the GS-validate set using the same 20 re-sampes (Figure 1). For each re-sample, individual comparisons in terms of AUPRs and AUROCs between different methods were made, and the number of times, out of 20, that the AUPR (or AUROC) of one method was higher than the other is shown in Table 1 (in terms of AUPR) and Table S2 (in terms of AUROC).

Note that the calculation of RNA half-lives differs between the Split **A** and Split **B** approaches. The Split **A** approach reports a single value for each bi-cluster, which is obtained by taking the median of optimal half-lives across the 20 re-samples, where the optimal half-life is the one that maximizes the AUPR calculated on the GS-fit that corresponds to the given re-sample. Taking the median predicted half-life reduces half-life prediction error due to noise and re-sample specific effects. In the Split **B** approach, each RNA half-life prediction is derived from only one GS re-sample, because we prioritized guaranteeing the statistical significance of comparisons between methods in terms of network inference accuracy over RNA half-life prediction accuracy. That is, comparisons between methods were made for each GS re-sample separately, which resulted in 20 points of comparison for each pair of methods, but the method that simultaneously inferred RNA half-life for each bi-cluster did so by maximizing AUPR over GS-fit produced from only one re-sample of the Gold Standard, in contrast with the median over 20 re-samples in the Split **A** approach. Therefore, the magnitude of the increase in inference accuracy due to half-life fitting that we calculate using the Split **B** approach (Table 1) is an underestimate of the actual increase in accuracy of our final predicted network, which is produced using the Split **A** approach.

All of the Inferelator and Genie3 runs that involved the 20x re-sampling of the Gold Standard were performed on the NYU High Performance Cluster (https://wikis.nyu.edu/display/NYUHPC).

