

Supplementary for:

## **A novel method of transcriptional response analysis to facilitate drug repositioning for cancer therapy**

**Guangxu Jin<sup>1,2</sup>, Changhe Fu<sup>1</sup>, Hong Zhao<sup>1,2</sup>, Kemi Cui<sup>1,2</sup>, Jenny Chang<sup>1,2,3</sup>,  
Stephen T.C. Wong<sup>1,2,3,\*</sup>**

<sup>1</sup>Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston; <sup>2</sup>NCI Center for Modeling Cancer Development, The Methodist Hospital Research Institute, Weill Medical College, Cornell University, Houston; and <sup>3</sup>Methodist Cancer Center, The Methodist Hospital, Houston, TX 77030, USA.

**\*Corresponding author:** Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Medical College, Cornell University, 6670 Bertner street, MC: R6-414, Houston, TX, 77030, USA. TEL.: +1 713 441 5883; FAX: +1 713 441 8696; Email: [stwong@tmhs.org](mailto:stwong@tmhs.org)

## Data

The protein-protein interaction data used in the study were collected from five public databases, IntAct (1), DIP (2), MINT (3), MIPS (4), and BioGrid (5). The high-throughput physical interactions, identified by Y2H (Yeast Two-hybrid) and Affinity Capture-MS, were filtered out for our analysis. The data includes 28,487 protein-protein interactions and 9,032 proteins. The signaling pathway data used in the analysis are the canonical signaling pathways in NCI-PID and BioCarta (6, 7). NCI-PID contains 3,519 proteins, 8,472 interactions, and 254 signaling pathways. BioCarta includes 2,394 proteins, 3,065 interactions, and 207 signaling pathways. The genes related to genetic disorder of cancer in OMIM (8, 9) were automatically processed by a Python script and followed by manually checking. The number of cancer genes is 531, in which there are 47, 53, and 8 genes for breast cancer, prostate cancer, and promyelocytic leukemia. The drug target information was obtained from Drugbank (10), TTD (11), and PharmGKB (12). The combined data contain 9,951 drugs and 4,946 drug-targets. The drugs of Drugbank that belong to 'Antineoplastic Agents' category were identified as known anti-cancer drugs. About 20,000 FDA-approved application labels were downloaded from Drugs@FDA (<http://www.accessdata.fda.gov/scripts/cder/drugsatfda/>) and DailyMed (<http://dailymed.nlm.nih.gov/dailymed/>). The FDA-approved information was automatically processed by a Python script and followed by manually checking. We identified 18 FDA-approved breast cancer drugs, 6 FDA-approved prostate cancer drugs, and 2 FDA-approved promyelocytic leukemia drugs. The clinical trial information was queried from ClinicalTrial.gov website (<http://clinicaltrials.gov/>) followed manually checking. There are 90,257 clinical trials in the ClinicalTrial.gov. We found that 180, 221, and 11 drugs are undergoing different stages for breast cancer, prostate cancer, and promyelocytic leukemia. The dose-response data of raloxifene, tamoxifen, paclitaxel, and fulvestrant for MCF7 were derived from Developmental Therapeutics Program (DTP) of NCI/NIH (13). All of the filtered data for CSB-BFRM model can be found in the website of the tools, R2D2-CSB, <http://r2d2drug.org/Software/csb/csb.aspx>.

**Supplementary Table 1. Data sources for the definition of CSBs.**

| Data types                           | Data sources <sup>¶</sup>        |
|--------------------------------------|----------------------------------|
| Protein-protein interaction networks | IntAct, DIP, MINT, MIPS, BioGrid |
| Canonical signaling Pathways         | NCI-PID, BioCarta                |
| Cancer proteins*                     | OMIM                             |

<sup>¶</sup>NCI-PID: NCI PathwayInteractionDatabase; OMIM: Online Mendelian Inheritance in Man; \*the proteins whose coding genes are those in OMIM that have a close relationship with cancer genetic disorder.







