Published in final edited form as:

Stat Med. 2019 March 30; 38(7): 1200–1212. doi:10.1002/sim.8027.

Identification of cancer omics commonality and difference via community fusion

Yifan Sun[1,2], **Yu Jiang**[3], **Yang Li**[1,2,4], and **Shuangge Ma**[2,5]

[1]Center for Applied Statistics, Renmin University of China, Beijing, China

[2]School of Statistics, Renmin University of China, Beijing, China

[3]School of Public Health, The University of Memphis, Memphis, Tennessee

[4]Statistical Consulting Center, Renmin University of China, Beijing, China

[5]Department of Biostatistics, Yale University, New Haven, Connecticut

Abstract

The analysis of cancer omics data is a “classic” problem; however, it still remains challenging. Advancing from early studies that are mostly focused on a single type of cancer, some recent studies have analyzed data on multiple “related” cancer types/subtypes, examined their commonality and difference, and led to insightful findings. In this article, we consider the analysis of multiple omics datasets, with each dataset on one type/subtype of “related” cancers. A Community Fusion (CoFu) approach is developed, which conducts marker selection and model building using a novel penalization technique, informatively accommodates the network community structure of omics measurements, and automatically identifies the commonality and difference of cancer omics markers. Simulation demonstrates its superiority over direct competitors. The analysis of TCGA lung cancer and melanoma data leads to interesting findings.

Keywords

commonality and difference; community fusion; multi-cancer analysis; network-based analysis

1 | INTRODUCTION

For most, if not all, cancer types, omics profiling studies have been extensively conducted. The early studies are usually focused on a single cancer type/subtype. More recently, as represented by the National Cancer Institute (NCI) pan-cancer study, more and more studies have conducted the joint analysis of data on multiple cancer types.[1,2] Such studies can be more challenging and more informative than those on a single cancer type.[3–5] On one hand,

Correspondence Shuangge Ma, Department of Biostatistics, Yale University, 60 College st, New Haven, CT 06520. shuangge.ma@yale.edu.

SUPPORTING INFORMATION
Additional supporting information may be found online in the Supporting Information section at the end of the article.

CONFLICT OF INTEREST
The authors declare no potential conflict of interests.

with the well-known heterogeneity, differences across cancer types are expected. On the other hand, many studies have suggested the shared omics basis of multiple cancers.[6] As such, a certain level of commonality is also expected. *It is important to acknowledge and accommodate both difference and commonality in analysis.*

In the literature, the heterogeneity (difference) across different types of cancers has been sufficiently acknowledged. In contrast, the difference among subjects with the same type of cancer has been noted to a less extent.[7–9] To fix idea, consider the TCGA (The Cancer Genome Atlas) melanoma data analyzed in study (for more details, refer to Section 4.1). In Figure 2, we show the plots of Breslow thickness (response variable) against the expressions of three genes. The three colors correspond to the three different cancer stages, and the lines are generated using lowess smoothing. It is easy to see that, for different stages of the same cancer, the effects of genes on the response variable are significantly different. It is noted that data on the three stages have been included in one single dataset, and this particular data[10] as well as those alike[11] have usually been analyzed as a whole, with insufficient attention to heterogeneity/difference. In short, accommodating omics difference associated with a third variable (stage in this particular example) is much needed but insufficiently studied.[8]

Many analytic approaches have been developed for analyzing a single cancer omics dataset.[12,13] Such methods can be used to analyze, for example, the TCGA melanoma dataset as a whole, which stresses commonality but cannot accommodate difference. They can also be applied to one part of data (which may correspond to one stage, subtype, etc) at a time, and then, results are compared across different data parts to draw conclusions on commonality and difference.[9] One major problem is that, since each data part often has a small sample size, results from single-part analysis and hence the final results can be unsatisfactory. Related discussions have been extensively provided in recent integrative analysis studies.[4,5,14] In the literature, the work that is the most relevant to this study is perhaps the contrasted penalization,[15] which encourages similarity in analysis results (especially regression coefficients) across datasets. Contrasted penalization and some other approaches can be limited by identifying similarity but not commonality, in the sense that parts of the analysis results (from multiple datasets) may be similar but not identical (and thus are not commonly shared). In addition, they may not sufficiently accommodate the coordination among omics variables.

The analysis scheme of this study is described in Figure 2. The left panel describes the true model structure. Here, the three columns correspond to three datasets (with each dataset corresponding to one cancer subtype, stage, etc), and the rows correspond to genes (or other omics units). The genes form three communities with sizes nine, seven, and six, which are generated from network analysis and describe the coordinated nature of genes. The genes that are associated with the responses are colored. For this schematic example, the orange community behaves the same for the three datasets (commonality), while the yellow and green communities behave differently (difference). *Our goal is to conduct community-based analysis (so as to sufficiently accommodate gene coordination) and identify commonality as well as difference in genes' associations with responses*. For such a purpose, a Community Fusion (CoFu) approach is developed. In the middle and right panels of Figure 2, the analysis results using an alternative and CoFu are provided, where CoFu has a more accurate identification (more definitive results are provided below in simulation).

For the analysis of cancer omics data, this study has a unique focus on identifying both commonality and difference and complements the existing studies. A novel CoFu approach is developed. Advancing from methods that pool all datasets together, it can flexibly allow difference across datasets. By jointly analyzing multiple datasets (and hence conducting integrative analysis), it can be more effective than analyzing multiple datasets individually and then pooling results. Advancing from contrasted penalization and other related approaches, it conducts community-based analysis and can identify commonality as opposed to similarity. Overall, this study can deliver a practically useful new venue for analyzing cancer omics data and may lead to important new findings.

2 | METHODS

Consider the analysis of K independent datasets. Each dataset can be on a different type of cancer or a different stage of the same cancer (or be stratified in another meaningful way). In dataset $k(=1, \ldots, K)$, there are n_k i.i.d. observations. Here, homogeneity is assumed within but not across datasets. Denote $\boldsymbol{y}^k = (y_1^k, \ldots, y_{n_k}^k)^\top$ as the response vector and $\boldsymbol{X}^k \in \mathbb{R}^{n_k \times p}$ as the design matrix for genes (or other omics units). To simplify notation, it is assumed that the same set of genes is measured in all K datasets. In addition, assume that data processing, for example, normalization, has been properly conducted.

