Bioinformatics, 34(10), 2018, 1733–1740
doi: 10.1093/bioinformatics/btx827
Advance Access Publication Date: 21 December 2017
Original Paper

Gene expression

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

Xi Chen[1], Jinghua Gu[1], Xiao Wang[1], Jin-Gyoung Jung[2], Tian-Li Wang[2], Leena Hilakivi-Clarke[3], Robert Clarke[3] and Jianhua Xuan[1,*]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA, [2]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD 21231, USA and [3]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on December 8, 2016; revised on December 10, 2017; editorial decision on December 19, 2017; accepted on December 20, 2017

Abstract

Motivation: NGS techniques have been widely applied in genetic and epigenetic studies. Multiple ChIP-seq and RNA-seq profiles can now be jointly used to infer functional regulatory networks (FRNs). However, existing methods suffer from either oversimplified assumption on transcription factor (TF) regulation or slow convergence of sampling for FRN inference from large-scale ChIP-seq and time-course RNA-seq data.

Results: We developed an efficient Bayesian integration method (CRNET) for FRN inference using a two-stage Gibbs sampler to estimate iteratively hidden TF activities and the posterior probabilities of binding events. A novel statistic measure that jointly considers regulation strength and regression error enables the sampling process of CRNET to converge quickly, thus making CRNET very efficient for large-scale FRN inference. Experiments on synthetic and benchmark data showed a significantly improved performance of CRNET when compared with existing methods. CRNET was applied to breast cancer data to identify FRNs functional at promoter or enhancer regions in breast cancer MCF-7 cells. Transcription factor MYC is predicted as a key functional factor in both promoter and enhancer FRNs. We experimentally validated the regulation effects of MYC on CRNET-predicted target genes using appropriate RNAi approaches in MCF-7 cells.

Availability and implementation: R scripts of CRNET are available at http://www.cbil.ece.vt.edu/software.htm.

Contact: xuan@vt.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next generation sequencing (NGS) technology continues to become ever more cost-effective. The era of 'big data', with large data sets of high quality and higher resolution, has clearly arrived (Schuster, 2008). In genetic and epigenetic studies, gene transcription is regulated through the integrated action of many cis-regulatory elements, including promoter-proximal bindings as well as various distal cis-regulatory modules functioning at enhancers (Spitz and Furlong,

2012). Promoter focused studies have shown that on average only 15% of target genes predicted from ChIP-seq data of a single TF are significantly differentially expressed when this TF is knocked down (Chen *et al.*, 2016; Cusanovich *et al.*, 2014). A large proportion of physical bindings are either not functional or do not act alone. Given a large number of TFs, it is not practical to knock down TFs individually. Therefore, computational efforts for inference of functional regulatory networks (FRNs) play an important role in large-scale TF-gene regulation analysis (Angelini and Costa, 2014). Prior binding information can be obtained from static ChIP-seq data, the condition of which is similar to the '0' time point when we start to generate time-course gene expression data with a certain stimulus. Integrating prior binding information with time-course gene expression data, we can infer which binding sites are functional during the cell response to the specific stimulus. FRN inference is a cost-effective and comprehensive approach to study the joint regulatory effects of multiple TFs. A variety of omics data types must be integrated, analyzed and interpreted (Angelini and Costa, 2014; Chen *et al.*, 2013; Karlebach and Shamir, 2008).

Early attempts for FRN inference solely used gene expression data (Huynh-Thu *et al.*, 2010; Zhang *et al.*, 2013). With the accumulation of binding information like binding motifs or ChIP-chip/seq data, integrative approaches are now being developed. For example, Sabatti *et al.* proposed a Bayesian network component analysis framework (BNCA) approach to simultaneously infer protein activities of TFs (TFAs) and functional bindings by integrating ChIP-chip and microarray gene expression (Sabatti and James, 2006). Chen *et al.* developed a similar Bayesian hierarchical model namely COGRIM, to infer regulatory gene clusters (Chen *et al.*, 2007). Large-scale ChIP-seq data in ENCODE database makes it possible to predict FRNs on many different cell types. Wang *et al.* developed a BETA package for functional gene prediction by integrating single TF ChIP-seq data with target gene RNA-seq data (Wang *et al.*, 2013). Qin *et al.* developed an LASSO based integrative approach (Qin *et al.*, 2014).

The importance of statistical integration of binding signals and gene expression data in understanding gene regulatory mechanisms was discussed by Angelina & Costa (Angelini and Costa, 2014). They proposed Bayesian integration models for causal inference of genome-wide FRNs. BNCA and COGRIM are two existing Bayesian approaches using Gibbs sampling to infer functional bindings. Both tools require sampling distributions for binding occurrence and non-occurrence (each is a Gaussian distribution). This step increases their computational cost greatly, since in each round the tools must sample two different distributions under two hypotheses. From our perspective, the key parameter of interest is the probability for binding occurrence. When the number of TFs is small, a linear regression model can be used to infer a FRN very efficiently (Liao *et al.*, 2003). Correlation coefficients of individual bindings and the regression error for each gene are reported and then jointly modeled as a variable following a Student's t distribution (Lange *et al.*, 1989). The false discovery rate (FDR) of each binding can be calculated from the Student's t distribution. Only if the FDR is lower than a threshold will the binding be accepted. This evaluation is very efficient because only one hypothesis (binding occurrence) is tested. With the accumulation of ChIP-seq data, the number of bindings on a gene is usually larger than the number of gene expression observations. Such a linear regression model cannot be directly used to infer large-scale FRNs. Efficient integrative methods are in need to overcome above limitations.

