Discovery of multi-dimensional modules by integrative analysis of cancer genomic data

Shihua Zhang[1,2], Chun-Chi Liu[3], Wenyuan Li[1], Hui Shen[4], Peter W. Laird[4] and Xianghong Jasmine Zhou[1,*]

[1]Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA, [2]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, [3]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taiwan 40227, Republic of China and [4]USC Epigenome Center, University of Southern California, Los Angeles, CA 90033, USA

Received April 14, 2012; Revised June 12, 2012; Accepted July 5, 2012

ABSTRACT

Recent technology has made it possible to simultaneously perform multi-platform genomic profiling (e.g. DNA methylation (DM) and gene expression (GE)) of biological samples, resulting in so-called 'multi-dimensional genomic data'. Such data provide unique opportunities to study the coordination between regulatory mechanisms on multiple levels. However, integrative analysis of multi-dimensional genomics data for the discovery of combinatorial patterns is currently lacking. Here, we adopt a joint matrix factorization technique to address this challenge. This method projects multiple types of genomic data onto a common coordinate system, in which heterogeneous variables weighted highly in the same projected direction form a multi-dimensional module (md-module). Genomic variables in such modules are characterized by significant correlations and likely functional associations. We applied this method to the DM, GE, and microRNA expression data of 385 ovarian cancer samples from the The Cancer Genome Atlas project. These md-modules revealed perturbed pathways that would have been overlooked with only a single type of data, uncovered associations between different layers of cellular activities and allowed the identification of clinically distinct patient subgroups. Our study provides an useful protocol for uncovering hidden patterns and their biological implications in multi-dimensional 'omic' data.

INTRODUCTION

Cells are complex systems with multiple levels of organization that interact and influence each other. The precise coordination among epigenetic status, transcriptions, translations, transportation and metabolic reactions are essential in maintaining the function and robustness of cellular systems. However, study of the coordination among such multilevel cellular activities has been hindered by a lack of appropriate data resources; most genomic research has focused on global profiling at only one level (e.g. profiling of gene expression (GE) or protein abundance).

The recent development of high-throughput genomics technologies, especially sequencing technology, has significantly facilitated the characterization of biological systems at multiple levels. For example, The Cancer Genome Atlas (TCGA) project is generating multi-dimensional maps of the key genomic changes (e.g. SNP, DNA methylation (DM), GE and microRNA [miRNA] expression (ME)) for the same set of tumour samples (1). The NCI60 project has profiled 60 human cancer cell lines in terms of drug responses (2–4), GE (5), protein expression (6) and ME. With the expected drop in sequencing cost, multi-dimensional genomics characterizations on the same set of samples will soon become a standard practice.

Emerging multi-dimensional genomics data pose new challenges for data analysis. In particular, because different types of genomics data have different scales and units, we cannot simply aggregate multiple datasets for analysis. For a specific type of two-dimensional genomics dataset consisting of single-nucleotide polymorphism (SNP) and expression data, various eQTL approaches have been developed to identify regulatory SNPs (7). However, eQTL approaches cannot be applied to datasets with more than

*To whom correspondence should be addressed. Tel: +1 213 740 7055; Fax: +1 213 740 2475; Email: xjzhou@usc.edu

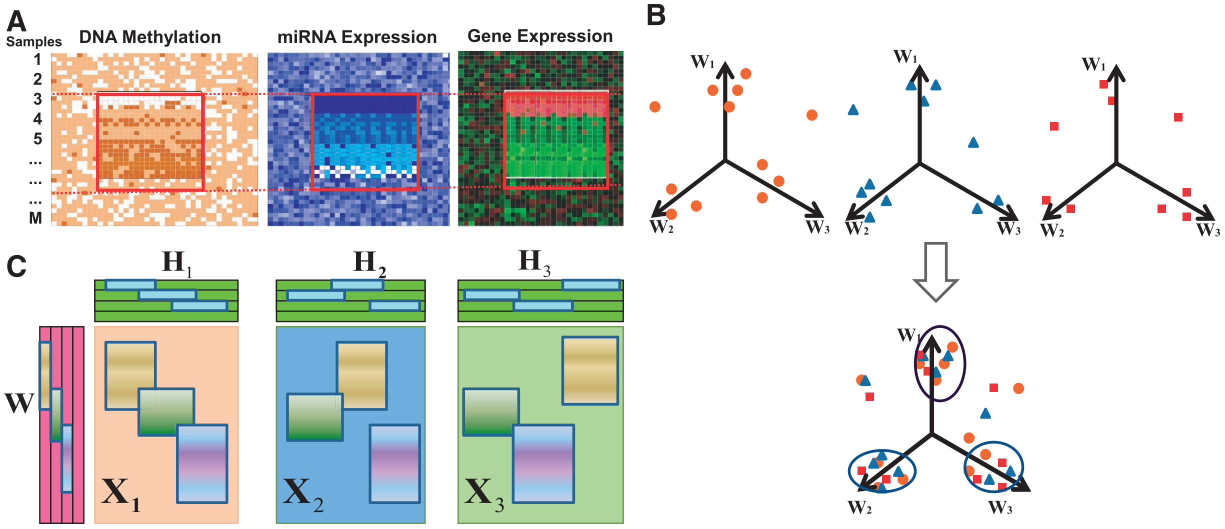

Figure 1. (**A**) An example of md-modules. In the three data matrices, rows correspond to the samples and columns correspond to different measurements. An md-module consists of r rows and n_I (I = 1,2,3) columns for GE, ME and DM data, respectively. These subsets of DMs, MEs and GEs exhibit correlated profiles across a subset of samples. (**B**) Rationale for the joint NMF approach. Input matrices of methylation, miRNA and GE data are projected onto a new common space, where the three correlated patterns containing different types of genomic measurements are uncovered. (**C**) Illustration of joint NMF factorization and the three identified md-modules.

two dimensions, nor can they be used for datasets with a moderate sample size, which include most future multi-dimensional datasets generated by individual laboratories (rather than consortiums). Multivariate regression is another analytical method applicable to two-dimensional genomics datasets to infer correlative relationships (e.g. between GE and transcription factor binding data (8) or between GE and proteomic data (9)). More recently, Kutalik *et al.* (10) proposed a powerful modular analysis approach, called the Ping-Pong algorithm, to uncover the 'co-modules' across GE and drug response data. Undoubtedly, these studies have identified important relationships between pair-wise genomics variables. We believe that the time has come to simultaneously explore the coordination patterns across more than two types of genomics variables.

In this article, we apply a powerful matrix factorization framework to identify correlative modules in multi-dimensional genomics data (Figure 1). As the testing system, we used data from the TCGA project, including DM, ME and GE profiles of 385 ovarian cancer samples. These three types of genomics variables are known to be highly dependent on each other. Our goal was to identify subsets of mRNAs, miRNAs and methylation markers for which all or a subset of the samples exhibit correlated profiles across different types of measurements (Figure 1A). These subsets are termed as '**m**ulti-**d**imensional **modules** (md-modules)'.

By identifying md-modules, we can break down the massive sets of data into smaller building blocks that exhibit similar patterns across certain rows and columns (Figure 1). This procedure provides two major advantages. First, representing coherent features across multiple datasets reduces the complexity of the data and facilitates a global overview of the inherent structure of the data. More importantly, this modular approach captures the associations among sets of different types of variables (mRNA, miRNA and methylation). The md-modules can identify vertical associations between multiple regulatory levels and can reveal significantly disrupted pathways that would be ignored if only data of the single dimensions were used. In addition, the md-modules can stratify patients (samples) into clinically distinct groups, which facilitate the identification of the complex molecular mechanisms that underlie different clinical phenotypes.

MATERIALS AND METHODS

Data preparation and preprocessing

The TCGA data were downloaded from the TCGA Data Portal on 27 April 2009. We used three types of data, as follows: GE data (Agilent G4502A), DM data (Illumina 27K) and ME data (Agilent H-miRNA_8x15K v2). In total, 385 samples are shared by the three datasets. We normalized the columns of the expression matrices, and then we scaled all the matrices so that sum of squares of each matrix is the same.

To make the input data fit the constraints of non-negativity, we used the method suggested by Kine and Tidor (11). We doubled the columns of each matrix, so that each variable (gene, miRNA) was represented with two columns in the respective matrix. If the original value of the variable was positive, then it was stored in the first column; otherwise, its absolute value was stored in the second column. The rest of matrices were filled with zeros.

