

# Supplementary Material

---

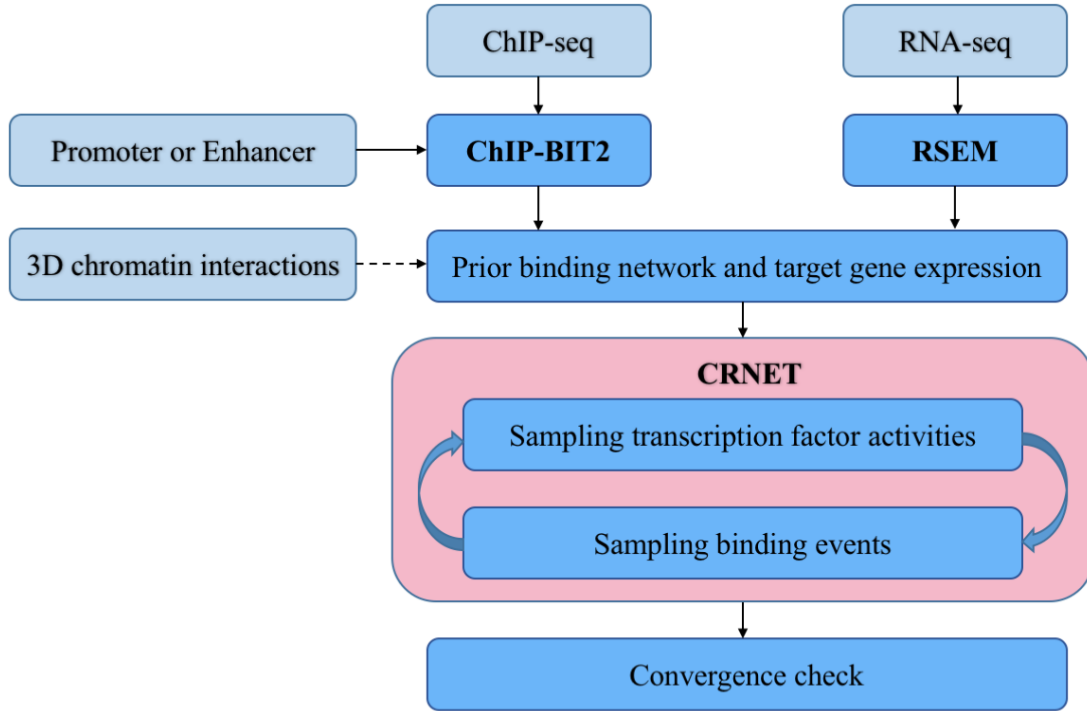
## CRNET: An efficient sampling approach to infer regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data

Xi Chen<sup>1</sup>, Jinghua Gu<sup>1</sup>, Xiao Wang<sup>1</sup>, Jin-Gyoung Jung<sup>2</sup>, Tian-Li Wang<sup>2</sup>, Leena Hilakivi-Clarke<sup>2</sup>, Robert Clarke<sup>2</sup>, and Jianhua Xuan<sup>1,\*</sup>

### Contents

<b>S1. Workflow of CRNET .....</b>	<b>2</b>
<b>S2. TFBS identification using ChIP-BIT2 .....</b>	<b>4</b>
<b>S3. CRNET model .....</b>	<b>6</b>
<b>S3.1 Enhancer-gene loop generation .....</b>	<b>6</b>
<b>S3.2 Gibbs sampling of TFA .....</b>	<b>6</b>
<b>S3.3 Logistic function parameter training for binding variable sampling .....</b>	<b>8</b>
<b>S3.4 Convergence check .....</b>	<b>9</b>
<b>S4. Simulation study .....</b>	<b>10</b>
<b>S4.1 Prior binding network simulation .....</b>	<b>10</b>
<b>S4.2 Gene expression data simulation .....</b>	<b>11</b>
<b>S4.3 Definitions of Precision, Recall and F-measure .....</b>	<b>11</b>
<b>S4.4 Convergence of Bayesian methods .....</b>	<b>12</b>
<b>S5. Large-scale FRN inference .....</b>	<b>13</b>
<b>S5.1 Prior binding network and candidate target gene expression .....</b>	<b>13</b>
<b>S5.2 Target gene validation for TFs (ATF3, EGR1 and SRF) in K562 cells .....</b>	<b>17</b>
<b>S6. Inference of FRNs in breast cancer MCF-7 cells .....</b>	<b>20</b>
<b>S6.1 Candidate target gene selection .....</b>	<b>20</b>
<b>S6.2 Prior binding matrix construction .....</b>	<b>21</b>
<b>S6.3 Convergence check of CRNET .....</b>	<b>24</b>
<b>S6.4 CRNET-estimated TFAs and their similarity to TF expression .....</b>	<b>27</b>
<b>S6.5 Validation of MYC's proximal or distal target genes .....</b>	<b>29</b>
<b>S7. Summary of data, tools and results .....</b>	<b>31</b>
<b>S8. Glossary of variables and parameters .....</b>	<b>32</b>

## S1. Workflow of CRNET



**Fig. S1.** A workflow of CRNET.

The main steps of the CRNET workflow (see Fig. S1) can be summarized as follows:

**Step 1:** CRNET can be run in either ‘promoter mode’ or ‘enhancer mode’. Promoter or enhancer regions should be pre-defined. We define a gene’s promoter region as  $\pm 10\text{k}$  bps around its transcription starting site (TSS) according to the UCSC hg19 RefSeq file. Enhancer regions can be identified using ChIP-seq data of enhancer markers like H3K27ac or DNase-seq data. For some cell types, their cell type-specific enhancer regions can be downloaded from ENCODE website (<https://www.encodeproject.org/data/annotations/>). We extend each enhancer region to  $\pm 1\text{k}$  bps around the middle point of the region.

**Step 2:** We use ChIP-BIT2 (<http://www.cbil.ece.vt.edu/software.htm>) to call transcription factor binding sites (TFBSs) at promoter or enhancer regions from ChIP-seq data of individual TFs. For each binding event, ChIP-BIT2 will report a probability denoting the possibility of binding occurrence. This probability is used as a prior in further functional regulatory network (FRN) inference. A brief description of ChIP-BIT2 can be found in Section S2. Alternatively, using ChIP-seq data with different peak callers or using other transcription factor database we may obtain binary binding information. CRNET can take binary prior input, too.

**Step 3:** Given time-course RNA-seq data, RSEM (<https://deweylab.github.io/RSEM/>) is used to align raw reads of each RNA-seq sample to the reference genome UCSC hg19 and estimate transcripts per million (TPM) value of each gene. We transfer the raw expression value to log2 format for further analysis. In this study, a candidate target gene is selected if at least at one time point its absolute fold change to the basal expression ('0' time point) is larger than 0.5. Further gene refinement like selecting a specific expression pattern or a biologically meaningful gene set is greatly helpful to narrow the candidate gene pool. To ensure reasonable computational time and parameter estimation accuracy, we suggest narrow the gene list to be shorter than 1000. CRNET has been tested with a few hundred to one thousand genes.

**Step 4:** For promoter study, binding events and target genes have already been linked together by ChIP-BIT2. If other peak calling is used, additional gene annotation needs to be done. We construct a prior binding matrix where each row represents a gene and each column represents a TF. For enhancer study, additional prior knowledge of 3D chromatin interactions is a must to map enhancer regions to target genes. Then, distal binding events can be linked together with target genes and a similar prior binding matrix as the promoter study can be established.

**Step 5:** To set the initial state of FRN, for each gene we select partial bindings (the number of selected bindings should be smaller than the number of total expression samples) according to their prior probabilities. And for each TF, we randomly select a Gaussian process with zero mean and unit variance as initial TF activity (TFA).

**Step 6:** CRNET is a two-stage Gibbs sampling approach. For each TF, we sample the mean and standard deviation of TFA (see Eqs. (7) and (8) of the main text) based on the conditional probability as defined in Eq. (5) in the main text. More details will be given in Section S3.2.

**Step 7:** For each binding event, we sample its state based on a t-score as defined in Eq. (7). Here, for sampling purpose, each t-score is transformed into a probability using logistic regression as Eq. (8). The logistic regression function will be further discussed later in Section S3.3.

**Step 8:** We repeat Steps 6 and 7 for sufficient times and collect Gibbs samples for each binding event. The sampling frequency (observed samples/total rounds) of each binding event represents the posterior probability for functional binding occurrence.

**Step 9:** As an optional step, users can run CRNET multiple times by varying initial settings in Step 5 and generate multiple Markov chains. As described in Section 3.4, sampling convergence on each variable can be monitored used a method proposed in (Gelman and Rubin, 1992).

## S2. TFBS identification using ChIP-BIT2

ChIP-BIT2 (an extended version of ChIP-BIT (Chen, et al., 2016)) uses a Gaussian mixture model (consisting of global and local Gaussian components) to capture both binding and background signals in the sample data. A unique feature of ChIP-BIT2 is that the Gaussian component modeling background signals is specially designed as a local Gaussian distribution that can be estimated accurately from the input data. Specific for promoter-focused studies, an exponential distribution is used to model the relative distance of TFBSs to TSS, which is further incorporated into the Bayesian approach of ChIP-BIT2 for target gene inference. Estimated by an expectation-maximization (EM) algorithm, a posterior probability is assigned to each TFBS under consideration, indicating the likelihood of a binding occurrence. A C++ package of ChIP-BIT2 can be downloaded from <http://www.cbil.ece.vt.edu/software.htm>.

Given ChIP-seq data of multiple TFs (the total number of TFs is denoted by  $T$ ) and annotation files of promoter or enhancer regions, a prior binding matrix  $\mathbf{B}$  can be constructed using ChIP-BIT2. Note that ChIP-BIT2 can be run in either ‘promoter mode’ or ‘enhancer mode’, but it cannot be run on both types of regions simultaneously because the mathematical models of the relative distance of TFBS to TSS are different under two cases. ChIP-BIT2 detects TFBSs reliably at promoter or enhancer regions by jointly modeling binding signals (read intensities) and binding locations of TFBSs.

Assuming that we have  $J$  promoter or enhancer regions and  $T$  TFs, for the  $j$ -th promoter or enhancer region, we partition it into small windows each with a size of 200 bps and calculate read intensities of the  $t$ -th TF’s ChIP-seq profile and its matched input data as  $s_{j,t,w}$  and  $s_{j,input,w}$  respectively for the  $w$ -th window (please see ChIP-BIT (Chen, et al., 2016) for read intensity calculation). For each window, the relative distance to the nearest TSS is recorded as  $d_{j,t,w}$ . Given the above observations, the conditional probability for binding event  $z_{j,t,w}$  is defined as

$$\begin{aligned} b_{j,t,w}(i) &= P(z_{j,t,w} = i | s_{j,t,w}, d_{j,t,w}) \\ &= \frac{1}{C} P(s_{j,t,w} | z_{j,t,w} = i) P(d_{j,t,w} | z_{j,t,w} = i) P(z_{j,t,w} = i), \quad i = 0, 1 \end{aligned} \quad (\text{S-1})$$

where  $C_{n,w} = \sum_{i=0,1} P(s_{j,t,w}, d_{j,t,w} | z_{j,t,w} = i) P(z_{j,t,w} = i)$  is a normalization factor.

The conditional probability  $P(s_{j,t,w} | z_{j,t,w})$  is a Gaussian mixture distribution with

$$\begin{cases} P(s_{j,t,w} | z_{j,t,w} = 1) \sim N(\mu_{TFBS}, \sigma_{TFBS}^2), \\ P(s_{j,t,w} | z_{j,t,w} = 0) \sim N(s_{j,input,w}, \sigma_{input}^2). \end{cases} \quad (\text{S-2})$$



For  $z_{j,t,w} = 1$ ,  $s_{j,t,w}$  is sequenced from a TFBS so it follows a global Gaussian distribution with mean  $\mu_{TFBS}$  and variance  $\sigma_{TFBS}^2$ ; for  $z_{j,t,w} = 0$ ,  $s_{j,t,w}$  is sequenced from background region so it follows a local Gaussian distribution with mean  $s_{j,input,w}$  (its input signal) and variance  $\sigma_{input}^2$ .  $\sigma_{input}^2$  is the variance of background signals, which can be directly calculated from the input data.