6.2 Supplementary Figures and Tables

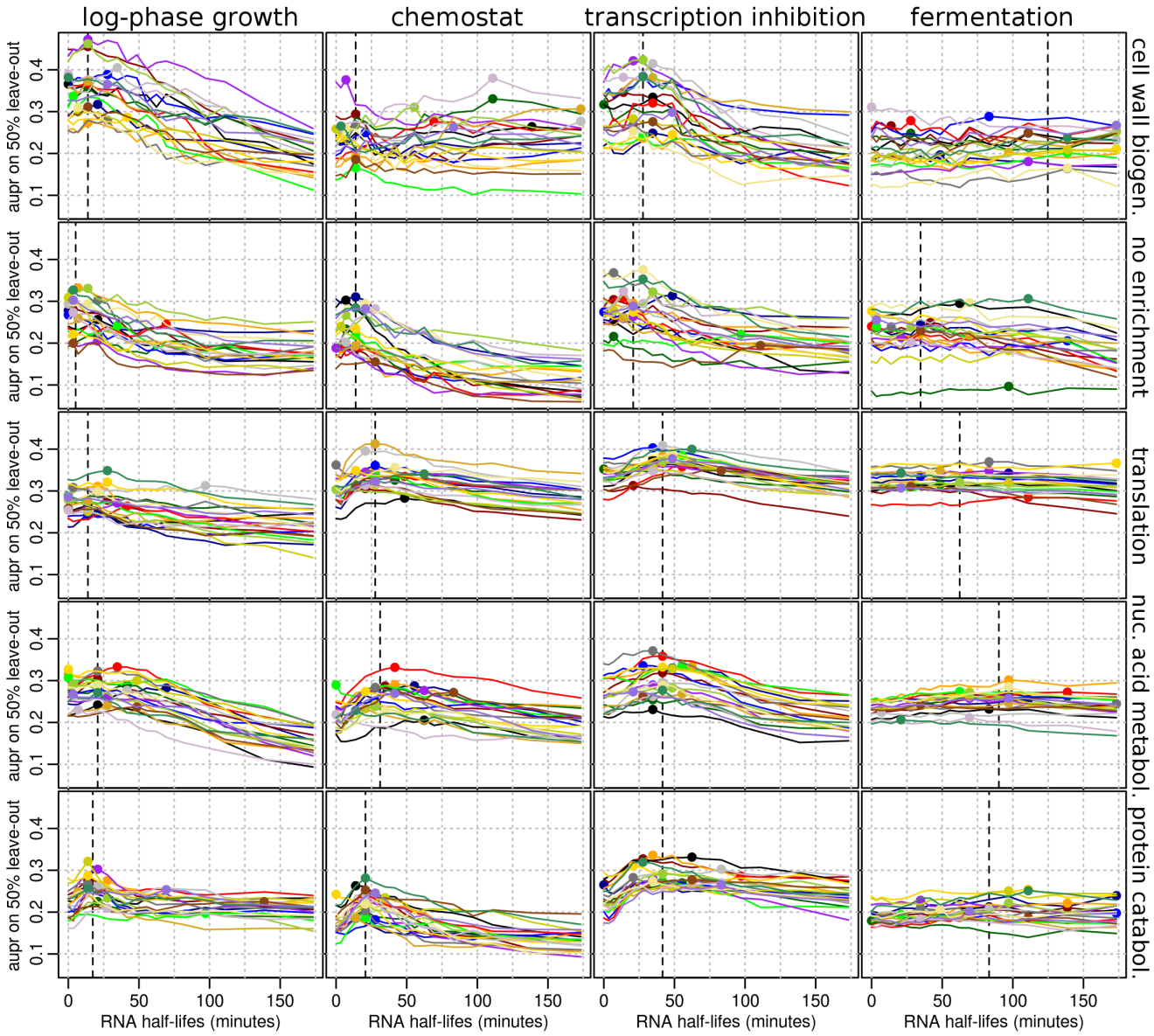

Figure S4: Network inference accuracy is sensitive to RNA half-lives in a condition-specific manner. Each panel shows 20 AUPR plots as as a function of pre-set RNA half-life, when the Inferelator is trained on the the genes specific to the gene cluster (shown on the right) and on conditions specific to the condition cluster (shown on top). Each of the 20 lines corresponds to the 20 Gold Standard re-samples, where 50% of the GS was used for training, and the remaining 50% for calculating AUPR.

