A Model-based Approach to Transcription Regulatory Network Reconstruction from Time-Course Gene Expression Data

Hong Hu* and Yang Dai*

Abstract— **Time-course gene expression profiling provides valuable data on dynamic behavior of cellular responses to external stimulation. Investigation of transcription factors (TFs) that regulate co-expressed genes in a dynamic process can reveal insights on the underlying molecular mechanisms. As the ChIP-seq technology is only suitable for a fraction of TFs in mammalian organisms, the computational identification of relevant TFs remains to be critical. We propose a regression-based model to infer the functional binding sites of TFs from time-course gene expression profiles. Our approach incorporates an association strength for each potential TF and target gene pair based on computational analysis of binding sites in promoter sequences of co-expressed genes. Our model further uses the Lasso-penalized technique to search for the most informative TF-target pairs. The application of our method to a gene expression study on E2-induced apoptosis in a variant of MCF-7 cells revealed that the findings are biologically meaningful.**

I. INTRODUCTION

Numerous studies have shown that transcription factors (TFs) play a key role in regulation of gene expression by binding to the DNA sequences of their target genes in a sequence-specific fashion [1]. These bindings might occur on either proximal or distal regulatory regions to the transcription start sites (TSSs) of the target genes. To investigate TF binding profiles at genome scale, experimental and computational approaches have been developed [2]. Chromatin immunoprecipitation followed by microarray (ChIP-chip) or massively parallel sequencing (ChIP-seq) are used as major experimental methods to capture global binding a given TF under specific condition [3], [4], [5]. However, they are limited by the availability of TF-specific antibody and prohibited by the high cost. On the other hand, Position Weight Matrices (PWMs) are employed to predict transcription factor binding sites (TFBSs) in potential regulatory regions [6], [7]. The established putative regulation relationship between TFs and target genes can be combined with gene expression time course data for the reconstruction of transcription regulation networks (TRNs) [8].

Earlier works attempt to model the expression of a specific gene as a function of the expression of other genes (possibly TF coding genes only) based on dynamic Bayesian networks or other statistical methods [9], [10], [11]. Other approaches include taking gene expression as linear combination of activities of other genes through network component analysis [12], [13], [14].

*Department of Bioengineering (MC563), University of Illinois at Chicago, 835 S. Wolcott St, W100 CSN Chicago, IL 60612, USA {hhu4,yangdai} at uic.edu

Our work employs a regression-based framework to model expression of a gene at a time point as the linear combination of expression of TF-coding genes and non-TF coding genes at the previous time point based on the putative TRN identified from predicted TFBSs in promotor regions of co-expressed genes. An association strength is further devised to provide a weight on each TF and target gene pair based on similarity scores of TFBSs of the TF. From the sparsity of a TRN and the limited number of time points in a time course experiment, LASSO penalized regression [15] is proposed for identifying a small set of connections in the putative TRN. The regression coefficients obtained from the model provide information on how a gene is regulated by its regulating TF coding genes and other non-TF genes. Therefore, the feature of our model is the ability to separate regulatory effects of TF coding genes from those of non-TF genes. The dissection of TF regulatory effect may provide better understandings on potential regulatory mechanism underlying the observed time-course gene expression data.

II. METHODS

A. Association strength of a TF to a target gene

To predict TFBSs in the promoter regions of a set of co-expressed genes, the cREMaG database [16] was queried. cREMaG are built upon searching promoter regions (10 kb upstream and 2 kb downstream of a TSS) of genes with the PWMs from TRANSFAC [17] or JASPAR [18]. For a query gene set, cREMaG provides the following information:

- Numbers of predicted TFBSs on individual promoter regions of the genes
- Similarity score of each predicted TFBS
- Enrichment fold p-value of a PWM in the gene set

For each PWM, the enrichment of the predicted TFBSs in the gene set are evaluated based on the distribution of random background sequences. A smaller enrichment fold p-value means that the number of TFBSs in the promoter sequences is higher than what is expected from the random sequences. Since a PWM may represent multiple TFs and each TF may have multiple PWMs, we remove PWMs representing multiple TFs and choose the PWM with the highest Information Content (IC) [19] for a TF with multiple PWMs. After applying this procedure, a one-to-one relationship between a TF and a PWM is ensured. In the subsequent description, TF and PWM are used alternatively.