The second likelihood function in Eq. (S-1),  $P(d_{j,t,w} | z_{j,t,w})$ , is determined by the relative distance  $d_{j,t,w}$  to TSS as well as the binding state  $z_{j,t,w}$ . Note that this is a special feature for TFBSs at promoter regions. Previous studies (Chen, et al., 2016) have shown that the average read enrichment in sample ChIP-seq data (binding signals) follows an exponential distribution along the gene promoter region, while this distribution in input ChIP-seq data (background signals) is relatively uniform. Therefore, for TFBSs located at promoter regions,  $P(d_{j,t,w} | z_{j,t,w})$  is also a mixture distribution with

$$\begin{cases} P(d_{j,t,w} | z_{j,t,w} = 1) = \text{Exp}(\lambda_t), \\ P(d_{j,t,w} | z_{j,t,w} = 0) = U(-d_p, d_p). \end{cases} \quad (\text{S-3})$$

where  $\lambda_t$  is the exponential distribution parameter and  $d_p$  (=10k bps) is the half length of a promoter region. For enhancer studies, we do not need to estimate parameter  $\lambda_t$ . Instead for TFBS prediction at enhancer regions, we assume that both binding signals and background signals follow a uniform distribution as defined in Eq. (S-3).

ChIP-BIT2 iteratively estimates distribution parameters using an Expectation-Maximization (EM) algorithm and finally estimates a probability  $b_{j,t,w}$  for each candidate binding event  $z_{j,t,w} = 1$ . It is possible that, for a promoter or enhancer region, there are multiple windows containing TFBSs of the same TF. In that case, we select the largest  $b_{j,t,w}$  as  $b_{j,t}$  to denote the probability of binding occurrence between the  $t$ -th TF and the  $j$ -th promoter or enhancer region. Finally, we report all binding events with a probability larger than 0.5 and construct a weighted prior binding matrix **B** with  $J$  rows and  $T$  columns.

ChIP-BIT2 is developed to detect TFBSs from a set of pre-selected enhancer or promoter regions. Therefore, users are required to identify or predict a list of active enhancer or promoter regions before using ChIP-BIT2. In this paper, we select two widely used human enhancer and promoter database: ENCODE cell type-specific enhancer regions and hg19 RefSeq promoter regions (+/- 10k bps around TSS). Note that optimal selection of enhancer or promoter regions for a certain cell type is helpful to improve the accuracy of functional binding inference by eliminating noisy prior bindings from false positive enhancer or promoter regions.

### S3. CRNET model

#### S3.1 Enhancer-gene loop generation

A distal enhancer region will physically interact with the promoter region of a target gene through a 3D chromatin interaction and then, TFs binding at enhancer regions will actively regulate the target gene expression (Sanyal, et al., 2012). However, there is no clear evidence for the functional relationship between TFs from different enhancers looping to the same gene, and also no clear relationship between TFs binding at the two ends in the same loop. Therefore, we cannot make a hard claim that those TFs work at the same time or hierarchically. Therefore, using CRNET we predict FRNs either at promoter or enhancer regions. In current framework, we do not integrate binding events at enhancer and promoter region together.

3D chromatin interactions between enhancers and target genes can be extracted from ChIA-PET data or Hi-C data. The difference between these two types of data is that the ChIA-PET technique is more like ChIP-seq technique and it can provide TF-specific enhancer-promoter loops, which can be extracted from ChIA-PET data using Mango (<https://github.com/dphansti/mango>); the Hi-C technique is more like the whole genome sequencing technique and it can provide topological association domains (TADs) across the whole chromosome, with which the enhancer-promoter loops can be extracted from Hi-C data using HiC-Pro (<https://github.com/nservant/HiC-Pro>). The resolution of ChIA-PET results is usually higher than those interactions provided by Hi-C and they are also much sparser. We map 3D chromatin interactions to enhancer and promoter regions used in ChIP-BIT2. If one end of an interaction is mapped to an enhancer region (2kbps long) with at least 500 bps overlap and the other end is mapped to a promoter region (20kbps long) with at least 500 bps overlap, it will be annotated as an enhancer-promoter loop. Based on annotated enhancer-promoter (gene) loops and binding observations at enhancer regions, we construct a prior binding matrix **B** specific for enhancer studies, where each row of **B** denotes an enhancer-gene association (enhancerID:geneID) and each column represent a TF.

#### S3.2 Gibbs sampling of TFA

The posterior probability of hidden variable **X** is defined as follows:

$$P(\mathbf{X} | \mathbf{Y}, \mathbf{A}, \mathbf{Z}) \propto \prod_t \prod_j \prod_m \frac{1}{\sigma} \frac{1}{\sigma_x} \exp \left( -\frac{1}{2\sigma^2} \left( y_{j,m} - \sum_t a_{j,t} z_{j,t} x_{t,m} - \eta_j \right)^2 \right) \exp \left( -\frac{1}{2\sigma_x^2} x_{t,m}^2 \right). \quad (\text{S-4})$$

We assume conditional independence among units in matrix **X** and hence the posterior probability of each variable  $x_{t,m}$ , the hidden activity of the  $t$ -th TF at the  $m$ -th time point or condition, can be calculated as follows:

$$P(x_{t,m} | \mathbf{Y}, \mathbf{A}, \mathbf{Z}) \propto \prod_{j=1}^J \frac{1}{\sigma} \exp \left( -\frac{1}{2\sigma^2} \left( y_{j,m} - \sum_{t'=1}^T a_{j,t'} z_{j,t'} x_{t',m} - \eta_j \right)^2 \right) \frac{1}{\sigma_x} \exp \left( -\frac{1}{2\sigma_x^2} x_{t,m}^2 \right) \quad (\text{S-5})$$

Equation (S-4) is a multiplication of two Gaussian distributions so the posterior probability of  $x_{t,m}$  is still a Gaussian distribution. Equation (S-5) can be further expanded as follows:

$$\begin{aligned} P(x_{t,m} | \mathbf{Y}, \mathbf{A}, \mathbf{Z}) &\propto \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^J \left( y_{j,m} - \sum_{t'=1}^T a_{j,t'} z_{j,t'} x_{t',m} - \eta_j \right)^2 - \frac{J}{2\sigma_x^2} x_{t,m}^2 \right) \\ &\propto \exp \left( -\frac{1}{2\sigma^2} \sum_{j=1}^J \left( \left( y_{j,m} - \sum_{t'=1, t' \neq t}^T a_{j,t'} z_{j,t'} x_{t',m} - \eta_j \right) - a_{j,t} z_{j,t} x_{t,m} \right)^2 - \frac{J}{2\sigma_x^2} x_{t,m}^2 \right) \\ &\propto \exp \left( -\left( \frac{1}{2\sigma^2} \sum_{j=1}^J (a_{j,t} z_{j,t})^2 + \frac{J}{2\sigma_x^2} \right) x_{t,m}^2 + \frac{1}{\sigma^2} \sum_{j=1}^J \left( \left( y_{j,m} - \sum_{t'=1, t' \neq t}^T a_{j,t'} z_{j,t'} x_{t',m} - \eta_j \right) a_{j,t} z_{j,t} \right) x_{t,m} \right) \end{aligned} \quad (\text{S-6})$$

We define two new variables  $\sigma_x'^2$  and  $\mu_x'$  as follows:

$$\sigma_x'^2 = \left( \frac{1}{\sigma^2} \sum_{j=1}^J (a_{j,t} z_{j,t})^2 + \frac{J}{\sigma_x^2} \right)^{-1} \quad (\text{S-7})$$

$$\mu_x' = \frac{\sum_{j=1}^J \left( \left( y_{j,m} - \sum_{t'=1, t' \neq t}^T a_{j,t'} z_{j,t'} x_{t',m} - \eta_j \right) a_{j,t} z_{j,t} \right)}{\sum_{j=1}^J (a_{j,t} z_{j,t})^2 + \frac{J\sigma^2}{\sigma_x^2}} \quad (\text{S-8})$$

Bring  $\sigma_x'^2$  and  $\mu_x'$  back to Equation (S-3) and then we can obtain a new equation as follows:

$$\begin{aligned} P(x_{t,m} | \mathbf{Y}, \mathbf{A}, \mathbf{Z}) &\propto \exp \left( -\frac{1}{2} \frac{1}{\sigma_x'^2} (x_{t,m}^2 - 2\mu_x' x_{t,m}) \right) \\ &\propto \frac{1}{\sqrt{2\pi\sigma_x'^2}} \exp \left( -\frac{(x_{t,m} - \mu_x')^2}{2\sigma_x'^2} \right) \end{aligned} \quad (\text{S-9})$$

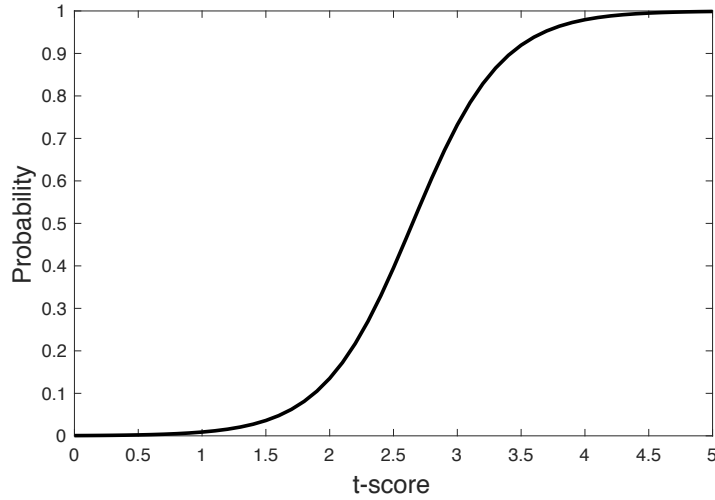
where  $\frac{1}{\sqrt{2\pi\sigma_x'^2}}$  is a constant for  $x_{t,m}$ . From Equation (S-9) it can be found that the posterior probability of  $x_{t,m}$  follows a Gaussian distribution with mean  $\mu_x'$  and variance  $\sigma_x'^2$ , which can be computed as in Equations (S-7) and (S-8).

### S3.3 Logistic function parameter training for binding variable sampling

Converting the Student's t-statistic value (t-score) for each binding into a probability we define a logistic function in Eq. (8) of the main text. Before running CRNET, we need to estimate the logistic function parameters via a training procedure. The training procedure is summarized as follows:

- 1) Randomly select a number of TFs (smaller than that of gene expression samples) and run Network Component Analysis (NCA) (Liao, et al., 2003) to estimate hidden TFAs and regulation strengths;
- 2) For each binding event, calculate the t-statistic ( $f(a_{j,t})$  as defined in Eq. (10) of the main text) using its regulation strength and regression error, and make a decision as functional binding ( $z=1$ ) if the t-statistic value is larger than  $t_{0.05}$  (the t-statistic value corresponding to a false positive rate $<0.05$ ); otherwise as non-functional binding ( $z=0$ );
- 3) Record t-statistic values and their associated binding decisions/labels ( $z=0$  or  $z=1$ );
- 4) Repeat Step 1) to Step 3) 100 times to test all TFs in a sufficient number of times;
- 5) Run a logistic regression using all t-statistic values and their binding labels.

From the training procedure, the logistic function parameters  $b_0$  and  $b_1$  are estimated as -15.15 and 5.72, respectively. The logistic function curve is shown in Fig. S2.



**Fig. S2.** Logistic function curve.

### S3.4 Convergence check

If benchmark network is available, convergence to those ‘true’ interactions can be directly monitored by examining Precision-Recall performance. For real data analysis without much knowledge about ‘true’ interaction, convergence of Gibbs sampling can also be monitored based on the ratio (R) of within variance and between variance using multiple sequences with different initial states (Gelman and Rubin, 1992). Here is a brief outline of the calculation of R for each variable estimated using Gibbs sampling.

