Supplemental Materials

Hierarchical Modularity in ERα Transcriptional Network Is Associated with Distinct Functions and Implicates Clinical Outcomes

Binhua Tang[1], Hang-Kai Hsu[2], Pei-Yin Hsu[2], Russell Bonneville[1], Su-Shing Chen[3], Tim H-M Huang[2], Victor X. Jin[1*]

1 Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA, **2** Department of Molecular Medicine, Institute of Biotechnology, University of Texas Health Science Center, San Antonio, TX 78245, USA, **3** CISE & Systems Biology Lab, University of Florida, Gainesville, FL 32611, USA

*Corresponding author:

*Dr. Victor Jin, Dept. of Biomedical Informatics, 460 W 12th Avenue, 212 BRT, The Ohio State University, Columbus, OH 43210. Tel: (614) 292-6931; Email: Victor.Jin@osumc.edu

Running title: Dynamic ERα Transcriptional Regulation Network

1. The experiment protocol of E2-stimulated time-series ERα ChIP-seq:

MCF7 cells were maintained in a hormone-free medium (phenol red–free MEM with 2 mmol/L L-glutamine, 0.1 mmol/L nonessential amino acids, 50 units/mL penicillin, 50 Ag/mL streptomycin, 6 ng/mL insulin, and 10% charcoal-stripped FBS) for three days.

MCF7 cells were treated with DMSO (as 0 hour time point) or E2 (108 mol/L) for 1, 4 and 24 hours. 5 x 10^7 cells were cross-linked with 1% formaldehyde for 10 min, at which point 0.125 M glycine was used to stop the crosslinking. In brief, after crosslinking, cells were treated by lysis buffers and sonicated to fragment the chromatin to a size range of 500bp-1kb. Chromatin fragments were then immunoprecipitated with 10ug of antibody/magnetic beads. The antibodies against ERα were purchased from Santa Cruz Biotechnology (Santa Cruz, sc-8005 X). After immunoprecipitation, washing, and elution, ChIP DNA was purified by phenol:chloroform:isoamyl alcohol and solubilized in 70 µl of water. The ChIP DNA sample was run in 12% PAGE and the 100-300bp DNA fraction was excised and eluted from the gel slice overnight at 4 °C in 300 µl of elution buffer (5:1, LoTE buffer (3 mM Tris-HCl, pH 7.5, 0.2 mM EDTA)-7.5 M ammonium acetate) and was purified using a QIAquick purification kit (Qiagen, Cat#28104). The library was constructed using Illumina genomic DNA prep kit by following its protocol (Illumina, cat# FC-102-1002), clusters were generated on the Illumina cluster station (Illumina, cat# FC-103-1002), DNA samples (20 nM per sample) quantified by an Agilent Bioanalyzer, were loaded onto Illumina Genome Analyzer IIx (GAIIx) for sequencing according to the manufacturer's protocol. Reads generated from the Illumina GAIIx pipeline were aligned to the Human Genome Assembly (NCBI build 36.1/hg18) using ELAND algorithm.

2. ChIP-seq data peak-calling analysis:

After mapped to the corresponding reference genome (HG18), the processed data sets need peak-calling and further identification of ER binding sites, currently there are quite a few programs and scripts available for such tasks, such as FindPeaks [1], Quest [2], and MACS [3], etc. Our ChIP-seq data sets were performed peak-calling by the in-house program, wBELT, developed in our laboratory [4].

And more details on performance comparison and implementation can be retrieved from the corresponding references.

3. Bayesian multivariate statistical modeling of genetic transcription rates:

In most cases, Bayesian statistical models are concerned with learning the parameter set θ=(θ_1, …, θ_l) of the dimension l, containing uncertain quantities (fixed and random effects), hierarchical parameters (hyperparameters), unobserved and/or latent variables, and even missing data [5].

In the Bayesian statistical modeling schema, prior information about the model parameter is denoted by the probability density function $p(\theta)$, the likelihood function is represented as $p(X|\theta)$, and the inference purpose is to derive the posterior density function $p(\theta|X)$. According to the Bayes theorem, the general inference can be formulated as,

$$p(\theta \mid X) = \frac{p(X \mid \theta) p(\theta)}{p(X)} \tag{1}$$

For the multivariate case, $p(X)$ can be specified by the decomposition,

$$p(X) = \prod_{i=1}^{n} p(X_i \mid Pa(X_i)) \tag{2}$$

which can be regarded as a normalized constant by integrating over all values of θ in the product $p(X|\theta)p(\theta)$. Thus, the equation above can be formulated as,

$$p(\theta \mid X) \propto p(X \mid \theta)p(\theta) \tag{3}$$

The strength of the observation data and corresponding prior knowledge influence those diverse weights on the beliefs inferred from the multiple sources.

Due to a relatively small sample size of E2-stimulated time-series gene expression data which contain no knowledge of transcription factors, binding information and direct target genes, the conventional Bayesian modeling has its limitation. In order to infer ERα-centered regulatory network, we integrate the time-series E2-stimulated ERα ChIP-seq data, where it can detect transcription factors and hubs, and facilitate the further reverse-engineering of the regulatory network by means of inferring parameters in the Bayesian statistical framework.

Herein we propose a Bayesian multivariate statistical approach for modeling the time-variant ERα transcriptional regulatory network. The basic model framework is illustrated as follows,

$$\dot{y}_i(t) = \sum_{i,j} \alpha_{ij} x_j(t) + \varepsilon,\ i = 1,\ldots,M, j = 1,\ldots,N \tag{4}$$

where $\dot{y}_i(t)$ denotes the ith gene's transcription rate, $x_j(t)$ for the jth gene's expression level at the investigated time, α_{ij} for the corresponding regulatory argument or strength of the jth gene which has any possible transcription regulatory activity on the ith gene, and ε represents the potential stochastic effects during the transcription regulatory process, which normally follows a normal distribution, *i.e.* $\varepsilon \sim N(0, \sigma^2)$.

Thus for a genetic regulatory network containing M transcription factors at T time points, the above equation can be organized as,

$$\dot{Y}_{M\times T} = [AX]_{M\times T} + \Xi_{M\times T} \tag{5}$$

where $\dot{Y}=(\dot{y}_1\,\dot{y}_2 \ldots\, \dot{y}_M)'$ denotes the transcription rate matrix of M transcription factors, $A=(a_1\; a_2\; \ldots\; a_M)'$ denote the regulatory coefficient matrix, $X=(x_1\; x_2\; \ldots\; x_N)'$ gene matrix and $\Xi=(\varepsilon_1\; \varepsilon_2\; \ldots\; \varepsilon_M)'$ the error term. Thus, inference of coefficient matrix A in the above equation is to acquire concrete knowledge about the transcription regulatory strength of transcription factors over diverse target genes under investigation. Following the Bayes theorem, the above equation can be formulated as,

$$(\dot{Y} \mid A, X)_{M\times T} = A_{M\times N} X_{N\times T} + \Xi_{M\times T} \tag{6}$$

and the error term Ξ is specified as an independent and normally distributed random vector with N-dimensional zero mean and a $N \times N$ covariance Σ. And the multivariate N-dimensional normal distribution for errors is represented as,

$$p(\varepsilon_i \mid \Sigma) \propto |\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}\varepsilon_i' \Sigma^{-1} \varepsilon_i} \tag{7}$$

where ε_i is the N-dimensional error vector. And from the multivariate normal error specification, the observation vector also follows a multivariate normally distribution, denoted as,

$$p(\dot{y} \mid A, x_i, \Sigma) \propto |\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}(\dot{y}-Ax_i)'\Sigma^{-1}(\dot{y}-Ax_i)} \tag{8}$$

then with extension to matrix model representation, the above equation can be formulated as,

$$p(\dot{Y} \mid A, X, \Sigma) \propto |\Sigma|^{-\frac{1}{2}}\, e^{-\frac{1}{2}tr(\dot{Y}-AX)'\Sigma^{-1}(\dot{Y}-AX)} \tag{9}$$

where the trace operator *tr* defines the sum of the diagonal entries of its matrix argument. Here, we denote the conditional prior distribution for A as $p(A|\Sigma)$, which follows normal distribution, denoted as,

$$p(A \mid \Sigma) \propto f(|M|, |\Sigma|) e^{-\frac{1}{2}trM^{-1}(A-A_0)'\Sigma^{-1}(A-A_0)} \tag{10}$$

