BMC Bioinformatics

Gene set-based module discovery in the breast cancer transcriptome

BMC Bioinformatics 2009, **10**:71 doi:10.1186/1471-2105-10-71

Atsushi Niida (aniida@ims.u-tokyo.ac.jp)
Andrew D Smith (asmith@cshl.edu)
Seiya Imoto (imoto@ims.u-tokyo.ac.jp)
Hiroyuki Aburatani (haburata-tky@umin.ac.jp)
Michael Q Zhang (mzhang@cshl.org)
Tetsu Akiyama (akiyama@iam.u-tokyo.ac.jp)

ISSN	1471-2105
Article type	Research article
Submission date	19 November 2008
Acceptance date	26 February 2009
Publication date	26 February 2009
Article URL	http://www.biomedcentral.com/1471-2105/10/71

Gene set-based module discovery in the breast cancer transcriptome

Atsushi Niida*[1], Andrew D. Smith[2], Seiya Imoto[3], Hiroyuki Aburatani[4], Michael Q. Zhang[2] and Tetsu Akiyama[1]

[1]Laboratory of Molecular and Genetic Information, Institute of Molecular and Cellular Biosciences, The University of Tokyo, 1-1-1, Yayoi, Bunkyo-ku, Tokyo, 110-0032, Japan
[2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11274, USA
[3]The Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
[4]Genome Science Division, Research Center for Advanced Science and Technology, The University of Tokyo,4-6-1 Komaba, Meguro, Tokyo, 153-8904, Japan

Email: Atsushi Niida*- niida@iam.u-tokyo.ac.jp; Andrew D. Smith - asmith@cshl.edu; Seiya Imoto - imoto@ims.u-tokyo.ac.jp; Hiroyuki Aburatani - haburata-tky@umin.ac.jp; Michael Q. Zhang - mzhang@cshl.org; Tetsu Akiyama - akiyama@iam.u-tokyo.ac.jp;

*Corresponding author

Abstract

Background: Although microarray-based studies have revealed global view of gene expression in cancer cells, we still have little knowledge about regulatory mechanisms underlying the transcriptome. Several computational methods applied to yeast data have recently succeeded in identifying expression modules, which is defined as co-expressed gene sets under common regulatory mechanisms. However, such module discovery methods are not applied cancer transcriptome data.

Results: In order to decode oncogenic regulatory programs in cancer cells, we developed a novel module discovery method termed EEM by extending a previously reported module discovery method, and applied it to breast cancer expression data. Starting from seed gene sets prepared based on *cis*-regulatory elements, ChIP-chip data, and gene locus information, EEM identified 10 principal expression modules in breast cancer based on their expression coherence. Moreover, EEM depicted their activity profiles, which predict regulatory programs in each subtypes of breast tumors. For example, our analysis revealed that the expression module regulated by the Polycomb repressive complex 2 (PRC2) is downregulated in triple negative breast cancers, suggesting similarity of transcriptional programs between stem cells and aggressive breast cancer cells. We also found that the activity of the PRC2 expression module is negatively correlated to the expression of EZH2, a component of PRC2 which belongs to the E2F expression module. E2F-driven EZH2 overexpression may be responsible for the repression of the PRC2 expression modules in triple negative tumors. Furthermore, our network analysis predicts regulatory circuits in breast cancer cells.

Conclusions: These results demonstrate that the gene set-based module discovery approach is a powerful tool to decode regulatory programs in cancer cells.

Background

In the last decade, microarray technology has produced exploding amounts of cancer transcriptome data; especially, breast cancer transcriptome has been intensively profiled. Human breast tumors show diversity in their histology, prognosis, and responsiveness to treatments. The microarray technology has demonstrated that transcriptomic diversity underlies phenotypic diversity, and brought great progress in our molecular understanding of breast cancer [1]. However, compared with the increasing knowledge about the transcriptome, little is yet known about regulatory programs generating the transcriptomic diversity.

To decode gene regulatory programs controlling the breast cancer transcriptome, we searched for *cis*-regulatory motifs associated with tumor phenotypes in our previous study [2]. One of the limitations of this method is that it takes a supervised approach and requires sample information. In this study, we introduce an alternative method which focuses on expression modules and does not require sample information. An expression module is defined as a set of coexpressed genes controlled by a common regulatory mechanism. Although expression modules were originally visualized by clustering analysis of microarray data [3], methods based only on expression data are insufficient to reveal regulatory programs controlling such expression modules. Recently, approaches that combine expression data and *cis*-regulatory information have succeeded in identify gene regulatory programs of lower organisms like *Saccharomyces cerevisiae* [4, 5]. However, such module discovery approaches have rarely been applied to cancer transcriptome data, although a number of analyses based on prescribed sets of genes have also been performed in order to analyze oncogenic regulatory programs [6, 7].

Our new computational method termed EEM (Extraction of Expression Modules) is constructed for extracting expression modules in the cancer transcriptome. Our approach is based on an integrative method by Bar-Joseph et al. [5], which successfully identified yeast expression modules by integrating ChIP-chip and expression data. By combining with gene set-based approaches [6, 7], we extended their approach and made it applicable to cancer transcriptome data. Starting from seed gene sets predicted based on *cis*-regulatory elements, ChIP-chip data, and gene locus information, EEM statistically evaluates their functionality and refines them based on their expression coherence. We analyzed breast cancer microarray data by EEM, and find 10 expression modules in the breast cancer transcriptome. Our additional bioinformatics analysis validated the 10 expression modules and demonstrated their significance in the pathophysiology of breast cancer.

Methods

Methods Overview

The EEM algorithm discovers an expression module by combining two types of data: seed gene sets and expression profile data. A set of genes whose expressions are considered to be regulated by the same molecular mechanism could be predicted based on various types of data, and prepared as a seed gene set. EEM assesses functionality of the seed gene set based on expression coherence. If seed gene set functions as expression module, all genes in it are ideally expressed coherently. Although a functional seed gene set might include false positives, or non-functional module genes in the biological context of interest, at least a significant fraction of seed genes should behave coherently in the expression data. This assumption is verified by the observation that putatively functional gene sets often harbors a large cluster of genes which behave coherently (see Additional File 1). EEM extracts only such a coherently expressed gene subset, filtering out false positive or non-functional module genes. Taking a geometric approach, EEM searches for the largest subset with a minimum degree of coexpression (specified by radius parameter r). Concurrently, EEM statistically evaluates the size of the retrieved coherent subset using a Z score based on randomization tests. If the Z score is greater than the prespecified cutoff value, we conclude that the seed gene set includes a functional expression module and the coherent subset is extracted as an expression module. We observe that the expression modules extracted by EEM are more functionally enriched than seed gene sets. This observation would justify our refinement procedure (see Additional File 1).