## Detection of network motifs

In the course of evolution, the protein modules were recombined into new patterns by genetic mutations. Innovations in signaling processing are brought about by unique combinations of existing building blocks, known as network motifs, rather than by invention of entirely new protein modules (14-16).

The network motifs were detected from the protein-protein interaction data by the FANMOD software (17). The detected network motifs were shown in Supplementary Table 1. One could notice that the numbers of nodes in network motifs are limited within 4. That is mainly due to the huge size of the data, which causes the difficulty in detecting the network motifs with more than 4 nodes. The time spent on searching the four-node motifs in the data is 120,812 seconds (about 34 hrs), and the estimated time for more node motifs (5 or 6) in the data is at least 10,000,000 (more than 100 days). The machine used in the detection is an 8-core (Intel®, Xeon®, CPU X5355@2.66GHz) Cluster with 4.0 GB of RAM.

The network motifs used in our analyses are Triangle and Square. We chose them by three criteria: (1) basic element interaction patterns; (2) relatively low frequencies in the original network; (3) being highly clustered. 13278 and 31710 contains triangle or squares as their subgraphs. 4958 and 4382 have relatively high frequencies. So they are excluded from our analyses. Triangle and Square satisfy all of the three criteria, and their clustering properties are shown in next section.



**Supplementary Table 2. The network motifs detected from protein networks**

| Network motifs | ID    | Adjacent Matrix              | Subgraph  | Frequency [Original] | Mean-Freq [Random] | Standard-Dev [Random] | Z-score | P value |
|----------------|-------|------------------------------|---|----------------------|--------------------|-----------------------|---------|---------|
| Triangle       | 238   | 011<br>101<br>110            |    | 1.4589%              | 0.0032589%         | 0.00011764            | 123.73  | 0       |
| Square         | 27030 | 0110<br>1001<br>1001<br>0110 |   | 0.39532%             | 0.22632%           | 7.6415e-005           | 22.116  | 0       |
| -              | 4958  | 0001<br>0011<br>0101<br>1110 |  | 2.982%               | 0.011686%          | 0.00041668            | 71.287  | 0       |
| -              | 13278 | 0011<br>0011<br>1101<br>1110 |  | 0.2052%              | 8.3246e-005%       | 3.6451e-006           | 562.7   | 0       |
| -              | 31710 | 0111<br>1011<br>1101<br>1110 |  | 0.020203%            | 1.0043e-007%       | 1.3836e-008           | 14602   | 0       |
| -              | 4382  | 0001<br>0001<br>0001<br>1110 |  | 60.119%              | 57.33%             | 0.0020323             | 13.723  | 0       |

### ***The clustering property of the identified network motifs***

We checked the clustering property of the identified triangles and squares. We examined the clustering property by the parameters of protein-protein interaction networks. Type I subgraph was proposed to identify the local structures of subgraphs by the global attributes of whole protein-protein interaction network (18). If we denote  $N_{nm}(k)$  is the average number of  $(n, m)$  subgraphs with  $n$  nodes and  $m$  interactions that pass by a node with degree  $k$ , then  $N_{nm}(k) \sim k^{n-\gamma-(m-n+1)\alpha}$ , where  $\alpha$  and  $\gamma$  are two exponents derived from the power law distributions of correlation coefficients and degrees, i.e.  $P(k) \sim k^{-\gamma}$  and  $C(k) \sim C_0 k^{-\alpha}$ . If the subgraph exponent satisfies  $n - \gamma - (m - n + 1)\alpha > 0$ , the  $(n, m)$  subgraphs were called

as Type I subgraph. And we found that the triangles and squares in our analyses belong to Type I subgraph, see following table.

| Bridge   | Subgraph  | Subgraph exponent |          |                                  |
|----------|---|-------------------|----------|----------------------------------|
|          |   | $\gamma$          | $\alpha$ | $n - \gamma - (m - n + 1)\alpha$ |
| Triangle |  | 1.836             | 0.7367   | 0.4273                           |
| Square   |  | 1.836             | 0.7367   | 1.4273                           |

This property of clustering facilitates the identified CSBs to better connect signaling pathways with cancer-related genes or proteins.

