Bioinformatics, 32(11), 2016, 1724–1732
doi: 10.1093/bioinformatics/btw059
Advance Access Publication Date: 1 February 2016
Original Paper

OXFORD

Gene expression

Integrative analysis for identifying joint modular patterns of gene-expression and drug-response data

Jinyu Chen and Shihua Zhang*

National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on 25 September 2015; revised on 8 January 2016; accepted on 27 January 2016

Abstract

Motivation: The underlying relationship between genomic factors and the response of diverse cancer drugs still remains unclear. A number of studies showed that the heterogeneous responses to anticancer treatments of patients were partly associated with their specific changes in gene expression and somatic alterations. The emerging large-scale pharmacogenomic data provide us valuable opportunities to improve existing therapies or to guide early-phase clinical trials of compounds under development. However, how to identify the underlying combinatorial patterns among pharmacogenomics data are still a challenging issue.

Results: In this study, we adopted a sparse network-regularized partial least square (SNPLS) method to identify joint modular patterns using large-scale pairwise gene-expression and drug-response data. We incorporated a molecular network to the (sparse) partial least square model to improve the module accuracy via a network-based penalty. We first demonstrated the effectiveness of SNPLS using a set of simulation data and compared it with two typical methods. Further, we applied it to gene expression profiles for 13 321 genes and pharmacological profiles for 98 anticancer drugs across 641 cancer cell lines consisting of diverse types of human cancers. We identified 20 gene-drug co-modules, each of which consists of 30 cell lines, 137 genes and 2 drugs on average. The majority of identified co-modules have significantly functional implications and coordinated gene-drug associations. The modular analysis here provided us new insights into the molecular mechanisms of how drugs act and suggested new drug targets for therapy of certain types of cancers.

Availability and implementation: A matlab package of SNPLS is available at http://page.amss.ac.cn/shihua.zhang/

Contact: zsh@amss.ac.cn

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

1 Introduction

The increasing amount of available high-throughput data sets at both levels of genomic data and drug response data provide us the opportunities for large-scale integrative analysis by computational methods (Barretina *et al.*, 2012; Garnett *et al.*, 2012; Lamb *et al.*, 2006; Shoemaker, 2006). This situation also enables us to study the underlying mechanisms of drug actions from the perspective of gene regulation. In general, drugs function in human body by binding to their targets or altering their targets activity to perturb biological systems (Drews, 2000; Hopkins and Groom, 2002; Penrod *et al.*, 2011; Zhao *et al.*, 2011). Previous studies suggested that 'one-drug-one-target' therapies cannot effectively treat complex diseases like

cancers which are caused by complex biological processes (Csermely *et al.*, 2005; Lu *et al.*, 2012; Medina-Franco *et al.*, 2013). In other words, drug molecules often interact with multiple targets, known as polypharmacology (Hopkins, 2008; Paolini *et al.*, 2006; Reddy and Zhang, 2013). Additionally, the same mechanism of action or target is shared by more than one drug (Takigawa *et al.*, 2011; Zhao *et al.*, 2011, 2014). Actually, in clinical practice, some drug combinations were adopted as valuable and promising therapies, such as thiazide diuretics and angiotensin-converting enzyme inhibitors for hypertension (Stanton and Reid, 2002), glyburide and metformin for type 2 diabetes (Bokhari *et al.*, 2003), saracatinib and trastuzumab for breast cancer (Zhang *et al.*, 2011,b). The multiple-to-multiple relationships between drugs and their targets imply that it is valuable to discover combinatorial gene-drug patterns to gain novel insights into molecular mechanisms and examine new drug targets for therapy.

The NCI-60 project employed an ensemble of 60 human cancer cell lines to screen over 100 000 chemical compounds and natural products, which greatly facilitated the pharmacological studies (Shoemaker, 2006). However, this project only used 60 cell lines, which limited further deep exploration. Fortunately, two large-scale pharmacogenomic studies-Cancer Cell Line Encyclopedia (Barretina *et al.*, 2012) and Cancer Genome Project (CGP) (Garnett *et al.*, 2012), published diverse types of genomic data such as gene expression, chromosomal copy number variation, mutation and pharmacological data across hundreds of cell lines, which provided valuable resources to reveal gene-drug associations. Both studies adopted a multivariate variable selection technique, the elastic net method, to systematically discover genomic markers of drug sensitivity in cancer cell lines. However, they only focused on uncovering genomic predictors of each drug independently and failed to determine coherent gene-drug patterns.

In recent years, a number of integrative methods were developed for identifying combinatorial patterns in multiple data sets for different purpose. For example, Chen *et al.* (2012) and Ma *et al.* (2014) proposed singular value decomposition-based statistical models to study the regulatory relationship between multi-dimensional predictors and responses. Zhang *et al.* (2012) developed a joint non-negative matrix factorization (NMF) framework and Li *et al.* (2012) introduced a sparse Multi-Block Partial Least Squares (sMBPLS) regression method to simultaneously analyze multiple types of genomic data to identify multi-dimensional regulatory modules. In addition, network structure such as pathways and gene-gene interactions with respect to input variables plays a complementary role in the integrative analysis (Zhang *et al.*, 2007). For example, Li and Li (2008, 2010) developed a network-constrained regularization procedure to analyze genomic data; Ma and Kosorok (2009) and Qiu *et al.* (2010) drew more attention towards pathway-based methods to identify differential gene pathways; Zhang *et al.* (2011a,b) adapted a network-regularized joint NMF method to discover miRNA-gene regulatory modules; Liu *et al.* (2012) designed a sparse group Laplacian shrinkage method to integrate multiple cancer prognosis data sets to select gene markers.

Particularly, Kutalik *et al.* (2008) developed the Ping-Pong Algorithm (PPA) to identify gene-drug co-modules by combining the gene-expression and drug-response data from NCI-60. However, this method tends to identify co-modules relating to a very limited number of cell lines (e.g. about 800 of all 859 identified co-modules only cover one or two cell lines), which is unreasonable and inconsistent with the definition of a co-module. In addition, the sizes of these co-modules are very large, majority of which contain thousands of genes and hundreds of drugs. This makes these co-modules impractical in clinical trials and leads to vast amounts of redundant information. Moreover, prior knowledge on gene interactions was not considered in this study, which could provide valuable combinational signals to improve the module discovery accuracy.

In this study, we adopted a sparse network-regularized partial least square (SNPLS) method to identify combinatorial gene-drug co-modules by integrating gene expression and drug response data across a set of cell lines as well as a gene interaction network (Fig. 1). The standard partial least square (PLS) (Boulesteix and Strimmer, 2007; Gelady and Kowalski, 1986; Rosipal and Kramer, 2006) is a class of methods for investigating the relations between two sets of observed variables by means of maximizing the covariance between their corresponding latent variables. However, it doesn't perform variable selection for high-dimensional phamacogenomic data which makes the results lack of biological interpretability. Hence, a few types of sparse PLS (SPLS) methods (Chun and Keles, 2010; Lê Cao *et al.*, 2008; Li *et al.*, 2012) were applied to the genomic data. To our knowledge, there was yet no study to incorporate network structure into the SPLS framework.

