TITLE PAGE

Title

Predicting transcription factor activity using prior biological information

Authors and affiliations

Joseph Estabrook[1,2,3,9], William M. Yashar[2,4,5,9], Hannah D. Holly[1,2], Julia Somers[1,2], Olga Nikolova[2,4,5], Özgün Barbur[7], Theodore P. Braun[2,4,6], Emek Demir[1,2,4,8,10]

[1] Department of Molecular and Medical Genetics, Oregon Health and Science University, 3181 SW Sam Jackson Park Rd; Portland, OR 97239, USA
[2] Knight Cancer Institute, Oregon Health & Science University; Portland, OR, 97239, USA
[3]Cancer Early Detection Advanced Research Center, Oregon Health & Science University; Portland, OR, 97239, USA
[4] Division of Oncologic Sciences, Department of Medicine, Oregon Health & Science University; Portland, OR, 97239, USA
[5] Department of Biomedical Engineering, Oregon Health & Science University; Portland, OR, 97239, USA
[6] Division of Hematology & Medical Oncology, Department of Medicine, Oregon Health & Science University; Portland, OR, 97239, USA
[7] Computer Science Department, University of Massachusetts; Boston, MA 02125, USA
[8] Pacific Northwest National Laboratories; Richland, WA 99354, USA
[9] These authors contributed equally to this work
[10] Corresponding author. Email: demire@ohsu.edu

Correspondence

Emek Demir
Department of Molecular and Medical Genetics
Oregon Health & Science University
3222 SW Research Drive
Portland, Oregon 97239
demire@ohsu.edu
ORCID ID 0000-0002-3663-7113

ABSTRACT

Detection of aberrant transcription factor activity is critical to understanding the mechanisms underlying disease pathogenesis. While many approaches have been developed, they often fail to produce models that reflect known biological systems. Here, we present Priori, which utilizes literature-supported regulatory relationship information to predict transcription factor activity. In our work, we evaluated the ability of Priori and 16 other methods to detect aberrant activity from 124 single-gene perturbation experiments. We show that Priori identifies perturbed transcription factors with greater sensitivity and specificity than other methods. Furthermore, our work demonstrates that Priori can be used to discover significant determinants of survival as well as mediators of drug response from breast cancer and leukemia primary patient samples.

INTRODUCTION

Dysregulation of transcription factor regulatory networks is a common driver of disease[1,2]. Transcription factors establish and maintain cellular gene expression programs by binding to *cis*-regulatory DNA sequences and recruiting cofactors that aid with gene transcription[3,4]. Many diseases are caused by mutations in these putative regulatory elements as well as in the transcription factors and co-factors that interact with them[2]. In cancer, overexpression of oncogenic transcription factors contributes to tumorigenesis through altered expression of downstream target genes[5–7]. Therefore, detection of abnormal transcription factor activity is valuable for better understating the mechanisms that underly disease pathogenesis.

Gene expression profiling, including RNA sequencing (RNA-seq) is commonly used to monitor dynamic changes in transcription factors and their gene regulatory networks. Initial studies to infer transcription factor activity used gene expression as a proxy for activity[8–10]. However, this approach has several shortcomings. Gene expression is only an indirect measurement of protein activity due to the complex mechanisms controlling protein synthesis and degradation[11–13]. Moreover, feedback loops may alter the expression of transcription factors in response to their regulatory activity[2,14–16]. Reliable predictions of transcription factor activity, therefore, cannot be limited to evaluating expression alone.

An alternative approach is to assess the expression of transcription factors as well as their downstream target genes[8–10]. This approach has two major benefits. First, evaluation of hundreds or thousands of downstream targets instead of a single transcription factor likely improves the prediction robustness. While some of these targets may be context-specific, analyzing them in aggregate likely improves the prediction generalizability across many contexts. Second, as target gene expression is downstream of transcription factor control, these signatures may more accurately reflect the actual transcriptional impact. Therefore, accounting for the downstream impact of transcription factors on its gene regulatory networks is important for activity inference.

Multiple methods have been developed to quantify transcription factor activity from gene expression data. These approaches can be grouped based on how they select gene expression features. Methods like Univariate Linear Model (ULM) and Multivariate Linear Model (MLM) use every gene in a dataset, nominating transcription factors as a covariate that best estimates the expression of all other genes[17]. However, these methods develop activity signatures using genes that may not have a true biological relationship to the transcription factor of interest. Gene set approaches like Over Representation Analysis (ORA), Fast Gene Set Enrichment (FGSEA), and Gene Set Variation Analysis (GVSA) infer activity using sets of either documented transcription factor targets or targets curated by experts[18–20]. These methods test for the enrichment of gene sets as compared to a control or cohort average. While gene set methodologies are simple and popular, they are susceptible to the quality and comparability of gene set signatures[20]. Network inference approaches like AUCell and Virtual Inference of Protein-activity by Enriched Regulon analysis (VIPER) infer gene regulatory networks based on the covariance of transcription factors and its putative targets. The challenge with these approaches is deconvoluting the relationships between many gene regulatory networks as target gene expression can be controlled by multiple transcription factors[2]. Additionally, the problem of network inference is often underdetermined as there are many possible solutions that can explain the underlying data[8,21]. Although other methodologies have constrained these networks using biological data generated from orthogonal modalities, grounding these predictions to actual transcription factor activity remains challenging[21–30].

Here, we propose an approach that uses prior, peer-reviewed biological information to infer transcription factor activity called Priori. Our method has three major advantages over the existing methods. First, Priori uses carefully extracted transcription factor gene regulatory networks from Pathway Commons[31,32]. This resource continually collects information on biological pathways including molecular interactions, signaling pathways, regulatory networks, and DNA binding. Pathway Commons currently contains data from 22 high-quality databases with over 5,700 detailed pathways and 2.4 million interactions. Second, Priori assesses not only the direct transcription factor targets, but secondary target genes in its regulatory network. Finally, Priori uses a two-component model to evaluate the expression of known transcription factors as well as their target genes. Our findings show that Priori detects aberrant transcription factor activity with greater sensitivity and specificity than other methods. Furthermore, we demonstrate that Priori can be deployed to discover significant predictors of survival in breast cancer as well as identify mediators of drug response in leukemia from primary patient samples.

RESULTS

Priori uses prior biological information to infer transcription factor activity

For each transcription factor in an RNA-seq dataset, Priori generates a Total Activity score (Figure 1a). The Total Activity score represents the degree of transcription factor regulation on its gene expression network. This score is composed of two sub-scores: the “Target” and “Regulator” components. The Target component captures the impact of transcription factor regulation on target gene expression, whereas the Regulator component

assesses the expression of transcription factors themselves. In order to calculate the Target component, Priori first identifies the known primary target genes from Pathway Commons (or from a network provided by the user)[31,32]. From this primary gene regulatory network, Priori infers the secondary target genes for each transcription factor as the downstream targets of its primary target genes. Priori can also extract the primary and secondary targets from a user-defined network. Priori then assigns weights to each target gene by correlating the expression of each target gene to its transcription factor. A positive weight indicates that the expression of the target gene is directly proportional to the transcription factor. To identify the targets most impacted by transcription factor regulation, Priori ranks all genes in the RNA-seq dataset. These ranks are subsequently scaled by the transcription factor-target gene weights. The Target component for each transcription factor is then calculated by summing the weighted ranks. Priori accounts for the abundance of the transcription factor itself with the Regulator component. Priori first ranks the expression of known transcription factors in the RNA-seq dataset. These ranks are then weighted by the transcription factor-target gene weights and summed, resulting in the Regulator component. The Total Activity score can subsequently be calculated by summing the Target and Regulator components.

In order to compare Priori's detection of aberrant transcription factor activity to other methods, we used the decoupleR benchmarking workflow (Figure 1b)[33]. decoupleR facilitated a common evaluation scheme to determine how often each method correctly identify perturbed transcription factors that have been knocked-down or over-expressed from RNA-seq data[34]. Using this workflow, we generated transcription factor activity scores for 17 methods, including Priori, using the default parameters. Several of these methods were developed by the authors of the decoupleR workflow, including the standard, normalized (Norm), and corrected (Corr) versions of Weighted Mean (WMEAN) and Weighted Sum (WSUM) as well as Univariate Decision Tree (UDT) and Multivariate Decision Trees (MDT)[33]. We ranked the transcription factor activity scores for each experiment and compared the ranks of perturbed and unperturbed transcription factors. Since the number of unperturbed transcription factors vastly outnumbered the perturbed transcription factors, we used a down-sampling strategy. Finally, we calculated area under the precision recall curve (AUPRC) and receiver operating characteristic (AUROC) metrics for each down-sampling permutation.

Priori identifies perturbed transcription factors with greater sensitivity and specificity than other methods

We generated activity scores for 124 knockdown (n = 94) and over-expression (n = 30) experiments across 62 transcription factors (Supplementary Table 1). *MYC* was the most frequently knocked-down transcription factor (11 experiments), whereas *SOX7* and *MYB* were the most frequently over-expressed transcription factors (3 experiments each; Supplementary Figure 1a, b). To understand the patterns in predicted activity across the methods, we correlated their activity (Figure 2a). The Priori activity scores and activity score p-values were dissimilar from the other methods, indicating that Priori identified a unique pattern of transcription factor activity (Supplementary Figure 1c). Using the decoupleR benchmarking workflow, we determined that Priori detects a distinct pattern of aberrant transcription factor activity.

To assess how well Priori identifies perturbed transcription factors, we calculated AUPRC and AUROC metrics for each method (Figure 2b-d). Priori had a greater AUPRC and AUROC than the other methods across all experiments (Supplementary Table 2). ULM, Norm WMEAN, Norm WSUM, and VIPER were the next closest methods by AUPRC and AUROC. However, Priori identified 10 perturbed transcription factors in 20 experiments that were not detected by these other methods (Figure 2e; Supplementary Figure 1d). Moreover, Priori and ULM also identified 8 other perturbed transcription factors in 19 experiments that were not identified by the other methods. Collectively, these experiments show that Priori detects perturbed transcription factors with greater sensitivity and specificity than other methods.

