Bioinformatics Advance Access published July 5, 2016

Bioinformatics, 2016, 1–12
doi: 10.1093/bioinformatics/btw295
Original Paper

Gene expression

SIFORM: shared informative factor models for integration of multi-platform bioinformatic data

Xuebei An, Jianhua Hu* and Kim-Anh Do

Department of Biostatistics, the University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA

*To whom correspondence should be addressed.

Associate Editor: Ziv Bar-Joseph

Received on November 9, 2015; revised on March 31, 2016; accepted on April 28, 2016

Abstract

Motivation: High-dimensional *omic* data derived from different technological platforms have been extensively used to facilitate comprehensive understanding of disease mechanisms and to determine personalized health treatments. Numerous studies have integrated multi-platform *omic* data; however, few have efficiently and simultaneously addressed the problems that arise from high dimensionality and complex correlations.

Results: We propose a statistical framework of shared informative factor models that can jointly analyze multi-platform *omic* data and explore their associations with a disease phenotype. The common disease-associated sample characteristics across different data types can be captured through the shared structure space, while the corresponding weights of genetic variables directly index the strengths of their association with the phenotype. Extensive simulation studies demonstrate the performance of the proposed method in terms of biomarker detection accuracy via comparisons with three popular regularized regression methods. We also apply the proposed method to The Cancer Genome Atlas lung adenocarcinoma dataset to jointly explore associations of mRNA expression and protein expression with smoking status. Many of the identified biomarkers belong to key pathways for lung tumorigenesis, some of which are known to show differential expression across smoking levels. We discover potential biomarkers that reveal different mechanisms of lung tumorigenesis between light smokers and heavy smokers.

Availability and Implementation: R code to implement the new method can be downloaded from http://odin.mdacc.tmc.edu/jhhu/

Contact: jhu@mdanderson.org

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

1 Introduction

The success of the Human Genome Project enables genome-wide studies of various biological activities in humans and other organisms. The extensive development of high-throughput biotechnologies have made data accessible from different platforms, including RNA sequencing, copy number variation, DNA methylation and protein lysate arrays. Although vital to the progress of biomedical research, these multi-platform data impose new challenges for data analysis. In addition to high dimensionality and complex correlations within and across platforms, different types of *omic* data also likely have different scales and distributions (Zhang *et al.*, 2012). Therefore, appropriate and effective analytical methods are desirable to extract useful information from the emerging data platforms.

Integrative analysis of multi-platform data has been proposed to study the interplay within and between different levels of biological data (Chari *et al.*, 2010; Kristensen *et al.*, 2014; Rhodes and Chinnaiyan, 2005). A large number of statistical methods have been discussed for different applications and purposes. For example, sequential analysis (e.g. multiple concerted disruption, (Chari *et al.*, 2010)) confirms or refines the findings obtained from one data type by analyzing additional *omic* data collected from the common set of samples. Pairwise correlation of genetic variables (e.g. weighted gene correlation network analysis, (Langfelder and Horvath, 2008)) is used to infer molecular network interactions. Network analysis (e.g. jActiveModules (Ha *et al.*, 2015; Ideker *et al.*, 2002)) identifies active or aberrant subsets in a certain biological system using molecular network interactions based on graphical models. Bayesian network approaches (Chekouo *et al.*, 2015; Ni *et al.*, 2014) identify biomarkers that are associated with clinical outcomes by incorporating another type of biomarkers collected from the same samples through prior distributions. Another line of work in canonical correlation analysis (CCA) intends to identify the linear combinations of the two sets of genetic variables which have maximum correlation with each other. More recent development based on the original CCA includes sparse CCA which identifies the linear combinations of the two high-dimensional genomic data by incorporating penalty function, supervised CCA which associates the linear combinations with some clinical outcome, and multiple CCA which extends to the case of more than two datasets (Witten and Tibshirani, 2009; Witten *et al.*, 2009).

We propose to simultaneously analyze multi-platform data to detect predictive biomarkers for a response variable of interest, which is typically a disease-associated phenotype. The widely used variable selection approach aggregates genetic variables derived from different platforms as covariates in regression models for the response variable (Bovelstad *et al.*, 2007; Mankoo *et al.*, 2011). To address the extremely high dimensionality of the covariate space, which is much larger than the number of samples, sparsity-induced penalization methods have been commonly used. This line of work includes the L_1-penalty-based *Lasso* (Tibshirani, 1996), its improved version, adaptive Lasso (Zou, 2006), L_1 and L_2 combined elastic net (Zou and Hastie, 2005) and smoothly clipped absolute deviation (*SCAD*) (Fan and Li, 2001; Fan and Peng, 2004). The *SCAD* method has been shown to have appealing theoretical oracle properties (unbiasedness, sparsity and continuity) (Fan *et al.*, 2001). However, the drawbacks of regression-based methods include a limitation on the number of selected variables according to the available sample size and the incapability to detect correlated variables.

To address these challenges, we propose a framework of shared informative factor models to use in integrating multiple types of *omic* data plus a disease-associated response variable. In contrast to the conventional factor models, we incorporate the disease phenotype information in the factor space to detect genetic variables that interact with the response variable. In addition, we assume a common structured factor across multiple data types for the purpose of detecting different levels of the important disease-associated genetic variables. The proposed framework lays the foundation for incorporating prior biological knowledge on functional pathways and networks, both within a data type and across different data types to improve the biological relevance of the results.

2 Methods

2.1 A framework of shared informative factor models

We describe a new model framework to explore associations between a disease phenotype and high-throughput genetic data generated from multiple (≥ 2) platforms. Herein, we focus on two-platform data for the purpose of demonstration. Let the $n \times p_1$ matrix $\mathbf{X}_1$ and $n \times p_2$ matrix $\mathbf{X}_2$ denote two data matrices containing the intensity measurements of p_1 genetic variables obtained from platform 1 and those of p_2 genetic variables obtained from platform 2, respectively. Correspondingly, we let the length-n vector $\mathbf{y}$ denote the phenotypes of n subjects (e.g. smoking status, cancer subtype). The system is built upon the generalized linear models, considering exponential family of distributions for genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ (e.g. continuous measurements, count data). We can express the transformed mean functions of the expression intensities of genetic variables via canonical link functions g_1 and g_2, respectively corresponding to $\mathbf{X}_1$ and $\mathbf{X}_2$, as follows,

$$\begin{aligned} g_1\{E(\mathbf{X}_1)\} &= \boldsymbol{\alpha}_1 + CA, \\ g_2\{E(\mathbf{X}_2)\} &= \boldsymbol{\alpha}_2 + CB. \end{aligned} \tag{1}$$

For the demonstration, we focus on continuously measured genetic variables (e.g. gene expression, protein expression, DNA copy number), for which the identity link $g_1(\mu) = g_2(\mu) = \mu$ is used. We also assume the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ follow normal distributions (posterior to transformation when appropriate). The $n \times 1$ parameter vectors $\boldsymbol{\alpha}_1$ and $\boldsymbol{\alpha}_2$ respectively correspond to baseline sample effects in two sets of genetic data. The jth columns of the $n \times p_1$ residual matrix $\boldsymbol{\varepsilon}_1$ of $\mathbf{X}_1$ and the $n \times p_2$ residual matrix $\boldsymbol{\varepsilon}_2$ of $\mathbf{X}_2$ follow mean-0 normal distributions with the respective variances σ_{1j}^2 and σ_{2j}^2. Our interest is drawn to the multiplicative terms CA and CB, which capture the associations between the phenotype and genetic variables. Our framework has two main distinctions from the conventional factor models that have similar forms. First, the $n \times 1$ parameter vector C is structured to deliver the phenotype information. In the case of a K-categorical phenotype (e.g. K levels of smoking status), $C = \{c_1, \ldots, c_1, \ldots, c_l, \ldots, c_l, \ldots, c_K, \ldots, c_K\}^T$, with c_l indicating the common effect that the subjects in the lth phenotype contribute to the associations between the phenotypes and genetic variables. Second, the same C is employed in these two models to represent the common intrinsic sample characteristics in the two sets of genetic data. The length-p_1 vector A and length-p_2 vector B contain the weights that the genetic variables contribute to the structured sample characteristics shared by the two sets of genetic data. The elements of A and B are called association scores. We call this new system shared informative factor models (SIFORM).

For model identifiability, we impose $\sum_{i=1}^{n} c_i^2 = 1$. We also follow the established practice of standardizing the intensity values of each genetic variable prior to the downstream analysis.

2.2 Model estimation

In the above-described framework, the log-likelihood of the two sets of observed genetic data is

$$\begin{aligned} l = &\sum_{j=1}^{p_1}\sum_{i=1}^{n}\left\{-\frac{(x_{1ij}-\alpha_{1i}-c_l\cdot a_j)^2}{2\sigma_{1j}^2}-\frac{1}{2}\log(2\pi\sigma_{1j}^2)\right\} \\ &+\sum_{j=1}^{p_2}\sum_{i=1}^{n}\left\{-\frac{(x_{2ij}-\alpha_{2i}-c_l\cdot b_j)^2}{2\sigma_{2j}^2}-\frac{1}{2}\log(2\pi\sigma_{2j}^2)\right\}. \end{aligned}$$

To identify important genetic variables associated with the phenotype of interest, we also impose a sparsity regularization on the elements of the parameter vectors A and B in model estimation. We adopt the *SCAD* penalization method (Fan *et al.*, 2001), which uses symmetric penalty functions that are non-concave on $(0, \infty)$ to simultaneously select the variables and estimate the coefficients. *SCAD* has been shown to have good theoretical properties and to offer promising empirical results.