We introduce *association strength* to quantify the aggregated effect of multiple TFBSs found in a promotor for a TF. Suppose there are m TF coding genes and n non-TF coding

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

genes with known gene expression profiles, and the TFBSs of these TFs are significantly enriched in the promoter regions of $n+m$ genes. This setting implies that a TF may also regulate itself.

The *association strength* of TF j $(j = n+1, \ldots, n+m)$ on the promoter of gene i $(i = 1, \ldots, n+m)$ is defined as

$$\alpha_{ij} = -\ln(p_j) \sum_{k=1}^{n_{ij}} s_{ijk} \tag{1}$$

where p_j is the enrichment fold p-value of TF j for the query gene set; n_{ij} is the number of predicted TFBSs of TF j on the promotor of gene i; s_{ijk} is the similarity score of predicted TFBS k of TF j on gene i. The enrichment fold p-value is used to weight the similarity score. If TF j has no predicted TFBS for gene i, $\alpha_{ij} = 0$.

B. LASSO regression model

Assume the expression of gene i at time point $t+1$ is proportion to the expression of all non-TF coding genes and the expression of all TF coding genes weighted by corresponding association strength defined in equation (1), at time point t. Let $y_{i,t}$ be the expression of gene i at time point t $(t = 1, ..., T)$. Denote $NTF = \{1, ..., n\}$ and $TF = \{n+1, ..., n+m\}$. The expression of gene i $(i = 1, ..., n+m)$ at time $t+1$ can be written as

$$y_{i,t+1} = \sum_{j \in NTF} \beta_{ij}^{NTF} y_{j,t} + \sum_{j \in TF} \beta_{ij}^{TF} \alpha_{ij} y_{j,t} + \epsilon_{i,t}$$

where β_{ij}^{NTF} is the regulatory effect of expression of non-TF coding gene j on expression of gene *i*; β_{ij}^{TF} is the regulatory effect of expression of TF coding gene j on expression of gene i; and $\epsilon_{i,t}$ is the error term. Obviously, if TF *j* has no predicted TFBS on gene *j*, $\alpha_{ij} = 0$ implies $\beta_{ij}^{TF} = 0$.

Our model learns the regulatory effects β_{ij}^{NTF} and β_{ij}^{TF} with the presence of the association strength α_{ij}. If a coefficient β_{ij}^{TF} is very close to zero, we conclude that no regulatory effect of TF j on its target gene i can be identified from the observed gene expression even with the presence of the predicted TFBSs. A similar statement can be made for those small coefficient β_{ij}^{NTF}.

The following notations are used to describe our regression model for time course. Given a temporal expression profile $\{y_{i,1}, ..., y_{i,T}\}$ of gene *i*, we define for $t = 1, ..., T$,

$$\boldsymbol{y}_{i,T_{-t}} = \begin{bmatrix} y_{i,1} & \cdots & y_{i,t-1} & y_{i,t+1} & \cdots & y_{i,T} \end{bmatrix}^T \tag{2}$$

$$\boldsymbol{\epsilon_{i,T_{-t}}} = \begin{bmatrix} \epsilon_{i,1} & \cdots & \epsilon_{i,t-1} & \epsilon_{i,t+1} & \cdots & \epsilon_{i,T} \end{bmatrix}^T \tag{3}$$

where $[\,]^T$ is vector transpose. The expression time course of all non-TF coding genes from time point $t = 1$ to $t = T-1$ is written in the following matrix form:

$$\boldsymbol{Y}_{T_{-T}}^{NTF} = \begin{bmatrix} \boldsymbol{y}_{1,T_{-T}} & \cdots & \boldsymbol{y}_{j,T_{-T}} & \cdots & \boldsymbol{y}_{N,T_{-T}} \end{bmatrix}$$

Similarly, the expression time course of all TF coding genes at same time points is written as

$$\boldsymbol{Y}_{T_{-T}}^{TF} = \begin{bmatrix} \boldsymbol{y}_{N+1,T_{-T}} & \cdots & \boldsymbol{y}_{j,T_{-T}} & \cdots & \boldsymbol{y}_{N+M,T_{-T}} \end{bmatrix}$$