Priori detects transcription factor perturbations using both the expression of the transcription factor and its target genes

We demonstrated that Priori detects transcription factors that were knocked-down or over-expressed with high sensitivity and specificity. However, it was unclear whether these signatures were driven by the direct effects of transcription factor perturbation on their expression or by the indirect effects on their target genes. To understand this, we used the benchmarking workflow to identify perturbed transcription factors with either the Target or Regulator component scores. The Total Activity scores were more strongly correlated with the Regulator component scores (R^2 = 0.97) than the Target component scores (R^2 = 0.60; Figure 3a). In addition, the Regulator component and Total activity scores had similar AUPRC and AUROC metrics (Figure 3b). However, the Target component alone was the most sensitive and specific in detecting perturbed transcription factors. These data suggest that while transcription factor expression was a strong driver of Priori's activity signatures, it detected aberrant activity by the indirect effects of transcription factor perturbation as well.

To understand the characteristics of perturbed transcription factors detected by each component, we compared their transcription factor activity scores. Across all experiments, the Total Activity scores were not significantly different than their component scores (Figure 3c). However, the Total Activity and Regulator component identified 5 transcription factors (GATA3, MYB, REST, SP3, and TFAP2C) across 10 experiments that was not detected by the Target component (Figure 3d; Supplementary Figure 2a). Despite this, the Total Activity and Regulator component scores were only significantly different from the Target component score in the GATA3 and MYB over-expression experiments (Figure 3e). On the other hand, the Total Activity and Target components identified 2 perturbed transcription factors (NFKB1 and TP63) that was not detected by the Regulator components. In these experiments, the Regulator component score was significantly less than the Regulator component score. Although not significantly different, the median Total Activity score was significantly less than the Regulator component score in these experiments. Together, these analyses highlighted contexts where an assessment of both transcription factor and target gene expression was critical to identify aberrant activity.

FOXA1 transcription factor activity is a significant determinant of survival among patients with breast cancer

Since we demonstrated that Priori identifies perturbed regulators with greater sensitivity and specificity than other methods, we sought to determine whether Priori could be used to understand aberrant transcription factor drivers of disease. We used Priori to generate transcription factor activity scores for 637 patients with invasive breast ductal carcinoma (BIDC)[35]. The Priori scores from these patients clustered by breast cancer subtypes (Figure 4a; Supplementary Figure 3a). BIDC is classified into three molecular subtypes: luminal, HER2, and basal cancers[36–40]. The most common types, luminal and basal breast cancers, are commonly distinguished by hormone receptor expression. Luminal breast cancers express estrogen and progesterone receptors, whereas basal breast cancers do not. Unsupervised clustering of the Priori scores revealed that the predicted activity of transcription factors that regulate the expression of estrogen (*ESR1*), progesterone (*NR3C1*), and HER2 receptors (*ERBB2*) were decreased in the basal breast cancer cluster (Figure 4b). Moreover, the basal breast cancer cluster had decreased predicted GATA3 activity, which is critical to luminal cell specification, as well as increased predicted MYC activity, which is consistent with previous studies[36,41–43]. However, predicted FOXA1 activity, which encodes a forkhead protein associated with mammary gland development, was most variable among all patients and was significantly decreased in the luminal breast cancer cluster (Figure 4c-d; Supplementary Table 3). High *FOXA1* expression is also associated with improved outcomes in patients with estrogen receptor-positive disease[44–46]. Together, these analyses show that we were able nominate known drivers of basal and luminal breast cancer by predicting transcription factor activity with Priori.

To understand the clinical impact of transcription factor activity in breast cancer, we evaluated survival differences among the patients in the BIDC cohort. Patients grouped by their molecular subtypes demonstrated no significant difference in survival (Figure 4e). Since breast cancer relapse usually occurs 5 years following initial treatment, we censored our survival analyses to 60 months[47,48]. When we grouped patients by predicted FOXA1 transcription factor activity scores, patients with low FOXA1 activity had a significantly decreased chance of survival (Figure 4f). Notably, there is no survival difference when patients were instead grouped by *FOXA1* expression or by the predicted activity of other transcription factors implicated in our previous analysis (Figure 4g; Supplementary Figures 3b-g). These data indicate that FOXA1 transcription factor activity, not just expression, is a critical determinant of survival for patients with BIDC. In order to understand the critical genes downstream of FOXA1, we generated a differential gene regulatory network (Figure 4h)[31,32,49]. This network revealed that FOXA1 target genes *ESR1* and *XBP1*, which promotes basal breast cancer progression, are significantly decreased among the patients with basal breast cancer (Supplementary Table 4)[50]. These analyses nominate FOXA1 regulation of *ESR1* and *XBP1* expression as important regulators of survival in BIDC.

FOXO1 transcription factor activity mediates venetoclax resistance in leukemia

Aberrant transcription factor activity is also an important regulator of drug resistance in multiple tumor types[1,51–54]. As we have shown that Priori nominates transcription factor regulation associated with breast cancer survival, we wanted to understand whether Priori could also be used to identify mediators of drug sensitivity. Therefore, we calculated Priori

scores from RNA-seq of 859 patients with acute myeloid leukemia (AML) with characterized *ex vivo* drug sensitivity data[55,56]. To nominate transcription factors mediators of drug sensitivity, we correlated Priori scores to drug response. We found 2,851 significant inhibitor-transcription factor activity relationships (Figure 5a; Supplementary Table 5). Among the strongest correlations was predicted FOXO1 transcription factor activity with venetoclax resistance (R^2 = -0.65; Figure 5b). While *FOXO* genes are frequently reported as tumor suppressors in leukemia and solid tumors, recent studies have highlighted *FOXO1's* role as an oncogene in AML[57–61]. Notably, venetoclax resistance is more highly correlated with predicted FOXO1 activity than *FOXO1* expression alone (Figure 5c). These findings suggest that FOXO1 activity is a mediator of venetoclax resistance in AML.

Prior work has also shown that AML with monocytic differentiation is intrinsically resistant venetoclax-based therapy[62]. We wanted to understand the extent to which *FOXO1* knockdown could sensitize monocytic AML to venetoclax treatment. We used CRISPR-Cas9 to knockdown *FOXO1* in THP-1 cells, a cell line model of monocytic leukemia (Supplementary Figure 4). Both *FOXO1* CRISPR guides significantly increased the venetoclax sensitivity in this cell line model of monocytic AML, suggesting *FOXO1* is an important mediator of venetoclax sensitivity (Figure 5d, e). These findings demonstrate that Priori can detect transcription factor mediators of drug resistance.

DISCUSSION

We have developed Priori, a computational algorithm that infers transcription factor activity using prior biological knowledge. We demonstrated that Priori detects perturbed transcription factors with higher sensitivity and specificity than other methods. Using Priori, we identified FOXA1 activity as a regulator of survival of BIDC and nominated important downstream targets that may contribute to this survival difference. Moreover, we used Priori to identify transcription factor regulators of drug sensitivity in AML. We found that predicted FOXO1 activity is highly associated with venetoclax resistance. We validated these findings in a cell line model of monocytic AML, which is resistant to venetoclax[62].

Priori leverages the Pathway Commons resource to identify known gene regulatory networks in gene expression data. While using the relationships from Pathway Commons enables Priori to ground its findings in peer-reviewed literature, this may limit the discovery of novel regulatory relationships. The analysis of BIDC patient samples revealed that Priori was able to identify known transcription factor drivers of BIDC. However, our findings also showed that Priori detected an under-appreciated relationship between FOXA1 activity and BIDC survival, providing a novel mechanistic hypothesis ready for experimental investigation.

Pathway Commons integrates publicly available RNA, DNA, and protein data sourced from a variety of tissue types. However, Pathway Commons is not designed to curate tissue-specific transcription factor gene regulatory networks. Cell context strongly influences regulatory interactions between transcription factors and their downstream target genes[2,63]. The single-gene perturbation gene expression dataset that we used to evaluate

the transcription factor activity methods were performed in numerous cell types. The size of this dataset precluded a definitive evaluation of tissue-context as a determinant of method performance. However, we designed Priori to allow researchers to include their own regulatory networks for context-specific evaluation of their experiments. Overall, Priori should generate robust predictions that are generalizable across many cellular contexts.

Collectively, Priori, identifies aberrant transcription factor activity with high sensitivity and specificity. Priori is therefore a useful tool for the identification of important regulators of disease pathogenesis or drug sensitivity.

ACKNOWLEDGEMENTS

We would like to thank all of the patients for their precious time and donation of samples supporting this research. We appreciate the OHSU core facilities ExaCloud Cluster Computational Resource and the Advanced Computing Center for their assistance. We would also like to thank Michal R. Grzadkowski for his advice in the early stages of Priori's development.

METHODS

Priori Algorithm

Pre-Processing

Priori normalizes and scales the input RNA-seq data prior to downstream analysis. Priori first filters out counts with a standard deviation less than 0.1 (controlled by the *thresh_filter* flag). Priori $\log_2$ normalizes the remaining counts (x_{gene}). Priori then scales the normalized counts (z_{gene}) using one of four methods: "standard"

$$z_{gene} = \frac{(x_{gene} - \mu_X)}{\sigma_X}$$

(where the mean and standard deviation of all normalized counts are indicated by μ_X and σ_X, respectively), "robust"

$$z_{gene} = \frac{(x_{gene} - X_{median})}{IQR}$$

(where the median normalized counts is indicated by X_{median} and the interquartile range is indicated by IQR), "minmax"

$$z_{gene} = \frac{(x_{gene} - X_{min})}{(X_{max} - X_{min})}$$

(where the minimum and maximum normalized counts are indicated by X_{min} and X_{max}, respectively), or "quant" where the normalized counts are scaled using the inverse of the cumulative distribution function, *F(x)*[64].

$$F(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Priori defaults to the "standard" scaling function.