For the kth dataset, consider the linear regression (LR) model

$$\boldsymbol{y}^k = \boldsymbol{X}^k \boldsymbol{\beta}^k + \epsilon^k, \; k = 1, 2, \ldots, K,$$

where $\boldsymbol{\beta}^k = (\boldsymbol{\beta}_1^k, \ldots, \boldsymbol{\beta}_p^k)^\top$ is the length-p vector of regression coefficients, and ϵ^k is the vector of random errors. Here, we first consider LR, which is the most popular and matches data analyzed in this article. The proposed method is applicable to other models, for example, the GLM model. More details are provided in Appendix A in the supporting information.

For analyzing cancer omics data, network-based analysis, which takes a system perspective and accommodates the interconnections among genes, has been shown to be more informative than individual-gene-based analysis. In our analysis, we accommodate the network community structure, with the understanding that genes in the same community tend to behave in a coordinated manner, while those in different communities tend to behave differently. In recent literature, there have been extensive studies on network and community construction.[16] Here, we adopt existing construction and assume that the community structure is available prior to analysis. Assume that the p genes belong to L non-overlapping communities, with $p_{(l)}$ genes in community l. Denote $\boldsymbol{\beta}_{(l)}^k$ as the coefficient vector for community l in dataset k.

We propose the CoFu estimate

$$\left\{\hat{\boldsymbol{\beta}}^k : k = 1, \ldots, K\right\} = \operatorname{argmin}\left\{\sum_{k=1}^{K} \frac{1}{2n_k}\left\|\boldsymbol{y}^k - \boldsymbol{X}^k\boldsymbol{\beta}^k\right\|_2^2 + \lambda_1 \sum_{k=1}^{K}\left\|\boldsymbol{\beta}^k\right\|_1 \right. \quad (1)$$

$$\left. + \lambda_2 \sum_{k=1}^{K-1}\sum_{l=1}^{L}\left\|\boldsymbol{\beta}_{(l)}^k - \boldsymbol{\beta}_{(l)}^{k+1}\right\|_2\right\},$$

where $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 > 0$ are data-dependent tuning parameters. The nonzero components of $\hat{\boldsymbol{\beta}}^k$ correspond to genes that are associated with the response in dataset k. If If $\hat{\boldsymbol{\beta}}_{(l)}^{k_1} = \hat{\boldsymbol{\beta}}_{(l)}^{k_2}$, then genes in community l behave the same in datasets k_1 and k_2.

That is, they represent the commonality shared by the two datasets. Otherwise, they represent the difference. As such, the proposed analysis can simultaneously identify both commonality and difference.

<u>Rationale</u>. In (1), with K independent datasets, the lack-of-fit measure is the sum of K individual ones. Normalization by sample size is taken to avoid domination by larger datasets. The first penalty is Lasso, which has been adopted in a large number of studies. It accommodates the high data dimensionality and conducts regularized estimation and selection of relevant genes. When desirable, it can be replaced by other, more complicated penalties. For example, when it is desirable to accommodate network adjacency (between any two nodes), then Laplacian-type penalties can be further added.

The main advancement is the second penalty, which conducts community-level analysis and has a fusion form. Tailored to data analyzed in Section 4, the penalty has been designed for datasets with a natural order. For example, dataset k may correspond to cancer stage k. For two adjacent datasets, the fusion penalty encourages equal regression coefficients, that is, commonality. However, it is flexible and does not reinforce commonality. Different from the "standard" fusion penalties, it is imposed on the regression coefficients of different datasets. Different from the contrasted penalties, it conducts community-based analysis: a whole community will be concluded as behaving the same (or differently) in multiple datasets. The results so generated can be more interpretable than those individual gene based. In addition, with a nonzero probability, the proposed penalty may generate $\hat{\boldsymbol{\beta}}_{(l)}^{k_1} = \hat{\boldsymbol{\beta}}_{(l)}^{k_2}$, differing from the contrasted penalization,[15] which generates similar but not identical estimates. For scenarios where a natural order of data does not exist, the second penalty can be revised as $\lambda_2 \sum_{k,j=1,\ldots,K} \sum_{l=1}^{L}\left\|\boldsymbol{\beta}_{(l)}^k - \boldsymbol{\beta}_{(l)}^j\right\|_2$. Loosely speaking, the newly proposed penalty has a group Lasso form (on the differences of coefficient vectors). It can be potentially replaced by other group penalties (as well as other techniques that can conduct group selection).

2.1. Computation

To tackle the nonseparability of the penalty, we introduce a new set of parameters: $\boldsymbol{\eta}^k = \boldsymbol{\beta}^k - \boldsymbol{\beta}^{k+1}$.In addition, we also define a set of parameters $\boldsymbol{\delta}^k = \boldsymbol{\beta}^k$ to separate the L_1

norm operation and the sum of squares operation defined on $\boldsymbol{\beta}^k$. Then, the minimization in (1) is equivalent to the following constrained optimization problem:

$$\min f(\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\delta}) \equiv \sum_{k=1}^{K}\frac{1}{2n_k}\left\|\boldsymbol{y}^k-\boldsymbol{X}^k\boldsymbol{\beta}^k\right\|_2^2+\lambda_1\sum_{k=1}^{K}\left\|\boldsymbol{\delta}^k\right\|_1+\lambda_2\sum_{k=1}^{K-1}\sum_{l=1}^{L}\left\|\boldsymbol{\eta}_{(l)}^k\right\|_2,$$

subject to

$$\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k+1}-\boldsymbol{\eta}^k=0,\ k=1,2,\ldots,K-1,$$

and

$$\boldsymbol{\beta}^k=\boldsymbol{\delta}^k,\ k=1,2,\ldots,K\,.$$

Here, $\boldsymbol{\beta}=((\boldsymbol{\beta}^1)^\top,\ldots,(\boldsymbol{\beta}^K)^\top)^\top,\boldsymbol{\eta}=((\boldsymbol{\eta}^1)^\top,\ldots,(\boldsymbol{\eta}^{K-1})^\top)^\top$, and $\boldsymbol{\delta}=((\boldsymbol{\delta}^1)^\top,\ldots,(\boldsymbol{\delta}^K)^\top)^\top$ .By the augmented Lagrangian method,[17] the estimates can be obtained by minimizing