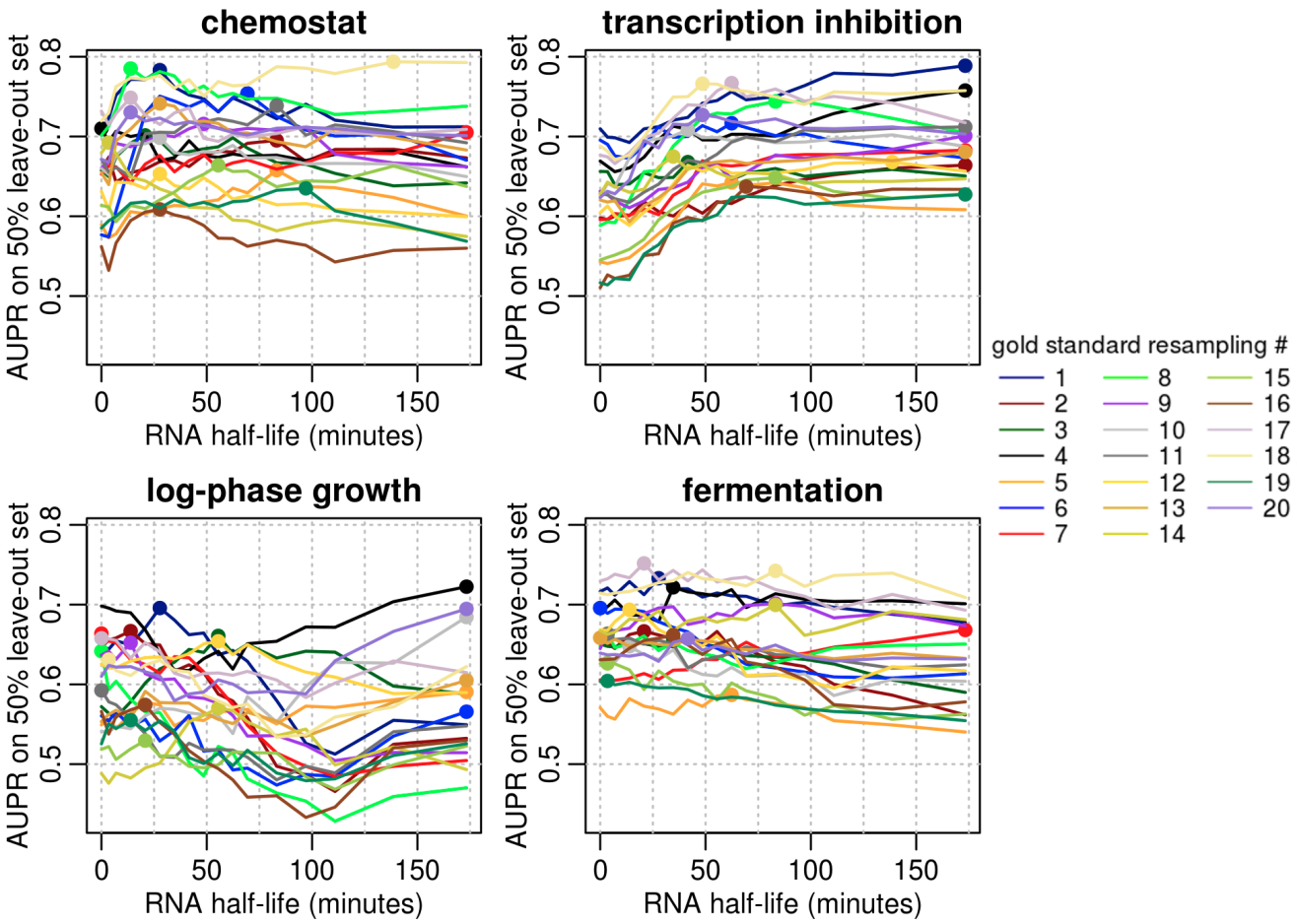

Figure S5: Network inference accuracy is sensitive to RNA half-lives in a condition-specific manner for translation genes. For each cluster, 169 known interactions with 115 "cytoplasmic translation" genes were used to calculate AUPR as a function of RNA half-life. Bold dots correspond to the optimal RNA half-lives for each GS re-sample.

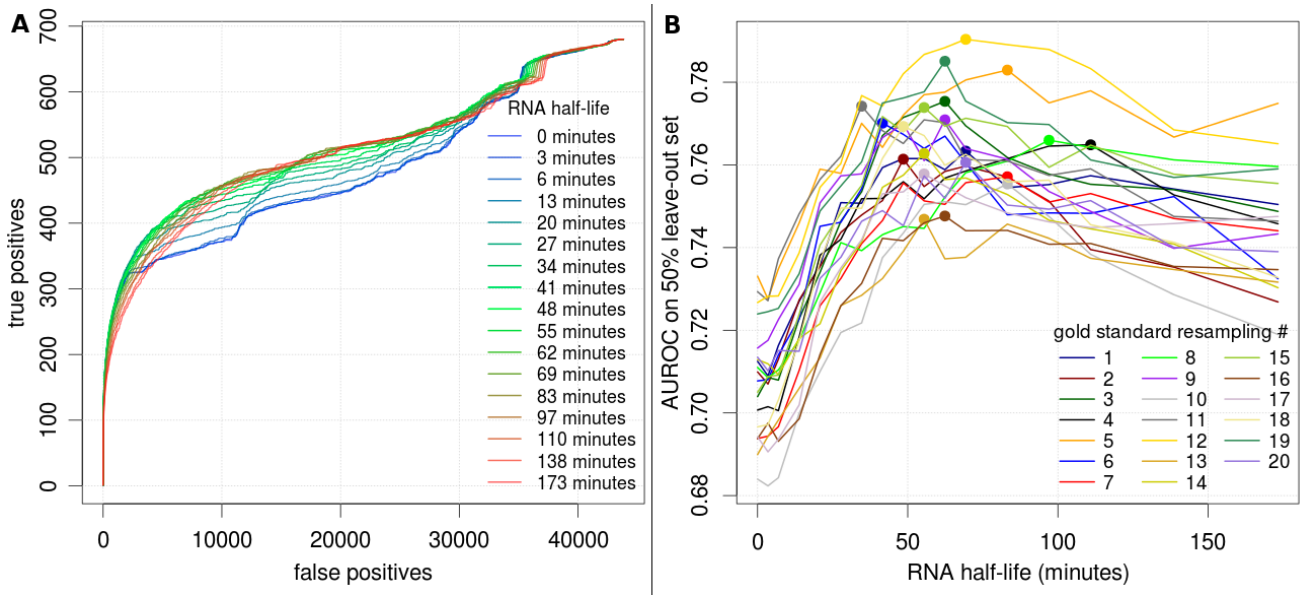

Figure S6: Network Inference is sensitive to RNA half-lives in terms of AUROC. A) True positives vs. false positives curves on the Inferelator output, with each line corresponding to a different pre-set value of RNA half-life. Each line displays the median number of true positives and true negatives across 20 GS re-samplings. B) shows AUROC as a function of pre-set RNA half-life. Different lines denote 20 independent GS re-samplings, and colored dots represent the maximum AUROC for a given GS re-sample.