We have developed a novel Bayesian method, namely CRNET, to integrate ChIP-seq and time-course RNA-seq data for functional regulatory NETwork inference. We model functional bindings as Bernoulli random variables with prior knowledge of physical bindings predicted from ChIP-seq data. CRNET uses a hybrid model of Gibbs sampling and linear regression to predict functional bindings with a significantly improved sampling efficiency. Several desirable properties of the proposed CRNET method are as follows: (i) using a Gibbs sampling framework, in each round CRNET only samples a subset of physical bindings as potentially functional; (ii) using linear regression model, regulation strength of each sampled binding and the regression error are jointly modeled as a variable following a Student's t-distribution and a logistic function is then used to transfer the Student's t-distributed variable to a probability; (iii) unlike conventional methods using known TF expression as the activated form of TFs (TFAs), CRNET models each TFA as a variable and use a two-stage Gibbs sampling procedure to sample functional bindings and TFAs iteratively. As an extension, if cell type-specific 3D chromatin interaction data are available [i.e. ChIA-PET (Li *et al.*, 2014) or Hi-C (van Berkum *et al.*, 2010)], CRNET can also integrate prior bindings observed from enhancer regions with target gene expression data to infer distal and functional bindings.

We show the advantages of CRNET on synthetic datasets with respect to different experimental settings of noise level and false connection rate. The performance of CRNET is further benchmarked on DREAM4 *in silico* regulatory networks with time-course gene expression data by varying the proportions of false positive/false negative perturbations in the prior networks. CRNET achieves a significant improvement on functional bindings prediction over existing methods. To demonstrate the capability of CRNET on large-scale FRN inference, we apply CRNET to K562 cell line data and GM12878 data, respectively. In terms of sampling speed, CRNET is five times faster than the conventional Bayesian approaches. Specific for the K562 study, we validate functional bindings of three selected TFs as ATF3, EGR1 and SRF. Compared to competing methods, a higher proportion of functional genes predicted by CRNET are validated. Finally, we apply CRNET to breast cancer MCF-7 data for FRN inference at promoter or enhancer regions. MYC is predicted as the most dominant TF and also a positive regulator in both FRNs. We transfect MCF-7 cells with siMYC for 24 h and successfully validate the positive regulatory effects of MYC on a significant set of target genes.

2 Materials and methods

CRNET is designed to use time-course RNA-seq data for the refinement of FRNs from initial candidate networks that can be constructed from ChIP-seq data. Specific for FRN inference at enhancer regions, additional prior information of enhancer-promoter interactions is needed, which can be obtained from cell type-specific ChIA-PET or Hi-C data. In Figure 1, using Gibbs sampling, CRNET iteratively samples hidden TFAs by assuming Gaussian random process, calculates the significance of regulatory strength for each binding based on Student's t statistics, and samples each functional binding as a Bernoulli random variable according to the conditional probability. After sufficient rounds of sampling, CRNET reports a posterior probability (sample frequency) for each binding that indicates the possibility that this connection is functional. A more detailed workflow of CRNET is shown in Supplementary Figure S1.

2.1 Prior binding network construction

Given promoter or enhancer annotation files and multiple TFs ChIP-seq data, a prior binding matrix can be constructed using a

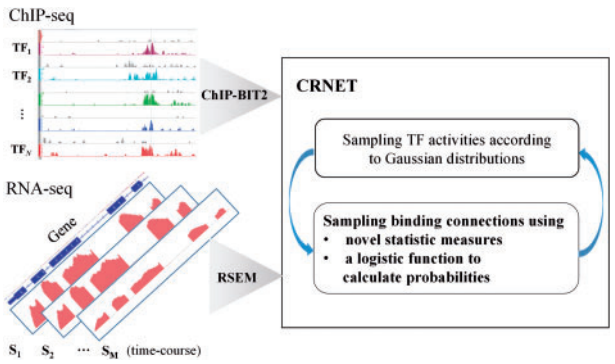

Fig. 1. Flowchart of CRNET for FRN inference. CRNET is built on a twostage Gibbs sampling procedure: (1) sampling hidden transcription factor activities (TFAs) and (2) sampling binding connections

probabilistic method, ChIP-BIT2 (an extended version of ChIP-BIT (Chen *et al.*, 2016) to detect binding sites at enhancer and promoter regions, respectively). An advantage of using ChIP-BIT2 is that we can detect both strong and weak bindings. A weak binding refers to a binding with a relatively low read intensity in the sample ChIP-seq data but that is still significantly higher than that of the matched input data. As demonstrated in (Chen *et al.*, 2016) and (Ramos and Barolo, 2013), weak bindings at promoter and enhancer regions could both result in functional regulation of target genes. More details about ChIP-BIT2 can be found in Supplementary Material S2.

For promoter study, a prior binding matrix between TFs and target genes can be directly constructed since each promoter region can be uniquely mapped to each gene. In this matrix, each row represents a unique gene and each column represents a TF. For enhancer study, an enhancer-promoter (gene) loop map (provided by 3D chromatin interactions) is needed to associate distal TF bindings at enhancer regions with target genes (Sanyal *et al.*, 2012). 3D chromatin interactions can be extracted from Hi-C data (Servant *et al.*, 2015) or ChIA-PET data (Phanstiel *et al.*, 2015). More details about how to construct the enhancer-gene loop map can be found in Supplementary Material S3.1. Using the map and distal bindings at enhancer regions, we can build a prior binding matrix, too, whereas each row represents a unique enhancer-gene loop (enhancer: gene). Note that the number of rows may be larger than the number of actual target genes or enhancers because one enhancer may regulate multiple genes through different loops and one gene may also be regulated by more than one enhancer. We define a prior binding matrix $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_j, \ldots, \mathbf{b}_J]^T$, where each row $\mathbf{b}_j = [b_{j,1}, \ldots, b_{j,t}, \ldots, b_{j,T}]$ represents prior binding probabilities of in total T TFs. If ChIP-seq data are not available, **B** can be constructed as a binary matrix from other regulatory network databases [e.g. TRED (Zhao *et al.*, 2005) and RegNetwork (http://www.regnetworkweb.org/)].