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\delta},\boldsymbol{u},\boldsymbol{v}) &= f(\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\delta})+\sum_{k=1}^{K-1}(u^k)^\top(\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k+1}-\boldsymbol{\eta}^k)+\sum_{k=1}^{K}(v^k)^\top(\boldsymbol{\beta}^k-\boldsymbol{\delta}^k)\\ &+\frac{\sigma}{2}(\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k+1}-\boldsymbol{\eta}^k\right\|_2^2+\sum_{k=1}^{K}\left\|\boldsymbol{\beta}^k-\boldsymbol{\delta}^k\right\|_2^2),\end{aligned}$$

where the dual variables $\boldsymbol{u}=((\boldsymbol{u}^1)^\top,\ldots,(\boldsymbol{u}^{K-1})^\top)^\top$ and $\boldsymbol{v}=((\boldsymbol{v}^1)^\top,\ldots,(\boldsymbol{v}^K)^\top)^\top$ are Lagrange multipliers, and $\boldsymbol{\sigma} > 0$ is the penalty parameter. Throughout this article, we set $\boldsymbol{\sigma} = 2$, which leads to satisfactory numerical results.

We compute the estimate of $(\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{\delta},\boldsymbol{u},\boldsymbol{v})$, denoted as $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}},\hat{\boldsymbol{\delta}},\hat{\boldsymbol{u}},\hat{\boldsymbol{v}})$, iteratively by using the alternating direction method of multipliers. For a given $(\boldsymbol{\eta},\boldsymbol{\delta},\boldsymbol{u},\boldsymbol{v})$, to obtain the update of $\boldsymbol{\beta}$, we set $\partial L/\partial\boldsymbol{\beta}$ to be zero, where

$$\begin{aligned}L &= \sum_{k=1}^{K}\frac{1}{2n_k}\left\|\boldsymbol{X}^k\boldsymbol{\beta}^k-\boldsymbol{y}^k\right\|_2^2\\ &+\frac{\sigma}{2}(\sum_{k=1}^{K-1}\left\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k+1}-\boldsymbol{\eta}^k+\boldsymbol{u}^k/\sigma\right\|_2^2+\sum_{k=1}^{K}\left\|\boldsymbol{\beta}^k-\boldsymbol{\delta}^k+\boldsymbol{v}^k/\sigma\right\|_2^2)\\ &=\frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}\right\|_2^2+\frac{\sigma}{2}(\|A\boldsymbol{\beta}-\boldsymbol{\eta}+\boldsymbol{u}/\sigma\|_2^2+\|\boldsymbol{\beta}-\boldsymbol{\delta}+\boldsymbol{v}/\sigma\|_2^2)\,.\end{aligned}$$

$\boldsymbol{X}=\ \text{diag}\ (\boldsymbol{X}^1/\sqrt{n_1},\ldots,\boldsymbol{X}^k/\sqrt{n_k}),\boldsymbol{y}=((\boldsymbol{y}^1/\sqrt{n_1})^\top,\ldots,(\boldsymbol{y}^K/\sqrt{n_k})^\top)^\top$ .e_j is the length-K column vector, with the jth element equal to 1 and the rest equal to 0. $\Delta=\left\{e_j-e_{j+1},j=1,\ldots,K-1\right\}^\top.A=\Delta\otimes I_p$, where $\boldsymbol{I}_p$ denotes the $p\times p$ identity matrix, and $\otimes$ denotes the Kronecker product.

Thus, for a given $(\boldsymbol{\eta}(t), \boldsymbol{\delta}(t), \boldsymbol{u}(t), \boldsymbol{v}(t))$ at the tth iteration,

$$\boldsymbol{\beta}(t+1) = \left[\boldsymbol{X}^{\top}\boldsymbol{X} + \sigma(A^{\top}A + I_{pK})\right]^{-1} \times \left\{\boldsymbol{X}^{\top}\boldsymbol{y} + \sigma\left[A^{\top}(\boldsymbol{\eta}(t) - \boldsymbol{u}(t)/\sigma) + \boldsymbol{\delta}(t) - \boldsymbol{v}(t)/\sigma\right]\right\}, \quad (2)$$

where $\boldsymbol{I}_{pK}$ denotes the $pK \times pK$ identity matrix.

The component of the Lagrange function $\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v})$ that depends on $\boldsymbol{\delta}$ is

$$\lambda_1\left\|\boldsymbol{\delta}\right\|_1 + \frac{\sigma}{2}\left\|\boldsymbol{\beta} - \boldsymbol{\delta} + \boldsymbol{v}/\sigma\right\|_2^2.$$

Hence, the closed-form solution for $\boldsymbol{\delta}$ is

$$\hat{\boldsymbol{\delta}} = \mathrm{ST}_{\lambda_1/\sigma}(\boldsymbol{\beta} + \boldsymbol{v}/\sigma),$$

where $\mathrm{ST}_{\lambda}(\boldsymbol{X}) = \operatorname{sign}(\boldsymbol{X})(|\boldsymbol{X}| - \lambda)_+$ is the soft thresholding function, and $(x)_+ = \boldsymbol{x}$ if $x > 0$, and $(x)_+ = 0$ otherwise. Then, the update of $\boldsymbol{\delta}$ at the $(t + 1)$th iteration is

$$\boldsymbol{\delta}(t+1) = \mathrm{ST}_{\lambda_1/\sigma}[\boldsymbol{\beta}(t+1) + \boldsymbol{v}(t)/\sigma]. \quad (3)$$

For $\boldsymbol{\eta}^k$, the relevant component in $\mathscr{L}(\boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\delta}, \boldsymbol{u}, \boldsymbol{v})$ is

$$\lambda_2\left\|\boldsymbol{\eta}^k_{(l)}\right\|_2 + \frac{\sigma}{2}\left\|\boldsymbol{\beta}^k_{(l)} - \boldsymbol{\beta}^{k+1}_{(l)} - \boldsymbol{\eta}^k_{(l)} + \boldsymbol{u}^k_{(l)}/\sigma\right\|_2^2. \quad (4)$$