And the prior distribution for the error covariance matrix Σ as $p(\Sigma)$, and normally the inverse Wishart distribution (IWD) [6,7] is considered as the conjugate prior for the covariance matrix Σ of the multivariate normal distribution, depicted as,

$$p(\Sigma) \propto |\Sigma|^{-\frac{d}{2}} e^{-\frac{1}{2} tr \Sigma^{-1} S} \tag{11}$$

where M, d and S are hyperparameters. According to the Bayes theorem, the joint posterior distribution of the coefficient matrix can be denoted as,

$$p(A, \Sigma \mid \dot{Y}, X) \propto p(\Sigma) p(A \mid \Sigma) p(\dot{Y} \mid A, X, \Sigma) \tag{12}$$

For the multivariate modeling, inference of the coefficient matrix contains two sub-problems, *i.e.* to determine the parameters' posterior distributions for the model, and then to apply the marginal posterior mean estimation for the parameter inference. Thus to determine the marginal posterior distributions for coefficient matrix A, the joint posterior distribution can be marginalized with integration on Σ, illustrated as,

$$\begin{aligned} p(A \mid \dot{Y}, X) &= \int_{\Sigma} p(A, \Sigma \mid \dot{Y}, X) d\Sigma \\ &\propto [(\dot{Y} - X)'(\dot{Y} - X) + (A - A_0) D^{-1} (A - A_0)' + Q]^{-\delta} \end{aligned} \tag{13}$$

where $D = (X_0 X_0')^{-1}$, X_0 is the prior input matrix, $\delta=(n-p+q)/2>0$ and Q is a high-order matrix independent of A. The integration for the posterior distribution is taken as the same form as an inverse Wishart distribution in Bayesian multivariate statistical estimation[6,7]. And denote $S(A) = (\dot{Y} - X)'(\dot{Y} - X) + (A - A_0) D^{-1} (A - A_0)' + Q$, where A_0 is the prior mean matrix. Thus the maximum of posterior coefficient mean estimation $\overline{A}$ should satisfy the following partial derivative,

$$\left. \frac{\partial S(A)}{\partial A} \right|_{A=\overline{A}} \triangleq 2(AX - Y)X' + 2D^{-1}(A' + A_0') \Big|_{A=\overline{A}} = 0 \tag{14}$$

Meanwhile, due to the unobservability of the transcriptional rate, $\dot{Y}$, hence the argument can be approximated as follows,

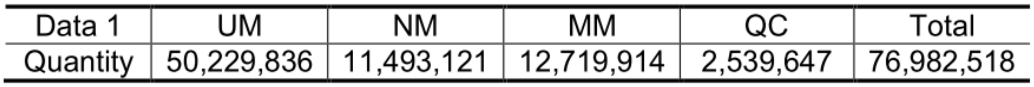
$$\dot{Y}(t) = \frac{dY(t)}{dt} \triangleq \frac{Y(t+\Delta t) - Y(t)}{\Delta t}, \quad \Delta t > 0, t \geq 0 \tag{15}$$

where Δt denotes the sampling interval. Thus the first-order differential of transcription rate can be replaced with the observable vector, $Y(t)$.

Furthermore, due to proportionality characteristics of coefficient matrix inferred by Bayesian statistical analysis, we normalize those coefficients by scaling them with the range within -1 to 1.

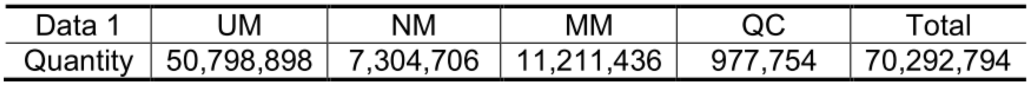
4. Summary of sequenced reads of the E2-stimulated time-series ERα ChIP-seq data:

Time 0 hour:



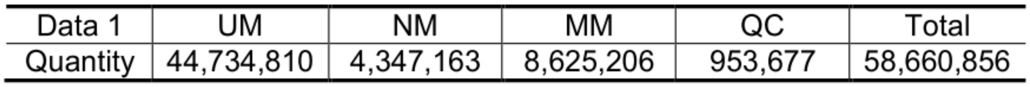
Data 1	UM	NM	MM	QC	Total
Quantity	50,229,836	11,493,121	12,719,914	2,539,647	76,982,518

Time 1 hour:



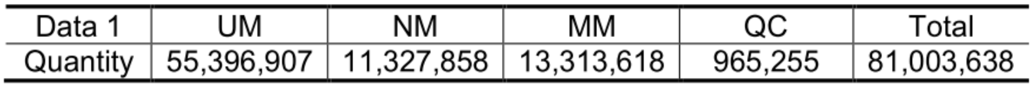
Data 1	UM	NM	MM	QC	Total
Quantity	50,798,898	7,304,706	11,211,436	977,754	70,292,794

Time 4 hours:



Data 1	UM	NM	MM	QC	Total
Quantity	44,734,810	4,347,163	8,625,206	953,677	58,660,856

Time 24 hours:



Data 1	UM	NM	MM	QC	Total
Quantity	55,396,907	11,327,858	13,313,618	965,255	81,003,638

Note: UM: Unique-Match, NM: Non-Match, MM: Multiple-Match, QC: Quality-Control.

5. Summary of identified ERα target genes:

For each time point, *i.e.* 0, 1, 4, 24 hours, we have identified the related gene lists at their corresponding regulatory regions, i.e. transcription start site (TSS), intragenic, proximal regions, etc.

A list of genes with annotation is provided in the **Supplemental file 1** (Excel files).

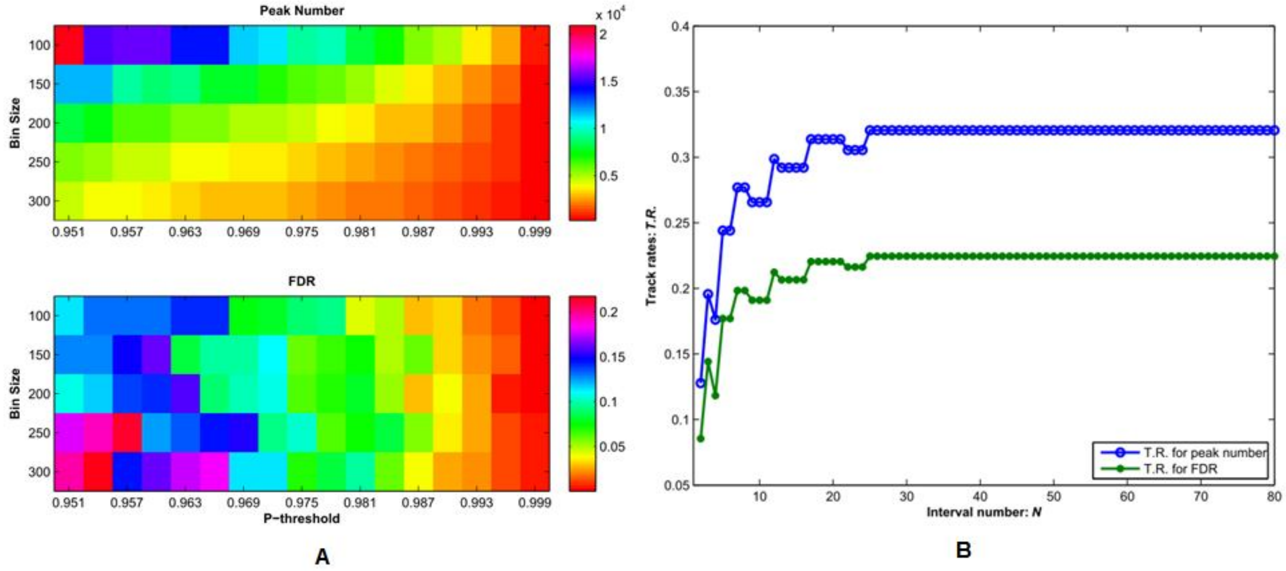


Figure S1. The global peak number and FDR distributions in the optimal argument selection for the ERα ChIP-seq data at time point 0 hour. (A) The upper and lower plots illustrate the peak number and FDR distributions as the related p-threshold varies, respectively; (B) The track rate plots for the peak number and FDR with respect to the interval number N.