Recall that in formula (2) the expression of TF coding gene j is weighted by its association strength on target gene i, i.e., α_{ij}. When expanded into the time course from time point 1 to $T-1$, it can be written as

$$\alpha_{ij}\boldsymbol{y}_{j,T_{-T}} = \begin{bmatrix} \alpha_{ij}y_{j,1} & \cdots & \alpha_{ij}y_{j,T-1} \end{bmatrix}^T$$

Take the association strength of all TF coding genes on target gene *i* as

$$\boldsymbol{\alpha}_i = \begin{bmatrix} \alpha_{i,n+1} & \cdots & \alpha_{i,n+m} \end{bmatrix}$$

The expression time courses of all TF coding genes from time point 1 to $T-1$ weighted by association strength on target gene i can be written as

$$\boldsymbol{\alpha}_i \otimes \boldsymbol{Y}_{T_{-T}}^{TF} = \begin{bmatrix} \alpha_{i,n+1}\boldsymbol{y}_{n+1,T_{-T}} & \cdots & \alpha_{i,n+m}\boldsymbol{y}_{n+m,T_{-T}} \end{bmatrix}$$

We further denote

$$\boldsymbol{\beta}_i^{NTF} = \begin{bmatrix} \beta_{i,1} & \cdots & \beta_{i,n} \end{bmatrix}^T$$
$$\boldsymbol{\beta}_i^{TF} = \begin{bmatrix} \beta_{i,n+1} & \cdots & \beta_{i,n+m} \end{bmatrix}^T$$

With all the notations defined above, the time course of expression of gene i $(i = 1, ..., n+m)$ can be modeled as

$$\boldsymbol{y}_{i,T_{-1}} = \boldsymbol{Y}_{T_{-T}}^{NTF}\boldsymbol{\beta}_i^{NTF} + \boldsymbol{\alpha}_i \otimes \boldsymbol{Y}_{T_{-T}}^{TF}\boldsymbol{\beta}_i^{TF} + \boldsymbol{\epsilon}_{i,T_{-T}}$$

Generally, in a time course microarray gene expression data, the number of time points is much smaller comparing to the number of genes. We propose a LASSO penalized regression model for model section:

$$\min \left\{ \sum_{i=1}^{n+m} \left[\left\| \boldsymbol{y}_{i,T_{-1}} - \boldsymbol{Y}_{T_{-T}}^{NTF}\boldsymbol{\beta}_i^{NTF} - \boldsymbol{\alpha}_i \otimes \boldsymbol{Y}_{T_{-T}}^{TF}\boldsymbol{\beta}_i^{TF} \right\|^2 + \lambda \Big(\sum_{j=1}^{n} |\beta_{ij}^{NTF}| + \sum_{j=n+1}^{n+m} |\beta_{ij}^{TF}| \Big) \right] \right\} \tag{4}$$

where λ is a positive parameter to be determined. The LASSO model aims at the balance between the prediction error of gene expression and the model size, i.e., the number of non-zero regulatory effects. By minimizing the object function for all the genes with a properly chosen λ, the TRN can be determined based on the $(n+m) \times (n+m)$ matrix of regulatory effects $\boldsymbol{\beta} = [\, \boldsymbol{\beta}^{NTF}\ \boldsymbol{\beta}^{TF}]$.

We determined λ by a K-fold cross-validation (CV) as illustrated in Figure 1. It is clear that the model for each gene i can be trained independently from formula (2) and (4). The expression of gene i at time point t depends on the expression of all genes at previous time point $t-1$ according to formula (2), which means that all time points in each subsample are required to be consecutive in the evenly divided split K subsamples over the time points. We randomly took 1 subsample as the testing set and the remaining of the subsamples as the training set, and iterated until all subsamples were tested.