The minimizer of (4) is

$$\hat{\boldsymbol{\eta}}^k_{(l)} = \left(1 - \frac{\lambda_2}{\sigma\left\|\boldsymbol{\beta}^k_{(l)} - \boldsymbol{\beta}^{k+1}_{(l)} + \boldsymbol{u}^k_{(l)}/\sigma\right\|_2}\right)_+ (\boldsymbol{\beta}^k_{(l)} - \boldsymbol{\beta}^{k+1}_{(l)} + \boldsymbol{u}^k_{(l)}/\sigma), \; k = 1, 2, \ldots, K-1.$$

We can thus obtain the update of $\boldsymbol{\eta}^k_{(l)}$ at the $(t + 1)$th iteration as

$$\boldsymbol{\eta}^k_{(l)}(t+1) = \left(1 - \frac{\lambda_2}{\sigma\left\|\boldsymbol{\beta}^k_{(l)}(t+1) - \boldsymbol{\beta}^{k+1}_{(l)}(t+1) + \boldsymbol{u}^k_{(l)}(t)/\sigma\right\|_2}\right)_+ \quad (5)$$
$$\times\left[\boldsymbol{\beta}^k_{(l)}(t+1) - \boldsymbol{\beta}^{k+1}_{(l)}(t+1) + \boldsymbol{u}^k_{(l)}(t)/\sigma\right].$$

Finally, the estimates of ***u*, *v*** are updated as

$$\boldsymbol{u}(t+1) = \boldsymbol{u}(t) + \sigma[A\boldsymbol{\beta}(t+1) - \boldsymbol{\eta}(t+1)], \quad (6)$$

$$\boldsymbol{v}(t+1) = \boldsymbol{v}(t) + \sigma[\boldsymbol{\beta}(t+1) - \boldsymbol{\delta}(t+1)]. \quad (7)$$

<u>Overall algorithm</u>. The overall algorithm proceeds as follows.

Step 1: Initialize $\boldsymbol{\beta}(0)$. Let $\boldsymbol{\delta}(0) = \boldsymbol{\beta}(0)$, $\eta(0) = \boldsymbol{A\beta}(0)$, $\boldsymbol{u}(0) = 0$, and $\boldsymbol{v}(0) = 0$. In our numerical study, we use the Lasso estimate (without fusion) as the initial value.

Step 2: At the $(t+1)$th iteration, compute $\boldsymbol{\beta}(t+1), \boldsymbol{\delta}(t+1), \boldsymbol{\eta}(t+1), \boldsymbol{u}(t+1),$ and $\boldsymbol{v}(t+1)$ according to Equations (2), (3), (5), (6), and (7), respectively.

Step 3: Repeat Step 2 until convergence. Specifically, convergence is concluded if all of the primal residuals and dual residuals are small enough, that is,

$$\max\left\{\|A\boldsymbol{\beta}(t+1) - \boldsymbol{\eta}(t+1)\|_2, \|\boldsymbol{\beta}(t+1) - \boldsymbol{\delta}(t+1)\|_2, \|\boldsymbol{\delta}(t+1) - \boldsymbol{\delta}(t)\|_2, \|\boldsymbol{\eta}(t+1) - \boldsymbol{\eta}(t)\|_2\right\} < \epsilon$$

In numerical study, ϵ is set to be 10^{-3}.

<u>Tuning parameter selection</u>. The proposed approach involves two tuning parameters: λ_1 and λ_2. In numerical study, they are selected using V-fold cross validation (CV). In this article, $\boldsymbol{V}$ = 5. More specifically, each dataset is partitioned randomly into five non-overlapping subsets with equal sizes. We apply a two-dimensional grid search for λ_1 and λ_2 with $\lambda_2 \in (0.001, 0.01, 0.1, 1)$. Let $\lambda_1^{\max}$ be the minimal λ such that all regression coefficients shrink to 0, ie, $\lambda_1^{\max} = \max_k \left\|\frac{(\boldsymbol{X}^k)^\top \boldsymbol{y}^k}{n^k}\right\|_\infty$ .We choose $\lambda_1^{\min}$ to be some small fraction (default value is 0.01 in our implementation) of $\lambda_1^{\max}$ and log-linearly interpolate between $\lambda_1^{\min}$ and $\lambda_1^{\max}$. For the CoFu method, the CV is computationally affordable. For example, the five-fold CV for a simulated dataset takes less than 15 minutes on a desktop PC.

3 | SIMULATION

In simulation, we set $K = 3$, $n_k = 200$, and $p = 1000$. Genes form $L = 50$ non-overlapping communities, with community sizes ranging from $\frac{3p}{4L}$ to $\frac{5p}{4L}$. As in the literature,[14] we simulate omics measurements from multivariate normal distributions, which may mimic gene expression data (as analyzed below). The normal distributions have marginal means 0, marginal variances 1, and correlation matrix Σ. The following correlation structures are considered: (a) structured correlation. Omics measurements within the same communities are more strongly correlated than those in different communities. More details are provided below; (b) unstructured correlation. The correlation coefficient between measurements i and j is randomly sampled from the uniform distribution $\mathcal{U}[0.2, 1]$; and (c) no correlation. That is, all omics measurements are independent. This may serve as a test of sensitivity and examine performance of the proposed approach when there is a lack of well-defined network structure.

Roughly speaking, the structured correlation matrix corresponds to the scenario where two genes within the same community are strongly correlated while two genes within different communities are weakly correlated. Since, in reality, it is possible that not all genes within the same community are correlated, we generate the structured correlation matrix via network-based analysis, which is considerably more complicated than in the literature and consists of the following steps: (a) generate an unweighted network with a community structure,* (b) add a weight to each edge, which quantifies the connection between two nodes (omics measurements), and (c) generate the correlation matrix Σ. More specifically, we adopt the degree-correlated stochastic block model[18] to generate an unweighted network with a community structure. To mimic networks in reality, we generate a degree sequence of nodes following the power-law distribution with exponent $\boldsymbol{y}$. For each pair of nodes $\boldsymbol{i}$ and j, an undirected edge is placed with probability