Here, we first proposed a SNPLS model to incorporate a gene interaction network. Moreover, we obtained the next co-module by subtracting the signals of a former one from the data matrices which can overcome the redundancy problem of PPA to some extent. To demonstrate its effectiveness, we applied SNPLS to a set of simulated data and compared it with two typical methods: SPLS and PPA. The results showed that SNPLS performed significantly better than SPLS and PPA in terms of specificity and sensitivity. We also applied SNPLS to a biological data set consisting of gene expression profiles of 13 321 genes and drug response data of 98 anticancer drugs across 641 cell lines derived from CGP (Garnett *et al.*, 2012). We identified 20 gene-drug co-modules and majority of them are significantly related to known functions, cancers, and coordinated gene-drug

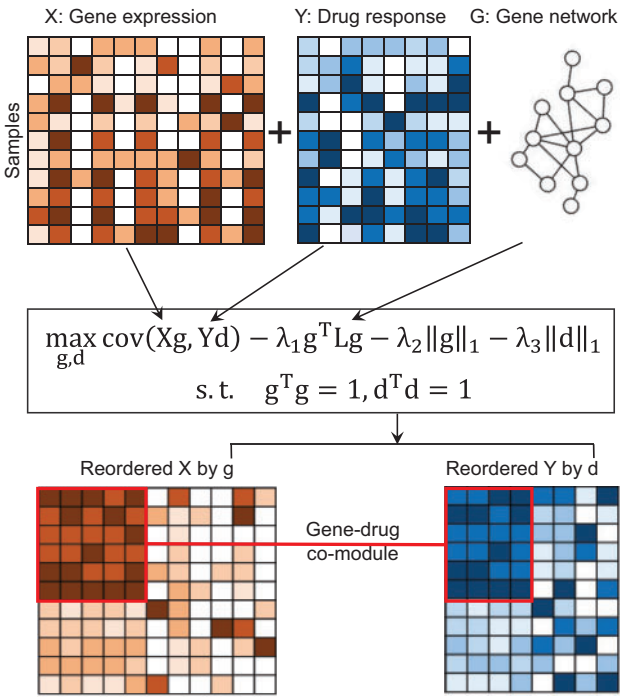

Fig. 1. Overview of the SNPLS for identifying gene-drug co-modules. A co-module is a subset of genes and drugs exhibiting similar profiles across a subset of samples determined by the weight variables *g* and *d* of SNPLS applied in pairwise gene expression data *X* and drug response data *Y*. A gene interaction network *G* is incorporated to enhance the modular characteristics

associations. These gene-drug patterns provide us insights into potential drug targets and drug combinations for cancer therapy.

2 Materials and methods

2.1 Data and preprocessing

We downloaded a large-scale pharmacogenomic dataset including pairwise gene expression data and drug response data from the CGP on the Genomics of Drug Sensitivity in Cancer website (http://www.cancerrxgene.org/downloads/) (Garnett *et al.*, 2012). We removed 40 drugs and 13 samples that only have a limited number of values across samples and drugs, respectively. Finally, we centered the gene expression and drug response data across samples, and obtained a normalized gene expression dataset of 13 321 genes and a drug response dataset of 98 drugs across 641 samples, which were represented in two matrices $X \in \mathbb{R}^{p\times n}$ and $Y \in \mathbb{R}^{p\times m}$ respectively (p, n, m represent the number of samples, genes and drugs, respectively).

We downloaded a gene interaction network from the PathwayCommons database (http://www.pathwaycommons.org) (Cerami *et al.*, 2011), which was once used by Hofree *et al.* (2013) and others. This network is compiled by integrating a variety of sources about gene or protein interactions including BioGrid, HPRD, IntAct and the NCI set of cancer specific pathways. It consists of 14 355 genes or proteins and 507 757 interactions. We filtered the genes which were absent from our genes expression data. For any gene that was in the input gene expression data X but not in this network, we added it to the network as an isolated node. Finally, we obtained a gene–gene interaction network with 13 321 genes and 262 462 interactions, which were denoted as a graph $G = (V, E)$, where $V = \{v_1, v_2, \cdots, v_n\}$ is the set of nodes (genes) and $E = \{e_1, e_2, \cdots, e_I\}$ is the set of undirected edges (interactions) in graph G.

2.2 Problem formulation

In this study, we aimed to identify coherent gene-drug co-modules via the SNPLS by integrating the gene expression data X and drug response data Y across a set of cell lines as well as a gene interaction network A. The standard PLS models the relations between two sets of variables by the covariance function, i.e.

$$\max_{g,d} \operatorname{cov}(Xg, Yd) \tag{1}$$
$$\text{s.t.} \quad g^{\mathrm{T}}g = 1, d^{\mathrm{T}}d = 1.$$

Here, let's denote $u = Xg$, $v = Yd$ as the latent variables which are the linear combination of n and m variables corresponding to X and Y, respectively, g and d are also named as weight vectors. This objective function indicates that the similarity between small blocks of X and Y is measured by the covariance of the two latent variables u and v. We can discover the corresponding blocks in X and Y which have similar or coherent patterns based on the absolute values of the optimal solutions of g and d.

However, this method doesn't perform variable selection and is likely to result in poor interpretation. Chun and Keles (2010) suggested to impose a sparsity penalty to the weight variables g and d and developed a SPLS regression method, which was also extended for multiple genomics data analysis recently (Li *et al.*, 2012).

$$\max_{g,d} \operatorname{cov}(Xg, Yd) - \lambda_1||g||_1 - \lambda_2||d||_1 \tag{2}$$
$$\text{s.t.} \quad g^{\mathrm{T}}g = 1, d^{\mathrm{T}}d = 1.$$

The SPLS produces sparse g and d, which can be used for selecting effective variables with better biological interpretation.