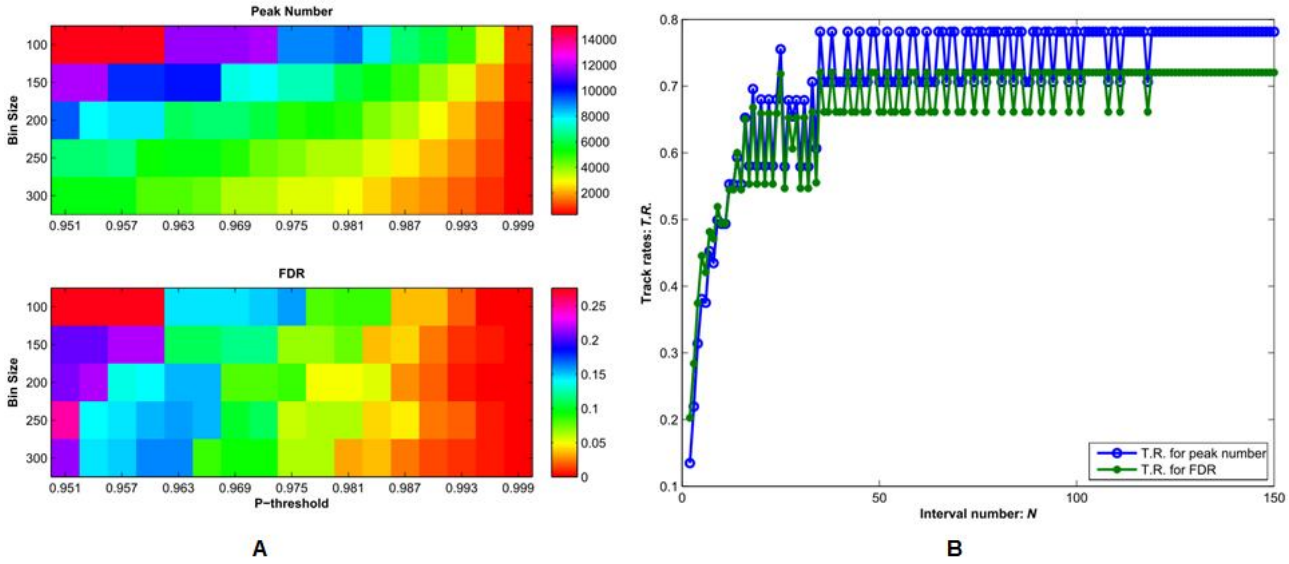


Figure S2. The global peak number and FDR distributions in the optimal argument selection for the ERα ChIP-seq data at time point 1 hour. (A) The upper and lower plots illustrate the

peak number and FDR distributions as the related p-threshold varies, respectively; (B) The track rate plots for the peak number and FDR with respect to the interval number N.

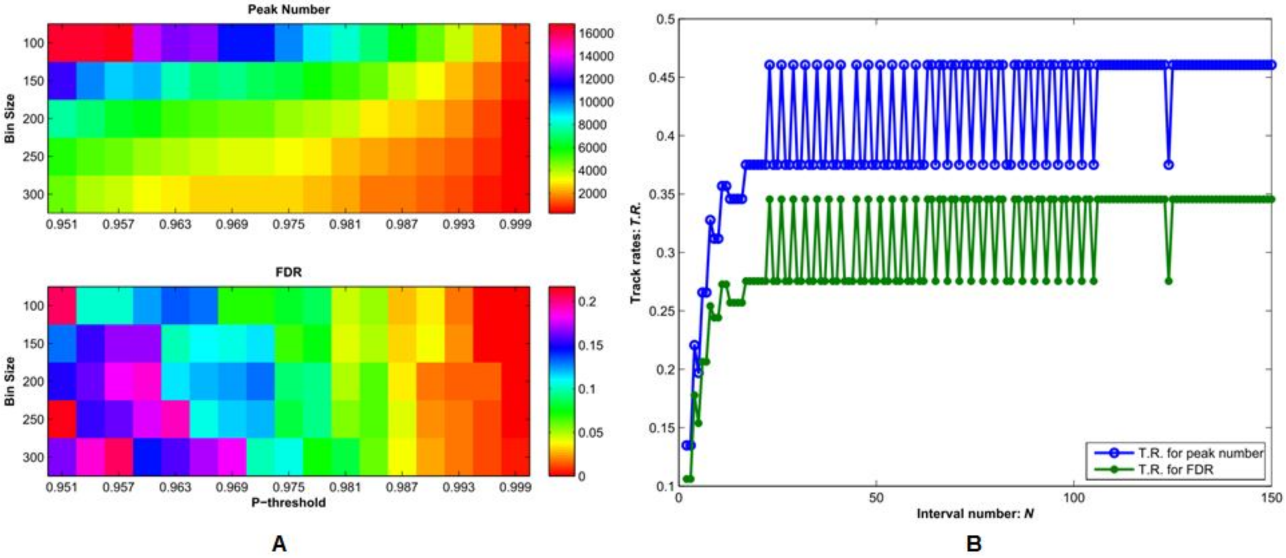


Figure S3. The global peak number and FDR distributions in the optimal argument selection for the ERα ChIP-seq data at time point 24 hours. (A) The upper and lower plots illustrate the peak number and FDR distributions as the related p-threshold varies, respectively; (B) The track rate plots for the peak number and FDR with respect to the interval number N.