### Enrichment analysis on the CSBs

We randomly sample one protein set  $S_1$  from the proteins in protein-protein interaction network, which satisfies that  $|S_1| = |S|$  and  $|S_1 \cap C| = |S \cap C|$ , and identify another instance subset  $\Pi^{S_1, C}$  of  $\Pi$  and each  $CSB_j$  ( $j = 1, 2, \dots, |\Pi^{S_1, C}|$ ) satisfies the criteria in (1) that  $|CSB_j \cap S_1| > 0$ ,  $|CSB_j \cap C| > 0$ , and  $|CSB_j| > |CSB_j \cap (S_1 \cap C)|$ .

Repeating the sample experiment for 10, 000 times, we can get a list of random numbers of CSBs in the identified instance sets,

$$L = \left\{ |\Pi^{S_1, C}|, |\Pi^{S_2, C}|, \dots, |\Pi^{S_{10,000}, C}| \right\} \quad (2)$$

Thus we can derive a random distribution  $f$  from (2). The corresponding P-value to evaluate the enrichment of CSBs in the connection between  $S$  and  $C$  is computed at  $|\Pi^{S, C}|$  by the ccdf (Complementary Cumulative Distribution Function) of  $f$ .

### E score

In Cmap 02, the connectivity score was proposed to measure the expression difference between sample genes (up- or down- regulated genes) and reference genes (all genes) based on a non-parameter test, i.e. Kolmogorov-Smirnov statistic (K-S test) (19, 20). By evaluating the enrichment of a gene set in a treatment instance, the connectivity score can reflect the treatment effect of the drug in the instance. Similarly, we apply K-S test to define an E score to evaluate the effects of drugs on an interested genes or proteins set in a treatment instance.

### Data preprocessing

The microarray data produced by drug treatment experiments were preprocessed and normalized by Partek software (21). In normalization configuration, we used the 'Partek Defaults'. Raw probe intensities were adjusted based on the number of G and C bases in the probe sequence, the background was corrected by using method

from RMA (22), ‘No log’ was used for probes, and quantile normalization and mean probeset summarization were adopted.

### **Fold change**

To evaluate the expression change of a gene in a treatment instance, we considered the fold-change of the gene. For each probe of the gene, the arithmetic mean of the values from the six individual control scans was first derived. Then the fold change for this probe is defined as the ratio of the corresponding treatment-to-control values.

### **Definition of E score**

E score is defined by the modified Kolmogorov-Smirnov test (K-S test) that is to test whether two underlying one-dimensional probability distributions differ. For each treatment instance  $i$ , the probe sets for the interested genes or proteins set and all genes in the microarray chip are denoted by  $\mathbf{A}$  and  $\mathbf{B}$  respectively. The probes in  $\mathbf{A}$  and  $\mathbf{B}$  were all sorted in ascending order according to their Fold-change. For a probe  $j$  in  $\mathbf{B}$  ( $j=1,2,\dots,b$ ), we denoted its position in  $\mathbf{A}$  as  $\mathbf{A}(j)$  (the positions for  $\mathbf{A}$  are  $1,2,\dots,n$ ), and then computed the following four values for up-regulated and down-regulated genes in the interested genes or proteins:

$$E_1^i = \max_{j \in \mathbf{U}} \left( \frac{j}{b} - \frac{\mathbf{A}(j)}{n} \right) \quad (3)$$

$$E_2^i = \max_{j \in \mathbf{U}} \left( \frac{\mathbf{A}(j)}{n} - \frac{j-1}{b} \right) \quad (4)$$

$$E_3^i = \max_{j \in \mathbf{D}} \left( \frac{j}{b} - \frac{\mathbf{A}(j)}{n} \right) \quad (5)$$

$$E_4^i = \max_{j \in \mathbf{D}} \left( \frac{\mathbf{A}(j)}{n} - \frac{j-1}{b} \right) \quad (6)$$

where  $\mathbf{U}$  and  $\mathbf{D}$  are the upper and lower quartile of the probes in  $\mathbf{B}$  respectively. (3-6) can determine whether the fold-change distributions of the interested gene set (up- and down- regulated) differ from that of all genes by checking the difference between the cumulative fraction functions for these two types of distributions.