Employing this EEM algorithm, we systematically searched for expression modules in the breast cancer transcriptome (Figure 1). In our search, a collection of seed gene sets are prepared based on *cis*-regulatory motifs, ChIP-chip data, and gene locus information. Since genes that possess a common *cis*-regulatory element in their promoters could be regulated by a common transcription factor (TF), we can predict an expression module based on the *cis*-regulatory motif. We searched human gene promoters and its mouse homolog promoters for 200 motif using PWMs obtained from the TRANSFAC and JASPAR databases [8, 9], and prepared 200 seed gene sets which include genes with common motifs in their promoters. We can also predict expression modules utilizing ChIP-chip data, which provide direct evidence of TF binding in the *cis*-regulatory regions. Published ChIP-chip results [10–14] are collected to prepare seed gene sets. DNA copy number alteration is known to have a significant effect on the cancer transcriptome as well as transcriptional regulation [15]. Hence, we also regarded it as one of expression regulatory mechanisms in cancer cells. Genes residing in a chromosomal region which is subjected to copy number alteration could be expressed coherently, and viewed as an expression module; taking sliding window approach, we prepared seed gene sets which consist of genes residing on the same chromosomal region. For each of the prepared seed gene sets, we tested the presence of coherently expressed gene subsets in breast cancer microarray data [16]. If such coherently expressed genes exist, we then extracted them as an expression module. Furthermore, the average expression profile of the predicted expression module can be considered as its activity in each of tumor samples. The expression module activity profiles were then analyzed using ordinary methods applied to gene expression profile data like clustering and survival analysis.

This approach is an extension of a module discovery method, GRAM, which is developed by Bar-Joseph et al. [5] for learning yeast expression modules from microarray and ChIP-chip data. In the first step, GRAM uses ChIP-chip data to find a small number of genes whose upstream regions are bound by common TFs with high confidence. In the second step, the microarray data is used to extract a coherently expressed subset of these genes. Finally, the resulting set is expanded by adding genes which are identified to be bound by the TFs with less strict criteria. Although GRAM and another similar method [17] require binding P values in ChIP-chip data, we relaxed this requirement by taking a gene set-based approach. Although gene set-based approaches are simpler than direct integration of ChIP-chip and expression data, they have shown substantial successes in cancer transcriptome analysis [6, 7]. From ChIP-chip data, we prepared gene sets as seed gene sets, by retrieving genes which have binding sites within specified *cis*-regulatory regions and with P values below a specified threshold. In addition to ChIP-chip data, our analysis utilized *cis*-regulatory motifs and locus information to generate gene sets, because available human

ChIP-chip data are insufficient for comprehensive analysis and they also provide clear evidence of transcriptional regulation or genomic alteration. Although EEM also takes a module discovery approach similar to GRAM, there are some clearly different points. In contrast to GRAM, EEM starts from a sufficient number of genes that are predicted to be under common regulatory mechanism, and refines them to produce a final expression module utilizing expression profiles. In this process, EEM evaluates statistical significance of the identified expression module by measuring how many of module genes are coherently expressed in the expression data. This statistical evaluation based on the expression coherence is a novel feature which is not implemented by other module discovery methods.

EEM is also regard as one of gene set screening methods like Gene Set Enrichment Analysis (GSEA) [18]. GSEA screens for gene gets that have a significant bias in a ranked list according to their differential expressions between two sample groups, while our approach searches for significant gene sets based on their expression coherence. However, because GSEA takes a supervised approach which uses sample labels, it potentially fails to identify expression modules which do not correlate with sample labels. By contrast, EEM realizes an unsupervised analysis, which does not depend on sample information and can search for expression modules more globally.

As we mentioned above, our method finally produces activity profiles of expression modules. Because microarray data usually include expression profiles of thousands of genes, it is difficult to understand the raw data intuitively. On the other hand, since our activity profiles consist of those of a small number of expression modules, they provide concise description of the transcriptome that allows it to be understood more easily. This problem can also be solved using dimension reduction approach of gene expression data. Dimension reduction is originally addressed by a study utilizing singular value decomposition [19], and can be performed by many other methods [20–23]. However, because most of them are based on purely mathematical framework, deduced components do not necessarily have biological meanings and are often difficult to understand biologically. By contrast, since expression modules deduced by EEM are derived from biologically meaningful seed gene sets, they can always be associated with molecular mechanisms. By extracting module modules as biologically meaningful components in expression data, EEM provides intuitively understandable views of transcriptomes.

EEM algorithm

Let $E = \{e_1, \ldots, e_n\}$ be a set of gene expression profiles such that each $e_i \in E$ is a vector $e_i = (e_{i1}, \ldots, e_{im})$ of values with e_{ij} giving the expression of the i-th gene in the j-th condition (or sample). Each $e_i \in E$ then exists as a point in a continuous m-dimensional gene expression space $\mathcal{S}$. Although the expression values can be obtained by any means, we may assume they are from gene expression microarray experiments. EEM operates on a subset $E_M \subseteq E$ called the seed gene set (we describe below how such seed gene sets are obtained). For a given radius 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 the coherently expressed gene set (or simply *coherent set*), and the point x is called the *center* of C_x. The objective of EEM is to find maximal sized coherent set C_B (and corresponding center B) for the genes in E_M. We remark that the center B may not necessarily correspond to any profiles in E_M. We also call B the *activity profile* for genes in C_B. As stated above, the distance measure we use to define a degree of co-expression between genes is Euclidean distance. In practice the expression profiles are normalized, so this is equivalent to measuring similarity using Pearson correlation. Our method is intended for large datasets (based on microarray expression profiles), and employs a heuristic modified from a previously proposed algorithm [5]. Similar geometric optimization problems arise in the context of clustering [24, 25]. EEM attempts to find an optimal center for E_M in two stages. The first stage identifies a candidate center B_1 from among the expression profiles in E_M. For each $e_i \in E_M$, the set C_{e_i} is constructed (see Equation 1). The profile $e_i \in E_M$ with maximal $|C_{e_i}|$ is retained as B_1. The second stage uses B_1 to find an improved center. Let $T \subseteq E_M$ denote the set containing the 9 profiles in E_M closest to B_1 along with B_1 itself (*i.e.* $|T| = 10$). For each triple $\{t_1, t_2, t_3\} \subset T$, the mean profile $t = (t_1 + t_2 + t_3)/3$ is constructed and C_t is evaluated. The mean profile t that maximizes $|C_t|$ over all triples from T is retained and returned by EEM as the optimal center B along with the identity of genes in C_B (see the Appendix section for a pseudocode for this optimization procedure).

EEM includes a critical step to estimate the statistical significance of the size of the coherent set, given the full set of expression profiles from the expression data set (recall that the procedure described above operates on a subset $E_M \subseteq E$ defined by a seed gene set). This is accomplished by sampling subsets of size k uniformly at random from the full set E of expression profiles, where $k = |E_M|$. The EEM optimization procedure (described above and summarized as in the Appendix section) is applied to each sampled subset to produce an empirical distribution for the sizes of coherent sets derived from E. The mean and standard deviation from this empirical distribution are used to obtain a Z score for $|C_B|$ corresponding to E_M, and Z score threshold is used to determine whether a particular coherent set is significant. Our results are based on Z scores estimated using 500 randomly sampled subsets of expression profiles.

Preparation of Expression data

From GEO database, we downloaded Affymetrix GeneChip data of 252 breast tumor samples [16] (the accession number is GSE3494). Absolute expression values of a data set were converted to the logarithmic scale and normalized so that the mean is equal to 0 and the variance is equal to 1 in each sample. The Probe set IDs were converted to Ensembl gene IDs. In cases that one gene ID matches multiple probe set IDs, the probe set which shows the most variance among the samples was mapped to the gene. A variation filter was then applied to the data, and we obtained 5000 genes with the highest variance. The expression profiles of the 5000 genes were normalized across samples and subjected to the following analysis.

