BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 19 2012, pages 2458–2466
doi:10.1093/bioinformatics/bts476

Gene expression

Advance Access publication August 3, 2012

Identifying multi-layer gene regulatory modules from multi-dimensional genomic data

Wenyuan Li[1,+], Shihua Zhang[2,+], Chun-Chi Liu[3] and Xianghong Jasmine Zhou[1,*]

[1]Program in Molecular and Computational Biology, Department of Biological Sciences, 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 and [3]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402, Taiwan, Republic of China

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Eukaryotic gene expression (GE) is subjected to precisely coordinated multi-layer controls, across the levels of epigenetic, transcriptional and post-transcriptional regulations. Recently, the emerging multi-dimensional genomic dataset has provided unprecedented opportunities to study the cross-layer regulatory interplay. In these datasets, the same set of samples is profiled on several layers of genomic activities, e.g. copy number variation (CNV), DNA methylation (DM), GE and microRNA expression (ME). However, suitable analysis methods for such data are currently sparse.

Results: In this article, we introduced a sparse Multi-Block Partial Least Squares (sMBPLS) regression method to identify multi-dimensional regulatory modules from this new type of data. A multi-dimensional regulatory module contains sets of regulatory factors from different layers that are likely to jointly contribute to a local 'gene expression factory'. We demonstrated the performance of our method on the simulated data as well as on The Cancer Genomic Atlas Ovarian Cancer datasets including the CNV, DM, ME and GE data measured on 230 samples. We showed that majority of identified modules have significant functional and transcriptional enrichment, higher than that observed in modules identified using only a single type of genomic data. Our network analysis of the modules revealed that the CNV, DM and microRNA can have coupled impact on expression of important oncogenes and tumor suppressor genes.

Availability and implementation: The source code implemented by MATLAB is freely available at: http://zhoulab.usc.edu/sMBPLS/.

Contact: xjzhou@usc.edu

Supplementary information: Supplementary material are available at *Bioinformatics* online.

Received on December 1, 2011; revised on July 16, 2012; accepted on July 24, 2012

1 INTRODUCTION

Eukaryotic gene expression (GE) is a complex process controlled at multiple levels, including epigenetic, transcriptional and post-transcriptional regulation. Dynamic and precise coordination of these regulatory processes is essential to maximize the efficiency and specificity in GE. Recent studies support the view that, rather than a simple 'step-by-step' production line, the GE machine is governed by multiple, complex and extensively coupled networks (Maniatis and Reed, 2002; Moore, 2005; Orphanides and Reinberg, 2002).

The development of high-throughput genomic technologies has enabled researchers to obtain a global view of gene regulation. Microarray and sequencing technologies can not only measure genome-wide GE levels but also profile DNA modifications (e.g. CNV [copy number variation]), epigenetic regulation (e.g. DNA methylation [DM] and histone modifications) and post-transcriptional regulation (e.g. microRNA expression [ME]). However, most genome-wide studies have been restricted to only one aspect of regulation such as studies based on GE profiles (Alter *et al.*, 2000; Omberg *et al.*, 2007; Tamayo *et al.*, 2007). Recently, a new type of large-scale multi-dimensional genomic dataset has been gaining in popularity. In these datasets, the same set of samples is profiled on several layers of genomic activity, e.g. CNV, DM, GE and ME. The Cancer Genomic Atlas (TCGA) (McLendon *et al.*, 2008) project and the NCI60 (Shoemaker, 2006) project provide this type of comprehensive genomic characterization for a cohort of cancer samples and cancer cell lines, respectively. Multi-dimensional datasets provide unprecedented opportunities to discover connections between the different layers of GE regulation.

The emerging large-scale multi-dimensional genomic data calls for novel computational methods. In fact, as the cost of sequencing falls, multi-dimensional characterization of samples will soon become standard practice. However, suitable analysis methods are currently sparse. In particular, since the different types of genomic data have different scales and units, we cannot simply aggregate them for analysis. Previous relevant effort has mostly focused on two-dimensional genomic datasets. For example, various eQTL methods can jointly analyze single-nucleotide polymorphism (SNP) and GE data to identify regulatory SNPs (Zhang *et al.*, 2010); multivariate regression can correlate GE and transcription factor (TF) binding data to associate TFs with their target genes (Gao *et al.*, 2004); and the Ping-Pong algorithm integrates GE and drug–response data (Kutalik *et al.*, 2008). Recently, several methods have been developed to analyze genomic datasets with more than two dimensions. For example, the multivariate model developed by Mankoo *et al.* (2011) and the sparse regression method proposed by Witten and Tibshirani (2009), both can learn multi-dimensional genomic

*To whom correspondence should be addressed.
+The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

data in the supervised manner. Another relevant method, cMonkey, is a multi-species biclustering method that was applied to analyze GE matrices from different species (Waltman *et al.*, 2010). In addition, there have been a series of multiple kernel learning methods designed for integrating heterogeneous genomic data (Alpaydin, 2011; Hamid *et al.*, 2012; Yu *et al.*, 2010). These methods combine multiple kernels (each of which is transformed from a data type) into a single kernel, known as kernel fusion, which is then used for prediction, regression or feature selection.

In this article, we used a novel approach for supervised module discovery from *R*-dimensional genomic data ($R \geq 3$), a topic that was not covered by the aforementioned methods. In particular, we aim to identify multi-dimensional gene regulatory modules, which from *R*-dimensional genomic data ($R \geq 3$). In this application, a regulatory module contains sets of regulatory factors from different layers that are likely to contribute jointly to a local 'gene expression factory'. Without losing generality, assume that we are given a four-dimensional dataset consisting of GE, copy number variation (CNV), DNA methylation (DM) and microRNA expression (ME) profiles measured on the same K samples (Figure 1). Considering CNV, DM and ME as the input variables and GE as the response variable, we represent this dataset as four matrices: $X_i \in \mathbb{R}^{K \times N_i}$, where $i = 1, 2, 3$ and $Y \in \mathbb{R}^{K \times M}$. In each matrix the rows correspond to the same samples. The columns of the matrices correspond to measurements of different types. We aim to identify subsets of the three types of variables (CNV, DM and ME) that jointly explain the expression of a subset of genes, in all or a subset of the samples. The union of these four subsets of the different types of variables are termed a 'Multi-Dimensional Regulatory Module (MDRM)' (Figure 1). In our matrix representation, such a module consists of k rows, and n_i ($i = 1, 2, 3$) and m columns for the CNV, DM, ME and GE data. This approach captures the association between different types of variables (CNV–DM–ME) in terms of their joint impact on GE and facilitates the reconstruction of the regulatory network across different layers.