Brief overview

Non-negative matrix factorization (NMF) is increasingly being used to analyse high-dimensional genomics data (11,12). NMF factorizes a matrix $X_{M \times N}$ into two non-negative matrices $X = WH$, where W is an M by K

matrix containing the basis vectors, and H is a K by N matrix containing the coefficient vectors. Each element in W and H must be ≥ 0. Thus, a key feature of NMF is the ability to identify nonsubtractive patterns that together explain the data as a linear combination of its basis vectors. The K basis vectors in W can be regarded as the 'building blocks' of the data, and the K coefficient vectors describe how strongly each 'building block' is present in the data.

Recently, an NMF-type method has been proposed to analyse pair-wise genomics data (13,14), including GE and transcription factor-binding data (13). We have developed a semi-supervised framework for combing miRNA/genes expression profiles and networked data to extract miRNA–gene regulatory programs (15). Here, we adopt the powerful NMF-type method for the discovery of md-modules by integrative analysis of cancer genomic data, all profiled on the same samples. We introduce the idea using a three-dimensional dataset, but it is applicable to higher dimensional datasets.

The NMF problem

Given a dataset consisting of N measurements of M non-negative scalar variables, we let the M-dimensional measurement vectors $x_{.j} (j = 1, \ldots, N)$ form the data matrix $X_{M \times N}$. For each column $x_{.j}$, a linear, non-negative approximation of the data is given by

$$x_{.j} = \sum_{k=1}^{K} w_{.k} h_{kj} = W h_{.j} \quad \text{or} \quad X = WH,$$

where W is an $M \times K$ matrix containing the basis vectors $w_{.k}$ as its columns and H is an $K \times N$ matrix containing the coefficient vector $h_{.j}$ corresponding to the measurement vector $x_{.j}$. Note that each measurement vector is written in terms of the same basis vectors. The K basis vectors $w_{.k}$ can be thought of as the 'building blocks' of the data, and the K-dimensional coefficient vector $h_{.j}$ describes how strongly each building block is present in the measurement vector $x_{.j}$.

Given a non-negative data matrix X, the optimal choices of matrices W and H are defined to be those non-negative matrices that minimize the reconstruction error between X and WH. Although several error functions have been proposed (16–18), the most widely used is the squared Euclidean error function:

$$F(W, H) = \|X - WH\|_F^2.$$

The resulting WH is called the non-negative matrix factorization of X. The choice of K is often problem-dependent. In most cases, K is chosen such that $K < \min(M, N)$ and WH represents a compressed form of the data in X. By not allowing negative entries in W and H, NMF enables a non-subtractive combination of parts to form a whole (17).

The joint NMF framework for integrative analysis

Let X_1, X_2 and X_3 be $M \times N_1$, $M \times N_2$ and $M \times N_3$ matrices representing three types of genomic profiling of the same samples, e.g. the methylation profiles of N_1 DNA markers and the expressions of N_2 genes and N_3 miRNAs of M samples. To extract md-modules across the three data matrices, the following joint factorization framework was used to decompose the three data matrices into a common basis matrix W and different coefficient matrices H_I $(I = 1, 2, 3)$:

$$X_I \approx WH_I,$$

with the non-negativity constraints:

$$W \geq 0, \quad H_I \geq 0, \quad I = 1, 2, 3,$$

where W is an $M \times K$ matrix, and each column of W represents a basis vector of the reduced system. H_I is a matrix of size $K \times N_I$, and each row of H_I represents a coefficient vector. Then, the objective function of joint NMF is formulated as

$$\min \sum_{I=1}^{3} \|X_I - WH_I\|_F^2.$$

Several algorithms have been developed to optimize the NMF problem (19). Lee and Seung (18) devised a multiplicative algorithm that is simple to implement and performs well. Like the standard NMF, we used the 'multiplicative update' equations to minimize the Euclidean error function. Specifically, given a desired rank K, the algorithm iteratively computes the approximations of X_1, X_2 and X_3 in the same manner. The method starts by randomly initializing matrices W and H_1, H_2 and H_3, which are iteratively updated to minimize the Euclidean distance function. Specifically, W, H_1, H_2 and H_3 are updated at each step by using the generalized multiplicative update rules as follows:

$$W_{ia} = W_{ia} \frac{(X_1 H_1^T + X_2 H_2^T + X_3 H_3^T)_{ia}}{(W(H_1 H_1^T + H_2 H_2^T + H_3 H_3^T))_{ia}},$$

$$(H_I)_{a\mu} = (H_I)_{a\mu} \frac{(W^T X_I)_{a\mu}}{(W^T W H_I)_{a\mu}}, \quad I = 1, 2, 3.$$

The above algorithm is a local optimization procedure, and thus, found only a local minimum. To address this limitation, we repeated the procedure for 50 times with different initial solution matrices. The factorization that leads to the lowest objective function value was used as the final solution for further analysis. The solutions found were reproducible, since that of different runs of the repeated algorithm showed strong correlations. The time complexity of the joint NMF decomposition is $O(tK(M + N_1 + N_2 + N_3)^2)$ which is similar to that of the original NMF model, where t is the number of iterations. The key to use this procedure is the computer memory. Generally, if we have enough memory space, it shall be applicable to even millions of features. If we do not have enough memory space, we can consider reducing the dimension of input data by data-reduction techniques such as the PCA-select tool used for decreasing the feature number in population structure studies (20).

In this way, the three data matrices are projected into a common coordinate system to explore the correlative relationships among the three types of variables (Figure 1B and C). Using this procedure, we obtained coefficient matrices H_1, H_2 and H_3 that can be used to identify memberships of DM markers, miRNAs and genes in md-modules, respectively. In the general application of NMF (11,12), researchers have used the maximum of each column of H (or row of W) to determine membership. In this way, each gene (or other object) can belong to one and only one module. However, some markers/miRNAs/genes may not be active in any module or may be active in multiple modules with multiple functions. Considering these facts, based on H_1, H_2 and H_3, we calculated the z-score for each element in each row of H by

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i},$$

where μ_i is the average value for feature j (DM markers/miRNA/gene) in H_I ($I = 1,2,3$) and σ_i is the standard deviation. We assigned feature j as a member of module k , if z_{ij} was greater than a given threshold T. Each DM marker/miRNA/gene may be assigned to md-modules, which allows the identification of multiple functional activities of DM markers/miRNAs/genes. We have implemented the method as a Matlab software package, which is available from the Supplementary Data. Mathematically, the multi-dimensional data of same samples are modeled using multiple matrices that share the same rows. Therefore, the technique cannot be applied to different types of data from different samples.

Statistical significance of vertical correlations in md-modules

We expect that, within an md-module, the profiles of genes, DM markers and miRNAs are highly (anti-) correlated. To determine whether such relationships are statistically significant, we performed the following assessment. We calculated the 'between-correlation' between two matrices with the same row dimensions as the sum of the absolute values of Pearson's correlations between any two columns (one column from each matrix). We derived the statistical significance (P-value) of the correlation between two matrices by comparing it with the distribution of between-correlations between 1000 random matrix pairs. Each pair is composed of two matrices with dimensions identical to the original ones, whose elements are extracted from randomly permuted matrices based on the original ones. P-values of $< 0.05/200$ were considered significant. For an md-module, if all three P-values for the pair-wise submatrices are significant, then the vertical correlation of this module is considered to be statistically significant.

Functional analysis of identified md-modules

For each md-module, we identified three gene sets, as follows: (i) genes from the GE dimension; (ii) genes in the 20-kb region around the methylation markers in the DM dimension and (iii) genes targeted by miRNAs in the ME dimension (based on the miRNA targets from the Microcosm database). For each gene set, we performed two types of enrichment analyses: gene ontology (GO) biological process and KEGG pathway analyses.

Cancer gene and protein interactions enrichment analysis

The protein–protein interaction network data were downloaded from BioGRID (release 2.0.54). The final network has 7682 proteins and 33 165 interactions. The cancer gene list was obtained from the Cancer Gene Census (CGC) web site (21). All cancer genes that are not included in our input gene list were excluded. The final list contains 290 cancer genes. We also collected an epigenetically regulated gene list of ovarian cancer, which includes 40 genes (22). All of the enrichment analyses for a gene set are assessed by the right-tailed Fisher's exact test.

'Vertical' implications of identified regulatory md-modules

For each md-module, we investigated the vertical associations between different dimensions by the following 'overlapping analysis:' we first identified overlapping genes between those from the GE dimension and those adjacent to methylation markers in the DM dimension, or between those from the GE dimension and those targeted by miRNAs in the ME dimension, and then performed the enrichment significance assessment.