Preparation of seed gene sets

Preparation based on cis-regulatory motifs

We prepared promoter data of human genes and mouse genes from the Ensembl database (Release 44). Assuming TSSs (transcription start sites) as gene starts registered in Ensembl, a repeat-masked promoter sequence covering the 500bp upstream and 100bp downstream of the TSS for each gene was retrieved from the genome sequences.

As *cis*-regulatory motif data, we prepared PWMs (position weight matrices). The value f_{ib} of a PWM represents the frequency of nucleotide base b at the i-th position in a motif. The frequencies of bases in each position are normalized so that $\sum_{b \in \{a,t,g,c\}} f_{ib} = 1$. If $f_{ib} = 0$, we reassigned $f_{ib} = 0.001$ to avoid errors in log calculations. We acquired a total of 601 PWMs, which consist of vertebrate 513 PWMs annotated as “good” in TRANSFAC 10.1 [8] and 88 PWMs from JASPAR core [9]. We then removed extremely simple or complex PWMs based on their information contents to produce a set of total 511 PWMs whose information contents range from 5 to 15. The information content R of a PWM is defined as follows:

$$R \quad = \quad 2w - \sum_{i=1}^{w} H_i,$$

where w is the width of the motif, and H_i is the information entropy at the i-th position defined by

$$H_i = -\sum_{b=a,c,g,t} f_{ib} \log_2 f_{ib}.$$

Since this set includes highly redundant PWMs, they were subjected to clustering to reduce the redundancy. For clustering, the dissimilarity between two PWMs A and B was calculated based on the Kullback-Leibler divergence. At every alignment offset, the PWMs were extended using a column representing the uniform base frequency ($f_{ib} = 0.25$ for all b) so that all position of two aligned motifs were matched. As for this alignment step, we followed a method used by Xie et al. [26] For every pair of the extended PWMs, A' and B', whose length are w', the dissimilarity $D_{A'B'}$ is calculated by:

$$D_{A'B'} = \sum_{i=1}^{w'} \sum_{b=a,t,g,c} (f_{ib}^{A'} - f_{ib}^{B'}) \log \frac{f_{ib}^{A'}}{f_{ib}^{B'}}.$$

We assumed the lowest score of $D_{A'B'}$ as the dissimilarity between A and B, D_{AB}. Note that $D_{AB} = D_{BA}$ holds. Using the partition around medoids algorithm, the 511 PWMs are divided into 200 clusters. We used 200 medoids of the clusters in the following analyses.

To predict expression modules, we searched promoter sequences for TF binding motifs based on the log odds ratio L between a PWM and background base frequency f_b^{bg}. Using the STORM program [27], we calculated log odds ratio L_s for every subsequence of each promoter s (including the complementary strand), whose length is equal to the width of the motif of interest, w:

$$L_s = \sum_{i=1}^{w} \log \frac{f_{ib_i}}{f_{b_i}^{bg}}.$$

In our analyses, f_b^{bg} is the base composition of each promoter, and the maximum of L_s in a human promoter sequence was taken as the motif score L^{human} for the sequence. For human genes whose mouse homolog is registered in Ensembl, L^{mouse} was also calculated. L^{human} and L^{mouse} were then averaged to produce the final motif score L. For human genes that do not have any homologs, we used L^{human} as L. Among all genes analyzed, genes which score the 5% highest L were assumed as a seed gene set regulated by the motif. For each of the 200 PWMs, we performed this procedure to produce 200 seed gene sets.

Preparation based on ChIP-chip data

We obtained TF-bound gene sets identified in ChIP-chip experiments; CREB1, FOXA2, HINF1, HNF4, HNF6 and USF1 in hepatocyte [13], and NFκB in U931 cells [11]. We assumed genes which are bound by

EED and SUZ12, and trimetylated at histone H3 lysin-27 in ES cells [12] as a PRC2-bound gene set. A p53-bound gene set was obtained from [14]. For ER-bound genes, we analyzed ChIP-chip results [10], and retrieved genes which are bound by ER in their promoters (from the TSSs to 3kbp upstream), 5'UTRs or first introns with P values of $< 10^{-50}$.

Preparation based on locus information

A sliding window was employed to prepare seed gene sets consisting of genes that reside on the same chromosomal region. A 10 Mb window was slid along each chromosome at interval of 1 Mb. At each position, if the window includes more than or equal to 10 genes, we pooled them in the gene set library and obtained 2766 seed gene sets.

Setting of parameters

The radius parameter

To specify the radius parameter r in EEM, we converted the absolute distance to a relative distance for each expression data set. Expression spaces specified by different data sets have different dimensions and different densities of points. Therefore, instead of the absolute distance, we used a relative distance which practically acts as an equal measure for different data sets. To convert the absolute distance, d^{absolute} to such a relative distance, d^{relative}, we define coherent set C_x^{all} for all $e_i \in E$, similarly to Equation 1:

$$C_x^{\text{all}} = \{e_i \in E : d(e_i, x) \leq r\},$$

where r is a given radius parameter and point $x \in \mathcal{S}$. The maximal sized coherent set C_B^{all} can also be found based on the above-described algorithm for radius $r = d^{\text{absolute}}$. d^{relative} is then defined as follows:

$$d^{\text{relative}} = \frac{|C_B^{\text{all}}|}{|E|}.$$

In our analysis, we assumed $r = 0.05$ in the relative distance. It should be should noted that we also tried to use $r = 0.03$ and 0.10 and observed that the identified expression modules are essentially the same; although the number of module genes increases as r increases, statistical significance, activity profiles and enriched GO terms were essentially unchanged. Exceptionally, we used a larger radius for the PRC2 expression module (see Additional File 1).

The threshold of Z scores

We assumed that a seed gene set includes a functional expression module, if its Z score is greater than a threshold. In this paper, we set 4.0 as the threshold. The reason why we set 4.0 as the threshold is that when we permuted gene labels in microarray data, no expression modules showed greater Z scores than 4.0. Therefore, we concluded that this threshold is sufficiently conservative and the resulting expression modules are expected to have high accuracy.

Evaluation of obtained expression modules

Gene ontology analysis

We evaluated the enrichment of GO categories in each identified expression module by using GO::TermFinder [28]. The GOA file we used was obtained from EBI. To predict biological function for each expression module, we also reported the GO category scoring the lowest P values as the most enriched GO.

Survival analysis

Kaplan-Meier survival curves were obtained for two patient groups with high or low activity of each identified expression module. The cutoff of the high and low groups was optimized to achieve the most significant P value in the Kaplan-Meier analysis with at least 20% patients at each group. Since the optimized P values in the Kaplan-Meier analysis overestimate the significance, we reported P values based on Cox regression analysis.

Network analysis

We evaluated enrichment of physically interacting gene pairs in the expression modules based on PPI data obtained from the Human Protein Reference Database (HPRD) [29]. To calculate a Z score for the number of interacting pairs in an expression module, we randomly sampled 500 gene sets with the same number of genes of the considered expression module. After finding expression modules having the significant number of PPIs, we constructed PPI subnetworks for elucidating molecular circuits. The PPI subnetworks were constructed by the interacting protein pairs and their first neighbor in PPI data. The network visualization was performed using CytoScape [30].