2.2 Time-course RNA-seq data processing

Considering gene expression data quality, we infer FRNs by integrating prior binding matrix **B** with time-course RNA-seq data. For each RNA-seq sample, we use RSEM (v1.3.0) (Li and Dewey, 2011) to estimate the transcripts per million (TPM) value of each gene. We then take a log2 transform of the TPM value. Candidate target genes should have a significant dynamic expression change (i.e. at least at one time point it is differentially expressed as compared to the basal expression ('0' time point)). Considering computational cost, pre-selection of biologically meaningful genes is preferred. We define a gene expression matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_j, \ldots, \mathbf{y}_J]^T$. **Y** has the same number of rows as **B** and each row $\mathbf{y}_j = [y_{j,1}, \ldots, y_{j,m}, \ldots, y_{j,M}]$ represents target gene expression under M conditions for the j-th gene or enhancer-gene loop in **B**. For simplicity, we treat each row of **Y** or **B** as a 'gene' and infer functional bindings for it.

2.3 Integrative modeling in the CRNET approach

In general, gene transcription is regulated by a set of TFs, whose activation via post-translational modification controls gene expression. The activated form of a TF [modeled as TF activity (TFA)], rather than its expression level, controls promoters or enhancers and dictates the physiological state of the cell (Liao *et al.*, 2003). Correspondingly, how each promoter or enhancer receives the signal reflects the relative contribution of each TF to the expression of the target gene, which can be quantified as regulation strength. For the j-th candidate target gene, we model gene expression $\mathbf{y}_j$ using a log-linear model (Liao *et al.*, 2003) as follows:

$$\mathbf{y}_j = \sum_{t=1}^{T} z_{j,t} a_{j,t} \mathbf{x}_t + \eta_j + \mathbf{n}_j, \tag{1}$$

where $\mathbf{x}_t = [x_{t,1}, \ldots, x_{t,m}, \ldots, x_{t,M}]^T$ is a TFA vector and each $x_{t,m}$ represents the activity of t-th TF under m-th sample. We define a TFA matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_T]^T$. $z_{j,t}$ and $a_{j,t}$ represent the binding state and associated regulation strength for the connection between t-th TF and j-th gene. Regulation strength $a_{j,t}$ is unknown and needs to be estimated for $z_{j,t} = 1$ ($a_{j,t}$ is 0 when $z_{j,t} = 0$). Accordingly, we define a binding matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_j, \ldots, \mathbf{z}_J]^T$ and a regulation strength matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_j, \ldots, \mathbf{a}_J]^T$. η_j represents the basal expression at '0' time point as $y_{j,0}$ and $\mathbf{n}_j = [n_{j,1}, \ldots, n_{j,m}, \ldots, n_{j,M}]$ is a Gaussian additive noise vector (with mean zero and variance σ^2).

2.3.1 Sampling hidden TFAs

We model each TFA vector $\mathbf{x}_t$ as a Gaussian random process where the prior distribution of each variable $x_{t,m}$ is a Gaussian distribution with mean zero and variance σ_x^2 (Sabatti and James, 2006). The joint probability of TFA matrix **X**, given gene expression matrix **Y**, regulatory strength matrix **A**, and binding matrix **Z**, is defined as follows:

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

We use Gibbs sampling to sample $x_{t,m}$ according to its conditional probability as defined in Equation (3), which is also a Gaussian distribution:

$$P(x_{t,m}|y_{j,m}, \mathbf{a}_j, x_{t' \neq t,m}) \propto \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_{x'}} \exp\left(-\frac{1}{2\sigma_{x'}^2}(x_{t,m} - \mu_{x'})^2\right), \tag{3}$$

where the mean and variance are defined in following equations:

$$\mu'_x = \frac{\sigma'^2_x}{\sigma^2}\frac{1}{J}\sum_j \left[\left(y_{j,m} - \sum_{t' \neq t} a_{j,t'} z_{j,t'} x_{t',m} - \eta_j\right) a_{j,t} z_{j,t}\right], \tag{4}$$

$$\sigma'^2_x = \frac{\sigma^2 J \sigma_x^2}{\sigma_x^2 \sum_j a_{j,t}^2 z_{j,t}^2 + \sigma^2 J}. \tag{5}$$

More details about the derivation of μ'_x and σ'^2_x can be found from Supplementary Material S3.2. If there are two or more gene

expression replicates under the same time point, we provide an option to assume TFA the same under the same time point and use the mean value of gene expression replicates to estimate μ'_x.

2.3.2 Sampling binding connections

It is difficult to model directly the distribution of regulation strength. In both BNCA and COGRIM, regulation strength is assumed to follow Gaussian distribution but the veracity of this assumption is not fully justified. For target gene identification, our final goal is to determine functional bindings rather than sample individual regulation strength. Therefore, we calculate the value of regulation strength using linear regression directly and determine whether a binding is likely to be functional (a true event). The basic idea is to use hypothesis testing: H_0 (null hypothesis)—no binding (regulation strength = 0); H_1 (alternative hypothesis)—functional binding (regulation strength ≠ 0).