We obtain parameter estimates via minimizing the following objective function, which is simply the negative penalized log-likelihood,

$$S = -l + \sum_{j=1}^{p_1} P_{\delta_1}(|a_j|) + \sum_{j=1}^{p_2} P_{\delta_2}(|b_j|),$$

where $P_\delta(.) = \delta^2 - (. - \delta)^2 I(. < \delta)$, and its first-order derivative is $P'_\delta(.) = \delta I(. \leq \delta) + \frac{(a\delta - .)_+}{a-1} I(. > \delta)$. We take the value of $a = 3.7$ as suggested by (Fan *et al.*, 2001). Herein, δ_1 and δ_2 are the penalty tuning parameters that respectively correspond to the two sets of genetic data.

We adopt the local linear approximation of Zou and Li (2008) for the sparsity penalty term and find a convex function to implement the efficient majorization–minimization algorithm (Hunter and Li, 2005). Using $P_{\delta_1}(|a_j|)$ as an example, the local linear approximation can be expressed as

$$P_{\delta_1}(|a_j|) \approx P_{\delta_1}(|a_j^{(k)}|) + P'_{\delta_1}(|a_j^{(k)}|)(|a_j| - |a_j^{(k)}|), \tag{2}$$

where $a_j^{(k)}$ is the value of a_j estimated at step k. Then the majorization function of the local linear approximated penalty is

$$G_{\delta_1}(|a_j|) = P'_{\delta_1}(|a_j^{(k)}|) \frac{a_j^2 + a_j^{(k)2}}{2|a_j^{(k)}|} + P_{\delta_1}(|a_j^{(k)}|) - P'_{\delta_1}(|a_j^{(k)}|)(|a_j^{(k)}|). \tag{3}$$

Similarly, we can obtain the majorization function $G_{\delta_2}(|b_j|)$ for $P_{\delta_2}(|b_j|)$.

We have thus obtained the parameter estimates by iteratively minimizing the following objective function,

$$S^* = -l + \sum_{j=1}^{p_1} G_{\delta_1}(|a_j|) + \sum_{j=1}^{p_2} G_{\delta_2}(|b_j|).$$

At the $(k+1)$th step, we can obtain the closed-form penalized log-likelihood estimators for the elements of A and B as

$$\widehat{a}_j^{(k+1)} = \frac{|a_j^{(k)}| \sum_{i=1}^{n} c_i (x_{1ij} - \alpha_{1i})}{|a_j^{(k)}| + \sigma_{1j}^2 P'_{\delta_1}(|a_j^{(k)}|)} \tag{4}$$

and

$$\widehat{b}_j^{(k+1)} = \frac{|b_j^{(k)}| \sum_{i=1}^{n} c_i (x_{2ij} - \alpha_{2i})}{|b_j^{(k)}| + \sigma_{2j}^2 P'_{\delta_2}(|b_j^{(k)}|)}. \tag{5}$$

All the other parameter estimates are obtained via their closed-form solutions in a straightforward manner at each step; therefore, we omit the details here.

The tuning parameters $\lambda = (\delta_1, \delta_2)$ are important for calibrating the goodness-of-fit for the data and model sparsity. We adopt the Bayesian information criterion (BIC) (Schwarz, 1978), which is widely used in high-dimensional studies (Fan, 2013; Zou *et al.*, 2007), and defined as

$$BIC = -2l + \log(n(p_1 + p_2)) \cdot df(\lambda),$$

where $df(\lambda)$ is proportional to the summation of the number of nonzeros in $\widehat{\boldsymbol{A}}$ and the number of nonzeros in $\widehat{\boldsymbol{B}}$. The pair of (δ_1, δ_2) that achieves the smallest BIC value is obtained by using a two-dimensional grid search over a pre-determined space. The value is typically between 0 and 10 for each tuning parameter.

2.3 Iterative algorithm

We provide the details of the iterative parameter estimation procedure in Algorithm 1. The convergence threshold ε is set to be 10^{-5}. We also learn from extensive simulations that the estimation seems to stabilize within 300 iterations. We stop the iterative procedure when either convergence is reached or 300 iterations are completed, whichever occurs first.

3 Simulation studies

We conducted extensive simulations to assess the performance of the proposed *SIFORM* and compared its performance to several popular regularized regression methods *SCAD*, *Lasso*, adaptive Lasso (*adaLasso*), and two canonical correlation analysis (CCA) based methods: multiple CCA (*mCCA*) and penalized collaborative regression (*pCollRe*) (Gross and Tibshirani, 2015; Witten and Tibshirani, 2009; Witten *et al.*, 2009). These methods can be implemented directly using the respective R packages *ncvreg*, *glmnet*, *parcor* and *PMA*. The tuning parameters are selected by the 5-fold cross-validation procedure.

We use seven simulation scenarios to study the various data generation models, inter-genetic marker dependence structures and residual variances. In scenarios 1–3 and 5–6, we consider a phenotype variable $\mathbf{Y}$ categorized into four groups and two high-dimensional genetic profiling matrices, $\mathbf{X}_1$ and $\mathbf{X}_2$. In the three penalized regression methods, all the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ are treated as the covariates to predict the phenotype. Note that multinomial logistic regression models are used for the response variable of smoking status in four categories. For SCAD, we take the union of all the nonzero coefficients identified in three separate logistic regression models, treating the category of non-smoker as the reference group. In contrast, *SIFORM* conveniently uses A and B to identify the important biomarkers despite the different disease phenotypes. In

Algorithm 1. Iterative parameter estimation procedure

```
1. Input: $X_1$, $X_2$, $y$
2. Initialization:
    2.1. $\alpha_1^{(0)} \leftarrow 0, \alpha_2^{(0)} \leftarrow 0$
    2.2. $\sigma_{1j}^{(0)} \leftarrow var(X_{1(\cdot j)}), \sigma_{2j}^{(0)} \leftarrow var(X_{2(\cdot j)})$
    2.3. Assign 1,···,K to $c_i$ according to the K-categorical
          phenotype y, then scale C to have $\sum_{i=1}^{n} c_i^2 = 1$. Use
          the scaled C as the initial value, $C^{(0)}$.
    2.4. Use least square estimators for $A^{(0)}, B^{(0)}$:
         $a_j^{(0)} \leftarrow \sum_{i=1}^{n} c_i x_{1ij}, b_j^{(0)} \leftarrow \sum_{i=1}^{n} c_i x_{2ij}$
3. Do-while loop:
    do {
        Update 3.1-3.3 using closed-form estimators
        3.1. $\alpha_{1,2}^{(k+1)} \leftarrow (X_1, X_2, \sigma_{1,2}^{2(k)}, C^{(k)}, A^{(k)}, B^{(k)})$
        3.2. $\sigma_{1,2}^{2(k+1)} \leftarrow (X_1, X_2, \alpha_{1,2}^{(k+1)}, C^{(k)}, A^{(k)}, B^{(k)})$
        3.3. $C^{(k+1)} \leftarrow (X_1, X_2, \alpha_{1,2}^{(k+1)}, \sigma_{1,2}^{2(k+1)}, A^{(k)}, B^{(k)})$
        Update 3.4 using formula (4)
        3.4. $A^{(k+1)} \leftarrow (X_1, C^{(k+1)}, \alpha_1^{(k+1)}, \sigma_1^{2(k+1)})$
        Update 3.5 using formula (5)
        3.5. $B^{(k+1)} \leftarrow (X_2, C^{(k+1)}, \alpha_2^{(k+1)}, \sigma_2^{2(k+1)})$
        3.6. For each estimate, compute the difference
              between current and previous iteration steps:
        $\Delta_1 = |\alpha_1^{(k+1)} - \alpha_1^{(k)}|, \Delta_2 = |\alpha_2^{(k+1)} - \alpha_2^{(k)}|,$
        $\Delta_3 = |\sigma_1^{(k+1)} - \sigma_1^{(k)}|, \Delta_4 = |\sigma_2^{(k+1)} - \sigma_2^{(k)}|,$
        $\Delta_5 = |A^{(k+1)} - A^{(k)}|, \Delta_6 = |B^{(k+1)} - B^{(k)}|,$
        $\Delta_7 = |C^{(k+1)} - C^{(k)}|$
    } while($\Delta_1 > \varepsilon || \Delta_2 > \varepsilon || \Delta_3 > \varepsilon || \Delta_4 > \varepsilon || \Delta_5 > \varepsilon || \Delta_6 > \varepsilon$
     $|| \Delta_7 > \varepsilon$)
4. Output:$\widehat{A} \leftarrow A^{(k+1)}, \widehat{B} \leftarrow B^{(k+1)}, \widehat{C} \leftarrow C^{(k+1)}$
```

addition, we investigated scenarios 4 and 7 wherein integration of three bioinformatics data platforms is considered.