Furthermore, prior knowledge on gene interactions is very useful and valuable to decipher the modular patterns among genes. Network-based penalty has been adopted for many different applications. For example, Li and Li (2008, 2010) and Liu *et al.* (2012) developed a network-constrained regularization procedure to realize variable selection and regression analysis for genomic data. In these studies, the network-based penalty is defined in the same way as a quadratic form of the Laplacian matrix associated with the genes interaction network. Zhang *et al.* (2011a) utilized predicted miRNA-gene interactions and gene-gene interactions to define network-regularized constraints for discovering the miRNA-gene regulatory modules. Chen *et al.* (2013) adopted the phylogenetic relationships among the taxa to construct Laplacian penalty function as before to study the association between nutrient intake and human gut microbiome composition. Although the network-based penalty functions are not completely identical, they all enforce the tightly connected nodes (genes) in the network tend to have more similar coefficients. Inspired by this technique, we introduced a SNPLS model to achieve our goal. Specifically, it can be formulated as follows:

$$\max_{g,d} \operatorname{cov}(Xg, Yd) - \lambda_1 g^{\mathrm{T}}Lg - \lambda_2||g||_1 - \lambda_3||d||_1 \tag{3}$$
$$\text{s.t.} \quad g^{\mathrm{T}}g = 1, d^{\mathrm{T}}d = 1.$$

where $cov(u, v)$ is the covariance of u and v ($u, v \in \mathbb{R}^p$), which approximates to $(u^{\mathrm{T}}v)/p$ if $\frac{1}{p}\sum_{i=1}^{p} u_i = \frac{1}{p}\sum_{i=1}^{p} v_i = 0$, and L is the symmetric normalized Laplacian matrix defined as

$$L = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}},$$

where $A = (a_{ij})_{n\times n}$ is the binary or weighted adjacency matrix of the gene-gene interaction network G, a_{ij} equals to 1 or a value ranging from 0 to 1, if gene i and j were linked in the network and $a_{ij} = 0$, otherwise; $D = (d_{ij})_{n\times n}$ is the degree matrix of graph G, where $d_{ii} = \sum_{j=1}^{n} a_{ij}$, and $d_{ij} = 0$, for $i \neq j$. The tuning parameters $\lambda_1, \lambda_2, \lambda_3$ control the amount of regularization for smoothness and sparsity. When $\lambda_1 = 0$, the model reduces to the SPLS.

If the matrices X and Y are normalized such that each column of X and Y is centered, the problem defined in Equation (3) is equivalent to

$$\max_{g,d} \frac{1}{p} g^{\mathrm{T}}X^{\mathrm{T}}Yd - \lambda_1 \sum\nolimits_{1\le i<j\le n} a_{ij}\left(\frac{g_i}{\sqrt{l_i}} - \frac{g_j}{\sqrt{l_j}}\right)^2 \tag{4}$$
$$-\lambda_2||g||_1 - \lambda_3||d||_1$$
$$\text{s.t.} \quad g^{\mathrm{T}}g = 1, d^{\mathrm{T}}d = 1.$$

The objective function consists of four key terms. The first one describes the covariance between the hidden components based on the genes expression data X and drugs response data Y. The second one captures the key prior knowledge which makes the connected genes in the network likely to be placed in the same co-modules. The last two ones enforce the sparsity of the variables g and d such that the results will have better biological interpretation.

2.3 The SNPLS algorithm

Obviously, the objective function in Equations (3) or (4) is not convex with respect to g and d. Thus, it is hard to obtain the global optimal solution by means of classical algorithms. In the following, we adopted a coordinate descent algorithm to find local maximum of this problem by updating variables g and d alternately (Supplementary Materials). For parameter selection, we have

Algorithm for SNPLS:

Step 1: Initialize g with the solution of Equation (1) and $u = Xg$.

Step 2: Update d and g alternately.

(1) Fix variable g and update variable d with

$$d \leftarrow \text{sign}\left(\frac{1}{p}Y^{\mathrm{T}}u\right)\left(\left|\frac{1}{p}Y^{\mathrm{T}}u\right| - \lambda_3\right)_+, \text{ norm } d.$$

$$v = Yd.$$

where

$$(x)_+ = \begin{cases} x, & \text{if } x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

(2) Fix variable d and update variable g with

$$g_j \leftarrow \frac{\text{sign}(z)(|z| - \lambda_2)_+}{2(\lambda_1 + \delta)}, \text{for } j = 1, 2, ..., n; \text{norm } g.$$

$$u = Xg.$$

where $z = t_j + 2\lambda_1 \sum_{i=1}^{n} \frac{a_{ij}g_i}{\sqrt{l_i l_j}}$, and t_j is the jth element of vector $t = \frac{1}{p}(X^{\mathrm{T}}Yd) = \frac{1}{p}(X^{\mathrm{T}}v)$. δ is a positive parameter for the constraint $g^{\mathrm{T}}g = 1$.

Step 3: Repeat **Step 2** until convergence of u

adopted a 5-fold cross-validation procedure to achieve it (Supplementary Materials). We can easily find that the computational complexity of one SNPLS iteration is $O(pm + pn + n^2)$. To speed up the convergence of this algorithm, we employed the solution of standard PLS as the initial solution of current algorithm. We provided this algorithm in the framework above. We implemented it in MATLAB R2013a as an user-friendly package (http://page.amss.ac.cn/shihua.zhang/).

2.4 Determining co-modules

The weight vectors g and d produced by the above algorithm will guide us to identify gene-drug co-modules. The main idea is to select the gene and drug variables with relatively large absolute values of weight variables g and d as the members of gene-drug co-modules. Specifically, we calculated the z-scores of g and d in the following way:

$$z_i = \frac{|x_i| - \bar{x}}{S_x} \tag{5}$$

where

$$\bar{x} = \frac{1}{n}\sum |x_i|, S_x^2 = \frac{1}{n-1}\sum (|x_i| - \bar{x})^2.$$

Based on this transformation, we obtained two vectors g^* and d^* and determined the co-module members if $g^*(i)$ (or $d^*(j)$) was larger than the given threshold T. Meanwhile, we updated g and d by setting the values of g_i and d_j, which were not selected as the members of a co-module be zeros. Moreover, we preferred to identify the gene-drug co-modules across certain subset of samples. To achieve this goal, we considered the latent vectors $u = Xg$ and $v = Yd$ and normalized u and v, such that $u^* = \frac{u}{||u||_2}$ and $v^* = \frac{v}{||v||_2}$. We applied formula (5) to the vector $(u^* + v^*)$, chose samples whose scores were larger than a given threshold T, and updated u and v as what we did to g and d. We set $T = 3$ for selecting genes and drugs, and $T = 2$ for selecting samples.

We obtained the first gene-drug co-module after running the SNPLS algorithm. Next, we subtracted the signal of current co-module from the input data as follows:

$$X := X - up^{\mathrm{T}}, \quad p = \frac{X^{\mathrm{T}}u}{u^{\mathrm{T}}u};$$

$$Y := Y - vq^{\mathrm{T}}, \quad q = \frac{Y^{\mathrm{T}}v}{v^{\mathrm{T}}v};$$

Then, we could continue to apply SNPLS algorithm to the updated input data X and Y to identify the next gene-drug co-module.

