NAR Genomics and Bioinformatics, 2023, **5**, 1–18
https://doi.org/10.1093/nargab/lqad106
Advance access publication date: 13 December 2023
Methods Article

A Bayesian noisy logic model for inference of transcription factor activity from single cell and bulk transcriptomic data

Argenis Arriojas [1,2,3], Susan Patalano [3], Jill Macoska [3,*] and Kourosh Zarringhalam [1,3,*]

[1]Department of Mathematics, University of Massachusetts Boston, Boston, MA 02125, USA
[2]Department of Physics, University of Massachusetts Boston, Boston, MA 02125, USA
[3]Center for Personalized Cancer Therapy, University of Massachusetts Boston, Boston, MA 02125, USA

[*]To whom correspondence should be addressed. Tel: +1 617 287 7486; Email: kourosh.zarringhalam@umb.edu
Correspondence may also be addressed to Jill Macoska. Email: Jill.Macoska@umb.edu

Abstract

The advent of high-throughput sequencing has made it possible to measure the expression of genes at relatively low cost. However, direct measurement of regulatory mechanisms, such as transcription factor (TF) activity is still not readily feasible in a high-throughput manner. Consequently, there is a need for computational approaches that can reliably estimate regulator activity from observable gene expression data. In this work, we present a noisy Boolean logic Bayesian model for TF activity inference from differential gene expression data and causal graphs. Our approach provides a flexible framework to incorporate biologically motivated TF–gene regulation logic models. Using simulations and controlled over-expression experiments in cell cultures, we demonstrate that our method can accurately identify TF activity. Moreover, we apply our method to bulk and single cell transcriptomics measurements to investigate transcriptional regulation of fibroblast phenotypic plasticity. Finally, to facilitate usage, we provide user-friendly software packages and a web-interface to query TF activity from user input differential gene expression data: https://umbibio.math.umb.edu/nlbayes/.

Introduction

Gene regulation plays an essential role in many cellular processes, including metabolism, signal transduction, development and cell fate (1,2). At the transcriptional level, regulation of genes is orchestrated by the concerted action between transcription factors (TFs), histone modifiers, and distal *cis*-regulatory elements to finely tune and modulate expression of genes (3). Sequence specific TFs, which have affinity for specific DNA sequences, may bind to *cis*-regulatory elements at the enhancer or promoter region of genes to either activate (upregulate) or repress (downregulate) the expression of genes. Aberration in TF activity and the dysregulation of target genes have been implicated in many pathological states and human disease (4). Activity of TFs can be triggered downstream of signaling events, which in turn may be activated in response to environmental and molecular perturbations (5). Perturbations in TF activity often result in modulation of gene expression. The technological advancements in high-throughput sequencing have made it possible to measure expression of genes at relatively low cost. However, direct measurement of regulatory mechanisms, such as TF protein expression and functional activity in a high-throughput manner is still not readily available. Consequently, there is a need for computational approaches that can identify active regulatory mechanisms from observable gene expression data.

The scientific community has proposed several computational algorithms and biophysical models to study the impact of TF activity on gene expression. Some of these algorithms use statistical and probabilistic approaches to infer TF activity and dynamics directly from gene expression data (6–11), and more recently (12–17), while others rely on biophysical approaches to model expression of genes based on known TF–gene interactions (18,19). Boolean networks and probabilistic extensions have also been used to model gene regulation (9,20–23). In logic models, genes are assumed to be either ON or OFF and Boolean logic (AND, OR, NOR, etc.) is utilized to model combinatorial regulation. For example, Bulashevska and Eils (9) introduced a Bayesian approach to generalize the Boolean logic to incorporate noise and utilized their approach to reconstruct gene regulatory networks in yeast.

Another class of algorithms use prior biological knowledge on biomolecular interactions to link a differential gene expression (DGE) profile to upstream regulators (e.g. TFs) (24–30). The essential ingredients of these type of algorithms are (i) an input DGE profile, (ii) a network of biomolecular interactions or pre-defined gene sets and (iii) an inference algorithm to query the network. The output is a set of candidate regulators, pathways, or biological processes with associated probabilities or significance *P*-values. The DGE profile as obtained from RNA-Seq or microarrays studies is the observable input that quantifies the difference in transcript abundance between two conditions (e.g. healthy versus disease, stimulated versus not stimulated, etc.). The network of biomolecular interactions encapsulates the prior biological knowledge. The inference algorithms typically map the DGE profile to the network to identify drivers (nodes, terms, or paths in the graph) of the observed transcriptional changes. SI Table 2 lists a few representative methods for inference of TF activity and reconstruction of gene regulatory networks and their corresponding descriptions (12,14,15,25,30–34).

Received: June 7, 2023. **Revised:** November 12, 2023. **Editorial Decision:** November 20, 2023. **Accepted:** November 24, 2023

Despite the popularity and success of these methods, several challenges remain to be sufficiently addressed. For example, biophysical models are computationally expensive and are suitable for small scale applications or simulation studies. Boolean logic models, although simple to implement, cannot directly account for noise in gene expression data. On the other hand, probabilistic models for inference of gene regulatory networks typically overlook the context of experiment. Regulatory networks may be noisy or contain interactions that are applicable in a specific context only. Properly modeling any dependencies on the biological context in active regulator inference and enrichment analysis algorithms can lead to more accurate inference of the regulatory mechanisms specific to that context. Utilizing causality and information on mode of regulation (activation versus repression) can also significantly reduce false positive predictions, resulting in more interpretable models. Moreover, biologically motivated TF–gene interaction logic models (e.g. combinatorial effects of activators and repressors on gene expression) must be taken into consideration when inferring transcriptional regulatory programs. To address these challenges, we have developed a Noisy-Logic Bayesian (NLBayes) TF activity inference model that accounts for these factors in a unified manner. Given an input DGE profile, our model incorporates the prior information on causal regulatory interactions and makes posterior adjustments to further account for noise and determine the context-specific posterior network structure and active regulators through a Gibbs sampling procedure.

We evaluate the performance of our model using simulation studies as well as over-expression datasets and demonstrate that our method can accurately identify active transcriptional regulators from gene expression data and causal graphs. We benchmark our algorithm against VIPER, a closely related method that is widely used for identification of regulon activity (15). Both methods are able to identify relevant TFs in the corresponding experiments, with several TFs identified by both methods at the intersection, indicating that the algorithms complement each other. Our method can be used for novel biological discoveries. To illustrate this, we apply our method to investigate transcriptional regulation of fibroblast phenotypic plasticity in response to signaling molecules TGFβ and CXCL12. This study utilizes differential expression profiles from bulk RNA-Seq experiment using prostate cell lines (35), stimulated with TGFβ and CXCL12. Our analysis recovers several TFs, including YAP1 as the top prediction, which has been identified as a driver of myofibroblast differentiation in multiple tissue phenotypes (36–45). Additionally, we present new single cell gene expression data from the same prostate cell lines as well as three additional human prostate fibroblast cell lines. We characterize the cell lines at the transcriptional level and apply our algorithm to identify major transcriptional regulators in each cell line and study the impact of immortalization on transcriptional regulation. Our algorithm provides a general framework and a widely applicable tool to study transcriptional regulators of differential gene expression. To facilitate wider use, we provide R and Python packages, and a web-interface for running inference experiments. Figure 1 summarizes the overall approach.

Materials and methods

Noisy logic-based gene regulation graphical model

As a starting point, we construct a causal graph from the TF–gene interaction network as follows. The causal graph is a triplet $(G,\ E,\ S)$, where G represent the nodes, E represent the edges (pairs of nodes) and S represent signs associated with each edge $(+,\ -)$. Figure 2 shows a graphical representation of the proposed model. The nodes in the graph consist of the following layers:

- Transcript nodes $Y = \{Y_1, \ldots, Y_m\}$: These are the terminal nodes in the network and represent the transcripts. The domain of these nodes is $\mathcal{D}(Y) = \{(-), (0), (+)\}$, representing *downregulated, not regulated and upregulated* respectively. The value of these nodes will be populated from the input gene expression data.
- True states nodes $H = \{H_1, \cdots, H_m\}$: These nodes represent the true unobserved state of the transcript nodes, and are incorporated to account for noise in input data. These nodes have domain $\mathcal{D}(H) = \mathcal{D}(Y)$ and are central in the implementation of noisy logic gates.
- Regulator state nodes $X = \{X_1, \cdots,\ X_n\}$: These nodes represent the activation state of TFs in the network. Here, we use $\mathcal{D}(X) = \{(0), (+)\}$, for no activation or activation respectively.
- TF activity noise nodes $\theta = \{\theta_1, \cdots,\ \theta_n\}$. For each X_i in the network, we assign a node θ_i, representing a continuous random variable with domain $\mathcal{D}(\theta_i) = [0, 1]$. These nodes represent the probability of activation for the corresponding node X_i and are modeled by a beta distribution.
- Mode of regulation $S = \{S_{11}, \ldots, S_{1m}, \ldots, S_{ij}, \ldots, S_{nm}\}$. These nodes represent the mode of action (activation versus repression) between parent TF node X_i and true state node H_j. These nodes have a domain $\mathcal{D}(S) = \{(I), (NA), (A)\}$, representing inhibition (I), non-applicable (NA) and activation (A) respectively. We use one-hot encoding for this variable, i.e. S_{ij} is a vector of size 3 with components $S_{ij}^{(I)}, S_{ij}^{(NA)}, S_{ij}^{(A)} \in \{0, 1\}$.

Transcriptional logic

In our model, we incorporate logic gates as in (9) to explicitly account for combinatorial effects using Boolean logic while accounting for uncertainty. In this work, we consider a combination of noisy OR and NOR gates. We consider two models as follows.

OR model

This model is used to describe the likelihood of downregulation of a gene by a set of TFs. In this model, the presence of one active inhibitor is sufficient to downregulate the gene. The probability mass function is modeled as a Bernoulli trial with probability of success

$$P\left(H_j = (-)\,|\Theta\right) = 1 - \prod_{i=1}^{n} (1 - \theta_i)^{X_i S_{ij}^{(I)}},$$

where $\Theta = \{X, \theta, S\}$ represents all model parameters involving H nodes. Although this model seems like a sensible choice, it assumes that all the genes targeted by a TF should strictly follow its influence. However, target regulation depends on many other factors, and we should expect only a fraction of targets to be effectively regulated by a given TF. To make our model more flexible, we incorporate a hyper-parameter ξ, that allows the likelihood model to tolerate more zero-genes in the evidence data. The OR model above is now

$$p\left(H_j = (-)\,|\Theta\right) = \left[1 - \prod_{i=1}^{n} \left[(1 - \xi)\,(1 - \theta_i)\right]^{X_i S_{ij}^{(I)}}\right]\left(1 - \xi^n\right) + \xi^n q_{-},$$

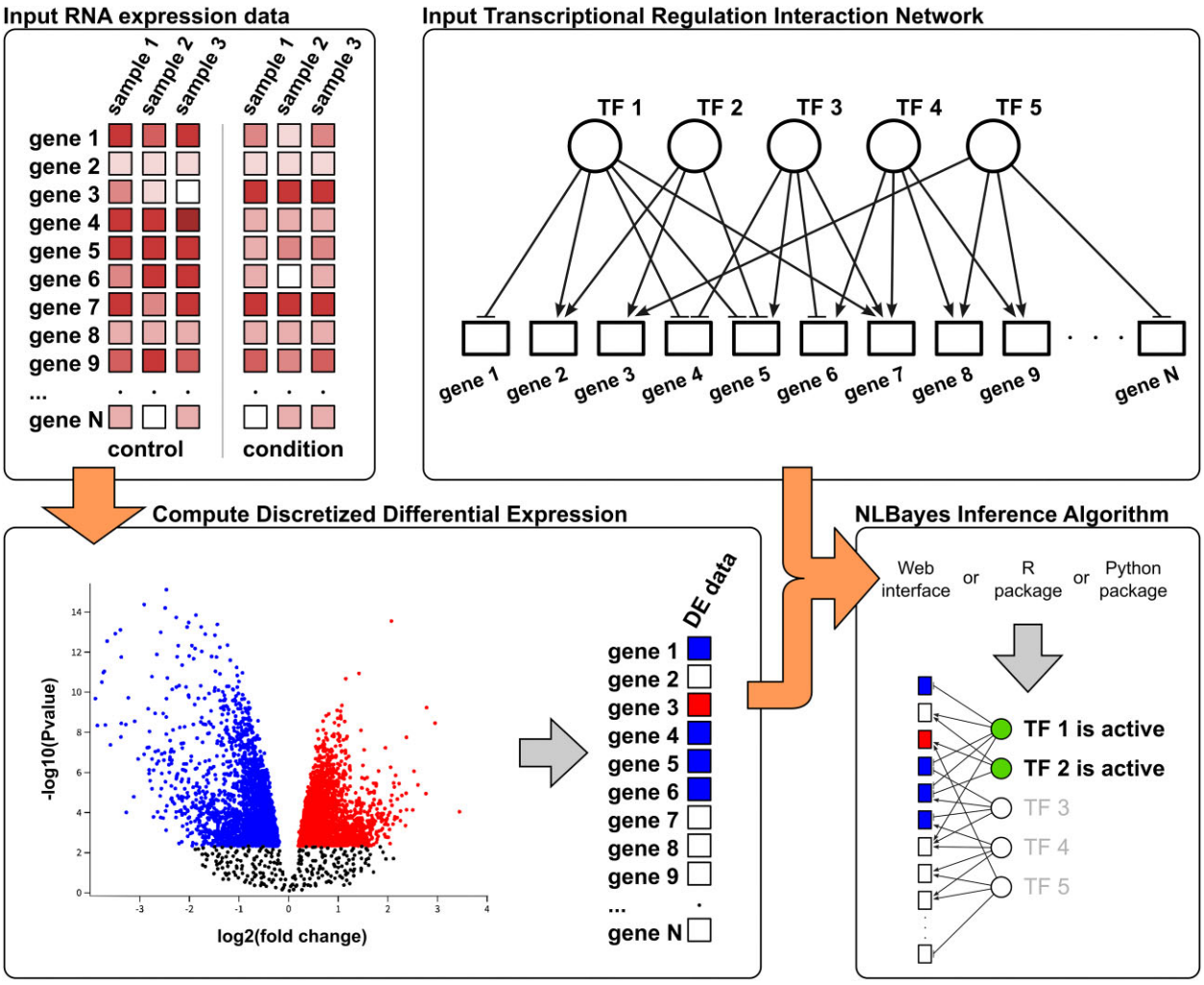

Figure 1. Schematic diagram of the NLBayes inference pipeline. The starting point is gene expression data from two conditions, from which a differential gene expression profile is calculated by discretizing gene values to –1 (down regulated), 0 (not regulated) or 1 (up regulated) using cutoff thresholds on *P*-values and/or foldchange. A TF–gene interaction network is used to build the graphical model. Values for gene nodes are populated from the differential expression profile. NLBayes runs a probabilistic query on the causal network and outputs the posterior distribution of TFs, from which the activation state of the TFs is determined.