To identify multi-dimensional regulatory modules, we introduced a sparse Multi-Block Partial Least Squares (sMBPLS) regression method. Partial least squares (PLS) is a class of regression methods for finding the fundamental relations between an input matrix (X) and a response matrix (Y). Instead of finding hyperplane of maximum variance between the input and response variables, it finds a linear regression model by projecting both variables to a new space (Boulesteix and Strimmer, 2007; Fornell and Bookstein, 1982; Lê Cao *et al.*, 2008; Liu and Rayens, 2007; Tenenhaus *et al.*, 2005). MBPLS method is an expansion of the PLS for the regression analysis of input variables that are blocked into multiple subsections (Wangen and Kowalski, 1988; Wold *et al.*, 1987). MBPLS was initially developed for the chemometrics analysis and has been rarely applied to Bioinformatics (Hwang *et al.*, 2004; Li and Chan, 2009). The multi-dimensional genomic data provide a new opportunity for its application. In this article, we further expanded the MBPLS method by imposing sparse constraints to identify multi-dimensional modules. In particular, we imposed sparse constraints in both genomic variables and the sample dimensions. Different from the original MBPLS whose objective is the regression analysis of the whole data blocks, the sMBPLS aims to decompose the whole data blocks into a collection of smaller building blocks—MDRMs.

We demonstrated the performance of our method on both simulated data and the multi-dimensional TCGA datasets. The simulation study showed that the sMBPLS method can accurately identify embedded multi-dimensional modules and remarkably outperforms the non-sparse approach. We applied sMBPLS method to a suite of TCGA data including the CNV, DM, ME and GE data on 230 ovarian cancer samples. We showed that majority of identified modules have significant functional and transcriptional enrichment, higher than that observed in modules identified using only a single type of genomic data. Our network analysis of the modules revealed that the multi-dimensional genomic components are tightly connected and the CNV, DM and microRNA can have combinatorial impact on expression of important oncogenes and tumor suppressor genes. Finally, we compared our sMBPLS approach to the commonly used approach in which all input data blocks were aggregated into a single block for module discovery. We found that almost half of modules from the single-block approach are not multi-dimensional, demonstrating the importance of our 'multi-block' approach in capturing functional relationships of variables from multiple dimensions.

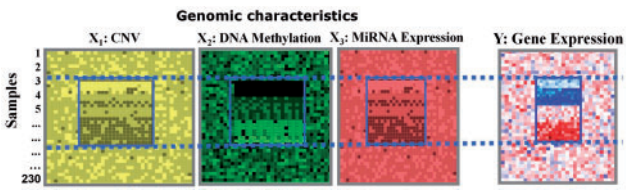

Fig. 1. Illustration of a 'multi-dimensional regulatory module'. A subsets of CNVs, DMs and MEs exhibit similar profiles as a subset of GEs across a subset of samples

2 METHODS

2.1 Definition of the MDRM

Given three input blocks X_1, X_2, X_3 and a response block Y, a multi-dimensional module is defined by satisfying the criterion "the profiles extracted from n_i columns across k rows of X_i ($i = 1, 2, 3$) has strong association with (or has similar and coherent pattern with) those from m columns across the same k rows of Y" (Fig. 1). Such association between two submatrices from X_i and Y can be measured by the covariance of their 'summary vectors'. The coincidence of these associations appearing in the same k samples establishes the strong signals that multiple types of input variables explain the response variables. Such distinct covariance structure in subspaces of multiple blocks can be identified by the sparse version of the MBPLS regression framework.

2.2 Objective function

We first introduce the covariance function of measuring the association between two matrices. Let X and Y be input and response matrices on the same K samples, respectively. To summarize columns of a matrix X, we introduce a 'summary' vector $\mathbf{t}$ which is a linear combination of all columns of X, i.e. $\mathbf{t} = X\,\mathbf{w}$ ($\mathbf{w}$ being the weights of input variables/columns). Similarly, $\mathbf{u}$ is a summary vector of Y columns, i.e. $\mathbf{u} = Y\,\mathbf{q}$ ($\mathbf{q}$ being the weights of response variable/columns). Thus, the larger the covariance of

two summary vectors t and u are, the more similar two matrices look like and the higher the association of two matrices is (Figure 2A). This association measure can be extended to multiple blocks of input variables.

Consider three input blocks $X_i \in \mathbb{R}^{K\times N_i}$ (where $i=1, 2, 3$), each of which contains a set of N_i centered (zero-mean) input variables on the same K samples, and let $Y \in \mathbb{R}^{K\times M}$ be the response data with M centered variables on the same K samples (Figure 2B). We used the weighted sum $\mathbf{t} = \sum_{i=1}^{3} b_i\mathbf{t}_i$ of 'summary vectors' $\mathbf{t}_i = X_i\mathbf{w}_i$ ($i=1, 2, 3$) of the three sets of input variables. The block weights $b_1,b_2,b_3>0$ indicate the contribution of each data block to the covariance structure of the input and response data. Therefore, the covariance between $\mathbf{t}$ and $\mathbf{u}=Y\mathbf{q}$ measures the association between three input data blocks and a response data block. The maximization of covariance between $\mathbf{t}$ and $\mathbf{u}$ can reveal the associations between from X_1,X_2,X_3 and Y, which lead to the discovery of a multi-dimensional module (Figure 2B). The problem is formally expressed as follows:

$$\max \mathrm{cov}\,(\mathbf{t},\mathbf{u}) \quad \text{with } \mathbf{t}_i = X_i\mathbf{w}_i,\ \mathbf{u} = Y\mathbf{q}, \text{ and } \mathbf{t} = \sum_{i=1}^{3} b_i\mathbf{t}_i \qquad (1)$$

subject to $\|\mathbf{w}_i\|^2=1$, $\|\mathbf{q}\|^2=1$, and $\|\mathbf{b}\|^2=1$. This is also the objective of the MBPLS regression problem (Wangen and Kowalski, 1988) which seeks to best explain the covariance structure between multiple groups of input variables and response variables. $\mathbf{t}_i$ and $\mathbf{u}$ are also called the 'latent variables' of the i^{th} block X_i and block Y, respectively, and $\mathbf{w}_i$ and $\mathbf{q}$ are their associated 'loading vectors'.

Although the solution to this objective can identify a multi-dimensional module by selecting those variables and samples with large absolute values from $\mathbf{w}_i,\mathbf{q},\mathbf{t}$, such module may not be the most distinct. As shown in our simulation study, the MBPLS regression approach often fails to identify distinct association signals of coherent structures. To address this problem, we added the sparsity penalties to the above objective.

2.3 Sparsity penalization

Sparsity penalization has recently attracted intense interest in regression analysis (Friedman, 2008), variable selection (Chun and Keles, 2010; Lê Cao *et al.*, 2008), matrix factorization (Kim and Park, 2007; Shen and Huang, 2008) and module discovery. The concept of sparsity (also called sparse coding in the literature) refers to a representational scheme (e.g. loading vector) where only a few elements are effectively used to represent data. Such sparsity is attractive from a data analysis viewpoint and makes representational scheme easy to interpret, as it selects the important elements and discards the rest. In effect, this implies most elements taking values close to zero while only few take significantly non-zero values. In our sparse version of the MBPLS problem, we searched the sparse representations of loading vectors whose non-zero elements can form a multi-dimensional module. It is achieved by adding sparsity penalties or regularizations to the optimization problem. Specifically, we adopted the widely used 'lasso penalization' (Tibshirani, 1996), which has been successfully applied in many fields. Let $\mathbf{x}$ be the vector to be computed in the optimization problem. The lasso regularization of $\mathbf{x}$, denoted $P_\lambda(\mathbf{x}) = \sum_i p_\lambda(x_i) = \sum 2\lambda|x_i|$ can be added to enforce sparsity on the solution of $\mathbf{x}$. Our maximization problem becomes