First, we simulate  $S \geq 2$  sequences independently, each of length  $K$ .  $S$  and  $K$  are set up to 10 and 1000, respectively. This relatively short sequence is mainly used to check when the chains start to converge.

Second, after each round of sampling, i.e.,  $k$ , for each variable  $v(s, k)$ , we update the target mean, which is the mean of the  $S \times k$  sampled values, as  $\bar{v}$ .

Third, we calculate between variance  $B$  and the average value of within variance  $W$  as follows:

$B/k$  = the variance between the  $K$  sequence means,  $\bar{v}(s)$ , each based on  $k$  values of  $v(s, k)$ ,

$$\frac{B}{k} = \frac{\sum_{s=1}^S (\bar{v}(s) - \bar{v})^2}{S-1}, \quad (\text{S-10})$$

$W$  = the average of  $S$  within-sequence variances,  $\gamma_k^2$ , each based on  $k-1$  degrees of freedom,

$$W = \frac{\sum_{s=1}^S \gamma_k^2}{S}. \quad (\text{S-11})$$

Fourth, we estimate the target variance by a weighted average of  $W$  and  $B$  as follows:

$$\hat{\sigma}^2 = \frac{k-1}{k} W + \frac{1}{k} B. \quad (\text{S-12})$$

Fifth, we estimate what is now known about  $v$ . The result is an approximate Student’s t-distribution for  $v$  with center  $\bar{v}$ , scale  $\sqrt{\hat{V}} = \sqrt{\hat{\sigma}^2 + B/Sk}$  and degrees of freedom  $df = 2\hat{V}^2 / \text{var}(\hat{V})$ .

Sixth, we monitor convergence of sampled Markov Chain by estimating the factor  $\sqrt{\hat{R}}$ , which the scale of the current distribution for  $x$  might be reduced if the simulations are continued in the limit  $n \rightarrow \infty$ . This potential scale reduction is estimated by

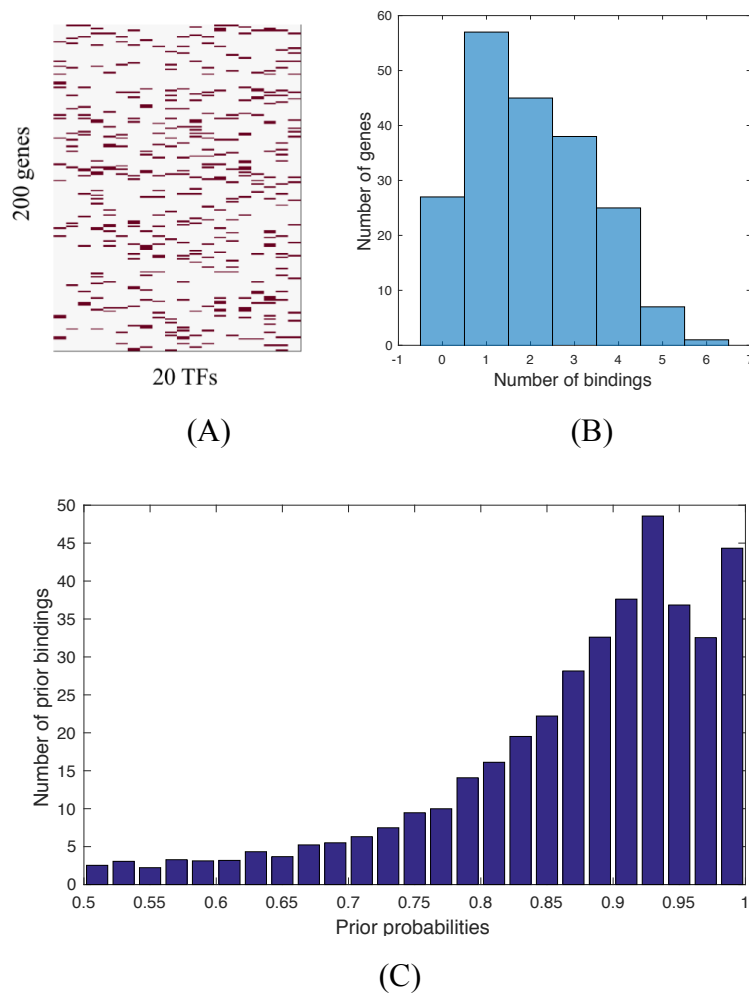
$$\hat{R} = \frac{\hat{V}}{W} \frac{df}{df-2} = \left( \frac{k-1}{k} + \frac{S+1}{kS} \frac{B}{W} \right) \frac{df}{df-2} \quad (\text{S-13})$$

which declines to 1 as  $k \rightarrow \infty$ .

## S4. Simulation study

### S4.1 Prior binding network simulation

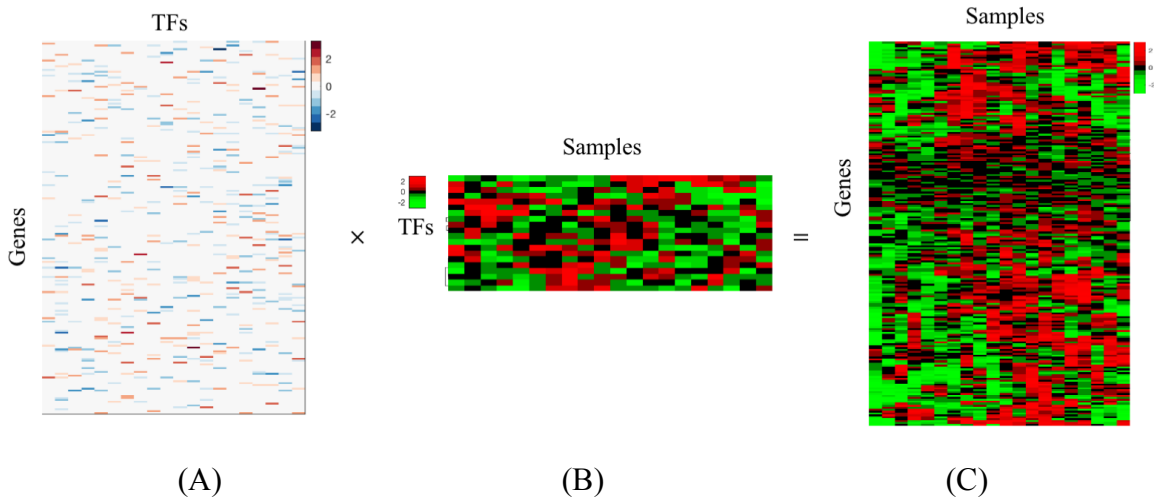
We simulated a regulatory network with 200 genes and 20 TFs as shown in Fig. S3(A). Each gene is regulated by a random number of TFs from 0 to 6 (on average 2), as shown in Fig. S3(B). The prior probability of each binding is a random value from 0.5 to 1 following a distribution as observed from ChIP-BIT2 results of analyzing ENCODE MCF-7 ChIP-seq data, as shown in Fig. S3(C).



**Fig. S3.** Simulated prior binding network. (A) Network structure ('red' color denotes bindings); (B) a histogram of binding degree on each gene; (C) a histogram of binding prior probabilities on simulated bindings (following a distribution of ChIP-BIT2 results of ENCODE MCF-7 cell ChIP-seq data).

## S4.2 Gene expression data simulation

According to the distribution assumption made on regulation strength by existing Bayesian methods like BNCA and COGRIM, we simulated the regulation strength (matrix A) on each binding connection by following a Gaussian distribution with zero mean and unit deviation. We simulated two different time course gene expression datasets: in Case 1, we simulated TFAs (matrix X) for individual TFs using Gaussian random process with zero mean and unit variance under 20 time points; gene expression (matrix Y) was then simulated based on the log-linear model introduced in Eq. (1) with X and A; in Case 2, we only generated 10 time course samples. Case 2 is more challenging because the number of TFs is larger than the number of gene expression samples.



**Fig. S4.** Simulated gene expression data. (A) Regulatory strength of prior binding events (matrix A); (B) a heatmap of simulated TFA (matrix X); (C) a heatmap of simulated gene expression data (matrix Y).

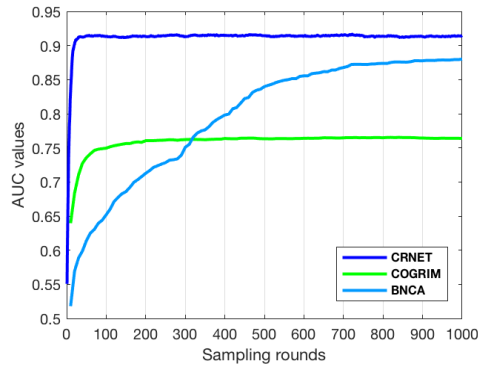
## S4.3 Definitions of Precision, Recall and F-measure

$$Precision = \frac{\text{Number of true positive edges}}{\text{Number of selected edges over threshold}}$$

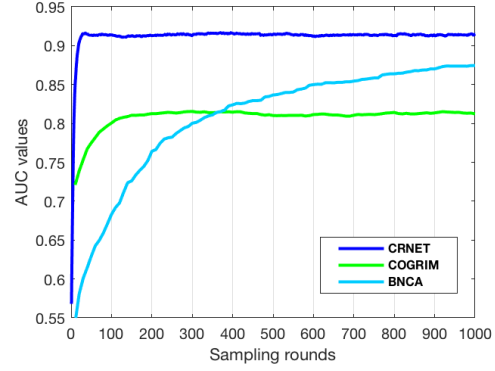
$$Recall = \frac{\text{Number of true positive edges}}{\text{Number of ground truth edges}}$$

$$F - \text{measure} = \frac{2 * precision * recall}{precision + recall}$$

#### S4.4 Convergence of Bayesian methods



(A)



(B)

**Fig. S5.** Convergence of competing methods. (A) AUC at different rounds of sampling for Case 1 with FPR = 15% and SNR = 3dB; (B) AUC performance at different rounds of sampling for Case 2 with FPR = 15% and SNR = 3dB.



## S5. Large-scale FRN inference

### S5.1 Prior binding network and candidate target gene expression

**K562 cell:** ChIP-seq data of 228 TFs generated from K562 cells was downloaded from ENCODE database. We used ChIP-BIT2 to process BAM files of each TF and its match input and further constructed a prior binding matrix at promoter region. A time-course gene expression dataset (GSE1036) of K562 cells was downloaded from GEO database. In this dataset, K562 cells were treated with 50 mM hemin for 72 hours and at each time point (0h, 6h, 12h, 24h, 48h, 72h) two replicates were generated using Affymetrix Human Genome U133A Array. We downloaded the normalized gene expression data as described in (Addya, et al., 2004) and estimated the Signal-to-Noise Ratio (SNR) of each gene expression sample relative to the basal expression at ‘0’ time point (control) using SNAGEE (Venet, et al., 2012), as shown in Table S1. On average the SNR of this gene expression dataset is 2.82dB. As shown in previous simulation studies, under this SNR range around 3dB CRNET can achieve a F-measure higher than 0.7. Following the gene selection procedure as described in (Addya, et al., 2004), 1,569 differentially expressed genes were selected. We selected genes with at least two binding observations from the prior binding matrix. Finally, 1,351 genes expression were selected. The prior binding matrix and normalized gene expression data can be found from Table S2. A heatmap of time-course gene expression is shown in Fig. S7(A).

**Table S1.** Relative SNRs of individual samples in the GSE1036 dataset.

Sample_ID	Relative SNR
0h_rep1	0
0h_rep2	0.19
6h_rep1	3.08
6h_rep2	3.92
12h_rep1	3.65
12h_rep2	3.51
24h_rep1	3.48
24h_rep2	3.99
48h_rep1	3.71
48h_rep2	2.32
72h_rep1	1.45
72h_rep2	1.74

**Table S2.** Prior binding matrix and normalized gene expression data of K562 cells.

Table S2 can be found in the supplementary file: “Supplementary Material Table S2.xlsx”.

