A NOVEL META-ANALYSIS APPROACH OF CANCER TRANSCRIPTOMES REVEALS PREVAILING TRANSCRIPTIONAL NETWORKS IN CANCER CELLS

ATSUSHI NIIDA aniida@ims.u-tokyo.ac.jp
SEIYA IMOTO imoto@ims.u-tokyo.ac.jp
MASAO NAGASAKI masao@ims.u-tokyo.ac.jp
RUI YAMAGUCHI ruiy@ims.u-tokyo.ac.jp
SATORU MIYANO miyano@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Although microarray technology has revealed transcriptomic diversities underlining various cancer phenotypes, transcriptional programs controlling them have not been well elucidated. To decode transcriptional programs governing cancer transcriptomes, we have recently developed a computational method termed EEM, which searches for expression modules from prescribed gene sets defined by prior biological knowledge like TF binding motifs. In this paper, we extend our EEM approach to predict cancer transcriptional networks. Starting from functional TF binding motifs and expression modules identified by EEM, we predict cancer transcriptional networks containing regulatory TFs, associated GO terms, and interactions between TF binding motifs. To systematically analyze transcriptional programs in broad types of cancer, we applied our EEM-based network prediction method to 122 microarray datasets collected from public databases. The data sets contain about 15000 experiments for tumor samples of various tissue origins including breast, colon, lung etc. This EEM based meta-analysis successfully revealed a prevailing cancer transcriptional network which functions in a large fraction of cancer transcriptomes; they include cell-cycle and immune related sub-networks. This study demonstrates broad applicability of EEM, and opens a way to comprehensive understanding of transcriptional networks in cancer cells.

Keywords: microarray data; cancer; expression module; transcriptional network.

1. Introduction

In the last decade, microarray technology has revealed transcriptomic diversities underlining various cancer phenotypes; on the other hand, we have yet little knowledge about transcriptional programs controlling them. In our previous study, we proposed a computational method termed EEM, which aims to identify expression modules. An expression module is defined as co-expressed genes under a common regulatory program (e.g., target genes of the same TF), and could be a key to understanding the regulatory program. EEM starts from prescribed gene sets defined by prior biological knowledge like TF binding motifs [10]; for each gene set, EEM first identifies the largest co-expressed subset of genes in an input microarray

dataset. Using the size of the subset as a test statistic, EEM then obtains significant co-expressed subsets of genes as expression modules.

Recently, the amount of transcriptome data deposited in public databases is exploding. For example, GEO [15], which is one of the principle repositories for microarray data, returned 3024 data sets when queried for "cancer" (as of Aug. 2009). Although even only separate analysis of each dataset has yielded many significant findings, it is expected that meta-analysis of the massive datasets will provide more global insights into regulatory programs encoded in cancer transcriptomes.

In this study, we present a new version of the EEM method which enables efficient screening for expression modules, aiming at meta-analysis of a large number of cancer microarray data sets. Although the original version of EEM employs z-score calculation to evaluate coherence of each gene set, we newly introduce an efficient p-value calculation method based on the extreme value distribution. Moreover, we extend our EEM approach to systematic prediction of cancer transcriptional networks. To reveal prevailing cancer transcriptional networks, we applied this new version of EEM to 122 microarray datasets including about 15000 experiments for tumor samples of various tissue origins. This analysis successfully revealed a prevailing cancer transcriptional network which functions in a large fraction of cancer transcriptomes. Taken together, this study provides a foundation for network-level meta-analysis based on our EEM method.