Results and Discussion

Our systematic search identified 10 expression modules in the breast cancer transcriptome (Table 1). Based on *cis*-regulatory motifs, we identified expression modules regulated by E2F, NFY, RUNX, IRF, and ETS family TFs. Hereafter, we use the TF name to refer to the family, e.g. the E2F module will refer to the module associated with the E2F family of TFs. ChIP-chip data led us to identify expression modules regulated by the estrogen receptor (ER), Polycomb repressive complex 2 (PRC2), and NFκB. In addition to these transcriptional modules, incorporation of locus information yielded two expression modules, which are located on the 17q12 and 8q24 locus. The reproducibility of these results was confirmed by analysis using other independent microarray data [31] (see Additional File 1).

Each expression module consists of dozens of genes. We performed Gene Ontology (GO) analysis to examine whether the obtained expression modules are enriched in genes involved in specific cellular activities (Table 1). The GO analysis showed that most of the expression modules deduced from *cis*-regulatory motifs and ChIP-chip data contain a significant number of genes sharing common GO terms, such as immune response and cell cycle. Thus, these transcriptional modules have the potential to function for specific cellular activities.

The EEM analysis predicted the activity profiles of the 10 expression modules; we performed hierarchical clustering analysis of them as performed for ordinary gene expression profiles (Figure 2). We found that clustering of tumor samples based only on these 10 expression modules succeeded in dividing samples into several subtypes that are consistent with clinical information and gene expression profiles. This observation suggests that a significant degree of diversity of the breast cancer phenotypes can be explained by only these 10 expression modules. In other words, this result demonstrates that the EEM analysis successfully reduced gene expression data of extremely high dimension to the 10 components. We also performed survival time analysis and found expression modules associated with prognosis (Figure 3, see below).

The E2F and NFY expression modules show similar activity profiles, which are activated in high grade breast tumors and strongly correlated with poor prognosis. They also share common modules genes and appear to cooperatively regulate the cell cycle. Similar expression profiles shared by the RUNX ETS, IRF and NFκB expression modules also suggest that these TFs regulate immune pathways cooperatively. These results are consistent with those of previous studies [32–36]. As expected, the ER expression module was found to be the most critical determinant of tumor subtypes. Their activity profiles are strongly correlated with ER status, demonstrating the validity of our approach. The 17q12 and 8q24 expression modules are derived from known amplified regions [15]. The 17q12 expression module contains the ERBB2 gene, while the 8q24 expression module contains genes residing near the Myc locus. The 17q12 expression is an important determinant of tumor subtypes and survival time. Although the 8q24 expression modules are not clearly associated with any subtypes, its upregulation is related to poor prognosis.

Triple-negative breast cancers characterized by a lack of the ER, progesterone receptor (PgR), and ERBB2 expression have attracted special attention in breast cancer research. In addition to their aggressive phenotype, they lack the benefit of specific therapy that targets these genes and, therefore, are associated with short survival. The sample cluster enriched for triple negative cancers has characteristic expression module activity profiles; the E2F and NFY expression modules are upregulated, while the ER, PRC2 and 17q12 expression modules are downregulated. Among them, the PRC2 expression module is especially intriguing. PRC2 is an epigenetic gene silencer, which plays a critical role in the maintenance of stem cells. They have also been reported to be implicated in neoplastic development. The PRC2 expression module was derived from gene sets bound by EED and SUZ12 and trimetylated at histone H3 lysin-27 in ES cells [12]. Therefore, our observation suggests similarity of transcriptional programs in both stem cells and the aggressive breast cancer. Triple-negative breast cancers are known to have poorly differentiated phenotype histology, which might be maintained by a PRC2-directed regulatory program. Recently a drug targeting PRC2 is developed [37]; PRC2 could be a therapeutic molecular target in triple negative breast cancers. Furthermore, we found that EZH2, a component of PRC2, belongs to the E2F expression module [38], while its expression is inversely correlated with profiles of the PRC2 expression module. This finding suggests that E2F-driven EZH2 overexpression is important for repression of the PRC2 expression modules in triple negative tumors. It should be noted that another independent study employing bioinformatics has also recently shown that PRC2 target genes are downregulated in malignant breast tumors, supporting our finding [39].

Inspection of individual genes in each expression module provided insights into regulatory networks in breast tumors. For example, our result suggests that auto-regulatory designs are prevailing in mammalian transcriptional networks. The E2F expression module contains three E2F family genes: E2F1, E2F7, and E2F8. The ER expression module also harbors ER itself and its interacting co-factor, FOXA1 [40]. Furthermore, RUNX3, one of the RUNX family genes, belongs to the RUNX expression module (see Additional File 1).

To obtain further insights into regulatory networks, we performed network analysis using protein-protein interaction (PPI) data in the HPRD [29]. Previous studies have demonstrated a significant correlation between yeast PPI and transcriptional networks [41, 42]. This observation prompted us to examine whether human expression modules identified by EEM also tend to contain genes involved in the same protein complex. We calculated the Z score for the count of module gene pairs which interact directly (nearest neighbor pairs), and those which interact directly or via the next node of each (nearest or next-to-nearest neighbor pairs) (Table 1). This analysis showed that the transcriptional modules that were identified based on *cis*-regulatory motifs or ChIP-chip data harbor a statistically significant number of physically interacting pairs, revealing a tight coupling of transcriptional and PPI networks in mammalian cells.

Finally, by extracting expression module genes and interacting partners from the PPI data, we depicted molecular circuits in breast tumors based on multiple lines of evidence: PPI, expression coherence, *cis*-regulatory motifs and ChIP-chip data (Figure 4). These network views revealed that E2F regulates cell cycle hub genes, such as CDC2 and Cyclins, in cooperation with NFY. The transcriptional sub-network involving RUNX, ETS, IRF, and NFκB regulates immune circuits involving various chemokines and chemokine receptors, some of which have been reported to be involved in tumor growth, invasion, and metastasis [43]. For example, a recent study showed that CCL5 can induce metastasis of breast tumors [44]. Consistent with a previous report [45], EEM predicted that ETS, IRF and NFκB transcriptionally control CCL5, suggesting that these TFs are responsible for CCL5-matiated metastasis. Human breast tumors are histologically complex and contain a variety of cell types in addition to the carcinoma cells. Hence, the transcriptional programs controlling these immune circuits could operate not in the carcinoma cells themselves, but in the tumor microenvironment. Indeed, CCL5 was reported to be secreted from mesenchymal stem cells. Also, many other module genes are known to be specifically expressed in immune cells such as lymphocytes and macrophages. On the other hand, we could also identify the IRF expression module in the transcriptome of breast cancer cell lines (see Additional File 1). Thus, the IRF expression module may function in the carcinoma cells. Recently, global expression profiling of distinct cell population in breast tumors has been attempted [46]. We expect that application of EEM to such data will clearly show cell type specific regulatory programs.