The  $\mathbf{E}$  scores for the up- and down- regulated genes of the interested gene set in the instance  $i$ ,  $E_{\mathbf{U}}^i$  and  $E_{\mathbf{D}}^i$ , were set as follows,

$$E_{\mathbf{U}}^i = \begin{cases} E_1^i & \text{if } E_1^i > E_2^i \\ -E_2^i & \text{if } E_2^i > E_1^i \end{cases}$$

$$E_{\mathbf{D}}^i = \begin{cases} E_3^i & \text{if } E_3^i > E_4^i \\ -E_4^i & \text{if } E_4^i > E_3^i \end{cases}$$

The  $\mathbf{E}$  score for the instance  $i$ , i.e.  $E^i$ , was defined as  $E_{\mathbf{U}}^i - E_{\mathbf{D}}^i$ , if  $E_{\mathbf{U}}^i > 0$  and  $E_{\mathbf{D}}^i < 0$ , and  $E_{\mathbf{U}}^i$ , if  $E_{\mathbf{U}}^i > 0$ ,  $E_{\mathbf{D}}^i > 0$ ;  $-E_{\mathbf{D}}^i$ , if  $E_{\mathbf{U}}^i < 0$ ,  $E_{\mathbf{D}}^i < 0$ .

Multiple treatment instances were designed for drug  $k$  due to the different cell lines, treated concentration and chips in Cmap 02. We denoted the instance set for the drug as  $\mathbf{I}_k$ . The median value for  $\{E^1, E^2, \dots, E^{I_k}\}$  was set to the  $\mathbf{E}$  score for the

drug, i.e.  $E_k$ . Set  $M = \max_k(E_k)$  and  $m = \min_k(E_k)$  across all drugs, then **E** score  $E_k$  was defined as  $E_k / M$  where  $E_k \geq 0$ , or  $-E_k / m$  where  $E_k < 0$ .

### Implementation of CSB-BFRM

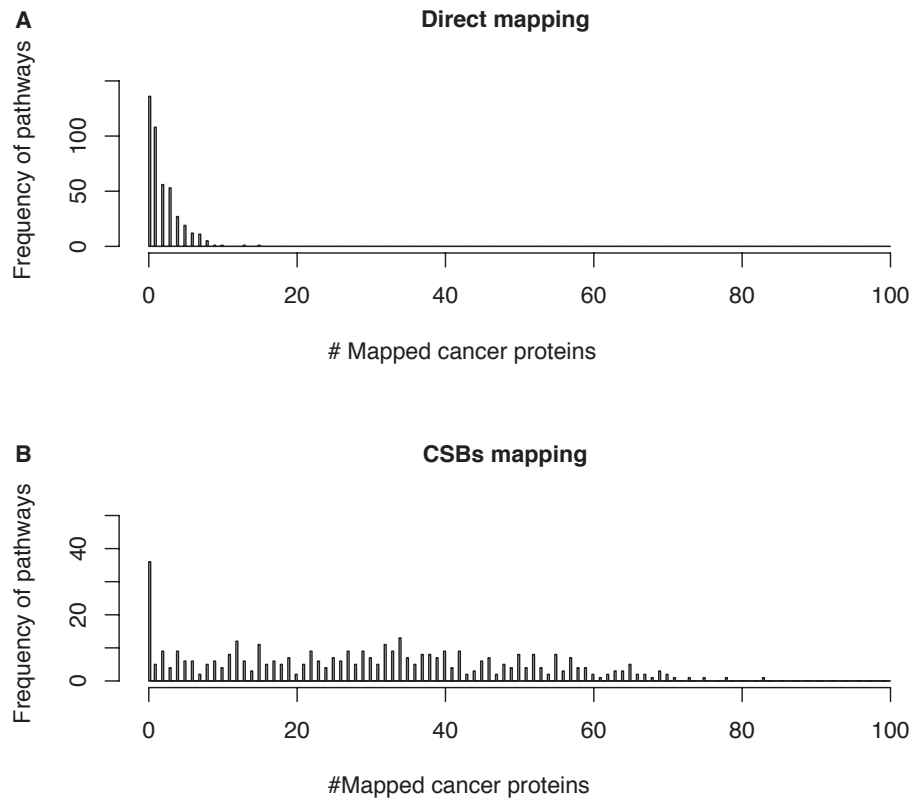
The cancer proteins for breast cancer, prostate cancer, and promyelocytic leukemia were manually identified from the OMIM database (17, 24) (Supplementary Table 4-6), and the expanded CSB proteins for the three cancer types are listed in Supplementary Table 7-9. The inputs of CSB-BFRM, i.e., treatment response matrices ( $\mathbf{X}$ ), for MCF7, PC3, and HL60 have 1,390, 1,215, and 1,099 columns (drugs) as shown in Supplementary Table 10-12. The number of signatures,  $k$ , was identified by the evolution algorithm in BFRM automatically. The numbers of the signatures for the three cancer types equal to 50, 46, and 40 respectively.