**GM12878 cell:** ChIP-seq data of 122 TFs generated from GM12878 cells was downloaded from ENCODE database. We used ChIP-BIT2 to process BAM files of each TF and its match input and construct a prior binding matrix at promoter region. A time-course gene expression dataset of GM12878 cells was downloaded from GEO data base with access number GSE51709. In this dataset, GM12878 cells were treated with 0.5  $\mu$ M doxorubicin (Calbiochem) for 0h, 4h and 18h and at each time point three replicates were generated using Affymetrix Human Exon 1.0 ST arrays. We downloaded the gene expression data analyzed through Affymetrix Expression Console using gene-level RMA summarization and sketch-quantile normalization methods (Su, et al., 2015). The relative SNR of each gene expression sample is shown in Table S3. SNRs vary from 1.55dB to 2.97dB with an average value of 2.04dB. Differentially expressed genes were identified at each time point using t-test with corrected  $p$ -value  $<0.05$  and absolute fold change  $>2$ . We refined gene list by selecting genes with at least two binding observations from the prior binding matrix. Finally, 925 genes were selected for FRN inference. The prior binding matrix and normalized gene expression data can be found in Table S4. A heatmap of time-course gene expression is shown in Fig. S7(B).

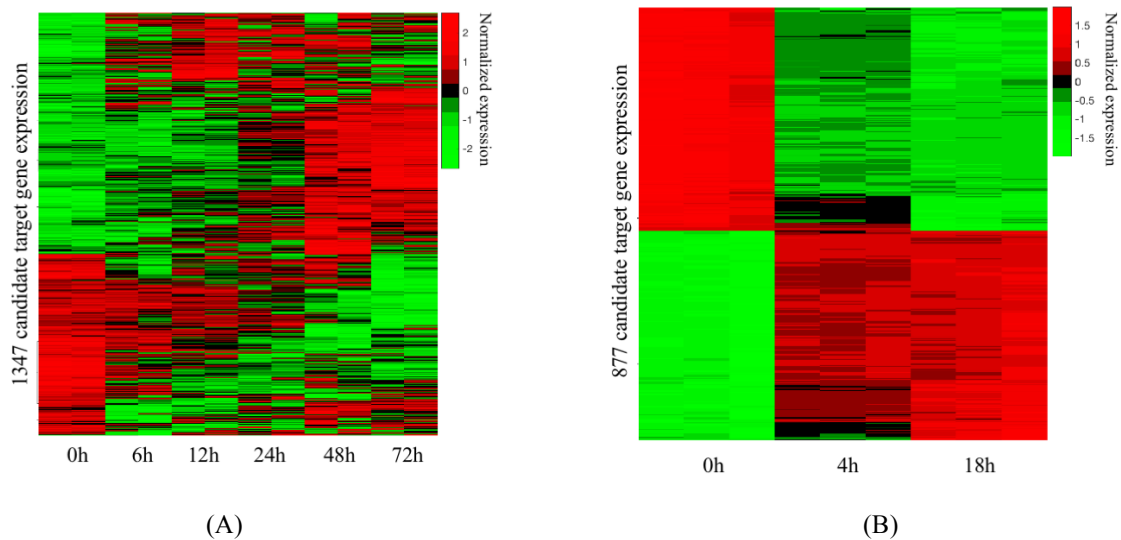
**Table S3.** Relative SNRs of individual samples in the GSE51709 dataset.

Sample_ID	Relative SNR
0h_rep1	0
0h_rep2	-0.4051867
0h_rep3	2.2919465
4h_rep1	2.6966258
4h_rep2	2.9733475
4h_rep3	2.9664372
18h_rep1	2.4840261
18h_rep2	1.5534577
18h_rep3	1.8033868

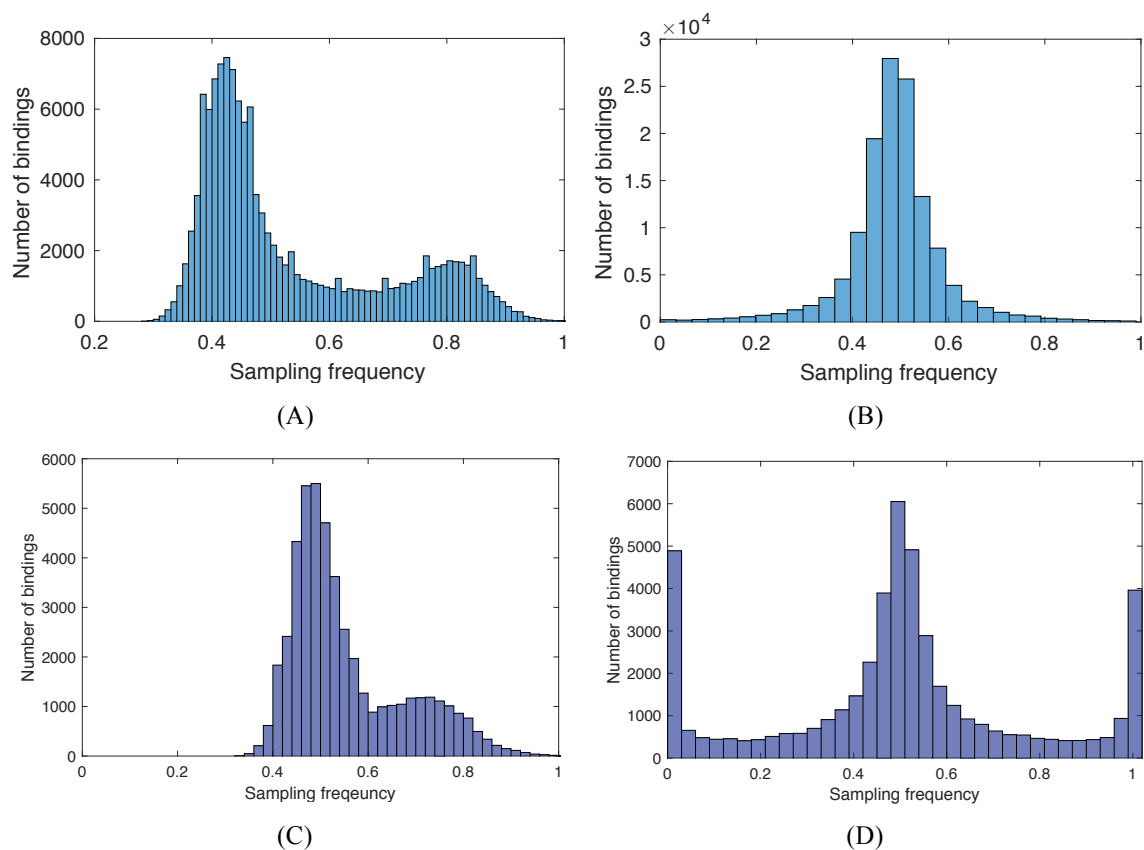
**Table S4.** Prior binding matrix and normalized gene expression data of GM12878 cells.

Table S4 can be found in the supplementary file: “Supplementary Material Table S4.xlsx”.

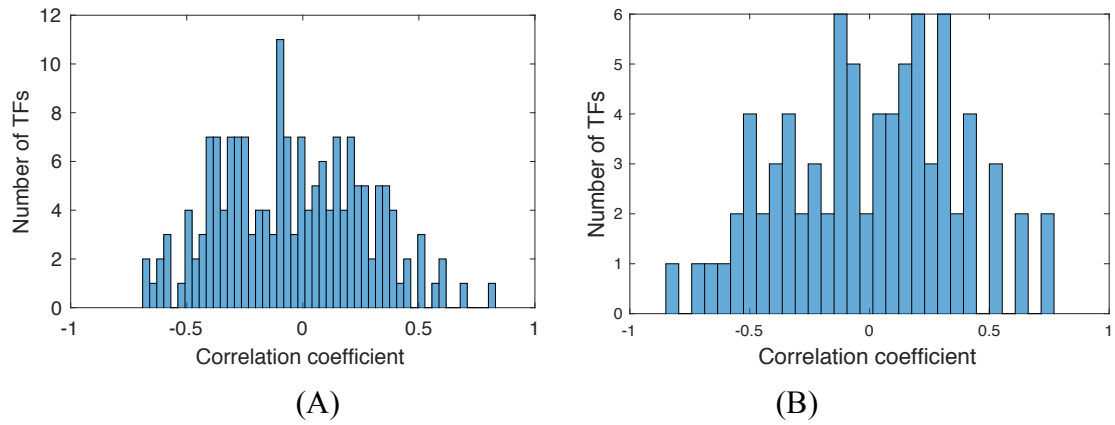
Weighted binding matrix



**Fig. S7.** Heatmaps of selected genes expression: (A) a heatmap of 1351 genes in the K562 GSE1036 dataset; (B) a heatmap of 925 genes in the GM12878 GSE51709 dataset.



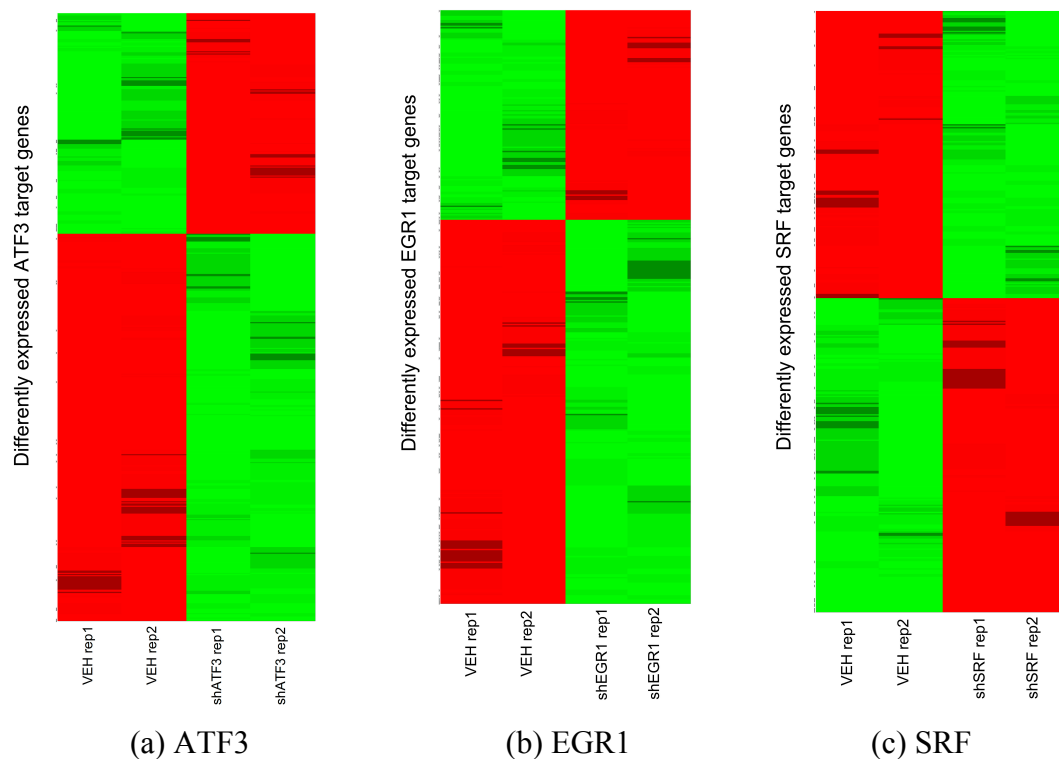
**Fig. S8.** Posterior distributions of Gibbs samples (sampling frequency) generated by CRNET and COGRAM: (A) CRNET results for K562 cells; (B) COGRIM results for K562 cells; (C) CRNET results for GM12878 cells; (D) COGRIM results for GM12878 cells.



**Fig. S9.** Correlation coefficient of CRNET estimated TFA and observed TF expression: (A) 173 TFs in K562 cells; (B) 80 TF in GM12878 cells.