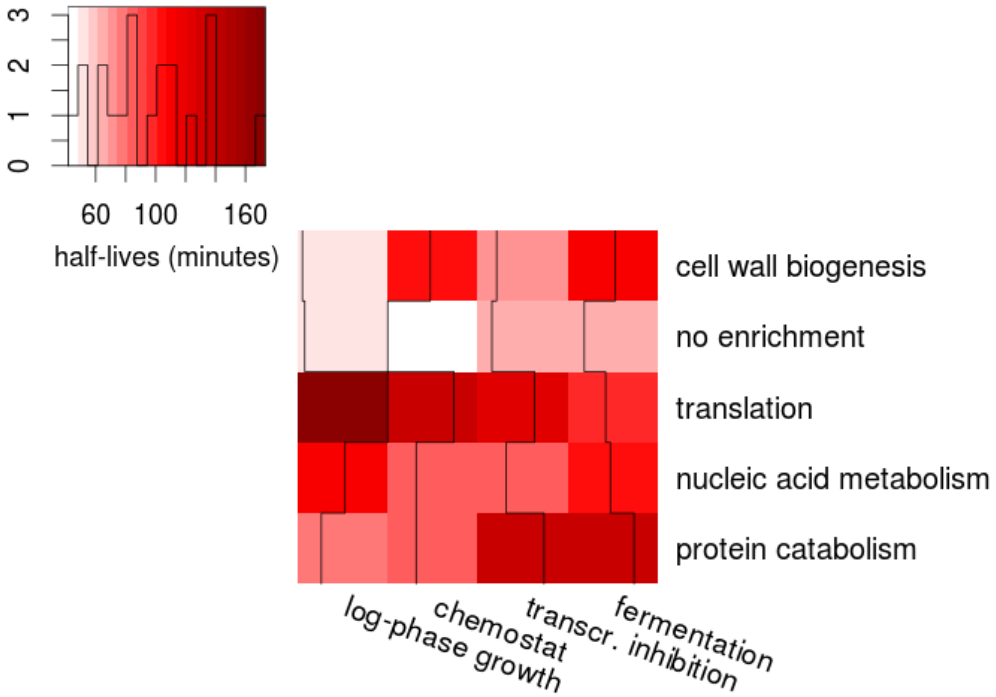

Figure S7: A heatmap of median RNA half-lives that maximized network inference performance with respect to AUROC, for every bi-cluster. This is a summary table of Figure S8.