In the identification of targetable signatures, the target information is indispensable. The targets of some drugs may not be included in the expanded CSB proteins of one specific cancer type. Our strategy is expanding them to the nearest CSB proteins using the shortest protein-protein interaction paths (in other words, using the smallest number of proteins linked head to tail). We used the expanded proteins to address the targetable signatures from the weight matrix  $\mathbf{A}$ . Still some other drugs do not have any known drug targets. For each of these drugs, our strategy is taking a randomized number of CSB proteins as its targets or off-targets. We repeated the randomized process for a thousand times to reduce computational bias.

Applying CSB-BFRM on the BFRM outputs, weight matrix  $\mathbf{A}$  and score matrix  $\Lambda$ , we identified one thousand repositioning profiles for the repositioned drugs after repeating the randomized process for 1,000 times for each cancer type. To further define the repositioning score, we applied a supervised regression model, Support Vector Regression (SVR), on the repositioning profiles. We used FDA-approval and publicly available clinical trial information as the prior knowledge and performed the regression between the identified repositioning profiles and the prior knowledge indicating which drugs are FDA-approved or under clinical trials. The information for FDA approved drugs and clinical trial drugs for breast cancer, prostate cancer, and promyelocytic leukemia, are shown in Supplementary Table 13.

**Supplementary Tables 3-18 are included in Supplementary Data.**

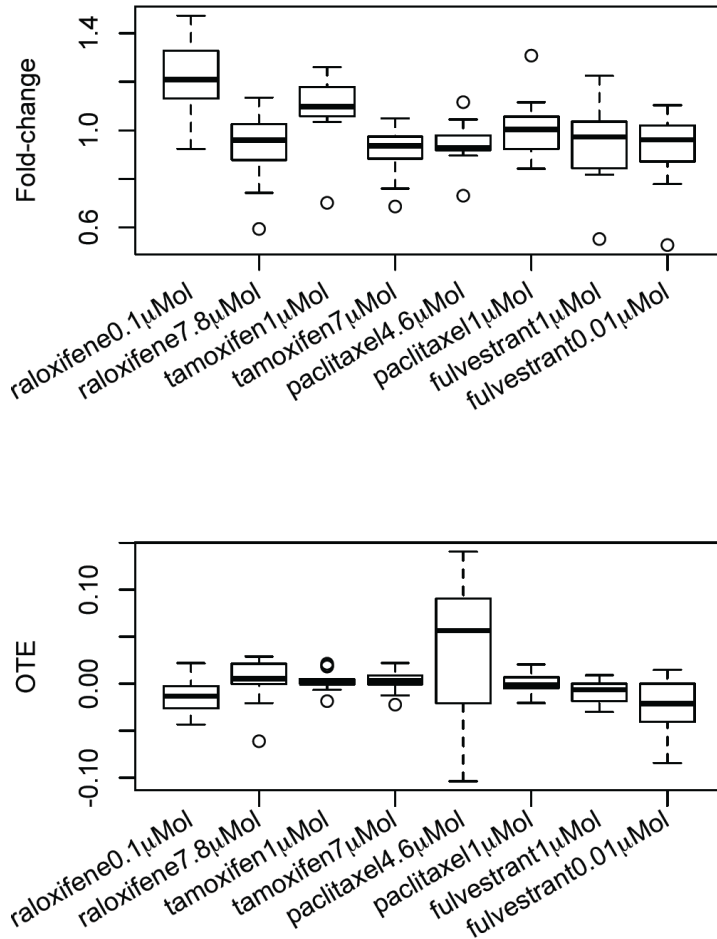
## Supplementary Figure 1.