In brief, scenarios 1–4 generate genetic data from model (1), and scenarios 5–7 simulate discrete phenotype data from multinomial logistic regression models. The data matrices $\mathbf{X}_k (k = 1, 2, 3)$ in scenarios 5–7 and residuals ε_k $(k = 1,2,3)$ in scenarios 1–4 are simulated from multivariate mean-0 normal distributions. The corresponding variances are set to be 1 for all the multivariate normally distributed variables in scenarios 1–4. In scenario 6, they are sampled from a lung cancer dataset obtained from The Cancer Genome Atlas (TCGA). In terms of the data dependence structure, the genetic variables in $\mathbf{X}_1$ and $\mathbf{X}_2$ are mutually independent in scenario 1; the genetic variables with nonzero coefficients are correlated with $\rho = 0.8$, and 14–32% of the remaining genetic variables are weakly correlated with $\rho = 0.2$ in scenarios 2–7. Throughout all the scenarios, sparse true biomarkers that are associated with the phenotype are assumed. Additional details for the five simulation scenarios follow.

- Scenario 1: We equally assign 120 samples to the four categories of the phenotype $\mathbf{y}$, which gives the corresponding $C = (0.033, \ldots, 0.033,\ 0.067, \ldots, 0.067, 0.100, \ldots, 0.100, 0.133, \ldots, 0.133)$. Each sample has intensity measurements collected over 100 genetic variables in dataset 1 ($\mathbf{X}_1$) and over 100 genetic variables in dataset 2 ($\mathbf{X}_2$). In each dataset, only the first 5 genetic variables are associated with the phenotype, with the sparse coefficient vectors $A_{1\times100} = (5, 7, 11, 15, 18, 0, \ldots, 0)$ and $B_{1\times100} = (4, 8, 10, 14, 20, 0, \ldots, 0)$. All the genetic variables are assumed to be mutually independent. The baselines α_1 and α_2 are set to be 0, and the intensity measurements in $\mathbf{X}_1$ and $\mathbf{X}_2$ are simulated from model (1).
- Scenario 2: This scenario is the same as scenario 1, except that we assume a blockwise compound symmetry correlation structure among the genetic variables to mimic real studies.
- Scenario 3: In this higher dimensional case, 160 samples are equally assigned to four categories, and there are 1000 variables in each of the two genetic datasets. Among the genetic variables, only the first 50 have nonzero coefficients in each dataset. The data are generated in the same way as in scenario 2.
- Scenario 4: The only difference from scenario 2 is that we include the third omics dataset $\mathbf{X}_3$ which contains additional 100 genetic variables and is generated from $\mathbf{X}_3 = \alpha_3 + \boldsymbol{CD} + \boldsymbol{\epsilon}_3$ with the corresponding sparse coefficient vectors $\boldsymbol{D}_{1\times100} = (3, 6, 9, 12, 17, 0, \ldots, 0)$ and $\alpha_3 = \mathbf{0}$.
- Scenario 5: We use a multinomial logistic regression model to generate the phenotype. We consider a total of 200 samples and 100 genetic variables, where the first 5 variables are true biomarkers, in each of $\mathbf{X}_1$ and $\mathbf{X}_2$. Letting the sparse coefficient vectors $\mathbf{A}_{(1)1\times100} = (6, 14, 14, 24, 20, 0, \ldots, 0)$, $\mathbf{B}_{(1)1\times100} = (4, 12, 10, 18, 18, 0, \ldots, 0)$, $\mathbf{A}_{(2)1\times100} = (3, 14, 11, 24, 24, 0, \ldots, 0)$ and $\mathbf{B}_{(2)1\times100} = (2, 10, 13, 24, 23, 0, \ldots, 0)$, $\mathbf{A}_{(3)1\times100} = (2, 14, 15, 20, 25, 0, \ldots, 0)$ and $\mathbf{B}_{(3)1\times100} = (2, 10, 11, 21, 23, 0, \ldots, 0)$, we have the logit transformed predictor $\eta_{il} = \log(\frac{P(Y_i=l)}{P(Y_i=4)}) = A_{(l)}X_{1i} + B_{(l)}X_{2i}, l = 1,2,3.$ The probability of $Y_i = lis\mathbf{p}_{\mathrm{il}} = \mathbf{e}^{\eta_{\mathrm{il}}}/(\mathbf{1} + \sum_{\mathbf{k}=\mathbf{1}}^{\mathbf{3}} \mathbf{e}^{\eta_{\mathrm{ik}}})$. Accordingly, we sample Y_i from the multinomial distribution $MN(1, p_{i1}, p_{i2}, p_{i3}, p_{i4})$. The genetic intensity values follow multivariate mean-0 normal distributions. The generated data are fairly balanced. For example, a simulated dataset contains 66, 40, 43 and 51 samples in the four respective categories.
- Scenario 6: We investigate a higher dimensional setting for the data generated under a multinomial logistic regression model, which is similar to scenario 4. We consider 230 samples, with $\mathbf{X}_1$ and $\mathbf{X}_2$ respectively containing 3500 and 150 genetic variables. This dimensionality is comparable to that of the TCGA lung cancer dataset we use to illustrate the real application of our method. Among the genetic variables, only the first 30 in $\mathbf{X}_1$ and the first 10 in $\mathbf{X}_2$ are the true predictors of the phenotype. As an example, a simulated dataset produces 59, 47, 49 and 75 samples in the four respective phenotype categories.
- Scenario 7: We investigate a 3-platform setting in which the data is generated from a multinomial logistic regression model. The same $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{A}_{(\mathrm{l})1\times100}$ $(l = 1, 2, 3)$, and $\mathbf{B}_{(\mathrm{l})1\times100}$ $(l = 1, 2, 3)$ in scenario 5 are used here. The only difference from scenario 5 is that we include the additional third omics dataset $\mathbf{X}_3$ which contains 100 genetic variables with the corresponding sparse coefficient vectors $\mathbf{D}_{(1)1\times100} = (3, 9, 12, 15, 22, 0, \ldots,\ 0)$, $\mathbf{D}_{(2)1\times100} = (4, 8, 12, 20, 23, 0, \ldots, 0)$, and $\mathbf{D}_{(3)1\times100} = (3, 10, 15,\ 19, 24, 0, \ldots, 0)$. Similar to scenario 5, Y_i is generated from the logit transformed mean model $\eta_{il} = \log(\frac{P(Y_i=l)}{P(Y_i=4)}) = A_{(l)}X_{1i} + B_{(l)}X_{2i} + D_{(l)}X_{3i}$, $l = 1, 2, 3$.

We ran 100 simulations for each scenario. We report the average value and standard error of the true positive rate (TPR), true negative rate (TNR) and false discovery rate (FDR) among the top genetic variables detected that have the largest values of $|A| or |B|$ for *SIFORM*, and the largest absolute values of the regression coefficients for the other methods. For scenario 3 with 100 true biomarkers, we focus on the top 10 biomarkers detected by different methods. For all the other scenarios, we focus on the top 5 biomarkers detected.

We report the variable selection results in Table 1. The first four scenarios correspond to the generation of data from the shared informative factor models, considering independent and dependent genetic variables, different levels of dimensionality and different number of platforms. The simulations demonstrate the consistently superior performance of the *SIFORM* method compared to those of the five other methods in terms of biomarker detection accuracy across the scenarios. *pCollRe* shows the second best performance, especially in high-dimensional scenarios, partially because both sparsity-induced L_1- and smoothing L_2-penalties are used to obtain better-behaved estimates (Gross and Tibshirani, 2015). In contrast, *mCCA*'s performance appears to be inferior and unstable in terms of TPR and FDR. Among the penalized regression methods, *SCAD* detects the highest number of true biomarkers while sacrificing false positives, and thus yields a high FDR. The remaining two penalized regression approaches, *Lasso* and *adaLasso*, detect fewer true biomarkers, and *adaLasso* produces a lower FDR than *Lasso* and *SCAD*.

We also considered generating data from a logistic regression model under scenarios 5, 6 and 7, which represent different dimensionalities and numbers of platforms of genetic data. While the differences between *SIFORM* and the other three methods are not as large as in the first three scenarios, the advantage of using *SIFORM* in terms of selecting true biomarkers is still clear. In the most scenarios, *SIFORM* detects the most true biomarkers with the smallest FDR among all the six evaluated methods. In contrast, *SCAD*, *Lasso* and *mCCA* yield high FDR, implying that large numbers of trivial genetic variables are selected into the models. *pCollRe* is again the second best performer, in fact, it discovers more true positives than *SIFORM* in scenario 6, but at the sacrifice of FDR. Furthermore, we evaluated the phenotype prediction performance of *SIFORM* and *pCollRe* via cross-validation, and the result suggests that *pCollRe* may not be well-suited for prediction, coherent with the observation in the literature (Gross and Tibshirani, 2015). For

Table 1. Sample means and standard errors of TPR, TNR and FDR for four methods under seven simulation scenarios