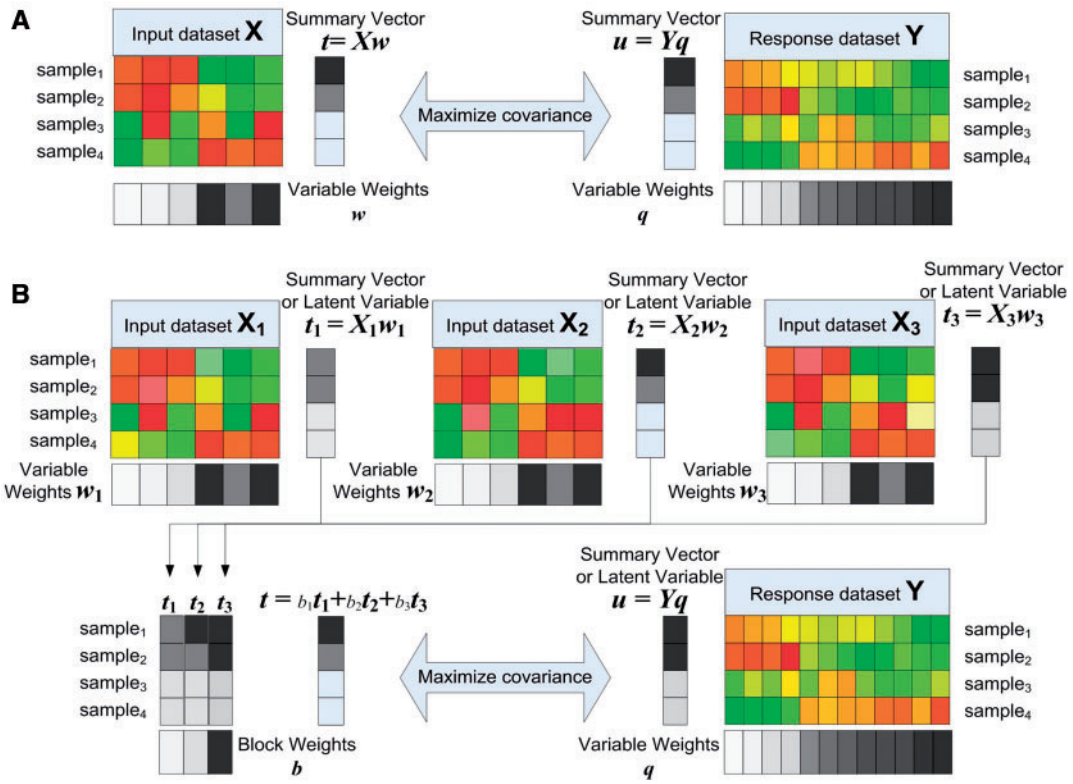

Fig. 2. Illustration of (**A**) the covariance function for measuring the association of two matrices and (**B**) the problem formulation of multi-dimensional module discovery. To search a multi-dimensional module, columns of each block are represented by a 'summary' vector, e.g. $\mathbf{t}_i$ summarizing X_i and $\mathbf{u}$ summarizing Y. Then the association between each input dimension X_i and the response dimension Y is measured by the covariance of their each summary vectors, i.e. $\mathrm{cov}(\mathbf{t}_i,\mathbf{u})$. The maximum covariance between summary vectors of X_i and Y reveals a distinct association representing the coherent profiles of X_i and Y. The maximization can be achieved by how we construct the summary vectors by weighting variables and samples. This discovery process is equivalent to the sparse version of the MBPLS problem

$$\max_{\mathbf{w}_i,\mathbf{q},\mathbf{t}_i,\mathbf{u}} \Omega(\mathbf{t},\mathbf{u},\mathbf{w}_i,\mathbf{q},\mathbf{b}) = \mathrm{cov}(\mathbf{t},\mathbf{u}) - \sum_{i=1}^{3} P_{\lambda_i}(\mathbf{w}_i) - P_{\lambda_4}(\mathbf{q})$$
$$\text{with } \mathbf{t}_i = X_i\mathbf{w}_i,\ \mathbf{u} = Y\mathbf{q}, \text{ and } \mathbf{t} = \sum_{i=1}^{3} b_i\mathbf{t}_i \qquad (2)$$
$$\text{subject to } \| \mathbf{w}_i \|^2 = 1,\ \| \mathbf{q} \|^2 = 1,\ \| \mathbf{b} \|^2 = 1$$

where the objective function $\Omega(\cdot)$ contains sparsity penalizations of the loading vectors $\mathbf{w}_i$ ($i = 1, 2, 3$) and $\mathbf{q}$.

2.4 Sparse multi-block PLS algorithm

To solve the problem in equation (2), we propose a sparse multi-block PLS (sMBPLS) regression algorithm. In this algorithm, sparse() is the soft thresholding function $\mathtt{sparse}_\lambda(x) = \mathtt{sign}(x)(|x| - \lambda)_+$ that is used to optimize the objective function with lasso penalties $p_\lambda(x)$. We prove that the sMBPLS algorithm can provide the maximizer of the sparse multi-block PLS problem (see Supplementary material).

sMBPLS Algorithm

(1) Initialize: Apply the MBPLS algorithm to X_1, X_2, X_3, Y and obtain the latent variable $\mathbf{u}^*$. Set $\mathbf{u} = \mathbf{u}^*$.

(2) Update:

(a) $\mathbf{w}_i = \mathtt{sparse}_{\lambda_i}(X_i^T\mathbf{u})$, norm $\mathbf{w}_i$ ($i = 1, 2, 3$)

(b) $\mathbf{t}_i = X_i\mathbf{w}_i$ (or $\mathbf{t}_i = \mathtt{sparse}_\mu(X_i\mathbf{w}_i)$) ($i = 1, 2, 3$)

(c) $T = [\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3]$

(d) $\mathbf{b} = T^T\mathbf{u}$, norm $\mathbf{b}$

(e) $\mathbf{t} = T\,\mathbf{b}$

(f) $\mathbf{q} = \mathtt{sparse}_{\lambda_4}(Y^T\mathbf{t})$, norm $\mathbf{q}$

(g) $\mathbf{u} = Y\mathbf{q}$ (or $\mathbf{u} = \mathtt{sparse}_\mu(Y\mathbf{q})$)

(3) Repeat Step 2 until convergence of $\mathbf{t}$.

In this algorithm, $\mathbf{u}$ is regressed on each block X_i to give the loading vector $\mathbf{w}_i$ of the block, which are then multiplied through the block to provide the latent variable $\mathbf{t}_i$. All three latent variables $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3$ are combined into the super block T and a classic PLS iterative cycle between T and Y is performed to give the block weights $\mathbf{b}$ and the combined latent variable $\mathbf{t}$. We repeat this until convergence on $\mathbf{t}$. The computational complexity of one sMBPLS iteration is $\mathbf{O}(\sum_{i=1}^{3} N_iK + MK)$. We are also interested in selection over the sample dimension. Specifically, we want to identify a multi-dimensional module whose input variables have the maximum covariance with response variables across a subset of samples. To achieve this goal, we impose the $\mathtt{sparse}_\mu$ function on Step 2(b) and Step 2(g) of the sMBPLS algorithm to select samples.