**(A)** Without facilitated by CSBs, most cancer proteins are isolated from the signaling pathways. The numbers of cancer proteins mapped to signaling pathways are less than are less than 10. More than half of signaling pathways don't involve in any cancer proteins. **(B)** Facilitated by CSBs, significantly more cancer proteins are involved in the signaling pathways mapping ( $P < 10^{-10}$ , Mann-Whitney *U* test). Most signaling pathways are expanded to cancer proteins (only less than 40 signaling pathways cannot be linked with cancer proteins).



## Supplementary Figure 2.



### Comparison between original fold-changes and the recognized OTEs

Every boxplot describes the fold-changes or OTEs of the drug's off-targets in the cell cycle G1/S checkpoint and P53 signaling pathways. The OTEs are recognized by the BFRM method, and they are the factorized values for targetable signatures. The data for original fold-changes are in the range of [0.4, 1.6] while those for OTEs are between -0.01 and 0.01. The OTEs are better to characterize the drug effects. Nearly all of the original fold-changes for raloxifene ( $0.1\mu M$ ) and tamoxifen ( $1\mu M$ ) are higher than 1. The response characterized by the fold-change cannot indicate the difference between the positive and negative effects on the molecules in cell cycle G1/S checkpoints and P53 signaling pathways.

1. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct--open source resource for molecular interaction data. *Nucleic Acids Res.* 2007;35:D561-5.
2. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 2002;30:303-5.
3. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, et al. MINT: the Molecular INTERaction database. *Nucleic Acids Res.* 2007;35:D572-4.
4. Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 2002;30:31-4.
5. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, et al. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* 2008;36:D637-40.
6. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.* 2009;37:D674-9.
7. Izmailov A, Yager TD, Zaleski H, Darash S. Improvement of base-calling in multilane automated DNA sequencing by use of electrophoretic calibration standards, data linearization, and trace alignment. *Electrophoresis.* 2001;22:1906-14.
8. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 2005;33:D514-7.
9. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. *Nat Biotechnol.* 2007;25:1119-26.
10. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 2008;36:D901-6.
11. Chen X, Ji ZL, Chen YZ. TTD: Therapeutic Target Database. *Nucleic Acids Res.* 2002;30:412-5.
12. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenetics and pharmacogenomics knowledge base. *Methods Mol Biol.* 2005;311:179-91.
13. Doh H, Roh S, Lee KW, Kim K. Response of primed human PBMC to synthetic peptides derived from hepatitis B virus envelope proteins: a search for promiscuous epitopes. *FEMS Immunol Med Microbiol.* 2003;35:77-85.
14. Jin G, Zhang S, Zhang XS, Chen L. Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. *PLoS One.* 2007;2:e1207.
15. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science.* 2002;298:824-7.
16. Shen-Orr SS, Milo R, Mangan S, Alon U. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet.* 2002;31:64-8.
17. Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics.* 2006;22:1152-3.

18. Vazquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabasi AL. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc Natl Acad Sci U S A*. 2004;101:17940-5.
19. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313:1929-35.
20. Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, et al. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell*. 2003;114:323-34.
21. Downey T. Analysis of a multifactor microarray study using Partek genomics solution. *Methods Enzymol*. 2006;411:256-70.
22. Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*. 2003;19:185-93.