For some genes, the number of prior bindings may be larger than the number of gene expression samples (M), especially when given tens or hundreds of TFs' ChIP-seq data. Binding signal differences as reflected by their prior probabilities are considered during the sampling process. To meet the requirement of linear regression, in each round of sampling, for each of such genes, we randomly select at most M bindings according to their prior probabilities in **B** and then, evaluate their functional effects (regression performance). Those bindings with higher prior probabilities will be selected more frequently. If binary prior is given, for each gene uniform prior will be assigned to all candidate bindings. Whether to accept each selected binding is determined by its regression performance on the target gene. To evaluate the functional effects of selected bindings on j-th gene, we estimate $\mathbf{a}_j$ directly using a least-squares method as:

$$\mathbf{a}_j = \left(\mathbf{X}\mathbf{X}^{\mathrm{T}}\right)^{-1}\mathbf{X}\left(\mathbf{y}_j - \eta_j\right)^T. \quad (6)$$

We define a t-score $f(a_{j,t})$ to help evaluate if $z_{j,t}$ is functional. $f(a_{j,t})$ (as defined in the following equation) takes into account both regulation strength and regression error, which follows a Student's t distribution under H_0 with a degree-of-freedom of $M - 1 - \sum_t z_{j,t}$ (Gu *et al.*, 2012):

$$f(a_{j,t}) = \frac{a_{j,t}\sqrt{M - 1 - \sum_t z_{j,t}}}{\sqrt{\sum_m \left(y_{j,m} - \sum_t a_{j,t}z_{j,t}x_{t,m} - \eta_j\right)^2 / \sum_m x_{t,m}^2}}. \quad (7)$$

With a significance level of α (e.g. 0.05), we can make a decision on any $f(a_{j,t}) \geq \mathrm{t}_\alpha$ with a α (=5%) risk to be a true binding (where t_α is the t-statistic value determined by the significance level α). We further transform $f(a_{j,t})$ to a probability with which to generate Gibbs samples. We propose to use a logistic function as defined in the following equation as a mapping function to calculate the probability of a t-score (Weaver *et al.*, 1999).

$$P\left(z_{j,t} = 1 | \mathbf{y}_j, \mathbf{X}\right) = \frac{1}{1 + \exp\left(-\left(b_1 | f(a_{j,t}) | + b_0\right)\right)}, \quad (8)$$

where b_1 and b_0 are trained following a procedure as described in Supplementary Material S3.3. A logistic function curve with trained b_1 and b_0 is shown in Supplementary Figure S2.

According to the conditional probability in Equation (8), we can sample $z_{j,t}$ by either accepting or rejecting $z_{j,t} = 1$. From Equation (7) it can be seen that $f(a_{j,t})$ calculation for $z_{j,t} = 1$ depends on other $a_{j,t'}$ ($t' \neq t$). In practice, we iteratively sample a new $z_{j,t}(i+1)$ in the $(i+1)$-th round of sampling as follows:

$$P\left(z_{j,t}(i+1) | \mathbf{Y}, \mathbf{X}, z_{j,1}(i+1), \ldots, z_{j,t-1}(i+1), 1, z_{j,t+1}(i), z_{j,T}(i)\right). \quad (9)$$

To ensure that the results from Equation (9) are not dependent of the TF order, we shuffle the TF order (columns of matrix B) in each round of sampling. After accumulating enough samples, we select functional bindings according to the sampling frequency of each binding, which represents the posterior probability of the binding. We also run multiple times of CRNET with different initial states and check algorithm convergence on variable estimation (Gelman and Rubin, 1992). More details about convergence check can be found in Supplementary Material S3.4.

3 Results

3.1 Benchmarking robustness using simulated and DREAM regulatory networks

We first simulated a weighted regulatory network with 200 genes and 20 TFs and multiple time-course gene expression data sets for performance evaluation (Supplementary Figs S3 and S4). We aimed to evaluate the effects of false connections in the prior binding matrix, the noise power of gene expression data, and the number of gene expression samples on final FRN inference. We mainly simulated two different cases as: in Case 1, we simulated TFAs for individual TFs using Gaussian random process with zero mean and unit variance under 20 time points; gene expression was then simulated based on the log-linear model introduced in Equation (1) with simulated TFA and regulation strength; 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. In each case, we varied the false-positive rate (defined by false-positive interactions/true interactions) from 5% to 25% in the prior binding network with a random prior probability between 0.5 and 1. We also varied the signal-to-noise ratio (SNR) of the gene expression data from 6 to −3 dB. More details about data simulation can be found from Supplementary Material S4.1 and S4.2.

For performance comparison, besides existing Bayesian models such as BNCA and COGRIM, we also included a LASSO-based integrative approach (Qin *et al.*, 2014) and two expression-based methods [i.e. NARROMI (Zhang *et al.*, 2013) and GENIE3 (Huynh-Thu *et al.*, 2010)]. F-measure (2/(1/precision + 1/recall)) of the competing methods was presented in Figure 2. In Case 1, the total number of expression samples is larger than the number of candidate TFs. Most integrative approaches work robustly against false-positive connections in the initial network, as shown in Figure 2A. CRNET has an improved performance over traditional Bayesian approaches. When the SNR of gene expression data decreases, performance of the competing methods degrades dramatically, as shown in Figure 2B. Robustness of BNCA and COGRIM against expression data noise is lower than CRNET. CRNET uses a joint measurement of regulation strength and regression errors so that overfitting can be effectively avoided as compared with existing Bayesian methods. CRNET also estimates hidden TFAs instead of using the noisy TF expression (as in COGRIM and LASSO). In Case 2, CRNET achieves the best performance among competing methods, as shown in Figure 2C and D. In this case, CRNET's robustness to network false connections over the other competing methods is more obvious when the number of gene expression samples is small.