Network

Priori uses known gene regulatory networks to predict transcription factor activity from RNA-seq data. By default, Priori extracts regulatory relationships from the Pathway Commons database to generate activity scores[31,32]. Users can also generate Priori scores using other gene regulatory networks with the *regulon* flag. The user-defined network must specify the transcription factor (*Regulator*) and their downstream target genes (*Target*). Pathway relationships in Pathway Commons are represented with the BioPAX language[31,65]. BioPAX abstracts major pathway relationships, including metabolic and signaling pathways, gene regulatory networks, and genetic and molecular interactions into a standardized format. However, BioPAX representations of gene regulatory networks cannot be interpreted directly. In order to identify the gene regulatory networks encoded in Pathway Commons, we extract transcription factors and their primary targets using the Patterns software package[66]. From these primary relationships, Priori identifies downstream secondary relationships for each transcription factor. Secondary relationships are the downstream target genes of the primary target genes. Priori then assigns these secondary targets as targets of the transcription factor. Primary and secondary targets are weighted equally. Once the full regulatory network is generated, Priori removes transcription factor regulons with less than 15 downstream targets. Users can control the number of targets needed for a regulon with the *regulon_size* flag.

Target Component

Once the network is prepared, Priori generates a Total Activity score, or a "Priori score", for each transcription factor in an RNA-seq dataset. The Total Activity score is composed of two sub-scores: the Target component, which captures the impact of the transcription factor regulation on its target genes, and the Regulator component, which accounts for the expression of the transcription factor itself. To calculate the Target component for a transcription factor (TF_i), Priori first calculates weights for its target genes ($w_{TF_i,target_j}$). The target gene weight is the summation the F-statistic ($F_{TF_i,target_j}$) and the Spearman correlation coefficient ($R_{TF_i,target_j}$):

$$w_{TF_i,target_j} = (F_{TF_i,target_j} + R_{TF_i,target_j}) * sign(R_{TF_i,target_j})$$

$$w_{TF_i,target_j} = \left(\frac{\left(z_{TF_i} - \mu_{TFs}\right) * \left(z_{target_j} - \mu_{targets}\right)}{\sigma_{TF_i} * \sigma_{targets}} + \left| \frac{cov\left(r_{TF_i}, r_{target_j}\right)}{\sigma_{r_{TF_i}} * \sigma_{r_{target_j}}} \right| \right) * sign\left(\frac{cov\left(r_{TF_i}, r_{target_j}\right)}{\sigma_{r_{TF_i}} * \sigma_{r_{target_j}}} \right)$$

where *i* represents the range of transcription factors in a dataset, *j* represents the range of target genes for a given transcription factor TF_i, and *r* represents the rank of the scaled counts relative to all features in the dataset. The F-statistic is used to control for collinearity in gene expression correlations.

The Target component of the Priori score also accounts for the expression of the target genes. Priori first uses the transcription factor-target gene weights to determine the the direction of regulation. Priori defines *k* down-regulated targets among all *j* target genes as those with $w_{TF_i,target_j} < 0$. Similarly, Priori identifies *l* up-regulated targets among all *j* target genes as those with $w_{TF_i,target_j} > 0$. Priori then ranks the scaled target counts for each transcription factor grouped by their direction of regulation ($r_{target_j,targets_k}$ or $r_{target_j,targets_l}$). Priori weighs the ranks of the scaled counts using the target gene weight $w_{TF_i,target_j}$:

$$r'_{target_j,targets_k} = w_{TF_i,target_j} * r_{target_j,targets_k}$$
$$r'_{target_j,targets_l} = w_{TF_i,target_j} * r_{target_j,targets_l}$$

The resulting Target component for a given transcription factor ($Target_{TF_i}$) is the summation of the weighted ranks:

$$Target_{TF_i} = -1\left(\sum_{k}^{down-regulated} r'_{target_j,targets_k} \right) + \left(\sum_{l}^{up-regulated} r'_{target_j,targets_l} \right)$$

where downregulated genes are scaled by -1 prior to summation. Finally, the Target components for each transcription factor are z-transformed relative to all other transcription factors and then to all other samples in the dataset.

Regulator Component

The Regulator component of the Total Activity score accounts for the expression of transcription factor themselves. To calculate the Regulator component for a transcription factor TF_i, Priori first ranks the transcription factor scaled counts relative to other transcription factors in the dataset, $r_{TF_i,TFs}$. Priori then scales these ranks by the total number of transcription factors, n_{TF}, and normalizes them using a normal distribution:

$$r'_{TF_i} = \frac{1}{\sqrt{2\pi}} e^{-\frac{\frac{r_{TF_i,TFs}{}^2}{n_{TF}+1}}{2}}$$

The Regulator Component $Regulator_{TF_i}$ is then calculated as the dot product of the normalized ranks and target gene weights:

$$Regulator_{TF_i} = w_{TF_i,target_j} \cdot r'_{TF_i}$$

Total Activity Score

The Total activity score, $Total_{TF_i}$, for each transcription factor is the summation of the Target and Regulator components.

$$Total_{TF_i} = Target_{TF_i} + Regulator_{TF_i}$$

<u>Benchmarking Workflow</u>

Pre-Processing

The decoupleR benchmarking workflow has been previously described[33]. The normalized RNA-seq data was linearly shifted by the minimum value so all values were non-negative[34].

Transcription factor activity scores and p-values

With the decoupleR workflow, we generated transcription factor activity scores for 11 methods including Priori: AUCell, FGSEA, GSVA, MDT, ORA, UDT, ULM, VIPER, WMEAN, and WSUM. decoupleR also generated normalized transcription factor activity scores for FGSEA, WMEAN, and WSUM as well as corrected scores for WMEAN and WSUM[33]. We generated transcription factor activity scores for each method using the default parameters. We also calculated p-values for the Priori Total Activity scores, Target component scores, and Regulator component scores using a Student's two-sided t-test.

Area under the receiver operating characteristic and precision recall curves

Using the decoupleRBench package, we generated AUROC and AUPRC[33]. In brief, we defined the "Condition Positives" as the perturbed transcription factors. We ranked the transcription factor activity scores for each method and determined whether the Condition Positives were among the top or bottom "n" scores, depending on the type of experiment. "n" was defined as the number of unique perturbed transcription factors in the dataset. Unlike in the published decoupleRBench method, we did not evaluate the absolute value of the transcription factor activity scores. Instead, we defined a "True Positive" as either an over-expressed regulator among the top "n" scores and knocked-down regulators among the bottom "n" scores. "Condition Negatives" were defined as the unperturbed transcription factors and "True Negatives" were unperturbed transcription factor not ranked among the top or bottom "n" scores. As the Condition Negatives substantially outnumbered the Condition Positives, we deployed a down-sampling strategy to compare an

equal number of Condition Positives and Negatives. AUROC and AUPRC values were calculated for each down-sampling permutation.

Data Analysis

Benchmarking Workflow

Custom scripts were used to evaluate the results of the decoupleR benchmarking workflow. In order to compare the scores across the different methods, we z-transformed the transcription factor activity scores. We calculated the Spearman correlation between the activity scores and p-values of each method.

TCGA BRCA

The normalized gene expression counts and metadata for the Breast Invasive Carcinoma (TCGA, PanCancer Atlas) study were downloaded from cBioPortal https://www.cbioportal.org[35,67]. Priori scores were generated from the normalized gene expression counts linearly shifted by the minimum value. Custom scripts were used to exclude patients without basal or luminal BIDC. The Seurat package was used to perform dimensional reduction on the Priori scores by PCA and UMAP[68]. Survival data was censored at 60 months and analyzed for significance using a log-rank Mantel-Cox test. We assigned patients to "High" or "Low" groups depending on whether they were among the top 75% or bottom 25% of the value of interest. Differential gene expression networks were generated from the normalized gene expression data using CausalPath[49]. An FDR threshold of 0.001 was used to evaluate significant relationships following 100 permutations.

BeatAML

Priori scores were generated from the normalized gene expression counts of baseline AML patient samples from the BeatAML cohort[55,56]. These counts were also linearly shifted by the minimum value. Custom scripts were used to exclude patients without a diagnosis of AML or those with a prior myeloproliferative neoplasm. Priori scores and single-inhibitor drug AUC values on the same patient sample were evaluated using a Spearman correlation. Significant correlations were those with an FDR < 0.05.

Cell Culture

Cell Lines

THP-1 cells (DSMZ) were cultured in RPMI (Gibco) supplemented with 10% fetal bovine serum (FBS, HyClone), 2 mM GlutaMAX (Gibco), 100 units/mL Penicillin, 100 µg/mL Streptomycin (Gibco), and 0.05 mM 2-Mercaptoethanol (Sigma Aldrich). All cells were cultured at 5% CO_2 and 37°C. Cell lines were tested monthly for mycoplasma contamination.

CRISPR

Two *FOXO1* CRISPR guide RNAs in a pLentiCRISPR v2 backbone were obtained from GenScript[69]. The target sequence for guide RNA (gRNA) 1 is GCTCGTCCCGCCGCAACGCG and the sequence for gRNA 2 is ACAGGTTGCCCCACGCGTTG. Additionally, a non-targeting CRISPR guide RNA (target

sequence CCTGGGTTAGAGCTACCGCA) generated by scrambling the target sequence of LentiCRISPRv2-ACTB-C1 in a pLentiCRISPR v2 backbone was obtained from Addgene (catalog #169795). Lentivirus was produced by transfecting Lenti-X 293T cells (Clontech) with the SMARTvector transfer plasmid and packaging/pseudotyping plasmids. psPAX2 was a gift from Didier Trono (Addgene plasmid #12260; http://n2t.net/addgene:12260; RRID:Addgene_12260). The supernatant containing lentivirus was collected after 48 hours of culture and filtered with a 0.45 µm filter. THP-1 cells were transduced with virus via spinnoculation in the presence of polybrene. Transduced cells were selected with 1 µg/mL puromycin to produce a stable cell line.