	Scenario	Method	TPR	TNR	FDR
The true model	Scenario 1	SIFORM	0.983(0.038)	0.985(0.008)	0.000(0.000)
		SCAD	0.676(0.129)	0.921(0.030)	0.342(0.319)
		Lasso	0.466(0.133)	0.987(0.013)	0.170(0.159)
		adaLasso	0.571(0.128)	0.991(0.011)	0.052(0.093)
		pCollRe	0.677(0.104)	1.000(0.000)	0.014(0.051)
		mCCA	0.603(0.243)	0.991(0.029)	0.048(0.090)
	Scenario 2	SIFORM	0.989(0.031)	0.987(0.008)	0.000(0.000)
		SCAD	0.266(0.178)	0.964(0.036)	0.350(0.320)
		Lasso	0.323(0.094)	0.958(0.024)	0.436(0.178)
		adaLasso	0.111(0.057)	0.995(0.001)	0.192(0.039)
		pCollRe	0.540(0.091)	1.000(0.000)	0.068(0.099)
		mCCA	0.688(0.277)	0.984(0.037)	0.022(0.142)
	Scenario 3	SIFORM	0.959(0.015)	0.994(0.002)	0.000(0.000)
		SCAD	0.195(0.036)	0.988(0.005)	0.441(0.222)
		Lasso	0.093(0.020)	0.998(0.002)	0.136(0.121)
		adaLasso	0.075(0.012)	0.999(0.001)	0.050(0.085)
		pCollRe	0.542(0.031)	0.996(0.002)	0.001(0.010)
		mCCA	0.068(0.014)	1.000(0.000)	0.000(0.000)
	Scenario 4	SIFORM	0.950(0.090)	1.000(0.000)	0.000(0.000)
		SCAD	0.364(0.102)	0.989(0.160)	0.335(0.299)
		Lasso	0.287(0.112)	0.995(0.242)	0.212(0.144)
		adaLasso	0.390(0.096)	1.000(0.147)	0.069(0.099)
		pCollRe	0.430(0.072)	1.000(0.000)	0.086(0.115)
		mCCA	0.316(0.161)	0.999(0.005)	0.529(0.069)
Logistic regression	Scenario 5	SIFORM	0.791(0.123)	0.987(0.007)	0.012(0.048)
		SCAD	0.754(0.105)	0.925(0.024)	0.334(0.203)
		Lasso	0.642(0.136)	0.903(0.028)	0.170(0.140)
		adaLasso	0.675(0.130)	0.984(0.020)	0.016(0.055)
		pCollRe	0.627(0.112)	0.994(0.000)	0.024(0.071)
		mCCA	0.398(0.286)	0.970(0.045)	0.376(0.216)
	Scenario 6	SIFORM	0.392(0.067)	0.999(0.005)	0.000(0.000)
		SCAD	0.246(0.062)	0.988(0.004)	0.598(0.208)
		Lasso	0.132(0.044)	0.990(0.003)	0.390(0.174)
		adaLasso	0.106(0.033)	0.999(0.001)	0.110(0.140)
		pCollRe	0.498(0.066)	0.996(0.001)	0.014(0.051)
		mCCA	0.206(0.220)	0.973(0.044)	0.740(0.280)
	Scenario 7	SIFORM	0.622(0.094)	1.000(0.000)	0.000(0.000)
		SCAD	0.262(0.094)	0.967(0.014)	0.346(0.170)
		Lasso	0.315(0.101)	0.929(0.019)	0.458(0.174)
		adaLasso	0.413(0.090)	0.999(0.007)	0.056(0.095)
		pCollRe	0.520(0.109)	1.000(0.000)	0.018(0.058)
		mCCA	0.270(0.200)	0.966(0.043)	0.726(0.119)

simplicity, in scenario 6 we focused on the samples in the two most extreme phenotype categories $y = 1$ versus 4. We used leave-one-out cross-validation and obtained the average misclassification rate for each dataset. To make a fair comparison, we focused on the predictive performance of the top 5 biomarkers identified by each method. The means (standard deviations) of misclassification rates of *SIFORM* and *pCollRe* are, respectively, 0.283 (0.085) and 0.409 (0.090) across 100 generated datasets. The result indicates that *SIFORM* also outperforms *pCollRe* in terms of prediction.

Figure 1depicts the boxplots of the number of TPRs and FDRs obtained by all six methods over 100 simulations under scenarios 3 and 7.

We also assessed the performance of the BIC in our proposed framework. We use scenario 3 as an example. Figure 2 shows the BIC values along the tuning parameters δ_1 and δ_2 in a simulated dataset. Reading from left to right in Figure 2, the subpanels correspond to δ_1 values increasing from 2.5 to 4.6. In each subpanel, the BIC values are plotted against the values of δ_2 in increments of 0.3. The good behavior of the BIC method is indicated by clear convex curves, with $(\delta_1 = 3.4, \delta_2 = 4)$ chosen as the optimal tuning parameter values for this dataset. In a common scenario, we note that the results are very similar across different simulated datasets, and thus, we use the tuning parameter values obtained from a single dataset for computational consideration.

4 Lung cancer applications

Lung cancer is the leading cause of cancer deaths in the United States (Sanchez-Cespedes *et al.*, 2001). Among the major histological types of lung cancer, adenocarcinoma is the most common form in non-smokers (Sun *et al.*, 2007). We investigated the applicability of *SIFORM* to the TCGA lung adenocarcinoma dataset available through the data portal hosted by the National Cancer Institute (http://cancergenome.nih.gov/). Particularly, we focused on two phenotypes, smoking status and survival, to demonstrate the applicability of *SIFORM* and explore new molecular biomarkers that may facilitate understanding of lung cancer.

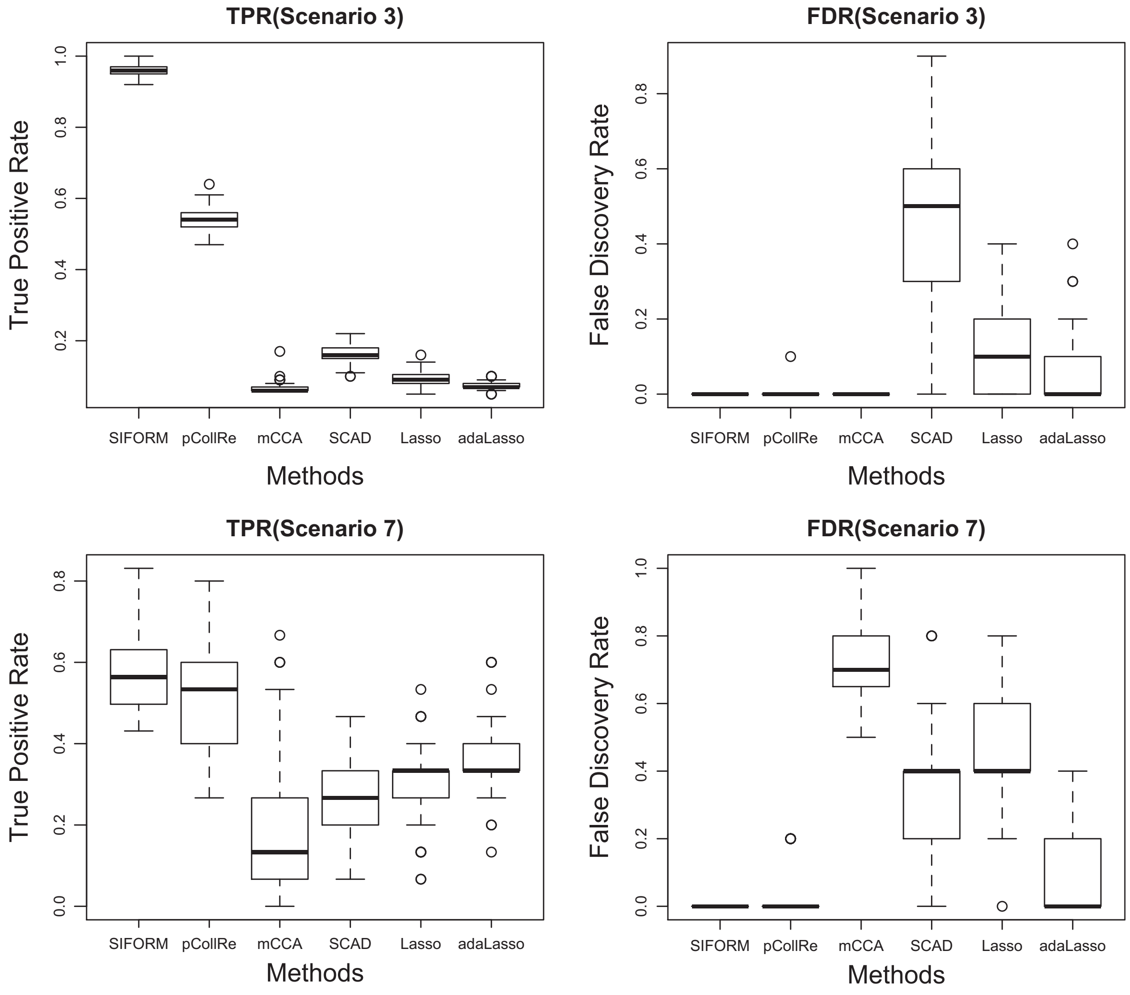

Fig. 1. Comparison of number of TPRs and FDRs among 6 methods under scenarios 3 and 7