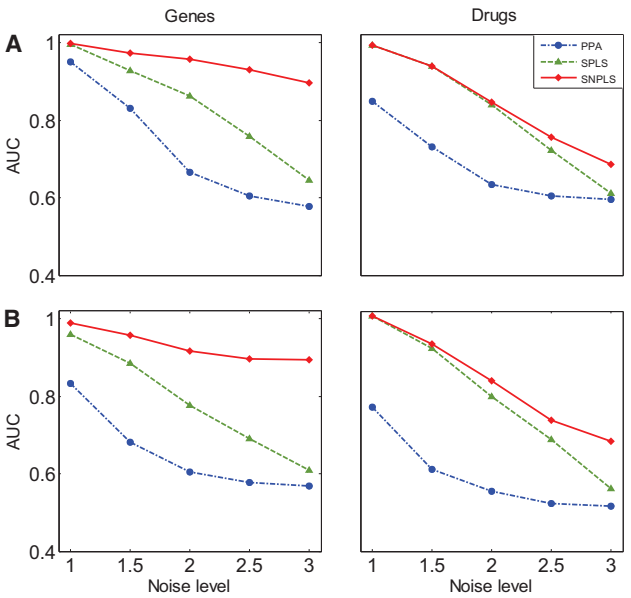

Fig. 2. Performance comparison of SNPLS, SPLS and PPA in terms of AUC in two simulation scenarios. In the two scenarios, the variables of simulated data X were **(A)** independent and **(B)** correlated, respectively. In each scenario, the average AUCs of 50 realizations on the simulated data with respect to different noise levels were shown

3 Results

3.1 Simulation study

In this section, we evaluated the performance of SNPLS and compared it with SPLS (Chun and Keles, 2010) and PPA (Kutalik *et al.*, 2008) by applying them to a set of simulated data. We used the area under receiver operating characteristic curves (AUC) as a measure to evaluate the performance of different methods. We ran each method on simulated data and repeated this procedure for 50 times (Supplementary Materials). We calculated the average AUCs of 50 realizations about genes and drugs, respectively.

We can clearly see that SNPLS always performs better than SPLS and PPA whenever the variables of simulated data X are independent (Fig. 2A) or correlated (Fig. 2B). In particular, when the data noises increase, the AUC values of SNPLS decrease much slower than the other two approaches regarding to both genes and drugs. For the gene module discovery, SNPLS even has much better

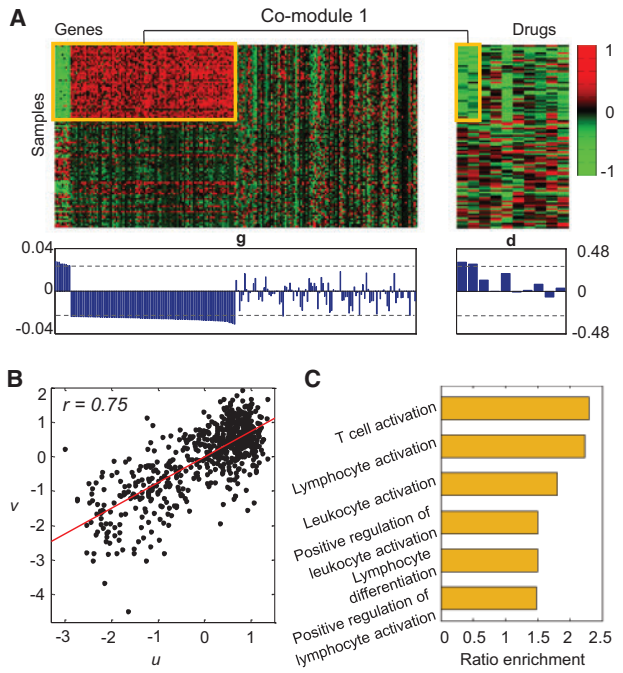

Fig. 3. Illustration of co-module 1. **(A)** Heat map of co-module 1 consisting of 104 genes and 2 drugs across 42 samples (squared boxes). We extended the heat map to cover more variables by randomly selecting 104 genes, 8 drugs and 58 samples for contrasting. We reordered the genes, drugs as well as samples in this co-module circled in squared boxes by the descending order of the weight variables g and d as shown with bar plots below the heat map. The horizontal lines over the bar plots indicate the thresholds used for selecting the co-module genes and drugs. **(B)** The scatter plot for normalized latent variables u and v of co-module 1 with Pearson correlation coefficient $r = 0.75$ indicating that they are highly correlated. **(C)** Top biological terms enriched by the genes of co-module 1. The ratio enrichment indicates the functional significance of a gene module with $-\log(P\text{-value})$ (Benjamini-corrected P-value). Similar setting was used in Figure 4

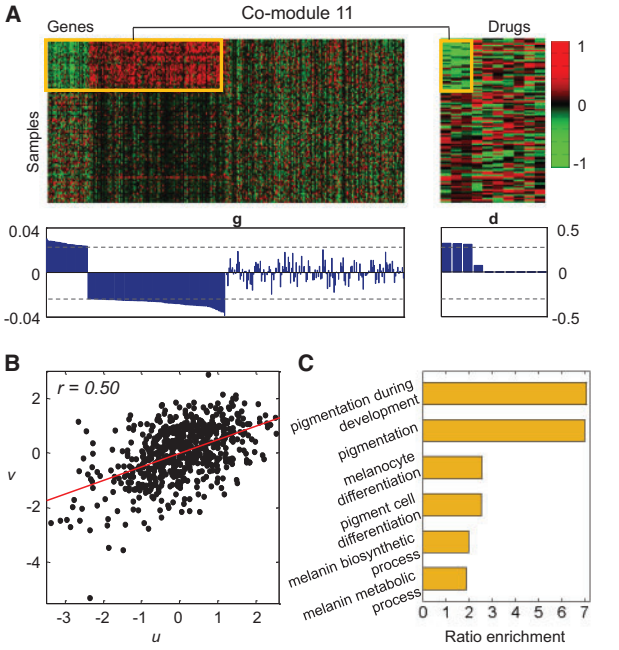

Fig. 4. Illustration of co-module 11 consisting of 160 genes and 3 drugs across 33 samples (squared boxes). We extended the heat map to cover more variables by randomly selecting 160 genes, 7 drugs and 67 samples for contrasting

performance, suggesting that the prior network knowledge is very useful to discover the underlying modular signal. Thus, SNPLS is more promising compared with SPLS and PPA for practical biological applications.

3.2 Identifying co-modules in a pharmacogenomics dataset

We applied SNPLS to a large-scale pharmacogenomic dataset derived from CGP (Garnett *et al.*, 2012) and obtained 20 gene-drug co-modules. The detailed lists for each co-module can be found in Supplementary Table S1. The 20 gene-drug co-modules cover about 30 cell lines, 137 genes and 2 drugs on average. We found that each of the three components occurred in about one to three co-modules (Supplementary Materials), indicating that the 20 co-modules are different with each other. We also used the hypergeometric test to assess the degree of overlap of any two co-modules (Supplementary Materials). As a result, only one pair of co-modules has significant overlap (FDR <0.05). Thus, almost all of the 20 co-modules are distinct. The co-module member genes and drugs also exhibit highly similar patterns across the same subset of samples (e.g., co-module 1 and 11 in Figs 3A and 4A and Supplementary Materials, respectively).