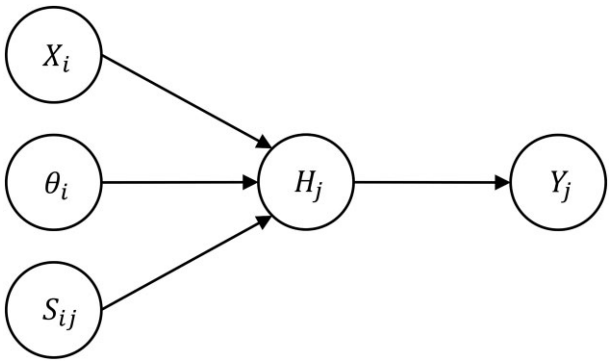

Figure 2. The proposed graphical model for each interaction $i \rightarrow j$.

where q_- represents a prior probability of finding a down-regulated gene in the evidence data. For convenience, let us define $\zeta = (1 - \xi)$, and let ζ^{deg} and $\zeta^{non-deg}$ denote two different values for the ζ parameter that are set depending on whether the current gene has been observed as modulated or not, respectively. The value of the parameter ζ is set close but greater than zero, for non-differentially expressed genes, e.g. $0 < \zeta^{non-deg} \leq 0.1$, while for differentially expressed genes it may be set close or equal to one. Additionally, an extra $(1 - \xi)$ term is now multiplying $(1 - \theta_i)$, effectively increasing the sensitivity of these TF–gene interactions. This has proven beneficial to improve specificity of the inference results. We set $\zeta^{deg} = 0.99$ and $\zeta^{non-deg}$ is set to be proportional to $N_{edges \rightarrow deg}$, i.e. the number of edges in the network that point to genes that are observed as differentially expressed. More specifically, we have used the following relation:

$$\zeta^{non-deg} = \frac{1}{10} \frac{N_{edges \rightarrow deg}}{N_{edges}},$$

where N_{edges} is the total number of edges in the network.

OR-NOR model for gene activation

This model offers a relatively simple way for describing combinatorial effects of both up- and down-regulation within the same interaction network. The rationale is that for a target gene to be activated, at least one of its upstream activators must be activated (OR gate), while at the same time none of

Table 1. Conditional probability of False positive and False negative in the true state given the observed state

$P(Y = \cdot \mid H = \cdot)$	$H = (-)$	$H = (0)$	$H = (+)$
$Y = (-)$	0.945	0.050	0.005
$Y = (0)$	0.050	0.900	0.050
$Y = (+)$	0.005	0.050	0.945

its inhibitors is active (NOR gate). In this case the target gene up-regulation event is modeled as a Bernoulli trial with probability of success

$$P\left(H_j = (+) \mid \Theta\right) = \left[\left[1 - \prod_{i=1}^{n} \left[(1-\xi)(1-\theta_i)\right]^{X_i S_{ij}^{(A)}}\right] \prod_{i=1}^{n} \left[(1-\xi)(1-\theta_i)\right]^{X_i S_{ij}^{(I)}}\right](1-\xi^n) + \xi^n q_+,$$

where q_+ is the prior probability of finding an upregulated gene in the evidence data. Target gene activation state is regarded as a multinomial trial with three possible outcomes: upregulation, downregulation, or not changed. This is congruent with discretized differential expression data and allows building a complete model likelihood. Correspondingly, the complementary outcome likelihood is then represented by a NOR–NOR model.

$$P\left(H_j = (0) \mid \Theta\right) = \left[\prod_{i=1}^{n} \left[(1-\xi)(1-\theta_i)\right]^{X_i S_{ij}^{(A)}} \prod_{i=1}^{n} \left[(1-\xi)(1-\theta_i)\right]^{X_i S_{ij}^{(I)}}\right](1-\xi^n) + \xi^n q_0,$$

Here, q_0 is the prior probability of finding a non-differentially expressed gene in the observed data.

The model likelihood

The posterior probability of model parameters given the observed data is given by:

$$P\left(\Theta \mid Y_j = y_j\right) = \frac{P\left(Y_j = y_j \mid \Theta\right) P\left(\Theta\right)}{P\left(Y_j = y_j\right)} = \frac{\sum_h P\left(Y_j = y_j \mid H_j = h\right) P(H_j = h \mid \vec{X}, \vec{\theta}, \vec{S}_j) P\left(\vec{X}\right) P\left(\vec{\theta}\right) P\left(\vec{S}_j\right)}{\sum_h P\left(Y_j = y_j \mid H_j = h\right) P\left(H_j = h\right)},$$

where $P(Y|H)$ is the conditional probability of the observed expression value Y given the true state. This conditional probability models the true positive and false positive rate in input gene expression data. Table 1 shows the values used in our implementation. These values are estimations based on typical experimental errors. This component of the model may encompass several sources of uncertainty, such as dropped reads during RNA sequencing or type I errors in the statistical analysis made for computation of differential gene expression. This may be treated as a prior probability, representing our belief that we would observe these rates.

Fitting the model

The next task is to find the set of model parameters that maximize the model likelihood. Given the large scale of the parameter space, this problem is intractable analytically. A widely used approach for inferring posterior probability in large scale Bayesian networks is Markov Chain Monte Carlo (MCMC) sampling. In particular, Gibbs sampling is a suitable MCMC method to approximate the posterior distribution of the model parameters given the observed data. In Gibbs sampling, we sequentially sample from each random variable, conditioned on the current state of its Markov blanket. To assess the convergence to the posterior distribution of the model parameters, we run at least three independent sampling chains and periodically compute the Gelman–Rubin statistic for each random variable (46). We stop the process after this diagnostic statistic is less than 1.1 for every random variable in the model. The core of the sampling algorithm has been implemented using C++, and user-friendly R and Python packages developed.

Algorithm

In this section, we present Algorithm 1, which was used throughout this work, along with a detailed description of each of the steps.

Algorithm 1. Sample model's posterior distribution.

1. Preprocess input data.
2. Populate the causal graph with observed evidence.
3. Generate N replicas of the graph and initialize random variables by sampling from their prior distributions.
4. Run the Gibbs sampling procedure:
   for each graph replica **do**
   for k iterations **do**
   for each random variable in the model **do**
   4.1. Sample a new value given the current state of its Markov blanket.
   4.2. Add the new value to this random variable's Markov chain.
   end for.
   end for.
   end for.
5. Compute Gelman-Rubin statistics for the Markov chains and get the maximum value among all random variables.
6. If the maximum Gelman-Rubin statistic is greater than 1.1, go back to step 4
7. Return posterior mean values for every random variable.

- Step 1: Data preprocessing. The input differential gene expression data should contain the computed P-values and fold-change scores for all genes. In this step, we select thresholds for P-values and fold change. The aim is to limit the input data to include only the most significant DEGs, totaling to $\Leftarrow 800$ as a rule of thumb.
- Step 2: Populate the causal graph. Here we assign the observed values (evidence) for the differential expression of each gene to the corresponding Y nodes in the graph. These values will remain fixed during the sampling process.
- Step 3: We create N independent copies of the graph, allowing us to store separate states of N Markov chains to sample. The use of multiple Markov chains provides a way to compare independent results and enables early stopping. Ideally, all N chains should converge to the same posterior distribution for all random variables in the model.
- Step 4: Gibbs sampling. For each random variable we retrieve the current state of its Markov blanket, which is given by the current values of the children nodes, the parent nodes, and the children of the parent nodes. Then, given this Markov blanket, we compute the conditioned probability of the random variable, from which a new sample is drawn. The sampled value is stored and we move the next random variable and repeat the process. After completing the sampling process for all variables, we start over from the first random variable. This process is repeated for each of the N graph replicas created in Step 3. To assess convergence of the posterior distribution, we pause the sampling process after completing k rounds to compute the Gelman-Rubin statistics described in the next Step of the algorithm. The choice of k, the period in which convergence is checked, is arbitrary and has been set to $k = 20$ in our implementation.

- Step 5: Maximum Gelman–Rubin statistic computation. During the Markov chains sampling process, we need to determine whether the sampled distributions for the random variables in the model have converged to their respective posterior distributions. For this purpose, we compare the N Markov chain distributions for every single variable in the model by means of the Gelman–Rubin statistic R, which is a dimensionless score that combines the between chain and within chain variances of the random variables. The statistic approaches $R = 1$ as the sampled distributions converge to a steady state. This score is computed for all random variables and we keep the maximum value obtained.
- Step 6: Convergence assessment. Convergence of the sampling process is called when $\max(R) < 1.1$ for every random variable. If this condition has not been satisfied, we continue the sampling process in Step 4.
- Step 7: At this point the sampling process is complete, and the mean value for each random variable is returned.

Simulations

To assess the impact of noise on model performance, we simulated interaction networks and input differential gene expression data (see section: Results, Simulation studies). First, a random interaction network was generated by selecting a total number of N_X TFs and a total number of N_Y genes in the network. For each TF, its number of target genes was sampled according to a negative binomial distribution. Target genes were assigned randomly using a uniform distribution among all genes in the network. Finally for each interaction edge a sign of regulation was assigned with 0.65 probability of being activation (+) and 0.35 probability for repression (–). Subsequently differential gene expression data was simulated as follows. First, we select a random set of TFs that are assigned as *active*. This set of TFs constitutes the ground truth. For each active TF, we select 10% of its target genes and assign differential expression according to the sign of regulation as predicted by the graph: +1 if target is upregulated by TF, –1 if downregulated. For genes targeted by multiple active TFs, we perform the algebraic sum of all incoming interactions and take the net sign.

The parameters for network simulation were chosen to mimic some characteristics of the biologically derived 'three-tissue network' described below. These parameters include the proportion of TFs to Genes, the node degree distribution for TFs, and the average number of TFs that target the same gene. To assess the robustness of our method against the network topology, we generated several additional networks with different combinations of these parameters (See Supplementary Information File, Supplementary Table S4).

TF–gene interaction networks

The TF–gene interaction network was obtained from (30) in which interaction network were constructed from direct experimental evidence, integrating data (47) from ChIP-Atlas and The Genotype-Tissue Expression (GTEx) databases (48,49). Integration was achieved through a regularized Gaussian Graphical model that softly integrated TF–gene interactions derived from ChIP-Seq data into gene expression derived from tissues, resulting in 15 tissue specific TF–gene interaction networks as well as a 'merged network' obtained by overlapping tissue-specific networks. In the present work we use the merged network containing interactions that are consistent in at least three tissue types, resulting in 338680 TF–gene interactions from 750 TF molecules (see section: Results, Overexpression datasets).

For the benchmark experiments (see section: Results, Benchmarks), we used the Human Breast Carcinoma context specific network termed 'regulonbrca', that has been constructed with ARACNe and made available by Lachmann et.al. (33). The regulon object has been filtered to retain only transcription factor molecules, similar to the approach followed by the authors of VIPER (15): DNA-binding transcription factors (GO:0003700), and DNA-binding molecules (GO:0003677) annotated as 'transcription regulator activity' (GO:0140110) or 'regulation of DNA-templated transcription' (GO:0006355). This results in a network with 1693 TF molecules and 200 336 TF–gene interactions. For use in the NLBayes inference, we further filtered out interaction edges with likelihood <0.5, resulting in 73 639 TF–gene interactions.