## S5.2 Target gene validation for TFs (ATF3, EGR1 and SRF) in K562 cells

We use the matched RNA-seq data generated before or after specific TF knockdown to validate genes identified by each method on each data set. The RNA-seq data is downloaded from GEO data base under access number GSE33816. For each TF, there are two replicates under control or treatment conditions (Vehicle vs. shRNA). We apply RSEM (<https://deweylab.github.io/RSEM/>) to the fastq files of each RNA-seq sample and estimate read counts and transcripts per million (TPM) values of genes across all samples. For each TF, we use DESeq2 (<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>) to identify its target genes (differentially expressed genes with a q-value cutoff as 0.05). In total, we identified 1133 genes for ATF3, 893 genes for EGR1 and 1011 genes for SRF, whose expression patterns are shown in Fig. S10. Note that as a baseline for comparison, TIP (Cheng, et al., 2011) (a probabilistic method using binding data alone) was applied to each ChIP-seq data set for target gene prediction. 13.61%, 9.42% or 11.45% of its predicted target genes are also differentially expressed when ATF3, EGR1 or SRF is knocked down.



**Fig. S10.** Heat map of differentially expressed target genes after knocking down each specific TF.

**Table S5.** Validation of ATF3, EGR1 and SRF's target genes using shRNA knockdown experiments

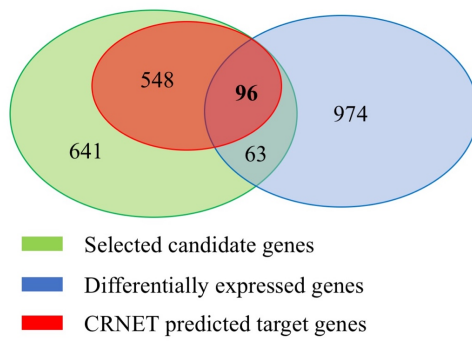
<b>TF symbols</b>		<b>ATF3</b>	<b>EGR1</b>	<b>SRF</b>
<b>CRNET (ChIP-BIT2 weighted prior)</b>	Number of predicted target genes	644	296	112
	Number of 'true' target genes	96	33	14
	Validation rate	<b>14.9%</b>	<b>11.15%</b>	<b>12.5%</b>
	Enrichment p-value	2.28e-4	5e-2	3.27e-2
<b>CRNET (confident binary prior)</b>	Number of predicted target genes	141	249	62
	Number of 'true' expressed target genes	20	26	7
	Validation rate	<b>14.8%</b>	<b>10.44%</b>	<b>11.3%</b>
	Enrichment p-value	2.92e-3	2.38e-2	3.25e-2
<b>COGRIM</b>	Number of predicted target genes	240	354	95
	Number of 'true' expressed target genes	20	26	7
	Validation rate	8.33%	7.34%	7.37%
	Enrichment p-value	0.35	0.528	0.197
<b>LASSO</b>	Number of predicted target genes	159	266	101
	Number of 'true' expressed target genes	20	27	7
	Validation rate	12.58%	10.15%	6.93%
	Enrichment p-value	1.19e-2	6.60e-2	0.242
<b>GENIE3</b>	Number of predicted target genes	426	427	154
	Number of 'true' expressed target genes	20	26	6
	Validation rate	4.69%	6.09%	3.90%
	Enrichment p-value	0.997	0.176	0.476

#### Hypergeometric p-value calculation:

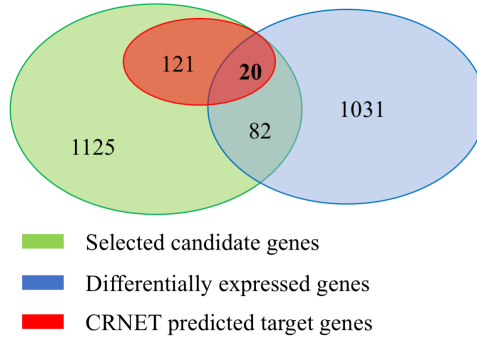
In total, we have 1348 candidate target genes. They are bound by at least one TF in K562 cells. 159 genes (96+63 in the middle) are validated as ATF3 target genes since they are significantly differentially expressed when ATF3 is knocked down. Using CRNET we finally identify 644 target genes for ATF3, where 96 of them are validated. We calculate the  $p$ -value using hypergeometric test as follows:

$$P(X \geq k) = \sum_{x=k}^K \binom{K}{x} \binom{N-K}{n-x} / \binom{N}{n},$$

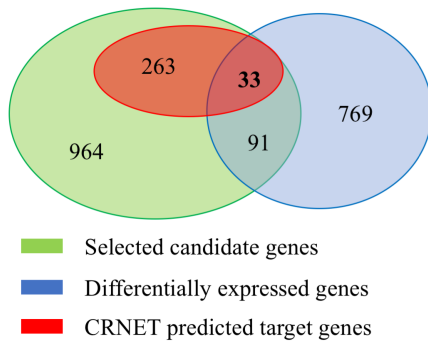
where  $N$  is the number of candidate target genes,  $K$  is the number of all validated genes,  $n$  is the number of genes identified by CRNET, and  $k$  is the number of validated genes in the CRNET results. In this case,  $N = 1348$ ,  $K = 159$ ,  $n = 644$  or  $k = 96$ . The significance  $p$ -value is 2.28e-4.



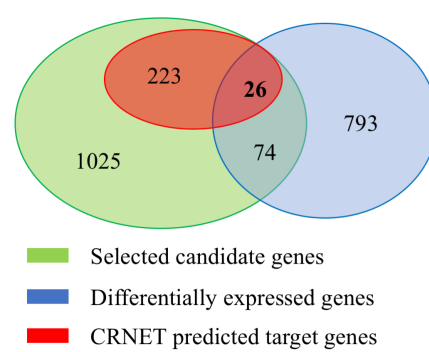
(A)



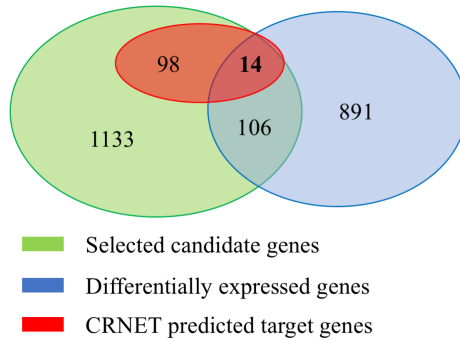
(B)



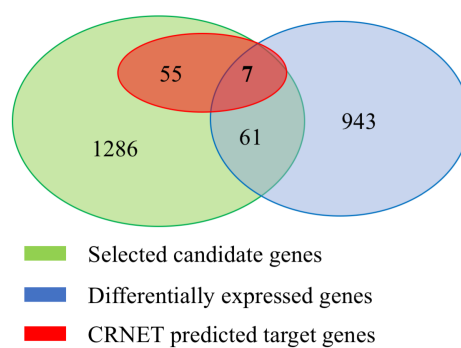
(C)



(D)



(E)



(F)

**Fig. S11.** Venn diagrams of selected candidate genes, differentially expressed genes and CRNET predicted genes. (A), (C) and (E) represent gene validation for ATF3, EGR1 and SRF, respectively, where all ChIP-BIT2-detected binding events are used; (B), (D) and (F) represent gene validation for ATF3, EGR1 and SRF, respectively, where ChIP-BIT2-detected bindings events of high confidence (probability > 0.85) are used. The overlap area between 'green' and 'blue' circles represents validated target genes for each study, while the overlap of three circles represents validated CRNET-predicted targets.

## S6. Inference of FRNs in breast cancer MCF-7 cells

### S6.1 Candidate target gene selection

A breast cancer MCF-7 RNA-seq dataset was downloaded from the GEO database (accession number: GSE62789). This is a time-course dataset including 10 samples generated within 24 hours of 10nM 17 $\beta$ -estradiol (E2) treatment (one sample for 0min, 5min, 10min, 20min, 40min, 80min, 160min, 320min, 640min or 1,280min). The transcripts per million (TPM) values of genes (see Supplementary Table S6) were estimated using RSEM (<https://deweylab.github.io/RSEM/>). Using the basal expression at 0min as control, the relative SNR of each sample is estimated by SNAGEE (Venet, et al., 2012), as shown in Table S7. The average SNR is 2.83dB.

**Table S6.** RSEM estimated TPM values for genes from UCSC RefSeq hg19.

Table S6 can be found in the supplementary file: “Supplementary Material Table S6.xlsx”.

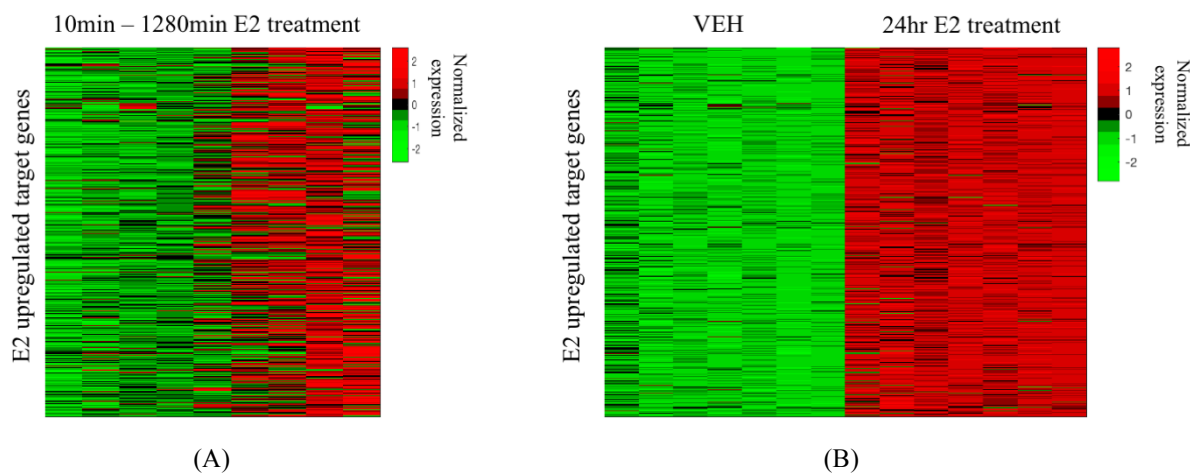
**Table S7.** Relative SNRs of individual samples in the GSE51709 dataset.

Sample ID	Relative SNR
0min	0
5min	4.23
10min	3.00
20min	3.12
40min	2.70
80min	2.33
160min	0.90
320min	1.74
640min	3.05
1280min	4.38

Following the RNA-seq processing pipeline described in the Methods section, we identified 3,258 significantly up/down regulated genes. Breast cancer MCF-7 cells are E2 sensitive so under E2 stimulation, MCF-7 cells will grow very fast with cell cycle, cell proliferation and cell growth signaling pathway activated. Using this E2 treated MCF-7 cell line model, it is reasonable to associate TFs with up activity and their target genes with up-regulation pattern to those E2 stimulated functional pathways, which have been demonstrated to be involved in breast cancer development. Furthermore, if we knockdown a TF with up activity using siRNA, theoretically its direct target genes should show a down-regulated gene expression pattern, as opposite to their original up-regulated expression under E2 treatment. Specific for MCF-7 cells and the E2 treatment condition, we were only focused on 1,907 up-regulated target genes and aimed to infer the FRNs regulating those genes.



To narrow the search space and gain more confidence on gene selection, we downloaded another steady-state RNA-seq dataset from GEO database (accession number: GSE51403). In this dataset, there are seven RNA-seq replicates generated under vehicle (VEH) condition and another 7 replicates generated after 24 hours of 10nM E2 treatment. Read counts of each gene under in total 14 samples were estimated using RSEM and differential expression analysis was performed using DeSeq2 (<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>) with a cut-off q-value 0.05 and fold change > 0.5 (higher expression after E2 treatment). Finally, we have identified 489 common target genes. Heatmaps of gene expression pattern in both datasets are shown in Fig. S12.