6. The inferred network structure and network properties at diverse time points

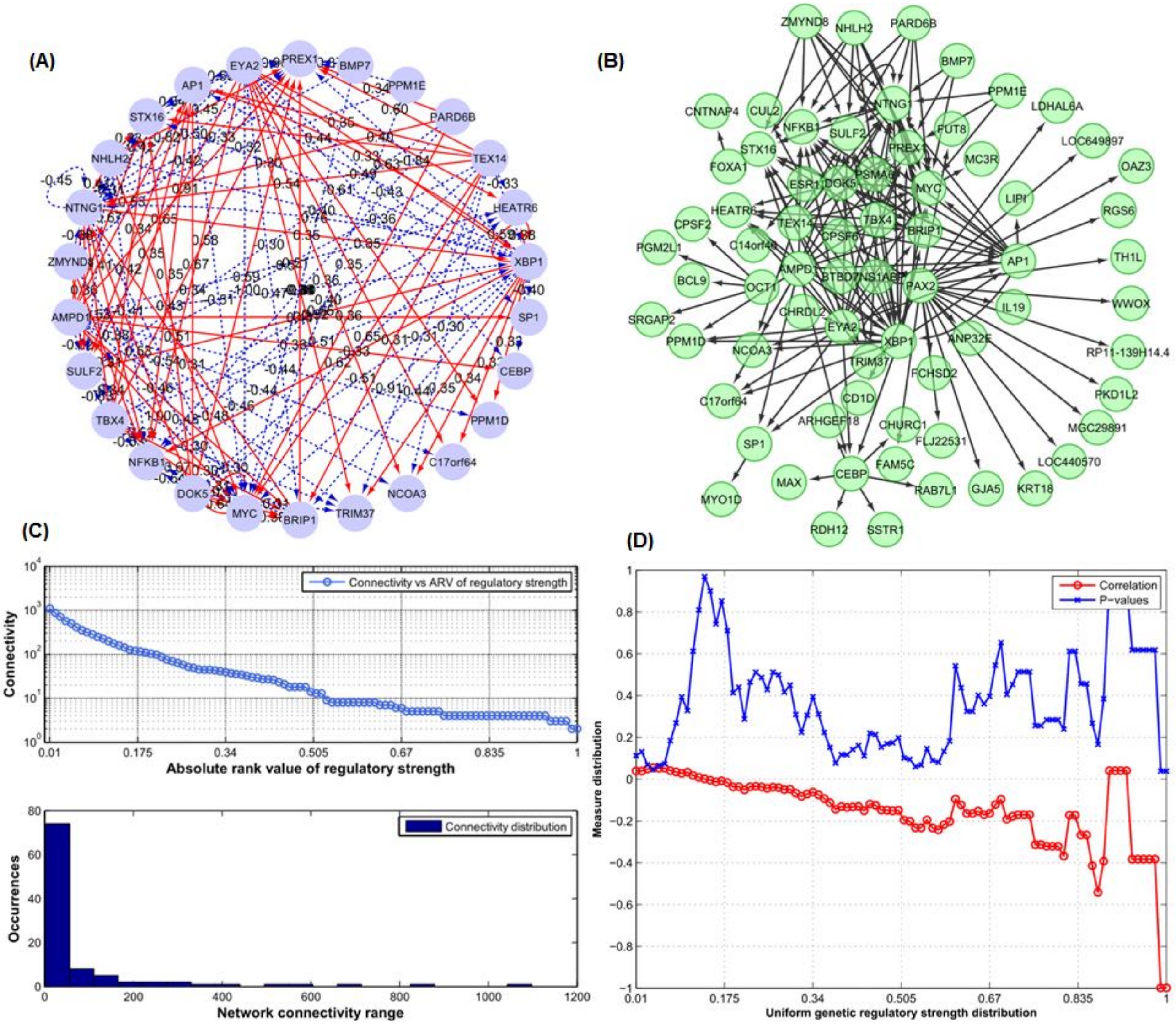


Figure S4. The ERα transcription regulatory network structure and related analysis at time 0 hour. (A) The ERα-centered regulatory network structure at the time 0 hour. The red edges denote positive activation, and dashed blue edges denote negative inhibition; **(B)** The hierarchical topological structure of the inferred ERα transcription regulatory network at time 0 hour; **(C)** and **(D)** illustrate the connectivity distribution, Pearson correlation and p-value distributions (between

the regulatory coefficients and SNRs) as the functions of uniform regulatory strength for the network structure at the time 0 hour.

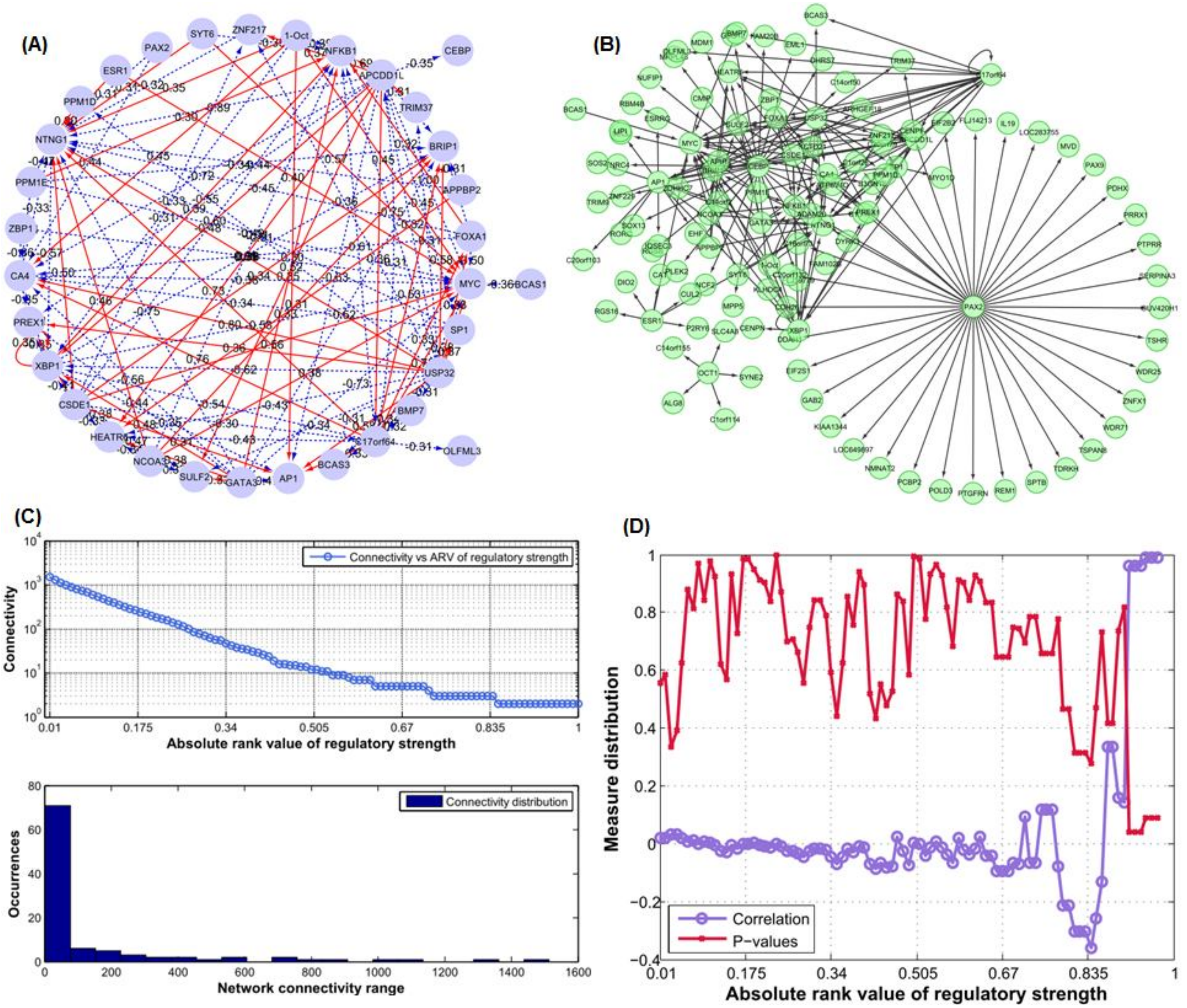


Figure S5. The ERα transcription regulatory network structure and related analysis at time 1 hour. **(A)** The ERα-centered regulatory network structure at the time 1 hour. The red edges denote positive activation, and dashed blue edges denote negative inhibition; **(B)** The hierarchical topological structure of the inferred ERα transcription regulatory network at time 1 hours; **(C)** and **(D)** illustrate the connectivity distribution, Pearson correlation and p-value distributions (between the regulatory coefficients and SNRs) as the functions of uniform regulatory strength for the network structure at the time 1 hour.