Clinical characterization

Based on the signals for all samples in each column of the common basis matrix W, we can characterize their level of association with the discovered md-modules. For each md-module, we divided the set of samples into two groups: module-specific and not module-specific, by using the z-score for each column of W with a threshold of 1. The clinical data were downloaded from the TCGA portal. Kaplan–Meier curves were computed by using R. Survival distributions between groups were computed through the log-rank test. Age differences between groups were compared with the Wilcoxon signed-rank test.

RESULTS

Figure 2 illustrates an example using simulated data (see the Supplementary Data). In a matrix representation, a md-module consists of r rows and n_I ($I = 1,2,3$) columns for mRNA, miRNA and methylation markers, respectively. Within these r rows (samples) in each matrix, the n_I ($I = 1,2,3$) columns exhibit correlated measurements (Figure 2). In biological applications, permutation tests are performed to evaluate the statistical significance of each md-module according to the 'between' correlations of different types of variables. Details and parameter selections are described in the 'Materials and Methods' section and in the Supplementary Data.

Before describing the application of this method, we briefly show how the md-module discovery is related to, but different from, several typical data mining tasks. Most existing techniques for module identification were applicable only to one or two matrices at a time.

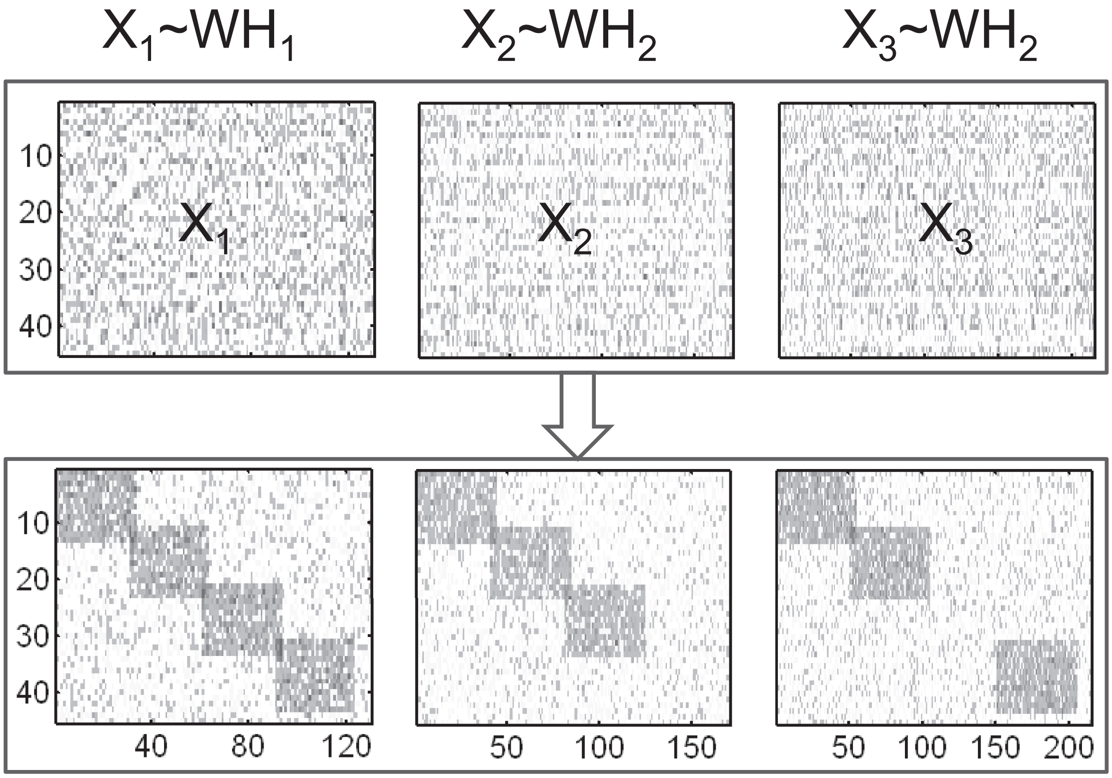

Figure 2. Illustration of the patterns (md-modules) identified by the adopted method. A simulated dataset with the same number of samples (rows) and different number of features (columns) was generated. The joint NMF method can accurately discover the patterns embedded in these data. A pattern may involve as many as all three datasets simultaneously or only cover two datasets. These different patterns may share the same samples (overlap) or/and the same features.

For example, the goal of clustering methods is to identify a group of relevant rows or columns in a data matrix. A more related task 'biclustering (co-clustering)' refers to a class of clustering techniques that perform simultaneous clustering of rows and columns in a data matrix (23). More recently, Kutalik *et al.* (10) extended the traditional modular analysis approach from one to two data matrices that share one common dimension, and applied their method to identifying drug–gene co-modules. We should note that this method is not directly applicable to three matrices. Shen *et al.* (24) have proposed a joint clustering model for multiple genomic datasets, but it was designed for sample clustering and subtype discovery and cannot identify modules comprising of correlated variables. We have previously proposed a NMF-type method to analyze paired matrices subjected to network constraints (15). However, it has not been applied to more than two data matrices and tested for md-modules analysis.

Identification of md-modules involved in ovarian cancer

The TCGA ovarian cancer dataset consisting of GE, DM and ME profiles across 385 samples (patients) was used as a testing system to show the discovery of md-modules. After parameter optimizations (details in Materials and Methods), the three large matrices were broken down into $K = 200$ basic building blocks, from which 200 md-modules were derived. The dimension reduction captures the major information embedded in the original data; the average sample-wise correlations of the reconstructed data using these building blocks (based on W and H_I) and the original data were 0.90, 0.92 and 0.91 in the methylation, miRNA and GE dimensions, respectively. The small variances of those correlations further demonstrate the robustness of the method (Figure 3A). The correlated profiles for the three samples are plotted in Figure 3B.

Each of the 200 md-modules comprises a set of genes, methylation markers and miRNAs. In total, the 200 md-modules cover 2985 genes, 2008 DM markers and 270 miRNAs. The average module sizes in the gene, methylation markers and miRNA dimensions are 239.6, 162.3 and 13.8, respectively. Size distribution and other characteristics of these modules are described in the Supplementary Data.

Md-modules reveal multilevel vertical associations and cooperative functional effects

To assess the biological relevance of the identified multi-dimensional modules, we first tested the functional homogeneity of members within individual dimensions. A set of genes is defined to be functionally homogenous if it is enriched in at least one GO biological process category (25), with a q-value of < 0.05 (the q-value is the P-value after a false-discovery rate multiple testing correction). Among the 200 md-modules, 80, 62.7, and 12.5% were functionally homogenous in the GE dimension with respect to member genes, in the DM dimension with