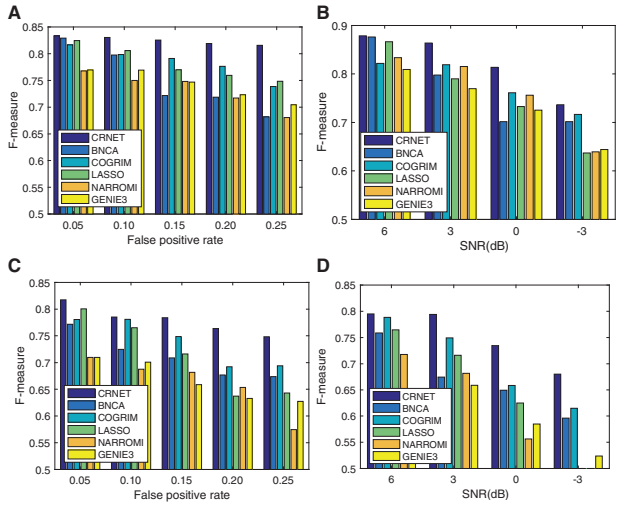

Fig. 2. F-measure performance comparison of competing methods using synthetic data with varying false positive connections and noise levels. (**A**) Case 1 with different FPRs; (**B**) Case 1 with different SNRs; (**C**) Case 2 with different FPRs; (**D**) Case 2 with different SNRs

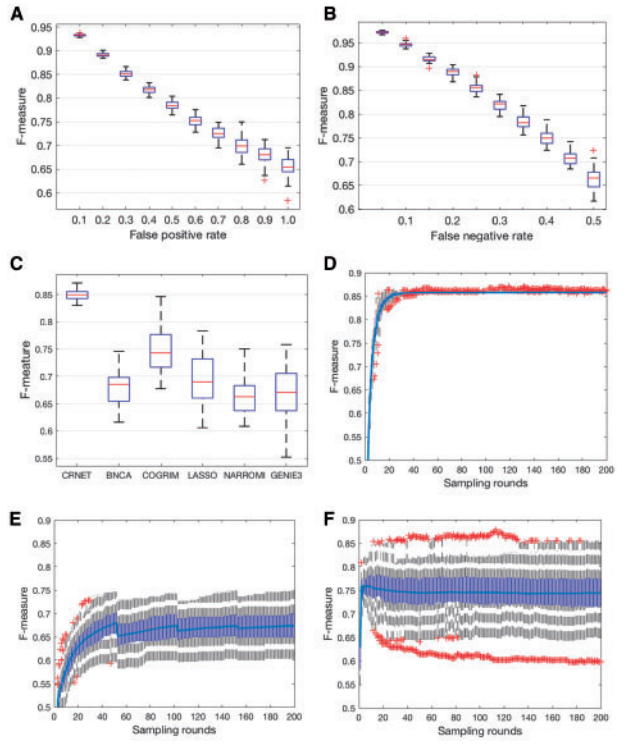

Fig. 3. Performance comparison using DREAM 4 in silico benchmark networks. (**A**) F-measure of CRNET by adding 10–100% false positive edges; (**B**) F-measure of CRNET by deleting 10–50% true edges (false negative); (**C**) A box plot of F-measure for competing methods (after 1000 rounds of Gibbs sampling for Bayesian methods); (**D**), (**E**) and (**F**) box plots of CRNET, BNCA and COGRIM F-measure performance during the sampling process

Furthermore, as shown in Supplementary Figure S5, CRNET converges much faster than competing Bayesian methods.

For further evaluation of CRNET's performance with non-ChIP-seq prior binary information, five benchmark regulatory networks and the matched time-course gene expression data were downloaded from (https://www.synapse.org/#! Synapse: syn3049712/files/), which are used in DREAM4 challenge (Marbach *et al.*, 2009). We first tested the robustness of CRNET on FRN inference by adding false-positive or -negative edges to the benchmark networks. A false-negative edge is a 'true' binding in the benchmark network but 'missed' by the prior binding matrix. We varied the false-positive rate from 5% to 100% or the false-negative rate (defined by false-negative interactions/true interactions) from 5% to 50% in the prior binding matrices. Box plots of F-measures of CRNET under different rates of false positive/false negative perturbations were presented in Figure 3A or B. It can be found from Figure 3A that when the false positive rate of the prior matrix is under 40%, CRNET can achieve an average Fmeasure higher than 0.8. To achieve a similar performance using a prior network with false negative interactions, as shown in Figure 3B, the rate should be below 30%.

We further tested competing methods using the same benchmark networks and time-course gene expression data, too. Here, we only add false positive interactions (30%) to the prior networks. GENIE3 is the winner of DREAM4 challenge; similar to NARROMI, it uses gene expression data as the only input. For a fair comparison, we only examined the performance of NARROMI and GENIE3 by focusing on observed interactions in the prior binding network. As shown in Figure 3C, F-measures are presented in box plots for all competing methods. CRNET provides the best performance and compared with BNCA or COGRIM, it converges much faster (as shown in Fig. 3D, E and F). It can be found from Figure 3C that integrative methods using prior binding information with a reasonable false connection rate and gene expression data can provide improved performance over traditional methods using gene expression data only.