**Fig. S12.** E2 upregulated candidate target gene expression pattern. (A) A heatmap of 489 common genes in E2 treated MCF-7 cell time-course RNA-seq data (GSE62789); (B) a heatmap of 489 common genes in E2 treated MCF-7 cell steady-state RNA-seq data (GSE51403);

## S6.2 Prior binding matrix construction

**Table S8.** 39 TF ChIP-seq profiles of breast cancer MCF-7 cells.

Data source	TF symbols
ENCODE	CEBPB, CTCF, E2F1, EGR1, ELF1, EP300, FOSL2, FOXM1, GABPA, GATA3, HDAC2, JUND, MAX, MYC, NR2F2, NRSF, PML, POLR2A, RAD21, SIN3AK20, SRF, TAF1, TCF7, TCF12, TEAD4, ZNF217
GSE26831	c-FOS, c-JUN, FOXA1
GSE41561	CREB1, ER- $\alpha$ , KLF4, RXRA, TLE3
GSE38901	HSF1
GSE44737	MBD3
GSE28008	PBX1
GSE22612	TDRD3

Promoter regions were extracted from human reference genome hg19 as  $\pm 10$ k bps around each TSS. In total, we obtained 25,802 promoter regions regardless of potential overlap. We downloaded breast cancer MCF7 enhancer like regions from ENCODE (<https://www.encodeproject.org/data/annotations/>). In total, we obtained non-overlap 33957 enhancer regions. We extended or pruned enhancer regions as  $\pm 1$ k bps around the original middle points. Then, we used ChIP-BIT2 to call TFBSs at promoter and enhancer regions, respectively. MACS2 was also applied to the same ChIP-seq data with default setting for narrow peak detection and detected peaks were further mapped to promoter or enhancer regions. We selected target genes with promoter binding events first. 464 genes (among the genes in Fig. S12) having at least one TF binding event were selected. Prior bindings and gene expression for these 464 genes can be found in Table S9, which were used for promoter FRN inference. We also presented MACS2 results together with ChIP-BIT2 in Fig. S13. We found that there were several TFs without many peaks predicted by MACS2. For example, TDRD3 had a very few MACS2-detected peaks. However, after evaluating read intensities using ChIP-BIT2, most of them were significantly higher than the matched input data and weak peaks (probability  $> 0.85$ ) could be identified on 444 candidate gene promoters. For another transcription factor MBD3, using MACS2-detected peaks we could only obtain 170 candidate target genes, but using ChIP-BIT2 we had obtained 398 genes whereas 268 of them were of a probability over 0.85.

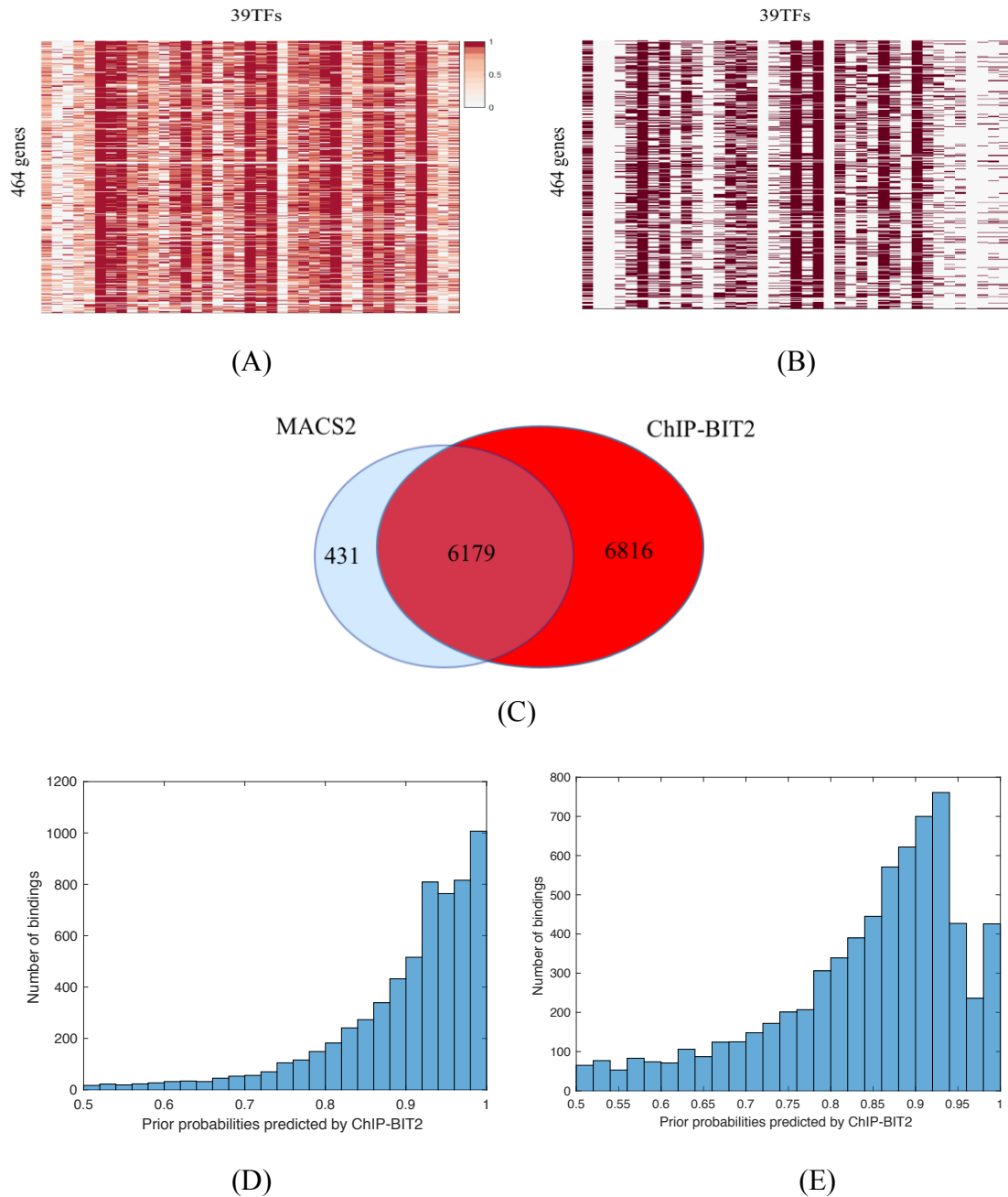
**Table S9.** Prior binding matrix (promoter) and normalized gene expression data of MCF7 cells.

Table S9 can be found in the supplementary file: “Supplementary Material Table S9.xlsx”.

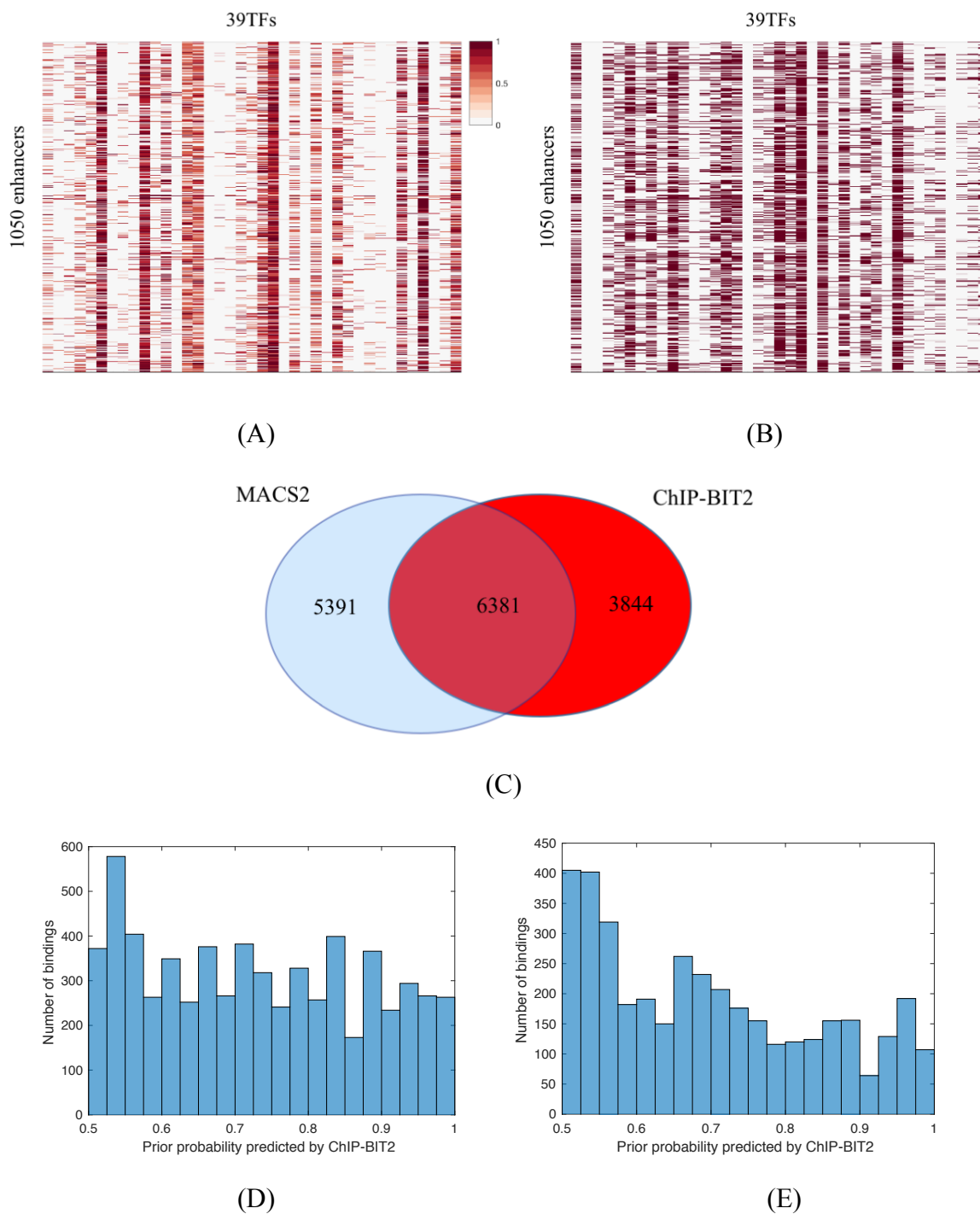
To map distal binding events at enhancer regions with target genes, we downloaded all MCF7 ChIA-PET data from ENCODE and used Mango (<https://github.com/dphansti/mango>) to extract significant 3D chromatin interactions with default setting. We annotated two ends of each interaction using enhancer or promoter regions (minimum 500 bps overlap). In total, 39,703 interactions were annotated as enhancer-promoter interactions including 9,977 enhancer regions and 9,651 target genes. Among those genes in Fig. S12, 318 genes had at least one enhancer-promoter interaction with 1,050 enhancers. In total, we obtained 1,122 enhancer-promoter interactions. Prior binding events at enhancer regions, enhancer-promoter interactions and target gene expression data can be found in Table S10, which were further used for enhancer FRN inference.

**Table S10.** Prior binding matrix (enhancer), enhancer-promoter interactions and normalized gene expression data of MCF7 cells.

Table S10 can be found in the supplementary file: “Supplementary Material Table S10.xlsx”.

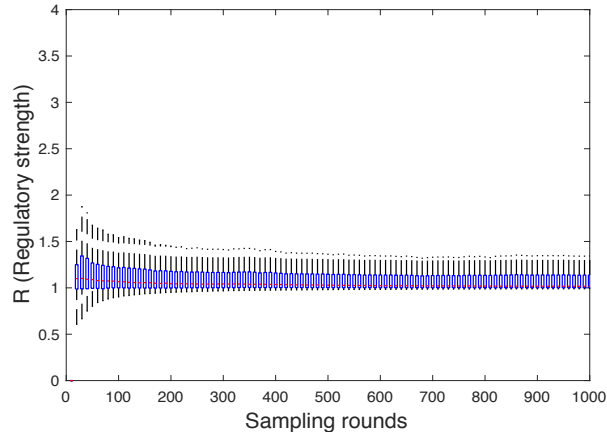


**Fig. S13.** Prior bindings at gene promoter regions. (A) ChIP-BIT2-generated prior binding matrix (weighted, 0~1); (B) MACS2-generated prior binding matrix (binary, 0 (white) or 1 (red)); (C) Similarity between binding events detected by ChIP-BIT2 and MACS2; (D) a distribution of ChIP-BIT2 probabilities for common binding events; (E) a distribution of ChIP-BIT2 probabilities for binding events detected by ChIP-BIT2 only.