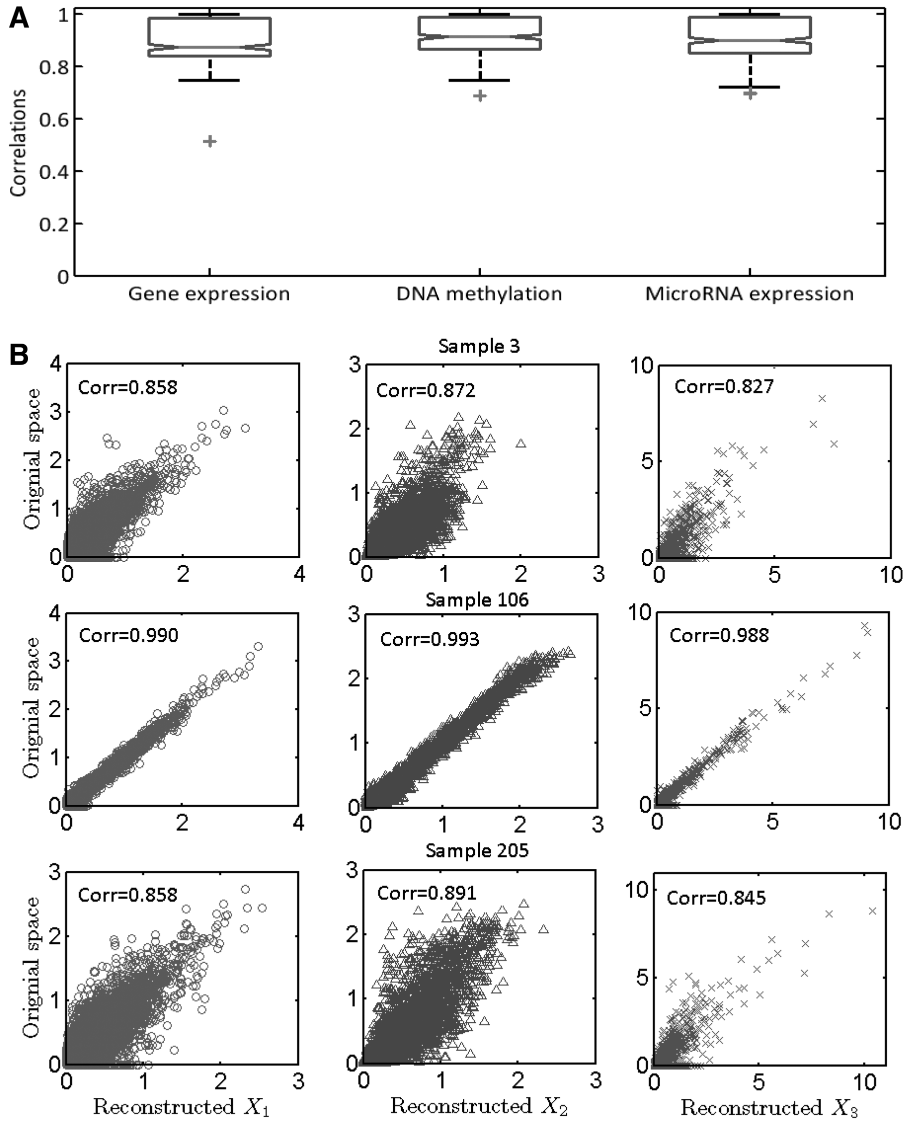

Figure 3. (**A**) Box-plot of sample-wise correlations of original and reconstructed methylation, miRNA and GE profiles across 385 samples. (**B**) Original data are plotted against the reconstructed methylation, miRNA and GE profiles for three samples.

respect to genes directly adjacent to the member DNA methylation markers and in ME dimension with respect to member miRNAs, respectively. The functions of the miRNAs were predicted based on the functions of the target genes. These values are significantly higher than those obtained after randomization (5, 13.1 and 3.9% for GE, DM and ME, respectively) (Figure 4A).

Although all three dimensions showed significant enrichment in developmental processes that are known to be tightly associated with cancer pathogenesis, this preference is most obvious in the DM dimension, with additional strong participation in embryonic development. This result is consistent with the previous report that polycomb complex targets in the embryonic stem cell are predisposed to cancer-specific hypermethylation (26). The most frequently activated biological processes in the GE dimension are responses to external stimuli (e.g. chemotaxis, locomotor behavior and inflammatory responses). This observation points to the flexibility of GE programs upon external perturbations. The ME dimension shows a distinct preference for participation in transcriptional regulation (as expected) and cell differentiation.

Although the individual dimensions of these modules exhibit a significant level of functional homogeneity, combining all dimensions reveals an even stronger functional synergy. When the GE dimension genes, methylation adjacent genes and miRNAs of a module were combined, 93% of the md-modules were functionally homogenous, compared with only 7.9% after randomization (Figure 4A). This result shows the power of current integrative analysis of muilti-dimensional data in identifying genomic variables of different natures that are involved in the same functional pathways.

The ability of the modules to capture multilevel synchronicity was also observed relative to perturbed

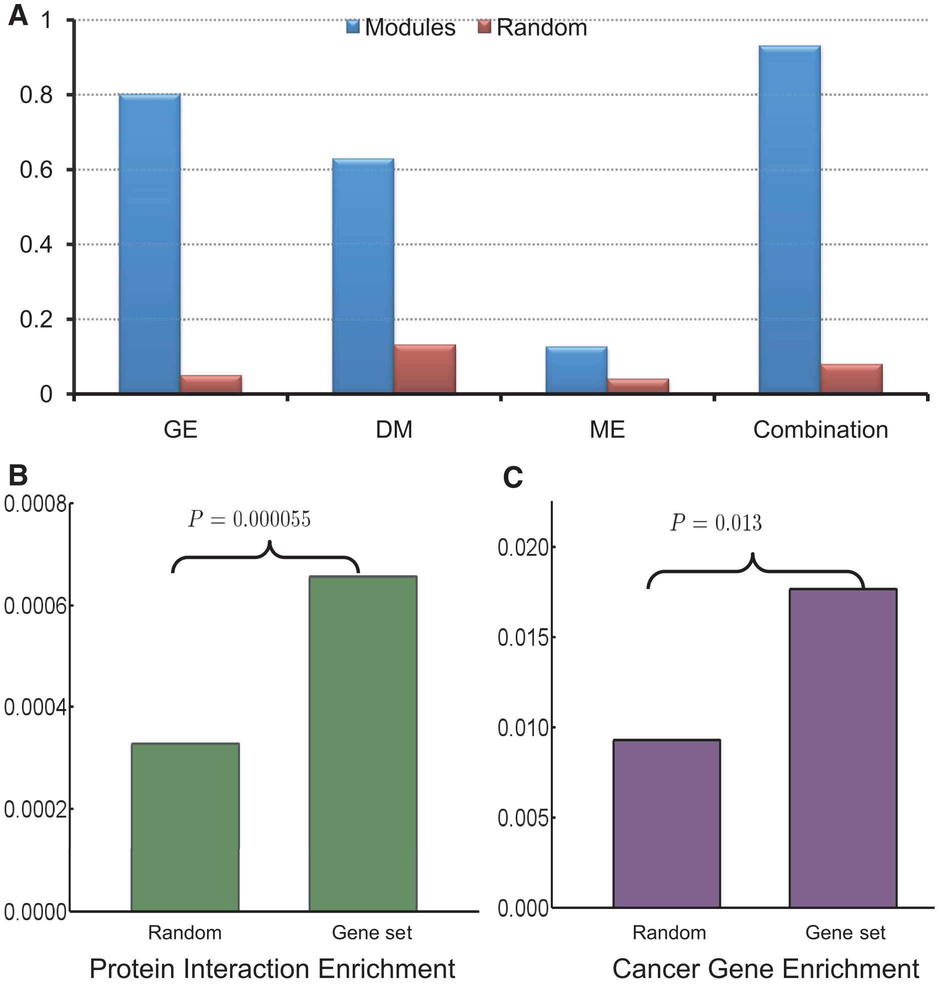

Figure 4. (**A**) Enrichment ratio of md-modules in each dimension (GE, DM and ME), with respect to the GO biological process terms. For comparison, the mean ratio of functional enrichment for 100 corresponding random runs is also plotted. (**B**) and (**C**) Examples of protein interaction enrichment and cancer gene enrichment, which were calculated for md-modules 173. The *P*-values were determined by right-tailed Fisher's exact test.

KEGG pathways. For example, simply by combining multiple dimensions, we observed that nine modules showed significant perturbations in at least one KEGG pathway (*P*-value < 0.05) that were not shown otherwise. These pathways include TGF-β signaling, Hedgehog signaling, bladder cancer and cytokine–cytokine receptor interaction pathways, all of which have been confirmed to be closely associated with ovarian cancer (27–32). For 11.5% of the md-modules, the pathway enrichment for combined members from all three dimensions are more significant than that for any individual dimension.

According to the model principle, a md-module should capture vertical associations, i.e. associations between variables of different dimensions (e.g. GE and DM) in it. Indeed, compared with randomly permuted modules, the Pearson's correlation coefficients between variables of any two of the GE, DM and ME dimensions are significantly high (*P*-value $< 0.05/200$) in 65.5% of the modules (for details see 'Materials and Methods' section). This result indicates that the probability of identifying these modules by chance is close to zero. The strong statistical correlations across different dimensions imply the coordinated activities of genes, methylations and miRNAs.

To explore further the biological implications of these vertical correlations, we tested whether genes in an md-module were likely to be located close to the methylation markers in the same module or/and targeted by miRNAs in the same module. At a significance level of 0.1, we found that 75 of the 200 md-modules showed significant overlap between genes adjacent to methylation markers and genes within the same module. This result confirms the strong influences of DNA methylation on the expression of adjacent genes. Likewise, 146 modules with *P*-value < 0.1 show significant overlap between genes targeted by miRNAs and genes within the same md-module. As the targeting relationship between miRNAs and genes is far from complete, our overlap assessment can only serve as an underestimate. These data show that the md-modules can elucidate the vertical association mechanisms between different layers of gene regulation. Table 1 showcases 12 of the md-modules, including the overlap between different dimensions within the same modules, and the over-represented functions and pathways of the modules.