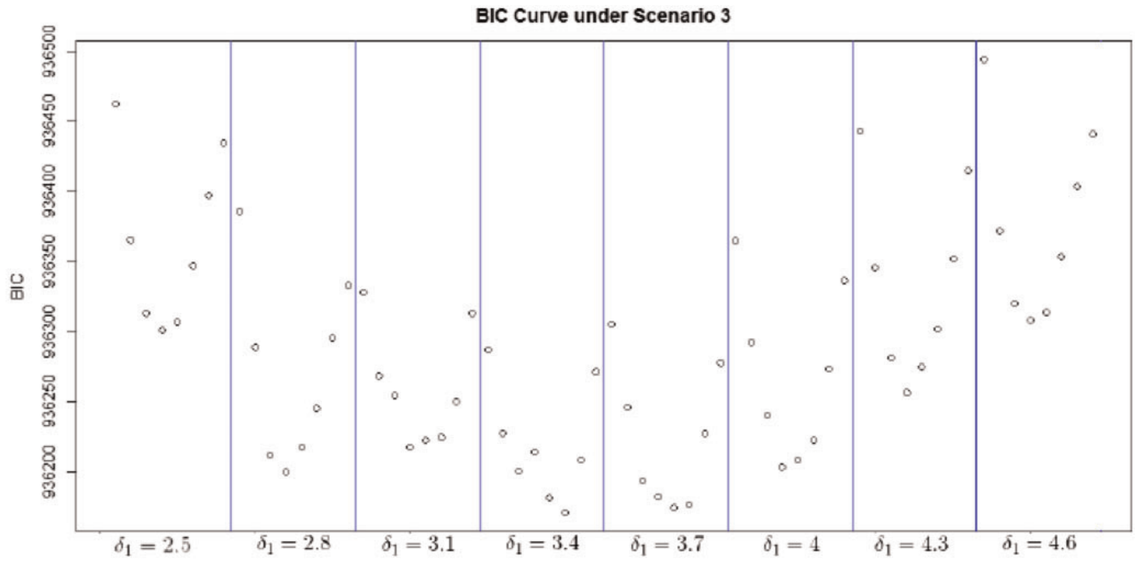

Fig. 2. BIC curve under scenario 3

4.1 The association of biomarkers with smoking status

Cigarette smoking is the most important risk factor for lung cancer; however, approximately 25% of lung cancer cases are not attributable to tobacco use. Striking differences in the epidemiological, clinical and molecular characteristics of lung cancer have been demonstrated when the cancer arises in a non-smoker versus a smoker (Samet *et al.*, 2009; Sun *et al.*, 2007; Yano *et al.*, 2008, 2011). This suggests that the lung cancers are likely caused by separate biological mechanisms for smokers and nonsmokers.

Substantial studies have identified the genes and pathways that contribute to lung tumorigenesis in smokers (e.g. *EGFR*, *KRAS* and *TP53*) (Ding *et al.*, 2008; Sun *et al.*, 2007), and these discoveries have led to the development of targeted therapies (e.g. *EGFR* tyrosine kinase inhibitor erlotinib; anti-VEGF antibody bevacizumab) (Serke, 2007). However, most of the studies have analyzed genetic profiles in a single assay, and the etiology of lung tumors that arise in non-smokers remains unclear (Sun *et al.*, 2007). Therefore, it is worthwhile to integrate multi-platform bioinformatic data to discover predictive biomarkers that are associated with the smoking status of patients diagnosed with lung adenocarcinoma.

4.1.1 Data description

The TCGA lung adenocarcinoma dataset includes samples from 225 patients, from whom expression intensities of 20 531 genes and 160 proteins were measured using the respective platforms of Illumina RNA sequencing and reverse-phase protein array (RPPA) technology. The variable of smoking status has four categories. Among the 225 patients represented in the samples, there were 49 with the status of 'current smoker', 86 with 'current reformed smoker for < or = 15 years', 58 with 'current reformed smoker for >15 years' and 32 with 'lifelong non-smoker'.

The gene expression data were normalized using the RNA-seq by Expectation–Maximization approach (Li and Dewey, 2011) and logarithm-transformed prior to downstream analysis. The protein concentration data were also normalized by subtracting the median, both column-wise and row-wise (Li *et al.*, 2013).

4.1.2 A subsample study

The six methods we investigated aim to simultaneously analyze genetic variables to detect predictive biomarkers for a given phenotype. We were further interested in investigating the capabilities of the methods to detect biomarkers that are likely marginally associated with the phenotype, using individual marker tests as the reference.

We subsampled 300 genes and 100 proteins from all the genetic variables. Because the truth is unknown in a real data study, we treated 30 genes as the true biomarkers; these genes were identified via gene shaving (Hastie *et al.*, 2000) or with the smallest P-values obtained from a univariate F-test. The remaining 270 'null' genes are genes that have the largest P-values obtained from the univariate F-test. Among the 100 proteins, 10 were treated as truly important as determined by gene shaving and univariate test results. The other 90 'null' proteins are proteins with the largest P-values based on the univariate F-test.

We report the results of our implementation of all six methods in Table 2(a), which lists the TPR, TNR and FDR values. Compared to the other methods, *SIFORM* has superior performance in terms of accurate biomarker detection. *pCollRe* has the second highest true positive rate among the remaining methods. *Lasso* performs the worst in both true and false biomarker detection, while *SCAD* and *adaLasso* have similar performances, with discovery rates better than *Lasso*. These results are consistent with the simulation results demonstrated in the previous section.

Table 2. Comparison of six methods in a subsampling study

Method	TPR	TNR	FDR
(a) Lung cancer study with smoking status as outcome			
SIFORM	0.95	1	0
SCAD	0.35	0.994	0.2
Lasso	0.1	0.986	0.4
adaLasso	0.275	1	0
pCollRe	0.775	0.994	0
mCCA	0.5	0.986	0
(b) Lung cancer study with survival as outcome			
SIFORM	0.95	1	0
SCAD	0.325	0.992	0
Lasso	0.325	0.958	0
adaLasso	0.25	1	0
pCollRe	0.8	0.992	0
mCCA	0.475	0.989	0

4.1.3 Full data analysis

We implemented the following steps to filter out trivial genes in the RNA-seq data. First, we removed 5% of the genes that had extremely small coefficient of variation values, where the coefficient of variation is defined as the ratio of the standard deviation to the absolute value of the mean of the gene expression intensities. Second, we removed genes for which the difference between the top 90% quantile and the bottom 10% quantile was no larger than 0.8. Then, we filtered out the genes with P-value ≥ 0.03 based on a univariate analysis of variance F-test. This procedure retained 3707 genes for further analysis.

We implemented all six methods for the integrative analysis of the genomic and proteomic data. To determine the tuning parameter values in *SIFORM*, we performed a grid search over $[4, 14]$ for δ_1 and $[1, 10]$ for δ_2. We obtained $(\delta_1, \delta_2) = (5.2, 4)$ as the local optimal values. *SIFORM* identified 17 genes that had nonzero coefficients in A: ATP13A4, BANK1, C20orf103, C5orf41, C7, DBF4B, FAM65C, LOC100132707, MACROD2, PCDHAC2, RSPO2, STOM, THOC4, TLR3, TMEM173, UGT1A4 and USP53. *SIFORM* also detected 13 proteins based on B: 4E-BP1, 4E-BP1_pT70, Akt_pS473, caspase-7_cleavedD198, Chk1, Chk2_pT68, EGFR_pY1068, JNK2, PCNA, PDK1_pS241, Ret_pY905, stathmin and tuberin.

We performed hierarchical clustering of the samples based on the Pearson correlation distance and applied Ward's linkage method to the selected genes and proteins. The clustering heatmaps using the 17 genes and 13 proteins, respectively, are displayed in Figure 3. In both panels, we observe that most 'non-smokers' (yellow) and 'current reformed smokers for >15 years' (orange) are clustered together. In addition, most of the 49 'current smokers' (blue) are mixed with some 'current reformed smokers for <=15 years' (green). However, the data for some of the 86 recently reformed smokers (green) behave differently from the rest of the group; these samples enlarge the overall difference between the green and blue groups. The observations can be somewhat reflected from the estimate of $C = (-0.064, -0.074, 0.015, 0.104)$. The close values (c_1, c_2) between 'non-smokers' and 'current reformed smokers for >15 years' indicate similarity between the major genetic profiling characteristics of these two categories of smoking status. The relatively large difference between (c_1, c_2), c_3 and c_4 manifests possible genetic profiling variations between (roughly) non-smokers, recent smokers and current smokers. In particular, the most different genetic profiles are those of non-smokers compared to current smokers, which agrees with our intuitive expectation.

We can also visually see that the most identified biomarkers show clear differential expression between light smoking and heavy smoking. For instance, the genes contained in the first 13 rows of the heatmap in the upper panel have lower expression levels in heavy smokers but higher expression levels in light smokers, and the bottom four genes show the opposite expression pattern. As an example, gene *UGT1A4* (the 14th gene in the upper panel of Fig. 3) appears to be up-regulated in the heavy smokers but down-regulated in the light smokers. Extensive evidence shows that *UGT1A4* can be induced by cigarette constituents (Bock *et al.*, 1994; Collier *et al.*, 2002).

We summarize the comparisons among the six methods in terms of variable selection as follows.