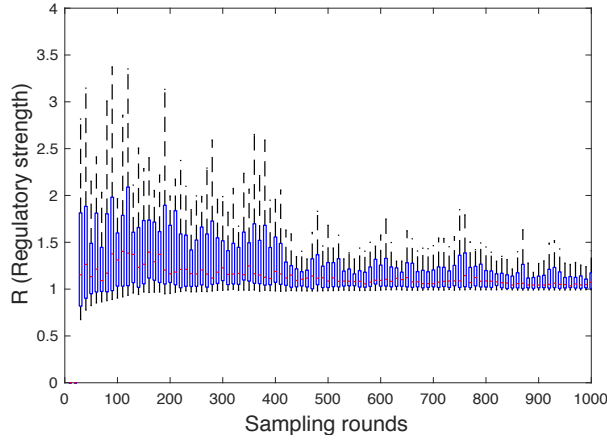


**Fig. S14.** Prior bindings at enhancer regions. (A) ChIP-BIT2-generated prior binding matrix (weighted, 0~1); (B) MACS2-generated prior binding matrix (binary, 0 (white) or 1 (red)); (C) Similarity between binding events detected by ChIP-BIT2 and MACS2; (D) a distribution of ChIP-BIT2 probabilities for common binding events; (E) a distribution of ChIP-BIT2 probabilities for binding events detected by ChIP-BIT2 only.

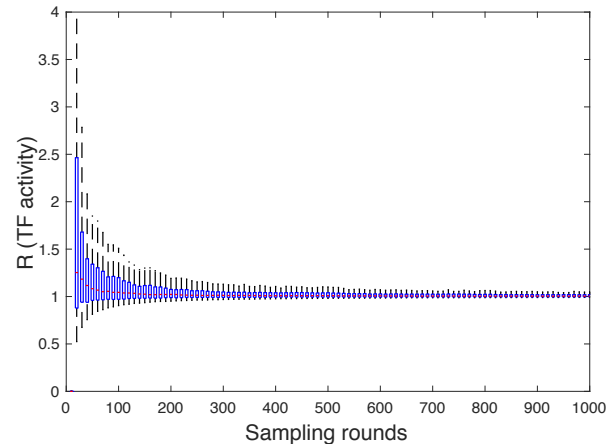
### S6.3 Convergence check of CRNET



(A)

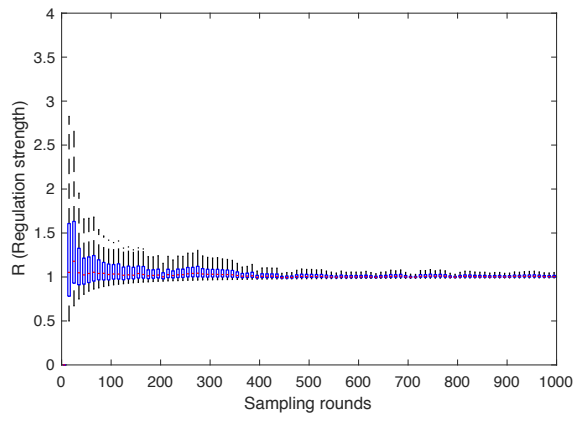


(B)

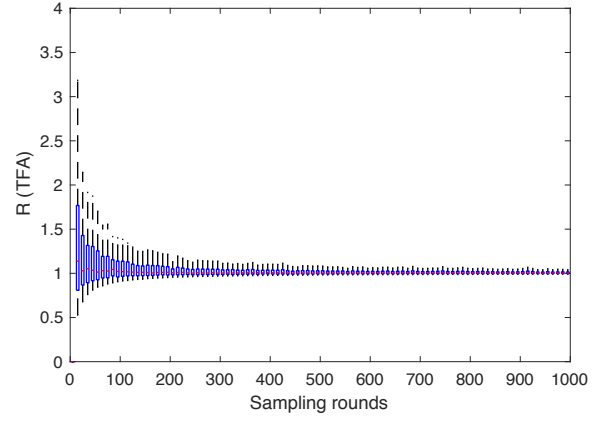


(C)

**Fig. S15.** Boxplots of  $\hat{R}$  values (convergence) for the inferred FRNs at promoter regions. (A)  $\hat{R}$  values of the regulatory strength sampled by CRNET; (B)  $\hat{R}$  values of the regulatory strength sampled by COGRIM; (C)  $\hat{R}$  values of TF activities sampled from time-course gene expression data using CRNET.



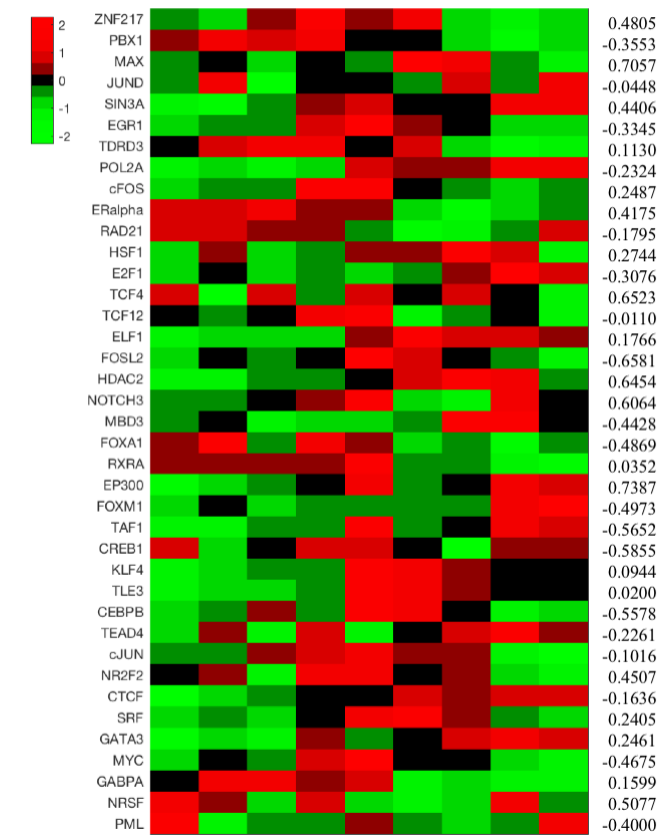
(A)



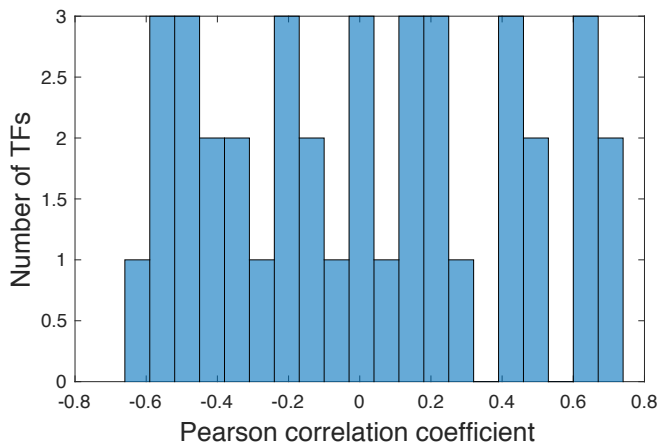
(B)

**Fig. S16.** Boxplots of  $\hat{R}$  values (convergence) for the inferred FRN at enhancer regions. (A)  $\hat{R}$  values of the regulatory strength sampled by CRNET; (B)  $\hat{R}$  values of TFAs sampled from time-course gene expression data using CRNET.

S6.4 CRNET-estimated TFAs and their similarity to TF expression

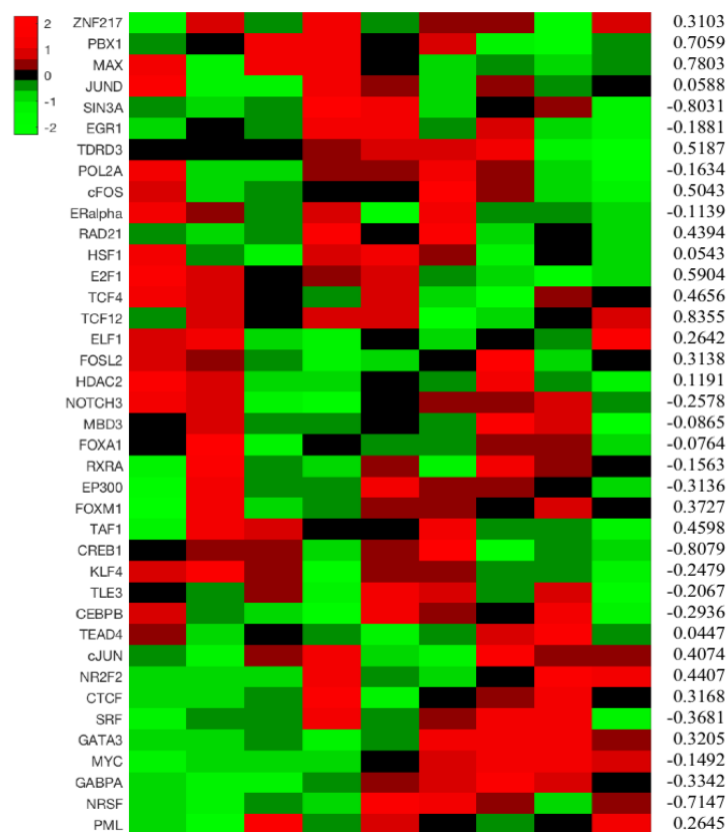


(A)

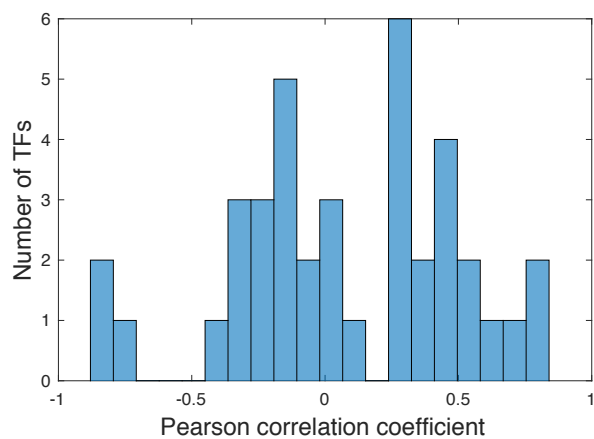


(B)

**Fig. S17.** CRNET-estimated TFAs for TFs functional at promoter regions and their similarity with original TF expression: (A) heatmap of TFAs; (B) histogram of Pearson correlation coefficients between TFAs and TF expression.



(A)



(B)

**Fig. S18.** CRNET-estimated TFAs for TFs functional at enhancer regions and their similarity with original TF expression: (A) heatmap of TFAs; (B) histogram of Pearson correlation coefficients between TFAs and TF expression.



### S6.5 Validation of MYC's proximal or distal target genes

Setting the threshold of fold change as 0.5, we obtained 2,720 differentially expressed genes among a total of 34,694 genes. A high proportion (2,271 genes, 83.5%) of them are down-regulated (fold change < -0.5) under siMYC condition. This suggests MYC is regulating most target genes positively in breast cancer MCF-7 cells under E2-induced condition. This is consistent to our observations of CRNET TFA estimation since the activities of MYC become stronger when MCF-7 cells are stimulated by E2, as shown in Figs. S17 and S18.