The iterative procedure of the sMBPLS algorithm can be used to obtain the first set of sparse loading vectors $\mathbf{w}_i$,$\mathbf{q}$ and latent variables $\mathbf{t}$. The non-zero elements of converged loading vectors and latent variable ($\mathbf{t}$) identify a multi-dimensional module that contains subsets of input and response variables and a subset of samples. After we identify a module, we deflate the matrix by subtracting the signal of current set of loadings and latent variables from the data matrices. Subsequent modules, or subsequent sets of sparse loadings and latent variables, can be obtained sequentially by maximizing covariance on the deflated matrices. We used the following deflation formula to remove the module's signal from each block:

$$\upsilon = f_{\lambda_i}\left(X_i^T\mathbf{t}_i/(\mathbf{t}_i^T\mathbf{t}_i)\right),\ X_i \leftarrow X_i - \mathbf{t}_i\upsilon^T,\ (i = 1,2,3)$$
$$\psi = f_{\lambda_4}\left(Y^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})\right),\ \ Y \leftarrow Y - \mathbf{u}\psi^T$$
$$\text{where } f_\lambda(x) = \begin{cases} x, & \text{for } |x| \geq \lambda \\ 0, & \text{otherwise} \end{cases}.$$

2.5 Tuning parameter selection

The sMBPLS algorithm is developed for fixed $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ (with the option of including μ). We tune these parameters using the cross-validation procedure that was also used in recently proposed methods such as sparse PCA (Shen and Huang, 2008) and sparse PLS (Lê Cao *et al.*, 2008). Tuning these parameters is equivalent to choosing the 'degree of sparsity', i.e. the number of non-zero components in each loading vector and latent variable. Note that setting the degree of sparsity to j ($1 < j < M$) (taking the loading vector $\mathbf{q}$ as an example) is equivalent to setting $\lambda_4 \in [|Y^T\mathbf{t}|_{(j)}, |Y^T\mathbf{t}|_{(j+1)}]$, where $|Y^T\mathbf{t}|_{(j)}$ is the j^{th}-order statistic of $Y^T\mathbf{t}$. Our computational framework allows different vectors to have different degrees of sparsity. To simplify notations, we use $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and μ to denote the degrees of sparsity of loading vectors and latent variables in the rest of the article. The cross-validation procedure is presented below.

(1) Randomly place the samples into L roughly equal groups. Each group has a corresponding matrix from each block of data. That is, the matrix of the i^{th} genomic block data with only samples in the l^{th} group is denoted X_i^l; let X_i^{-l} be the matrix composed of data from all other samples. The same notation applies to the response block data Y, which is divided into Y^l and Y^{-l} for each group.

(2) For each combination of degrees of sparsity, $\lambda_i \in \{2,3,\ldots,N_i\}$ ($i = 1,2,3$), $\lambda_4 \in \{2,3,\ldots,M\}$ and $\mu \in \{2,3,\ldots,K\}$

(a) For $l = 1,\ldots,L$, apply the sMBPLS algorithm on $X_1^{-l}, X_2^{-l}, X_3^{-l}, Y^{-l}$, to derive loading vectors of independent variables: $\mathbf{w}_1^{-l}, \mathbf{w}_2^{-l}, \mathbf{w}_3^{-l}$ and the loading vector of response variables $\mathbf{q}^{-l}$. Next, project X_1^l, X_2^l, X_3^l, Y^l onto $\mathbf{w}_1^{-l}, \mathbf{w}_2^{-l}, \mathbf{w}_3^{-l}$ and $\mathbf{q}^{-l}$ to obtain the projection coefficients as $\xi_i^l = X_i^l\mathbf{w}_i^{-l}$ and $\zeta^l = Y^l\mathbf{q}^{-l}$, respectively.

(b) Calculate the L-fold CV score defined as

$$CV = \sum_{i=1}^{3}\sum_{l=1}^{L} \frac{\| X_i^l - \xi_i^l\mathbf{w}_i^{-l^T} \|^2}{K_lN_i} + \sum_{l=1}^{L} \frac{\| Y^l - \zeta^l\mathbf{q}^{-l^T} \|^2}{K_lM} \qquad (3)$$

where K_l is the number of samples in the lth group.

(3) Select the combination of degrees of sparsity $\{\tilde{\lambda}_1, \tilde{\lambda}_2, \tilde{\lambda}_3, \tilde{\lambda}_4, \tilde{\mu}\}$ whose CV score is the smallest.

L and the number of combinations of degrees of sparsity will impact the computational efficiency. In practice, L is usually chosen to be 5 or 10 for large datasets. Naturally, a different random grouping in Step 1 may result in different degrees of sparsity. Usually, the larger the value of L, the more stable the degrees of sparsity selected by CV. In our study, we use $L = 5$, which is large enough for a stable selection of parameters. The number of combinations of degrees of sparsity for the thresholds λ_i ($i = 1$, 2, 3, 4) and μ is large for large-scale data. Therefore, in practice, we used a subset of combinations in the cross-validation procedure.

3 RESULTS

To assess the performance of sMBPLS, we first applied the sMBPLS to a variety of simulated data, generated with various complexities. We then applied sMBPLS to the multi-dimensional TCGA ovarian cancer data to gain insights into the multi-layer coordination of GE regulations. To reveal the advantages of sMBPLS, we also compared our methods to several competing methods including MBPLS, sparse PLS and biclustering methods.

3.1 Simulation study

The simulation data were generated by extending scenarios proposed in the recent literature on sparse PLS methods (Chun and

Keles, 2010; Lê Cao *et al.*, 2008) to multi-dimensional data (see Supplementary material for details). The scenario follow the general model of $Y = \sum_{i=1}^{3} b_i X_i \beta_i + \Xi$, where Ξ is the Gaussian noise and the settings of b_i and β_i refer to the Supplementary material. X_i is generated with various complexity: different sizes, hidden component structures, correlation structures and even the presence of multicolinearity (simulated from a multivariate normal with a first-order autoregressive process's covariance matrix with auto-correlation $\rho = 0.9$).

In order to discover the multi-dimensional modules by using MBPLS, an intuitive two-step procedure can be performed: first applying MBPLS to the data, then selecting the top-ranking input and response variables to form a module by ordering the absolute values of the loadings and weight vectors. A more intuitive comparison way is to sort the loadings and latent variable from either MBPLS or sMBPLS, then to visualize X_1, X_2, X_3, Y whose rows and columns are reordered by the sortings of their loadings and weight vectors. The multi-dimensional module would appear in the left-top and right-bottom corners (whose corresponding variables and samples have large absolute values of weights) in the reordered blocks. For example, the reordered blocks by MBPLS as shown in Figure S1(B)-panel2 (in Supplementary material) are observed to have no clear modular submatrices in corners, while a multi-dimensional module can be observed in reordered blocks by sMBPLS in Figure S1(B)-panel3 and zoomed out in Figure S1(C).