- *SCAD*, *Lasso*, *adaLasso*, *pCollRe* and *mCCA* identify 60, 22, 33, 27 and 12 genes, respectively. *SCAD* detects more genes than the other methods; however, it likely yields the largest number of false positives, as indicated by the simulation studies. No common gene is identified across the six methods. This implies an ongoing challenge for genetic studies of the association between smoking and lung cancer, which is also reflected in the very limited medical literature in this area of research. *SCAD*, *Lasso* and *mCCA*, all of which are known to produce high false positives, identified 5 genes in common. *AdaLasso* and *pCollRe*, which are known to have appealing empirical and theoretical properties (Gross and Tibshirani, 2015; Zou, 2006), identified very few genes that overlap with those identified by the other comparative methods but substantially overlap with those identified by *SIFORM* (13 by *adaLasso* and 10 by *pCollRe*).
- In terms of proteomic profiling, *SCAD*, *Lasso*, *adaLasso* and *mCCA* identified very few proteins, 3, 0, 1 and 1, respectively, that are associated with smoking status. In contrast, our proposed method, *SIFORM* detected 13 proteins and pCollRe detected 10. *SIFORM* and *pCollRe* identified 9 proteins in common, which could be more convincing due to the good performance of *pCollRe* in simulation studies.

Next, we studied the biological relevance of the genes and proteins detected by *SIFORM*. Conducting pathway analysis with a web-based tool, Pathway Commons (http://www.pathwaycommons.org/pcviz/), we discovered that 13 genes likely interact with each other through functional pathways and networks. This is displayed in Figure 4(a), where these genes are highlighted in green.

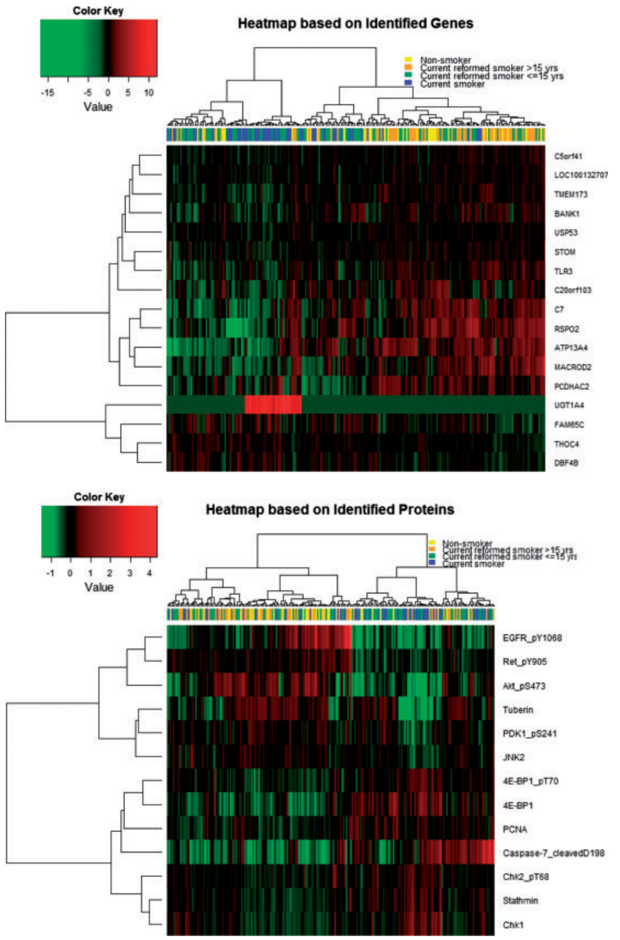

Fig. 3. Sample clustering based on genes and proteins selected by SIFORM

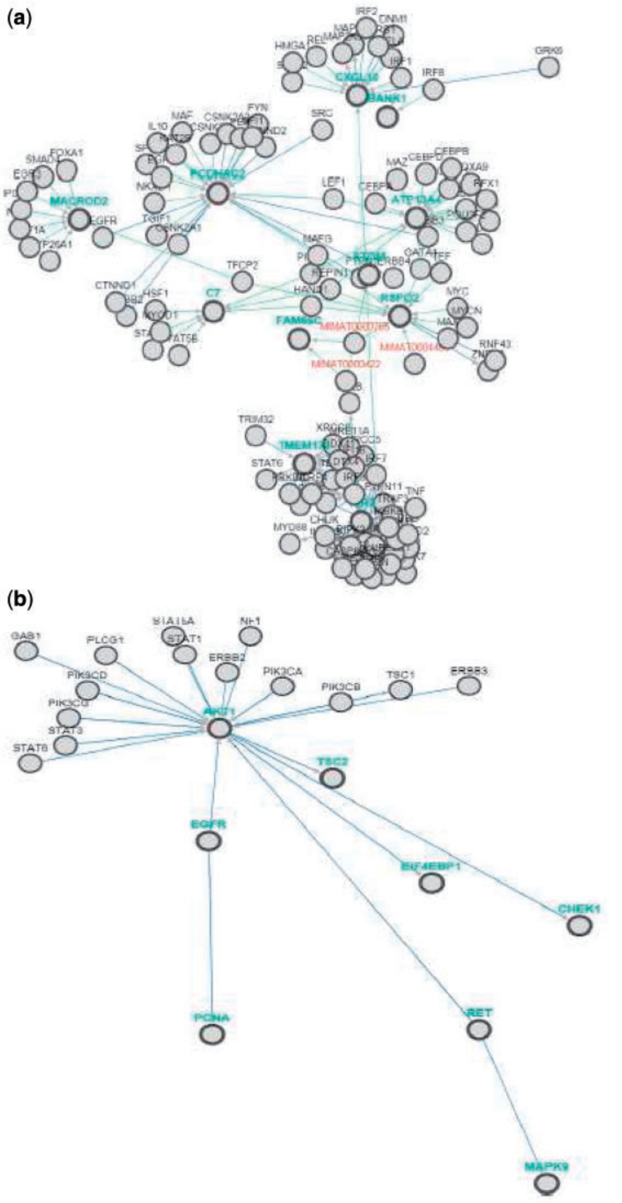

Fig. 4. Identified biomarker networks

The important miRNAs associated with a common pathway are shown in red. The 8 detected proteins appear to be in several common pathways, as highlighted in green in Figure 4(b). Several of them are known to be associated with lung cancer.

EGFR, which was detected in our study, has been found to play a key role in lung tumorigenesis. Approximately 10% of non-small-cell lung carcinoma (NSCLC) cases in the U.S. and 35% in East Asia have tumor-associated *EGFR* mutations. *EGFR* mutations are more often found in tumors from non-smokers with adenocarcinoma histology (http://www.mycancergenome.org/content/disease/lung-cancer/egfr/; Lovly *et al.*, 2015), which supports the biological relevance of our finding.

Activated *EGFR* can regulate the activity of another detected gene, *PCNA*, and trigger its important downstream signaling pathway AKT/PI3K (Sarris *et al.*, 2012). *PCNA*, an index of tumor cell proliferation, has high expression in poorly differentiated lung adenocarcinomas (Brand *et al.*, 2011; Ma *et al.*, 2008). Another detected biomarker, *AKT1*, is an isoform of *AKT* (Fumarola *et al.*, 2014). The signal transduction pathway AKT/PI3K is involved in the regulation of cell proliferation, survival, differentiation, adhesion, motility and invasion. Aberrations of this pathway have been implicated in lung cancer development and progression (Fumarola *et al.*, 2014).

Most of the proteins selected by *SIFORM* are associated with the tumorigenesis of NSCLC through the AKT pathway, as shown by Figure 4(b). First, the activated AKT pathway regulates tuberin (*TSC2*), which inhibits the *mTOR* nutrient signaling input through the tuberous sclerosis complex. *mTOR* has been correlated with NSCLC tumor progression (Fumarola *et al.*, 2014; Sarris *et al.*, 2012). Second, the activated *AKT* phosphorylates *CHEK1*, an integral component of the DNA damage response. The overexpression of *CHEK1* is associated with poor tumor differentiation and significantly worse patient survival in NSCLC (Grabauskiene *et al.*, 2014). In addition, activated *AKT* induces the phosphorylation and inactivation of *EIF4EBP1*, and the increased phosphorylation of *EIF4EBP1* is found to be associated with progression of several types of cancer, including lung adenocarcinoma (Dumstorf *et al.*, 2010; Gingras *et al.*, 1998; Sekia *et al.*, 2010). In addition to *EGFR*, we identified another activator of the AKT/PI3K pathways—*RET*. The alteration of *RET* has key roles in cell growth, differentiation and survival, and has been associated with NSCLC (Fumarola *et al.*, 2014). In addition to *PI3K/AKT*, *RET* signaling activates the *MAPK* family, including *MAPK9 (JNK2)* (Giunti *et al.*, 2013). Some studies have found that *MAPK9* is frequently activated in NSCLC (Nitta *et al.*, 2011).

In summary, *EGFR* and its downstream *PI3K/AKT* pathway are among the most important molecular therapeutic targets for NSCLC (Cooper *et al.*, 2013; Ren *et al.*, 2012). Existing studies suggest differences in *EFGR* mutations between smokers and non-smokers, and thus implicate the *AKT/PI3K* pathway and its downstream targets as playing nontrivial roles that differ in patients who develop lung cancer and have a history of smoking versus those who have never smoked. In contrast, other methods have identified only one protein, collagen VI, which is an extracellular matrix protein. Although collagen VI has been correlated with tumor progression, its role in NSCLC is rarely discussed (Chen *et al.*, 2013; Voiles *et al.*, 2014).