CRISPR Validation

FOXO1 knockdown was validated using Tracking of Indels by Decomposition (TIDE)[70]. Briefly, cellular DNA was PCR-amplified using primers upstream (sequence AAGTAGGGCACGCTCTTGAC) and downstream (sequence CGTTCCCCCAAATCTCGGAC) of the *FOXO1* gRNA target sequences. The primers were designed in Geneious Prime and synthesized by Integrated DNA Technologies. Paramagnetic beads were used to purify the PCR DNA fragments (MagBio Ref #AC-60001) and subsequently sequenced by EuroFins. Inference of CRISPR Edits (ICE) was performed using the Synthego web tool (https://ice.synthego.com/#/).

Drug Sensitivity Assay

Cells were cultured for 72 hours along a 7-point dose curve with venetoclax. Cell viability was assessed by CellTiter Aqueous colorimetric assay.

Availability of Data and Materials

Priori is free to use and is publicly accessible on GitHub (https://github.com/ohsu-comp-bio/regulon-enrichment). The forked versions of decoupleR (https://github.com/JEstabrook/decoupleR) and the benchmarking workflow (https://github.com/ohsu-comp-bio/decoupleRBench) used in this study can be accessed on GitHub. The benchmark analysis pipeline is freely available on GitHub[71,72] (https://github.com/ohsu-comp-bio/decoupler_workflow). The single-gene perturbation RNA-seq data from *Holland et al.* analyzed in this study is accessible at: https://zenodo.org/record/3564179[34]. The TCGA BRCA RNA-seq data from BIDC patients analyzed during the current study is available on cBioPortal[35,67]. The BeatAML RNA-seq data from AML patients evaluated in this work is available on Vizome[55,56].

AUTHOR CONTRIBUTIONS

J Estabrook: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-review and editing. **WM Yashar**: conceptualization, software, formal analysis, validation, investigation, visualization, methodology, writing-original draft, writing-review and editing. **H Holly**: conceptualization, formal analysis, investigation, writing-review and editing. **J Somers**: formal analysis, investigation, writing-review and editing. **O Ninkolova**: conceptualization, investigation, writing-review and editing. **Ö. Barbur**: conceptualization, investigation, writing-review and editing. **TP Braun**: conceptualization, resources, supervision, funding acquisition, validation, investigation,

writing-review and editing. **E Demir**: conceptualization, resources, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing-review and editing. The co-first authors may identify themselves as lead authors in their respective CVs.

DISCLOSURE OF CONFLICTS OF INTEREST

WMY is a former employee of Abreos Biosciences, Inc. and was compensated in part with common stock options. Pursuant to the merger and reorganization agreement between Abreos Biosciences, Inc. and Fimafeng, Inc., WMY surrendered all of his common stock options in 03/2021. TPB has received research support from AstraZeneca, Blueprint Medicines as well as Gilead Sciences and is the institutional PI on the FRIDA trial sponsored by Oryzon Genomics. The authors certify that all compounds tested in this study were chosen without input from any of our industry partners. The other authors do not have competing interests, financial or otherwise.

REFERENCES

1. Bushweller, J. H. Targeting transcription factors in cancer — from undruggable to reality. *Nat Rev Cancer* **19**, 611–624 (2019).
2. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).
3. Fuda, N. J., Ardehali, M. B. & Lis, J. T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* **461**, 186–192 (2009).
4. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
5. Sanda, T. *et al.* Core Transcriptional Regulatory Circuit Controlled by the TAL1 Complex in Human T Cell Acute Lymphoblastic Leukemia. *Cancer Cell* **22**, 209–221 (2012).
6. Lin, C. Y. *et al.* Transcriptional Amplification in Tumor Cells with Elevated c-Myc. *Cell* **151**, 56–67 (2012).
7. Nie, Z. *et al.* c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68–79 (2012).
8. Khatri, P., Sirota, M. & Butte, A. J. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology* **8**, e1002375 (2012).
9. Nguyen, T.-M., Shafi, A., Nguyen, T. & Draghici, S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol* **20**, 203 (2019).
10. Essaghir, A. *et al.* Transcription factor regulation can be accurately predicted from the presence of target gene signatures in microarray gene expression data. *Nucleic Acids Research* **38**, e120 (2010).

11. de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M. & Vogel, C. Global signatures of protein and mRNA expression levels. *Mol Biosyst* **5**, 1512–1526 (2009).

12. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227–232 (2012).

13. Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J. & Smith, V. A. Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system. *Sci Rep* **5**, 10775 (2015).

14. Kiełbasa, S. M. & Vingron, M. Transcriptional Autoregulatory Loops Are Highly Conserved in Vertebrate Evolution. *PLOS ONE* **3**, e3210 (2008).

15. Benito, J., Zheng, H., Ng, F. S. & Hardin, P. E. Transcriptional feedback loop regulation, function and ontogeny in Drosophila. *Cold Spring Harb Symp Quant Biol* **72**, 437–444 (2007).

16. Bornstein, C. *et al.* A negative feedback loop of transcription factors specifies alternative dendritic cell chromatin states. *Mol Cell* **56**, 749–762 (2014).

17. Teschendorff, A. E. & Wang, N. Improved detection of tumor suppressor events in single-cell RNA-Seq data. *npj Genom. Med.* **5**, 1–14 (2020).

18. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).

19. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 1–15 (2013).

20. Hung, J.-H., Yang, T.-H., Hu, Z., Weng, Z. & DeLisi, C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics* **13**, 281–291 (2012).

21. Olsen, C. *et al.* Inference and validation of predictive gene networks from biomedical literature and gene expression data. *Genomics* **103**, 329–336 (2014).

22. Zhu, M., Liu, C.-C. & Cheng, C. REACTIN: regulatory activity inference of transcription factors underlying human diseases with application to breast cancer. *BMC Genomics* **14**, 504 (2013).

23. Gao, F., Foat, B. C. & Bussemaker, H. J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics* **5**, 31 (2004).

24. Tarca, A. L. *et al.* A novel signaling pathway impact analysis. *Bioinformatics* **25**, 75–82 (2009).

25. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).

26. Babur, O., Demir, E., Gönen, M., Sander, C. & Dogrusoz, U. Discovering modulators of gene expression. *Nucleic Acids Res* **38**, 5648–5656 (2010).

27. Walhout, A. J. M. What does biologically meaningful mean? A perspective on gene regulatory network validation. *Genome Biol* **12**, 109 (2011).

28. Barbuti, R., Gori, R., Milazzo, P. & Nasti, L. A survey of gene regulatory networks modelling methods: from differential equations, to Boolean and qualitative bioinspired models. *J Membr Comput* **2**, 207–226 (2020).

29. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. Bioinformatics challenges for personalized medicine. *Bioinformatics* **27**, 1741–1748 (2011).

30. Yngvadottir, B., Macarthur, D. G., Jin, H. & Tyler-Smith, C. The promise and reality of personal genomics. *Genome Biol* **10**, 237 (2009).

31. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, D685–D690 (2011).

32. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res* **48**, D489–D497 (2020).

33. Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).

34. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol* **21**, 36 (2020).

35. Berger, A. C. *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* **33**, 690-705.e9 (2018).

36. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* **98**, 10869–10874 (2001).

37. Perou, C. M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci U S A* **96**, 9212–9217 (1999).

38. Brenton, J. D., Carey, L. A., Ahmed, A. A. & Caldas, C. Molecular classification and molecular forecasting of breast cancer: ready for clinical application? *J Clin Oncol* **23**, 7350–7360 (2005).

39. Tamimi, R. M. *et al.* Comparison of molecular phenotypes of ductal carcinoma in situand invasive breast cancer. *Breast Cancer Res* **10**, 1–9 (2008).

40. Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418–8423 (2003).

41. Kouros-Mehr, H. & Werb, Z. Candidate regulators of mammary branching morphogenesis identified by genome-wide transcript analysis. *Dev Dyn* **235**, 3404–3412 (2006).

42. Kouros-Mehr, H., Slorach, E. M., Sternlicht, M. D. & Werb, Z. GATA-3 Maintains the Differentiation of the Luminal Cell Fate in the Mammary Gland. *Cell* **127**, 1041–1055 (2006).

43. Sotiriou, C. *et al.* Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* **100**, 10393–10398 (2003).

44. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406**, 747–752 (2000).

45. Lupien, M. *et al.* FoxA1 Translates Epigenetic Signatures into Enhancer-Driven Lineage-Specific Transcription. *Cell* **132**, 958–970 (2008).

46. Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D. & Carroll, J. S. FOXA1 is a key determinant of estrogen receptor function and endocrine response. *Nat Genet* **43**, 27–33 (2011).

47. Wangchinda, P. & Ithimakin, S. Factors that predict recurrence later than 5 years after initial treatment in operable breast cancer. *World J Surg Onc* **14**, 1–8 (2016).

48. Sestak, I. *et al.* Factors Predicting Late Recurrence for Estrogen Receptor–Positive Breast Cancer. *JNCI: Journal of the National Cancer Institute* **105**, 1504–1511 (2013).

49. Babur, Ö. *et al.* Causal interactions from proteomic profiles: Molecular data meet pathway knowledge. *PATTER* **0**, (2021).