2. Methods

2.1. *Overview of the EEM Algorithm*

First we briefly review the EEM algorithm, which is described in detail in Niida *et al.* [10]. EEM takes two types of input data: a microarray dataset and a gene set (we call a *seed gene set*). Let $E = \{e_1, \ldots, e_n\}$ be the set of gene expression profiles in the input microarray dataset, where $e_i = (e_{i1}, \ldots, e_{im})$ is a vector of the expression values of the i-th gene, i.e., e_{ij} is the expression value of the i-th gene in the j-th sample. We then assume that e_i exists as a point in a continuous m-dimensional gene expression space $\mathcal{S}$. EEM operates on a subset $E_M \subseteq E$, which is the expression profiles of the genes in the seed gene set. For a given radius parameter r and point $x \in \mathcal{S}$, define

$$C_x = \{e_i \in E_M : d(e_i, x) \leq r\}, \tag{1}$$

where d is the Euclidean distance. We call C_x a *coherent set*, and the point x the *center* of C_x. First, EEM finds the maximal sized coherent set C_B (and corresponding center B) for the genes in E_M. Next, by assuming the size of the maximal sized coherent set $|C_B|$ as a test statistics (we call the *EEM statistics*), EEM evaluates expression coherence of E_M. If it was judged to be significant, EEM obtains C_B as an expression module.

Using EEM, we can screen for expression modules in the transcriptome of the

input microarray dataset. Prior to the screening, we prepare a set of gene sets that could contain expression module, based on prior biological knowledge like TF binding motifs. EEM then applies the above procedure to each of the gene sets, and finally obtains expression modules from gene sets of significant coherence.

2.2. *Calculation of P-Values*

In the original version, EEM calculated a Z score of an EEM statistic $|C_B|$ from an empirical null distribution of the statistics. The empirical null distribution can be generated by repeatedly calculating test statistics for randomly sampled gene sets whose size is equal to that of the input gene set. Based on the empirical null distribution, we can also calculate p-values; the empirical approach enables us to calculate p-values for any statistics without theoretical models for null distributions. However, if it relies only on the empirical null distribution, precise calculations of small p-values need prohibitive computational time for generation a large number of null statistics. Especially, speed of computation is essential in meta-analysis which deals with a large amount of transcriptome data.

To overcome this limitation, we propose a novel efficient p-value calculation method based on based on an extreme value distribution fitted to the empirical null distribution. Prior to the p-value calculation based on the extreme value distribution, we roughly calculate p-values to filter out non-significant genes. EEM calculates a p-value from an empirical null distribution from 100 randomly sampled genes sets. The seed gene sets with p-values > 0.05 are filtered out in this step. This step works well enough to filter out non-significant gene set before proceeding to the following computer intensive steps which calculate a larger number of null statistics.

Extreme value distributions are used to model the minimum or the maximum of the collection of random observations from the same arbitrary distribution [4]. Because an EEM test statistic is the size of the largest coherent subset among many possible candidates, it is reasonable to model the null distribution based on an extreme value distribution. A family of extreme value distributions is represented as the generalized extreme value distribution, which has the following cumulative distribution function:

$$\mathrm{F}(x;\mu,\sigma.\xi) = \exp\left[-\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{-1/\xi}\right]$$

for $1+\xi(x-\mu)/\sigma > 0$, where $\mu \in \mathbf{R}, \sigma > 0$, and $\xi \in \mathbf{R}$ are the location, scale, and shape parameters, respectively.

We can fit the above equation to the empirical distribution, and use it to calculate p-values ($p = 1 - \mathrm{F}(|C_B|)$). In practice, the empirical distribution is generated from 10^3 randomly sampled genes sets and the fitting was performed using R library *ismev*. As shown in Fig.1, we confirmed that p-values based on the fitted extreme value distribution correspond well to those based on the empirical

distribution. Note that we can calculate p-value $< 10^{-3}$ by the fitted distribution, though it only requires the same computational time for sampling 10^3 null statistics. Combined with the first filtering step, the computational time is further boosted; for a microarray dataset of typical size in this study, EEM takes a few hours using a workstation with Intel Core i7 and 4G RAM.