The range of λ was determined as follows. First, we ran a trial LASSO regression with all time points to obtain the ranges of λ for each individual gene. Let $\lambda_{\max}$ be the maximum among all λ values. Then λ range was reset as

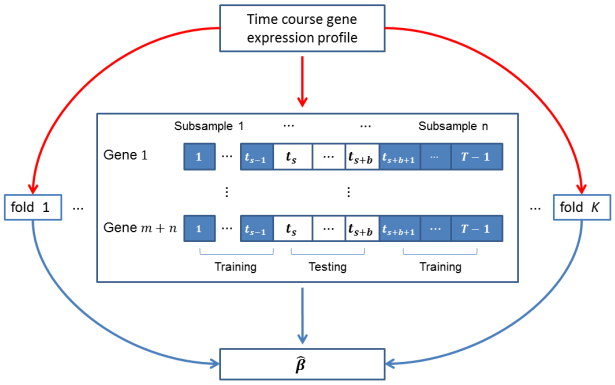

Fig. 1. **Procedure of K-fold cross validation.** The time course are evenly divided into K subsamples by time points. In each subsample, the time points are consecutive. For each testing subsample including expression from time point t_s to t_{s+b} for some positive integer s and b, the rest of K-1 subsamples is used to train the model (2) for individual λ values. After solving the optimization problem and obtaining the optimal solution $\boldsymbol{\beta}^{NTF}(\lambda)$ and $\boldsymbol{\beta}^{TF}(\lambda)$, they are used to predict the gene expression levels for all genes at time point t_s to t_{s+b} based on equation (2). Iterating this process for all K subsamples, the total cross-validated sum-of-square errors for entire time course (except time point 1) is calculated.

$[0, \lambda_{max}/5]$. All values in the range starting from 0 with an increment of $\lambda_{max}/300$ were used in the subsequent K-fold CV procedure. In our experiment, we set $K = 10$. We repeated this procedure for all possible λ values in the above range and determined the best λ value as the one with the smallest CV validated sum-of-square errors. Finally, the $\hat{\boldsymbol{\beta}}^{TF}$ and $\hat{\boldsymbol{\beta}}^{NTF}$ values obtained from the K training folds were averaged to identify the network learned from the LASSO model. The "parallel" package in R was used to solve the model (4) for all genes simultaneously.

III. Results

A. Dataset

We applied our method to a microarray gene expression dataset [20] (GEO access id: GSE29917), in which the apoptosis of MCF-7 variants in response to stress was investigated. These variants are either estrogen-dependent for growth (MCF-7:WS8), or resistant to estrogen deprivation and refractory (MCF-7:2A) or sensitive (MCF-7:5C) to E2-induced apoptosis. Each cell line was treated with 10^{-9}M E2 or vehicle control over a 96h time course consisting of 7 time points. cRNA probes from individual E2-treated samples were competitively hybridized against time-matched pooled control probes using 2-color Agilent 4×44k human oligonucleotide microarrays. Gene expression at 2h, 6h, 12h, 24h, 48h, 72h, and 96h for 6 biological replicates per condition were observed.

B. TRN reconstruction

There are 1,142 genes identified as differentially expressed (DE) for the MCF-7:5C cell line in the original paper [20]. We extracted the expression of DE genes data and grouped them into 5 clusters using hierarchical clustering with Pearson correlation metric and Ward linkage. The Gene Ontology enrichment analysis shows that Cluster 5 is enriched for TF binding. Therefore we chose this cluster for our modeling.

We queried the genes of Cluster 5 in cREMAG for TFBS prediction as described in Method section. From the result we selected 25 TF coding and 110 non-TF coding genes to control the network size. To better reconstruct the TRN, we used the B-spline technique [21] to interpolate the original dataset and expanded the expression time course into 25, 32 and 48 time points.

The number of interactions with non-zero β^{TF} coefficients obtained from our model are listed in Table I. The sub-TRN representing the interactions between TF coding and target genes are presented in Figure 2. The result shows the need of increasing sampling time points by interpolation of the original data.

TABLE I

Performance Characteristics on the Experimental Data

Cluster	TFBS predicted (#edges)	#Time	Learned (#edges)	%learned
C5	1,131	25	143	12.51
		32	140	12.38
		48	154	13.61

TFBS prediction stands for the connected edges with non-zero association strength. Learned stands for the learned connected edges through regression. The percentage stands for the ratio between the number of learned connected edges and the connected edges given by association strength matrix.