We further systematically compared two methods by simulating data 50 times. We found that MBPLS always failed for all 50 simulation data in that it assigns totally different variables and samples to the discovered module. In contrast, all modules identified by sMBPLS have significant overlaps with predefined modules. (Details of overlap significance test for modules can be found in Section S3 of Supplementary material.) An example is detailed in Figure S1. Since the MBPLS method maximizes the covariance between all input and response variables across all samples as shown in Figure S1(B), it overlooks embedded modules when the module's signal is overwhelmed by background noise. On the contrary, the sparsity penalty forces sMBPLS to focus on 'local' (i.e. across a small subsets of variables and samples) peaks in the covariance, which correspond to (multi-dimensional) modules of relatively small size. Our results show that the sMBPLS method is more accurate at identifying modules in noisy data, and thus more suitable for biological applications.

We also compared the sMBPLS with the sparse version of PLS which is applied to the combined single input block (i.e. X_1, X_2, X_3 merged to a single block X). The same 50 times of simulation data were used. The result showed that all modules identified by sparse PLS have at least one dimension missing, 56% modules have two dimensions missing and none of these modules module has significant overlap with predefined module. The lack of such power for the single-block approach may attribute to the unbalanced covariance structures across multiple blocks, i.e. the covariance signals of some blocks may be overwhelmed by those of other blocks. The result is even worse when we compared with a popular biclustering algorithm 'SAMBA' (Tanay *et al.*, 2002) which is applied to the single block by merging X_1, X_2, X_3, Y. All modules identified by SAMBA have at least one dimension missing, and 84%/34% modules have at least two/three dimensions missing. None of these modules has overlaps with predefined modules.

3.2 Identifying MDRM in TCGA ovarian cancer data

We applied the sMBPLS method to the TCGA ovarian cancer genomic data. We included four types of genomic data profiled on the same 230 samples: CNV, DM, ME and GE. The data were downloaded from http://cancergenome.nih.gov/. Detailed preprocessing procedures can be found on the TCGA website (we used the Level 3 data). We filtered out genomic variables with little variation across the whole sample ($|\mu/\sigma| < 0.5$, where μ and σ are the mean and standard deviation of each variable), resulting in a final dataset with the expressions of 799 microRNAs and 15 846 genes, the CNV profiles of 31 324 loci and the DMs of 14 735 marks (see Section S5 of Supplementary material). We repeated the iterative cycle of sMBPLS to identify the modules one by one, until no further significant gain is achieved on the covariances between input and response variables of the identified modules. In the following, we perform further analysis on the top 100 modules. The details of the 100 modules can be found on the supplementary website.

On average, each module contains 30 samples, 45 CNV loci, 42 methylation marks, 5 microRNAs and 44 genes (see the figure of the distribution of module sizes in Supplementary material). We used the overlap significance test on the identified 100 modules to investigate how distinct these modules are. Our results showed that only one pair of modules has significant overlap at the level of P-value after Bonferroni correction < 0.05 (see Section S8 of Supplementary material). Figure 3 shows the heat maps of two example modules, demonstrating the high degree of (anti-)correlation between the four dimensions. As expected, genomic profiles of most CNV marks are positively correlated with the expression levels of genes. But we can also see that genomic profiles of the methylation marks are sometimes partially anti-correlated with the expression levels of genes. We should note that our problem formulation shown in Figure 2(B) is the covariance maximization between $\mathbf{u}$ (from Y) and the weighted sum of $\mathbf{t}_i$ (from X_i), so it can well capture the holistic

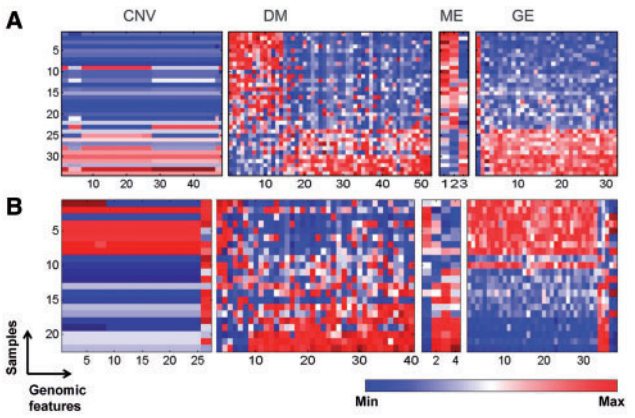

Fig. 3. Heat map for feature profiles of CNV, DM, microRNA and GE in modules across the same small set of samples for (**A**) Module 23 and (**B**) Module 83. Each row represents a sample and each column represents a genomic feature

correlations between response block and all input blocks. This formulation can work well but cannot guarantee to identify those modules in which each of input dimensions is highly correlated with the response dimension. Therefore in some cases such as shown Figure 3(B), there does exist some input dimension that may not have expected correlation with the response dimension. In addition to identifying multiple types of variables that jointly explain the expression of a set of genes, sMBPLS also provides the relative weights (i.e. the block weights $\mathbf{b} = [b_1, b_2, b_3]$) of each dimension in contributing to the observed covariance. The weight b_i ($i = 1, 2, 3$) is proportional to the correlation of $\mathbf{t}_i$ and $\mathbf{u}$ in the sMBPLS model. Among the 100 modules, we observed a significant correlation between the latent variables $\mathbf{t}_i$ of the CNV, DM and ME dimensions and that $\mathbf{u}$ of the GE dimension in 63, 100 and 91 modules, respectively (P-value <0.01, computed from a Student's t distribution for a transformation of the Pearson's correlation). In reality, it is not necessary for all dimensions to be equally important. The block weight information can help to identify important regulatory factors from the selected variables. To evaluate how robust these identified modules are, we randomly remove 10% samples from the dataset and then performed the same procedure to identify 100 modules. By using the overlap significance test, we checked the overlaps between them and the 100 modules identified from full set of samples. The 74 modules identified from 90% samples show significant overlaps over at least three dimensions with 79 modules identified from full set of samples. This result indicates good robustness of our method.

Because none of conventional methods can solve our problem, we could only resort to comparing with those well-established methods that can approximately arrive at our target. We used two classes of conventional methods: (1) biclustering algorithms, which identify correlated subsets of features across subsets of samples from a single block (e.g. combined X and Y) and (2) sparse PLS methods, which make regression analysis over two blocks X and Y. The former is unsupervised learning for exploratory data analysis while the latter is supervised learning. We used SAMBA, a popular adopted biclustering algorithm (Tanay *et al.*, 2002), to perform on the block X combining all four dimensions (CNV, DM, ME and GE). Out of the top 100 modules identified by SAMBA, 59% of the modules missed at least one type of variables and 22% of the modules missed at least two types of variables. We further compared our sMBPLS approach to the sPLS approach in which we combined three types of input data, CNV, DM and ME into a single combined block X. Out of the top 100 modules identified by sPLS, 47% of the modules missed at least one type of input variables and 17% of the modules were only one-dimensional. This result showed the unique advantage of multi-block modeling in capturing modules that elucidate relationships of variables from multiple dimensions.