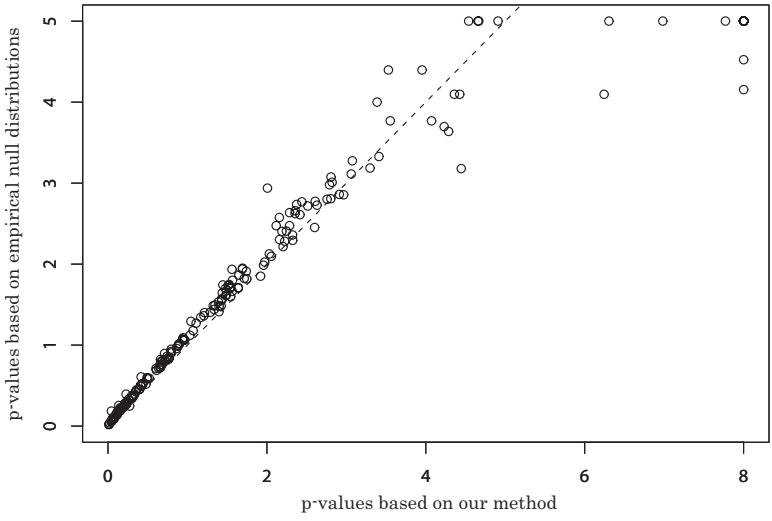

Fig. 1: Comparison of p-values based on fitted extreme distributions and empirical null distributions. we plot minus log p-values calculated by our method and 100000 random samplings of null EEM statistics. Although this result was obtained from the microarray dataset GSE3494 and TF binding motif-based gene sets, similar results were also obtained from different inputs.

2.3. *Input Data*

2.3.1. *Microarray Datasets*

In this study, we applied EEM to a large number of cancer microarray data sets. To obtain the cancer microarray data sets, we first collected the datasets from GEO [1] and ArrayExpress [11] using "cancer" as the keyword. The datasets were then excluded if the platform is not the the Affymetrix U133 series and the number of samples is less than 50. We finally obtained 122 datasets and applied the following preprocessing steps to them. Assuming the 5% lowest expression value as a floor, we rounded up expression values below the floor, and then converted expression values to the logarithmic scale. The Probe set IDs were converted to gene symbols. In cases that one gene symbol matches multiple probe set IDs, the probe set which shows the largest variance across the samples was mapped to the gene. A variation

filter was then applied to the data, and we obtained the top half of genes with the highest variance. The expression profiles of them were normalized across samples and subjected to EEM analysis.

2.3.2. *Seed Gene Sets*

As the other type of input data, EEM needs seed gene sets that could contain coherent subset of genes. In this study, we prepared seed gene sets based on common TF binding motifs in their promoters [10]. Briefly, we prepared human and mouse promoter sequences encompassing the 500 bp upstream and 100 bp downstream of the transcription start sites. We also obtained 200 PWMs After motif clustering was applied to all vertebrate PWMs in TRANSFAC 2009.1 [8] to remove redundant motifs. For each PWM, we scored every human and mouse promoter sequence based on maximum log odds scores, and obtained the average of human and mouse homolog promoter scores as the PWM score for each gene. We assumed genes which record the 5 % highest PWM scores as seed gene sets sharing common TF binding motifs associated with the PWM.

2.4. *Transcriptional Network Prediction*

When given a microarray data set and seed gene sets based on TF binding motifs, EEM identifies not only expression modules, but also functional TF binding motifs regulating them. Starting from these EEM results, we can predict a transcriptional network in the transcriptome of the input microarray dataset. The predicted transcriptional network consists of three types of nodes: TF binding motifs (*motif nodes*), TFs (*TF nodes*), and GO terms (*GO nodes*). We assume that there are four types of edges between them.

(1) *MMI edges*: Undirected edges between motif nodes which represent motif-motif interactions
(2) *TF-binding edges*: Directed edges from TF nodes to motif nodes which represent TF biding to motifs
(3) *cis-regulatory edges*: Directed edges from motif nodes to TF or GO nodes which represents *cis*-transcriptional regulation
(4) *PPI edges*: Undirected edges between TF nodes which represent protein-protein interactions

Next, we explain how to derive these four types of edges from EEM results and various other data sources.