Previously, Segal et al. [47] also reported expression modules in cancer transcriptome. However, their method identifies functional modules based on relative up or down-regulation of module genes in each sample, which contrasts with our method that takes into account expression coherence across all samples. In this study, our method succeeded in reducing the expression profiles of thousands of genes to the activity profiles of 10 expression modules which explains a significant degree of diversity of the breast cancer phenotypes. It should be noted that the expression module activity profiles show some similarity with oncogenic pathway activity profiles depicted by Bild et al. [48]. Their Bayesian regression-based method learns signatures that can predict pathway activities of clinical tumor samples from microarray experiments using cell line models. By contrast, our method searches for coexpressed gene sets under common regulatory mechanisms using prescribed gene sets. Hence, these two methods are considered to be complementary to each other. In addition to clustering analysis, we applied survival analysis to the expression module activity profiles and succeeded in identifying expression modules associated with prognosis. While a number of studies have identified signature genes associated with prognosis in breast cancer [6], the result suggested that our approach is also useful to search for such signature genes. Based on EEM-deduced expression modules, we also predicted transcriptional regulatory networks in breast tumors. Although some previous studies have addressed reverse engineering problems of regulatory networks in cancer cells [49], they are only based on correlation or conditional independence of expression profiles. On the other hand, our method incorporates evidence of direct TF regulation. Collectively, we can say that EEM is a powerful module discovery method that provides various types of information essential for a deeper understanding of cancer transcriptomes.

As well as these notable advantages, our approach has several limitations. EEM assumes that module genes behave coherently across all samples. However, because gene regulatory programs are usually functional in specific contexts, it might be more appropriate to assume that module genes are assumed to behave coherently in only a subset of samples. Under the current assumption, we might fail to find tumor subtype-specific expression modules. It is also probable that different genes in a common module are controlled by different modes of a regulatory program (e.g., it is known that some TFs act as both activators and repressors, depending on target genes). Although current version of EEM cannot detect expression modules which show such complex patterns of expression profiles, future studies will improve the algorithm to overcome these limitation. Also, it should be noted that EEM uses the size of a coherent subset as an index of expression coherence. We use the arbitrary parameter r to specify the minimum degree of coexpression of coherent gene subsets. It is possible that EEM misses tightly coexpressed small modules or loosely coexpressed large modules, depending the values of r. In such a case, optimization of r based on Z scores will improve results. Recently, another gene set screening method based on a different index of expression coherence [50] was also reported. Comparison of different coherence indexes should be addressed in future studies.

To search for expression modules utilizing EEM, we prepared a collection of seed gene sets based on *cis*-regulatory motifs and ChIP-chip data. Therefore, comprehensiveness of our method depends on coverage of these data. Although a large number of motifs are already registered in databases, the quality and coverage seem to be incomplete. However, because several promising methods that are suitable for high-throughput determination of TF binding specificity have been devised [51, 52], more accurate and comprehensive data of regulatory motifs are expected to be available soon. Furthermore, instead of ChIP-chip, a more high-throughput and cost-effective alternative, the ChIP-Seq technique has recently emerged [53, 54]. It is expected that a great deal of TF binding site data will be produced by ChIP-Seq in the next decade. These increasing amounts of data will enable more global analysis in the near future. In this study, we assumed that each expression module is regulated by a single TF. However, combinatorial regulations by multiple TFs are known to be essential in mammalian regulatory networks. Combinatorial analysis will be enabled by constructing expression modules based on *cis*-regulatory information about multiple TFs. We will focus on this problem in future studies.

Conclusions

We apply a new gene-set based module discovery method, EEM, to breast cancer microarray data, and revealed 10 principal expression modules in the breast cancer transcriptome. The subsequent analyses of expression module activity profiles and predicted regulatory networks demonstrated their importance in the pathophysiology of breast cancer. We believe that our method will be a powerful tool to decode gene regulatory programs in cancer transcriptomes.

Authors' contributions

A. N., T. A., and H. A. designed research; A. N. performed research; A. N., A. D. S. and M. Q. Z. contributed new analytic tools; A. N., T. A., and S. I. wrote the paper: all authors read and approved the final manuscript.

Appendix

A pseudocode for the optimization procedure in EEM is as follows:

```
comment: $E_M$ is the set of expression profiles for a seed gene set (i.e. some subset of available genes).
comment: $r$ is the given radius value.
```

```
$C_{B_1} \Leftarrow \emptyset$
**for all** $e_i \in E_M$ **do**
  $C_{e_i} \Leftarrow \{e_j \in E_M : d(e_j, e_i) < r\}$
  **if** $|C_{e_i}| > |C_{B_1}|$ **then**
    $B_1 \Leftarrow e_i$
    $C_{B_1} \Leftarrow C_{e_i}$
  **end if**
**endfor**
$C_B \Leftarrow \emptyset$
$T \Leftarrow \{B_1$ and the 9 profiles in $G_{B_1}$closest to $B_1\}$
**for all triples** $\{t_1, t_2, t_3\} \subset T$ **do**
  $t \Leftarrow (t_1, t_2, t_3)/3$
  $C_t \Leftarrow \{e_i \in E_M : d(e_i, t) < r\}$
  **if** $|C_t| > |C_B|$ **then**
    $B \Leftarrow t$
    $G_B \Leftarrow C_t$
  **end if**
**end for**
**return** $(B, C_B)$
```

Acknowledgements

We thank S. Mori, Y. Hoshida, and Y. Suzuki for their valuable comments and helpful suggestions on earlier version of this manuscript. We appreciate comments from the two anonymous referees that improve the quality of our article considerably. Computation time was provided by the Super Computer System, Human Genome Center, Institute of Medical Science, University of Tokyo. This work was supported by Grants-in-Aid for Scientific Research on Priority Areas.

References

1. Sotiriou C, Piccart MJ: **Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?** *Nat Rev Cancer* 2007, **7**:545-553.

2. Niida A, Smith AD, Imoto S, Tsutsumi S, Aburatani H, Zhang MQ, Akiyama T: **Integrative bioinformatics analysis of transcriptional regulatory programs in breast cancer cells.** *BMC Bioinformatics* **9**:404.

3. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**:14863-14868.

4. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117**:185-198.

5. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21**:1337-1342.

6. Nevins JR, Potti A: **Mining gene expression profiles: expression signatures as cancer phenotypes.** *Nat Rev Genet* 2007, **8**:601-609.

7. Segal E, Friedman N, Kaminski N, Regev A, Koller D: **From signatures to models: understanding cancer using microarrays.** *Nat Genet 37* 2005, **Suppl**:S38-S45.

8. Matys V, Kel-Margoulis OV, 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 AE, Wingender E: **TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.** *Nucleic Acids Res* 2006, **34**:D108-D110.

9. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, Wasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34**:D95-D97.

10. Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, Wang Q, Bekiranov S, Sementchenko V, Fox EA, Silver PA, Gingeras TR, Liu XS, Brown M: **Genome-wide analysis of estrogen receptor binding sites.** *Nat Genet* 2006, **38**:1289-1297.

11. Schreiber J, Jenner RG, Murray HL, Gerber GK, Gifford DK, Young RA: **Coordinated binding of NF-kappaB family members in the response of human cells to lipopolysaccharide.** *Proc Natl Acad Sci U S A* 2006, **103**:5899-5904.

12. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, Koseki H, Fuchikami T, Abe K, Murray HL, Zucker JP, Yuan B, Bell GW, Herbolsheimer E, Hannett NM, Sun K, Odom DT, Otte AP, Volkert TL, Bartel DP, Melton DA, Gifford DK, Jaenisch R, Young RA: **Control of developmental regulators by Polycomb in human embryonic stem cells.** *Cell* 2006, **125**:301-313.

