**S1 Text. Supplemental Information.** Additional overview and analyses.

**Network inference methods**

Several methods have been developed to estimate directed graphs of genes from transcription time-series data (Fig S1). Broadly, these methods estimate directed networks in which the directed edges between nodes---representing genes---indicate a cause-effect relationship between those genes, such that perturbing the expression levels of the causal gene would lead to changes in expression of the effect gene [20].

Let *G* be the set of all genes and *g* be a single gene. Let *¬ g* be *G* with *g* removed. Let there be *T* time points total, and let *t* be a single time point ranging from {1, 2, …, T}. Let be the expression of gene *g* at time *t*. Let be the residual noise at time *t*. Let : denote sequencing through values, for example would denote all the values through . Let refer to the causal parents of gene *g* at time *t* in dynamic Bayesian Networks. For example, may include . Let *g'* be the gene we are testing to be causal for gene *g*. Let *l* be the time lag of the causal interaction. We are testing the existence of the edge at lag *l*.

**Mutual information**

Mutual information (MI) methods assess the MI between the expression of *g'* at the *l-*th previous time point and the expression of *g* at the current time point (Fig S1A) [21-25]:

A causal edge is included if exceeds a threshold. MI methods have the advantage of being simple and fast. However, they do not give insight into the sign of two genes' relationship (i.e., activation or repression) because MI is an unsigned metric [24, 26].

**Granger causality**

Granger causality methods determine if including the expression of *g'* at the previous time point improves our ability to predict the expression of *g* at the current time point above using the expression of *g* at the previous time point [27]. A common way to implement a Granger causality approach uses a vector autoregression (VAR) model, which usually assumes a linear relationship between all genes' previous expression and *g*'s current expression. (Fig S1B) [87].

A causal edge is included in the network if is significantly different from 0 for some *l*. While older VAR analyses did not fit the causal predictors simultaneously [87-89], newer analyses fit them simultaneously, using regularization techniques such as lasso [13, 90] or ridge regression [91] to handle the high dimensionality of genome-wide sequencing assays. Ensemble methods have also been used to improve the performance of lasso [47]. Random forests and nonlinear, kernel-valued functions have also been used to implement ideas in Granger causality [47, 92].

**Ordinary differential equations**

Ordinary differential equations (ODEs) fit the derivative of the expression of *g* as a function of all genes' expression at a single time point (Fig S1C) [11, 28, 29]:

Although complex dynamics are often nonlinear, ODE methods typically assume linearity, as small sample sizes make it challenging to infer the parameters of nonlinear functions. A causal edge is included in the network if *g'* has a significant coefficient in the ODE.

These methods are often combined with additional methods such as spline interpolation and piecewise linear functions to improve performance [28, 29].

**Decision trees**

Decision trees (DT) are a type of nonparametric function based on partitioning the data [30, 31]. DT methods fall either under VAR or ODE methods. Either the DTs fit the expression of *g* at the current time as a function of all genes' expression at the previous time point (VAR), or they fit the derivative of the expression of *g* as a function of all genes' expression at a single time point (ODE) (Fig S1D) [32, 33].

A causal edge is included in the network when an importance score for *g'*--- typically, the reduction in variance of *g* from including *g'* as a predictor--- exceeds some threshold. One limitation of DT methods is that they only produce a ranking of edges, without specifying the sign of the relationship between the genes [32, 33].

**Dynamic Bayesian networks**

Dynamic Bayesian networks (DBNs) search the space of possible directed acyclic graphs between previous and current expression levels and identify the network structure with the highest posterior probability of each edge given the data (Fig S1E) [34-38]. DBNs typically assume a linear relationship between previous expression values and current expression values. A causal edge is included in the network when its marginal posterior probability of existence exceeds some threshold. The joint probability of all the genes' expression across the time points 1 : *T* factorizes as:

Each gene's expression has a linear relationship with its parents:

While DBNs have been shown to be effective on smaller data sets [93], they scale poorly due to the superexponential growth of possible causal graph structures [58, 91]. Even after limiting the number of possible parents per gene to two, this results in cubic scaling of the search space. One exception is ScanBMA, which uses a pruning method based on Occam's window to limit the search space and gain a speedup [36].

**Gaussian process**

The Gaussian process (GP) is a distribution over continuous, nonlinear functions. GPs are often used in the context of nonlinear DBNs, where GP regression is used to model a nonlinear relationship between previous expression levels and current expression levels (Fig S1F) [39, 40]. A causal edge is included in the network based on its posterior probability of existence, i.e., the sum of the posterior probabilities of those networks that contain the edge. Each gene's expression has a nonlinear relationship with its parents:

By allowing nonlinear relationships between genes, GPs have proven highly effective. However, like DBN, they perform a search over causal graphs, and therefore suffer from the same scalability issues [39, 40].