2.4.1. *MMI Edges*

MMI edges were predicted from overlaps between motif-associated expression modules. If two different motifs share a significantly large number of target genes, they should interact functionally, possibly, via physical interactions of binding TFs.

We tested overlaps of expression modules based on the hypergeometric distribution, and assumed that edges exists if p-values are smaller than 10^{-10}. Note that the TRANSFAC database includes redundant PWMs which can recognize the same binding motif. In such cases, expression modules derived from redundant PWMs can also have significant overlap with each other. Although we reduced the redundancy by clustering in the preparation step of PWMs, we found that the input PWM still includes some "cognate PWMs" which are annotated as bound to the same TF. Therefore, apparently significant MMIs might be "pseudo-MMIs" between a pair of redundant PWMs. To discriminate real MMIs from pseudo-MMIs, we measured similarity between each pair of motifs, and we assumed that the pair constitutes real MMI if the KL-distance exceeds 15 [10]. This cutoff value is based on the observation that it can filter out most of cognate PWM pairs.

2.4.2. *TF Binding Edges*

TF binding edges are predicted based on two information sources: TRANSFAC and expression profiles. Based on binding TF information attached to PWM entries in TRANSFAC, we first obtain binding TFs of each motif. Because the TRANSFAC database is based only on the primary literature and unsatisfactory from the viewpoint of comprehensiveness, we expanded TF binding information for each PWM as follows:

(1) If an associated binding TF is not human genes, we converted it to human homologs based on HomoloGene data.
(2) Because it is known that most members of the same TF family recognize similar binding motifs, we assumed that the other members which belong to the same family with a binding TF are also associated with the same motif. Gene family information was prepared from Ensembl.
(3) Note that input PWMs are representatives of clusters based on motif similarity. We also assigned TF associated with cluster members to the representatives.

To find binding TFs which functions in the transcriptome of the input microarray dataset, we measured correlation between the expression profile of each binding TF and the mean expression profile of module genes regulated by the motif, using sample label permutation tests; we assumed that a TF binding edges exists between the motif and a TF, if a p-value is less than 10^{-8}. Note that there are two types of TF binding edges; positive correlations are assumed as activating TF binding edges while negative correlations are assumed as repressive TF binding edges.

2.4.3. *Cis-Regulatory Edges*

Cis-regulatory edges can be obtained from results from EEM analysis. If an expression module associated with a motif includes a TF, the motif was assumed to regulate expression of the TF. GO terms enriched motif-associated expression modules were connected to motifs (p-value $< 10^{-6}$).

2.4.4. *PPI Edges*

PPI edges are based on PPI data registered by public databases. The data were prepared as described in Tamada *et al.* [14]

2.4.5. *Construction of Transcriptional Networks*

Starting from motif nodes with p-values $< 10^{-6}$ in EEM analysis, we construct a transcriptional network for an input microarray dataset. These "seed" motif nodes is connected to each other by MMI edges, and to TF nodes by TF-binding edge. We depicted *cis*-regulatory edges between the seed motifs and the added TF nodes, and PPI edges between the added TFs. GO nodes are added to the seed motifs via *cis*-regulatory edges. The network predicted in this way can be assumed as a transcriptional network which functions in the cancer transcriptome of the input microarray dataset.

Moreover, to reveal a functional transcriptional networks across multiple microarray datasets, we can obtain a "meta-network" by superimposing networks predicted for each microarray dataset. We counted the number of nodes and edges across networks for all the microarray datasets, and constructed a network by assembling nodes and edges which are predicted more frequently than a prespecified cutoff (3 out of 122 in this study). Network visualizations are performed so that the node size and the edge width proportionally reflect frequency of their prediction.