13. Odom DT, Dowell RD, Jacobsen ES, Nekludova L, Rolfe PA, Danford TW, Gifford DK, Fraenkel E, Bell GI, Young RA: **Core transcriptional regulatory circuitry in human hepatocytes.** *Mol Syst Biol* 2006, **2**:2006.0017.

14. Wei CL, Wu Q, Vega VB, Chiu KP, Ng P, Zhang T, Shahab A, Yong HC, Fu Y, Weng Z, Liu J, Zhao XD, Chew JL, Lee YL, Kuznetsov VA, Sung WK, Miller LD, Lim B, Liu ET, Yu Q, Ng HH, Ruan Y: **A global map of p53 transcription-factor binding sites in the human genome.** *Cell* 2006, **124**:207-219.

15. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci U S A* 2002, **99**:12963-12968.

16. Miller LD, Smeds J, George J, Vega VB, Vergara L, Ploner A, Pawitan Y, Hall P, Klaar S, Liu ET, Bergh J: **An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival.** *Proc Natl Acad Sci U S A* 2005, **102**:13550-13555.

17. Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K: **Inferring transcriptional modules from ChIP-chip, motif and microarray data.** *Genome Biol* 2006, **7**:R37.

18. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102**:15545-15550.

19. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci U S A* 2000, **97**:10101-10106.

20. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CW, Mesirov JP: **Metagene projection for cross-platform, cross-species characterization of global transcriptional states.** *Proc Natl Acad Sci U S A* 2007, **104**:5959-5964.

21. Moloshok TD, Klevecz RR, Grant JD, Manion FJ, Speier WF 4th, Ochs MF: **Application of Bayesian Decomposition for analysing microarray data.** *Bioinformatics* 2002, **18**:566-575.

22. Dueck D, Morris QD, Frey BJ: **Multi-way clustering of microarray data using probabilistic sparse matrix factorization.** *Bioinformatics Suppl* 2005, **1**:i144 - i151.

23. Kim PM, Tidor B: **Subsystem identification through dimensionality reduction of large-scale gene expression data.** *Genome Res* 2003, **13**:1706-1718.

24. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9**:1106-1115.

25. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: **Adaptive quality-based clustering of gene expression profiles.** *Bioinformatics* 2002, **18**:735-746.

26. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**:338-345.

27. Schones DE, Smith AD, Zhang MQ: **Statistical significance of cis-regulatory modules.** *BMC Bioinformatics* 2007, **8**:19.

28. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G: **GO::TermFinder?open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.** *Bioinformatics* 2004, **20**:3710-3715.

29. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK,

Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JG, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A, Pandey A: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13**:2363-2371.

30. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13**:2498-504.

31. Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, Gray JW: **Genomic and transcriptional aberrations linked to breast cancer pathophysiologies.** *Cancer Cell* 2006, **10**:529-541.

32. Zhu W, Giangrande PH, Nevins JR: **E2Fs link the control of G1/S and G2/M transcription.** *EMBO J* 2004, **23**:4615-4626.

33. Sharrocks AD: **The ETS-domain transcription factor family.** *Nat Rev Mol Cell Biol* 2001, **2**:827-837.

34. Blyth K, Cameron ER, Neil JC: **The RUNX genes: gain or loss of function in cancer.** *Nat Rev Cancer* 2005, **5**:376-387.

35. Perkins ND: **Integrating cell-signalling pathways with NF-kappaB and IKK function.** *Nat Rev Mol Cell Biol* 2007, **8**:49-62.

36. Honda K, Taniguchi T: **IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors.** *Nat Rev Immunol* 2006, **6**:644-658.

37. Tan J, Yang X, Zhuang L, Jiang X, Chen W, Lee PL, Karuturi RK, Tan PB, Liu ET, Yu Q: **Pharmacologic disruption of Polycomb-repressive complex 2-mediated gene repression selectively induces apoptosis in cancer cells.** *Genes Dev* 2007, **21**:1050-1063.

38. Bracken AP, Pasini D, Capra M, Prosperini E, Colli E, Helin K: **EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer.** *EMBO J* 2003, **22**:5323-5335.

39. Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, Weinberg RA: **An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors.** *Nature Genet* 2008, **40**:499-507.

40. Carroll JS, Liu XS, Brodsky AS, Li W, Meyer CA, Szary AJ, Eeckhoute J, Shao W, Hestermann EV, Geistlinger TR, Fox EA, Silver PA, Brown M: **Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1.** *Cell* 2005, **122**:33-43.

41. Simonis N, Gonze D, Orsi C, van Helden J, Wodak SJ: **Modularity of the transcriptional response of protein complexes in yeast.** *J Mol Biol* 2006, **363**:589-610.

42. Manke T, Bringas R, Vingron M: **Correlating protein-DNA and protein-protein interaction networks.** *J Mol Biol* 2003, **333**:75-85.

43. Balkwill F: **Cancer and the chemokine network.** *Nat Rev Cancer* 2004, **4**:540-550.

44. Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, Richardson AL, Polyak K, Tubo R, Weinberg RA: **Mesenchymal stem cells within tumour stroma promote breast cancer metastasis.** *Nature* 2007, **449**:557-563.

45. Liu J, Ma X: **Interferon regulatory factor 8 regulates RANTES gene transcription in cooperation with interferon regulatory factor-1, NF-kappaB, and PU.1.** *J Biol Chem* 2006, **281**:19188-19195.

46. Allinen M, Beroukhim R, Cai L, Brennan C, Lahti-Domenici J, Huang H, Porter D, Hu M, Chin L, Richardson A, Schnitt S, Sellers WR, Polyak K: **Molecular characterization of the tumor microenvironment in breast cancer.** *Cancer Cell* 2004, **6**:17-32.

47. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36**:1090-1098.

48. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, Joshi MB, Harpole D, Lancaster JM, Berchuck A, Olson JA Jr, Marks JR, Dressman HK, West M, Nevins JR: **Oncogenic pathway signatures in human cancers as a guide to targeted therapies.** *Nature* 2006, **439**:353-357.

49. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37**:382-390.

50. Kim TM, Chung YJ, Rhyu MG, Jung MH: **Inferring biological functions and associated transcriptional regulators using gene set expression coherence analysis.** *BMC Bioinformatics* 2007, **8**:453.

51. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays.** *Nat Genet* 2004, **36**:1331-1339.

52. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities.** *Nat Biotechnol* 2006, **24**:1429-1435.

53. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K: **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129**:823-837.

54. Johnson DS, Mortazavi A, Myers RM, Wold B: **Genome-wide mapping of in vivo protein-DNA interactions.** *Science* 2007, **316**:1497-1502.

Figures

Figure 1 - Schema of systematic search for expression modules

We prepared a collection of seed gene sets based on *cis*-regulatory motifs, ChIP-chip and gene locus information. We next statistically evaluated whether each seed gene set includes a significant large number of coherently expressed genes in expression profile data. If such a coherently expressed gene subset exists, we assumed it as an expression module, and obtained its averaged expression profile as an activity profile.

Figure 2 - Clustering analysis of expression module activity profiles in breast tumors.