Differential gene expression data

DNA microarray-based gene expression profiles for overexpression studies in human primary mammary epithelial cell cultures (50), were obtained from the GEO repository (GSE3151) by using the GEO2R tool for sample selection (51). Differential expression was computed using the limma R package (52). We limited the number of differentially expressed genes by applying cutoff thresholds for the adjusted P-values ($P \leq 0.01$) and log2-foldchange ($fc \geq 1$) for E2F3 and MYC datasets. For the RAS dataset these cutoff values produced 2226 differentially expressed genes. To further limit the number of DEGs in this experiment, we increased the $\log_2$-foldchange cutoff threshold to 2.

Additionally, we used data from a study on fibroblast-to-myofibroblast phenotypic conversion in response to profibrotic signaling molecules TGFβ and CXCL12 (35,53,54) (see section: Results, Fibroblast phenotypic plasticity). Differential expression for TGFβ and CXCL12 treated fibroblasts were generated using the R package edgeR (55).

Single cell RNA sequencing

Following Trypsin digestion, cells were collected in a 50mL conical tube and washed and resuspended in PBS to a final concentration of ∼700 cells/ul. After gentle resuspension, 2.3 ul (∼1610 cells) per sample was combined with nuclease free water and master mix per 10× recommendations for a targeted recovery of ∼1000 cells per sample.

The samples were loaded onto a 10× Chromium Chip A (PN230027, deprecated) (N1 and SFT1) or Chip G (PN 2000177) (pHPF and iHPF) and run through the 10X Chromium controller instrument and manufacturer protocols for RNA recovery and library preparation. Briefly, cells were partitioned into individual lipid droplets, lysed to release RNA and tagged with UMIs (unique molecular identifiers) for cell of origin. mRNA was isolated with dT oligo beads then reverse transcribed to DNA, ligated with Illumina-compatible sequencing adapters with multiplex capable barcodes (PN-120262 for N1/SFT1 (deprecated), PN-1000213 for pHPF/iHPF), for sample of origin, and PCR amplified for 13 cycles. N1 and SFT1 samples were prepared with Chromium v2 Library and gel bead kit (PN-120267,

deprecated) while pHPF and iHPF were prepared with v3 (PN 1000128).

Following library preparation, samples were assessed by Agilent 2100 Bioanalyzer using High Sensitivity chips and reagents (PN 5067–4626) to confirm a normal size distribution to minimize bias, quantified by qPCR with Illumina adapter compatible primers and Sybr Green (Kapa ROX Low Universal Library Quant kit PN KK4873) and molarity calculated by size-correcting to the bioanalyzer average size. N1 and SFT1 samples were pooled together while pHPF and iHPF were pooled together separately.

The samples were sequenced on a Hiseq 2500 in Rapid Run mode in the CPCT Genomics Core, using paired-end on board clustering (PN PE-402-4002) and sequencing by synthesis (SBS) reagents (PN FC-402-4021). Twenty-eight bases were sequenced in read 1 to capture the UMIs, 8 bases for the single indexes, and 91 bases in Read 2 to capture transcripts, yielding ∼100 million total sequencing reads per sample.

RNA-Seq alignment

Reads were aligned to human genome version GRCh38 using the 10× cellranger v4.0.0 pipeline (cellranger mkref, and cellranger count) using default parameters. Over 93% of reads were mapped to the genome for all 4 datasets. Mapping rates were 93.6%, 93.3%, 96.3% and 96.3% for N1, SFT1, iHPF and pHPF, respectively. To improve the quality of the data, the resulting count matrices were reanalyzed to force the number of cells accepted, to those with highest UMI counts.

Single cell data analysis

Count data from single cell alignment were processed using the Seurat R package (56,57). Seurat objects were created from the filtered matrices resulting from the alignment step. Low quality cells were filtered out, by removing those with large mitochondria contamination and cells with either too few or too many unique genes or total RNA count. Cell cycle scores were assigned to each cell with the method CellCycleScoring using default parameters and cell cycle genes provided by Seurat package. Datasets N1, SFT1, pHPF and iHPF were combined by using Seurat's merge function and normalized with SCTransform using 3000 variable features and no centering. Batch effect was removed by considering the number of genes, RNA counts and mitochondrial RNA contamination as unwanted sources of variation. Dimensionality reduction was performed through principal component analysis (PCA) and Uniform Manifold Approximation & Projection (UMAP) as implemented in Seurat package, with functions RunPCA (30 PCs) and RunUMAP respectively. Cells were then filtered to work with G1 cells only. Differential expression was computed with respect to pHPF, by using the FindMarkers method. All Seurat methods were used with default parameters, unless otherwise stated.

Prostate tissue single cell data was retrieved from (58). This corresponds to FACS sorted cells, containing fibroblasts, smooth muscle, endothelial and epithelial cells. Here we ignore FACS sorting labels as we run our own classification process. This dataset was preprocessed using Seurat and its SCTransform pipeline, same as with cell lines data. Cells were clustered by using methods FindNeighbors and FindClusters. Each cluster was classified by looking into cell type markers taken from (58), to assign labels: fibroblast, smooth muscle, endothelia, basal epithelia, luminal epithelia, and other epithelia 1 and 2. Cells in G1 cell cycle phase were retained for downstream analysis. Table 2 shows the markers used for the tissue cells classification.

Table 2. Gene markers used for classification of single cells in the prostate tissue single cell dataset by Henry *et al.* (58)

Tissue type	Gene markers
Basal Epithelia	KRT14, DST, KRT15, KRT5, RGCC
Luminal Epithelia	MSMB, KLK3, ACPP, PLA2G2A, KLK2
Other Epithelia 1	SCGB3A1, LCN2, PIGR, WFDC2, FCGBP
Other Epithelia 2	KRT13, APOBEC3A, CSTB, LYPD3, SERPINB1
Fibroblast	APOD, FBLN1, PTGDS, CFD, DCN
Smooth muscle	TPM2, ACTA2, RGS5, MT1A, MYH11

These are the same markers used in that study.

Cell lines data and tissue data were combined by using Seurat's integration pipeline with SCTransform, using 3000 features, k.anchor = 6, and reduction = 'rpca'.

Results

To test our approach, we performed a series of experiments including simulations studies, benchmarks against an alternative approach, and inference of TF activity on novel datasets.

The core of the algorithm uses an OR-NOR transcriptional regulation logic to predict TF activity. In this logic two conditions must be satisfied for gene activation: 1) at least one activator is targeting the gene and 2) no inhibitor is targeting the gene. On the other hand, for down regulation of genes we use a simple OR transcriptional regulation logic model: at least one inhibitor must target the genes.

The algorithm outputs posterior probabilities for each TF activation state. The prior probability for TF activation is set to a small value ($P_0 = 0.01$), and as such we consider a TF with posterior probability $P \geq 0.2$ as potentially relevant. We then define three thresholds to classify inferred active TFs: High-confidence ($P \geq 0.8$) Mid-confidence ($P \geq 0.5$), and Low-confidence ($P \geq 0.2$).

Simulation studies

We performed several simulation studies to assess the ability of our algorithm in recovering active transcriptional regulators from gene expression data and to test the robustness of the inference process to noise in gene expression data and the causal graph of TF–gene interactions. For this analysis we generated a random interaction network consisting of 250 TFs and 5000 downstream target genes. The downstream targets of each TF were picked at random from a binomial distribution, resulting in ∼30000 interactions (edges) in the network. The edges of the network were randomly set as activation (65%) and inhibition (35%).

We randomly selected 10 TFs and assigned them as active (the ground truth) and simulated downstream differential gene expression data by assigning + 1 or -1 to 10% of the target genes according to the causal graph. For genes targeted by multiple active TFs, we calculated the algebraic sum of all incoming interactions and took the net sign. This produced an average of 120 differentially expressed genes. Each experiment was repeated 20 times.

Impact of data randomization

For this analysis, we randomized a fraction of the input data (0, 0.25, 0.5, 0.75, 1.00) by randomly toggling the values. At 0% the data is not randomized, while at 100% the input data

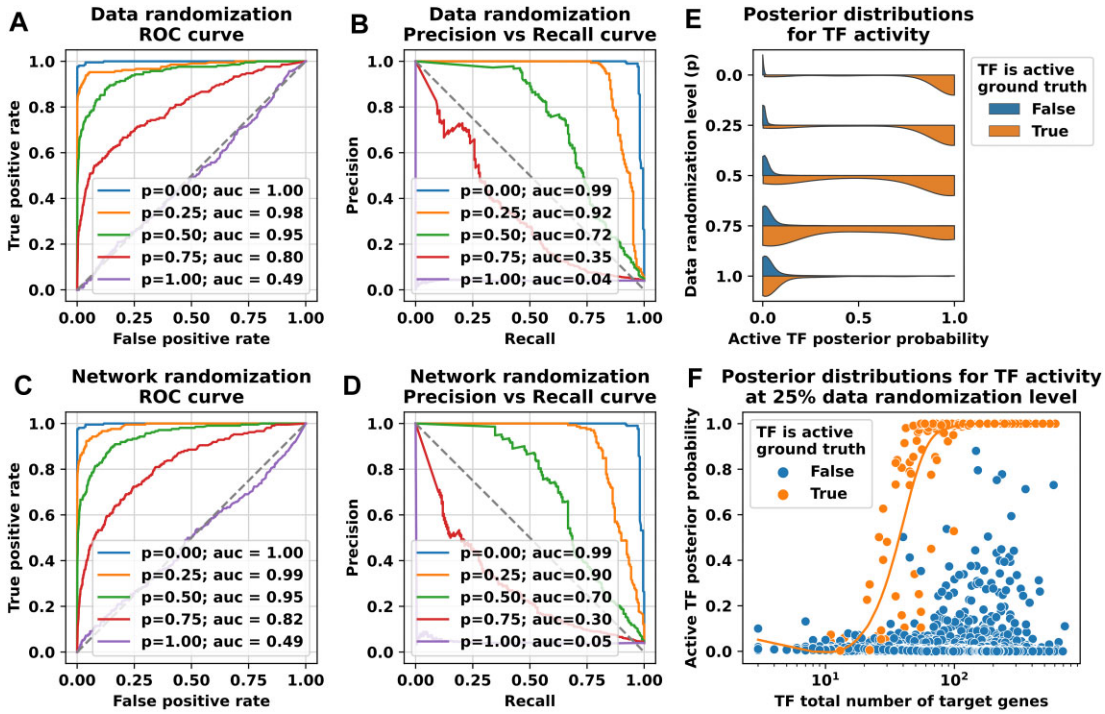

Figure 3. Performance evaluation of the model before and after randomization of data. ROC curves and Precision vs Recall curves are shown for randomization simulation in input gene expression **(A, B)** and TF–gene interaction network **(C, D)**. p indicates the corresponding fraction of randomization used. AUC scores are displayed in the legend. **(E)** The posterior distributions for TF activity is shown. The colors indicate the ground truth, i.e. whether the TF was set to active in the simulation. **(F)** The impact of the number of target genes on the inference results for the case of 25% randomization of data.

is completely random. The inference procedure was run on each input data set. Figure 3A and B shows the ROC and precision-recall curves for each randomization experiment respectively. Figure 3E illustrates the sampled posterior distributions for active TFs at different randomization levels. As expected, at 100% the model fails to recover active TFs. However, up to 75% randomization the algorithm is still able to recover active TFs, with significant increase in accuracy at lower levels of noise.

To assess the impact of the number of target genes regulated by a TF on the posterior probability of the TF's activity, we ran multiple simulations and plotted the posterior probability vs the number of genes regulated by TFs, color coded by TF activity (Figure 3F). For these simulations, we used noisy data at 25% randomization. We observe that for active TFs with less than 30 target genes, the posterior probability is low. This is expected as in these simulations, only 10% of target genes are set as differentially expressed, yielding an average of only three target genes that are modulated. This information is too small to shift the posterior probability. However, as the total number of target genes increase, we observe a threshold effect, where posterior probabilities stabilize, and the total number of targets does not have an impact on the inference. These experiments were repeated for several other randomly generated networks with consistent results (SI File Supplementary Table S4).

Impact of network randomization

For this analysis, we select a fraction (0, 0.25, 0.5, 0.75, 1.00) of the edges in the network and randomly reassign them to different target genes. The inference was run using unperturbed input gene expression data and randomized networks. Figure 3C and D shows the ROC and precision-recall curves for each randomization experiment. A similar picture as in data randomization emerges.

These results demonstrate the robustness of our algorithm at 25% randomization level in both the gene expression data as well as the input network, while still retaining some prediction power at the 50% and 75% randomization levels.

Over-expression datasets

To test the ability of our algorithm in recovering known active regulators, we used three publicly available over-expression datasets (GSE3151), performed on human primary mammary epithelial cell cultures, each generated by over-expression of an oncogene: E2F3, c-Myc and H-Ras (50). For these inference experiments, we utilized a TF–gene interaction network generated by Farahmand *et al.* (30). The network was generated using several high-throughput datasets, and a Gaussian Graphical Model. We chose to use this network as it showed consistently high predictive power across several datasets (30). In each experiment, differentially expressed genes were determined compared to the control sample and inputted into the TF activity inference algorithm. Tables 3 and 4 summarize the inference results. Inferred active regulators are split into three categories based on posterior probability p of inferred activity: High-confidence ($P \geq 0.8$) Mid-confidence ($P \geq 0.5$), and Low-confidence ($P \geq 0.2$). Percentage of differentially expressed genes targeted (explained) by at least one inferred active regulator is also presented in the table.