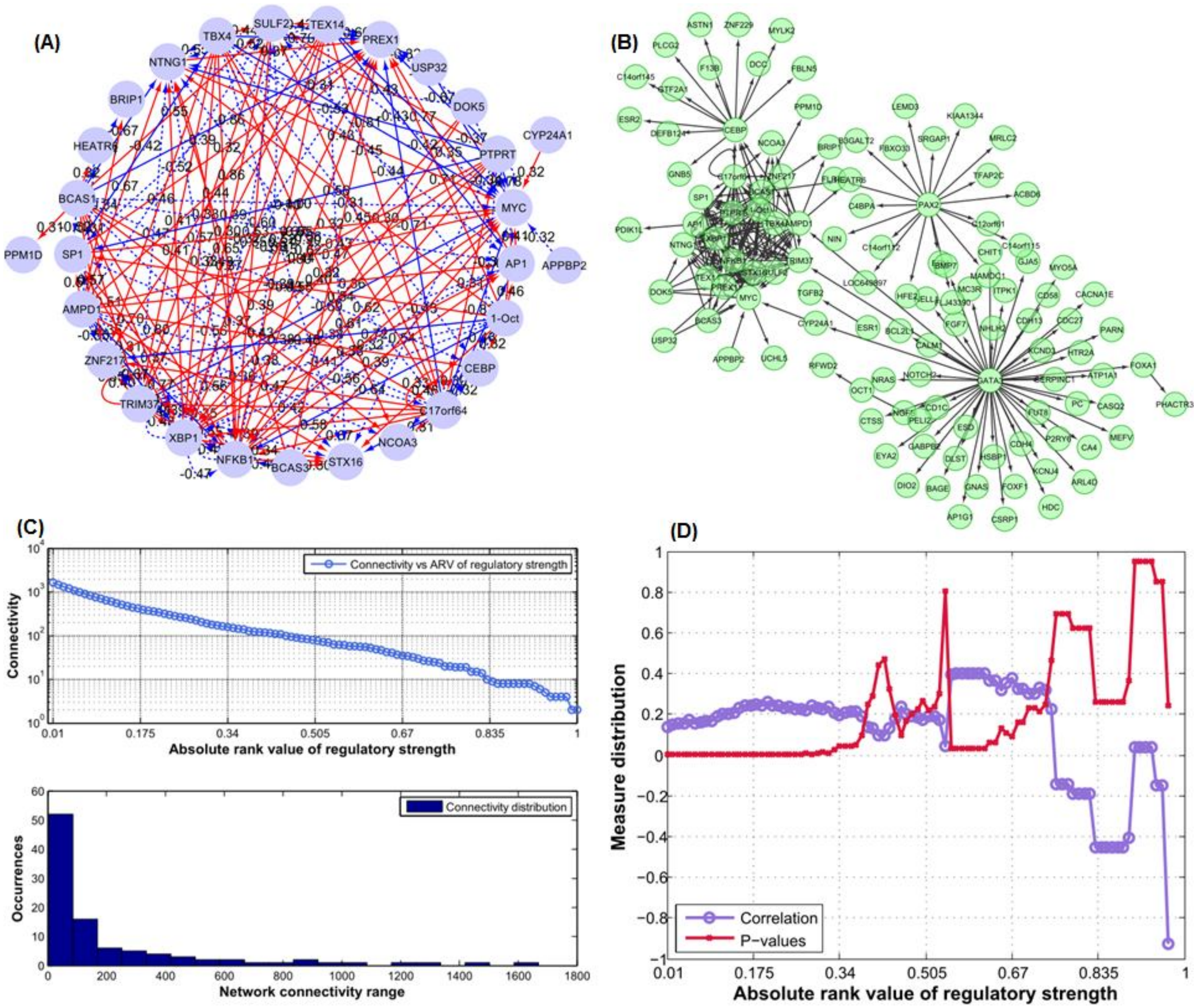


Figure S6. The ERα transcription regulatory network structure and related analysis at the time 24 hours. (A) The ERα-centered regulatory network structure at the time 24 hours. The red edges denote positive activation, and dashed blue edges denote negative inhibition; **(B)** The hierarchical topological structure of the inferred ERα transcription regulatory network at time 24 hours; **(C)** and **(D)** illustrate the connectivity distribution, Pearson correlation and p-value distributions (between the regulatory coefficients and SNRs) as the functions of uniform regulatory strength for the network structure at the time 24 hours.

7. Identification of recurrent regulatory motif patterns:

From the those regulatory modularity analysis, we mainly find the four recurrent regulatory motif patterns, *i.e.* self-loop, interactive loop, feed-forward loop, feedback loop and bi-span loop (**Figure S7**).

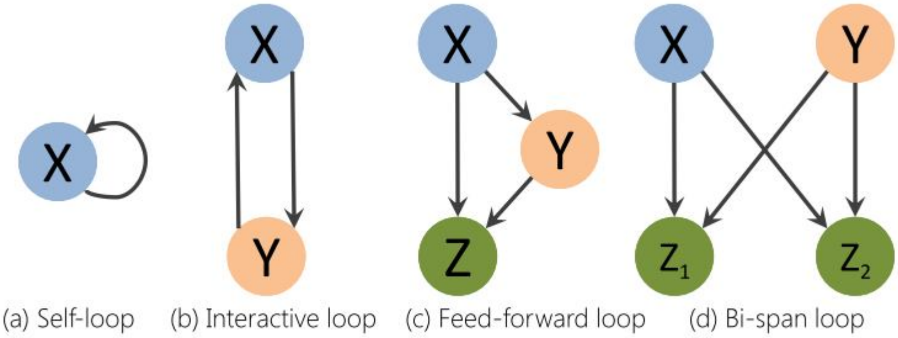


Figure S7. The four recurrent motif patterns identified from the ERα-centered regulatory module, *i.e.*, (a) Self-loop, (b) Interactive loop, (c) Feed-forward loop and (d) Bi-span loop patterns.

Within those recurrent motif patterns, the regulatory edge may denote activation or inhibition activity. For example the (a) self-loop may denote self-activation and self-inhibition, and the (b) interactive loop may represent three cases, i.e. both nodes activate each other, both inhibit each other, or one activates and one inhibits.

Actually, such motif patterns implement specific biological functions in genetic regulatory processes, i.e. (i) Self-regulation: the regulatory activities include simple self-activation or self-inhibition on those genes under investigation, see **Figure S7** (a) self-loop; (ii) Cooperative activation and inhibition regulation: more than two regulators cooperatively regulate (activate or inhibit) one or several target genes, see **Figure S7** (c) feed-forward loop and (d) bi-span; (iii) Feedback mechanism: one or more feedback loops exist in source regulators and target genes.

Such feedback loops directly reflect the regulatory effects on target genes to source regulators, thus simultaneously the source regulators self-tune their regulatory activities. Normally there exists a dynamic equilibrium between source and target genes' activities, see **Figure S7** (b) interactive loop.

8. Gene expression characteristics regulated by Modules I and II:

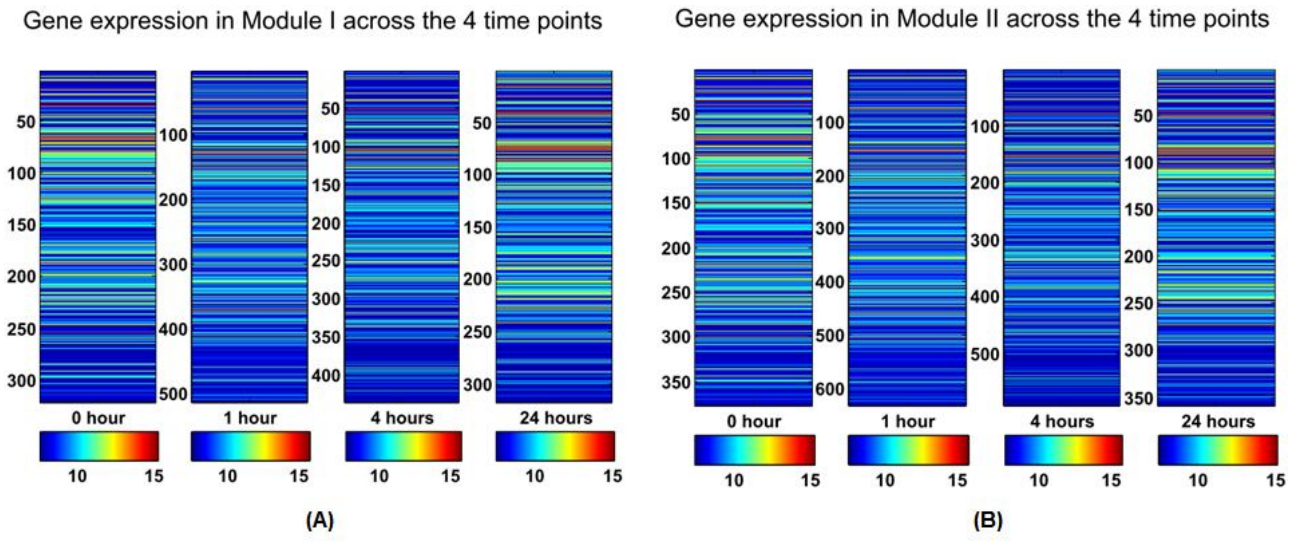


Figure S8. The expression profile plots for the regulated genes by Modules I and II across the four time points. (A) At the time point 0 hour, Module I directly regulates 321 genes; at the time point 1 hour, 512 genes are regulated by Module I; at the time point 4 hours, the regulated genes fall down to 435, and 317 at the time point 24 hours. (B) Module II regulates 376 genes at the time point 0 hour, 632 at the time point 1 hour, 591 at the time point 4 hours and 358 at the time point 24 hours.