Activity profiles of 10 expression modules extracted from breast cancer expression data were analyzed by hierarchical clustering. Red indicates increased activity and blue indicates decreased activity. The upper color bars indicate clinical and gene expression information of each tumor sample; histological grades (G1: red, G2: yellow G3: blue), p53 status (wildtype: red, mutant: blue), expression subtypes, ER, PgR and nodal status (positive: red, negative: blue), ERBB2 and EZH2 expression (increased expression: red, decreased expression: blue). The expression subtypes are based on the five major branches in the clustering dendrogram of the gene expression profiles. In the upper dendrogram, red branches represent a sample cluster which is enriched for triple negative breast tumors.

Figure 3 - Survival time analysis of expression module activity profiles in breast tumors.

Associations between survival time of patients and expression module activities were evaluated. Kaplan-Meier curves for two patient groups with different activities of each expression module were displayed using a color code (high survival rate: red, low survival rate: blue). P values are calculated for coefficients in Cox regression analysis.

Figure 4 - PPI and transcriptional sub-networks in breast tumors.

(A) The E2F and NFY expression modules and an overlapping PPI sub-network. (B) The ETS, IRF, NFκB, and Runx expression modules and an overlapping PPI sub-network. For clarity, we displayed nodes which have more than one links. Red, yellow and blue nodes denote transcriptional regulators, regulated genes, and their interacting partners, respectively. Red links denote transcriptional regulations predicted by our analysis, and blue links denote PPIs registered in the HPRD.

Tables

Table 1 - Motif associated with histological grades or prognosis identified based on independent datasets

Module ID	Size	Z score	The most enriched GO	P value for the most enriched GO	Z score for nearest neighbor pair	Z score for nearest or next to nearest pairs
ETS	47	10.0	immune response	4.55×10^{-15}	10.9	9.85
IRF	47	7.59	immune response	6.19×10^{-12}	N.S.	3.69
E2F	37	6.67	cell cycle	1.98×10^{-20}	32.5	24.5
RUNX	34	5.57	immune response	2.57×10^{-11}	12.1	6.08
NFY	30	4.22	cell cycle	9.48×10^{-14}	15.4	12.3
NFκB	29	9.53	immune response	1.51×10^{-7}	17.9	4.65
ER	17	9.45	-	-	N.S.	N.S.
PRC2	61	5.87	multicellular organismal development	4.90×10^{-7}	N.S.	5.07
8q24	10	7.80	-	-	N.S.	N.S.
17q12	11	7.78	-	-	N.S.	N.S.

N.S. denotes ‘not significant’.

Additional Files

Additional File 1

File format: PDF

Title: Supplementary text

Description: supplementary discussions, tables and figures.

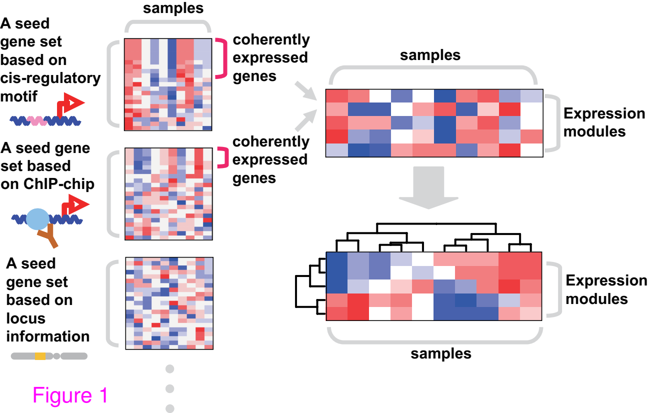
samples
A seed gene set based on cis-regulatory motif
coherently expressed genes
samples
Expression modules
A seed gene set based on ChIP-chip
coherently expressed genes
A seed gene set based on locus information
Expression modules
samples

Figure 1

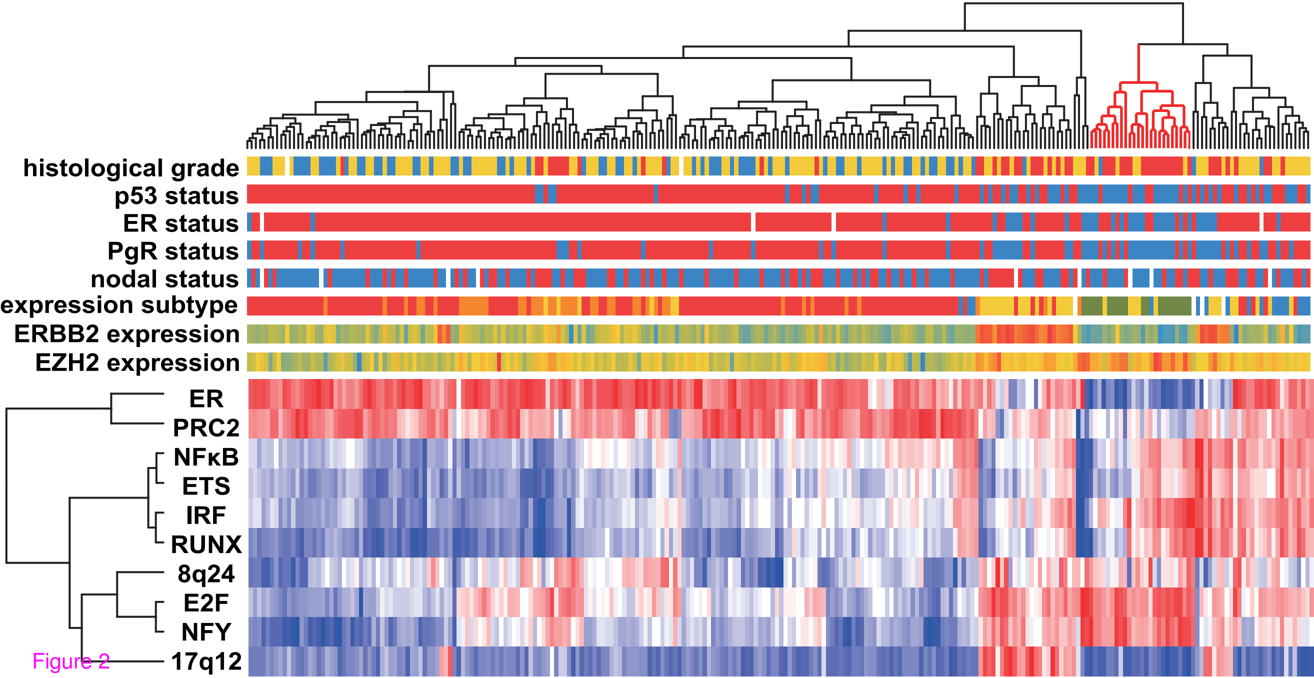
histological grade
p53 status
ER status
PgR status
nodal status
expression subtype
ERBB2 expression
EZH2 expression
ER
PRC2
NFκB
ETS
IRF
RUNX
8q24
E2F
NFY
17q12