Cross-validation: We evaluated the prediction performance of each method used in simulation. For simplicity, we focused on the 81 patients whose habits placed them in the two most extreme categories related to smoking: 49 who were current smokers and 32 who were lifelong non-smokers. We used leave-one-out cross-validation to predict the smoking status of each patient. This procedure was repeated 81 times to obtain the overall misclassification rate. To make a fair comparison, we focused on the predictive performance of the top 5 biomarkers with the largest absolute coefficient estimates. The misclassification rates of *SIFORM*, *SCAD*, *Lasso*, *adaLasso*, *mCCA* and *pCollRe* are respectively 0.407, 0.469, 0.593, 0.556, 0.605 and 0.593. These results indicate that *SIFORM* has the best prediction accuracy. The overall high misclassification rates across the six methods might arise from the limited sample sizes and the recognized difficulty of differentiating the genetic mechanisms of lung cancer attributable to smoking versus not smoking.

4.2 The association of biomarkers with survival

The 5-year survival rate for lung cancer is only approximately 15%, and the conventional treatments (e.g. chemotherapy) have limited impact on improving survival of advanced NSCLC. Biological functions of biomarkers may play important roles in disease formation and progression. Therefore, identification of such biomarkers can likely advance targeted cancer therapies towards the direction of personalized medicine (Kuykendall and Chiappori, 2014). The integrative analysis of multi-platform bioinformatic data aims at understanding the underlying relationship between survival and genetic activities at different levels in patients with lung adenocarcinoma.

We investigate the same set of gene intensity and protein expression data as described in Section 4.1. The only difference is that the survival phenotype is considered here. We divide 219 patients into two groups (Long-Term Survival (LTS) and Short-Term Survival (STS)) according to the length of survival time. By using an extreme discordant phenotype design (Nebert, 2000), we define the top 25% (55 patients, surviving >896 days) of the patients as LTSs and the bottom 65% (29 patients, surviving <593 days) as STSs. We include these 84 patients in data analysis.

4.2.1 A subsample study

We again subsampled 300 genes and 100 proteins, among which 30 genes and 10 proteins are selected as the true biomarkers. The sampling and selection procedure is described in Section 4.1.2. We report the results obtained by the six methods in Table 2(b). Compared to the other methods, *SIFORM* shows superior performance in terms of biomarker detection accuracy, in particular, TPR. *pCollRe* performs the second best, while *Lasso* produces the smallest true negative rate. The variables detected by each method are all the true biomarkers (FDR = 0) for this particular case. Overall, the result is also consistent with the simulation studies.

4.2.2 Full data analysis

We filtered out trivial genes in the RNA-seq dataset in the similar fashion as described in Section 4.1.3. This procedure results in 2529 genes included in the following data analysis. We implemented all the six methods. For *SIFORM*, the tuning parameters (δ_1, δ_2) are set to (4.7, 1.9) using the BIC method. Consequently, *SIFORM* identified 12 proteins: 4E-BP1, CD49b, Cyclin_B1, ER-alpha, ER-alpha_pS118, Fibronectin, INPP4B, LCN2a, Napsin-A, PAI-1, PCNA and TTF1. In comparison, *SCAD*, *Lasso*, *adaLasso*, *mCCA* and *pCollRe* detected 1, 1, 1, 1 and 4 proteins, respectively; these five methods identified only one common protein, estrogen receptor (ER)-alpha. The prognostic values of ER expressions in lung cancer have been extensively discussed. ER proteins, including ER-alpha, is found to be associated with a poorer prognosis among NSCLC patients (Kawai *et al.*, 2005; Olivo-Marston *et al.*, 2010). This protein is also discovered by *SIFORM*.

We also notice that the most proteins identified by *SIFORM* are found biologically relevant to lung cancer prognosis. 4E-BP1 is a eIF4E binding protein, which is known to be related to reduced survival in a variety of cancers including lung adenocarcinoma. High 4E-BP1 expression has been found to be correlated with worse overall survival of lung cancer patients (Dumstorf *et al.*, 2010; Lee *et al.*, 2015; Lv *et al.*, 2015). Cyclin B1 plays a key role in the G2-M phase transition of the cell cycle, the elevated expression level of which has been shown as an indicator of poor prognosis in NSCLC (Arinaga *et al.*, 2003; Cooper *et al.*, 2009; Yoshida *et al.*, 2004). Fibronectin, which plays an important role in cell adhesion, migration, growth and differentiation by mediating cellular interactions with the extracellular matrix, stimulates NSCLC cell growth and survival through activation of Akt/mTOR/p70S6K and inactivation of LKB1/AMPK signal pathways (Han et al., 2006; Pankov and Yamada, 2002). Napsin-A and TTF1 have been extensively studied as the correlated prognostic factors for lung cancer patients' survival. Patients with high expression levels of TTF-1 and Napsin A have better survival rates than those with low levels of expression (Ma *et al.*, 2015). Finally, high expression levels of PAI-1 and PCNA proteins indicate a shorter survival for patients diagnosed with lung adenocarcinoma (Di Bernardo *et al.*, 2009; Robert *et al.*, 1999).

In terms of gene expression profile, *SCAD*, *Lasso*, *adaLasso*, *mCCA* and *pCollRe* identified 24, 40, 17, 215 and 29 genes, respectively, compared to none detected by *SIFORM*. However, no common genes are identified across these five methods. Moreover, we observe that only eight genes are identified by any four methods simultaneously, but none of them has been reported to be correlated with lung cancer survival in the literature.

We also used the Kaplan–Meier method to further investigate if the selected proteins can differentiate patients in terms of length of survival time. Specifically, we fit logistic regression to the binary survival phenotype ('STS' versus 'LTS') treating the 12 proteins as the covariates, and then call the linear prediction value, based on the estimated coefficients, as the single score for a patient. Each patient is reclassified into two subgroups (called 'long-term survivors' versus 'short-term survivors') based on their calculated scores by using 0.5 as the cutoff. We display the Kaplan–Meier curves of the two subgroups based on the patients' original survival data in Figure 5, which clearly shows the distinctive survival curves between the two subgroups (P-value = 0.045). This result suggests that the 12 proteins selected by *SIFORM* can potentially be considered as prognostic biomarkers for lung adenocarcinoma.

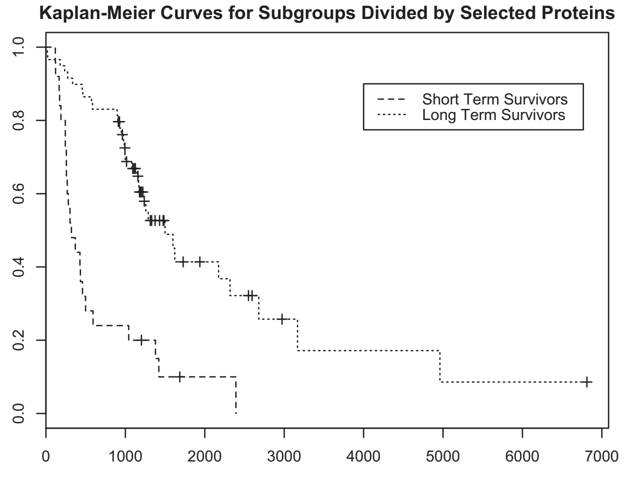

Fig. 5. Kaplan–Meier Curves for Subgroups Divided by SIFORM-Selected Proteins

Furthermore, we evaluated the prediction performance of all the methods via leave-one-out cross-validation. The misclassification rates of *SIFORM*, *SCAD*, *Lasso*, *adaLasso*, *mCCA* and *pCollRe* are, respectively, 0.322, 0.369, 0.631, 0.560, 0.417 and 0.607. SIFORM again outperforms all the others in terms of prediction accuracy for the survival phenotype.

5 Discussion

We have proposed a generalized statistical framework *SIFORM* to jointly model high-throughput *omic* data produced by multiple platforms and to discover associations between these genetic variables and a disease-associated phenotype. The new method conveniently produces direct rankings of genetic variables in terms of the strength of association with the response variable. Extensive simulation studies demonstrated the superior performance of *SIFORM* in terms of biomarker detection accuracy, regardless of inter-variable correlations, compared to the performances of other penalized variable selection methods. Biological meaningfulness of the proposed method is supported by two TCGA lung adenocarcinoma studies that investigate the association of the integrated mRNA expression and RPPA protein concentration data and two phenotypes, smoking status and discretized survival. In particular, we discovered that most of the identified proteins either belong to known key pathways for NSCLC tumorigenesis or are prognostic biomarkers in NSCLC. We also discovered some proteins that are potentially associated with the different ways in which NSCLC develops in smokers versus non-smokers.

The proposed framework appears to fit the lung cancer data reasonably well. However, a more comprehensive version of this framework can be developed to improve the goodness of fit of the model if appropriate. A possible solution is to include a nonparametric mean component, for example, with the structure of singular value decomposition, and use model selection methods (e.g. BIC) to determine the rank of the additional component. This idea requires extensive investigation and, therefore, is deferred to our future research. Another ongoing research project is to modify *SIFORM* by incorporating the prior biological pathway information to improve biomarker detection accuracy. Specifically, further regularization of the parameters, for example, a graphic model based penalization method (Kim *et al.*, 2013), will be investigated.

Funding

This work was supported in part by the National Institute of Health Grants R01CA158113, CCSG P30 CA016672, and the Moonshot Programs Funding at M.D. Anderson.

Conflict of Interest: none declared.

References

Arinaga,M. *et al.* (2003) Clinical implication of cyclin B1 in non-small cell lung cancer. *Oncol. Rep.*, **10**, 1381–1386.