3.3 MDRMs reveal synergistic functions across multiple dimensions

To evaluate the biological relevance of those identified multi-dimensional modules, we first test the functional homogeneity for each dimension of them. A set of genes is defined as functionally homogenous if it is enriched in at least one GO category (Ashburner *et al.*, 2000) with a q-value <0.05 (the q-value is the P-value after a false discovery rate multiple testing correction). This was often the case. The GE dimension is functionally homogenous with respect to genes in 36% modules; the CNV dimension with respect to CNV-harbored genes in 24% modules; the DM dimension with respect to Methylation mark adjacent genes in 13% modules and the ME dimension with respect to microRNA in 9% modules (microRNA function was predicted based on the functions of their target genes), which are significantly higher than the 1.24%, 1.48%, 2.32% and 0.35% modules after randomization (Figure 4A), respectively.

Moreover, in 17 of those miRNA modules, miRNAs were enriched with members from the same miRNA clusters (q-value <0.1 after multiple testing correction), where a miRNA cluster is defined as a set of miRNAs located within 50 kb in the genome (Baskerville and Bartel, 2005). miRNAs in a cluster are expected to play related functional roles (Yuan *et al.*, 2009). For example, the Module 96 includes five miRNAs (miR-let-7e, miR-125a, miR-150, miR-200c and miR-141). Two of these, let-7e and 125a, are members of a miRNA cluster in chromosome 19, while miR-200c and miR-141 belong to another miRNA cluster in chromosome 12. It can also be shown that many modules contain genes targeted by miRNAs in the same modules (see Section S6 in Supplementary material). Members of the let-7-family have been extensively reported to suppress ovarian cancers (Koturbash *et al.*, 2010). Also, miR-125a, miR-200c and miR-141 were reported to be dys-regulated in ovarian cancer. To take another example, Module 45 covers three miRNAs (miR-27a, miR-23a and miR-205). Interestingly, miR-27a and miR-23a are clustered in the genome and both have been reported to be up-regulated in ovarian cancer (Koturbash *et al.*, 2010). In addition, miR-205 has been extensively studied in

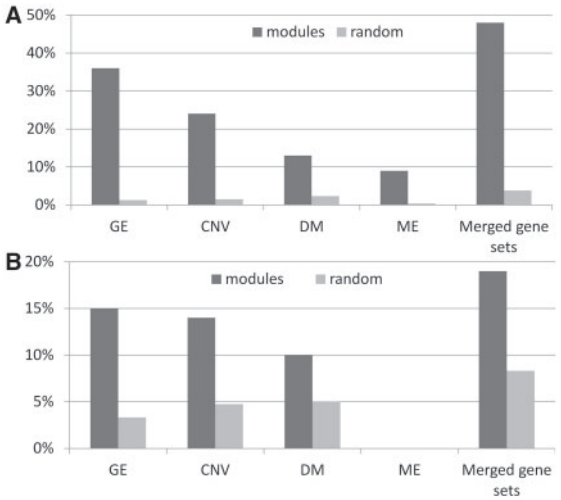

Fig. 4. Comparison of **(A)** functional homogeneity and **(B)** transcriptional homogeneity between gene sets from identified modules (blue bars) and randomized gene sets (red bars). The gene set of an identified module is either from each individual dimension (GE, CNV, DM or ME) of the module or from genes combining all dimensions of the module. Shown are percentages of gene sets that are functionally or transcriptionally enriched with q-value <0.05

relation to cancers of the bladder, lung, pancreatic, breast, esophagus and prostate.

While the individual dimensions of those modules already exhibit significant level of functional homogeneity, combining all dimensions reveals an even stronger functional synergy. When we consider all genes in the GE dimension, CNV-harbored genes, methylation adjacent genes and microRNAs, 48 out of the 100 modules were found to be functionally homogenous (Figure 4), compared to 3.8% modules in randomized data. This result highlights the power of multi-dimensional modules in grouping functionally relevant factors from different regulatory layers. In addition, we further compared the *q*-values of GO terms enriched for modules identified by either our method or the sparse PLS method by following the comparison procedure described in (Costa *et al*., 2008; Ernst *et al*., 2005). The result indicates that our method can identify more functionally homogeneous modules with more diverse GO terms (see Section S7 of Supplementary material). Many of the identified modules are enriched of the biological processes such as cell cycle, cell activation, immune system process and so on, implying their possible involvement in cancer. Also, they contain many important genes (or microRNAs) known to be related to ovarian cancer. For example, Module 37 includes four *HOX* family genes (*HOXB2*, *HOXB4*, *HOXB6* and *HOXB7*) that all have been extensively reported to be related with ovarian cancer (Cheng *et al*., 2005; Ota *et al*., 2009; Widschwendter *et al*., 2009; Wu *et al*., 2007). In addition, the module contains a DM mark adjacent to *HOXA9* that was reported to be significantly hypermethylated in ovarian cancer patients (Widschwendter *et al*., 2009; Wu *et al*., 2007). In addition, the member genes *FGF19*, *GAS2*, *BMP7*, *TNFSF11*, and *FGFR3* are all known to play important roles in tumor genesis and progression. In the Module 76, miR-214, known to be involved in cell cycle, has been reported to be dys-regulated in ovarian cancer (Iorio *et al*., 2007; Nam *et al*., 2008; Yang *et al*., 2008), and *CDH13* is a potential epigenetic biomarker for ovarian cancer (Wu *et al*., 2007).

We then test the transcriptional homogeneity for each dimension of the identified multi-dimensional modules. We used the 191 ChIP-seq profiles generated by the Encyclopedia of DNA Elements (ENCODE) consortium (Thomas *et al*., 2007). This dataset includes the genome-wide binding of 40 TFs, 9 histone modification marks and 3 other markers (DNase, FAIRE and DM) on 25 different cell lines (see Supplementary material). These data provide a set of potential targets of regulatory factors. A set of genes is defined as 'transcriptional homogenous' if it is enriched in the targets for any regulatory factor with a *q*-value <0.05. We achieved similar results as those of functional homogeneity analysis (Figure 4B). These modules are enriched of the TFs such as *SRF*, *STAT1* and *H3K27me3*. *SRF* regulates the activity of many immediate-early genes and thereby participates in cell cycle regulation, apoptosis, cell growth and cell differentiation. *STAT1* enhances inflammation and innate and adaptive immunity, triggering in most instances anti-proliferative and pro-apoptotic responses in tumor cells (Pensa *et al*., 2009). Particularly, *STAT1* negatively regulates the cell cycle by inducing *p21 WAF1/CIP1* in ovarian cancer (Burke *et al*., 1999). *H3K27me3* has been evaluated as a prognostic indicator for clinical outcome in patients with breast and ovarian cancers (Wei *et al*., 2008).

3.4 MDRMs facilitate the regulatory analysis

Our method has provided sets of genomic features from different regulatory layers that are likely to be synergistic in their impact on GE. To further elaborate the relationships between those implicated features, we used the Ingenuity Pathway Analysis (IPA) system (Redwood City, CA, USA) to build molecular interaction networks. From each multi-dimensional module, we formed a set consisting of genes involved in the GE dimension, CNV-harbored genes, methylation-adjacent genes and microRNAs. Using this set as the input, IPA constructs networks based on literature-derived relationships between genes (or microRNAs) and computes a ranking score $-log(p)$ for each network. The *P*-value indicates the likelihood that the genes in the input network would be found together due to random chance. All of the multi-dimensional modules lead to statistically significant interaction networks (*P*-value $<$ 1.0E-20) by this analysis, which indicates the significant associations among them.