9. Gene Ontology (GO) analysis on Modules I and II:

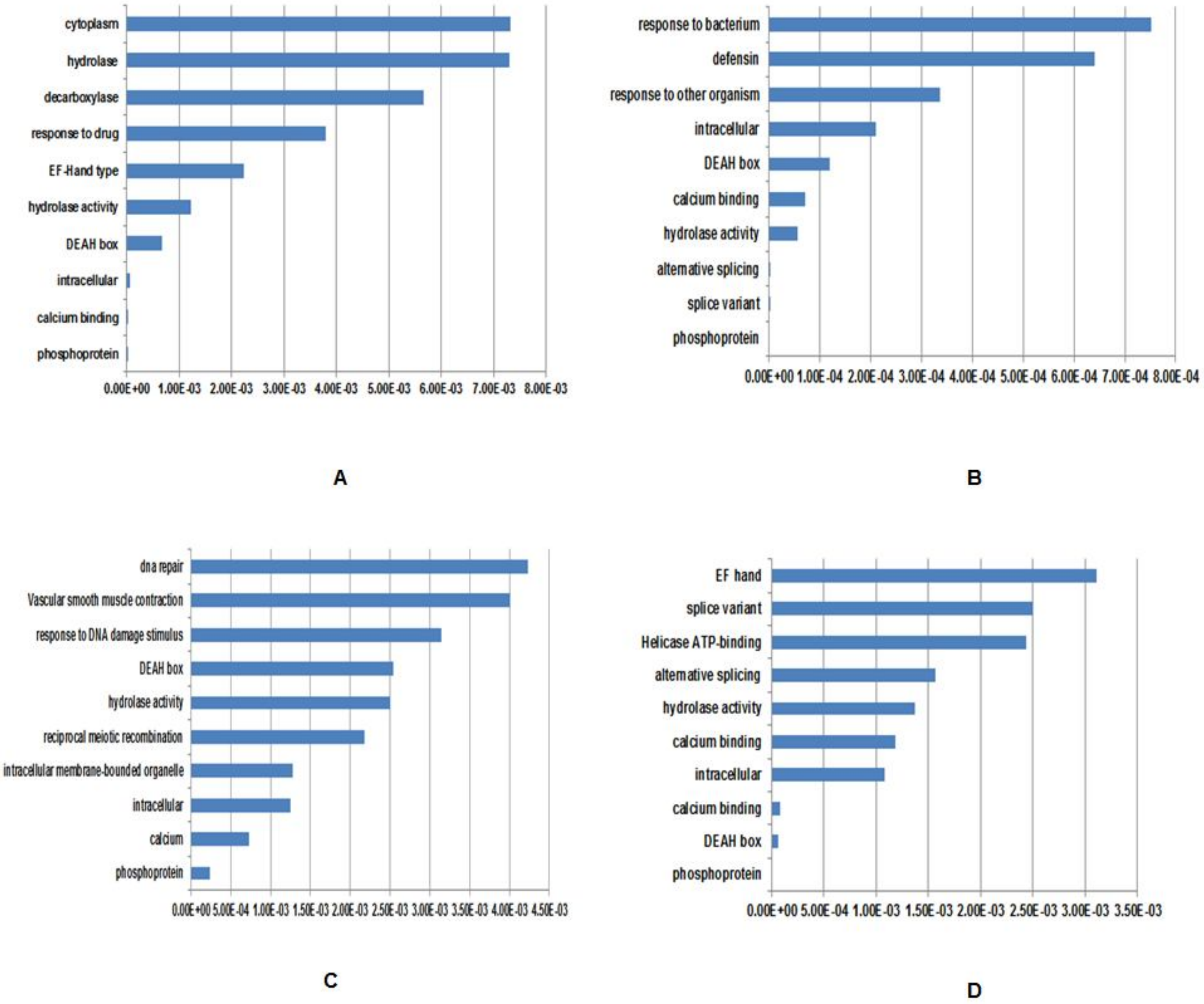


Figure S9. The GO (Gene Onotology) annotation results for Module I across the four time points. On these plots, the horizontal axes denot the calculated p-values, and the vertical axes illustrate the GO annotation terms for each time points, (A) for time 0 hour, (B) for 1 hour, (C) for 4 hours and (D) for 24 hours.

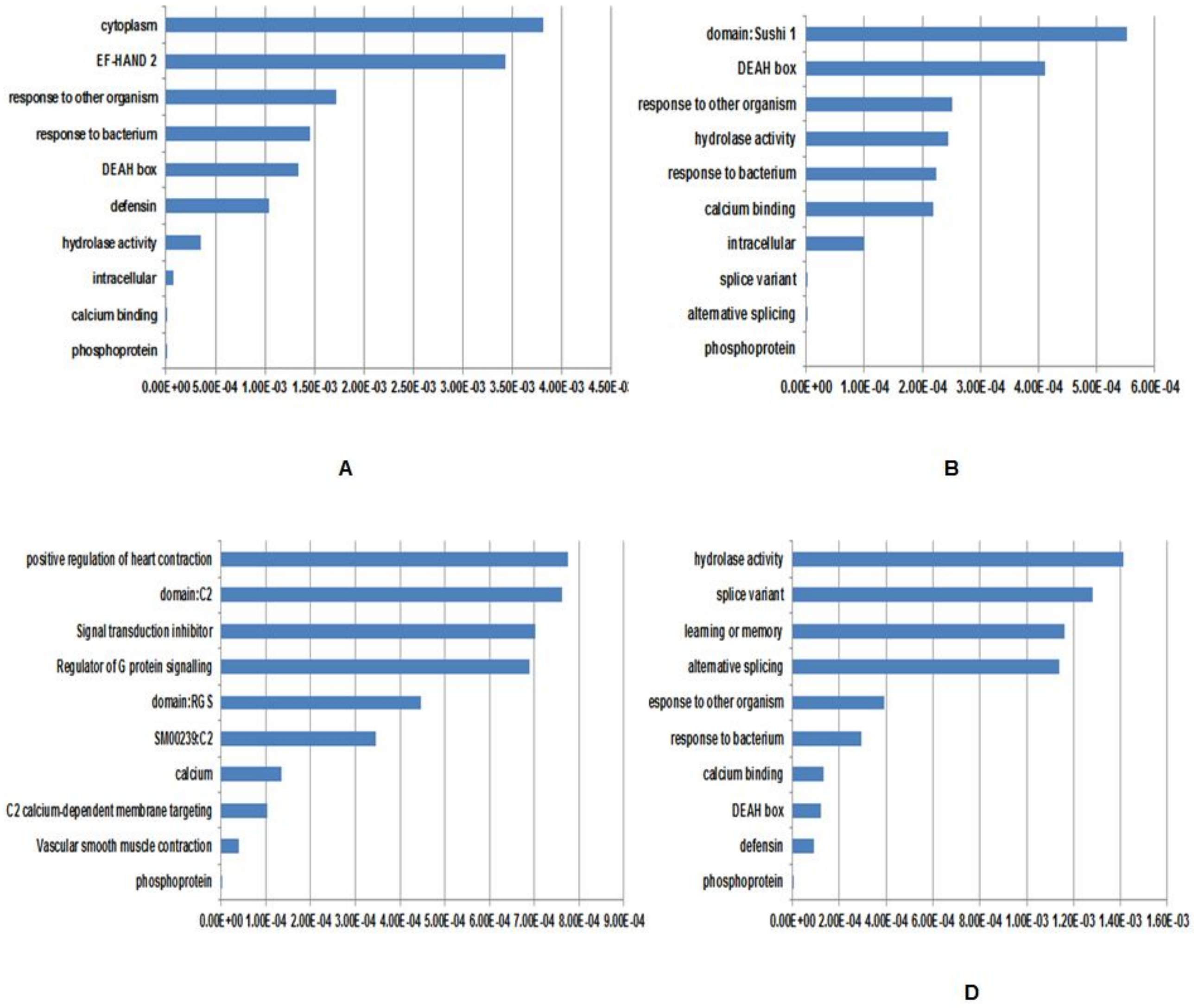
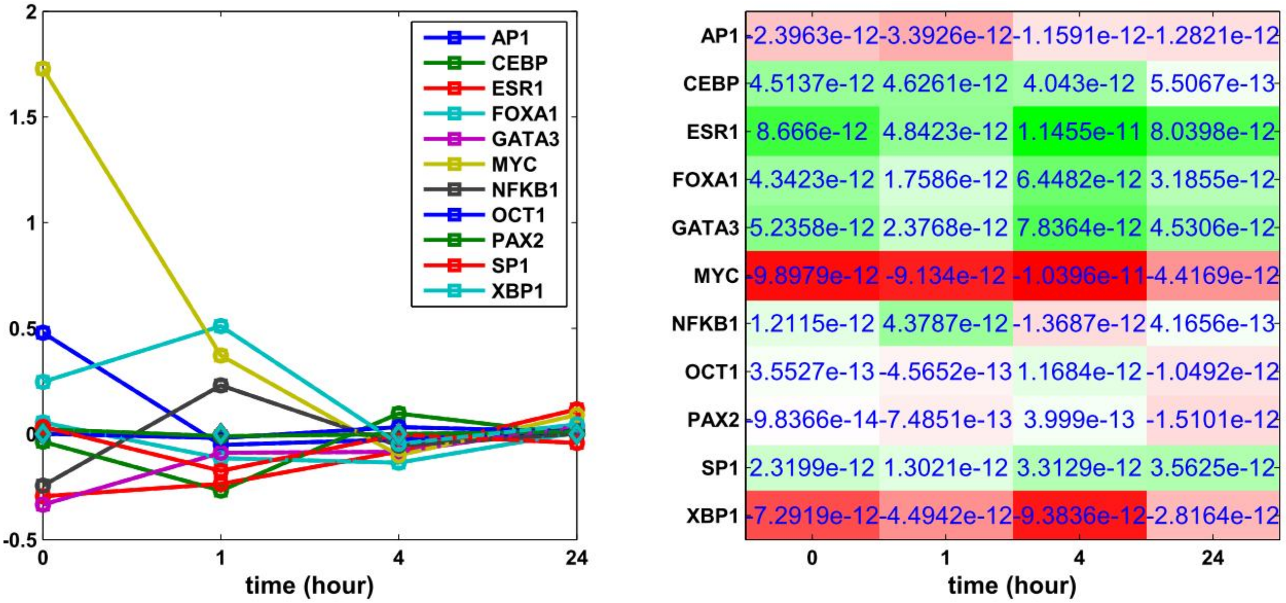