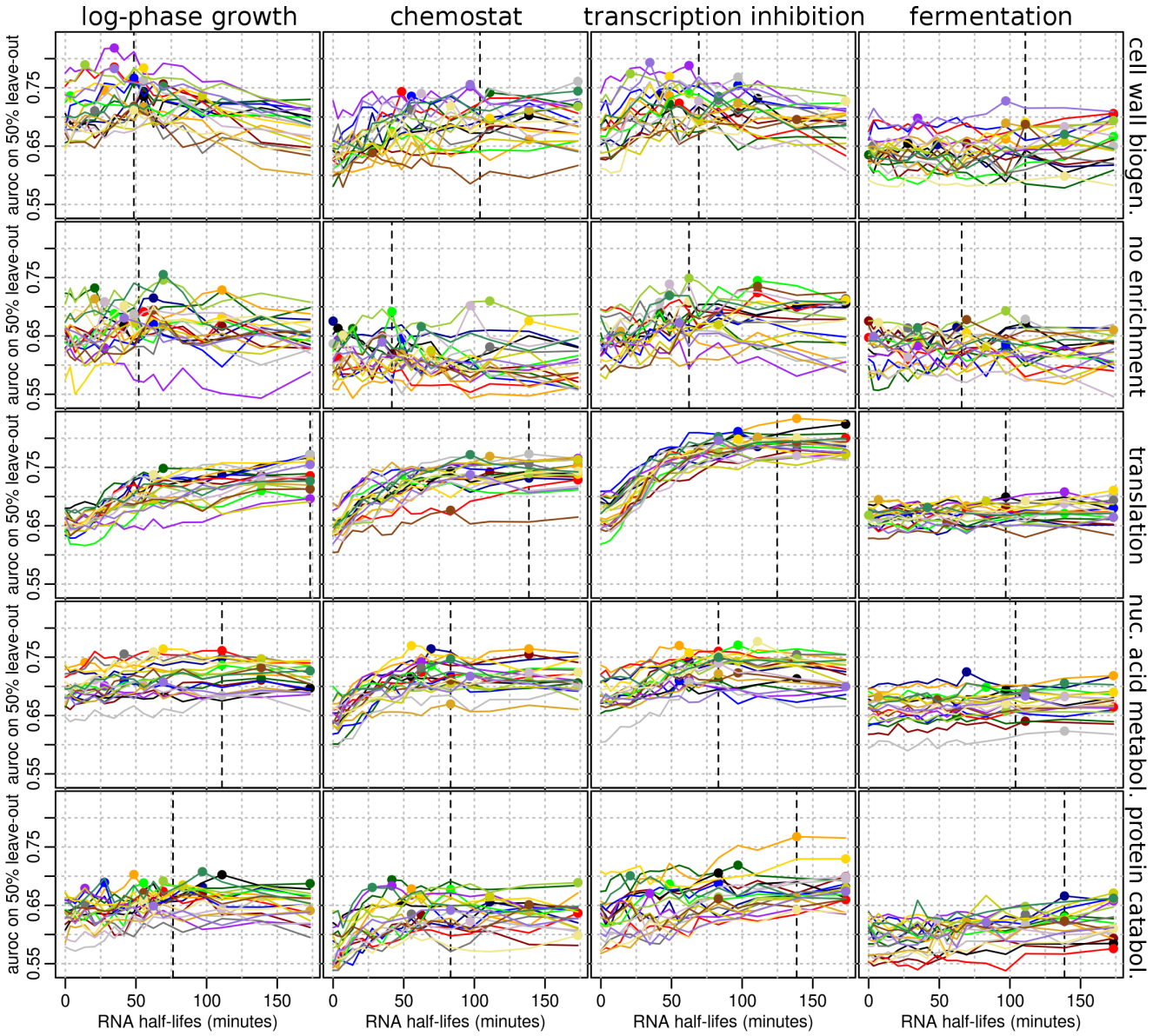

Figure S8: Network inference accuracy is sensitive to RNA half-lives in a condition-specific manner, in terms of area under the ROC curve (AUROC). Each panel shows 20 AUROC plots as as a function of pre-set RNA half-life, when the Inferelator is trained on the the genes specific to the gene cluster (shown on the right) and on conditions specific to the condition cluster (shown on top). Each of the 20 lines corresponds to the 20 Gold Standard re-samples, where 50% of the GS was used for training, and the remaining 50% for calculating AUROC.

Table S2: Both expression data bi-clustering and RNA half-life fitting independently improve Inferelator performance (second column). Furthermore, combining the two modifications improves performance as compared to using each of them separately (third column). Columns 2 and 3 show the number of times the AUROC measured on GS-fit from the same re-sample was higher for one method than the other, given a pair of methods specified by the row and the column. Median AUROC for each approach is reported in the fourth column. See Sections 2.5, 2.6, and 6.1.7 for further details.

Method	Re-samples outperforming no clustering + fitting	Re-samples outperformed by clustering + fitting	Median AUROC
Clustering + Fitting	20/20	-	**0.775**
Clustering	20/20	20/20	0.745
Fitting	19/20	19/20	0.746
none	-	20/20	0.718
Genie3	0/20	20/20	0.573
iRafNet	0/20	20/20	0.555

Table S3: Top 10 new predicted targets for various TFs. The top half of the table lists the 10 TFs that have the highest number of targets in the Gold Standard. For each TF, the number of known targets in the GS is shown in the parentheses, and the top 10 most likely new (i.e. not in the GS) targets are listed. The bottom half displays the median 10 TFs in terms of the number of known targets, out of the 97 TFs with at least one target in the GS. For each TF, targets are listed in the order of decreasing prediction confidence. Superscripts denote whether a given interaction was also present in any of the four large-scale collections of interactions: Yeastract-Direct (1), Yeastract-Indirect (2), SGD (3), and Kemmeren et al. 2014 (4). Bold targets correspond to interactions that are seen in at most one of these four data bases, highlighting completely new interactions. Table 2 displays the precision-based confidence ranks of these interactions, and Supplementary File SuppNetwork3.tsv lists precision and orthogonal validation scores for all 97 TFs in the GS.