$$p_{ij} = \frac{< d > p}{Z} d_i d_j q_{m_i m_j}, \quad (8)$$

where $Z = \sum_{i,j} d_i d_j q_{m_i m_j}$ is the normalization constant, d_i is the degree of node i, $< d >$ is the average degree of nodes, and m_i represents the community that node i belongs to. $q_{m_i m_j}$ represents the connection probability between communities m_i and m_j. If $m_1 = m_2$, $\boldsymbol{q}_{m_1 m_2}$ is sampled from an uniform distribution $\mathcal{U}[0.3, 0.5]$, otherwise $\boldsymbol{q}_{m_1 m_2} = 0.02$. In all simulations, we set $\gamma = 2.5$ and $< d > = 10$. Next, we add a weight to each edge. Specifically, for edges between nodes in the same community and edges between nodes in different communities, weights are sampled independently from uniform distributions $\mathcal{U}[0.5, 1]$ and $\mathcal{U}[0.2, 0.5]$, respectively. Denote the adjacency matrix of the weighted network as Σ_0, with all diagonal elements equal to 1. Note that Σ_0 is not necessarily positive definite. To guarantee positive definiteness, we set $\Sigma = \Sigma_0 - (\lambda_{\min} - \frac{1}{p})\mathbf{I}_p$, where $\lambda_{\min}$ is the smallest eigenvalue of Σ_0, and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. For a more intuitive presentation, we plot one simulated structured correlation matrix in Figure B1 (Appendix B in the supporting information) with $p = 200$, $L = 10$, and other parameters as described above. As shown in the Figure, only genes that are connected by an edge are correlated. The correlation between two connected genes within the same community is strong, whereas that between two connected genes from different communities is weak.

Each dataset has r important omics measurements with $r = 100$ and 150, which are distributed uniformly across communities. All communities are divided randomly into three categories: (a) all-overlapping, where the three datasets have the same sparsity structure and also the same regression coefficients; (b) half-overlapping, where datasets 1 and 2 share half of the important effects, for which the regression coefficients are identical. The same is true for datasets 2 and 3. Half of the important effects in datasets 2 and 3 are dataset specific; and (c) non-overlapping, where there is no important effect shared by any two datasets. Denote $\rho_a, \rho_h, \rho_n (\rho_a + \rho_h + \rho_n = 1)$ as the proportions of all-, half-, and non-overlapping communities, respectively. For the important effects, their regression coefficients are (a) all set to be 0.5; (b) sampled from the uniform distribution $\mathscr{U}[0.2, 1]$, and (c) sampled from different distributions. Specifically, the nonzero regression coefficients in dataset 2 are from $\mathscr{U}[0.4, 0.7]$, those specific to dataset 1 are from $\mathscr{U}[0.1, 0.3]$, and those specific to dataset 3 are from $\mathscr{U}[0.8, 1]$. The random errors are simulated independently from$\mathscr{N}$(0,1). The response variables are computed from the linear models.

We compare CoFu with two closely related alternatives: (a) P.Lasso, which pools all datasets together and applies Lasso. This approach emphasizes commonality but cannot detect difference across datasets; and (b) S.Lasso, which applies Lasso to each dataset separately, and then, the results are combined and compared. This is virtually a meta-analysis strategy, allows difference, but cannot encourage commonality. For CoFu and the alternatives, we are mainly interested in two identification accuracy. The first, as in many other studies, is the identification of nonzero effects. The second, which is unique to this study, is the identification of communities that behave the same/differently in different datasets. More precisely, for a given community, if the L_2 norm of the difference between the regression coefficient vectors of two adjacent datasets is less than 0.01, we conclude commonality; otherwise, difference is concluded. The identified commonality/difference is compared against the true. For the proposed as well as alternatives, tuning parameter values affect identification performance. To get a more comprehensive view and minimize the (possibly different) impact of tuning on different methods, we follow the literature, consider a sequence of tunings, and evaluate using the receiver operating characteristic approach, under which the area under the curve (AUC) is the measure of identification accuracy (more details in Appendix C in the supporting information). Considering that, in practice, a definitive set of results may be desirable, we also evaluate identification results using true/false positive rates with tunings selected using five-fold CV. In addition, we evaluate estimation and prediction performance. Specifically, estimation is quantified by using estimation root mean square error (RMSE), which is defined as $\sqrt{\sum_k \left\| \boldsymbol{\beta}^k - \hat{\boldsymbol{\beta}}^k \right\|_2^2}$, and prediction is quantified by using prediction RMSE, which is defined as $\sqrt{\sum_k \left\| \boldsymbol{y}^k - \boldsymbol{X}^k \hat{\boldsymbol{\beta}}^k \right\|_2^2}$.

We simulate 100 replicates for each setting. For the settings with all nonzero coefficients equal to 0.5, we show the identification of nonzero effects in Table 1 and differentiation of communities with commonality/difference in Table 2, where the AUC summaries are presented. Results for the other settings are shown in Appendix B in the supporting information. The proposed CoFu is observed to have favorable performance in identifying nonzero effects. Consider for example Table 1. With $r = 100, (\rho_a, \rho_h, \rho_n) = (0.4, 0.1, 0.5)$, and the structured correlation, the mean AUCs are 0.746 (P.Lasso), 0.775 (S.Lasso), and 0.823 (CoFu), respectively. Performance of S.Lasso is not strongly affected by the overlapping across datasets, as it analyzes each dataset separately. In general, S.Lasso has good performance; however, it is inferior to CoFu under most settings. P.Lasso has superior performance when the sets of important effects in the three datasets almost entirely overlap, for example, when $(\rho_a, \rho_h, \rho_n) = (0.9, 0, 0.1)$. However, as expected, its performance deteriorates significantly when there are large differences across datasets. It is noted that CoFu still has competitive performance even under the worst case scenario, thus providing a "safe" choice in practice. Similar observations are also made under the other settings as presented in Appendix B in the supporting information. In the differentiation of community commonality/difference, note that as P.Lasso is not capable of identifying differences across datasets, its results are not presented. Simulation shows that CoFu significantly outperforms S.Lasso. For example in Table 2 with $r = 100$ and $(\rho_a, \rho_h, \rho_n) = (0.1, 0, 0.9)$, CoFu has AUCs 0.739 (structured), 0.762 (unstructured), and 0.724 (independence), respectively, while S.Lasso has AUCs below 0.6. For the settings presented in Appendix B in the supporting information, the observed patterns are similar. The results with tunings selected using CV are presented in Table B5-B7 (Appendix B in the supporting information). The CoFu method is observed to have superior performance in terms of identification, estimation, and prediction.