**Sensitivity analysis of inferred network**

To assess the robustness of our inferred network to the choice of lag, we ran BETS with lag *L* = 1 on the glucocorticoid data. The discovered network has 2098 edges and shares 1286 edges with the lag *L* = 2 network's 31,945 edges. The overlap is significant (odds ratio (OR) = 393.9, Fisher's exact test (FET) ). We validated the discovered network on the GTeX eQTL study and find 76 trans-eQTLs in all samples corresponding to 24 network edges. 74 of the trans-eQTLs pairs are in lung samples, corresponding with 22 network edges (q-value FDR ). The remaining 2 trans-eQTLs were in adipose samples, corresponding to the 2 remaining network edges. There were 5 shared trans-eQTLs (4 network edges) between the 76 trans-eQTLs (24 network edges) discovered by lag *L* = 1 and 873 trans-eQTLs (341 network edges) discovered by lag *L* = 2. Thus, changing from lag 2 to lag 1, BETS still infers similar edges and discovers similar trans-eQTLs in lung tissue.

**Validation of inferred network on overexpression data**

Our analyses regressed network edge signs as predictors against the VAR model edge coefficients from the overexpression data as response (Fig 6). We sought to assess the strength of these associations across the 10 data sets, compared against shuffled edges. We compared the effect sizes of all 10 regressions of positive edge one-hot encodings on the overexpression coefficients with the effect sizes estimated similarly after shuffling the edge labels; we did the same for negative edges. At FDR , there was a substantial enrichment of effect sizes of positive edges among the original network (Common Language Effect Size (CLES) = 0.93, two-sided Mann-Whitney U-test (MWU) adjusted ); there was no enrichment of effect sizes for negative edges in the original network (CLES = 0.55, two-sided MWU adjusted ). Thus, the positive edges inferred by BETS validate on the overexpression data, but the negative edges do not, indicating repressive effects may have inconsistent signs.

**Validation of BETS' prediction of targets of overexpression genes**

We generated a set of pseudo-targets for each overexpressed transcription factor . For every gene , we use *nsgp* [94] to detect whether *g* is differentially expressed in the TF overexpression data set compared to the original exposure data set. The degree of *g*'s differential expression is measured by

Here, is the marginal log-likelihood of fitting one nonstationary GP for in the TF overexpression data set, and fitting an independent nonstationary GP for in the original exposure data set. is the marginal log-likelihood of fitting the same nonstationary GP over jointly in the TF overexpression data set and original exposure data set.

*nsgp* does not provide a framework for significance thresholding. We developed a procedure for generating a null distribution based on the null hypothesis is that *g* is not differentially expressed in the TF overexpression data set compared to the original overexpression data set. We replace the in the TF overexpression data set with a null version . is created from by randomly substituting an original exposure measurement for an overexpression measurement ofat every time point *t* in the overexpression data set. We then compute as if was the measurement in the TF overexpression data. We do this ten times, generating where n {1, 2, …, 10}.

We control the FDR at 0.01 by finding the threshold such that

For every gene , a directed edge exists (i.e., *g* is differentially expressed between the TF overexpression data set and original exposure data set) if .

Using this thresholding procedure, we find 443 targets of *CEBPB*, 318 targets of *CEBPD*, 223 targets of *FOSL2*, 342 targets of *FOXO1*, 19 targets of *FOXO3*, 767 targets of *KLF15*, 19 targets of *KLF6*, 212 targets of *KLF9*, 522 targets of *OCT4*, and 20 targets of *TFCP2L1*.

BETS is somewhat accurate at predicting the targets of *KLF15* (AUPR = 0.28), OCT4 (AUPR = 0.19), *CEBPB* (AUPR = 0.17), *FOXO1* (AUPR = 0.13), and *CEBPD* (AUPR = 0.11). It does not accurately predict the targets of *FOSL2*, *FOXO3*, *KLF6*, *KLF9*, and *TFCP2L1* (AUPR < 0.1).

Thus, BETS weakly predicts targets of 5 of the 10 over-expression TFs called by *nsgp*.

We generated a null by ranking every gene with a random score drawn from the uniform distribution on [0,1). We did this 10 different random seeds. The null has very similar predictive performance as BETS. The null somewhat accurately predicts the targets of *KLF15* (AUPR = 0.27-0.3), *OCT4* (AUPR = 0.18-0.21), *CEBPB* (AUPR = 0.15-0.17), *FOXO1* (AUPR 0.11-0.13), and *CEBPD* (AUPR = 0.11-0.12). It does not accurately predict the targets of *FOSL2*, *FOXO3*, *KLF6*, *KLF9*, and *TFCP2L1* (AUPR < 0.1).

Thus, random guessing also weakly predicts targets of 5 of the 10 over-expression TFs called by *nsgp*. BETS does not substantially improve over random guessing at FDR 0.1.