Bock,K.W. *et al.* (1994) The influence of environmental and genetic factors on CYP2D6, CYP1A2 and UDP-glucuronosyltransferases in man using sparteine, caffeine, and paracetamol as probes. *Pharmacogenetics*, **4**, 209–218.

Bovelstad,H.M. *et al.* (2007) Predicting survival from microarray data - a comparative study. *Bioinformatics*, **23**, 2080–2087.

Brand,T.M. *et al.* (2011) The nuclear epidermal growth factor receptor signaling network and its role in cancer. *Discov. Med.*, **12**, 419–432.

Chari,R. *et al.* (2010) An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Syst. Biol.*, **4**, 67.

Chekouo,T. *et al.* (2015) miRNA-target gene regulatory networks: a Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*, **71**, 428–438.

Chen,P. *et al.* (2013) Collagen VI in cancer and its biological mechanisms. *Trends Mol. Med.*, **19**, 410–417.

Collier,A.C. *et al.* (2002) Metabolizing enzyme localization and activities in the first trimester human placenta: the effect of maternal and gestational age, smoking and alcohol consumption. *Hum. Reprod.*, **17**, 2564–2572.

Cooper,W.A. *et al.* (2009) Expression and prognostic significance of cyclin B1 and cyclin A in non-small cell lung cancer. *Histopathology*, **55**, 28–36.

Cooper,W.A. *et al.* (2013) Molecular biology of lung cancer. *J. Thorac. Dis*, **5**, S479–S490.

Di Bernardo,M.C. *et al.* (2009) Plasminogen activator inhibitor variants PAI-1 A15T and PAI-2 S413C influence lung cancer prognosis. *Lung Cancer*, **65**, 237–241.

Ding,L. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.

Dumstorf,C.A. *et al.* (2010) Modulation of 4E-BP1 function as a critical determinant of enzastaurin-induced apoptosis. *Mol. Cancer Ther.*, **9**, 3158–3163.

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Fan,J. and Peng,H. (2004) Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Stat.*, **32**, 928–961.

Fan,Y. (2013) Tuning parameter selection in high dimensional penalized likelihood. *J. R. Stat. Soc. B*, **75**, 531–552.

Fumarola,C. *et al.* (2014) Targeting PI3K/AKT/mTOR pathway in nonsmall cell lung cancer. *Biochem. Pharmacol.*, **90**, 197–207.

Gingras,A.C. *et al.* (1998) 4E-BP1, a repressor of mRNA translation, is phosphorylated and inactivated by the Akt(PKB) signaling pathway. *Genes Dev.*, **12**, 502–513.

Giunti,S. *et al.* (2013) Cellular signaling pathway alterations and potential targeted therapies for medullary thyroid carcinoma. *Int. J. Endocrinol.*, **2013**, 803171.

Grabauskiene,S. *et al.* (2014) Checkpoint kinase 1 protein expression indicates sensitization to therapy by checkpoint kinase 1 inhibition in non-small cell lung cancer. *J. Surg. Res.*, **187**, 6–13.

Gross,S.M. and Tibshirani,R. (2015) Collaborative regression. *Biostatistics*, **16**, 326–338.

Ha,M.J. *et al.* (2015) DINGO: differential network analysis in genomics. *Bioinformatics.*, **31**, 3413–3420.

Han,S. *et al.* (2006) Fibronectin stimulates non-small cell lung carcinoma cell growth through activation of Akt/mammalian target of rapamycin/S6 kinase and inactivation of LKB1/AMP-activated protein kinase signal pathways. *Cancer Res.*, **66**, 315–323.

Hastie,T. *et al.* (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, research0003.1–research0003.21.

Hunter,D.R. and Li,R. (2005) Variable selection using MM algorithm. *Ann. Stat.*, **33**, 1617–1642.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Kawai,H. *et al.* (2005) Estrogen receptor alpha and beta are prognostic factors in non-small cell lung cancer. *Clin. Cancer Res.*, **11**, 5084–5089.

Kim,S. *et al.* (2013) Network-based penalized regression with application to genomic data. *Biometrics*, **69**, 582–593.

Kristensen,V. *et al.* (2014) Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer*, **14**, 299–313.

Kuykendall,A. and Chiappori,A. (2014) Advanced EGFR mutation-positive non-small-cell lung cancer: case report, literature review, and treatment recommendations. *Cancer Control*, **21**, 67–73.

Langfelder,P. and Horvath,S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee,H.W. *et al.* (2015) Prognostic significance of phosphorylated 4E-binding protein 1 in non-small cell lung cancer. *Int. J. Clin. Exp. Pathol.*, **8**, 3955–3962.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Li,J. *et al.* (2013) TCPA: a resource for cancer functional proteomics data. *Nat. Methods*, **10**, 1046–1047.

Lovly,C. *et al.* (2015) EGFR mutations in non-small cell lung cancer (NSCLC). *My Cancer Genome*. https://www.mycancergenome.org/content/disease/lung-cancer/egfr/.

Lv,T. *et al.* (2015) Twist1-mediated 4E-BP1 regulation through mTOR in non-small cell lung cancer. *Oncotarget*, **6**, 33006–33018.

Ma,J. *et al.* (2008) Clinicopathological significance of E-cadherin and PCNA expression in human non-small cell lung cancer. *Chin. J. Clin. Oncol.*, **5**, 87–92.

Ma,Y. *et al.* (2015) The expression of TTF-1 and Napsin A in early-stage lung adenocarcinoma correlates with the results of surgical treatment. *Tumour Biol.*, **36**, 8085–8092.

Mankoo,P.K. *et al.* (2011) Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS ONE*, **6**, e24709.

Nebert,D.W. (2000) Extreme discordant phenotype methodology: an intuitive approach to clinical pharmacogenetics. *Eur. J. Pharmacol.*, **410**, 107–120.

Ni, Y. *et al.* (2014) Integrative Bayesian network analysis of genomic data. *Cancer Inform.*, **13**, 39–48.

Nitta,R.T. *et al.* (2011) The role of the c-Jun N-terminal kinase 2-alpha-isoform in non-small cell lung carcinoma tumorigenesis. *Oncogene*, **30**, 234–244.

Olivo-Marston,S. *et al.* (2010) Serum estrogen and tumor-positive estrogen receptor-alpha are strong prognostic classifiers of non-small-cell lung cancer survival in both men and women. *Carcinogenesis*, **31**, 1778–1786.

Pankov,R. and Yamada,K.M. (2002) Fibronectin at a glance. *J. Cell Sci.*, **115**, 3861–3863.

Ren,J.H. *et al.* (2012) EGFR mutations in non-small-cell lung cancer among smokers and non-smokers: a meta-analysis. *Environ. Mol. Mutagen.*, **53**, 78–82.

Rhodes,D.R. and Chinnaiyan, A.M. (2005) Integrative analysis of the cancer transcriptome. *Nat. Genet*, **37**, S31–S37.

Robert,C. *et al.* (1999) Expression of plasminogen activator inhibitors 1 and 2 in lung cancer and their role in tumor progression. *Clin. Cancer Res.*, **5**, 2094.

Sarris,E.G. *et al.* (2012) The biological role of PI3K pathway in lung cancer. *Pharmaceuticals (Basel)*, **5**, 1236–1264.

Samet,J. *et al.* (2009) Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clin. Cancer Res.*, **15**, 5626–5645.

Sanchez-Cespedes,M.S. *et al.* (2001) Chromosomal alterations in lung adenocarcinoma from smokers and nonsmokers. *Cancer Res.*, **61**, 1309–1313.

Schwarz,G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Sekia,N. *et al.* (2010) Prognostic significance of expression of eukaryotic initiation factor 4E and 4E binding protein 1 in patients with pathological stage I invasive lung adenocarcinoma. *Lung Cancer*, **70**, 329–334.

Serke,M. (2007) Lung cancer: targeted therapy. *Pneumologie*, **61**, 162–170.

Sun,S. *et al.* (2007) Lung cancer in never smokers – a different disease. *Nat. Rev. Cancer*, **7**, 778–790.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B*, **58**, 267–288.

Voiles,L. *et al.* (2014) Overexpression of type VI collagen in neoplastic lung tissues. *Oncol. Rep.*, **32**, 1897–1904.

Witten,D.M. and Tibshirani,R. (2009) Extensions of sparse canonical correlation analysis with applications to genomic data. *Stat. Appl. Genet. Mol. Biol.*, **8**, Article 28.

Witten,D.M. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Yano,T. *et al.* (2008) Never-smoking nonsmall cell lung cancer as a separate entity — the clinico-pathologic features and survival. *Cancer*, **113**, 1012–1018.

Yano,T. *et al*. (2011) Non-small cell lung cancer in never smokers as a representative “non-smoking-associated lung cancer”: epidemiology and clinical features. *Int. J. Clin. Oncol*., **16**, 287–293.

Yoshida,T. *et al*. (2004) The clinical significance of Cyclin B1 and Wee1 expression in non-small-cell lung cancer. *Ann. Oncol*., **15**, 252–256.

Zhang,S. *et al*. (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucl. Acids Res*., **40**, 9379–9391.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B*, **67**, 301–320.

Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc*., **101**, 1418–1429.

Zou,H. *et al*. (2007) On the “degrees of freedom” of the lasso. *Ann. Stat*., **35**, 2173–2192.

Zou,H. and Li,R. (2008) One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Stat*., **36**, 1509–1533.