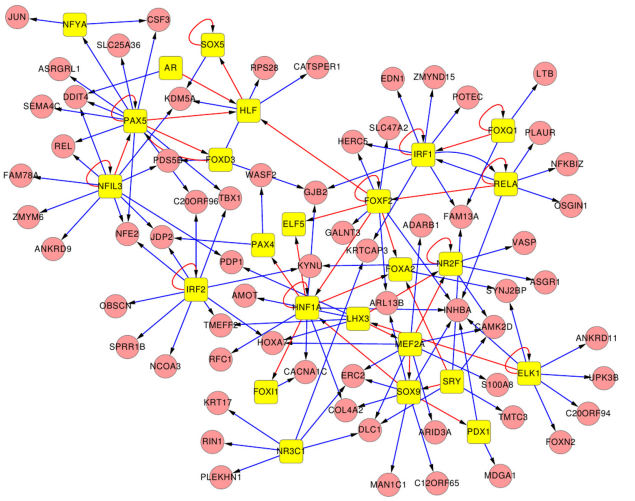

Fig. 2. **The sub-network for TF coding genes and their targets in Cluster 5**. The red pies are non-TF coding genes. The yellow squares are TF coding genes. The red arrows indicate the connection between TF coding genes. The blue arrows show the connection between TF coding genes and non-TF targets.

C. Verification of predicted TF-target interactions

Several learned interactions have been verified in previous studies. For example, the detected regulation targets

of RELA. In human, RELA (v-rel reticuloendotheliosis viral oncogene homolog A) gene encodes Nuclear factor NFκB $p65$ subunit, which is a TF expressed in growth plate chondrocytes where it facilitates chondrogenesis. $p50$ (encoded by NFκB1) binds to $p65$ (RELA) and $p50/p65$ heterodimer is the most abundant form of NFκB complex. NFκB is a generic name for an volitionally conserved TF system that contributes to the mounting of an effective immune response and is also involved in the regulation of cell proliferation, development, and apoptosis. IRF1 (Interferon regulatory factor 1) was the first identified member in interferon regulatory transcription factor (IRF) family. IRF1 regulates expression of a variety of target genes by binding to an interferon stimulated response element (ISRE) in their promoters. Previous studies have shown $p50/p65$ heterodimer and IRF1 physically interact with each other and cooperatively induce MHC class I gene expression [22]. Studies also revealed that increased IRF1 activation will suppress the NFκB $p65$ (RELA) activity and inhibits the expression of prosurvival (BCL2, BCL-W), and induces the expression of proapoptotic members (BAK, mitochondrial BAX) of the BCL2 family. This molecular signaling is associated with activation of STAT1, and leads to increased mitochondrial membrane permeability, activation of CASP7, CASP8, and CASP9, and induction of apoptosis in MCF-7 cells [23].

Another example is the self regulation of NFIL3 (nuclear factor, interleukin 3 regulated). NFIL3 binds to the promoter region of IL3 (interleukin-3) gene thus initiates its transcription. IL3, which encodes cytokine, is capable of supporting the proliferation of a broad range of hematopoietic cell types, and involves in a variety of cell activities such as cell growth, differentiation and apoptosis [24]. A recently study investigated basic region-leucine zipper (bZIP) transcription factors and quantified bZIP dimerization networks for five metazoan and two single-cell species, measuring interactions in vitro for 2,891 protein pairs. The interaction of NFIL3 on itself has been verified using fluorescence resonance energy transfer (FRET) [25]. Other studies have shown when BRAC1 gene is knocked out in MCF-7 cells, the expression of NFIL3 will be up-regulated [26].

These verified TF interactions in previous studies suggest that our method may be capable of identify "functional" transcription regulatory effects. However, a comprehensive evaluation of the proposed model is needed.

REFERENCES

[1] Latchman DS (1997) Transcription factors: An overview. International Journal of Biochemistry & Cell Biology 29: 1305-1312.

[2] Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, et al. (2005) Identification of transcription factor binding sites with variable-order bayesian networks. Bioinformatics 21: 2657-66.

[3] Barski A, Cuddapah S, Cui KR, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

[4] Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-dna interactions. Science 316: 1497-1502.