3.2 Benchmarking performance using hundreds of TFs in K562 and GM12878 cells from ENCODE

We continued examination of CRNET efficiency on large-scale FRN prediction with hundreds of TFs. ChIP-seq data of K562 and GM12878 cells were downloaded from the ENCODE database. Matched time-course gene expression datasets were downloaded from the GEO database (with access numbers: GSE1036 and GSE51709) for K562 and GM12878 cells, respectively. We estimated the SNR (signal-to-noise ratio) of each gene expression dataset using SNAGEE (Venet *et al.*, 2012). As shown in Supplementary Tables S1 or S3, using baseline expression at '0' time point as control (0 dB), the average SNR is 2.82 dB (or 2.04 dB) for K562 (or GM12878) data. Prior binding matrix construction and candidate gene selection can be found from Supplementary Material S5.1. Prior binding matrixes and gene expression data can be found in Supplementary Tables S2 and S4. Heatmaps of gene expression are shown in Supplementary Figure S7. In total, we selected 1351 candidate genes and 228 TFs to infer FRNs in K562 cells; 925 genes and 122 TFs for FRN inference in GM12878 cells.

We also applied competing methods to each prior binding network and the matched gene expression data. Due to the high density of the prior networks (Supplementary Fig. S6), BNCA, a method exhaustively searching and testing TF combinations, does not work. COGRIM and LASSO require sparse binary binding events as input. For a fair comparison, we set the prior probability threshold as 0.85 and selected binary binding events from 1348 genes and 173 TFs in K562 cells and 877 genes and 80 TFs in GM12878 cells for further analysis. GENIE3 uses gene expression data only and theoretically has no limitations on the number of candidate TFs. After running 1000 rounds of Gibbs sampling, the average speed of CRNET is 498 or 58 sec/round for K562 or GM12878 FRN inference (R 3.3.1,

MAC OS X, CPU 2.8 GHz, RAM 16GB). While the sampling speed of COGRIM is 2611 or 322 sec/round under the same condition. CRNET is five times faster than COGRIM. Learned distributions of sampling frequency of CRNET and CORGIM are depicted in Supplementary Figure S8. It can be found that the sample frequency of CRNET follows a bimodal distribution. It is easy to define the cut-off threshold (0.6 in this case) and further extract confident edges. COGRIM tested two hypotheses (functional binding vs. non-functional binding) and only recorded samples meeting the requirement of 'functional binding' hypothesis. Its sampling frequency distribution is a one-component Gaussian like distribution. Similarity of CRNET-estimated TFAs and original mRNA expression for each TF was checked using Pearson correlation coefficient. As shown in Supplementary Figure S9, there is no clear correlation between them and for a number of TFs, the correlation is negative.

Specific for the FRN inferred using K562 data we validated functional target genes for three selected TFs: ATF3, EGR1 and SRF. For each TF, a 'true' target gene is defined as: (i) there should be at least one functional binding site at promoter region; (ii) this gene should be significantly differentially expressed when the TF is knocked down. We downloaded RNA-seq data with shRNA TF knockdown for each specific TF (GEO access number: GSE33816). For each TF, two RNA-seq replicates were generated under control or treatment conditions (Vehicle vs. shRNA). We applied RSEM to individual RNA-seq samples and estimated read counts and TPM values of each gene across all samples. The knockdown efficiency (differential expression of each TF) is shown in Figure 4A. We then used DeSeq2 (Love *et al.*, 2014) to identify differentially expressed target genes with a q value cutoff as 0.05. In total, we identified 1133 differential genes for ATF3, 893 genes for EGR1 and 1011 genes for SRF, whose expression patterns are shown in Supplementary Figure S10. Using ChIP-BIT2 weighted prior binding information, CRNET predicted 141, 249, and 62 functional target genes for each of the three TFs; the validation success rates are 14.9%, 11.15% and 12.5%, respectively. While using confident binding events only, the validation success rates are 14.8%, 10.44% and 11.3%. Venn diagrams of gene validation are shown in Supplementary Figure S11 and validation success rates of competing methods are shown in Figure 4B. In Supplementary Table S5, we further listed the detailed number of genes being predicted by each method containing the same number of validated genes as CRNET with binary input (achieving the same recall performance). It can be found that the FRN predicted by CRNET has the highest validation rate for each of the three TFs.

3.3 Real application using breast cancer MCF-7 cells

We finally applied CRNET to breast cancer MCF-7 cell ChIP-seq and time-course RNA-seq data and inferred two FRNs at promoter and enhancer regions, respectively. A 17 b-estradiol (E2) treated

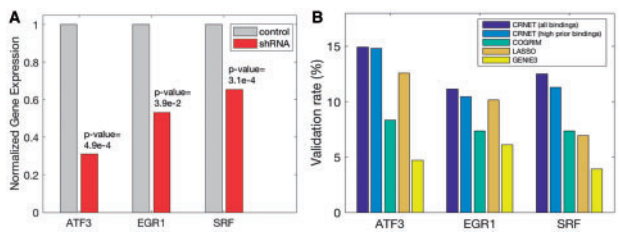

Fig. 4. True positive rate of differentially expressed genes in inferred FRN by competing methods. (**A**) TF knockdown efficiency for ATF3, EGR1 or SRF; (**B**) validation success rate for each TF

MCF-7 time-course RNA-seq dataset was downloaded from the GEO database (accession number: GSE62789) and further processed using RSEM. (Supplementary Table S6). We estimated the SNR of each RNA-seq sample using SNAGEE. As shown in Supplementary Table S7, the average SNR is 2.83 dB. Candidate target gene selection can be found from Supplementary Material S6.1. A heatmap of gene expression is shown in Supplementary Figure S12. 39 TFs ChIP-seq data were downloaded from ENCODE or GEO database (Supplementary Table S8) and processed using ChIP-BIT2. More details about data processing can be found from Supplementary Material S6.2 To associate prior bindings at the enhancer regions with target genes, 3 D chromatin interactions were extracted from a set of ECNODE MCF-7 ChIA-PET data using Mango. In total, we selected 464 genes for promoter FRN inference; 1050 enhancers and 318 genes (1122 loops) for enhancer FRN inference (Supplementary Material S6.2). A majority of enhancers have only one target gene but on average, each gene is regulated by four enhancers. Prior binding matrixes, enhancer-gene loops and candidate gene expression data were provided in Supplementary Tables S9 and S10.