Simulation is also conducted under the Logit model, a representative of generalized linear models. Details of the estimation procedure are presented in Appendix A in the supporting information. In simulation, covariates are generated in the same way as described above. The response values are generated from the Logit model and Bernoulli distribution. The identification results measured by AUCs are presented in Tables B8 and B9 (Appendix B in the supporting information). The CoFu method is observed to have similar superior identification performance as under the LR model. It is noted that the improvement over the alternatives may not be as large as under the LR model. To this end, we conduct a nonparametric test on the paired AUC values and find that the improvement is statistically significant for all scenarios. For example, for $r = 100$, $(\rho_a, \rho_h, \rho_n) = (0.1, 0, 0.9)$,and from $\mathscr{U}[0.2, 1]$, the differences between the CoFu's and S.Lasso's AUC values have p-values 0.002 and $< 10^{-9}$ for nonzero effect and community identification, respectively.

4 | DATA ANALYSIS

TCGA is a collaborative effort organized by NCI and has recently published high-quality profiling data on multiple cancer types. The analysis of TCGA data has led to interesting findings. In our analysis, both clinical and genetic data are downloaded from the cBioPortal website.

4.1. Analysis of cutaneous melanoma data

We first consider the SKCM (cutaneous melanoma) data.[10] As in the literature,[19] the inclusion criteria are (1) white patients, (2) no neo-adjuvant therapy before tumor sample collection, (3) the type of skin upon which melanoma arose is nonglabrous skin, (4) no missing values in Breslow thickness and AJCC pathologic tumor stage, and (5) with gene expression measurements. In our analysis, we are interested in the regulation of Breslow thickness, which is an important prognostic marker, by gene expressions. In published studies,[20] similar analysis has been conducted, however, *using samples of all tumor stages* *and with insufficient attention to the potential difference across stages*. Partly motivated by Figure 2, our “hypothesis” is that the regulation relationships for different stages have commonality as well as difference. There are a total of 240 samples, with 70 in stage I, 60 in stage II, and 110 in stages III and IV.

A total of 18 947 gene expression measurements are available. From the KEGG pathway database downloaded from the Broad Institute, we identify 5266 unique genes, representing 186 pathways. Matching those gene names with those in the SKCM dataset, we identify 4243 genes for downstream analysis. Although, in principle, it is possible to directly apply the proposed approach to these genes, to obtain more reliable analysis results, we further conduct a supervised screening. Specifically, we compute the Pearson correlation coefficient of each gene with the response variable and identify those with p-values less than 0.05. A total of 973 genes are identified. We construct a dense network based on correlations and then generate a sparse one by filtering out edges that are not statistically significant at the 0.05 level.[21] The resulted network has 15 891 edges (detailed structure available from the authors). The Louvain method,[22] which performs a greedy optimization of community identification in a hierarchical manner, is applied and identifies 46 communities.

The analysis results of CoFu are summarized in Figure 3. Briefly, a total of 21 communities, with 126 genes, are identified as associated with the response. Among them, eight communities behave the same across the three stages, and the rest behave differently. More detailed estimation results are available from the authors. Simply eyeballing Figure 3 suggests that some communities, for example, 22, 39, and 45, demonstrate significantly different stage-specific properties. Such differences have not been well noted in the literature and deserve additional attention. Literature search suggests that the CoFu-identified genes may have important implications. For example, the gene that has the strongest signal in all three stages is ARMC, armadillo repeat containing 2. It belongs to a family of Armadillo repeat proteins, which play important roles in cell-cell adhension, cytoskeletal regulation, and intracellular signaling. Other genes that also have strong signals in all three stages include C10orf114 and KTGNR. C10orf114 is also known as CASC10. Over expression of C10orf114 is associated with poor survival in glioma and urothelial cancer (www.proteinatlas.org/ENSG00000204682-CASC10/cell). KTGNR (DNAH5) encodes an axonemal heavy chain of dynein proteins. It works as a force-generating protein with ATPase activity. The TRA2B-DNAH5 fusion has been identified as a novel oncogenic driver in lung cancer.[23] In addition to the genes that are associated with Breslow depth in all three stages, we have also identified genes that have strong signals only in the earlier or later stages. The top three genes that only have strong signals in stage I are DKFZP434B094, AK3, and ANKS1A. AMN and PDCR are found to have strong signals only in stage II. FAM219B and BSDC1 are found to be stage III and IV specific. In addition, genes that are more associated with Breslow depth in later stages are LOC392331, ANKRD20A20P, CRELD2, CEBPG, and others. CEBPG is one of the C/EBP transcription factors that regulate cell growth and differentiation of various tissues. One study has shown that CEBPG is a suppressor of myeloid differentiation in acute myeloid leukemia.[24]

Different findings are generated by the alternatives. The estimated coefficients are plotted in Figures B3 (S.Lasso) and B4 (P.Lasso) in Appendix B in the supporting information.

S.Lasso identifies 38 communities, with 176 genes, as associated with the response variable, and P.Lasso identifies 39 communities, with 105 genes. With their particular properties, S.Lasso identifies all communities (with nonzero effects) as behaving differently across stages, while P.Lasso identifies all communities as behaving the same. Biologically speaking, the CoFu results, which have both commonality and difference, are more sensible.

We also evaluate prediction performance and stability of each method. Specifically, each dataset is randomly divided into a training set and a testing set, with sizes 2:1. The regression parameters are estimated only using the training set and used to make prediction for the testing set. We use the RMSE of the response variable to measure prediction. For the three methods, the RMSEs are calculated as 1.077 (P.Lasso), 1.095 (S.Lasso), and 1.005 (CoFu). In addition, for each gene, we compute the proportion of being identified in 100 resamplings, which has been referred to as the Observed Occurrence Index (OOI) in the literature, to measure the stability of this gene. The highest 20 OOIs are shown in Figure B7 (Appendix B in the supporting information). CoFu has OOIs (0.823) higher than P.Lasso (0.782) and S.Lasso (0.711).