Figure S10. The GO (Gene Onotology) annotation results for Module II across the 4 time points. On these plots, the horizontal axes denot the calculated p-values, and the vertical axes illustrate the GO annotation terms, (A) for time 0 hour, (B) for 1 hour, (C) for 4 hours and (D) for 24 hours.

10. Time-series network inference error plots:



	0	1	4	24
AP1	-2.3963e-12	-3.3926e-12	-1.1591e-12	-1.2821e-12
CEBP	4.5137e-12	4.6261e-12	4.043e-12	5.5067e-13
ESR1	8.666e-12	4.8423e-12	1.1455e-11	8.0398e-12
FOXA1	4.3423e-12	1.7586e-12	6.4482e-12	3.1855e-12
GATA3	5.2358e-12	2.3768e-12	7.8364e-12	4.5306e-12
MYC	-9.8979e-12	-9.134e-12	-1.0396e-11	-4.4169e-12
NFKB1	1.2115e-12	4.3787e-12	-1.3687e-12	4.1656e-13
OCT1	3.5527e-13	-4.5652e-13	1.1684e-12	-1.0492e-12
PAX2	-9.8366e-14	-7.4851e-13	3.999e-13	-1.5101e-12
SP1	2.3199e-12	1.3021e-12	3.3129e-12	3.5625e-12
XBP1	-7.2919e-12	-4.4942e-12	-9.3836e-12	-2.8164e-12

Figure S11. The inference errror plot for the 11 TF hubs across the four time points. It is based on Equation (6), i.e. regulatory coefficients, gene expression and transcription rates. The left panel gives the transcript rates for each TF hubs and the left one illustrates the inference errors for those TF hubs across the four time points.

11. The Kaplan-Meier survival analysis based on diver patient cohorts:

For the clinical outcome analysis, the gene signatures are the specific genes that characterized the diverse clustered patient subgroups.

In determining the gene signature and clinical outcome analysis, we firstly adopted the *k*-means clustering approach to analyze the three patient cohorts containing 337, 251 and 137 patients, respectively. Since initially we have no concrete group information about the signatures, actually the *k*-means clustering here is an unsupervised approach.

A too small subgroup number may cluster diverse clinical pathological features into a single group and a too large subgroup number may render the similar clinical pathological features in two or more subgroups, thus the both situations cannot ensure the statistically sounding results. In the exploratory analysis we selected diverse cluster numbers (i.e. 2, 3, 4 and 5, etc.), then each clustered subgroups containing gene signature information were further processed with the Kaplan-Meier survival probability analysis and corresponding log-rank test. Based on the analysis results on the three diverse patient cohorts, we found the clustered group number 4 was significantly capable of rendering statistically meaningful results (with each log-rank test p-value < 0.05), thus we reported the analysis results with the group number 4 in the work.