50. Chen, X. *et al.* XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. *Nature* **508**, 103–107 (2014).

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

52. Alessandrini, F., Pezzè, L., Menendez, D., Resnick, M. A. & Ciribilli, Y. ETV7-Mediated DNAJC15 Repression Leads to Doxorubicin Resistance in Breast Cancer Cells. *Neoplasia* **20**, 857–870 (2018).

53. Neel, D. S. & Bivona, T. G. Resistance is futile: overcoming resistance to targeted therapies in lung adenocarcinoma. *NPJ Precis Oncol* **1**, 3 (2017).

54. Matkar, S. *et al.* An Epigenetic Pathway Regulates Sensitivity of Breast Cancer Cells to HER2 Inhibition via FOXO/c-Myc Axis. *Cancer Cell* **28**, 472–485 (2015).

55. Tyner, J. W. *et al.* Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).

56. Bottomly, D. *et al.* Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. *Cancer Cell* **40**, 850-864.e9 (2022).

57. Hou, T., Li, Z., Zhao, Y. & Zhu, W.-G. Mechanisms controlling the anti-neoplastic functions of FoxO proteins. *Semin Cancer Biol* **50**, 101–114 (2018).

58. Kornblau, S. M. *et al.* Highly Phosphorylated FOXO3A Is an Adverse Prognostic Factor in Acute Myeloid Leukemia. *Clinical Cancer Research* **16**, 1865–1874 (2010).

59. Obrador-Hevia, A., Serra-Sitjar, M., Rodríguez, J., Villalonga, P. & Fernández de Mattos, S. The tumour suppressor FOXO3 is a key regulator of mantle cell lymphoma proliferation and survival. *Br J Haematol* **156**, 334–345 (2012).

60. Gurnari, C., Falconi, G., De Bellis, E., Voso, M. T. & Fabiani, E. The Role of Forkhead Box Proteins in Acute Myeloid Leukemia. *Cancers* **11**, 865 (2019).

61. Lin, S. *et al.* A FOXO1-induced oncogenic network defines the AML1-ETO preleukemic program. *Blood* **130**, 1213–1222 (2017).

62. Pei, S. *et al.* Monocytic Subclones Confer Resistance to Venetoclax-Based Therapy in Patients with Acute Myeloid Leukemia. *Cancer Discov* **10**, 536–551 (2020).

63. MacNeil, L. T. *et al.* Transcription Factor Activity Mapping of a Tissue-Specific In Vivo Gene Regulatory Network. *cels* **1**, 152–162 (2015).

64. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

65. Demir, E. *et al.* BioPAX – A community standard for pathway data sharing. *Nat Biotechnol* **28**, 935–942 (2010).

66. Babur, Ö. *et al.* Pattern search in BioPAX models. *Bioinformatics* **30**, 139–140 (2014).

67. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401–404 (2012).

68. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).

69. Sanjana, N. E., Shalem, O. & Zhang, F. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods* **11**, 783–784 (2014).

70. Brinkman, E. K., Chen, T., Amendola, M. & van Steensel, B. Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Research* **42**, e168 (2014).

71. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at https://doi.org/10.12688/f1000research.29032.2 (2021).

72. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **34**, 3600–3600 (2018).

FIGURES

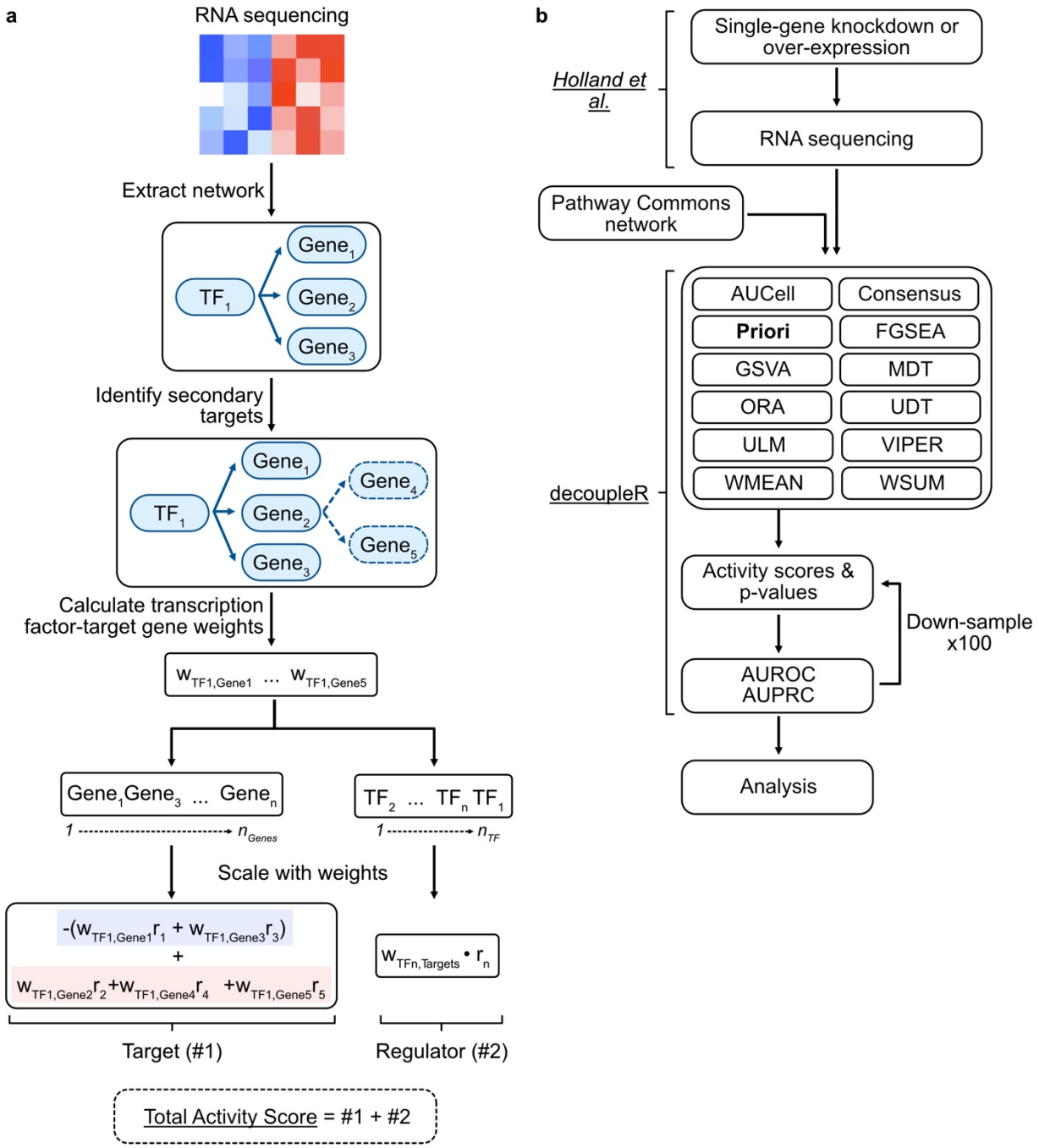


Figure 1: Overview of the Priori methodology and benchmarking workflow. a. Priori generates a Total Activity score for each transcription factor in an RNA-seq dataset. There are two components to the Total Activity score: the Target component, which assesses target gene expression, and the Regulator component, which evaluates the expression of the transcription factor itself. For each transcription factor, Priori first extracts their downstream target genes from Pathway Commons. After determining secondary targets from the regulatory network, Priori then calculates weights for each target gene. The

weights are calculated by correlating of the expression of each transcription factor to its target genes. Priori then ranks all genes in the dataset and scales these ranks by the transcription factor-target gene weights. The summation of the weighted ranks is the Target component. The Regulator component instead ranks transcription factor expression relative to the other transcription factors in the dataset. The Regulator component is the dot product of the transcription factor-target weights and the transcription factor expression ranks. The Total Activity score is calculated by summing the Target and Regulator component scores. b. Schematic overview of the benchmarking workflow. We generated transcription factor activity scores for each method using normalized RNA-seq counts following single-gene knockdown or over-expression. Priori, along with other methods that use regulatory networks, generate scores using the Pathway Commons network. AUROC and AUPRC values are calculated for each down-sampling permutation. 100 down-sampling permutations are performed to compare an equal number of perturbed and unperturbed genes.