4.2. Analysis of lung cancer data

In TCGA, there are two lung cancer datasets, on Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC), respectively. In the literature, they have been separately analyzed,[25,26] and differences have been acknowledged.[27] However, as they are both non-small cell lung carcinomas, certain commonality is expected. For both LUAD and LUSC, the inclusion criteria are (1) no neo-adjuvant therapy before tumor sample collection, (2) in stage I of the AJCC pathologic tumor stage measurement, and (3) with FEV1 (forced expiratory volume in 1 second, prebroncholiator) and gene expressions measured. More details on sample selection are provided in Figure B2 (Appendix B in the supporting information). In this analysis, the response variable is FEV1, a critical measure of lung function. The final sample sizes are 142 (LUAD) and 89 (LUSC), respectively. For gene expressions, we conduct a similar processing as described above. Specifically, 20 531 gene expressions are initially available for analysis. Matching with the KEGG pathway information leads to 4243 genes. The p-value based marginal screening further reduces the number of gene expressions to 901. Using the same approach as described above, 41 communities are constructed.

The analysis results of CoFu are summarized in Figure 4. A total of 26 communities, with 54 genes, are identified as associated with the response. Among them, 13 communities behave the same for both LUAD and LUSC. Figure 4 suggests that some communities, for example, 9 and 13, behave significantly differently for LUAD and LUSC. Literature review again suggests that the findings are biologically sensible. Specifically, among the identified genes, CCNG2 has the strongest signal for both LUAD and LUSC. CCNG2 encodes protein cyclin G2, which has been shown to be a tumor suppressor in several studies.[28,29] A recent study shows that lung cancer patients with higher CCNG2 expressions had longer overall survival. Another gene that has a strong signal in both cancers is BRI3BP, BRI3-binding protein. Over expressions of BRI3BP are found to promote drug induced apoptosis via cross talking between mitochondria and endoplasmic reticulum. We also identify 24 genes that are associated with the response in LUAD but not LUSC. Among them, Chemokine-like receptor 1 (CMKLR1) has the strongest signal. CMLKLR1 is a transmembrane multifunctional receptor and found to play an important role in inflammatory. One genetic variant of CMLKLR1, rs1878022, is found to be significantly associated with poorer survival in advanced stage non-small cell lung cancer.[30] Another gene that we find to be associated with lung function in LUAD only is DLGAP5, DLG associated protein 5. DLGAP5 is a mitotic spindle protein and involved in mitosis processes. A study by Schneider et al[31] shows that DLGAP5 expression is higher in lung tumor tissues. Lung cancer patients with the overexpression of DLGAP5 tend to have poorer survival. Genes that have strong signals in LUSC but not LUAD include CALHM2, BTBD3, CLPP, and others. CALHM2 encodes protein calcium homeostasis modulator family member 2, which plays a critical role in the modulation of neural activity via ATP-releasing channel. The BTB domain-containing 3 (BTBD3) gene is found to be upregulated in Hepatocellular carcinoma tissues.[32] The genetic variant of BTBD3 is also found to be significantly associated with survival in non–small cell lung cancer patients.[33] CLPP, caseinolytic mitochondrial matrix peptidase proteolytic subunit, belongs to the peptidase family S14. Its function is to hydrolyze proteins into small peptides in the mitochondria matrix. Increased protein expression is found in type I endometrial cancer patients.[34]

The alternative analysis results are presented in Figures B5 (S.Lasso) and B6 (P.Lasso) in Appendix B in the supporting information. S.Lasso identifies 27 communities with 44 genes. No commonality is identified. P.Lasso identifies 31 communities with 44 genes, all of which behave the same for the two cancers. In terms of prediction, the CoFu RMSE is 0.917, lower than S.Lasso (0.999) and P.Lasso (1.023). The OOI results are represented in Figure B7 (Appendix B in the supporting information). CoFu has the highest OOIs among the three methods.

4.3. Simulation

It has been recognized in some studies that simulated data may be “simpler” than real data. Here, we conduct an additional set of simulation based on the SKCM data analyzed above. Specifically, the observed gene expression data and community structure are used in simulation. The structure of important covariate effects is the same as described above in Section 3. The identification results for community and individual effects are summarized in Table B10 (Appendix B in the supporting information). Although there are some small numerical differences, the observed patterns are similar to those in Section 3, providing a strong support to the effectiveness of the proposed method.

5 | DISCUSSION

In the literature, although the commonality and difference of “related” cancers have been noted, effective analysis methods are still lacking. This study fills this knowledge gap by developing a novel community fusion method. The CoFu method has an intuitive formulation and can be effectively realized. Although sharing some similar spirits with the existing fused and contrasted penalization methods, it also has significant advancements by conducting the integrative analysis of multiple datasets, promoting commonality as opposed to similarity, and accommodating the community structure among genes. Numerical studies have demonstrated its superiority over the direct competitors. We have mostly described the proposed method under the LR model. As described in Appendix A in the supporting information as well as in the simulation, the proposed method can be extended to the Logit model, a representative of generalized linear models. A closer examination of the penalty function and computation suggests that the CoFu method can be potentially coupled with others, for example prognosis, outcomes and models. More complicated penalties can take the place of Lasso. For example, when it is desirable to accommodate network adjacency, Laplacian penalties can be further imposed. The group Lasso-type fusion penalty can also be replaced by other group penalties. In the first data analysis, the "stratification" variable is cancer stage, which has a sound biological basis. It should be noted that, as demonstrated in simulation, the proposed method can accommodate different degrees of commonality/ difference. As such, the choice of the stratification variable is not critical. In fact, the proposed method can be applied to "test" whether the response-omics relationships are the same with respect to a specific stratification variable.

Limitations of this study may include a lack of theoretical investigation and deeper bioinformatics analysis of the data analysis results. Such pursuit will be deferred to future research. We will also defer possible extensions as discussed above to future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank the associate editor and two reviewers for careful review and insightful comments, which have led to a significant improvement of this article. We thank Pan Zhang for sharing computer code, which helps data generation in simulation. This study was supported by the National Natural Science Foundation of China (grants 11605288 and 71771211), by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (16YJCZH088), by the Fund for building world-class universities (disciplines) of Renmin University of China, by the National Bureau of Statistics of China (2016LD01), and by the National Institute of Health (CA204120 and CA216017).