For comparison, we also used MACS2 (v2.1.0) (Zhang *et al.*, 2008) to call genome-wide peaks using the same ChIP-seq dataset and then, mapped peaks to candidate gene promoter or enhancer regions. As shown in Supplementary Figure S13, most (~94%) MACS2 bindings at gene promoter regions were captured by ChIP-BIT2 and 78% of them had a ChIP-BIT2-estimated probability larger than 0.85 (the default threshold of ChIP-BIT2 for peak prediction). For those bindings predicted by ChIP-BIT2 only, 58% of them still had a ChIP-BIT2-estimated probability over 0.85. Obviously, there were a number of weak bindings missed by MACS2. Prior bindings detected from enhancer regions were shown in Supplementary Figure S14. About 60% of MACS2 bindings were still captured by ChIP-BIT2 and the latter provided additional binding events for functional exploration.

We predicted FRNs at promoter regions using CRNET and COGRIM, respectively. The average speed of CRNET is 8.8 s/round as compared with 47.6 s/round for COGRIM. Convergence check (Gelman and Rubin, 1992) was carried out using results of five replicated Markov chains generated independently using each method. As shown in Supplementary Figure S15, after 100 rounds of sampling, CRNET starts to converge, while for COGRIM the number is 500. Similarity between TFA estimated by CRNET and mRNA expression for each TF was shown in Supplementary Figure S17. It can be found that there is no clear dependency between TFA and TF expression. The pattern of TFA is, however, more consistent with that of candidate gene expression. Analyzing the top 500 edges in the CRNET-predicted FRN, we identified a key module including TFs MYC, TDRD3, E2F1, MBD3, SIN3A and MAX. Using COGRIM, a slightly different module was identified including MAX, FOXM1, RAD21, SRF and E2F1. We also checked the top 500 edges in the FRN predicted by CRNET, but with the MACS2 prior binding matrix, a module including MAX, SIN3A, MYC and E2F1 was identified. TDRD3 and MBD3 were missed because they had much fewer prior bindings predicted by MACS2, as discussed in Supplementary Material S6.2.

We then predicted another FRN at enhancer regions using CRNET. The convergence curve in Supplementary Figure S16 shows that CRNET starts to converge after 100 rounds. The average sampling speed is 16 s/round. A key TF module including MYC, TDRD3, ER-α, GABPA, GATA3 and E2F1 was identified using top ranked edges. Here, three well-known breast cancer specific enhancer activators: ER-α, GABPA and GATA3 were specifically

enriched in this enhancer FRN. Again, we further examined the enhancer FRN predicted by CRNET but using prior bindings from MACS2. Still using the top 500 edges in the inferred FRN, we identified a TF module, whereas TFs TDRD3 and GABPA in the previous module were missed. In this MCF-7 cell study, for the 39 selected TFs, prior bindings predicted by MACS2 may not provide a 'complete' candidate space for functional binding exploration.

Since in both promoter and enhancer FRN analyses MYC was identified as the 'top' TF, we used siMYC to knock down MYC in MCF-7 cells for 24 h, followed by Western blotting to confirm the knockdown efficiency (Fig. 5A). The estimated TFA of MYC goes up after E2 treatment (Supplementary Figs S17 and S18), suggesting a positive regulation relationship of MYC with those over-expressed target genes under E2 condition. If this active regulation relationship is true, it can be expected that MYC target genes will be down-regulated when MYC is knocked down. We therefore performed microarray mRNA profiling using Illumina HumanHT-12 v4 Expression BeadChip. Two replicates were generated under wild type (siSCR) or siMYC condition and processed and normalized using beadarray (v2.26.0) (Dunning *et al.*, 2007). As shown in Figure 5B, MYC has been efficiently knocked down (differential mRNA expression *P* value 2.5e-2). Setting fold change threshold as 0.5, in total we selected 2720 differentially expressed genes. A heatmap of selected gene expression is shown in Figure 5C. 2271 (83.5%) MYC target genes are significantly down-regulated after MYC knockdown. Among 101 predicted genes in the promoter FRN, there are 40 genes significantly down-regulated ('true' MYC targets). Among 92 enhancer target genes, the number of 'true' MYC targets is 44. There are 49 common predicted targets and 16 common validated targets between two studies. We calculated the hypergeometric *P* value (in –log10 format) of the enrichment of MYC valid target genes in the FRN predicted by each competing method. As shown in Supplementary Table S11, using the same prior from ChIP-BIT2, CRNET performs better than COGRIM. Since ChIP-BIT2 can capture more weak but still functional bindings, even for a strong TF like MYC, CRNET performs better if combined with ChIP-BIT2 than MACS2. Numbers of validated or non-validated MYC target and their gene expression heatmap were shown in Supplementary Figure S19.