[5] Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao YJ, et al. (2007) Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods 4: 651-657.

[6] Stormo GD (2000) Dna binding sites: representation and discovery. Bioinformatics 16:16-23.

[7] Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics 5: 276-287.

[8] Bar-Joseph Z, Gitter A, Simon I (2012) Studying and modelling dynamic biological processes using time-series gene expression data. Nat Rev Genet 13: 552-564.

[9] Mathavan S, Lee SGP, Mak A, Miller LD, Murthy KRK, et al. (2005) Transcriptome analysis of zebra fish embryogenesis using microarrays. PLoS Genet 1: e29.

[10] Almasri E, Larsen P, Chen G, Dai Y (2008) Incorporating literature knowledge in bayesian network for inferring gene networks with gene expression data. In: Bioinformatics Research and Applications, Lecture Notes in Computer Science 4983. Springer Berlin Heidelberg, pp. 184-195.

[11] Zhu J, Chen Y, Leonardson AS, Wang K, Lamb JR, et al. (2010) Characterizing dynamic changes in the human blood transcriptional network. PLoS Comput Biol 6: e1000671.

[12] Liao JC, Boscolo R, Yang YL, Tran LM, Sabatti C, et al. (2003) Network component analysis: Reconstruction of regulatory signals in biological systems. Proceedings of the National Academy of Sciences 100: 15522-15527.

[13] Bansal M, Gatta GD, di Bernardo D (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics 22: 815-822.

[14] Seok J, Xiao W, Moldawer L, Davis R, Covert M (2009) A dynamic network of transcription in lps-treated human subjects. BMC Systems Biology 3: 78.

[15] Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B-Methodological 58: 267-288.

[16] Piechota M, Korostynski M, Przewlocki R (2010) Identi
cation of cis-regulatory elements in the mammalian genome: The cREMaG database. Plos One 5.

[17] Wingender E, Dietze P, Karas H, Knuppel R (1996) Transfac: A database on transcription factors and their dna binding sites. Nucleic Acids Research 24: 238-241.

[18] Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) Jaspar: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research 32: D91-D94.

[19] Schneider TD, Stormo GD, Gold L, Ehrenfeucht A (1986) Information-content of binding-sites on nucleotide-sequences. Journal of Molecular Biology 188: 415-431.

[20] Ariazi EA, Cunliffe HE, Lewis-Wambi JS, Slifker MJ et al. (2011) Estrogen induces apoptosis in estrogen deprivation-resistant breast cancer through stress responses as identified by global gene expression across time. Proceedings of the National Academy of Sciences 108(47):18879-86.

[21] Carl de Boor (1978). A Practical Guide to Splines. Springer-Verlag. ISBN 3-540-90356-9.

[22] Drew PD, Franzoso G, Becker KG, Bours V, Carlson LM, et al. (1995) Nf kappa b and interferon regulatory factor 1 physically interact and synergistically induce major histocompatibility class i gene expression. Journal of Interferon and Cytokine Research 15: 1037-1045.

[23] Ning Y, Riggins B, Mulla J, Chung H, Zwart A, Clarke R, et al. (2011) Interferon Gamma Restores Breast Cancer Sensitivity to Fulvestrant by Regulating STAT1, IRF1, NFB, BCL2 Family Members, and Signaling to Caspase-dependent Apoptosis. Mol Cancer Ther. 9(5): 12741285

[24] Zhang W, Zhang J, Kornuc M, Kwan K, Frank R, et al. (1995) Molecular-cloning and characterization of nf-il3a, a transcriptional activator of the human interleukin-3 promoter. Molecular and Cellular Biology 15: 6055-6063.

[25] Reinke AW, Baek J, Ashenberg O, Keating AE (2013) Networks of bzip protein-protein interactions diversified over a billion years of evolution. Science 340: 730-734.

[26] Bae, I., Rih, J. K., Kim, H. J., Kang, H. J., Haddad, B., Kirilyuk, A., Rosen, E. M. (2005). Report BRCA1 Regulates Gene Expression for Orderly Mitotic Progression. Cell Cycle, 4(11), 1641-1666.