TF Name	Targets
Rap1 (192)	**SEC14**^{2}, **YEL1**^{1}, PMT4^{2,3}, FEN1^{2,3}, ALG7^{2,3}, **RBD2**, RPL18B^{1,2,3}, RPL16B^{1,2,3}, **YLR412C-A**^{1}, HXK2^{2,3}
Gcn4 (143)	LYS1^{1,2,3}, **HOM2**^{2}, **BAT1**^{2}, **HIS2**^{2}, **STR2**^{2}, **SDT1**^{2}, **RIB3**^{2}, PSF2^{1,2,3}, **POS5**^{2}, **SRY1**^{2}
Sfp1 (128)	RPS24A^{1,2,3}, RPL18B^{1,2,3,4}, RPS16A^{1,2,3}, RPL8A^{2,3}, RPS18B^{1,2,3}, RPS18A^{2,3}, RPL40B^{2,3}, RPL42B^{2,3}, RPS22A^{2,3,4}, RPL9A^{2,3}
Msn2 (61)	YDR391C^{2,3,4}, CMK2^{2,3}, **YLR257W**^{2}, STF2^{2,4}, **RCN2**^{2}, HAL5^{1,2}, **OXR1**^{2}, PNC1^{2,4}, **DOA1**^{1}, **MRP8**^{2}
Sok2 (60)	**YRO2**^{2}, **UIP**^{2}, YNL194C^{1,2,4}, **GAD1**^{2}, JIP4^{2,4}, MSC1^{2,4}, **TFS1**^{2}, YNL195C^{2,4}, **YJR096W**^{2}, **OM45**^{2}
Yap1 (51)	**GAC1**, TAH18^{1,2} **ISF1**^{2}, YNR034W-A^{1,2}, **MBR1**^{2}, **MRK1**, TCB2^{1,3}, **GAT2**^{2}, YLR460C^{1,2,3}, ATR1^{1,2}
Hsf1 (48)	OPI10^{2,3}, **GGA1**^{2}, RTC3^{2,3}, YRO2^{2,3}, **APJ1**^{1}, **LSB1**, **YGR127W**, YGR250C^{1,2,3}, **CUR1**^{1}, **SAF1**
Rpn4 (44)	RPN12^{2,4}, RPT2^{1,4}, PRE10^{2,4}, RPT5^{1,2,4}, RPN7^{1,2,3,4}, RPN1^{2,4}, RPN10^{2,4}, PRE7^{2,4}, YBR062C^{2,4}, PUP1^{1,2,3,4}
Abf1 (33)	**YER010C**, **COA3**^{3}, **GET2**^{2}, BSD2^{1,2}, **YAH1**, **TDA5**, **MIM1**^{2}, **YDR541C**^{1}, **MGR2**, **SLC1**^{2}
Tec1 (31)	**TRX2**, CRH1^{1,2}, **YOL019W**^{1}, RAX2^{1,3}, **YOL014W**^{2}, **TGL2**^{2}, **YPS3**^{2}, **PHM8**^{2}, **YBL111C**, **RIB4**
Mga2 (6)	PLB2^{2,4}, TDA4^{2,4}, YLR413W^{2,4}, **ALE1**, FAS1^{1,2}, **HEM13**, SUR2^{2,4}, **NEM1**, **PEX31**, **FHN1**
Mot3 (6)	**FIG2**^{4}, **PRM1**^{2}, **SAG1**^{4}, **TIR1**^{2}, PAU24^{2,4}, AAC3^{2,3,4}, TIR4^{2,4}, **EUG1**^{4}, TIR2^{2,3}, FIG1^{2,4}
Ino2 (5)	HNM1^{1,4}, SAH1^{1,2,4}, FAS1^{1,2,3,4}, **SAM2**^{4}, CHO1^{2,4}, EPT1^{2,4}, ADO1^{1,3,4}, OPI3^{2,4}, EHT1^{2,4}, **YIP3**^{4}
Msn4 (5)	YHR022C^{1,2}, **CRS5**^{2}, **JLP1**, **SNO4**^{2}, **FRE7**^{2}, **PUG1**^{2}, YEL073C^{1,2}, PDC6^{2,3}, **GCY1**^{2}, FMP48^{2,3}
Fkh2 (5)	AIM20^{1,2}, HOF1^{1,2}, ALK1^{1,2,3}, CLB1^{1,2}, **YMR030W-A**, BUD4^{1,2,3}, **CDC5**^{2},**YMR001C-A**, **KIN3**^{1}, **HST3**^{1}
Flo8 (5)	**FLO9**^{1}, **TIR4**, **DED81**, **PAU7**, **MSC7**, **PAU5**, **ERG24**, **PAU24**, **NCP1**, **YGL108C**
Rtg3 (4)	IDH2^{1,2,3,4}, PDH1^{2,4}, **AAT2**^{2}, **CPR4**, **ARP3**, **WTM1**^{2}, IDP1^{1,3,4}, **URA8**, **ADE3**, **SLP1**
Stp1 (4)	VHR2^{2,3,4}, **MET17**, **MET3**^{2}, **SUL1**, MET10^{2,3}, MET5^{2,3}, **MEP2**^{4}, **PET8**, GNP1^{1,2,3,4}, DAL80^{2,4}
Leu3 (4)	OAC1^{2,3,4}, BAT1^{2,3,4}, **FRS2**^{2}, ILV5^{2,3}, LEU1^{2,3,4}, **SSB1**, MAE1^{2,3}, **PAB1**^{3}, **YPL260W**, **ILV1**
Nrg1 (3)	**PDE2**, **PRM7**, **VTS1**, **YLR407W**^{2}, **GFD2**, **YGR079W**, **YNL095C**, **TRM5**, **PHO90**, **MCH5**^{1}