3. Results and Discussion

EEM analysis identifies expression modules that could be keys to understanding transcriptional programs in the transcriptome of the input microarray dataset. We predicted transcriptional networks, starting from expression modules identified by EEM and TF binding motif associated with them; we constructs a transcriptional network by connecting motifs based on module overlap, and adding regulatory TFs to the motifs based on the TRANSFAC database and expression correlations. PPI and GO terms enriched in each module were also incorporated to predicted networks. Furthermore, we attempted network-level meta-analysis to reveal prevailing cancer transcriptional networks, which possibly regulate fundamental oncogenic processes commonly employed by various types of cancers. We downloaded 122 microarray datasets associated the "cancer" keyword from GEO and ArrayExpress, and predicted transcriptional networks for each dataset. We then depicted a "meta-network" by assembling nodes and edges which were repeatedly predicted across the microarray datasets.

Fig. 2 shows the meta-network, where node size and edge width reflect frequency of their prediction. The meta-network contains two major sub-networks. One sub-network (lower right) includes E2F and DP family TFs, NFYA, and their binding motifs. It has been known, and also confirmed by our GO enrichment analysis, that the E2F and DP families are master regulators of cell-cycle [9]; it

is very reasonable that they drive one of the principal transcriptional programs in cancer transcriptomes. This sub-network harbors positive auto-regulatory loops mediated by several E2F family genes, suggesting that they act as an engine in cell-cycle regulation. The other sub-network (upper left) contains various TF and motif nodes in addition to two main nodes, IRF and PU.1 binding motifs; they apparently act as hubs in this network and might be drug targets in cancer therapy. This sub-network includes known cancer genes like IRF, ETS, and RUNX [3, 6, 13] and is linked to immune-system by GO analysis. Although this result suggests that this immune transcriptional program is responsible for oncogenesis in broad types of cancer cells, it is also possible that a part of this network is due to infiltrating immune cells in the tumor microenvironment.

Our EEM-based network prediction takes as input various types of fragmented knowledge deposited in the databases (e.g. TF binding motifs, binding TF information associated with motifs, GO, PPI, etc.) and systematically assembles them into networks which presumably functions in the transcriptome of an input microarray dataset. Although many other types of network prediction methods have been proposed for transcriptome analysis, our approach possesses outstanding features as compared to the others. We can predict direct casual relationships based on TF binding information, which cannot be predicted by co-expression or Bayesian network approaches which rely on only transcriptome data [2, 5, 7]. Our predicted network also includes motif-motif interaction, PPI, GO terms associated with motifs; they make the predicted networks information-rich, highly interpretable and easy to extract biological knowledge. In this study, we applied EEM-based network prediction to more than a hundred cancer microarray datasets and, by superimposing networks predicted for each dataset, constructed a "meta-network". This meta-network approach should increase reliability of predicted networks; in fact, the obtained meta-network includes many TF associated with cancer, and are highly consistent with the accumulated knowledge of molecular biology. Although a number of studies have performed meta-analysis of cancer transcriptomes [12, 16], they have mainly focused on individual genes or a set of genes associated with oncogenic phenotypes. In contrast, our meta-analysis focused for the first time on cancer transcriptional networks, and successfully revealed prevailing transcriptional network in cancer cells.

Our analysis successfully identified two major transcriptional programs employed by various types of cancer cells: cell-cycle and immune-related transcriptional programs. While our previous study also demonstrated that the breast cancer transcriptome harbors similar cell-cycle and immune-related transcriptional programs, this study expanded the finding to many other types of cancer, suggesting their generality in various types of cancer. Based on our approach developed in this study, our next challenge should be to find transcriptional programs specific to each types of cancer. To address this problem, we need an enough number of datasets for each type of cancer; we are currently collecting microarray datasets more comprehensively for future studies.