For the E2F3 expression data, the E2F1 is returned as the top regulator. E2F1 and E2F3 have a similar function

Table 3. Predicted active TFs from 3 over-expression experiments: E2F, MYC and RAS

E2F3			c-MYC			H-RAS		
Name	Enrichment	Inference	Name	Enrichment	Inference	Name	Enrichment	Inference
E2F1	3.7E-06	1.00	TCF3	4.9E-01	1.00	E2F3	6.1E-01	1.00
FLI1	2.7E-02	1.00	HDAC2	1.0E + 00	0.90	PPARG	4.0E-01	1.00
FOXA2	7.0E-01	1.00	TP53	1.2E-02	0.85	TEAD4	4.9E-01	1.00
HOXA4	7.0E-01	1.00	–	–	–	KLF5	2.2E-02	1.00
MXI1	2.6E-01	1.00	–	–	–	FOSL1	2.4E-06	1.00
BCOR	2.2E-02	1.00	–	–	–	RELA	1.4E-01	1.00
AR	1.4E-02	1.00	–	–	–	ASCL1	1.0E + 00	0.99
FOXF1	7.9E-01	1.00	–	–	–	STAT3	6.1E-01	0.99
TCF3	8.6E-01	1.00	–	–	–	PRKDC	1.0E + 00	0.97
MBD3	1.0E + 00	0.99	–	–	–	JUND	1.4E-01	0.97
SOX2	7.3E-01	0.97	–	–	–	NFKB2	1.2E-02	0.96
–	–	–	–	–	–	SOX2	5.3E-01	0.96
–	–	–	–	–	–	CTCFL	1.0E + 00	0.94
–	–	–	–	–	–	GATA3	9.0E-01	0.92
–	–	–	–	–	–	KDM6B	3.2E-03	0.86
HNF1B	1.0E + 00	0.77	SIX2	1.0E + 00	0.77	MYC	2.1E-01	0.75
FOXM1	1.7E-02	0.76	ETV1	1.0E + 00	0.73	NME2	1.0E + 00	0.66
NR5A2	9.2E-01	0.72	KLF11	1.0E + 00	0.54	ESRRA	1.0E + 00	0.65
HNF4A	1.0E + 00	0.71	ILF3	2.9E-01	0.51	SREBF1	1.8E-01	0.64
POU5F1	7.0E-01	0.69	–	–	–	MBD2	9.7E-01	0.64
GATA3	1.5E-01	0.68	–	–	–	–	–	–
TCF21	1.0E + 00	0.50	–	–	–	–	–	–
PGR	8.6E-01	0.39	PAX5	1.0E + 00	0.49	PBX4	1.0E + 00	0.41
TEAD1	1.0E + 00	0.39	TFAP4	2.0E-01	0.49	HOXB13	1.0E + 00	0.37
GATA6	1.0E + 00	0.39	ZC3H8	5.0E-01	0.45	FOXF1	3.6E-03	0.36
VEZF1	1.0E + 00	0.30	PPARG	7.6E-01	0.44	KLF6	1.8E-02	0.34
EGR1	7.0E-01	0.29	PDX1	6.9E-01	0.42	BCOR	9.7E-01	0.33
NKX2-1	1.0E + 00	0.29	TEAD4	2.9E-01	0.37	HNF1B	1.0E + 00	0.32
HOXA9	4.3E-01	0.26	BATF	7.3E-01	0.37	SRF	9.0E-01	0.31
HOXC5	1.0E + 00	0.26	THAP11	1.0E + 00	0.36	GATA4	1.0E + 00	0.27
–	–	–	SMARCC2	1.0E + 00	0.33	CEBPB	4.1E-02	0.24
–	–	–	NCOA1	1.0E + 00	0.33	MAFF	3.6E-03	0.22
–	–	–	TRIM28	2.9E-01	0.24	–	–	–
–	–	–	MYB	1.0E + 00	0.21	–	–	–

The panel lists the inferred regulators, split into three categories, high-confidence (top), mid-confidence (middle) and low-confidence (bottom). For each TF, enrichment scores are computed from the number of target genes that are differentially expressed.

Table 4. Summary of over-expression experiments

E2F3			c-MYC			H-RAS		
max *P*-value = 0.01 min abs(log2 fold-change) = 1.00			max *P*-value = 0.01 min abs(log_2 fold-change) = 1.00			max *P*-value = 0.01 min abs(log_2 fold-change) = 2.00		
DEG	Increased	Decreased	DEG	Increased	Decreased	DEG	Increased	Decreased
418	321	97	127	75	52	409	163	246
Confidence level	DEG explained		Confidence level		DEG explained	Confidence level	DEG explained	
0.80	53%		0.80		32%	0.80	58%	
0.50	61%		0.50		49%	0.50	64%	
0.20	65%		0.20		70%	0.20	71%	

Top: Thresholds for *P*-values and log-foldchange for each experiment are shown, and their respective total number of Differentially Expressed Genes (DEGs) and the number of genes with increased and decreased RNA expression. For the RAS over-expression experiment, a higher log-foldchange threshold was used to limit the number of DEGs for inference. Bottom: proportion of explained DEGs by at least one inferred regulator at the indicated confidence level.

in control of the cell cycle and are similarly implicated in cancer (59). E2F3 also regulates expression of FLI1, an ETS domain transcription factor and proto-oncogene (60), as well as the FOXA2 transcription factor that promotes aggressive prostate cancer (61) and HOXA4, a transcription factor important for embryonic development but often over-expressed in human colorectal (62) and ovarian (63) cancers.

The MYC and RAS oncogenic proteins transcriptionally activate multiple genes associated with tumor progression and prognosis. MYC transcriptionally activates TCF3, HDAC2 and TP53. The TCF3 transcription factor was recently found to promote gastric (64) and endometrial (65) cancers, among others. HDAC2 has been reported to promote metastasis in pancreatic (66) and breast (67) cancers. MYC also regulates transcription of the well-known TP53 tumor suppressor gene.

Table 5. Predicted active TFs upon TGFβ (left) and CXCL12 induction (right)

	TGFβ					CXCL12			
	Top DE genes					Top DE genes			
	200	400	600	800		200	400	600	800
Max *P*-value	4.7E-133	2.6E-95	1.6E-71	8.0E-53	Max *P*-value	6.0E-132	4.6E-94	1.2E-72	4.1E-58
YAP1	**1.00**	**1.00**	**1.00**	**1.00**	YAP1	**1.00**	**1.00**	**1.00**	**1.00**
RBPJ	**1.00**	**1.00**	**0.53**	**1.00**	BCLAF1	**1.00**	**1.00**	**1.00**	**1.00**
KMT2C	**1.00**	**1.00**	**0.96**	**0.61**	BPTF	**1.00**	**1.00**	**0.82**	0.04
ELF1	**1.00**	**1.00**	0.07	**0.75**	RBPJ	**0.99**	**1.00**	**1.00**	**1.00**
STAT1	**0.96**	**0.96**	**1.00**	**0.80**	KMT2C	**0.98**	**0.91**	0.07	**0.62**
BPTF	**0.92**	0.02	**1.00**	**0.95**	GABPA	**0.98**	**0.28**	**1.00**	**1.00**
HIF1A	**0.90**	**1.00**	0.05	0.12	STAT1	**0.96**	**0.99**	**1.00**	**0.82**
BRCA1	**0.63**				RB1	**0.91**	**0.95**	**0.99**	**1.00**
YY2	**0.50**				LIN9	**0.89**	**0.57**	**0.83**	**0.99**
JMJD1C	0.08	**0.90**			PRDM1	**0.33**	0.10	0.02	
VEZF1	0.06	**0.89**	0.09	**0.22**	CBFB		**1.00**	**0.78**	0.17
ERG		**0.84**			AHR	0.01	**0.68**		
TCF12		**0.82**	**0.86**	**0.91**	SP4		**0.45**		
ZNF12		**0.57**	**0.90**	**0.25**	ATF2		**0.37**	0.02	**1.00**
LIN9	0.15	**0.33**	**0.97**	**0.90**	MYBL2	0.02	**0.30**	0.01	
BACH1			**0.98**	**0.85**	BACH1		**0.27**	**0.24**	**0.45**
RB1	0.02		**0.97**	**1.00**	ELF1	0.05	0.13	**0.94**	**0.38**
GABPA		0.09	**0.96**	**1.00**	CHD1			**0.40**	0.10
BCLAF1		0.15	**0.96**	**0.97**	ELK4		0.01	0.10	**0.74**
TCF4			**0.93**		TCF4				**0.63**
MEF2A	0.01	0.01	**0.48**	**0.38**	PIAS1			0.04	**0.59**
SETX		0.02	0.03	**0.94**	SMAD4		0.01	0.17	**0.47**
ATF2			0.16	**0.34**	SETX			0.02	**0.29**
PIAS1			0.05	**0.25**					
IRF1	0.04	0.04		**0.20**					

The top row shows total number of DEGs using four different *P*-value cutoff thresholds (Max *P*-value). Bottom panel lists the inferred regulators. Posterior probabilities greater than or equal to 0.20 are highlighted in bold font face, and those equal to zero are left blank.

RAS proteins activate transcription of E2F3 itself, as well as PPARG and TEAD4. Modulation of PPARG activity has been intensively examined as an anti-cancer therapeutic target (68), and TEAD4 which is known to modulate different cellular processes in cancer via its transcriptional output (69).

Taken together, these results demonstrate that the algorithm can accurately detect modulated transcriptional signals from DNA binding proteins.

Benchmarks

We compared the performance of our algorithm against VIPER (15), a widely used method for inference of regulon activity from input gene expression data and causal graphs. For this benchmark, we used the Human Breast Carcinoma context specific network from ARACNe interactome, named 'regulonbrca' (33). This network is appropriate for the overexpression datasets as they used human breast cell cultures and allows a fair comparison. Figure 3 summarizes the overlap between the two algorithms. For all three experiments, a fisher exact test reveals positive correlation between NLBayes and VIPER with *P*-values 0.034, 0.001 and 0.09 for the E2F3, MYC and H-RAS overexpression experiments respectively. Overall, there is good agreement between the methods as well as regulators recovered only by one algorithm, demonstrating the viability of both methods in recovering modulated TFs. Each algorithm predicts TF activity that is not shared by the other algorithm. While individually predicted TFs by each method may be important and can provide useful clues to the underlying biology, there is higher confidence in predictions at the intersection of both methods.

As an additional benchmark, we have compared the inference results by NLBayes and VIPER on the MYC overexpression experiment, with two other methods for TF activity inference: Gene Set Enrichment Analysis (GSEA) (26,70) and Univariate Linear Model (ULM) (70). To run these two methods, we used the *DecoupleR* Python package (70), to which we provided the log-fold changes of the corresponding overexpression experiment. Genes with differential expression *P*-values over 0.05 were set as not differentially expressed (fold-change set to 0). Resulting *P*-values for each inference method were FDR corrected and TFs with adjusted *P*-values below 0.05 were labeled as active. Figure 5 highlights the overlaps between the 4 different methods. Subsets with TFs inferred active by NLBayes are shown in blue. Here we observe that of the 23 TFs inferred as active by NLBayes, 12 are supported by at least one other method. This analysis illustrates how NLBayes tends to provide a shorter list of active TFs when compared to other methods. This is because the proposed method seeks to find only the best combination of TFs that explain the observed differential expression pattern.

Fibroblast phenotypic plasticity

To demonstrate the utility of our methodology in discovery of novel biology, we applied our algorithm to study fibroblast-to-myofibroblast phenotypic conversion in response to pro-fibrotic signaling molecules TGFβ and CXCL12 (35,53,54). In this experiment, patient derived, immortalized prostate N1 cells were treated with the pro-fibrotic proteins TGFβ and CXCL12, both of which are known to promote

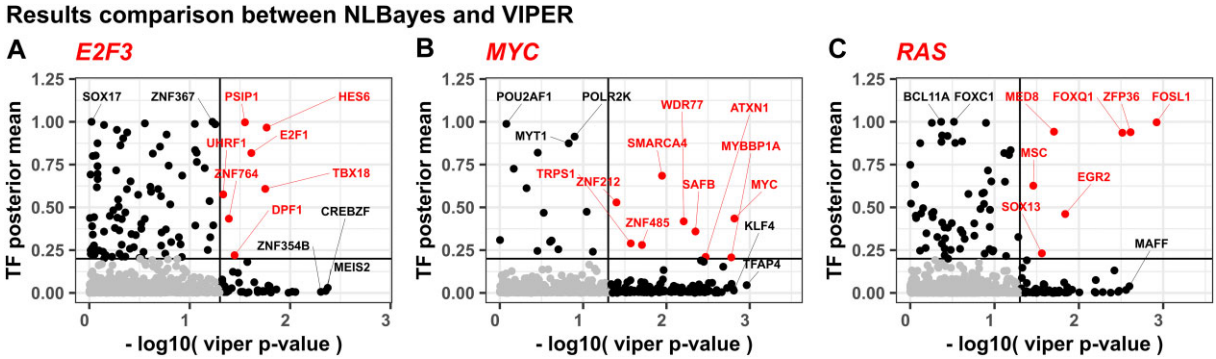

Figure 4. Comparison of active TF predictions by our method (y axis) and VIPER (x axis), in three separate overexpression experiments. Input network used is the BRCA derived regulon from (15). Jointly predicted regulators are colored in red. Top predictions specific to one algorithm are labeled in black. Gray dots show low confidence predictions by both algorithms.