4 Discussion

CRNET is designed as an efficient sampling approach to integrate ChIP-seq and time-course RNA-seq data for large-scale FRN inference. It aims to identify functional bindings among observed physical TF-gene interactions. Compared with other peer methods, CRNET has a faster convergence speed and an improved performance in identifying FRNs from noisy binding and gene expression data. A summary of data and tools used in this paper can be found in Supplementary Material S7. Note that CRNET is developed based on an assumption that TFs work parallel on gene regulation, thus inferring FRNs, respectively, for promoter regions and enhancer regions. However, evidence starts to emerge that TFs binding at enhancers and at promoters could work collaboratively or hierarchically. Enhancer TFs may activate TFs at promoter regions and then the latter will regulate gene expression, or enhancer TFs may directly regulate genes through enhancer-promoter interactions. If such prior information is given, the proposed CRNET can be further extended and used to infer a joint FRN by properly merging prior binding observations from promoter and enhancer regions. Another potential extension of current CRNET framework is to predict cis-regulatory modules (TF-associations), especially for enhancer studies. Prior knowledge of TF-associations may be needed to define the candidate search space for module prediction. CRNET can predict functional TF modules by sampling candidate TF-associations instead of individual TF bindings.

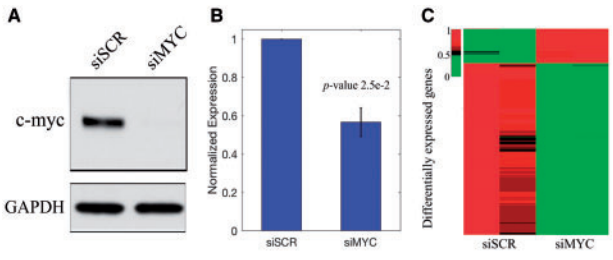

Fig. 5. Experimental validation of MYC target genes. (**A**) Western blot of MYC protein expression after transfecting MCF-7 breast cancer cells with siRNA of MYC and scramble (SCR) for 24 h; (**B**) mRNA expression levels of MYC across siRNA samples; (**C**) the heatmap of differentially expressed MYC target genes

Funding

This work was supported in part by the National Institutes of Health (CA149653 to J.X., CA149147 & CA184902 to R.C., CA164384 to L, H.-C., CA148826 & CA187512 to T.-L.W.).

Conflict of Interest: none declared.

References

Angelini,C., and Costa,V. (2014) Understanding gene regulatory mechanisms by integrating ChIP-seq and RNA-seq data: statistical solutions to biological problems. *Front. Cell. Dev. Biol.*, **2**, 51.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Chen,X. *et al.* (2016) ChIP-BIT: Bayesian inference of target genes using a novel joint probabilistic model of ChIP-seq profiles. *Nucleic Acids Res.*, **44**, e65.

Chen,X. *et al.* (2013) Reconstruction of transcriptional regulatory networks by stability-based network component analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 1347–1358.

Cusanovich,D.A. *et al.* (2014) The functional consequences of variation in transcription factor binding. *PLoS Genet.*, **10**, e1004226.

Dunning,M.J. *et al.* (2007) beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–2184.

Gelman,A., and Rubin,D.B. (1992) Inference from iterative simulation using multiple sequences. *Stat. Sci.*, **7**, 457–472.

Gu,J. *et al.* (2012) Robust identification of transcriptional regulatory networks using a Gibbs sampler on outlier sum statistic. *Bioinformatics*, **28**, 1990–1997.

Huynh-Thu,V.A. *et al.* (2010) Inferring regulatory networks from expression data using tree-based methods. *PLoS One*, **5**, e12776.

Karlebach,G., and Shamir,R. (2008) Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell. Biol.*, **9**, 770–780.

Lange,K.L. *et al.* (1989) Robust statistical modeling using the t distribution. *J. Am. Stat. Assoc.*, **84**, 881–896.

Li,B., and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,G. *et al.* (2014) Chromatin interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. *BMC Genomics*, **15** (Suppl. 12), S11.

Liao,J.C. *et al.* (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl. Acad. Sci. USA*, **100**, 15522–15527.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*., **15**, 550.

Marbach,D. *et al*. (2009) Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *J. Comput. Biol*., **16**, 229–239.

Phanstiel,D.H. *et al*. (2015) Mango: a bias-correcting ChIA-PET analysis pipeline. *Bioinformatics*, **31**, 3092–3098.

Qin,J. *et al*. (2014) Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods. *Methods*, **67**, 294–303.

Ramos,A.I., and Barolo,S. (2013) Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci*., **368**, 20130018.

Sabatti,C., and James,G.M. (2006) Bayesian sparse hidden components analysis for transcription regulation networks. *Bioinformatics*, **22**, 739–746.

Sanyal,A. *et al*. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.

Schuster,S.C. (2008) Next-generation sequencing transforms today's biology. *Nat Methods*, **5**, 16–18.

Servant,N. *et al*. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol*., **16**, 259.

Spitz,F., and Furlong,E.E. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet*., **13**, 613–626.

van Berkum,N.L. *et al*. (2010) Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp*., **39**, e1869.

Venet,D. *et al*. (2012) A measure of the signal-to-noise ratio of microarray samples and studies using gene correlations. *PLoS One*, **7**, e51013.

Wang,S. *et al*. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc*., **8**, 2502–2515.

Weaver,D.C. *et al*. (1999) Modeling regulatory networks with weight matrices. *Pac. Symp. Biocomput*, **4**, 112–123.

Zhang,X. *et al*. (2013) NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics*, **29**, 106–113.

Zhang,Y. *et al*. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*., **9**, R137.

Zhao,F. *et al*. (2005) TRED: a Transcriptional Regulatory Element Database and a platform for in silico gene regulation studies. *Nucleic Acids Res*., **33**, D103–D107.