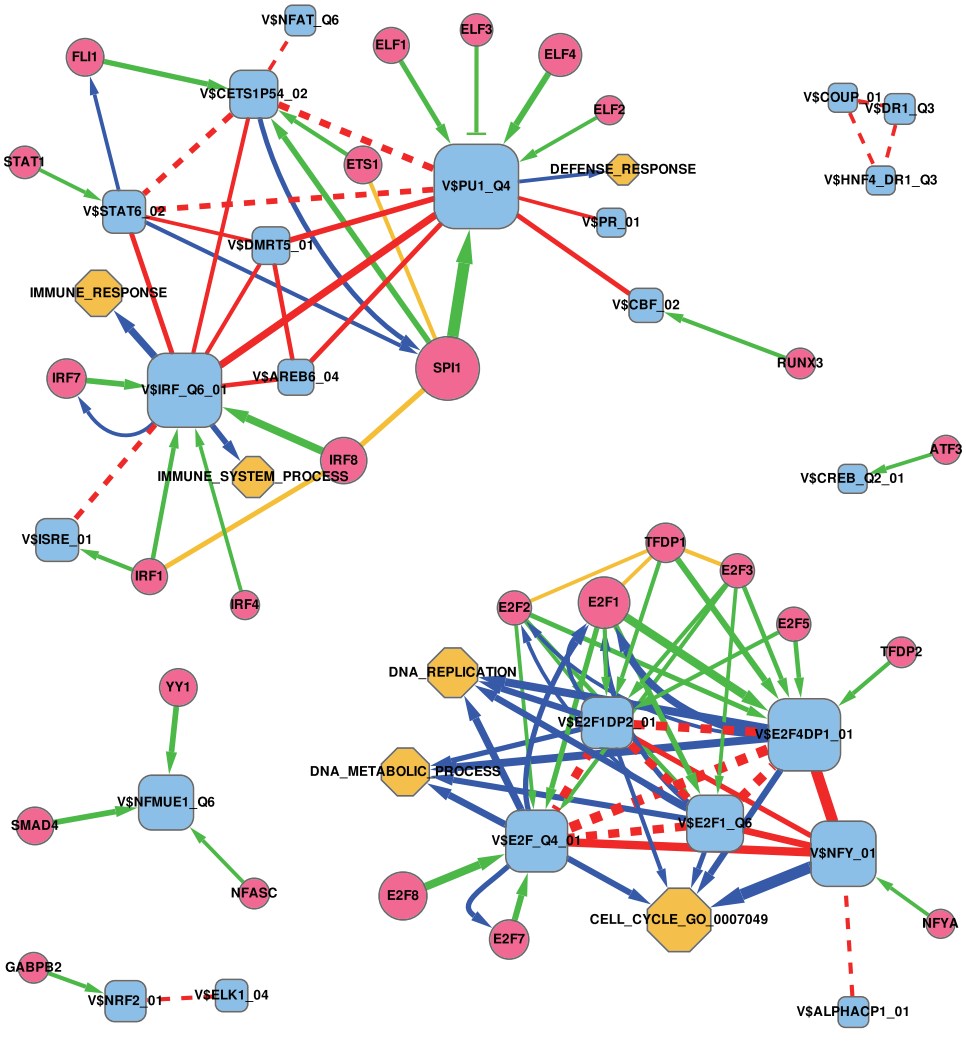

Fig. 2: Prevailing transcriptional networks in cancer cells. EEM based network prediction were applied to 122 cancer microarray datasets and frequently predicted nodes and edges were displayed as a "meta-network". Node size and edge width proportionally reflect frequency of their prediction. Pink, light blue, and yellow nodes denote TF, motif, and GO nodes, respectively. Solid red edges indicate MMI edges while dashed red edges indicate "pseudo-MMI edges" which might be bound by the same TF. Green normal directed edges indicate activating TF binding edges while green directed edges terminated with small bars indicate repressive TF binding edges. Blue and yellow edges represent *cis*-regulatory and PPI edges, respectively.