When compared with random generated ones, our co-modules demonstrate high degree of (anti-) correlation between genes and drugs across the same subset of samples. In addition, we computed the Pearson correlation coefficients between the latent variables u and v for all the 20 co-modules and they all show significantly high correlations (see two examples in Figs 3B and 4B, respectively).

3.3 Co-modules reveal distinct biological relevance

To evaluate the biological relevance of the 20 co-modules, we performed systematic enrichment analysis for genes, drugs and cell lines of each co-module with known knowledge, and investigated the drug targets, drug effector pathways or biological processes and their connections to known tumors. Using DAVID tools (https://david.ncifcrf.gov/) for gene enrichment analysis (Huang *et al.*, 2009), we found that 12 (60%) and 11 (55%) of the gene modules respectively have at least one significant GO biological process and KEGG pathway (Benjamini-corrected P value < 0.05). In total, these modules are enriched in 193 distinct GO biological processes and 29 KEGG pathways. Among them, the most frequent biological processes are biological adhesion, chromosome organization, cell cycle and mitosis. The most frequently enriched KEGG pathways are focal adhesion and cell cycle. Adhesion-related processes play important roles in cancer progression and metastasis. For example, cell adhesion to the extracellular matrix (ECM) provides the tractions for cell motility and invasion, which affects the metastasis of cancer cells. Thus, the integrins, cell surface receptors to mediate cell adhesion to ECM, have been key targets of cancer therapy for test (Desgrosellier and Cheresh, 2010). Chromosome organization, cell cycle and mitosis are three closely related processes, which all occur in cell division and proliferation. During these processes, there appear to be some alterations and modifications, such as genetic variations and epigenetic events, which are likely to cause cancer cell initiation and progression (Veltri and Christudass, 2014). We summarized the key messages for seven co-modules (Table 1) and all co-modules (Supplementary Table S2). Besides, the overlap of gene sets for each pair of co-modules is provided in Supplementary Table S3.

Table 1. Biological function analysis of seven co-modules

ID	#G	#D	#S	Top enriched GO biological process	Drug name	Drug target	Drug effector	Tumor type
1	104	2	42	T-cell activation; positive regulation of leukocyte activation; translational elongation; lymphocyte activation; positive regulation of cell activation	Methotrexate	DHFR	Replication, transcription	Lymphoblastic T-cell leukaemia; lymphoblastic leukemia; Burkitt lymphoma; AML; CML
					ATRA	Retinoic acid and retinoid X receptor agonist	Transcription	
2	110	2	29	mitotic sister chromatid segregation; M phase of mitotic cell cycle; microtubule cytoskeleton organization; mitotic cell cycle; M phase	RDEA119	MEK1/2	ERK signaling	Small cell lung carcinoma
					Docetaxel	Microtubules	Cytoskeleton, mitosis	
6	134	2	24	dorsal/ventral pattern formation; regulation of intracellular transport; regulation of protein import into nucleus; regulation of nucleocytoplasmic transport; regulation of establishment of protein localization	Camptothecin	TOP1	Replication and repair	Ewings sarcoma; Rhabdomyosarcoma
					AZD-2281	PARP1/2	DNA repair	
7	142	2	31	blood vessel development; MAPKKK cascade; vasculature development; cell activation; antigen processing and presentation of peptide or polysaccharide antigen via MHC class II	Metformin	AMPK agonist	AMPK, metabolism	Myeloma; B cell lymphoma
					ATRA	Retinoic acid and retinoid X receptor agonist	Transcription	
8	138	1	29	peroxisome organization; cell adhesion; biological adhesion	BX-795	TBK1, PDK1, IKK, AURKB/C	Mitosis, NFkappB, PI3K/MTOR	Breast
11	160	3	33	melanin metabolic process; melanocyte differentiation; melanin biosynthetic process; pigmentation; pigmentation during development	CI-1040	MEK1/2	ERK signaling	Malignant melanoma
					PLX4720	BRAF	ERK signaling	
					SB590885	BRAF	ERK signaling	
17	178	6	30	ectoderm development; epidermis development; negative regulation of peptidase activity; regulation of cell proliferation; regulation of peptidase activity	Gefitinib	EGFR	ERK signaling, PI3K/MTOR	Upper aerodigestive tract; pancreas
					RDEA119	MEK1/2	ERK signaling	
					CI-1040	MEK1/2	ERK signaling	
					BIBW2992	EGFR, ERBB2	ERK signaling, PI3K/MTOR	
					PD-0325901	MEK1/2	ERK signaling	
					AZD6244	MEK1/2	ERK signaling	

ID, co-module ID; #G/#D/#S, number of genes/drugs/samples; drug target, a molecule to which a given drug binds; drug effector, the biological process or pathway which a given drug has an effect on; tumor type: the enriched tumor types in the samples.

For the drugs in each co-module, we analyzed their targets and effector pathways or biological processes. 12 of the 20 co-modules include more than one drug. Among 6 (50%) of these 12 drug modules, their drug members share the same targets or effector pathways (P-value = 0.0008 by permutation test). For example, the three drugs (CI-1040, PLX4720 and SB590885) in co-module 11 all target *ERK* signaling pathway. Moreover, for samples in the 20 co-modules, we tested whether they tended to belong to the same or similar tumor types. As a result, 13 (65%) modules are enriched in certain tumor types (FDR < 0.05, hypergeometric test). For example, the co-module 1 is enriched in several types of blood diseases including lymphoblastic T-cell leukemia, lymphoblastic leukemia, acute myelogenous leukemia (AML), Burkitt lymphoma, chronic myelogenous leukemia (CML) and so on.

3.4 Co-modules reveal significant drug-gene connections

We found that the co-modules reveal significant drug-gene connections from different angles (Table 1 and Fig. 5). We took co-module 1, 6 and 11 as examples. The co-module 1 consists of 104 member genes with a significant number (16) of genes in the human cancer genes census (FDR = 2.1456×10^{-6}, hypergeometric test), and a significant number (6) of genes (*BCOR, BLM, IKZF1, PTPRC, SEPT6, SFRS2*) relating to leukemia (Futreal *et al.*, 2004). Moreover, 8 of 12 enriched GO biological processes are about leukocyte, lymphocyte or T cell, which are all closely related to leukemia (Fig. 3C). Surprisingly, the sample-enriched tumor types exactly refer to this kind of disease, including lymphoblastic T-cell leukemia, lymphoblastic leukemia, AML, Burkitt lymphoma and CML, indicating the distinct biological relevance of the identified co-modules. On the other hand, its two member drugs both have effects on transcription, which is a key part of cell activation. This is consistent with the enriched biological functions: cell activation, translational elongation and ribosome pathway which play a leading role during transcription. As to these two drugs, methotrexate is an antineoplastic antimetabolite with immunosuppressant properties (Knox *et al.*, 2011; Law *et al.*, 2014; Wishart *et al.*, 2006, 2008). It competitively inhibits dihydrofolate reductase (DHFR), an enzyme participating in the tetrahydrofolate synthesis, which is necessary for synthesis of purines, thymidylate and several amino acids (Pajagopalan *et al.*, 2002). Therefore, methotrexate is able to inhibit cellular replication and was approved by the Food and Drug Administration for acute lymphoblastic leukemia. Another drug, ATRA, also known as tretinoin, is an important regulator of cell