Previously we collected 464 E2 up-regulated genes for FRN prediction at gene promoter regions. Here, 119 of them are validated as MYC targets (significantly down regulated under siMYC condition). The success rate of validation is 25.7%. Also, among 317 genes used for enhancer FRN prediction, 85 are validated MYC's target genes. The success rate of validation is 26.8%. Using CRNET or COGRIM, we integrated prior binding information with time-course gene expression data to predict functional bindings for a set of TFs. We calculate the success rate of validation on MYC's target genes in promoter FRN predicted by CRNET (ChIP-BIT2), CRNET (MACS2) or COGRIM (ChIP-BIT2), and that of enhancer FRN predicted by CRNET (ChIP-BIT2) or CRNET (MACS2) (as shown in Table S11).

**Table S11.** Summary of validated genes in the FRNs predicted by competing methods.

Region	Promoter			Enhancer	
Raw validation rate	25.7% (119/464)			26.8% (85/317)	
Method	CRNET (ChIP-BIT2)	CRNET (MACS2)	COGRIM (ChIP-BIT2)	CRNET (ChIP-BIT2)	CRNET (MACS2)
Predicted MYC targets	101	87	55	92	78
Validated MYC targets	40	34	23	44	35
p-value	1.3e-4	6.5e-4	1.5e-3	3.4e-8	2.9e-6

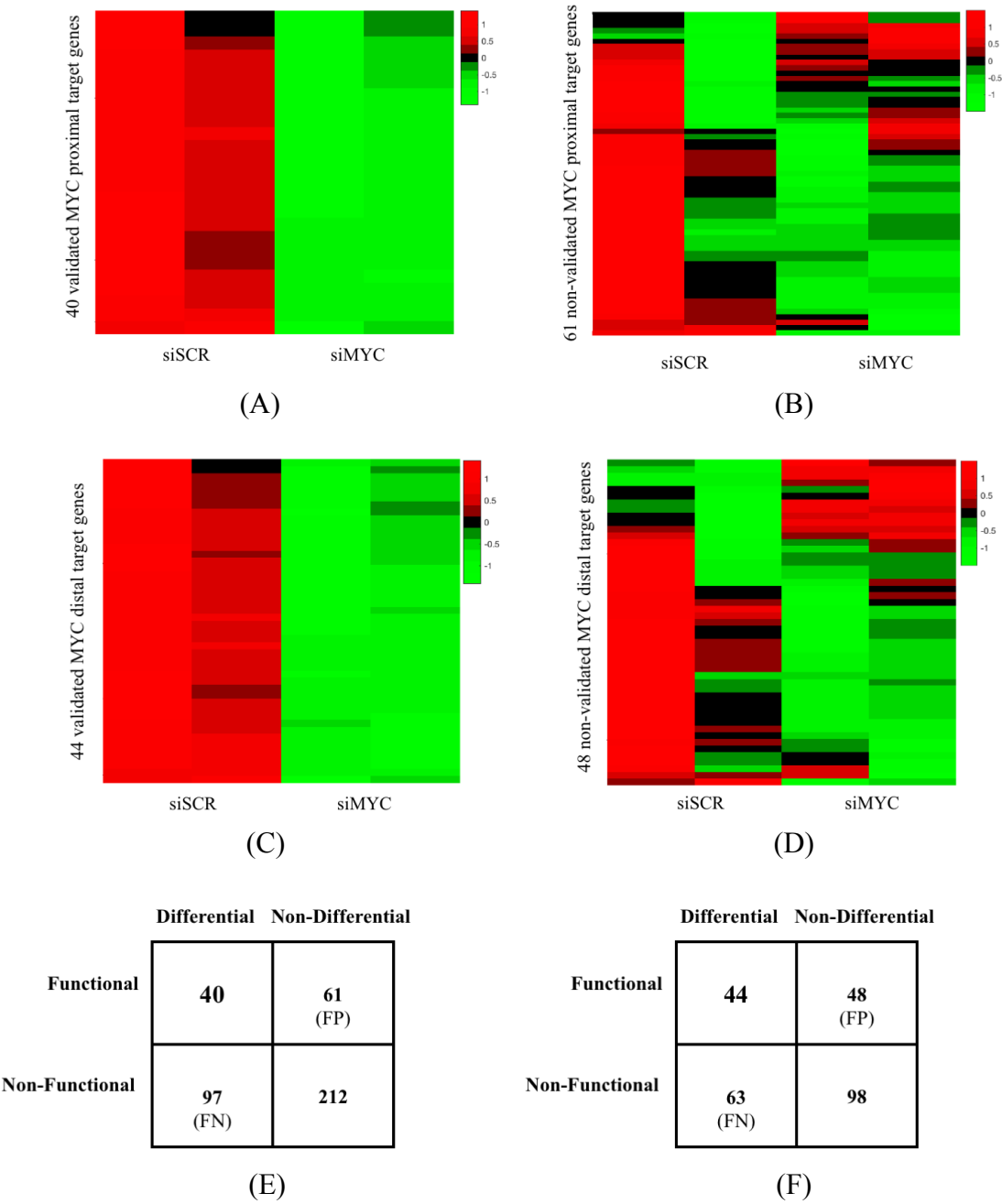
A true (validated) target gene is a gene with predicted MYC's functional binding and differentially expressed when MYC is knocked down. In the FRN of CRNET (ChIP-BIT2), we have validated 40 proximal genes (with proximal MYC bindings) and 44 distal genes (with distal MYC bindings).

A false positive (non-validated) target gene is a gene with predicted MYC's functional binding but non-differentially expressed when MYC is knocked down. In the FRN of CRNET (ChIP-BIT2), we have 61 false positive proximal genes and 48 false positive distal genes.

A false negative target gene is a gene with predicted non-functional binding but differentially expressed when MYC is knocked down. In the FRN of CRNET (ChIP-BIT2), we have 97 false negative proximal genes and 63 false negative distal genes.

A negative target gene is a gene without a predicted non-functional binding and non-differentially expressed when MYC is knocked down. In the FRN of CRNET (ChIP-BIT2), we have 212 negative proximal genes and 98 negative distal genes.

Heatmap of gene expression of validated true target genes and false positive genes, and Venn diagrams of false positive/false negative predictions are shown in Fig. S19.



**Fig. S19.** Target gene validation of MYC in the CRNET-predicted FRN: (A) and (B) validated and non-validated MYC’s target genes, respectively, in the CRNET (ChIP-BIT2)-predicted promoter FRN; (C) and (D) validated and non-validated MYC’s target genes, respectively, in the CRNET (ChIP-BIT2)-predicted enhancer FRN; (E) and (F): Venn diagrams showing false positive/negative MYC functional predictions in promoter and enhancer FRNs.

## S7. Summary of data, tools and results

	Data	Competing Methods	Results
Robustness test	Simulated regulatory network with weighted binding prior and time-course gene expression	CRNET BNCA (Sabatti and James, 2006) COGRIM (Chen, et al., 2007) LASSO (Qin, et al., 2014) NARROMI (Zhang, et al., 2013) GENIE3 (Huynh-Thu, et al., 2010)	(a) CRNET is more robust against false positive bindings and noise in gene expression data than competing methods.  (b) CRNET converges much faster than existing Gibbs sampling based methods.
	DREAM4 regulatory network with binary binding prior and time-course gene expression		
Large scale network inference	ENCODE ChIP-seq data of 228 TFs from K562 cells and time-course gene expression data (GSE1036)	CRNET COGRIM LASSO NARROMI GENIE3	(c) CRNET can be used to jointly analyze hundreds of TFs and is running much faster than COGRIM.  (d) CRNET has higher validate rates on ‘true’ target genes of ATF3, EGR1 and SRF than competing methods.
	RNA-seq data (GSE33816) from K562 cells with shRNA targeting to ATF3, EGR1 or SRF		
	ENCODE ChIP-seq data of 122 TFs from GM12878 cells and time-course gene expression data (GSE51709)		
Real application	39 TFs ChIP-seq data from MCF-7 cells and time-course gene expression data (GSE62789)	ChIP-BIT2+CRNET MACS2+CRNET ChIP-BIT2+COGRIM	(e) CRNET can also be used to infer regulatory networks from enhancer regions.  (f) CRNET has a better performance on predicting MYC target genes in both promoter and enhancer studies if ChIP-BIT2 results are used as binding prior.  (g) CRNET is better than COGRIM even if ChIP-BIT2 results are given to COGRIM as binding prior.
	ECNODE MCF-7 ChIA-PET data		
	Gene expression data from MCF-7 cells treated by siMYC		

## S8. Glossary of variables and parameters

<b>B</b>	prior TF-gene binding network (weighted (0~1) or binary [0,1])
$T$	the total number of TFs
$J$	the total number of genes (or enhancer-gene loops)
$M$	the total number of gene expression samples (or conditions)
$z_{j,t}$	a binary binding event of $t$ -th on $j$ -th gene
$b_{j,t}$	a prior probability for physical binding event $z_{j,t}=1$
$s_{j,t}$	read intensity of $t$ -th TF at $j$ -th gene promoter region
$\mu_{TFBS}$	mean of the global Gaussian distribution component of ChIP-BIT
$\sigma_{TFBS}^2$	variance of the global Gaussian distribution component of ChIP-BIT
$s_{j,input}$	read intensity of input ChIP-seq data at $j$ -th gene promoter region; which is also the mean of the local Gaussian distribution component
$\sigma_{input}^2$	variance of the local Gaussian distribution component
$d_{j,t}$	relative distance of binding site of $t$ -th TF to $j$ -th gene TSS
$\lambda_t$	binding distance exponential distribution parameter for $j$ -th TF
$d_p$	length of one side gene promoter region
<b>Y</b>	a $J \times M$ matrix of all $J$ genes expression under all $M$ conditions (time points)
$y_j$	a gene expression vector of $j$ -th gene under $M$ conditions
$y_{j,m}$	gene expression of $j$ -th gene under $m$ -th condition
$n$	gene expression data noise
$\sigma^2$	variance of noise
<b>X</b>	a $T \times M$ matrix of all $T$ TFs' activities under $M$ conditions
$x_t$	a TFA vector of $t$ -th TF under $M$ conditions
$x_{t,m}$	a TFA variable of $t$ -th TF under $m$ -th condition
$\mu'_x$	mean of sampled TFA
$\sigma'^2_x$	variance of sampled TFA
$\eta_j$	base line gene expression of $j$ -th gene

<b>A</b>	a $J \times T$ matrix of regulation strength of $T$ TFs on $J$ genes
$a_{j,t}$	regulation strength of $t$ -TF on $j$ -th gene
<b>Z</b>	functional TF-gene binding network
$b_0$	logistic regression parameter
$b_1$	logistic regression parameter

## References

- Addya, S., *et al.* Erythroid-induced commitment of K562 cells results in clusters of differentially expressed genes enriched for specific transcription regulatory elements. *Physiol Genomics* 2004;19(1):117-130.
- Chen, G., Jensen, S.T. and Stoeckert, C.J., Jr. Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol* 2007;8(1):R4.
- Chen, X., *et al.* ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles. *Nucleic Acids Res* 2016;44(7):e65.
- Cheng, C., Min, R. and Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* 2011;27(23):3221-3227.
- Gelman, A. and Rubin, D.B. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* 1992;7(4):457-472.
- Huynh-Thu, V.A., *et al.* Inferring regulatory networks from expression data using tree-based methods. *PLoS One* 2010;5(9).
- Liao, J.C., *et al.* Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A* 2003;100(26):15522-15527.
- Qin, J., *et al.* Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods* 2014;67(3):294-303.
- Sabatti, C. and James, G.M. Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics* 2006;22(6):739-746.
- Sanyal, A., *et al.* The long-range interaction landscape of gene promoters. *Nature* 2012;489(7414):109-113.
- Su, D., *et al.* Interactions of chromatin context, binding site sequence content, and sequence evolution in stress-induced p53 occupancy and transactivation. *PLoS Genet* 2015;11(1):e1004885.
- Venet, D., Detours, V. and Bersini, H. A measure of the signal-to-noise ratio of microarray samples and studies using gene correlations. *PLoS One* 2012;7(12):e51013.
- Zhang, X., *et al.* NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* 2013;29(1):106-113.