Table 1. Summary of the 12 md-modules detected by the joint NMF method in TCGA Ovarian data

No.	Ge	Me(Ge)	Mi(Ge)	Oa	Ob	Selected overrepresented functional sets
34	248	195(174)	11(715)	6**	12*	Embryonic morphogenesis; glutamate signaling pathway; growth factor activity
48	243	209(171)	18(437)	10****	11**	Pattern specification process; embryonic morphogenesis
67	220	189(153)	12(1561)	6**	19*	Endopeptidase inhibitor activity; G-protein-coupled receptor binding; cell communication
68	297	179(170)	9(320)	8***	9**	Embryonic morphogenesis; positive regulation of transport; regulation of cytokine secretion
71	215	207(171)	18(1946)	7***	23*	Homophilic cell adhesion; cell–cell adhesion; calcium-dependent cell–cell adhesion
81	195	77(62)	16(1261)	4***	16**	Cell–cell adhesion
112	239	217(201)	16(697)	6**	12*	Cytokine activity; inflammatory response
116	235	217(175)	16(545)	5*	10*	Keratinization; calcium-dependent cell–cell adhesion; homophilic cell adhesion
123	217	216(176)	16(459)	6**	8*	Proteinaceous extracellular matrix; embryonic morphogenesis; homophilic cell adhesion
154	238	192(162)	15(1030)	6**	17*	Cytokine activity; heparin binding; inflammatory response; tumor necrosis factor receptor binding
169	204	245(218)	15(1065)	8***	14*	Organ morphogenesis; regulation of leukocyte chemotaxis; reproductive structure development
193	200	178(146)	21(809)	4*	15**	Homophilic cell adhesion; calcium-dependent cell adhesion; embryonic morphogenesis

No.: the index of the md-module. Ge: number of genes in GE dimension. Me(Ge): number of DM markers and their adjacent genes. Mi(Ge): number of miRNAs and their targeting genes. Oa: overlap between gene set and DM markers adjacent gene set; Ob: overlap between gene set and miR target gene set. Where *(0.05–0.1), **(0.01–0.05), ***(1.0e-03–0.01) and ****(0–1.0e-03) represent the *P*-value ranges for the hypergeometic test, respectively.

Interestingly, among the 3733 genes overlapping at least two dimensions from all md-modules, genes related to ovarian cancer are significantly enriched (*P*-value = 0.000087). Note that the overlapping genes are those on which regulatory perturbations were observed at multiple levels. It is not surprising that those genes are especially concentrated in the biological processes of ‘positive regulation of developmental processes,’ ‘positive regulation of cell differentiation,’ ‘inflammatory response’ and ‘regulation of cell development.’ Md-module 173 contained six, nine and nine genes overlapping the GE and DM, GE and ME and DM and ME dimensions, respectively. Among these genes, *NID2* (Nidogen-2) was overlapped by all three dimensions. *NID2* recently was defined as a new biomarker for ovarian cancer by comparing its concentration in the serum of healthy women with that in women with ovarian carcinoma (33). More interestingly, *NID2* gene promoters are aberrantly methylated in human gastrointestinal cancer (34), and methylated *NID2* has been defined as a marker for primary bladder cancer (35).

In 44 modules, the genes from the GE and DM dimensions are enriched in protein–protein interactions (Figure 4B) (*P*-value < 0.05; we skipped the ME dimension, due to the large number of potential miRNA targets). Among these 44 modules, 18 are enriched in protein–protein interactions bridging the GE and DM dimensions (i.e. one protein belongs to the GE dimension and another belongs to the DM dimension) (*P*-value < 0.05 with right-tailed Fisher’s exact test). This finding highlights the different regulatory effects on closely adjacent molecules of the same pathway.

Finally, we hypothesized that the identified md-modules might play a role in cancer. Indeed, 22 combined sets of genes (from the GE and DM dimensions) are enriched with the cancer gene reference set (*P*-value < 0.05 with right-tailed Fisher’s exact test) (Figure 4C) (i.e. the CGC list (21)). The results of the large-scale enrichment analysis support the biological relevance of the regulatory programs detected by our method.

Md-modules capture multilevel synchronized disruptions on pathways: two case studies

This section provides in-depth descriptions of two case studies (modules 119 and 5) to demonstrate how multilevel regulatory changes cooperatively perturb pathways.

Md-module 119. The individual dimensions of module 119 do not show significant enrichment in any KEGG pathway. However, when all three dimensions were considered, the bladder cancer pathway emerged as a significantly disrupted pathway. This pathway, which is frequently altered in bladder cancer, shares a set of known oncogenes and tumor suppressors with many other cancers (e.g. prostate, ovarian and lung cancers). Module 119 overlaps with the bladder cancer pathway in three genes in the GE dimension (*MMP1, MYC* and *RB1*), three genes adjacent to markers in the DM dimension (*CDKN2A, RASSF1* and *TYMP*), and four miRNAs in the ME dimension (*mir-130b, mir-149, mir-196b* and *mir-218*). Figure 5A provides a snapshot of perturbation positions for some of these molecules along the pathway. Promoter hypermethylations of two identified tumor suppressor genes, *CDKN2A* and *RASSF1*, are thought to be involved in the development and progression of ovarian cancer (22). The DNA methylation markers adjacent to *RASSF1* and *CDKN2A* were negatively correlated with the expression of these two genes (*P*-value < 0.0001). Figure 5A shows that *RASSF1* could be additionally inhibited by the increased expression of its predicted post-transcriptional regulator *mir-130* in ovarian tumors, compared with normal tissues. *RASSF1* encodes a protein similar to the effector proteins of the oncogene *HRAS*. Thus, promoter hypermethylation will silence *RASSF1*, thereby upregulating the activity of *HRAS*. The effector of *HRAS* on the pathway, *ARAF*, is also a potential target of *mir-218* and, thus, would also be activated.

The next two neighbors on the pathway, *MAPK1* and *RPS6KA5*, are potentially targeted by another onco-miRNA *mir-130b*, whose elevation is known to be associated with a variety of cancers (36–38). *Mir-130b* is predicted to target several downstream molecules in this

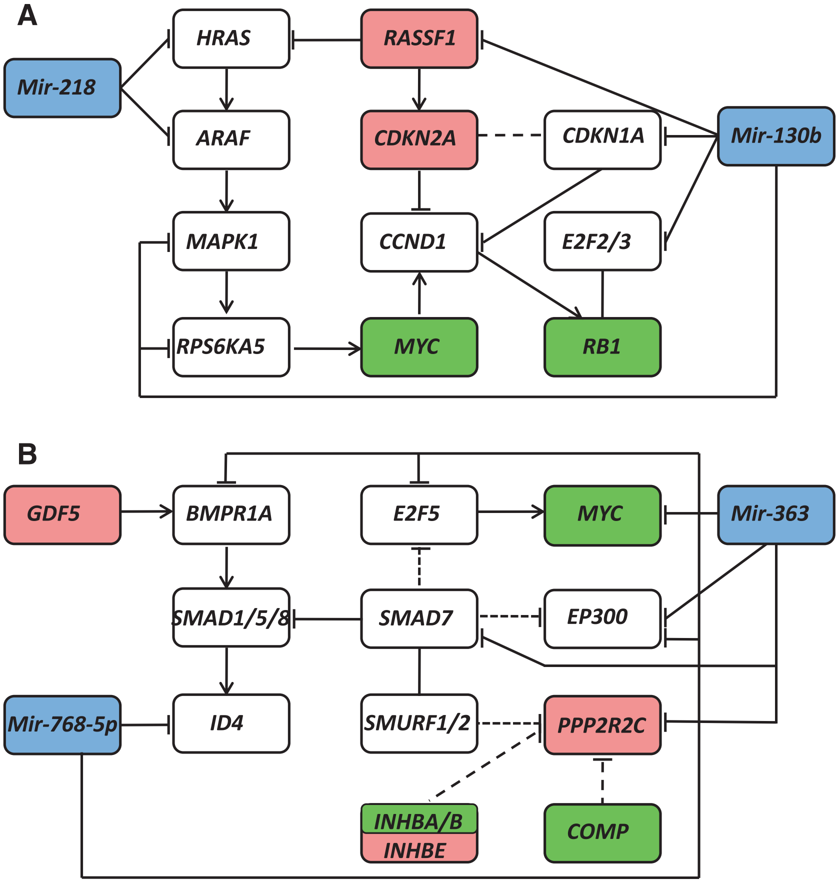

Figure 5. Multilevel factors cooperatively perturb pathways. (**A**) Bladder cancer pathway and (**B**) TGF-β signaling pathway, which are enriched in the combination of molecules in all three dimensions, but not in each dimension. In both subfigures, molecules in this module participating in the corresponding pathways include those from the gene expression dimension (in green), DNA methylation dimension (red), miRNA expression dimension (blue) and miRNA targets (white).

pathway, including *CDKN1A* and *E2F2/3*, both of which are reported to be critically involved in the pathogenesis of ovarian cancer (39,40). In fact, the multiple potential targets of *mir-130b* in the bladder cancer pathway suggest that *mir-130b* could be a key regulatory factor of this dysfunctional pathway in ovarian cancer. The GE dimension of our module includes two important genes on this pathway, the oncogene *MYC* and the tumor suppressor *RB1*. An interesting gene, *CCND1*, connects *MYC, RB1* and another tumor suppressor gene, *CDKN1A*, in this pathway. Mutations, amplification or overexpression of *CCND1*, which alter the cell cycle progression, are observed frequently in a variety of tumors (41,42). Thus, *CCND1* may be an important contributor to tumorigenesis. This example clearly shows that the md-modules capture the associations among epigenetic regulation, gene expression and post-transcriptional regulation on various parts of the pathway. Such synchronized effects from multiple regulatory levels are otherwise difficult to identify.