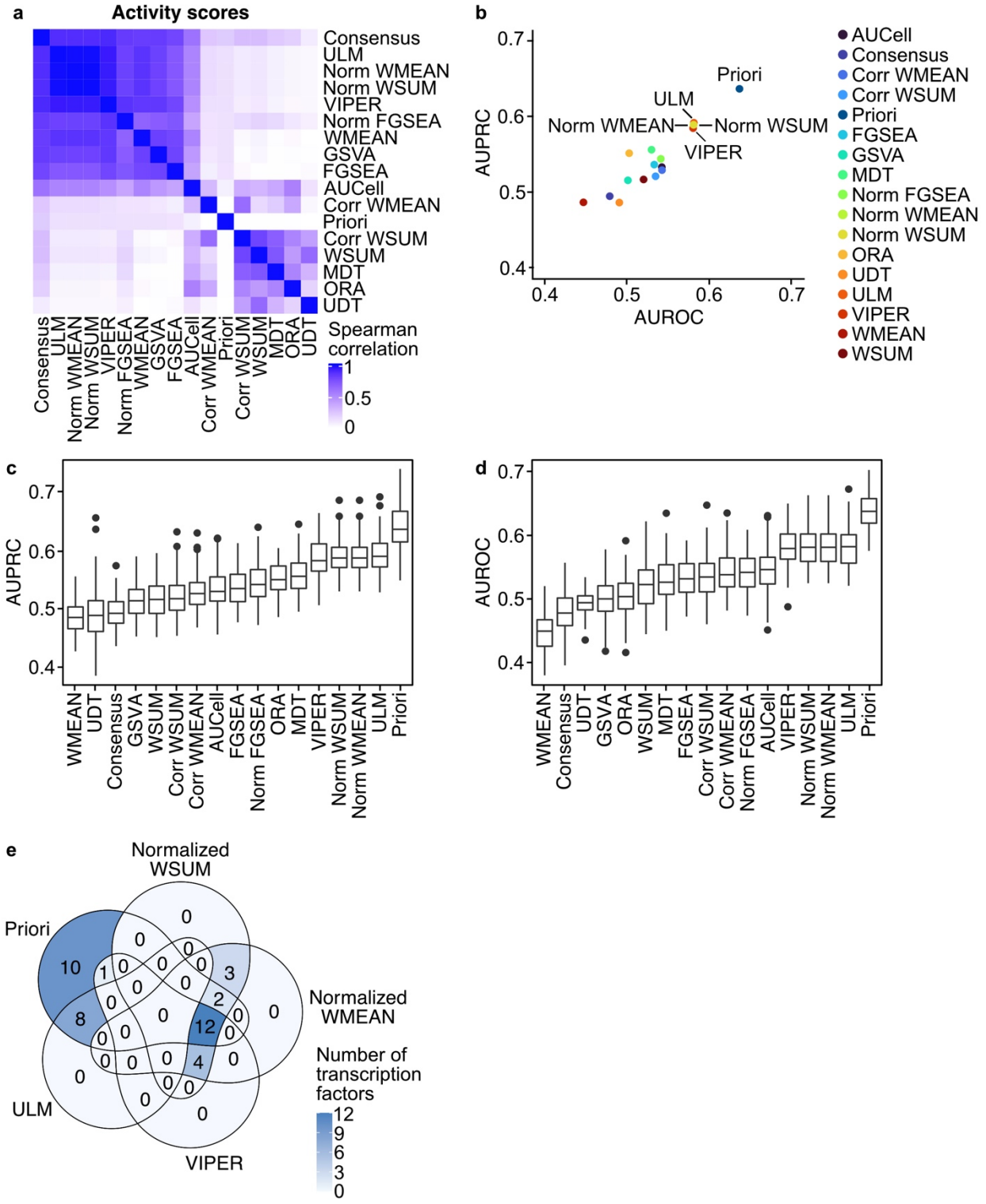


Figure 2: Priori detects aberrant transcription factor activity with high sensitivity and specificity. a. Spearman correlation of transcription factor activity scores for each method. b. Median AUPRC and AUROC values across the 100 down-sampling permutations for each method. c, d. The distribution of (c) AUPRC and (d) AUROC values across

the down-sampling permutations. e. Overlap of perturbed transcription factors identified by the top 5 methods by AUPRC and AUROC from (b).

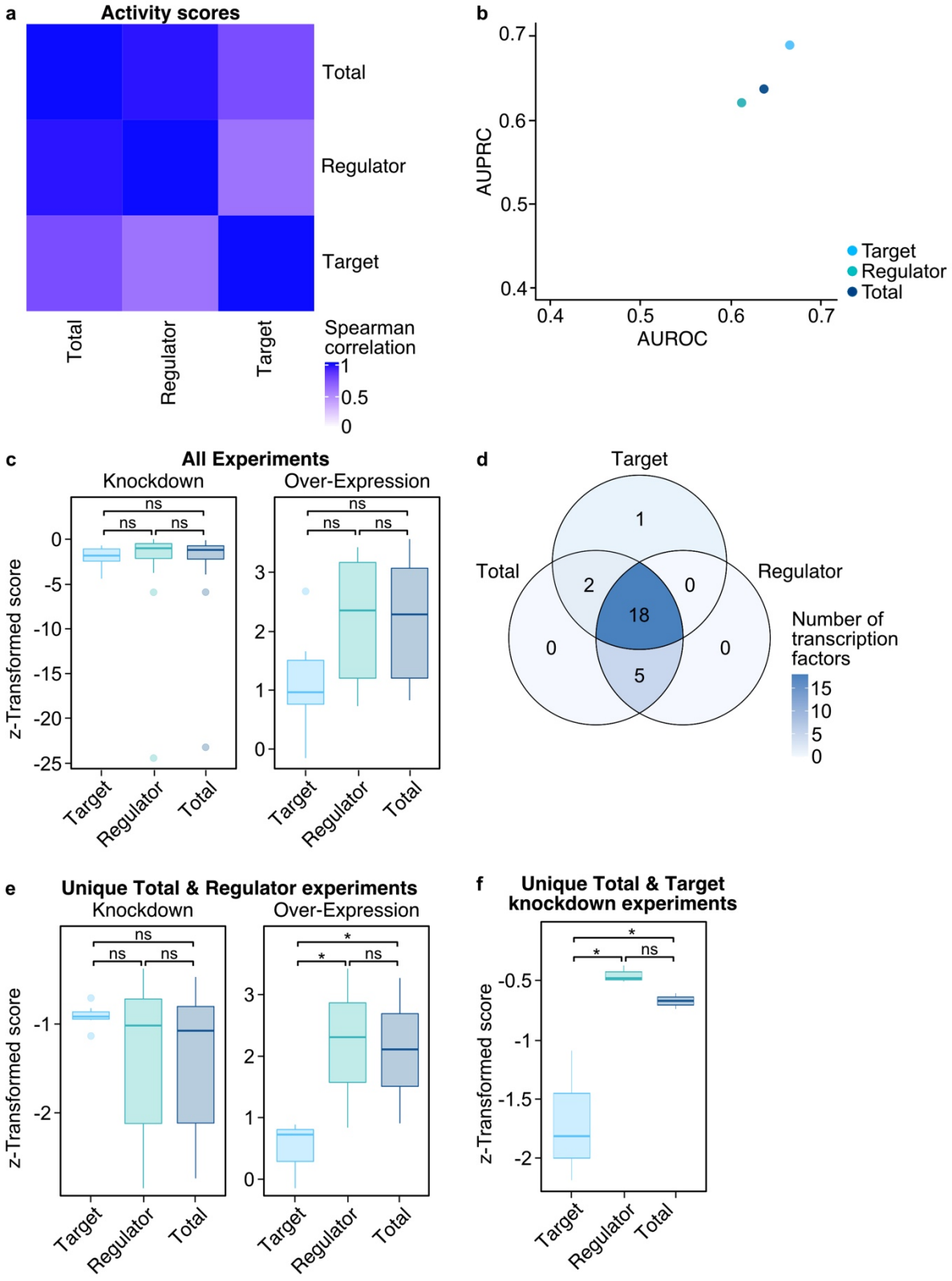


Figure 3: Both transcription factor and target gene expression are important drivers of Priori activity signatures. a. Spearman correlation of transcription factor activity

scores for the Total Activity score as well as the Target and Regulator components. b. Median AUPRC and AUROC values across the 100 down-sampling permutations for each method. c. z-Transformed Total Activity, Target component, and Regulator component scores across all experiments. d. Overlap of perturbed transcription factors identified by each method. e,f. z-Transformed Total Activity, Target component, and Regulator component scores across (e) experiments where the perturbed transcription factor was only identified by the Total Activity score and the Regulator component from (d) and (f) experiments where the perturbed transcription factor was only identified by the Total Activity score and the Target component from (d). ns = not significant, * = $p < 0.05$.