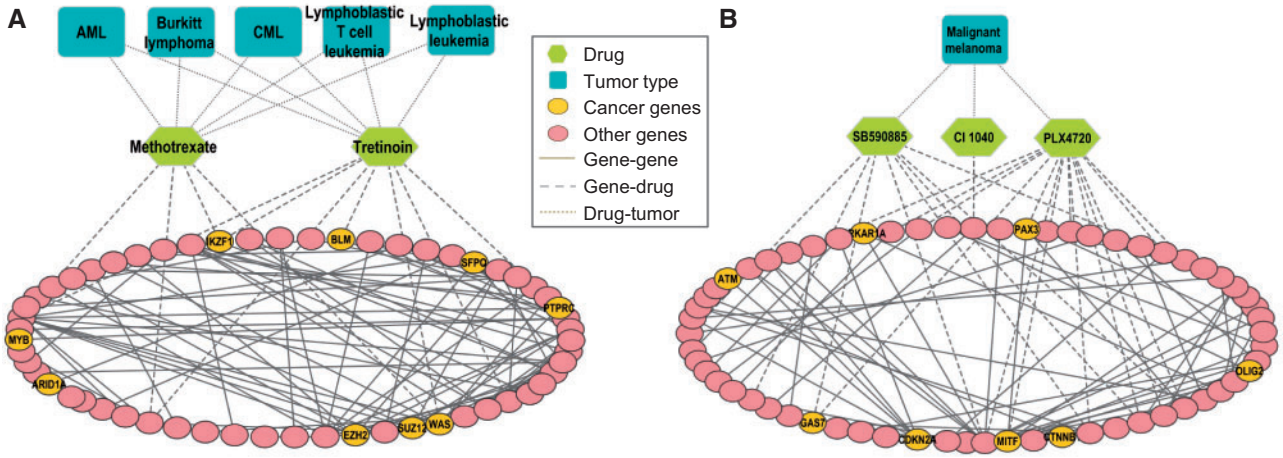

Fig. 5. Two networks of three levels of genes, drugs and tumor types in co-module 1 **(A)** and 11 **(B)**

reproduction, proliferation and differentiation. It was used in the treatment of acute promyelocytic leukemia (Castaigne *et al.*, 1990; Huang *et al.*, 1988; Law *et al.*, 2014; Sanz, 2006). Although these two anticancer drugs were not reported to be used to treat the sample enriched diseases together, the high degree of correlated structure with genes enriched in crucial biological functions and the similar drug effectors indicated their functional similarity for leukemia. We constructed a network of three levels for co-module 1 (Fig. 5A and Supplementary Materials) to demonstrate their close connections among genes, drugs and sample-enriched tumor types. The drugs directly link to three cancer genes (*BLM, IKZF1, WAS*), which are all associated with leukemia or lymphoma (Warde-Farley *et al.*, 2010). Moreover, the drugs are both connected with *VAV1*, proteins encoded by what are important in hematopoiesis, playing a role in T- and B-cell development and activation (Safran *et al.*, 2010). In the network, *VAV1* is interacting with four cancer genes: *EZH2* (related to diffuse large B-cell lymphoma), *PTPRC* (related to T-cell acute lymphoblastic leukemia), *WAS* (related to lymphoma) and *SUZ12* (related to endometrial stromal tumor) (Futreal *et al.*, 2004), indicating the potential of this gene as a new drug target for treatment of lymphoma-related diseases.

The co-module 6 consists of 134 genes with a significant number (12) of cancer genes (FDR = 0.0041, hypergeometric test). The genes of this co-module are mainly involved in intracellular substance transport which plays a leading role in DNA repair and replication. The two drugs exactly target these biological processes (Table 1). Moreover, the 24 samples of this co-module are enriched in Ewings sarcoma and rhabdomyosarcoma (putative Ewings). One drug AZD-2281 in this co-module, also named olaparib, is an inhibitor of poly ADP ribose polymerase (*PARP*). Brenner *et al.* (2012) reported that *PARP-1* inhibition could be used as a targeted strategy to treat Ewings sarcoma. On the other hand, Lee *et al.* (2013) proposed that combing *PARP-1* inhibition and radiation in Ewings sarcoma resulted in lethal DNA damage, which increased apoptosis and cell death and finally blocked the development of Ewings sarcoma. More interestingly, it was reported that the combination of olaparib (AZD-2281) and camptothecin could be promising to improve the clinical therapy for Ewings sarcoma comparing with using olaparib alone (Miura *et al.*, 2012).

The last example, co-module 11 exhibits distinct biological relevance with malignant melanoma in terms of genes, drugs and samples respectively. First, the 33 samples of this co-module are enriched in this tumor type. Second, the 160 genes of this co-module are enriched in melanin or pigment (Fig. 4C) with two melanoma oncogenes–*CDKN2A* and *MITF* (Futreal *et al.*, 2004). Third, the three drugs of this co-module affect the same pathway–*ERK* signaling pathway, which plays a central role in the mediating growth-promoting signals for a diverse group of upstream stimuli (Allen *et al.*, 2003). These three drugs target two genes: one is *BRAF* which has been an attractive target for melanoma drug development (Villanueva *et al.*, 2010); the other is *MEK1/2* which is one of the key components in the *ERK* signaling pathway. The inhibitors of *MEK* could effectively block the phosphorylation of *ERK* and continuous signal transduction through *ERK* signaling pathway, thereby they have important clinical benefit in the treatment of cancers (Allen *et al.*, 2003). Moreover, a V600E mutation of the *BRAF* serine/threonine kinase (S/T kinase) was found occurred in >50% of all melanoma (Puzanov and Flaherty, 2010). Combination of *BRAF* and *MEK* inhibition in melanoma with *BRAF* V600 mutation, comparing with *BRAF* inhibition alone, can delay the emergence of resistance and reduce toxic effects in patients, thereby improves the rate of progression-free survival (Flaherty, 2012; Georgina *et al.*, 2014). Interestingly, for co-module 11, the most significantly enriched mutation type is BRAF V600E (*P*-value = 2.4e-18) (Supplementary Table S4), which indicates the closely relationships between drugs and melanoma samples in this co-module. We also demonstrated the close relationships among these components in a three-level network (Fig. 5B). Particularly for drug SB590885 and PLX4720, they tend to link to the same genes. All these observations demonstrated the high connections of genes, drugs and sample-enriched tumor types in these co-modules, suggesting the effectiveness of SNPLS for discovering biologically relevant co-modules.