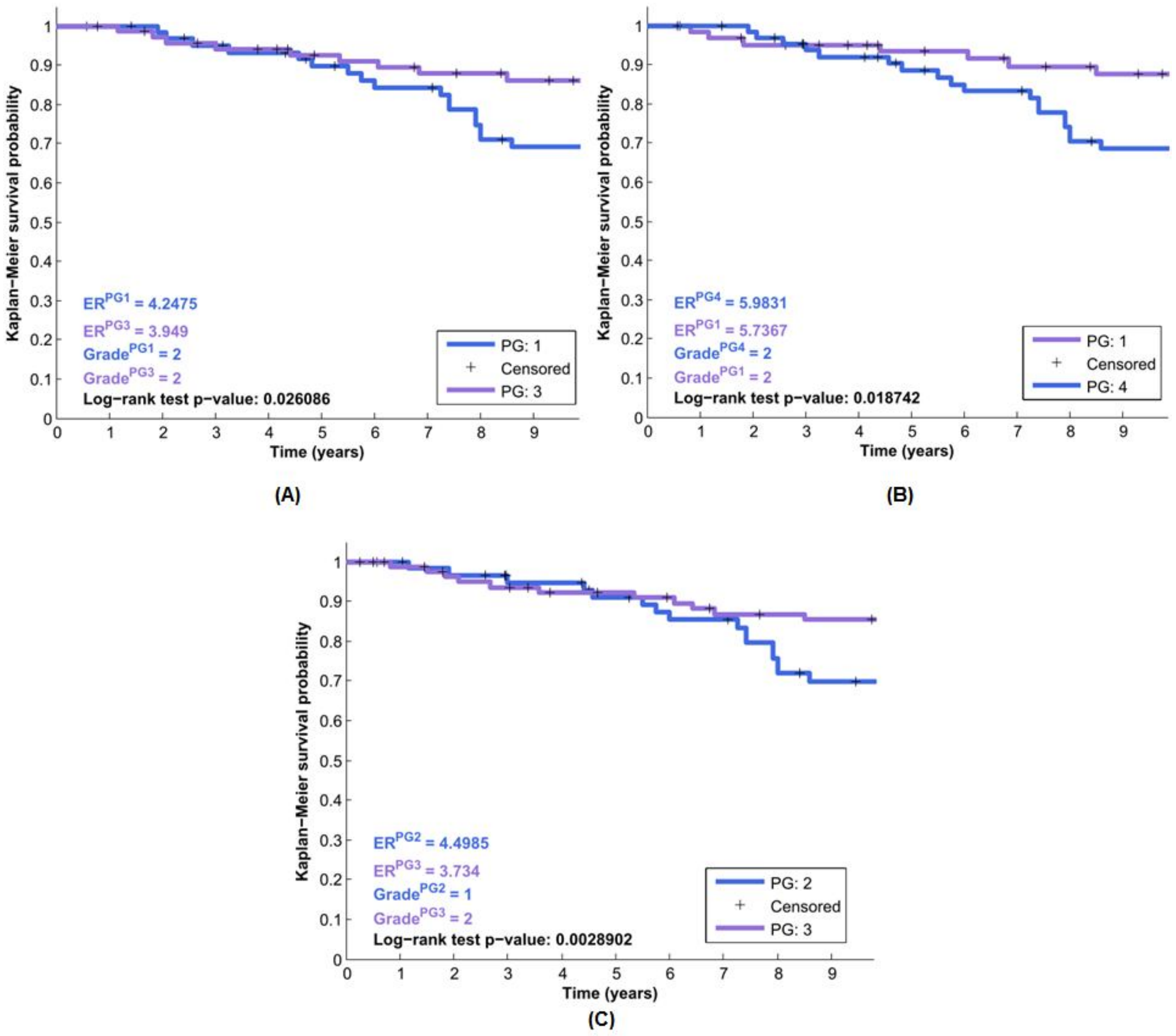
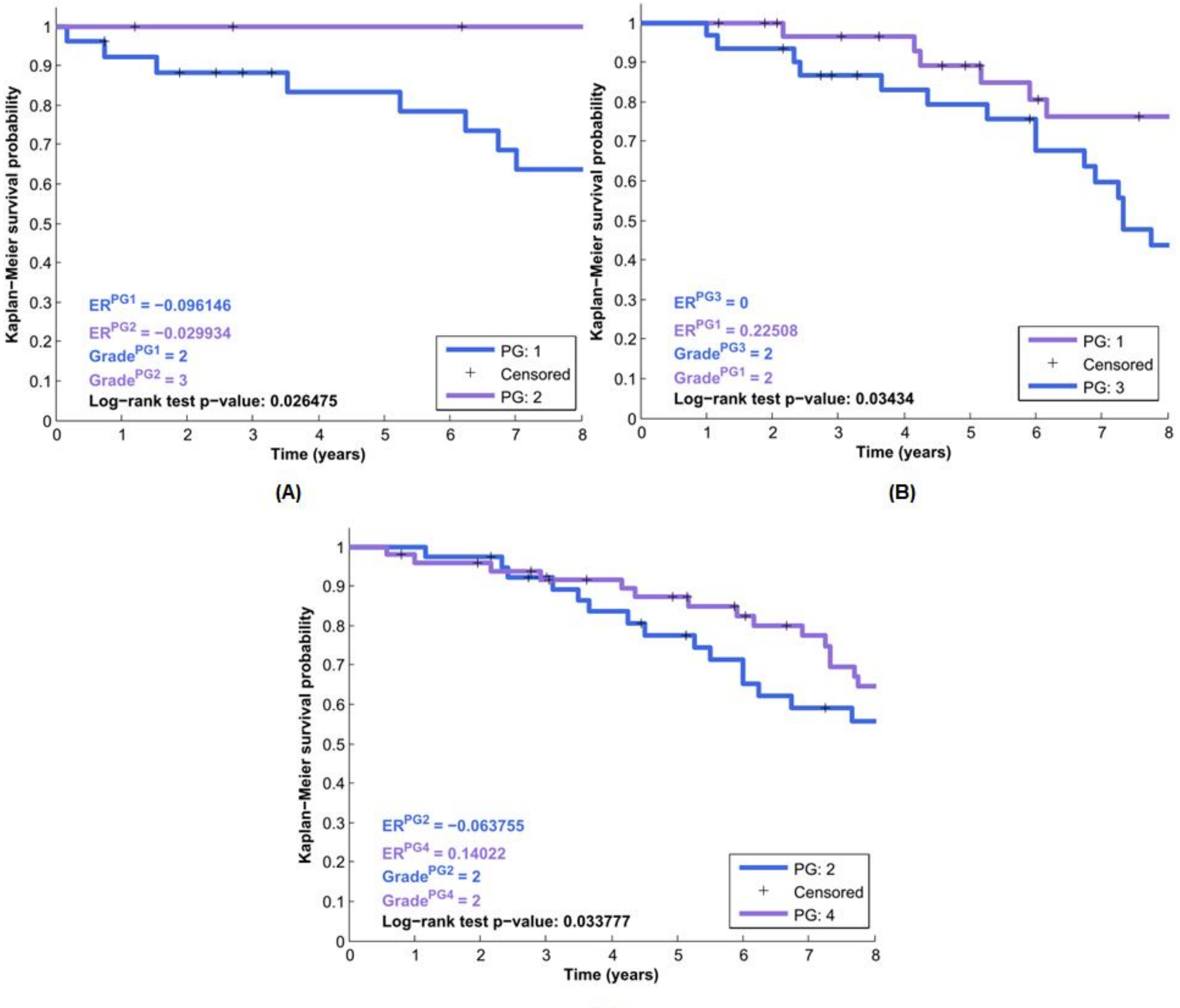


Figure S12. The Kaplan-Meier survival analysis based on the regulated target genes by the three modules. The clinical survival information of 251 breast cancer patients is selected from Miller *et al.* [8]. The subplot **(A)** gives the statiscally siginificant result between the patient groups PG:1 vs PG:3 (Module I), and their corresponding group estrogen receptor status and survival stage (grade) information are also provided on the left bottom (log-rank test p-value: 0.026086). The subplots **(B)** and **(C)** depict the analysis results on the patient groups PG:1 vs PG:4 (Module II), and PG:2 vs PG:3 (Module III), respectively.



(A)

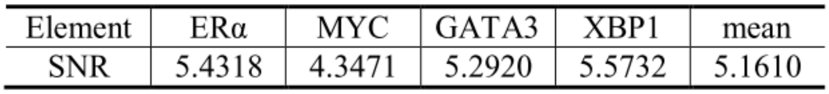
(B)

(C)

Figure S13. The clinical survival analysis based on the regulated target genes by three modules. The clinical survival information of 137 breast cancer patients is selected from Sotiriou *et al.* [9]. The subplot **(A)** gives the statiscally siginificant result between the patient groups PG:1 vs PG:2 (Module I), and their corresponding group estrogen receptor status and survival stage (grade) information are also provided on the left bottom (log-rank test p-value: 0.026475). The subplots **(B)** and **(C)** depict the analysis results on the patient groups PG:1 vs PG:3 (Module II), and PG:2 vs PG:4 (Module III), respectively.

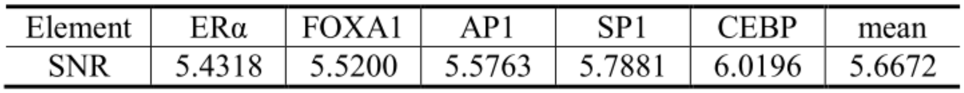
12. Supplemental Tables:

Table S1. The network elements' SNR statistics (dB) in Module I.



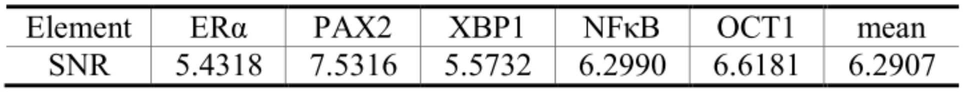
Element	ERα	MYC	GATA3	XBP1	mean
SNR	5.4318	4.3471	5.2920	5.5732	5.1610

Table S2. The network elements' SNR statistics (dB) in Module II.



Element	ERα	FOXA1	AP1	SP1	CEBP	mean
SNR	5.4318	5.5200	5.5763	5.7881	6.0196	5.6672

Table S3. The network elements' SNR statistics (dB) in Module III.



Element	ERα	PAX2	XBP1	NFκB	OCT1	mean
SNR	5.4318	7.5316	5.5732	6.2990	6.6181	6.2907

References

1 **Fejes, A. P. *et al.* FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics* 24, 1729-1730 (2008).**

2 **Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth* 5, 829-834 (2008).**

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

4 **Lan, X., Bonneville, R., Apostolos, J., Wu, W. & Jin, V. X. W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics* 27, 428-430 (2011).**

5 **Gelman, A. & Rubin, D. B. Markov chain Monte Carlo methods in biostatistics. *Statistical Methods in Medical Research* 5, 339-355 (1996).**

6 **O'Hagan, A. & Forster, J. J. *Kendall's advanced theory of statistics: Bayesian inference*. 2nd edn, (Wiley, John & Sons, 2004).**

7 **Rowe, D. B. *Multivariate Bayesian statistics: Models for source separation and signal unmixing.*, (CRC Press, 2003).**

8 **Miller, L. D. *et al.* An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS* 102, 13550-13555 (2005).**

9 **Sotiriou, C. *et al.* Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute* 98, 262-272 (2006).**