To pursue the comprehensiveness of analysis, we need to expand not only microarray datasets, but also seed gene sets. In this study, seed gene sets are

based on only TF binding motifs on the proximal promoter regions. Clearly, this point is a major drawback of this study; we cannot obtain a real global view of oncogenic transcriptional programs without considering distal *cis*-regulation. In previous study, we demonstrated that seed gene sets prepared from ChIP-chip data can be also input to EEM analysis; one of solutions of this problem is preparing ChIP-chip-derived seed gene sets by literature curation. As new technologies recently start to reveal a genomic view of *cis*-regulatory regions [15], another possible solution is expected; more global prediction of transcriptional targets will be feasible in the near future by combining TF binding motifs with comprehensive information about of *cis*-regulatory regions.

4. Conclusion

In this study, we presented a novel meta-analysis approach of cancer transcriptome data. Application of an extended version of EEM to more than a hundred microarray datasets revealed prevailing transcriptional networks which functions in various types of cancer. This study is the first meta-analysis to focus on cancer transcriptional networks, and has opened a way to comprehensive understanding of transcriptional networks in cancer cells.

References

[1] Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I., Soboleva, A., Tomashevsky, M., Marshall, K., Phillippy, K., Sherman, P., Muertter, R., Edgar, R. Ncbi geo: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37:D885–90, 2009.

[2] Basso, K., Margolin, A., Stolovitzky, G., Klein, U., Dalla-Favera, R., Califano, A. Reverse engineering of regulatory networks in human b cells. *Nat Genet*, 37:382–90, 2005.

[3] Blyth, K., Cameron, E., Neil, J. The runx genes: gain or loss of function in cancer. *Nat Rev Cancer*, 5:376–87, 2005.

[4] Coles, S. An introduction to statistical modeling of extreme values. *London: Springer-Verlag*, 2001.

[5] Friedman, N., Linial, M., Nachman, I., Pe'er, D. Using bayesian networks to analyze expression data. *J Comput Biol*, 7:601–20, 2000.

[6] Honda, K., Taniguchi, T. Irfs: master regulators of signalling by toll-like receptors and cytosolic pattern-recognition receptors. *Nat Rev Immunol*, 6:644–58, 2006.

[7] Lee, H., Hsu, A., Sajdak, J., Qin, J., Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res*, 14:1085–94, 2004.

[8] Matys, V., Kel-Margoulis, O., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A., Wingender, E. Transfac and its module transcompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34:D108–10, 2006.

[9] Nevins, J. The rb/e2f pathway and cancer. *Hum Mol Genet*, 10:699–703, 2001.

[10] Niida, A., Smith, A., Imoto, S., Aburatani, H., Zhang, M., Akiyama, T. Gene set-based module discovery in the breast cancer transcriptome. *BMC Bioinformatics*, 10:71, 2009.

[11] Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T., Rezwan, F., Sharma, A., Williams, E., Bradley, X., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S., Rocca-Serra, P., Sansone, S., Sklyar, N., Zhao, M., Sarkans, U., Brazma, A. Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37:D868–72, 2009.

[12] Rhodes, D., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A., Chinnaiyan, A. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc Natl Acad Sci U S A*, 101:9309–14, 2004.

[13] Sharrocks, A. The ets-domain transcription factor family. *Nat Rev Mol Cell Biol*, 2:827–37, 2001.

[14] Tamada, Y., Araki, H., Imoto, S., Nagasaki, M., Doi, A., Nakanishi, Y., Tomiyasu, Y., Yasuda, K., Dunmore, B., Sanders, D., Humphreys, S., Print, C., Charnock-Jones, D., Tashiro, K., Kuhara, S., Miyano, S. Unraveling dynamic activities of autocrine pathways that control drug-response transcriptome networks. *Pac Symp Biocomput*, 14:251–63, 2009.

[15] Visel, A., Blow, M., Li, Z., Zhang, T., Akiyama, J., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E., Pennacchio, L. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 12;457:854–8, 2009.

[16] Xu, L., Geman, D., Winslow, R. Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, 8:275, 2007.