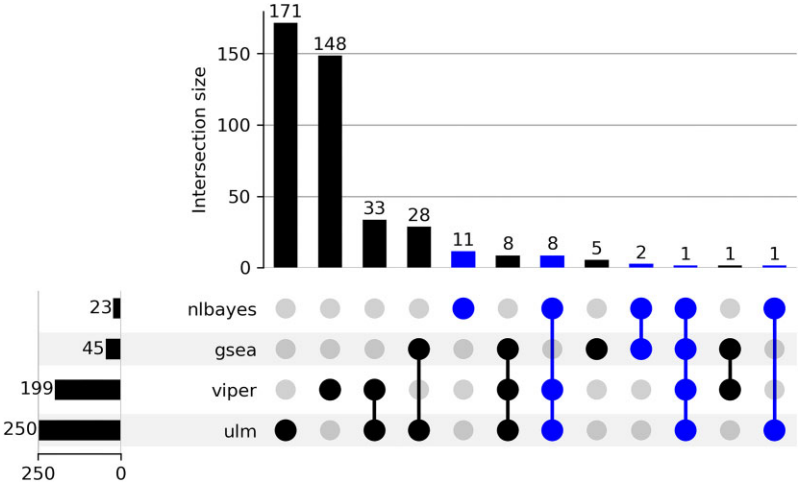

Figure 5. Overlap of inference results by four different methods. Subsets containing TFs inferred active by NLBayes are highlighted in blue.

collagen expression (53). TGFβ acts upon TGF/TGFR signaling axis and activates multiple Smad proteins, while CXCL12 acts upon CXCL12/CXCR4-axis, which transactivates EGFR and downstream signaling through MEK/ERK and PI3K/Akt pathways. Both signaling axes converge in the nucleus and promote the expression of multiple collagen genes (35,54). RNA-Seq data was processed and compared to the background model to generate differential gene expression profiles as previously described (53). Differential gene expression data from TGFβ and CXCL12 treated cells were identified at fold change ≥ 2 and four cutoff thresholds for *P*-value. The different cutoffs for *P*-value were applied to examine the impact of stringency in significance on the inference results, and were chosen such that only 200, 400, 600 or 800 top differentially expressed genes are considered. To achieve this, we sorted the table of differentially expressed genes by *P*-values in ascending order and took the top rows for the analysis. Both datasets were used as input to the TF activity inference algorithm. For these experiments (and the remaining experiments), we used the 3-tissue TF–gene interaction network generated by Farahmand et. al. This network contains interaction edges that are common in at least three of the tissues used in that work and showed consistent performance across multiple datasets (30). Table 5 summarizes the results. Inferred active regulators upon CXCL12 induction are largely similar to that of TGFβ. This is expected as transcriptional profiles induced by TGFβ and CXCL12 are 75% similar (53). The top predicted regulators for TGFβ and CXCL12 for the top 200 DEGs are YAP1, RBPJ, KMT2C, ELF1, STAT1 and BPTF. YAP1 is known to play a role in the development and progression of multiple cancers as a transcriptional regulator of this signaling pathway and may function as a potential target for cancer treatment (71). Moreover, YAP1 has been identified as a driver of myofibroblast differentiation in several tissue phenotypes, like skin, heart, lung, pharynx, liver and kidney (36–45). STAT1 is a member of the STAT protein family. In response to cytokines and growth factors, STAT family members are phosphorylated by the receptor associated kinases, and then form homo- or heterodimers that translocate to the cell nucleus where they act as transcription activators. The protein encoded by this gene can be activated by various ligands including interferon-alpha, interferon-gamma, EGF, PDGF and IL6. This protein mediates the expression of a variety of genes, which is thought to be important for cell viability in response to different cell stimuli (72). RB1 is a negative regulator of the cell cycle. The active, hypophosphorylated form of the protein binds transcription factors from the E2F family, which forms a transcription factor heterodimer that controls the transcription of cell cycle regulatory genes. RB-E2F and MuvB complexes (which contain LIN9) regulate the expression of G1/S and G2/M genes. G1/S genes are repressed during G0 and early G1 by RB-E2F and DREAM complexes that interact

through the DNA binding domains (DBD) of E2F/DP with E2F promoter sites (73).

We note that all these predicted active TFs appear as top regulators for the top 400, 600 and 800 DEGs, indicating the robustness of the algorithm to the number of DEGs used as input.

Fibroblast heterogeneity: single cell experiments

We performed several scRNA-seq experiments to further investigate the phenotypic plasticity of human prostate fibroblasts and characterize heterogeneity in cell populations. For this study, we utilized 4 human prostate cell lines: N1, SFT1, pHPF and iHPF (For more information, see SI File Supplementary Table S1). N1 cells are HPV E6/E7-immortalized prostate stromal fibroblasts originally explanted and grown from a stromal Nodule of benign prostatic hyperplasia (74). They exhibit a fibroblastic morphology, and express fibroblastic markers vimentin and calponin. These cells demonstrate secretion and proliferation profiles consistent with aging primary prostate fibroblasts. SFT1 cells are spontaneously immortalized prostate fibroblasts grown from a prostate of a patient with a Solitary Fibrous Tumor of the prostate (75). These cells carry an uncommon NAB2/STAT6 fusion gene that is associated with solitary fibrous tumors and likely accounted for cellular immortalization. pHPF cells are primary Human Prostate Fibroblasts, purchased at passage 3 from Lifeline Cell Technology, harvested from young adult male. Finally, iHPF cells are created through transduction from pHPF cells with an EF1α-driven hTERT Lentivirus construct and have grown continuously in culture $>$ 30 passages.

We applied scRNA-seq to all 4 cell lines. Figure 6A shows the UMAP projection of the cell lines. The N1 and SFT1 form distinct clusters in close proximity. Most of the cells are in G1 phase. pHPF cells also form a single cluster, mostly consisting of cells in G1 phase. Interestingly the iHPF cells cluster in two groups (A and B) that surround the primary pHPF cells. The majority of cells in iHPF_A cluster are in G1 phase, while iHPF_B consists of a mix of cells in G1, G2, S and M phases. To further investigate the identity of these cells, we merged the data with FACS sorted single cell expression data derived from prostate tissue generated by Henry et al. (58). Figure 6B shows that the RNA expression profiles of the five human prostate fibroblast cell lines N1, SFT1, iHPF_A and iHPF_B, and primary pHPF cells, cluster as expected with that of tissue-derived human prostate fibroblasts. As seen in Figure 6C, the five human prostate fibroblast cell lines share a large signature of highly and commonly expressed genes, likely reflecting their common fibroblastic cell type. In particular, all five cell lines express COL1A1 (collagen 1) and VIM (vimentin). Examination of the top 10 differentially expressed genes in the four immortalized cell lines compared to primary human fibroblasts shows that N1 and SFT1 demonstrate a high degree of overlap and commonly express several inflammation-associated genes (CXCL1, ZNFAS1, CHI3L1). The iHPF_A and iHPF_B share a common gene signature as well that includes gene encoding signaling proteins (BEX1, WNT5A), growth factors and pathways (EREG, IGFBP5), and a gene over-expressed in the autoimmune disease, rheumatoid arthritis (TGM2). However, iHPF_B cells also highly express genes that are not expressed by iHPF_A, including several associated with vasculogenesis or angiogenesis (ANGPT1, F3, ADAMT1) or connective tissue and bone growth (TNFRSF11B). This suggests that iHPF_B cells may phenotypically resemble endothelial cells, which can differentiate from fibroblasts (76). Figure 6E quantifies the average log FC of top expressed markers (Figure 6C) compared to the background (pHPF). Taken together, these data suggest that a seemingly homogenous culture of primary stromal prostate fibroblasts may comprise several subpopulations as have recently been shown for dermal fibroblasts (77).

Next, we sought to quantitatively characterize similarities between cell lines at the transcriptional level. We first performed a differential gene expression analysis using the pHPF cell line as the background. Figure 7A shows a bar plot of total number of upregulated and downregulated genes in each cell line compared to pHPF cells. We performed a gene set enrichment analysis on up & down regulated genes (Figure 7B). As expected, the N1 and SFT1 cells demonstrate a high level of similarity, as we have previously shown that they respond similarly to stimulation with pro-fibrotics (54). Conversely, although immortalized from the pHPF cells, iHPF_A and iHPF_B demonstrate a higher-than-expected dissimilarity, potentially reflecting fibroblast heterogeneity in the primary cell culture from which they were derived.

Next, we applied our algorithm to the DEG profiles from each cell line to quantify similarity in transcriptional gene regulation. Figure 8 shows top inferred regulators in each cell line (left panel bar plots), along with their RNA expression level (middle panel bar plots), and the corresponding enrichment of the differentially expressed target genes (right panel bar plots). The enrichment analysis was performed by quantifying the overlap between the targets of the TF and DEGs using Fisher's exact test. This analysis was performed for comparison of enrichment-based methods with our approach. Enrichment-based approaches do not consider the global topology of the TF–gene interaction network into consideration and yield results that are purely based on the local overlap of TF targets and the set of DEGs. In Figure 8, we observe that many TFs inferred by our method have low enrichment scores. Moreover, some of these show no RNA expression, which suggests that these are false positives.

Among the TF regulators identified, HAND2 was shared across 2 cell lines. The protein encoded by this gene belongs to the basic helix-loop-helix family of transcription factors and, among many other development-related functions, is required for vascular development and regulation of angiogenesis, possibly through a VEGF signaling pathway (78).

The N1 and SFT1 cell lines shared expression of CEBPD, GTF2F1 and MXI1. CEBPD is an intron-less gene that encodes a bZIP transcription factor which can bind as a homodimer to certain DNA regulatory regions. It can also form heterodimers with the related protein CEBP-alpha. The encoded protein is important in the regulation of genes involved in immune and inflammatory responses and may be involved in the regulation of genes associated with activation and/or differentiation of macrophages. It may also be involved in the early stages of adipogenesis (79,80).

GTF2F1 encodes TFIIF, a general transcription initiation factor that binds to RNA polymerase II and helps to recruit it to the initiation complex in collaboration with TFIIB. It is also a JNK1/3-binding partner and may modulate c-JUN-mediated MAPK signaling in cell proliferation, differentiation, migration, senescence and apoptosis (81). MXI1 encodes a basic helix-loop-helix protein that inhibits the