Md-module 5. As another example, md-module 5 captures the significant dysfunction of the TGF-β signaling pathway in ovarian cancer, which, again, only become obvious by combining perturbations in all three dimensions. Genes in this module that participate in the TGF-β pathway include *INHBA, INHBB, COMP* and *MYC* in the GE dimension, *PPP2R2C, INHBE* and *GDF5* adjacent to markers in the DM dimension, and *mir-363, mir-768-5p* and *mir-451* in the ME dimension (Figure 5B shows a snapshot of perturbation positions for some of these molecules among the pathway). The TGF-β signaling pathway normally exerts anticancer activities by arresting the G1–S transition. However, its abnormal function reverts to promote tumorigenesis, especially in terms of metastatic progression, a functional switch known as the 'TGF-β paradox' (43). In fact, in this module, 60% of tumors with characterized recurrences sites have metastasized.

The 'core' metastasis-associated gene expression signature is manifested in this module, mainly through the increased expressions of *COMP* and *INHBA* (44). This finding further confirms the strong metastasis characteristics of samples in the module. Interestingly, *mir-363, mir-768-5p* and *mir-451* all potentially target *EP300*, a metastasis suppressor whose decreased expression and protein abundance have been detected in many highly metastatic cancer tissues (45). Another tumor suppressor, *PPP2R2C*, not only appeared in the methylation dimension of the module, but also may be a potential target of

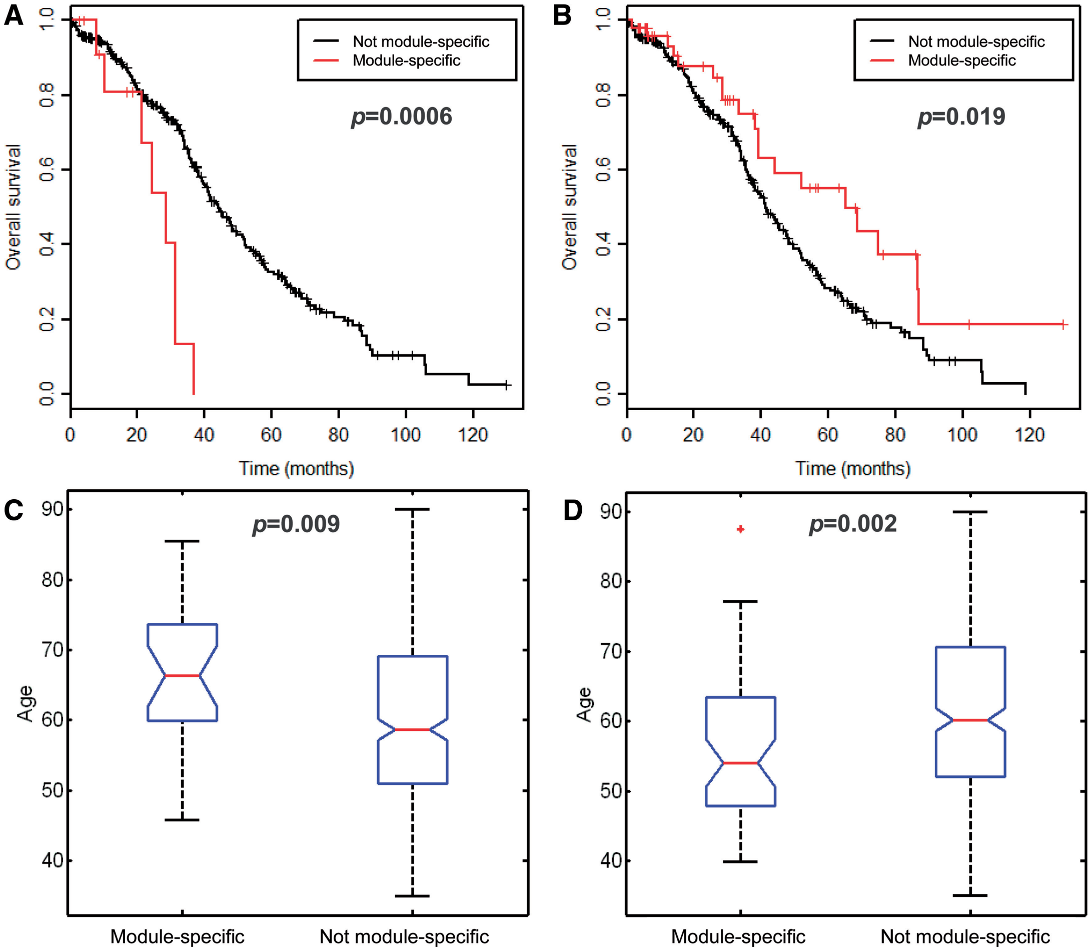

Figure 6. (**A**) and (**B**) Kaplan–Meier survival analysis for patients associated with module 166 (A) or module 3 (B) compared with other patients. The *P*-values of the log-rank test were *P* = 0.0006 and *P* = 0.019, respectively. Median survivals for patients in module 166 or module 3 compared with other patients were 26.4 versus 36.1 years and 38.2 versus 33.8 years, respectively. (**C**) and (**D**) Box-plot for the ages of patients associated with module 28 (C) or module 78 (D) compared with other patients. The *P*-values of the rank-sum test were *P* = 0.009 and *P* = 0.002, respectively. Median ages for patients in module 28 or module 78 compared with other patients were 66.3 versus 58.7 years and 54.1 versus 60.2 years, respectively.

mir-363. In addition, *mir-363* targets a set of *SMAD* molecules, which play important roles in the metastasis transition contributed by TGF-β (46–48). Furthermore, *mir-768-5p* is predicted to inhibit *E2F5* and *BMPR1A*, both of which support the original anticancer activities of TGF- β pathway (50,49).

The TGF-β signaling pathway has been regarded as a potential therapeutic target in ovarian cancer metastases (27). More interestingly, a recent study suggests that the accumulation of epigenetic modifications, including DNA methylation, leads to the suppression of TGF-β signaling and contributes to ovarian carcinogenesis (51). Our md-module facilitates the discovery of the abnormal functions of this pathway at multiple regulatory levels. Thus, this method can aid a holistic approach to drug interventions that can simultaneously correct the effects of various types of dysfunctions.

Clinical associations of the md-modules

In the NMF framework, the decomposed component vector (i.e. column of the *W* matrix) can provide information on the association of each sample/patient with an individual module. This information, combined with the available clinical characterizations of each patient, can aid in the discovery of phenotype-specific md-modules. An md-module that stratifies patients into clinically distinct groups can shed light on the molecular mechanisms of the respective clinical phenotypes.

Based on the information from the *W* matrix, we compared the survival time of ovarian cancer patients that are strongly associated with a specific md-module versus those that are not. We found patients in several md-modules who showed significantly shorter or longer median survival time (log-rank test $P < 0.05$; Supplementary Data). For example, 13 patients are strongly associated with md-module 166. They show significantly worse outcome, with a median survival of 26.4 months compared with 34.1 months for other patients (P = 0.0006, log-rank test) (Figure 6A). In fact, in all three dimensions of this md-module, these 13 patients show distinct characteristics compared with the rest of the patients. For example, genes/miRNAs in this module are over/under-expressed in these 13 samples compared with other samples, as are the methylation levels of the markers. The module contains numerous cell cycle check-point genes (e.g. *BUB1B*, *CENPF, MAD2L1, CCNB1, BUB1, CCNA2, CHEK1* and *TTK*) and is significantly enriched in genes from the 'nuclear division' functional category (P-value $< 10^{-8}$). In another case, the patients in md-module 3 are associated with an improved survival, with a median survival of 38.2 months versus 33.8 months in the remaining patients ($P < 0.02$, log-rank test). This module reveals the significant perturbation of the endometrial cancer pathway with several key genes related to tumorigenesis, e.g. *EGFR*, *CTNNA2* and *ARAF*.