Below, we provide in-depth descriptions of the heterogeneous regulatory networks that affect a key tumor suppressor gene (*EGR1*) and an oncogene (*AKT*) in ovarian cancer. *EGR1* is a cancer-suppressing gene known to be down-regulated in ovarian cancer (Lamber *et al*., 2010). The network derived from the Module 4 reveals that complex connections of heterogeneous factors control the expression of *EGR1* (Figure 5A). A direct regulation on *ERG1* comes from *PITX2*, a gene adjacent to a DM mark in our module and known to be essential for the expression of *EGR1* in rat (Suh *et al*., 2002). Multiple indirect influences on *EGR1* are transmitted by the TF *NFkB*, by *GNA15* and *GNA11* (genes hosted by CNVs in the module) and by *FCER2* (genes adjacent to the methylation marks in the module). In particular, *WT1* (regulated by DM) is known to positively regulate *AMH* (Nachtigal *et al*., 1998) (which is also regulated by a CNV), which in turn positively regulates *EGR1* via *NFkB*. In addition, *EGR1* is regulated by both *TMEM173* and *ENTD1* (adjacent to methylations) via *ERK* and *P38 MARK*, respectively. The disruption of multiple neighbor nodes to *EGR1* by different regulatory mechanisms highlights the complex nature of the controls on this key suppressor gene for ovarian cancer. As another example, from Module 61 we derived a multi-layer coordinated network (Figure 5B) regulating *AKT*, a key oncogene for ovarian cancer (Altomare *et al*., 2004;

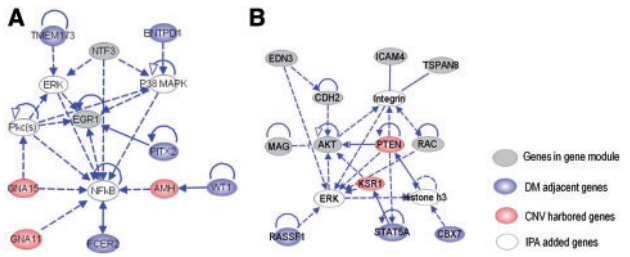

Fig. 5. The molecular interaction networks (constructed by IPA) that center around the genes (**A**) *EGR1* and (**B**) *AKT*. The networks consist of affected genes (gray nodes), CNV-sharbored genes (red nodes) and DM-adjacent genes (blue nodes). The solid lines represent direct interactions and the dashed lines represent indirect interactions

Yuan *et al*., 2000) and an important component of the *PI3-kinase/AKT* pathway. The expression of *AKT* is positively regulated by the loss of a CNV-containing *PTEN*, which would otherwise down-regulate the expression of *AKT* (Pore *et al*., 2003). Moreover, *AKT* is activated by *CDH2* (Rieger-Christ *et al*., 2004), the expression of which, in turn, is increased by *EDN3* (Bagnato *et al*., 2004). *AKT* is further up-regulated due to the methylation of *RASSF1* (a tumor suppressor gene), which is consistent with the previously observed over-expression of *AKT* in *RASSF1A*-depleted cells (Dallol *et al*., 2005). Additional regulations are exerted by the loss of *KSR1* (CNV), the methylated *STAT5A* and the expression of *MAG*, among others. These factors, intertwined together, paint complex mechanisms leading to the activation of the important oncogene *AKT* in ovarian cancer.

In addition to the two examples detailed above, we observed multi-layer coupled regulatory networks around multiple oncogenes or tumor suppressor genes (e.g. *PIK3CA* and *CCNE1*). The results of this analysis clearly illustrate and illuminate the complex genetic origins of ovarian cancer. In fact, CNV, methylation and microRNA regulation are only few of the many expression regulatory mechanisms, and much more sophisticated coordination likely exists in the origins of this illness. The algorithm proposed in this study takes the first steps along the path to effectively integrate multi-dimensional data to explore the complex regulatory mechanisms.

4 CONCLUSIONS

In this study, we developed a sMBPLS regression method to identify multi-dimensional regulatory modules in diverse types of genomic data measured on the same set of samples. Classical eQTL analysis can only be applied to relate one type of genomic marker (e.g. SNP) to GE. In contrast, sMBPLS can identify combinations of multiple types of genomic markers that jointly impact the expression of a set of genes. We have applied the sMBPLS method to a suite of genomic profiles from 230 ovarian cancer samples, including CNV, DM, microRNA and GE data. The algorithm identified 100 modules, many of which display a high degree of functional homogeneity in at least one genomic dimension. If all dimensions of data are considered together, the modules exhibit an even greater degree of functional synergy. A detailed network view of individual modules reveals that many genomic features would remain isolated if we only considered one type of data. By combining diverse types of data, sMBPLS links the different regulatory layers and thus discovers more coherent and connected regulatory networks. Furthermore, our method derives weights for the dimensions of CNV, methylation and microRNA in each module, which indicate their relative contributions to the expression of individual sets of genes. We have demonstrated that multiple heterogeneous factors in a module can have combinatorial effects on GE. We should note that (1) this does not necessarily reflect the direct causal mechanisms for GE, but the revealed modules can be a good start point to further study the underlying mechanisms; (2) sMBPLS outperforms most existing algorithms in analyzing more than two data blocks, although it may not possibly improve the results when applied to only two blocks X and Y.

In summary, we expect that there will soon be a rapid increase of multi-dimensional data, and developing methods for such data will become an active research area. We have proposed a promising tool to extract coherent substructures from large-scale, complex datasets, greatly facilitating downstream biological analysis. Interpreting such complex modules is still a major challenge, given our limited knowledge of multi-layer coordination in biological systems. However, the rapid accumulation of multi-dimensional data and the knowledge derived from them will definitely accelerate a positive cycle of the knowledge discovery.

ACKNOWLEDGEMENTS

We thank the reviewers for their constructive comments, which we used to improve the manuscript.

Funding: This work was supported by the National Institutes of Health Grants R01GM074163, the National Science Foundation Grant 0747475 and the Alfred P. Sloan Fellowship to X.J.Z.; the National Natural Science Foundation of China, No. 11001256, 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.

Conflict of Interest: none declared.

REFERENCES

Alter,O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci.*, **97**, 10101–10106.

Altomare,D.A. *et al.* (2004) AKT and mTOR phosphorylation is frequently detected in ovarian cancer and can be targeted to disrupt ovarian tumor cell growth. *Oncogene*, **23**, 5853–5857.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Bagnato,A. *et al.* (2004) Endothelin B receptor blockade inhibits dynamics of cell interactions and communications in melanoma cell progression. *Cancer Res.*, **64**, 1436–1443.

Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.

Boulesteix,A. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinformatics*, **8**, 32–44.

Burke,F. *et al.* (1999) Cytotoxic response of ovarian cancer cell lines to IFN-gamma is associated with sustained induction of IRF-1 and p21 mRNA. *Br. J. Cancer*, **80**, 1236–1244.