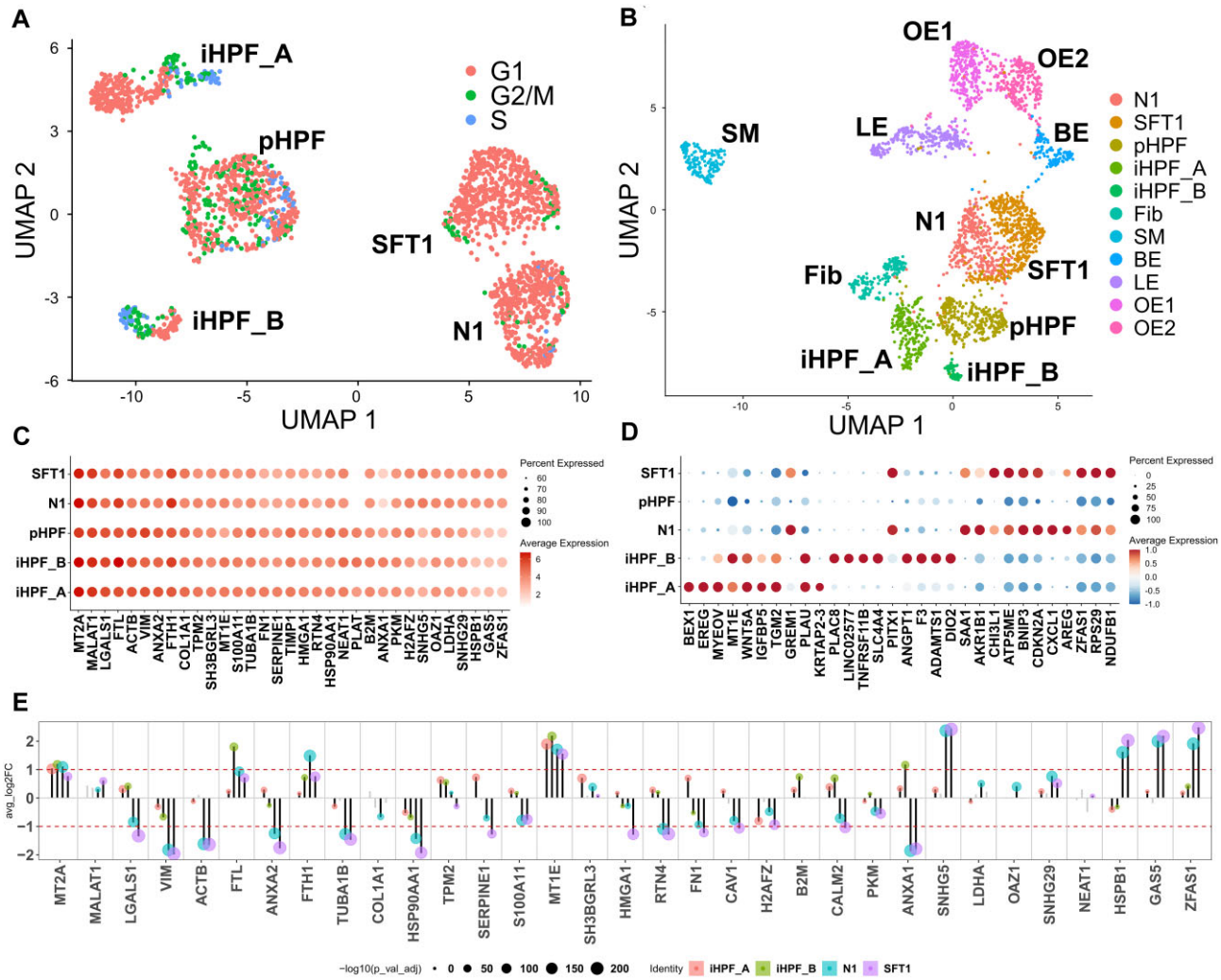

Figure 6. Single cell gene expression data from four prostate cell lines. **(A)** UMAP projection of cell lines. iHPF shows two separate phenotypes here termed iHPF_A and iHPF_B. Cell cycle phase G1 is prominent in N1 and SFT1, but pHPF cells show significant number of cells in S and G2/M phases. **(B)** Integration single prostate tissue data (58), showing clusters for Fibroblasts (Fib), Smooth Muscle (SM), Basal Epithelia (BE), Luminal Epithelia (LE) and Other Epithelia (OE1, OE2). All 4 cell lines, N1, SFT1, pHPF and iHPF, appear as interconnected clusters, lying between epithelial and fibroblast cells. **(C)** Top expressed genes. Overall, all 5 cell classes share the same highly expressed genes including COL1A1. **(D)** Top 10 differentially expressed genes with respect to pHPF as a reference. Smallest and largest *P*-values are 1E-128 and 1E-26 respectively. See SI File Tables S2 and S3 for a list of top genes for each cell line, as shown in (C) and (D); full list of differentially expressed genes is available in SI Table 1. **(E)** Differential expression with respect to pHPF, for the highly expressed genes shown in (C). FC of genes in panel C compared to pHPF. Notably, even though collagen expression high in across all groups, it is downregulated in N1 when compared to pHPF.

transcriptional activity of MYC by sequestering MAX, thus preventing the formation of MYC-MAX heterodimers, and by competing with MYC-MAX heterodimers for binding to target sites (82).

The iHPF_A and iHPF_B cell lines shared expression of the TF regulators PPARG and TCF21. PPARG encodes a member of the peroxisome proliferator-activated receptor (PPAR) subfamily of nuclear receptors. PPARs form heterodimers with retinoid X receptors (RXRs) and these heterodimers regulate transcription of various genes that regulate adipocyte differentiation and, pathologically, the development or progression of obesity, diabetes, atherosclerosis and cancer (83,84). TCF21 encodes a transcription factor of the basic helix-loop-helix family. The TCF21 product is mesoderm specific, and expressed in embryonic epicardium, mesenchyme-derived tissues of lung, gut, gonad, and both mesenchymal and glomerular epithelial cells in the kidney. It is involved in the differentiation of mesenchymal cells to fibroblasts (85).

Of note, many of these transcriptional regulators are basic helix-loop-helix TFs, and three TF regulators in particular—CEBPD, TCF21 and HAND2 have been identified as promotors of mesenchymal cell differentiation towards the fibroblast lineage (as opposed to the smooth muscle cell lineage). This suggests that the immortalized fibroblast cell lines express TF regulators that function to maintain the fibroblast phenotype as well as those that may extend this phenotype towards that of immune/inflammatory cells (CEBPD), adipocytes (CEBPD, PPARG) or vascular cells (HAND2). This suggests that fibroblast phenotypic plasticity is perhaps a common rather than exceptional cellular state that may be identified by the expression of particular TF regulators.

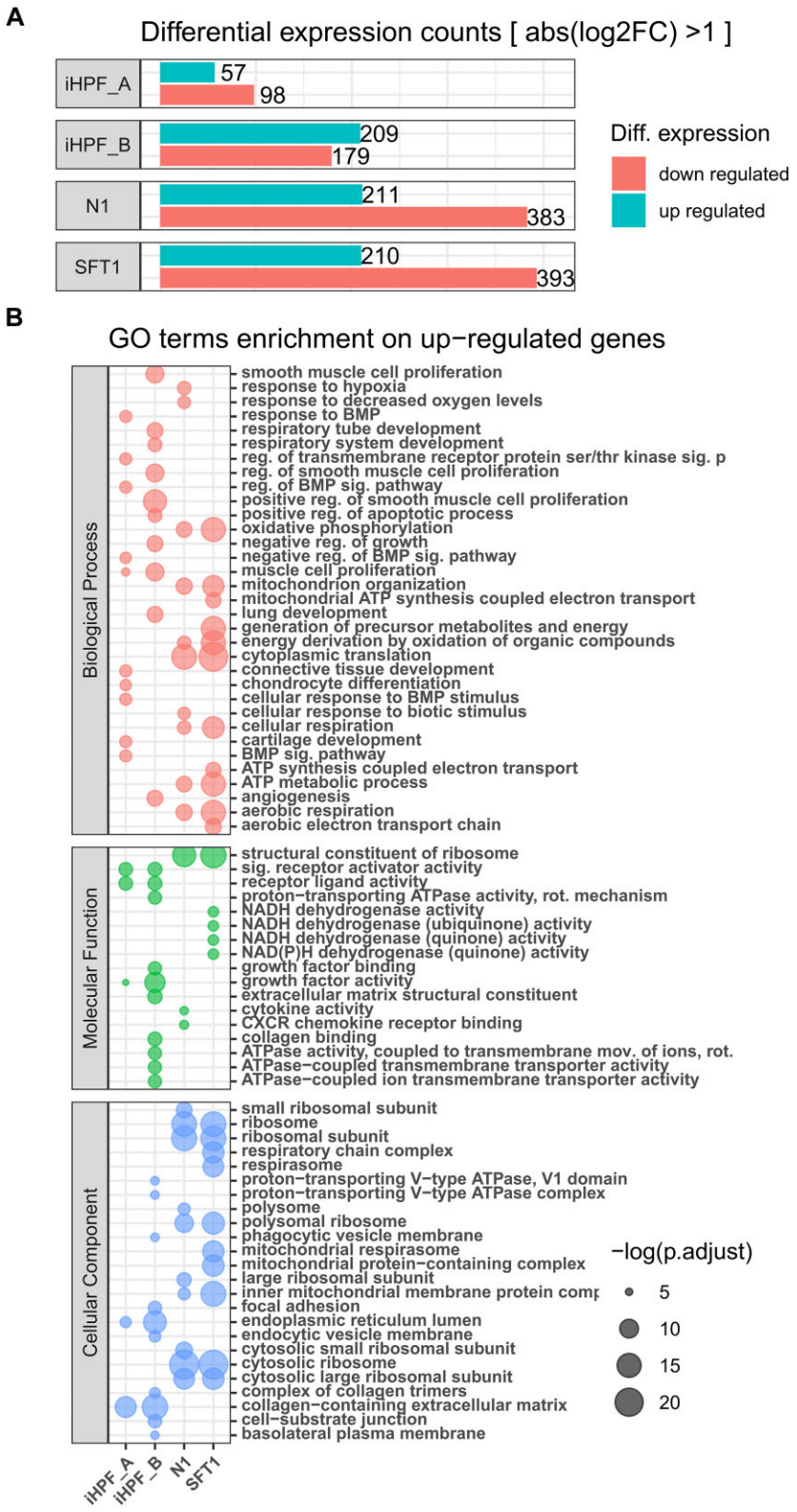

Figure 7. (A) Total number of DEGs compared to the background model (pHPF). **(B)** GO term Enrichment analysis of up regulated genes in each cell line (columns). See SI File Fig S1 for the GO term Enrichment analysis on down regulated genes.

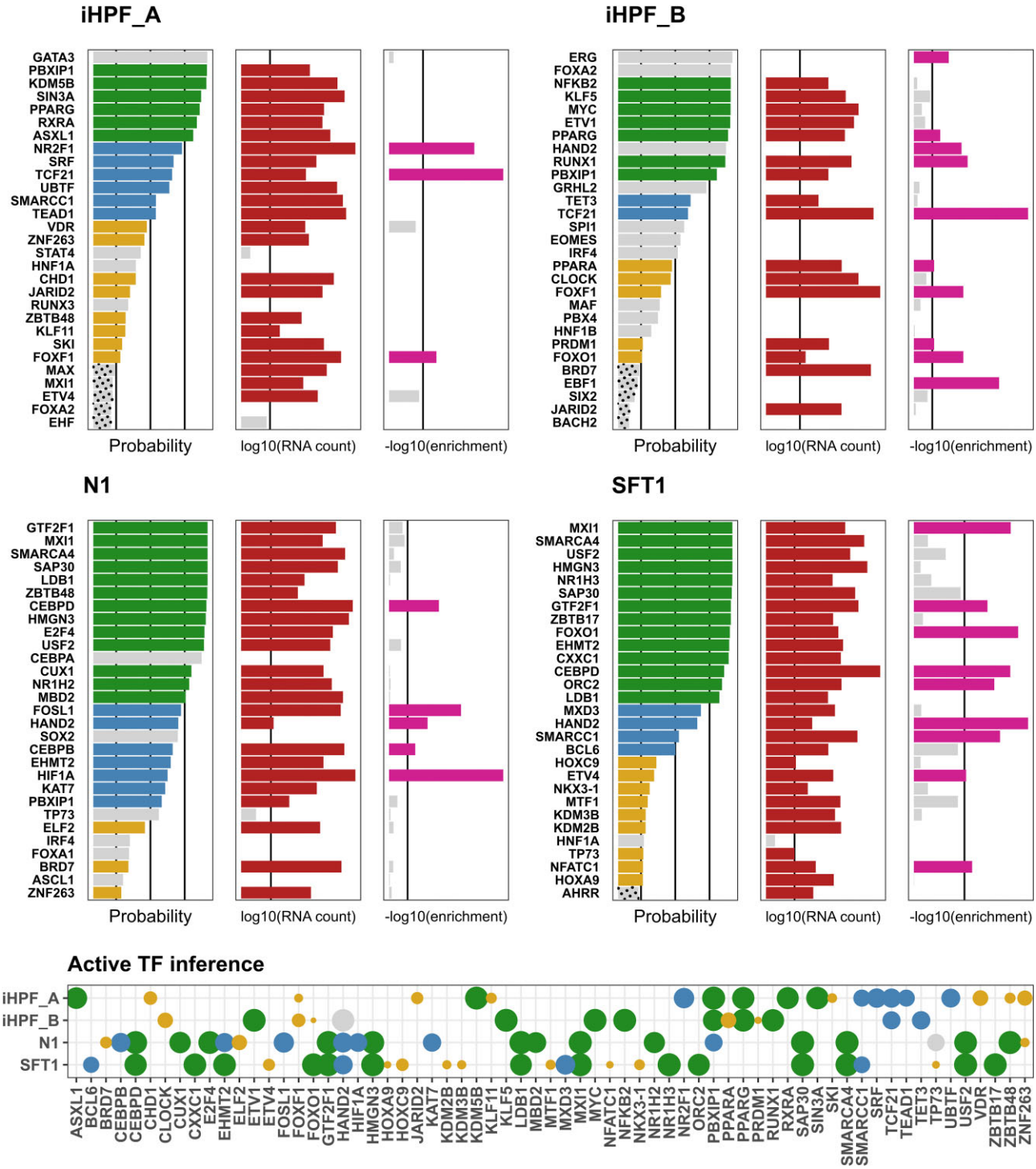

Figure 8. Active TF inference for each cluster of immortalized cells. Gene markers were identified by comparison with pHPF and fed into the inference algorithm. In each data set, the left panel bar plots show the inferred probability of regulator activity by the algorithm. Vertical lines mark 0.2, 0.5 and 0.8 probability thresholds. The right panel bar plots show the Enrichment Analysis of DE targets (Fisher's Exact test) for comparison. Significant *P*-values (<0.05) are highlighted in pink. The middle panel bar plots show mean RNA expression across single cells. TFs with expression above 25-percentile are highlighted in red. TFs with inference posterior probability $P > 0.2,\ 0.5,\ 0.8$ that also show significant enrichment and RNA expression are highlighted in yellow, blue, and green respectively. TFs with posterior probability $P > 0.2$, but the expression level below the 25 percentile are highlighted in gray. TFs with $P < 0.2$ are shown with a dotted pattern. The bottom panel shows active TFs inferred in each cell line (rows) by the algorithm. N1 and SFT show a similar pattern of TF activity.