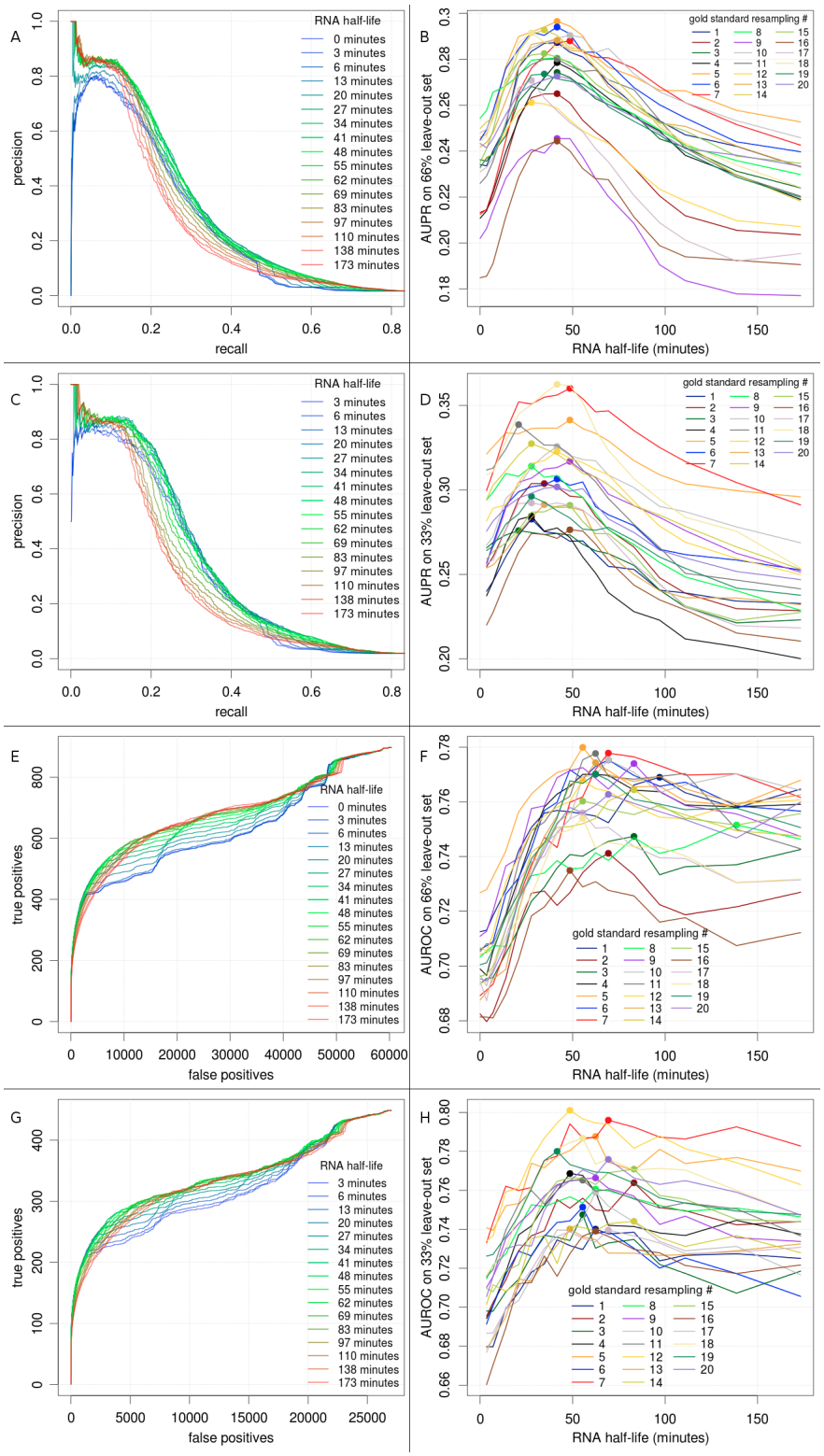

Figure S9: Network inference accuracy is sensitive to RNA half-lives in terms of AUPR and AUROC, independently of the percentage of the Gold Standard used for training. A) and B) The precision-recall curves and the AUPRs as a function of RNA half-life, respectively, for the regime in which 34% of the GS was used for training the Inferelator, and the rest - as a leave-out for validation. C) and D) The same metrics but for a regime in which 67% of the GS was used for training, and the rest for validation. E) and F) The true positives vs. false positives curves and the AUROCs as a function of RNA half-life, respectively, for the regime in which 34% of the GS was used for training, and the rest - for validation. G) and H) The same metrics as E and F, but with 67% of GS as the training set.

6.3 Supplementary Data

SuppNetwork1.tsv is the final network of predicted interactions at a 0.5 precision cutoff. Each row corresponds to a target gene, each columns corresponds to a TF, and a value of 1 represents a predicted regulatory interaction between the TF and the target gene, whereas a value of 0 represents a lack of regulatory interaction.
SuppNetwork2.tsv is a list of all possible TF-gene pairs, ranked in the decreasing order of confidence that there is a regulatory interaction between the pair, according to our final network prediction. The first and the third column list the TF, using its Systematic name and Common name, respectively. The second and fourth columns list the target gene (OLN), using Systematic name and Common name, respectively. The fifth column lists precision value of predictions up to the corresponding interaction, calculated using the Gold Standard (GS). The sixth column represents an interaction's value in the GS (1 if activating, -1 if repressing, 0 if neither or unknown). Columns 7-10 represent the interaction's value in Yeastract-Direct, Yeastract-, SGD (all values, not signed), and Kemmeren, respectively.