3.5 Comparison with other methods

To demonstrate the effectiveness of SNPLS, we also applied SPLS (Chun and Keles, 2010) and PPA (Kutalik *et al.*, 2008) to the CGP data and identified 20 co-modules, respectively. The global distributions for the number of cell lines, genes and drugs of 20 co-modules produced by each method are very different (Supplementary Materials). Especially for PPA, its co-modules contain a very large number of genes and drugs, but relatively few cell lines comparing to those of SPLS and SNPLS (i.e. on average 5 cell lines, 1281 genes

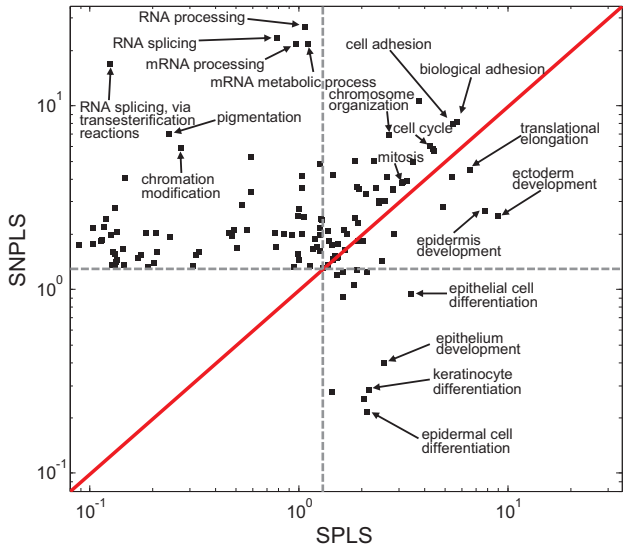

Fig. 6. Comparison of all the enriched GO biological processes of gene modules detected by SNPLS and SPLS. For each GO biological process, we computed enrichment scores (−log10(*q*-value) with Benjamini-corrected *q*-values) and the highest scores among all modules were taken as the final score of this GO biological process for each method. The scores for SNPLS were plotted against those of SPLS. Dash lines along the horizontal and vertical directions both indicate the significance threshold. The points in the top-left part represent the significantly enriched GO biological process exclusively by SNPLS co-modules, whereas the points in the bottom-right part are ones exclusively by SPLS co-modules. A set of representative terms are shown. Points above the central diagonal line represent terms that are more significantly enriched using SNPLS than SPLS, indicating that SNPLS can identify more biological relevant gene modules than SPLS by incorporating the prior gene network

and 23 drugs per co-module for PPA versus 31 cell lines, 128 genes and 2 drugs for SPLS and 30 cell lines, 137 genes and 2 drugs for SNPLS). The large size of co-modules of PPA makes it difficult to extract essential information for practical usage, and too few cell lines in each seems unreasonable. It is hard to image to treat a set of genes and drugs, which play similar or coherent tendency across only two or three samples as a joint modular pattern. Moreover, among these 20 co-modules detected by PPA, 7 pairs of co-modules have significant overlap whereas only one pair for SNPLS and two pairs for SPLS have implying co-modules detected by PPA are very redundant.

We also analyzed the interaction enrichment in each co-module of SPLS and SNPLS based on the gene-gene interaction network. 14 (70%) co-modules of SNPLS are enriched with gene interactions (FDR < 0.05), whereas only 11 (55%) co-modules of SPLS are enriched, indicating the strong biological relevance of co-modules of SNPLS than those of SPLS. Actually, 14 co-modules of SNPLS are enriched in at least one GO biological process or KEGG pathway (Benjamini-corrected q-value < 0.05) and only 11 co-modules of SPLS are. We also found that the enriched biological processes of SNPLS have more significant P-value than those of SPLS, suggesting that SNPLS indeed have improvement in identifying more biologically relevant genes (Fig. 6 and Supplementary Table S5). Moreover, we applied SNPLS and SPLS to NCI60 data with a large number of compounds and a small number of samples as used by PPA (Kutalik *et al.*, 2008) (Supplementary Materials). Both methods showed very similar performance as applied to CGP data.

4 Discussion

Deciphering the multiple-to-multiple relationships between drugs and their targets is crucial for studying the mechanisms of drug actions and developing effective treatment for patients. Meanwhile, the dramatic accumulation of large-scale genomic data and drug response data from the same cell lines provides us the unprecedented opportunities to identify gene-drug joint modular patterns to decode these relationships from the perspective of gene regulation. In this study, we developed a method SNPLS to integrate these two data as well as a gene interaction network to identify gene-drug co-modules. When compared with SPLS, SNPLS employs the network structure as prior knowledge such that genes in each co-module tend to be closely connected in the network, which makes such a co-module more biologically interpretable.

We applied SNPLS to a pairwise gene-expression and drug-response data from 641 cell lines covering a wide range of tumor types and identified 20 gene-drug co-modules. Most of these co-modules displayed significant functional connections from functional interpretation of gene, drug and cell line views. For drug members of the same co-module, they often have the same related targets, or have effects on the same biological processes and pathways in some related diseases. These observations suggested that the co-modules can help us to find new drug combinations or similarities for treatment of certain cancer or provide new drug target candidates.

We may study the following aspects to extend SNPLS in future studies. First, we used a coordinate descend algorithm to solve the objective function, which may be improved with more computationally efficient algorithms. Second, besides using a gene interaction network, we can also incorporate drug-drug similarities or interactions to improve the accuracy of module discovery. Third, SNPLS may be further adapted to consider the potential negative correlation of two genes in the network-regularized term, i.e. taking the sign of weight variable g into account. Finally, with the accumulation of abundant data in biology and pharmacology, we can extend this method to similar pairwise biological data for joint modular analysis.

Funding

This work was supported by the National Natural Science Foundation of China, No. 61379092, 61422309, 61171007 and 11131009, the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (XDB13040600), the Outstanding Young Scientist Program of CAS and the Key Laboratory of Random Complex Structures and Data Science, CAS.

Conflict of Interest: none declared.

References

Allen,L.F. *et al.* (2003) CI-1040 (PD184352), a targeted signal transduction inhibitor of MEK (MAPKK). *Semin. Oncol.*, **30**, 105–116.

Barretina,J. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.

Bokhari,S.U. *et al.* (2003) Beneficial effects of a glyburide/metformin combination preparation in type 2 diabetes mellitus. *Am. J. Med. Sci.*, **325**, 66–69.

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

Brenner,J.C. *et al.* (2012) PARP-1 inhibition as a targeted strategy to treat Ewing's sarcoma. *Cancer Res.*, **72**, 1608–1613.

Castaigne,S. *et al.* (1990) All-trans retinoic acid as a differentiation therapy for acute promyelocytic leukemia. I. Clinical results. *Blood*, **76**, 1704–1709.