Discussion

In this work, we presented an algorithm for inference of TF activity from differential gene expression profiles and causal graphs. The algorithm incorporates transcriptional logic in the context of Bayesian networks, allowing for probabilistic deviation from deterministic logic rules. The probabilistic framework provides the flexibility for 'plug-and-play' integration of various logic models. In this study, we focused on one such model (OR-NOR logic). As a future direction we plan to extend the packages so users can choose the logic prior to running the inference.

The queries are run on causal graphs of TF–gene interactions. We provide several options for such graphs assembled from small-scale curated databases (86,87), large-scale public databases (88–90), as well as *de novo* reconstructed graphs from high-throughput experiments (30). We note that the quality and the coverage of the causal graph has a major impact on the ability of regulator activity inference models. Most curated publicly available network of transcriptional regulation with annotation on mode of regulation are small and very limited in their coverage while other higher coverage networks may consist of noisy inferred interactions. Unlike standard enrichment analysis methods, our framework has been designed to account for noise (applicability of interactions and noise in direction of regulation).

Bayesian Networks are Directed Acyclic Graphs (DAGs) and as such, feedback loops cannot be directly modeled in this context, which is a limitation of this approach. Another limitation of our approach is that it is designed to detect TF activation, but not TF deactivation. Moreover, since we only consider the OR-NOR transcriptional regulatory logic, results produced by this approach may miss TFs with alternative regulatory relationships. Since the approach is Bayesian and takes the entire topology of the network into account, by design it outputs a minimal number of TFs whose activation can explain the gene expression data. This is an advantage of our algorithm over enrichment analysis methods that typically contain a large proportion of false positives. The disadvantage may be that sometimes not all true positives get high posteriors probabilities, especially if there are many active regulators present. This can be observed in the validation and benchmark subsections, where we used overexpression experiments as input for the inference algorithm (see Table 4 and Figure 4). In the MYC overexpression data set, MYC itself was recovered as an active regulator. Although this was not the case for other over-expression data sets, the predicted results collectively point to the relevant biology. It should be noted that over-expression experiments may contain many off-target effects that will confound the inference results. Moreover, the network may not sufficiently capture the regulatory interaction in the context of the biological experiment. Indeed, in the case of MYC, we observe that the 'regulonbrca' network contains many interactions between MYC and downstream genes that are not present in the three-tissue network. On the other hand, the three-tissue network provides better performance in other experiments. The main difference between the three-tissue and 'regulonbrca' networks is that the former attempts to represent a universal non-context specific network, while the latter is a breast cancer-specific network. The ability of regulator inference algorithms to recover upstream regulators depends on multiple factors, including sufficiency of evidence in the input gene expression data, noise, and importantly encapsulation of the regulatory interaction in the network. Since regulatory networks are dynamic and context-dependent, it is unlikely for any universal network or methodology to capture the exact modulators. However, tools such as ours can be tried with multiple networks and the predictions made by the network should point to the relevant biology. We have provided multiple networks for users to try in our web application. Moreover, users can try custom networks within the web app, or by using the provided Python and R packages. It is worth noting that predicted TFs that show no expression in the data (see Figure 8), should be considered false positives. If desired, information about the expression of TFs may be used to filter the network prior to the inference analysis, which might reveal an alternative set TFs that explain the input data.

For the inference process, we utilized Gibbs Sampling, an MCMC algorithm that is widely used in Bayesian networks. A drawback from MCMC models in Bayesian networks is the convergence time. We implemented the core of the inference in C++ to reduce the wait time. Convergence time mainly depends on the number of interactions in the network used. For the 3-tissue network which has approximately 250 thousand TF–gene interactions, the run time is of the order of 20 minutes. For the used ARACNe network 'regulonbrca', which after filtering contains around 74 thousand TF–gene interactions, the run time is about 2 minutes. Other strategies can be taken to speed up processing time. For instance, an enrichment-based test can be run a priori to exclude TFs with insufficient differentially expressed targets, effectively reducing the network size. This will result in a significant speed up in convergence time, albeit some border line cases may be lost.

Our tool is an exploratory discovery tool that provides a narrow list of potentially relevant TFs, summarizing the observed differential gene expression data. This is similar to standard GO term and pathway enrichment analysis that are also typically applied to summarize differential gene expression data. The focus of our tool is transcriptional regulation and our algorithm can be used as a complementary tool in conjunction with enrichment analysis methods.

To increase the utility of our algorithm, we provide user-friendly R and Python packages as well as a web-based platform with integrated interactive visualization. The preprocessing steps for speeding up the algorithm are implemented as default in the webserver. As databases of causal transcriptional regulatory interactions become more available, we will integrate them in the web-platform and accordingly optimize the inference algorithm for each network.

Data and software availability

'scRNAseq of Primary and Immortalized Human Prostate Fibroblast Cell Lines'

BioProject Accession number: PRJNA881605

Study Accession number: SRP397809

SRA Accession numbers: SRX17617080, SRX17617081, SRX17617082, SRX17617083

Use NCBI's SRA toolkit to download the 4 datasets above. For further instructions, see:

https://www.ncbi.nlm.nih.gov/sra/docs/sradownload

We make our inference algorithm available to use through the following web application:

https://umbibio.math.umb.edu/nlbayes

Open-source R and Python packages are available at Github:

https://github.com/umbibio/nlbayes-r (doi:10.5281/zenodo.7105306)

https://github.com/umbibio/nlbayes-python (doi:10.5281/zenodo.7105233)

We have used R version 4.1.3 and Python 3.10 to develop and test the corresponding packages. Detailed instructions and examples are available in each corresponding repository.

All figures presented in this work can be reproduced by following instructions available at the GitHub repository:

https://github.com/argearriojas/nlbayes-reproducibility (doi: 10.5281/zenodo.10116763)

Corresponding files needed for generating the figures are made available at Zenodo:

https://zenodo.org/records/10116664 (doi: 10.5281/zenodo.10116663).

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

This work was supported by grants AI150090 (KZ), R01AI167570 (KZ), DK104310 (JAM), and CA156734 (to J.A.M., K.Z.) from the National Institutes of Health; A.A. was supported in part by College of Science and Mathematics Dean's Doctoral Research Fellowship through fellowship support from Oracle [ID R20000000025727]. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Conflict of interest statement

None declared.

References

1. Lelli,K.M., Slattery,M. and Mann,R.S. (2012) Disentangling the many layers of eukaryotic transcriptional regulation. *Annu. Rev. Genet.*, **46**, 43–68.
2. de Jong,H. (2002) Modeling and simulation of genetic regulatory systems: a literature review. *J. Comput. Biol.*, **9**, 67–103.
3. Wilkinson,A.C., Nakauchi,H. and Gottgens,B. (2017) Mammalian transcription factor networks: recent advances in interrogating biological complexity. *Cells*, **5**, 319–331.
4. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
5. Barolo,S. and Posakony,J.W. (2002) Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev.*, **16**, 1167–1181.
6. Bonneau,R., Reiss,D.J., Shannon,P., Facciotti,M., Hood,L., Baliga,N.S. and Thorsson,V. (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.
7. Asif,H.M.S. and Sanguinetti,G. (2011) Large-scale learning of combinatorial transcriptional dynamics from gene expression. *Bioinformatics*, **27**, 1277–1283.
8. Ocone,A. and Sanguinetti,G. (2011) Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics*, **27**, 2873–2879.
9. Bulashevska,S. and Eils,R. (2005) Inferring genetic regulatory logic from expression data. *Bioinformatics*, **21**, 2706–2713.
10. Veber,P., Guziolowski,C., Le Borgne,M., Radulescu,O. and Siegel,A. (2008) Inferring the role of transcription factors in regulatory networks. *BMC Bioinf.*, **9**, 228.
11. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.
12. Schacht,T., Oswald,M., Eils,R., Eichmüller,S.B. and König,R. (2014) Estimating the activity of transcription factors by the effect on their target genes. *Bioinformatics*, **30**, i401–i407.
13. Jiang,P., Freedman,M.L., Liu,J.S. and Liu,X.S. (2015) Inference of transcriptional regulation in cancers. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7731–7736.
14. Fröhlich,H. (2015) biRte: bayesian inference of context-specific regulator activities and transcriptional networks. *Bioinformatics*, **31**, 3290–3298.
15. Alvarez,M.J., Shen,Y., Giorgi,F.M., Lachmann,A., Ding,B.B., Ye,B.H. and Califano,A. (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, **48**, 838–847.
16. Yu,B., Zhang,K., Milner,J.J., Toma,C., Chen,R., Scott-Browne,J.P., Pereira,R.M., Crotty,S., Chang,J.T., Pipkin,M.E., *et al.* (2017) Epigenetic landscapes reveal transcription factors that regulate CD8+ T cell differentiation. *Nat. Immunol.*, **18**, 573–582.
17. Keenan,A.B., Torre,D., Lachmann,A., Leong,A.K., Wojciechowicz,M.L., Utti,V., Jagodnik,K.M., Kropiwnicki,E., Wang,Z. and Ma'ayan,A. (2019) ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res.*, **47**, W212–W224.
18. Honkela,A., Girardot,C., Gustafson,E.H., Liu,Y.-H., Furlong,E.E.M., Lawrence,N.D. and Rattray,M. (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 7793–7798.
19. Djordjevic,M., Sengupta,A.M. and Shraiman,B.I. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.
20. Hashimoto,R.F., Kim,S., Shmulevich,I., Zhang,W., Bittner,M.L. and Dougherty,E.R. (2004) Growing genetic regulatory networks from seed genes. *Bioinformatics*, **20**, 1241–1247.
21. Friedman,N. (2003) Probabilistic models for identifying regulation networks. *Bioinformatics*, **19**, ii57.
22. Bar-Joseph,Z., Gerber,G.K., Lee,T.I., Rinaldi,N.J., Yoo,J.Y., Robert,F., Gordon,D.B., Fraenkel,E., Jaakkola,T.S., Young,R.A., *et al.* (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, **21**, 1337–1342.
23. Segal,E., Taskar,B., Gasch,A., Friedman,N. and Koller,D. (2001) Rich probabilistic models for gene expression. *Bioinformatics*, **17**, S243–S252.
24. Zarringhalam,K., Enayetallah,A., Gutteridge,A., Sidders,B. and Ziemek,D. (2013) Molecular causes of transcriptional response: a Bayesian prior knowledge approach. *Bioinformatics*, **29**, 3167–3173.
25. Fakhry,C.T., Choudhary,P., Gutteridge,A., Sidders,B., Chen,P., Ziemek,D. and Zarringhalam,K. (2016) Interpreting transcriptional changes using causal graphs: new methods and their practical utility on public networks. *BMC Bioinf.*, **17**, 318.
26. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
27. Chindelevitch,L., Ziemek,D., Enayetallah,A., Randhawa,R., Sidders,B., Brockel,C. and Huang,E.S. (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.
28. Chindelevitch,L., Loh,P.-R., Enayetallah,A., Berger,B. and Ziemek,D. (2012) Assessing statistical significance in causal graphs. *BMC Bioinf.*, **13**, 35.