SuppNetwork3.tsv lists the top 1000 newly predicted targets (i.e. these interactions are not in the GS) for the 97 TFs that have known targets in the GS. The first column for each TF lists the targets, the second column lists the precision value of the corresponding TF-target interaction, and the third column represents ":"-delimited numbers with the value of this interaction in Yeastract-Direct (YD), Yeastract-Indirect (YI), SGD, and Kemmeren (K), respectively. For example, "1:0:1:-1" means that the interaction was present in YD, SGD, and Kemmeren, but not YID, and furthermore that the interaction was marked as a repression event in the Kemmeren colleciton of interacitons.
SuppDoc1.pdf describes the results of a manual inspection of the Gold Standard targets of the top 15 TFs, listed in the order of decreasing number of known targets in the GS. For each TF, we include the number of positive and negative targets in the GS, known biological attributes of the TF according to SGD, and a discussion of whether the interactions contained in the GS, which were obtained automatically by combining several large data sets, correspond to the established knowledge about the function and the regulation of that TF.
SuppData1.zip lists all of the input files required for the inference procedure, including the RNA expression data, the Gold Standard, the meta data, TF names, gene membership and GO enrichments for each gene cluster, cluster membership for each condition cluster, all of the collections of interactions listed in Figure S3, and the list of translation and NCSM metabolism genes used. For more information, see note.txt inside the zip file.
SuppData2.zip lists all terms enriched in condition cluster annotation analysis, including a document that describes a manual inspection and confirmation of these automatically-generated assignments. For more information, see note.txt inside the zip file.
SuppData3.zip lists all gene-specific and condition-specific τ's predicted in our analysis. The τ parameter is defined in Equation 2, and can be converted to RNA half-lives by being multiplied by ln(2).
The R code is available upon request.