Cerami,E.G. *et al.* (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Chen,J. *et al.* (2013) Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, **14**, 244–258.
Chen,K. *et al.* (2012) Reduced rank stochastic regression with a sparse singular value decomposition. *J. R. Stat. Soc. B Stat. Method.*, **74**, 203–221.
Chun,H. and Keles,S. (2010) Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B Stat. Method.*, **72**, 3–25.
Csermely,P. *et al.* (2005) The efficiency of multi-target drugs: the network approach might help drug design. *Trends Pharmacol. Sci.*, **26**, 178–182.
Desgrosellier,J. and Cheresh,D. (2010) Integrins in cancer: biological implications and therapeutic opportunities. *Nat. Rev. Cancer*, **10**, 9–22.
Drews,J. (2000) Drug discovery: a historical perspective. *Science*, **287**, 1960–1964.
Flaherty,K.T. *et al.* (2012) Combined BRAF and MEK inhibition in melanoma with BRAF V600 mutations. *N. Engl. J. Med.*, **367**, 1694–1703.
Futreal,P.A. *et al.* (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
Garnett,M.J. *et al.* (2012) Systematic identification of genomicmarkers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
Gelady,P. and Kowalski,B. (1986) Partial least square regression: a tutorial. *Anal. Chim. Acta*, **185**, 1–17.
Georgina,V.L. *et al.* (2014) Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *N. Engl. J. Med.*, **371**, 1877–1888.
Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.
Hopkins,A.L. and Groom,C.R. (2002) The druggable genome. *Nat. Rev. Drug. Discov.*, **1**, 727–730.
Hopkins,A.L. (2008) Network pharmacology: the next paradigm in drug discovery. *Nat. Chem. Biol.*, **4**, 682–690.
Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
Huang,M.E. *et al.* (1988) Use of all-trans retinoic acid in the treatment of acute promyelocytic leukemia. *Blood*, **72**, 567–572.
Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res.*, **39**, 1035–1041.
Kutalik,Z. *et al.* (2008) A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nat. Biotechnol.*, **26**, 531–539.
Lamb,J. *et al.* (2006) The Connectivity Map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, 1091–1097.
Lê Cao,K.A. *et al.* (2008) A sparse PLS for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.*, **7**, 1544–6115.
Lee,H.J. *et al.* (2013) Combining PARP-1 inhibition and radiation in Ewing sarcoma results in lethal DNA damage. *Mol. Cancer Ther.*, **12**, 2591–2600.
Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.
Li,C. and Li,H. (2010) Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl Stat*, **4**, 1498–1516.
Li,W. *et al.* (2012) Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*, **28**, 2458–2466.
Liu,J. *et al.* (2012) Incorporating network structure in integrative analysis of cancer prognosis data. *Genet. Epidemiol.*, **37**, 173–183.
Lu,J.J. *et al.* (2012) Multi-target drugs: the trend of drug research and development. *PLoS One*, **7**, e40262.
Ma,S. and Kosorok,M.R. (2009) Identification of differential gene pathways with principal component analysis. *Bioinformatics*, **25**, 882–889.
Ma,X. *et al.* (2014) Learning regulatory programs by threshold SVD regression. *Proc. Natl. Acad. Sci. USA*, **111**, 15675–15680.
Medina-Franco,J.L. *et al.* (2013) Shifting from the single- to the multitarget paradigm in drug discovery. *Drug Discov. Today*, **18**, 495–501.
Miura,K. *et al.* (2012) The combination of olaparib and camptothecin for effective radiosensitization. *Radiat Oncol.*, **7**, 62.
Pajagopalan,P.T.R. *et al.* (2002) Interaction of dihydrofolate reductase with methotrexate: Ensemble and single-molecule kinetics. *Proc. Natl. Acad. Sci. USA*, **99**, 13481–13486.
Paolini,G.V. *et al.* (2006) Global mapping of pharmacological space. *Nat. Biotechnol.*, **24**, 805–815.
Penrod,N.M. *et al.* (2011) Systems genetics for drug target discovery. *Trends Pharmacol. Sci.*, **32**, 623–630.
Puzanov,I. and Flaherty,K.T. (2010) Targeted molecular therapy in melanoma. *Semin. Cutan. Med. Surg.*, **29**, 196–201.
Qiu,Y.Q. *et al.* (2010) Detecting disease associated modules and prioritizing active genes based on high throughput data. *BMC Bioinf.*, **11**, 26.
Reddy,A.S. and Zhang,S.X. (2013) Polypharmacology: drug discovery for the future. *Expert Rev. Clin. Pharmacol.*, **6**, 41–47.
Rosipal, R. and Kramer, N. (2006) Overview and recent advances in partial least squares. In Saunders, C. *et al.* (eds), *Subspace, Latent Structure and Feature Selection*. Springer Berlin Heidelberg, Vol. 3940, pp. 34–51.
Safran,M. *et al.* (2010) GeneCards version 3: the human gene integrator. *Database(Oxford)*, **2010**, baq020.
Sanz,M.A. (2006) Treatment of acute promyelocytic leukemia. *Hematology Am Soc Hematol Educ Program*, **2006**, 147–155.
Shoemaker,R. (2006) The NCI60 human tumour cell line screen. *Nat. Rev. Cancer*, **6**, 813–823.
Stanton,T. and Reid,J.L. (2002) Fixed dose combination therapy in the treatment of hypertension. *J Hum Hypertens*, **16**, 75–78.
Takigawa,I. *et al.* (2011) Mining significant substructure pairs for interpreting polypharmacology in drug-target network. *PLoS One*, **6**, e16999.
Veltri,R. and Christudass,C. (2014) Nuclear morphometry, epigenetic changes, and clinical relevance in prostate cancer. *Adv. Exp. Med. Biol.*, **773**, 77–99.
Villanueva,J. *et al.* (2010) Acquired resistance to BRAF inhibitors mediated by a RAF kinase switch in melanoma can be overcome by cotargeting MEK and IGF-1R/PI3K. *Cancer Cell*, **18**, 683–695.
Warde-Farley,D. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
Wishart,D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, 901–906.
Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, 668–672.
Zhang,S. *et al.* (2007) Discovering functions and revealing mechanisms at molecular level from biological networks. *Proteomics*, **7**, 2856–2869.
Zhang,S. *et al.* (2011a) A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. *Bioinformatics*, **27**, i401–i409.
Zhang,S. *et al.* (2011b) Combating trastuzumab resistance by targeting SRC, a common node downstream of multiple resistance pathways. *Nat. Med*, **17**, 461–469.
Zhang,S. *et al.* (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.
Zhao,J. *et al.* (2014) Predicting cooperative drug effects through the quantitative cellular profiling of response to individual drugs. *CPT Pharmacometrics Syst. Pharmacol.*, **3**, e102.
Zhao,X.M. *et al.* (2011) Prediction of drug combinations by integrating molecular and pharmacological data. *PLoS Comput. Biol.*, **7**, e1002323.