We identified 20 md-modules, each of which contains patients with significantly different age characteristics from patients outside the module. For example, patients in module 28 had an older median age compared with other patients (66.3 years versus 58.7 years; $P = 0.009$, rank-sum test) (Figure 6C), and md-module 78 was associated with significantly younger patients (median age of 54.1 years versus 60.2 years for the rest of patients) ($P = 0.002$, rank-sum test) (Figure 6D).

Finally, in addition to tumor samples, our study samples include eight normal fallopian tube samples. Md-module 120 contains six samples, all of which are normal fallopian tube samples (enrichment $P = 6.4 \times 10^{-12}$ based on Fisher's exact test). This is an extreme example demonstrating that our modules can distinguish phenotypically distinct patient groups. A number of miRNAs, e.g. *mir-143, mir-145, mir-224* and *mir-424*, are reported to be down-regulated in ovarian carcinoma cells 53–54). Not surprisingly, all of them show high expression values in this module containing only normal samples.

DISCUSSION

Recent technology has enabled the simultaneous multi-platform genomic profiling of biological samples, resulting in so-called multi-dimensional genomic data. With the rapid decline of sequencing costs, such data will soon accumulate rapidly. However, systematic analysis of such multi-dimensional data for discovering biologically relevant combinatorial patterns are currently lacking. A great number of tools designed for one- or, at most, two-dimensional data have been developed, and many of which have been applied for genomic data analysis in the past. In this article, we attempted to adopt powerful data analysis technique to address the sophisticated modular structures embedded in multi-dimensional genomics data. We proposed the novel concept of md-modules.

Using the TCGA ovarian cancer dataset comprising gene expression, DNA methylation and miRNA expression in 385 samples, we showed that md-modules provide several unique insights. (i) By considering several different aspects of genomic modulation, md-modules can reveal perturbed pathways that would be overlooked with only a single type of data. (ii) An md-module identifies associations between different layers of cellular activity (e.g. DNA methylation, gene or miRNA expression), even if these associations exist only in a subgroup of samples. (iii) An md-module can identify clinically distinct patient (sample) subgroups that share subsets of multi-dimensional genomic features (methylations, gene expressions, etc). Cancer in particular is characterized by the existence of many subtypes with heterogeneous genetic origins, and one type of genomic feature is often not sufficient to characterize the clinical subgroup. We should note that the md-modules were constructed based on variable correlations/associations, which do not necessarily imply causal relationships among the variables. However, since many identified md-modules are of significant biological relevance, we believe that such modules can be a good start to uncover further underlying causal mechanisms of gene regulation.

Identifying coordinated patterns across multiple regulatory layers is a vital step toward revealing the high-order organization of complex gene regulatory systems. In this study, we attempted to reveal the coordinated subspace patterns comprising the epigenetic, transcription and post-transcription levels, yet the real picture can be much more complex, given the many other levels of regulatory controls (e.g. copy number changes, SNPs, protein transport and localization). For example, gene copy number losses of *miR-210* have been found in ovarian carcinomas (55), and mutations in *p53* are the most common gene mutations in human cancer, including ovarian cancers (56). In future studies, it will be worthwhile to apply the proposed method to more data sources simultaneously, to uncover more sophisticated 'factories' that comprise many layers of regulatory factors.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figures 1–3, Supplementary Package and Supplementary Dataset.

FUNDING

National Institutes of Health [R01GM074163]; the National Science Foundation [0747475] and the Alfred P. Sloan Fellowship (to X.J.Z.); the National Natural Science Foundation of China [11001256]; Innovation Project of Chinese Academy of Sciences (CAS), kjcx-yw-s7; the 'Special Presidential Prize'—Scientific Research Foundation of the CAS; the Special Foundation of President of AMSS at CAS for 'Chen Jing-Run' Future Star Program and the Foundation for Members of Youth Innovation Promotion Association, CAS (to S.Z.). Funding for open access charge: NSF07474750.

Conflict of interest statement. None declared.

REFERENCES

1. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

2. Weinstein,J.N., Myers,T.G., O'Connor,P.M., Friend,S.H., Fornace,A.J. Jr, Kohn,K.W., Fojo,T., Bates,S.E., Rubinstein,L.V., Anderson,N.L. *et al.* (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.
3. Bussey,K.J., Chin,K., Lababidi,S., Reimers,M., Reinhold,W.C., Kuo,W.L., Gwadry,F., Ajay., Kouros-Mehr,H., Fridlyand,J. *et al.* (2006) Integrating data on DNA copy number with gene expression levels and drug sensitivities in the NCI-60 cell line panel. *Mol. Cancer Ther.*, **5**, 853–867.
4. Scherf,U., Ross,D.T., Waltham,M., Smith,L.H., Lee,J.K., Tanabe,L., Kohn,K.W., Reinhold,W.C., Myers,T.G., Andrews,D.T. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
5. Staunton,J.E., Slonim,D.K., Coller,H.A., Tamayo,P., Angelo,M.J., Park,J., Scherf,U., Lee,J.K., Reinhold,W.O., Weinstein,J.N. *et al.* (2001) Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA*, **98**, 10787–10792.
6. Shankavaram,U.T., Reinhold,W.C., Nishizuka,S., Major,S., Morita,D., Chary,K.K., Reimers,M.A., Scherf,U., Kahn,A., Dolginow,D. *et al.* (2007) Transcript and protein expression profiles of the NCI-60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.*, **6**, 820–832.
7. Zhang,W., Zhu,J., Schadt,E.E. and Liu,J.S. (2010) A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.*, **6**, e1000642.
8. Gao,F., Foat,B.C. and Bussemaker,H.J. (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.
9. Fagan,A., Culhane,A.C. and Higgins,D.G. (2007) A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, **7**, 2162–71.
10. Kutalik,Z., Beckmann,J.S. and Bergmann,S. (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.*, **26**, 531–539.
11. Kim,P.M. and Tidor,B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, **13**, 1706–1718.
12. Brunet,J.P., Tamayo,P., Golub,T.R. and Mesirov,J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.
13. Badea,L. (2007) Combining gene expression transcription factor regulation data using simultaneous non-negative matrix factorization. In: *Proceedings of the BIOCOMP07*. CSREA Press, Las Vegas, Nevada, USA.
14. Badea,L. (2008) Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous non-negative matrix factorization. *Pac. Symp. Biocomput.*, pp. 267–278.
15. Zhang,S., Li,Q., Liu,J. and Zhou,X.J. (2011) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**, i401–i409.
16. Paatero,P. and Tapper,U. (1994) Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**, 111–126.
17. Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.
18. Lee,D. and Seung,H. (2001) Algorithms for non-negative matrix factorization. *Adv. Neural Inform. Process. Syst.*, **13**, 556–562.
19. Shahnaz,F., Berry,M., Pauca,P. and Plemmons,R. (2006) Document clustering using non-negative matrix factorization. *J. Inform. Process. Manage.*, **42**, 373–386.
20. Paschou,P., Ziv,E., Burchard,E.G., Choudhry,S., Rodriguez-Cintron,W., Mahoney,M.W. and Drineas,P. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.*, **3**, e160.
21. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
22. Kwon,M.J. and Shin,Y.K. (2011) Epigenetic regulation of cancer-associated genes in ovarian cancer. *Int. J. Mol. Sci.*, **12**, 983–1008.
23. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinform.*, **1**, 24–45.
24. Shen,R., Olshen,A.B. and Ladanyi,M. (2009) Integrative Clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **25**, 2906–2912.
25. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
26. Widschwendter,M., Fiegl,H., Egle,D., Mueller-Holzner,E., Spizzo,G., Marth,C., Weisenberger,D.J., Campan,M., Young,J. and Jacobs,I. (2007) Epigenetic stem cell signature in cancer. *Nat. Genet.*, **39**, 157–158.
27. Yamamura,S., Matsumura,N., Mandai,M., Huang,Z., Oura,T., Baba,T., Hamanishi,J., Yamaguchi,K., Kang,H.S., Okamoto,T. *et al.* (2012) The activated transforming growth factor-beta signaling pathway in peritoneal metastases is a potential therapeutic target in ovarian cancer. *Int. J. Cancer*, **130**, 20–28.
28. Chou,J.L., Chen,L.Y., Lai,H.C. and Chan,M.W. (2010) TGF-β: friend or foe? The role of TGF-β/SMAD signaling in epigenetic silencing of ovarian cancer and its implication in epigenetic therapy. *Expert. Opin. Ther. Targets*, **14**, 1213–1223.
29. Bhattacharya,R., Kwon,J., Ali,B., Wang,E., Patra,S., Shridhar,V. and Mukherjee,P. (2008) Role of hedgehog signaling in ovarian cancer. *Clin. Cancer Res.*, **14**, 7659–7666.
30. Rapberger,R., Perco,P., Sax,C., Pangerl,T., Siehs,C., Pils,D., Bernthaler,A., Lukas,A., Mayer,B. and Krainer,M. (2008) Linking the ovarian cancer transcriptome and immunome. *BMC Syst. Biol.*, **2**, 2.
31. Marks,J.R., Davidoff,A.M., Kerns,B.J., Humphrey,P.A., Pence,J.C., Dodge,R.K., Iglehart,J.D., Bast,R.C. Jr and Berchuck,A. (1991) Overexpression and mutation of p53 in epithelial ovarian cancer. *Cancer Res.*, **51**, 2979–2984.
32. Dinulescu,D.M., Ince,T.A., Quade,B.J., Shafer,S.A., Crowley,D. and Jacks,T. (2005) Role of K-ras and Pten in the development of mouse models of endometriosis and endometrioid ovarian cancer. *Nat. Med.*, **11**, 63–70.
33. Kuk,C., Gunawardana,C.G., Soosaipillai,A., Kobayashi,H., Li,L., Zheng,Y. and Diamandis,E.P. (2010) Nidogen-2: a new serum biomarker for ovarian cancer. *Clin. Biochem.*, **43**, 355–361.
34. Ulazzi,L., Sabbioni,S., Miotto,E., Veronese,A., Angusti,A., Gafà,R., Manfredini,S., Farinati,F., Sasaki,T., Lanza,G. *et al.* (2007) Nidogen 1 and 2 gene promoters are aberrantly methylated in human gastrointestinal cancer. *Mol. Cancer*, **6**, 17.
35. Renard,I., Joniau,S., van Cleynenbreugel,B., Collette,C., Naômè,C., Vlassenbroeck,I., Nicolas,H., de Leval,J., Straub,J., van Criekinge,W. *et al.* (2010) Identification and validation of the methylated TWIST1 and NID2 genes through real-time methylation-specific polymerase chain reaction assays for the noninvasive detection of primary bladder cancer in urine samples. *Eur. Urol.*, **58**, 96–104.
36. Lai,K.W., Koh,K.X., Loh,M., Tada,K., Subramaniam,M.M., Lim,X.Y., Vaithilingam,A., Salto-Tellez,M., Iacopetta,B., Ito,Y. *et al.* (2010) MicroRNA-130b regulates the tumour suppressor RUNX3 in gastric cancer. *Eur. J. Cancer*, **46**, 1456–1463.
37. Ma,S., Tang,K.H., Chan,Y.P., Lee,T.K., Kwan,P.S., Castilho,A., Ng,I., Man,K., Wong,N., To,K.F. *et al.* (2010) MiR-130b promotes CD133+ liver tumor-initiating cell growth and self-renewal via tumor protein 53-induced nuclear protein 1. *Cell Stem Cell*, **7**, 694–707.
38. Lui,W.O., Pourmand,N., Patterson,B.K. and Fire,A. (2007) Patterns of known and novel small RNAs in human cervical cancer. *Cancer Res.*, **67**, 6031–6043.
39. Reimer,D., Hubalek,M., Riedle,S., Skvortsov,S., Erdel,M., Concin,S., Fiegl,H., Müller-Holzner,E., Marth,C., Illmensee,K. *et al.* (2010) E2F3a is critically involved in epidermal growth factor receptor-directed proliferation in ovarian cancer. *Cancer Res.*, **70**, 4613–4623.
40. Siu,M.K., Chan,H.Y., Kong,D.S., Wong,E.S., Wong,O.G., Ngan,H.Y., Tam,T.F., Zhang,H., Li,Z., Chan,Q.K. *et al.* (2010) p21-activated kinase 4 regulates ovarian cancer cell proliferation,