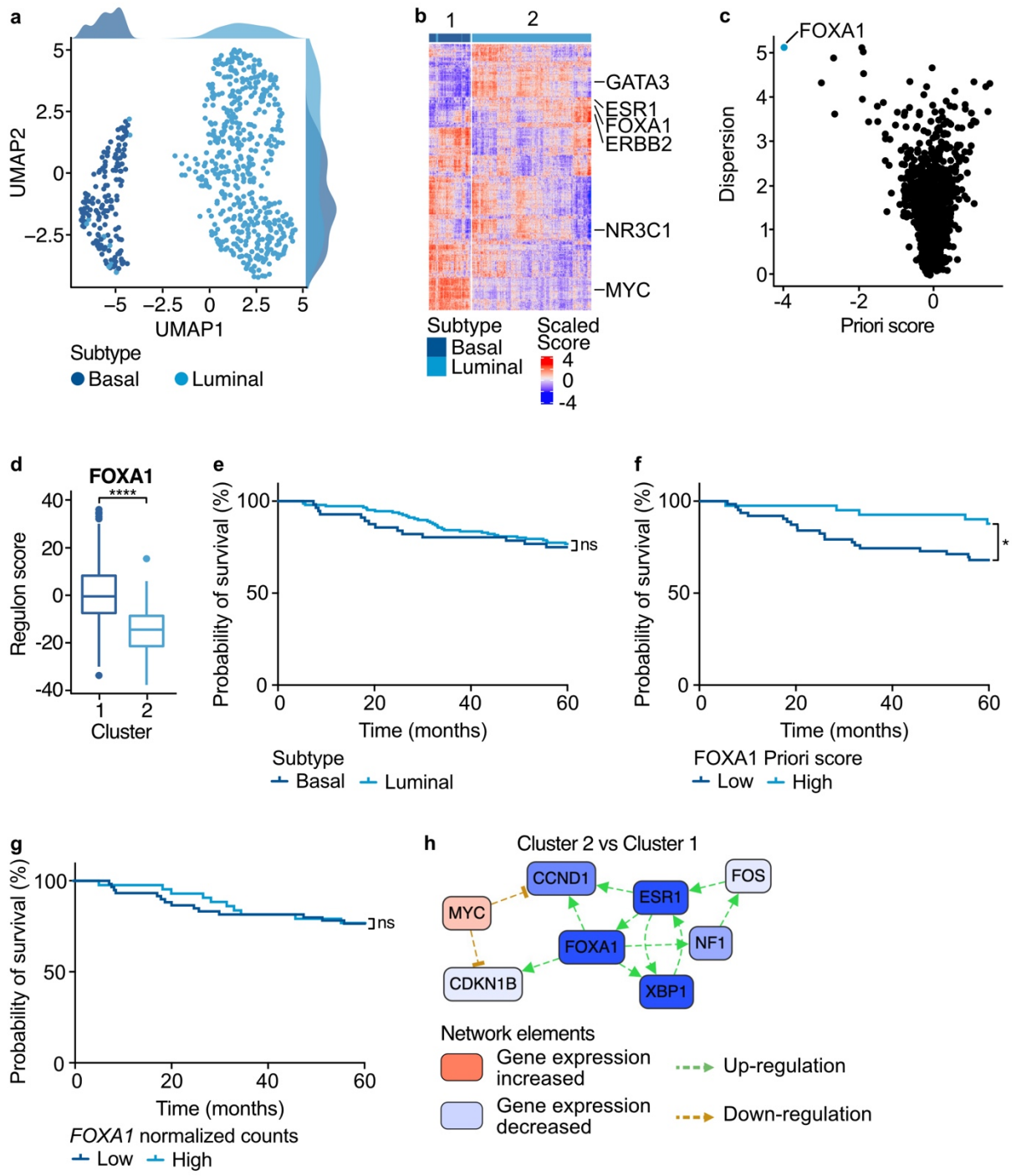


Figure 4: FOXA1 transcription factor activity is a significant determinant of survival for patients with BIDC. a. Priori scores were generated from RNA-seq of 637 patients with BIDC. UMAP dimensional reduction and projection of Priori scores. Dots are colored by the breast cancer molecular subtype. b. Unsupervised hierarchical clustering of Priori scores generated in (a). c. Dispersion, or the $\log_{10}$(variance), of the Priori scores from (a). d. Distribution of Priori scores among patients in clusters 1 and 2 defined in (c). e-g. Kaplan-Meier survival analysis of patients grouped by (e) molecular subtype, (f) FOXA1

Priori scores, or (g) *FOXA1* normalized gene expression counts. Patients among the top 75% of Priori scores or counts were grouped into "High" and those in the bottom 25% were grouped into "Low". Survival was censored at 60 months. h. Differential gene expression network enrichment between clusters defined in (c). Select significantly enriched nodes are shown. ns = not significant, * = $p < 0.05$.

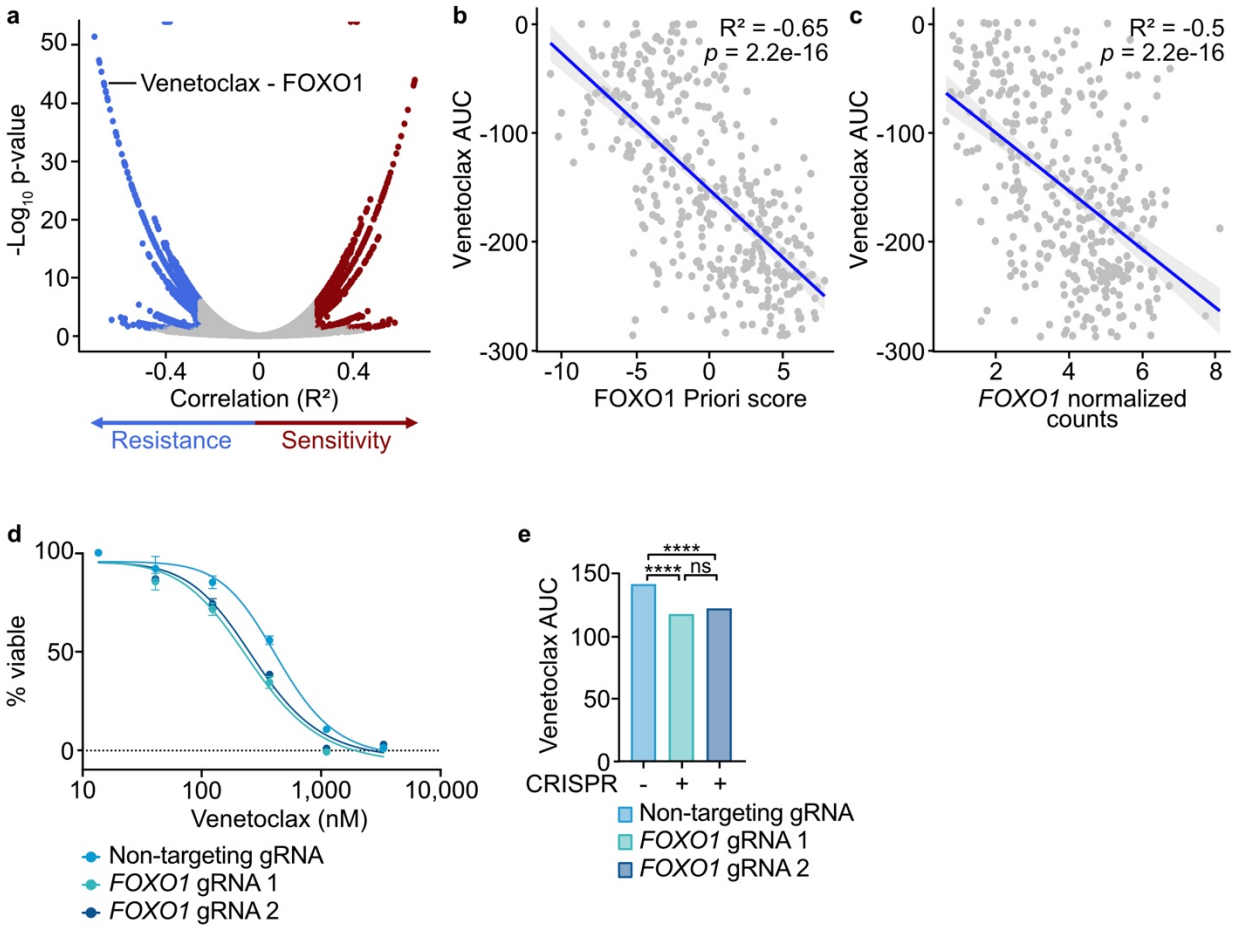
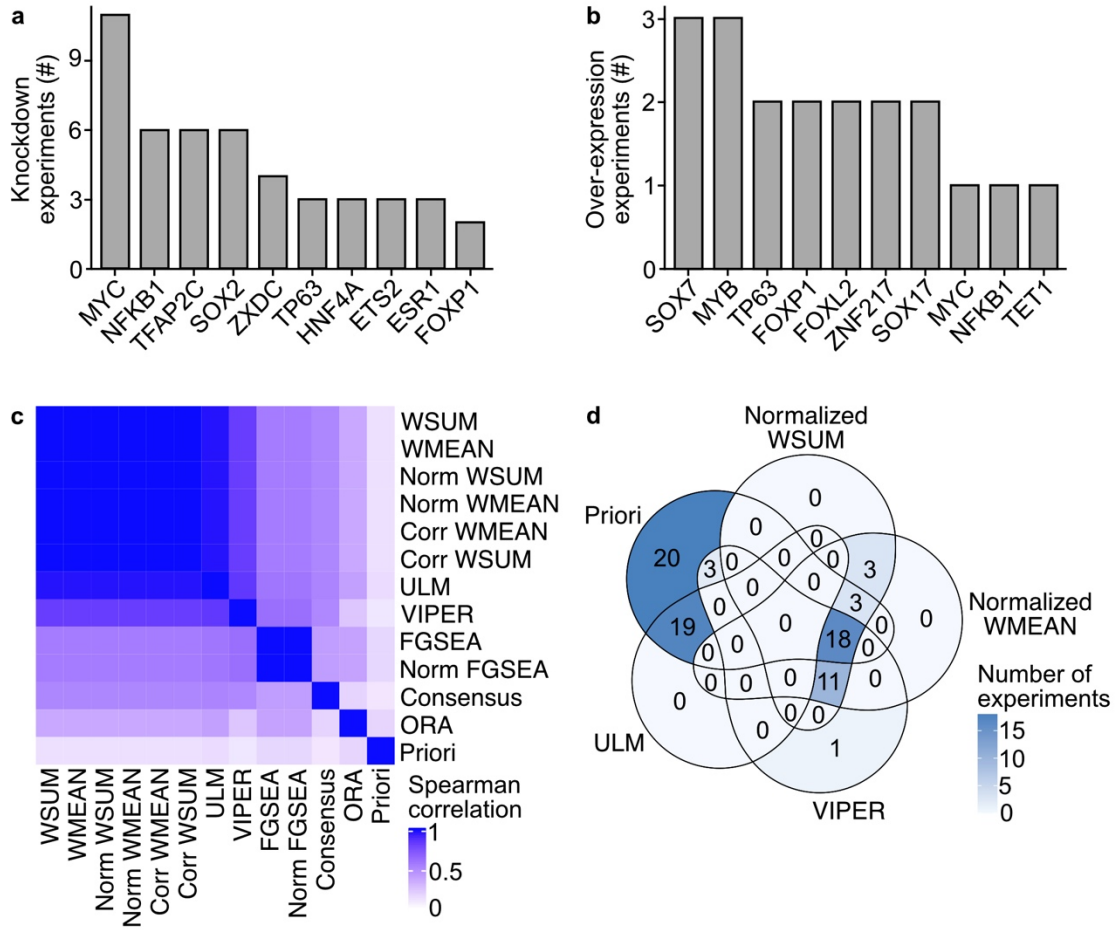
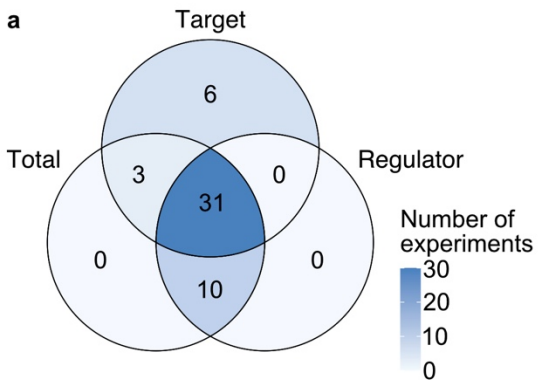


Figure 5: FOXO1 is a critical mediator of response to venetoclax in AML. a. Priori scores generated from RNA-seq of 859 patients with AML. Spearman correlation of Priori scores and *ex vivo* drug response AUC data. b, c. Spearman correlation of venetoclax activity and (b) FOXO1 Priori scores or (c) *FOXO1* normalized counts. D, e. THP-1 cells were transduced with lentiviral particles harboring expression cassettes for hSpCas9 and a non-targeting or *FOXO1* guide RNA. Cells were cultured for 3 days along a 7-point curve with venetoclax. Cell viability was assessed by CellTiter Aqueous colorimetric assay. ns = not significant; * = p < 0.05, ** = p < 0.01, *** = p < 0.001, **** = p < 0.0001.