Cheng,C. *et al.* (2011) Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. *PLoS Comput. Biol.*, **7**, e1002190.

Cheng,W. *et al.* (2005) Lineage infidelity of epithelial ovarian cancers is controlled by HOX genes that specify regional identity in the reproductive tract. *Nat. Med.*, **11**, 531–537.

Chun,H. and Keles,S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. Roy. Stat. Soc. B (Stat. Methodol.)*, **72**, 3–25.

Costa,I.G. *et al.* (2008) Inferring differentiation pathways from gene expression. *Bioinformatics*, **24**, i156–i164.

Dallol,A. *et al.* (2005) Involvement of the RASSF1A tumor suppressor gene in controlling cell migration. *Cancer Res.*, **65**, 7653–7659.

Ernst,J. *et al.* (2005) Clustering short time series gene expression data. *Bioinformatics*, **21** (Suppl. 1), i159–i168.

Fornell,C. and Bookstein,F. (1982) Two structural equation models: Lisrel and pls applied to consumer exit-voice theory. *J. Market. Res.*, **19**, 440–452.
Friedman,J. (2008) Fast sparse regression and classification. In *Technical report*. Department of Statistics, Starford University.
Gao,F. *et al.* (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.
Gönen,M. and Alpaydin,E. (2011) Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, **12**, 2211–2268.
Hamid,J.S. *et al.* (2012) Weighted kernel fisher discriminant analysis for integrating heterogeneous data. *Comput. Stat. Data Anal.*, **56**, 2031–2040.
Hwang,D. *et al.* (2004) Inverse modeling using multi-block PLS to determine the environmental conditions that provide optimal cellular function. *Bioinformatics*, **20**, 487–499.
Iorio,M.V. *et al.* (2007) MicroRNA signatures in human ovarian cancer. *Cancer Res.*, **67**, 8699–8707.
Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.
Koturbash,I. *et al.* (2011) Small molecules with big effects: the role of the microRNAome in cancer and carcinogenesis. *Mutat. Res*, **722**, 94–105.
Kutalik,Z. *et al.* (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotech*, **26**, 531–539.
Lamber,E.P. *et al.* (2010) BRCA1 represses amphiregulin gene expression. *Cancer Res.*, **70**, 996–1005.
Lê Cao,K.-A. *et al.* (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol*, **7,** Article No. 35.
Li,Z. and Chan,C. (2009) Systems biology for identifying liver toxicity pathways. *BMC Proc.*, **3** (Suppl. 2), S2.
Liu,Y. and Rayens,W. (2007) PLS and dimension reduction for classification. *Comput. Stat.*, **22**, 189–208.
Maniatis,T. and Reed,R. (2002) An extensive network of coupling among gene expression machines. *Nature*, **416**, 499–506.
Mankoo,P.K. *et al.* (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*, **6**, e24709.
McLendon,R. *et al.* (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
Moore,M. (2005) From birth to death: the complex lives of eukaryotic mRNAs. *Science*, **309**, 1514.
Nachtigal,M.W. *et al.* (1998) Wilms' tumor 1 and dax-1 modulate the orphan nuclear receptor SF-1 in sex-specific gene expression. *Cell*, **93**, 445–454.
Nam,E.J. *et al.* (2008) MicroRNA expression profiles in serous ovarian carcinoma. *Clin. Cancer Res.*, **14**, 2690–2695.
Omberg,L. *et al.* (2007) A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proc. Natl. Acad. Sci.*, **104**, 18371–18376.
Orphanides,G. and Reinberg,D. (2002) A unified theory of gene expression. *Cell*, **108**, 439–451.
Ota,T. *et al.* (2009) Expression and function of HOXA genes in normal and neoplastic ovarian epithelial cells. *Differentiation*, **77**, 162–171.
Pensa,S. *et al.* (2009) STAT1 and STAT3 in tumorigenesis: two sides of the same coin? In *JAK-STAT Pathway in Disease*, Landes Bioscience, Austin, Texas, pp. 100–121.
Pore,N. *et al.* (2003) PTEN mutation and epidermal growth factor receptor activation regulate vascular endothelial growth factor (VEGF) mRNA expression in human glioblastoma cells by transactivating the proximal VEGF promoter. *Cancer Res.*, **63**, 236–241.
Rieger-Christ,K.M. *et al.* (2004) Novel expression of N-cadherin elicits *in vitro* bladder cell invasion via the Akt signaling pathway. *Oncogene*, **23**, 4745–4753.
Shen,H. and Huang,J.Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.*, **99**, 1015–1034.
Shoemaker,R.H. (2006) The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer*, **6**, 813–823.
Suh,H. *et al.* (2002) Pitx2 is required at multiple stages of pituitary organogenesis: pituitary primordium formation and cell specification. *Develop. (Cambridge, England)*, **129**, 329–337.
Tamayo,P. *et al.* (2007) Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proc. Natl. Acad. Sci.*, **104**, 5959–5964.
Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18** (Suppl. 1), S136–S144.
Tenenhaus,M. *et al.* (2005) PLS path modeling. *Comput. Stat. Data Anal.*, **48**, 159–205.
Thomas,D.J. *et al.* (2007) The ENCODE project at UC santa cruz. *Nucleic Acids Res.*, **35** (Database issue), D663–D667.
Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B (Methodological)*, **58**, 267–288.
Waltman,P. *et al.* (2010) Multi-species integrative biclustering. *Genome Biol.*, **11**, R96.
Wangen,L.E. and Kowalski,B.R. (1988) A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometr.*, **3**, 3–20.
Wei,Y. *et al.* (2008) Loss of trimethylation at lysine 27 of histone H3 is a predictor of poor outcome in breast, ovarian, and pancreatic cancers. *Mol. Carcinogen.*, **47**, 701–706.
Widschwendter,M. *et al.* (2009) HOXA methylation in normal endometrium from premenopausal women is associated with the presence of ovarian cancer: a proof of principle study. *Int. J. Cancer*, **125**, 2214–2218.
Witten,D.M. and Tibshirani,R.J. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. **8,** Article No. 28.
Wold,S. *et al.* (1987) PLS modeling with latent variable in two or more dimensions. *Proc. Symp. on PLS Model Building: Theory and Application*, Frankfurt am Main.
Wu,Q. *et al.* (2007) DNA methylation profiling of ovarian carcinomas and their in vitro models identifies HOXA9, HOXB5, SCGB3A1, and CRABP1 as novel targets. *Mol. Cancer*, **6**, 45.
Yang,H. *et al.* (2008) MicroRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. *Cancer Res.*, **68**, 425–433.
Yu,S. *et al.* (2010) L2-norm multiple kernel learning and its application to biomedical data fusion. *BMC Bioinformatics*, **11**, 309.
Yuan,X. *et al.* (2009) Clustered microRNAs' coordination in regulating protein-protein interaction network. *BMC Syst. Biol.*, **3**, 65.
Yuan,Z.Q. *et al.* (2000) Frequent activation of AKT2 and induction of apoptosis by inhibition of phosphoinositide-3-OH kinase/Akt pathway in human ovarian cancer. *Oncogene*, **19**, 2324–2330.
Zhang,W. *et al.* (2010) A bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Comput. Biol.*, **6**, e1000642.