migration, and invasion and contributes to poor prognosis in patients. *Proc. Natl Acad. Sci. USA*, **107**, 18622–18627.

41. Diehl,J.A. (2002) Cycling to cancer with cyclin D1. *Cancer Biol. Ther.*, **1**, 226–231.
42. Musgrove,E.A., Caldon,C.E., Barraclough,J., Stone,A. and Sutherland,R.L. (2011) Cyclin D as a therapeutic target in cancer. *Nat. Rev. Cancer*, **11**, 558–572.
43. Wendt,M.K., Tian,M. and Schiemann,W.P. (2012) Deconstructing the mechanisms and consequences of TGF-β-induced EMT during cancer progression. *Cell Tissue Res.*, **347**, 85–101.
44. Kim,H., Watkinson,J., Varadan,V. and Anastassiou,D. (2010) Multi-cancer computational analysis reveals invasion-associated variant of desmoplastic reaction involving INHBA, THBS2 and COL11A1. *BMC Med. Genomics*, **3**, 51.
45. Mees,S.T., Mardin,W.A., Wendel,C., Baeumer,N., Willscher,E., Senninger,N., Schleicher,C., Colombo-Benkmann,M. and Haier,J. (2010) EP300–a miRNA-regulated metastasis suppressor gene in ductal adenocarcinomas of the pancreas. *Int. J. Cancer*, **126**, 114–124.
46. Kang,Y. (2006) Pro-metastasis function of TGFbeta mediated by the Smad pathway. *J. Cell Biochem.*, **98**, 1380–1390.
47. Kakonen,S.M., Selander,K.S., Chirgwin,J.M., Yin,J.J., Burns,S., Rankin,W.A., Grubbs,B.G., Dallas,M., Coi,Y. and Quise,T.A. (2002) Transforming growth factor- β stimulates parathyroid hormone-related protein and osteolytic metastases via Smad and mitogen-activated protein kinase signaling pathways. *J. Biol. Chem.*, **277**, 24571–24578.
48. Kang,Y., He,W., Tulley,S., Gupta,G.P., Serganova,I., Chen,C.R., Manova-Todorova,K., Blasberg,R., Gerald,W.L. and Massaqué,J. (2005) Breast cancer bone metastasis mediated by the Smad tumor suppressor pathway. *Proc. Natl. Acad Sci. USA.*, **102**, 13909–13914.
49. Edson,M.A., Nalam,R.L., Clementi,C., Franco,H.L., Demayo,F.J., Lyons,K.M., Pangas,S.A. and Matzuk,M.M. (2010) Granulosa cell-expressed BMPR1A and BMPR1B have unique functions in regulating fertility but act redundantly to suppress ovarian tumor development. *Mol. Endocrinol.*, **24**, 1251–1266.
50. Gaubatz,S., Lindeman,G.J., Ishida,S., Jakoi,L., Nevins,J.R., Livingston,D.M. and Rempel,R.E. (2000) E2F4 and E2F5 play an essential role in pocket protein-mediated G1 control. *Mol. Cell*, **6**, 729–735.
51. Matsumura,N., Huang,Z., Mori,S., Baba,T., Fujii,S., Konishi,I., Iversen,E.S., Berchuck,A. and Murphy,S.K. (2011) Epigenetic suppression of the TGF-beta pathway revealed by transcriptome profiling in ovarian cancer. *Genome Res.*, **21**, 74–82.
52. Nam,E.J., Yoon,H., Kim,S.W., Kim,H., Kim,Y.T., Kim,J.H., Kim,J.W. and Kim,S. (2008) MicroRNA expression profiles in serous ovarian carcinoma. *Clini. Cancer Res.*, **14**, 2690–2695.
53. Zhang,L., Volinia,S., Bonome,T., Calin,G.A., Greshock,J., Yang,N., Liu,C.G., Giannakakis,A., Alexiou,P. *et al.* (2008) Genomic and epigenetic alterations deregulate microRNA expression in human epithelial ovarian cancer. *Proc. Natl Acad. Sci. USA*, **105**, 7004–7009.
54. Dahiya,N. and Morin,P.J. (2010) MicroRNAs in ovarian carcinomas. *Endocr. Relat. Cancer*, **17**, F77–89.
55. Iorio,M.V., Visone,R., Di Leva,G., Donati,V., Petrocca,F., Casalini,P., Taccioli,C., Volinia,S., Liu,C.G., Alder,H. *et al.* (2007) MicroRNA signatures in human ovarian cancer. *Cancer Res.*, **67**, 8699–8707.
56. Bast,R.C. Jr, Hennessy,B. and Mills,G.B. (2009) The biology of ovarian cancer: new opportunities for translation. *Nat. Rev. Cancer*, **9**, 415–428.