29. Kramer,A., Green,J., Pollard,J. Jr and Tugendreich,S. (2014) Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, **30**, 523–530.
30. Farahmand,S., O'Connor,C., Macoska,J.A. and Zarringhalam,K. (2019) Causal inference engine: a platform for directional gene set enrichment analysis and inference of active transcriptional regulators. *Nucleic Acids Res.*, **47**, 11563–11573.
31. Gao,S., Dai,Y. and Rehman,J. (2021) A Bayesian inference transcription factor activity model for the analysis of single-cell transcriptomes. *Genome Res.*, **31**, 1296–1311.
32. Markowetz,F., Bloch,J. and Spang,R. (2005) Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, **21**, 4026–4032.
33. Lachmann,A., Giorgi,F.M., Lopez,G. and Califano,A. (2016) ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, **32**, 2233–2235.
34. Liu,A., Trairatphisan,P., Gjerga,E., Didangelos,A., Barratt,J. and Saez-Rodriguez,J. (2019) From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL. *Npj Syst. Biol. Appl.*, **5**, 40.
35. Gharaee-Kermani,M., Kasina,S., Moore,B.B., Thomas,D., Mehra,R. and Macoska,J.A. (2012) CXC-type chemokines promote myofibroblast phenoconversion and prostatic fibrosis. *PLoS One*, **7**, e49278.
36. Piersma,B., Rond,S.de, Werker,P.M.N., Boo,S., Hinz,B., Beuge,M.M. van and Bank,R.A. (2015) YAP1 is a driver of myofibroblast differentiation in normal and diseased fibroblasts. *Am. J. Pathol.*, **185**, 3326–3337.
37. Pang,P., Si,W., Wu,H., Wang,C., Liu,K., Jia,Y., Zhang,Z., Zhang,F., Kong,X., Yang,Y., *et al.* (2023) The circular RNA circHelz enhances cardiac fibrosis by facilitating the nuclear translocation of YAP1. *Transl. Res.*, **257**, 30–42.
38. Li,T.-Y., Su,W., Li,L.-L., Zhao,X.-G., Yang,N., Gai,J.-X., Lv,X., Zhang,J., Huang,M.-Q., Zhang,Q., *et al.* (2022) Critical role of PAFR/YAP1 positive feedback loop in cardiac fibrosis. *Acta Pharmacol. Sin.*, **43**, 2862–2872.
39. Li,S., Zhou,X., Zeng,R., Lin,L., Zou,X., Yan,Y., Lu,Z., Xia,J., Zhang,L., Ni,S., *et al.* (2022) YAP1 silencing attenuated lung injury/fibrosis but worsened diaphragmatic function by regulating oxidative stress and inflammation response in mice. *Free Radic. Biol. Med.*, **193**, 485–498.
40. Xu,W., Song,W., Wang,Y.U., Zan,Y., Zhang,M., Li,M., Huang,Q., Zhao,W., Sun,Y.U., Hoffman,R.M., *et al.* (2021) Efficacy of YAP1-gene knockdown to inhibit alveolar-epithelial-cell senescence and alleviate idiopathic pulmonary fibrosis (IPF). *Cancer Genomics Proteomics*, **18**, 451–459.
41. Lee,P.-J., Sui,Y.-H., Liu,T.-T., Tsang,N.-M., Huang,C.-H., Lin,T.-Y., Chang,K.-P. and Liu,S.-C. (2022) Epstein-Barr viral product-containing exosomes promote fibrosis and nasopharyngeal carcinoma progression through activation of YAP1/FAPα signaling in fibroblasts. *J. Exp. Clin. Cancer Res.*, **41**, 254.
42. Salloum,S., Jeyarajan,A.J., Kruger,A.J., Holmes,J.A., Shao,T., Sojoodi,M., Kim,M.-H., Zhuo,Z., Shroff,S.G., Kassa,A., *et al.* (2021) Fatty acids activate the transcriptional coactivator YAP1 to promote liver fibrosis via p38 mitogen-activated protein kinase. *Cell. Mol. Gastroenterol. Hepatol.*, **12**, 1297–1310.
43. Wang,X., Wang,G., Qu,J., Yuan,Z., Pan,R. and Li,K. (2020) Calcipotriol inhibits NLRP3 signal through YAP1 activation to alleviate cholestatic liver injury and fibrosis. *Front. Pharmacol.*, **11**, 200.
44. Li,X., Zhang,F., Qu,L., Xie,Y., Ruan,Y., Guo,Z., Mao,Y., Zou,Q., Shi,M., Xiao,Y., *et al.* (2021) Identification of YAP1 as a novel downstream effector of the FGF2/STAT3 pathway in the pathogenesis of renal tubulointerstitial fibrosis. *J. Cell. Physiol.*, **236**, 7655–7671.
45. Allison,S. (2021) A SOX-9-NAV3-YAP1 axis in kidney fibrosis. *Nat. Rev. Nephrol.*, **17**, 297.
46. Gelman,A. and Rubin,D.B. (1992) Inference from Iterative Simulation Using Multiple Sequences. *Statist. Sci.*, **7**, 457–472.
47. Wang,Q., Armenia,J., Zhang,C., Penson,A.V., Reznik,E., Zhang,L., Minet,T., Ochoa,A., Gross,B.E., Iacobuzio-Donahue,C.A., *et al.* (2018) Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data*, **5**, 180061.
48. Oki,S., Ohta,T., Shioi,G., Hatanaka,H., Ogasawara,O., Okuda,Y., Kawaji,H., Nakaki,R., Sese,J. and Meno,C. (2018) ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.*, **19**, e46255.
49. Consortium,Gte., Aguet,F., Brown,A.A., Castel,S.E., Davis,J.R., He,Y., Jo,B., Mohammadi,P., Park,Y., Parsana,P., *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
50. Bild,A.H., Yao,G., Chang,J.T., Wang,Q., Potti,A., Chasse,D., Joshi,M.-B., Harpole,D., Lancaster,J.M., Berchuck,A., *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.
51. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2012) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
52. Smyth,G.K. (2005) limma: linear Models for Microarray Data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer New York, NY, pp. 397–420.
53. Patalano,S., Rodríguez-Nieves,J., Colaneri,C., Cotellessa,J., Almanza,D., Zhilin-Roth,A., Riley,T. and Macoska,J. (2018) CXCL12/CXCR4-mediated procollagen secretion is coupled to cullin-RING ubiquitin ligase activation. *Sci. Rep.*, **8**, 3499.
54. Rodríguez-Nieves,J.A., Patalano,S.C., Almanza,D., Gharaee-Kermani,M. and Macoska,J.A. (2016) CXCL12/CXCR4 axis activation mediates prostate myofibroblast phenoconversion through non-canonical EGFR/MEK/ERK signaling. *PLoS One*, **11**, e0159490.
55. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
56. Stuart,T., Butler,A., Hoffman,P., Hafemeister,C., Papalexi,E., Mauck,W.M., Hao,Y., Stoeckius,M., Smibert,P. and Satija,R. (2019) Comprehensive integration of single-cell data. *Cell*, **177**, 1888–1902.
57. Hafemeister,C. and Satija,R. (2019) Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.*, **20**, 296.
58. Henry,G.H., Malewska,A., Joseph,D.B., Malladi,V.S., Lee,J., Torrealba,J., Mauck,R.J., Gahan,J.C., Raj,G.V., Roehrborn,C.G., *et al.* (2018) A cellular anatomy of the normal adult human prostate and prostatic urethra. *Cell Rep.*, **25**, 3530–3542.
59. Chen,H.-Z., Tsai,S.-Y. and Leone,G. (2009) Emerging roles of E2Fs in cancer: an exit from cell cycle control. *Nat. Rev. Cancer*, **9**, 785–797.
60. Li,Y., Luo,H., Liu,T., Zacksenhaus,E. and Ben-David,Y. (2015) The ets transcription factor Fli-1 in development, cancer and disease. *Oncogene*, **34**, 2022–2031.
61. Qi,J., Pellecchia,M. and Ronai,Z.A. (2010) The Siah2-HIF-FoxA2 axis in prostate cancer – new markers and therapeutic opportunities. *Oncotarget*, **1**, 379–385.
62. Bhatlekar,S., Addya,S., Salunek,M., Orr,C.R., Surrey,S., McKenzie,S., Fields,J.Z. and Boman,B.M. (2014) Identification of a developmental gene expression signature, including HOX genes, for the normal human colonic crypt stem cell niche: overexpression of the signature parallels stem cell overpopulation during colon tumorigenesis. *Stem Cells Dev.*, **23**, 167–179.
63. Yamashita,T., Tazawa,S., Yawei,Z., Katayama,H., Kato,Y., Nishiwaki,K., Yokohama,Y. and Ishikawa,M. (2006) Suppression of invasive characteristics by antisense introduction of overexpressed HOX genes in ovarian cancer cells. *Int. J. Oncol.*, **28**, 931–938.

64. Xie,G., Dong,P., Chen,H., Xu,L., Liu,Y., Ma,Y., Zheng,Y., Yang,J., Zhou,Y., Chen,L., *et al.* (2021) Decreased expression of ATF3, orchestrated by β-catenin/TCF3, miR-17-5p and HOXA11-AS, promoted gastric cancer progression via increased β-catenin and CEMIP. *Exp. Mol. Med.*, **53**, 1706–1722.
65. Gui,T., Liu,M., Yao,B., Jiang,H., Yang,D., Li,Q., Zeng,X., Wang,Y., Cao,J., Deng,Y., *et al.* (2021) TCF3 is epigenetically silenced by EZH2 and DNMT3B and functions as a tumor suppressor in endometrial cancer. *Cell Death Differ.*, **28**, 3316–3328.
66. Krauß,L., Urban,B.C., Hastreiter,S., Schneider,C., Wenzel,P., Hassan,Z., Wirth,M., Lankes,K., Terrasi,A., Klement,C., *et al.* (2021) HDAC2 facilitates pancreatic cancer metastasis. *Cancer Res.*, **82**, 695–707.
67. Huang,W., Chen,J., Liu,X., Liu,X., Duan,S., Chen,L., Liu,X., Lan,J., Zou,Y., Guo,D., *et al.* (2021) MIER3 induces epithelial-mesenchymal transition and promotes breast cancer cell aggressiveness via forming a co-repressor complex with HDAC1/HDAC2/Snail. *Exp. Cell. Res.*, **406**, 112722.
68. Chi,T., Wang,M., Wang,X., Yang,K., Xie,F., Liao,Z. and Wei,P. (2021) PPAR-γ modulators as current and potential cancer treatments. *Front. Oncol.*, **11**, 737776.
69. Chen,M., Huang,B., Zhu,L., Chen,K., Liu,M. and Zhong,C. (2020) Structural and Functional Overview of TEAD4 in Cancer Biology. *Ott.*, **13**, 9865–9874.
70. Badia-i-Mompel,P., Vélez Santiago,J., Braunger,J., Geiss,C., Dimitrov,D., Müller-Dott,S., Taus,P., Dugourd,A., Holland,C.H., Ramirez Flores,R.O., *et al.* (2022) decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinform. Adv.*, **2**, vbac016.
71. Astudillo,P. (2022) An emergent Wnt5a/YAP/TAZ regulatory circuit and its possible role in cancer. *Semin. Cell Dev. Biol.*, **125**, 45–54.
72. Harrison,A.R. and Moseley,G.W. (2020) The dynamic interface of viruses with STATs. *J. Virol.*, **94**, e00856-20.
73. Fischer,M. and Müller,G.A. (2017) Cell cycle transcription control: dREAM/MuvB and RB-E2F complexes. *Crit. Rev. Biochem. Mol. Biol.*, **52**, 638–662.
74. Begley,L.A., Kasina,S., MacDonald,J. and Macoska,J.A. (2008) The inflammatory microenvironment of the aging prostate facilitates cellular proliferation and hypertrophy. *Cytokine*, **43**, 194–199.
75. Gharaee-Kermani,M., Mehra,R., Robinson,D.R., Wei,J.T. and Macoska,J.A. (2014) Complex cellular composition of solitary fibrous tumor of the prostate. *Am. J. Pathol.*, **184**, 732–739.
76. Junker,J.P.E., Lönnqvist,S., Rakar,J., Karlsson,L.K., Grenegård,M. and Kratz,G. (2013) Differentiation of human dermal fibroblasts towards endothelial cells. *Differentiation*, **85**, 67–77.
77. Hu,M.S., Moore,A.L. and Longaker,M.T. (2018) A fibroblast is not a fibroblast is not a fibroblast. *J. Invest. Dermatol.*, **138**, 729–730.
78. Yamagishi,H., Olson,E.N. and Srivastava,D. (2000) The basic helix-loop-helix transcription factor, dHAND, is required for vascular development. *J. Clin. Invest.*, **105**, 261–270.
79. Balamurugan,K. and Sterneck,E. (2013) The many faces of C/EBPδ and their relevance for inflammation and cancer. *Int J Biol Sci*, **9**, 917–933.
80. Hishida,T., Nishizuka,M., Osada,S. and Imagawa,M. (2009) The role of C/EBPδ in the early stages of adipogenesis. *Biochimie*, **91**, 654–657.
81. Sun,Y., Liu,W.-Z., Liu,T., Feng,X., Yang,N. and Zhou,H.-F. (2015) Signaling pathway of MAPK/ERK in cell proliferation, differentiation, migration, senescence and apoptosis. *J. Recept. Signal Transduct.*, **35**, 600–604.
82. Zervos,A.S., Gyuris,J. and Brent,R. (1993) Mxi1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell*, **72**, 223–232.
83. Rosen,E.D. (2005) The transcriptional basis of adipocyte development. *Prostaglandins Leukotrienes Essent. Fatty Acids*, **73**, 31–34.
84. Cataldi,S., Costa,V., Ciccodicola,A. and Aprile,M. (2021) PPARγ and diabetes: beyond the genome and towards personalized medicine. *Curr. Diab. Rep.*, **21**, 18–15.
85. Lighthouse,J.K. and Small,E.M. (2016) Transcriptional control of cardiac fibroblast plasticity. *J. Mol. Cell. Cardiol.*, **91**, 52–60.
86. Kolchanov,N.A., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Stepanenko,I.L., Merkulova,T.I., Pozdnyakov,M.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic. Acids. Res.*, **30**, 312–317.
87. Han,H., Shim,H., Shin,D., Shim,J.E., Ko,Y., Shin,J., Kim,H., Cho,A., Kim,E., Lee,T., *et al.* (2015) TRRUST: a reference database of human transcriptional regulatory interactions. *Sci. Rep.*, **5**, 11432.
88. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic. Acids. Res.*, **28**, 27–30.
89. Jensen,L.J., Kuhn,M., Stark,M., Chaffron,S., Creevey,C., Muller,J., Doerks,T., Julien,P., Roth,A., Simonovic,M., *et al.* (2009) STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.
90. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,Ö., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Received: June 7, 2023. **Revised:** November 12, 2023. **Editorial Decision:** November 20, 2023. **Accepted:** November 24, 2023