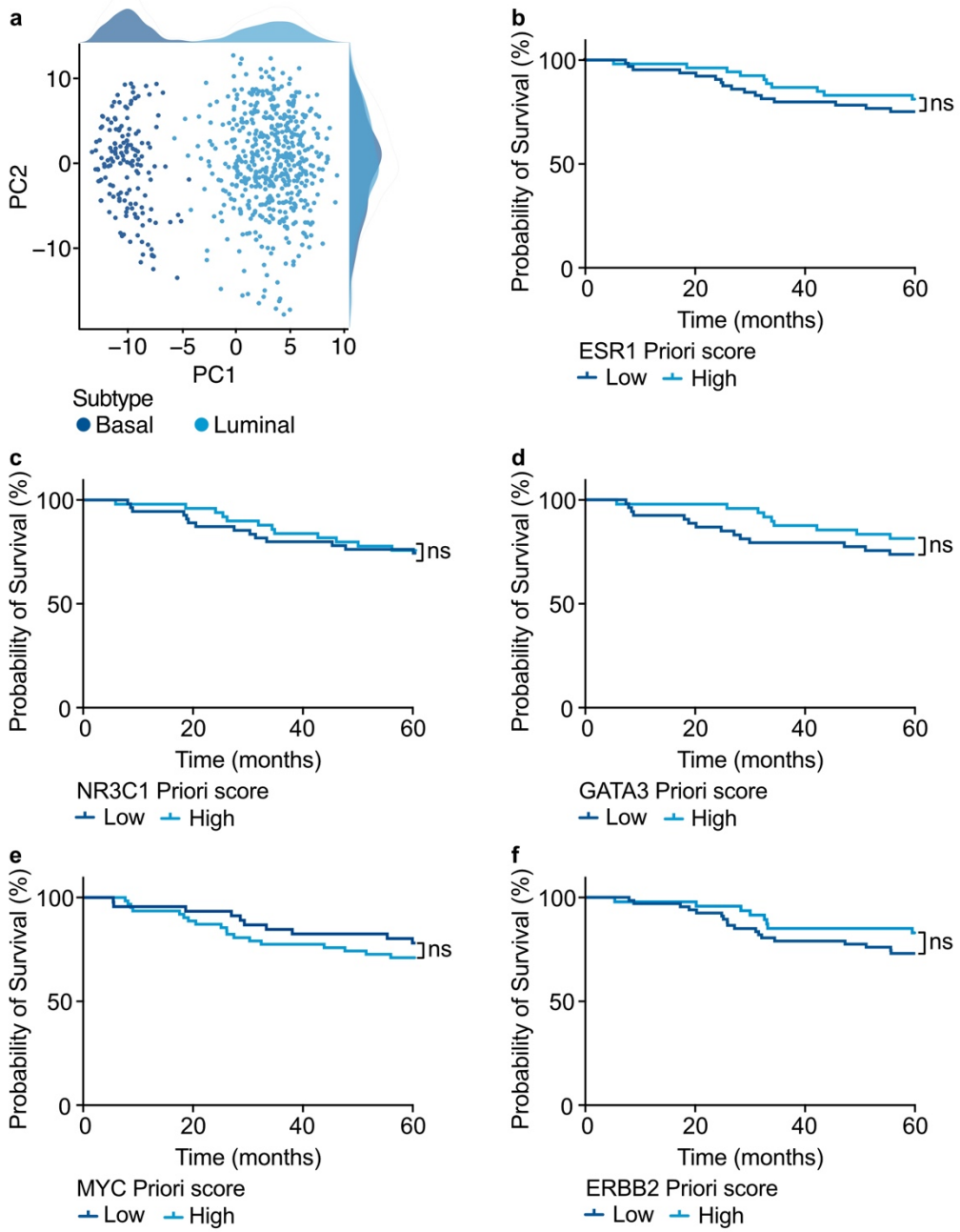
SUPPLEMENTAL FIGURES



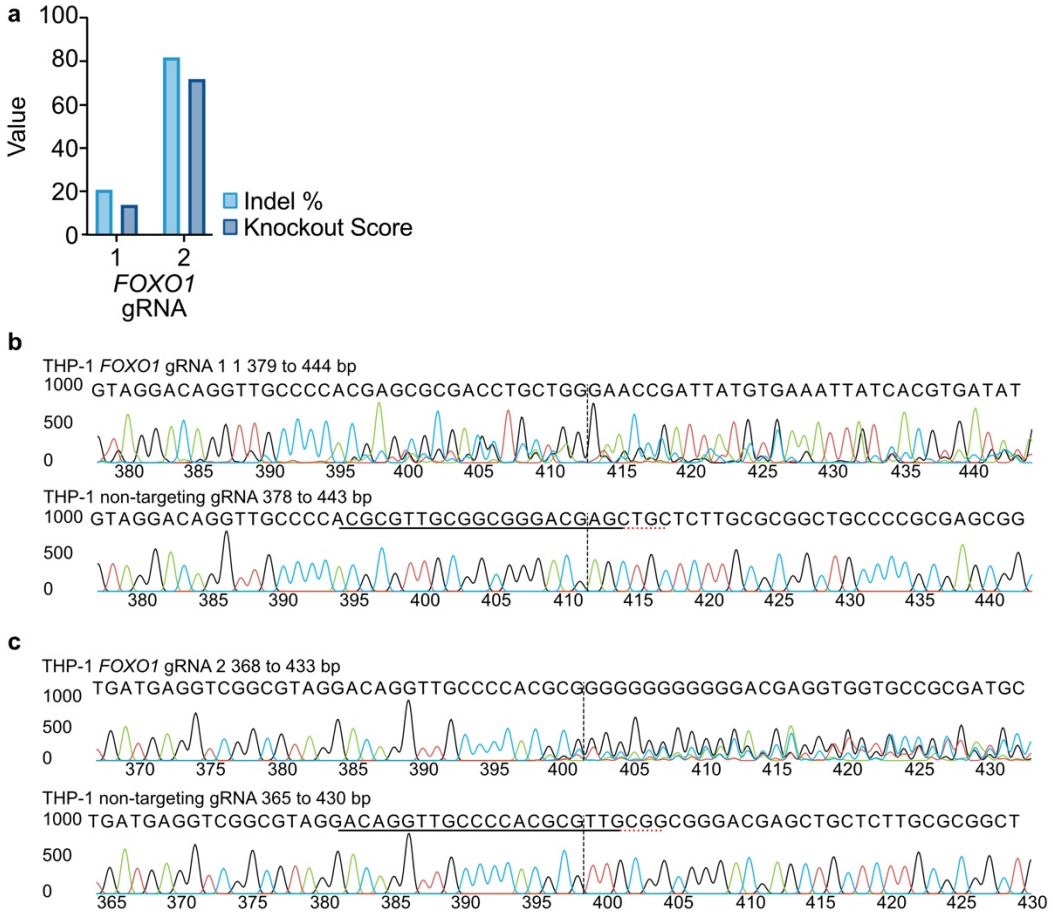
Supplementary Figure 1: Priori detects a perturbed transcription factors not identified by other methods. a, b. Number of experiments that a transcription factor was (a) knocked-down or (b) over-expressed in the single-gene perturbation RNA-seq dataset from Figure 1b. Only the top ten transcription factors are shown. c. Spearman correlation of transcription factor activity score p-values for each method. AUCell, GVSA, MDT, and UDT did not report activity score p-values. d. Overlap of experiments where the perturbed transcription factor was identified by one of the top 5 methods by AUPRC and AUROC from Figure 2b.



Supplementary Figure 2: Both the Target and Regulator components were important contributors to the Total Activity. a. Overlap of experiments where the perturbed transcription factor was identified by the Total Activity score, Target component, and Regulator component.



Supplementary Figure 3: Predicted activity of transcription factors with known roles in BIDC did not predict survival. a. Priori scores were generated from RNA-seq of 637 patients with BIDC. PCA dimensional reduction and projection of Priori . Dots are colored by the breast cancer molecular subtype. b-g. Kaplan-Meier survival analysis of patients grouped by (b) ESR1, (c) NR3C1, (d) GATA3, (e) MYC, or (f) ERBB2 Priori scores. Patients among the top 75% of Priori scores were grouped into "High" and those in the bottom 25% were grouped into "Low". Survival was censored at 60 months.



Supplementary Figure 4: Validation of *FOXO1* knockdown by CRISPR. a. THP-1 cells were transduced with lentiviral particles harboring expression cassettes for hSpCas9 as well as a non-targeting or *FOXO1* gRNA. Cells were evaluated by the percentage of indels and knockout score, which is the proportion of indels that are a frameshift or are greater than 21 bp in length. b, c. Sanger sequencing traces showing edited and control (cells with non-targeting gRNA) sequences in the region around the guide sequence. The horizontal black underlined region represents the guide sequence. The horizontal red underline represents the PAM site. The vertical black dotted line represents the cut site.