Figure 2

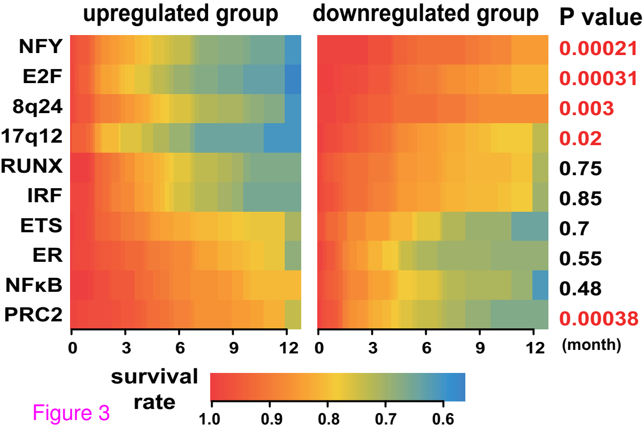
upregulated group
downregulated group
P value
NFY
E2F
8q24
17q12
RUNX
IRF
ETS
ER
NFκB
PRC2
0.00021
0.00031
0.003
0.02
0.75
0.85
0.7
0.55
0.48
0.00038
0
3
6
9
12
0
3
6
9
12
(month)
survival rate
1.0
0.9
0.8
0.7
0.6

Figure 3

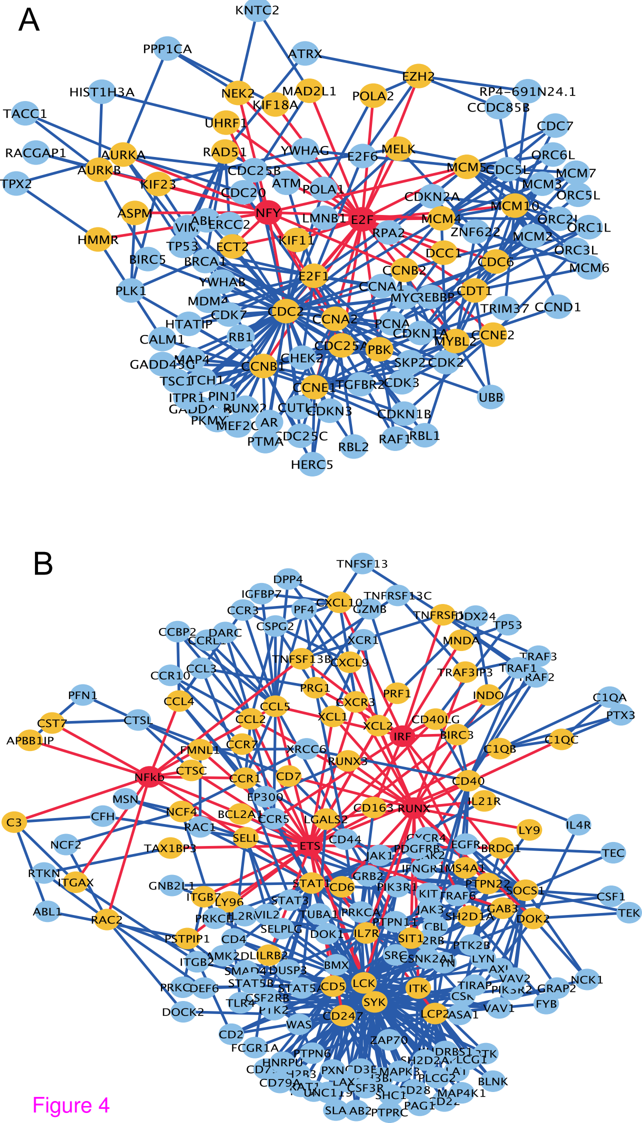
A
KNTC2
PPP1CA
ATRX
EZH2
HIST1H3A
NEK2
MAD2L1
KIF18A
POLA2
RP4-691N24.1
CCDC85B
TACC1
UHRF1
CDC7
RACGAP1
AURKA
RAD51
YWHAG
MELK
E2F6
ORC6L
AURKB
CDC25B
MCM5
CDC5L
MCM7
TPX2
KIF23
ATM
POLA1
MCM3
CDC20
ORC5L
CDKN2A
MCM10
ASPM
NFY
ABL
ERCC2
LMNB1
E2F
MCM4
ORC2L
VIM
ORC1L
HMMR
ZNF622
TP53
RPA2
MCM2
ECT2
KIF11
DCC1
ORC3L
BIRC5
BRCA1
CDC6
MCM6
CCNB2
E2F1
YWHAB
CCNA1
CDT1
PLK1
MYC
CREBBP
MDM4
TRIM37
CCND1
CDK7
CDC2
HTATIP
CCNA2
PCNA
CALM1
RB1
CDKN1A
CCNE2
MYBL2
CDC25A
PBK
GADD45G
MAP4
CHEK2
SKP2
CDK2
CCNB1
TSC1
TCH1
CCNE1
TGFBR2
CDK3
UBB
ITPR1
PIN1
RUNX2
CUTL
CDKN3
CDKN1B
PKMY
MEF2C
AR
RBL1
RAF1
PTMA
CDC25C
RBL2
HERC5
B
TNFSF13
DPP4
IGFBP7
CXCL10
TNFRSF13C
CCR3
PF4
GZMB
TNFRSF
DDX24
TP53
CSPG2
CCBP2
DARC
XCR1
MNDA
CCRL
TRAF3
TNFSF13B
CXCL9
TRAF1
CCL3
TRAF3IP3
TRAF2
CCR10
PFN1
PRG1
PRF1
INDO
C1QA
CCL4
CXCR3
CCL5
PTX3
CST7
XCL1
CTSL
CCL2
CD40LG
XCL2
IRF
APBB1IP
BIRC3
C1QC
CCR7
FMNL1
XRCC6
C1QB
RUNX3
NFkb
CTSC
CCR1
CD7
CD40
EP300
IL21R
MSN
CD163
RUNX
NCF4
BCL2A1
CCR5
LGALS2
C3
CFH
IL4R
LY9
RAC1
NCF2
CXCR4
SELL
ETS
CD44
TAX1BP3
BRDG1
EGFR
PDGFRB
TEC
JAK1
JAK2
RTKN
IFNGR1
MS4A1
ITGAX
GRB2
PTPN22
GNB2L1
STAT1
CD6
PIK3R1
SOCS1
KIT
TRAF6
CSF1
ITGB7
LY96
STAT3
ABL1
JAK3
GAB3
TEK
TUBA1
PRKCA
SH2D1A
DOK2
PRKCI
IL2RVIL2
PTPN11
RAC2
CBL
SELPLG
DOK1
IL7R
CD4
SIT1
PSTPIP1
2RB
PTK2B
AMK2D
LILRB2
SRC
CSNK2A1
ITGB2
LYN
SMAD4
DUSP3
BMX
AXL
YN
STAT5B
VAV2
PRKC
DEF6
CD5
LCK
ITK
TIRAP
NCK1
STAT5A
PIK3R2
GRAP2
TLR4
CSF2RB
SYK
CSK
FYB
VAV1
DOCK2
PTK2
CD247
LCP2
ASA1
CD2
WAS
ZAP70
FCGR1A
PTPN6
DRBS1
TK
HNRPU
SH2D2A
LCG1
CD7
H2B3
PXN
CD3E
MAPK8
LAT
CD79A
PLCG2
BLNK
LAX
CSF3R
CD28
UNC119
MAP4K1
SHC1
PAG1
CD24
SLA
AB2
PTPRC

Figure 4

Additional files provided with this submission:

Additional file 1: texts1.pdf, 866K
http://www.biomedcentral.com/imedia/4848861792501869/supp1.pdf