Funding information

National Natural Science Foundation of China, Grant/Award Number: 11605288 and 71771211; MOE (Ministry of Education in China) Project of Humanities and Social Sciences, Grant/Award Number: 16YJCZH088; Fund for building world-class universities (disciplines) of Renmin University of China; National Bureau of Statistics of China, Grant/Award Number: 2016LD01; National Institutes of Health, Grant/Award Number: CA204120 and CA216017

REFERENCES

1. Weinstein JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–1120. [PubMed: 24071849]
2. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Sci Rep. 2013;3:2650. [PubMed: 24084849]
3. Guerra R, Goldstein DR. Meta-analysis and Combining Information in Genetics and Genomics. Boca Raton, FL: Chapman & Hall/CRC; 2009.
4. Ma SG, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. Stat Med. 2011;30(28):3361–3371. [PubMed: 22105693]

5. Huang Y, Zhang Q, Zhang S, Huang J, Ma S. Promoting similarity of sparsity structures in integrative analysis With penalization. J Am Stat Assoc. 2016;112(517):342–350.

6. Rhodes DR, Yu J, Shanker K, et al. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. ProcNatlAcad Sci USA. 2004;101(25):9309–9314.

7. Habermann JK, Paulsen U, Roblick UJ, et al. Stage-specific alterations of the genome, transcriptome, and proteome during colorectal carcinogenesis. Genes Chromosom Cancer. 2007;46(1):10–26. [PubMed: 17044061]

8. Cheng Y, Lu J, Chen GD, et al. Stage-specific prognostic biomarkers in melanoma. Oncotarget. 2015;6(6):4180. [PubMed: 25784655]

9. Palaniappan A, Ramar K, Ramalingam S. Computational identification of novel stage-specific biomarkers in colorectal cancer progression. PloS one. 2016;11(5). Article ID e0156665.

10. Akbani R, Akdemir KC, Aksoy BA, et al. Genomic classification of cutaneous melanoma. Cell. 2015;161(7):1681–1696. [PubMed: 26091043]

11. Krauthammer M, Kong Y, Ha BH, et al. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. Nat Genet. 2012;44(9):1006–1014. [PubMed: 22842228]

12. Chadeau-Hyam M, Campanella G, Jombart T, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen. 2013;54(7):542–557. [PubMed: 23918146]

13. Ang JC, Mirzal A, Haron H, Hamed HNA. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. IEEE/ACM Trans Comput Biol Bioinform. 2016;13(5):971–989. [PubMed: 26390495]

14. Huang Y, Liu J, Yi HD, Shia BC, Ma S. Promoting similarity of model sparsity structures in integrative analysis of cancer genetic data. Stat Med. 2017;36(3):509–559. [PubMed: 27667129]

15. Shi XJ, Liu J, Huang J, Zhou Y, Shia B, Ma S. Integrative analysis of high-throughput cancer studies with contrasted penalization. Genet Epidemiol. 2014;38(2):144–151. [PubMed: 24395534]

16. Zhang W, Chien J, Yong J, Kuang R. Network-based machine learning and graph theory algorithms for precision oncology. NPJ Precis Oncol. 2017;1(1):25. [PubMed: 29872707]

17. Hestenes MR. Multiplier and gradient methods. J Optim Theory Appl. 1969;4(5):303–320.

18. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Phys Rev E. 2011;83(1). Article ID 016107.

19. Jiang Y, Shi XJ, Zhao Q, Krauthammer M, Rothberg BEG, Ma S. Integrated analysis of multidimensional omics data on cutaneous melanoma prognosis. Genomics. 2016;107(6):223–230. [PubMed: 27141884]

20. Chai H, Shi XJ, Zhang QZ, Zhao Q, Huang Y, Ma S. Analysis of cancer gene expression data with an assisted robust marker identification approach. Genet Epidemiol. 2017;41(8):779–789. [PubMed: 28913902]

21. Serrano MA, Boguna M, Vespignani A. Extracting the multiscale backbone of complex weighted networks. Proc Natl Acad Sci USA. 2009;106(16):6483–6488. [PubMed: 19357301]

22. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;2008(10). Article ID P10008.

23. Li F, Fang Z, Zhang J, et al. Identification of TRA2B-DNAH5 fusion as a novel oncogenic driver in human lung squamous cell carcinoma. Cell Res. 2016;26(10):1149–1164. [PubMed: 27670699]

24. Alberichjorda M, Wouters B, Balastik M, et al. C/EBPγ deregulation results in differentiation arrest in acute myeloid leukemia. J Clin Invest. 2012;122(12):4490–4504. [PubMed: 23160200]

25. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511(7511):543–550. [PubMed: 25079552]

26. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012;489(7417):519–525. [PubMed: 22960745]

27. Sun FH, Yang XD, Jin YL, et al. Bioinformatics analyses of the differences between lung adenocarcinoma and squamous cell carcinoma using The Cancer Genome Atlas expression data. Mol Med Rep. 2017;16(1):609–616. [PubMed: 28560415]

28. Wang S, Zeng Y, Zhou JM, et al. MicroRNA-1246 promotes growth and metastasis of colorectal cancer cells involving CCNG2 reduction. Mol Med Rep. 2016;13(1):273–280. [PubMed: 26573378]

29. Yao D, Cui H, Zhou S, Guo L. Morin inhibited lung cancer cells viability, growth, and migration by suppressing miR-135b and inducing its target CCNG2. Tumour Biol. 2017;39(10). Article ID 101042831771244.

30. Wu XF, Ye YQ, Rosell R, et al. Genome-wide association study of survival in non-small cell lung cancer patients receiving platinum-based chemotherapy. JNatl Cancer Inst. 2011;103(10):817–825. [PubMed: 21483023]

31. Schneider MA, Christopoulos P, Muley T, et al. AURKA, DLGAP5, TPX2, KIF11 and CKAP5: Five specific mitosis-associated genes correlate with poor prognosis for non-small cell lung cancer patients. IntJ Oncol. 2017;50(2):365–372. [PubMed: 28101582]

32. Xiao W, Zhao W, Li L, et al. Preliminary investigation of the role of BTB domain-containing 3 gene in the proliferation and metastasis of hepatocellular carcinoma. Oncol Lett. 2017;14(2):2505–2510. [PubMed: 28789460]

33. Wu C, Xu B, Yuan P, et al. Genome-wide examination of genetic variants associated with response to platinum-based chemotherapy in patients with small-cell lung cancer. Pharmacogenet Genomics. 2010;20(6):389–395. [PubMed: 20463552]

34. Cormio A, Musicco C, Gasparre G, et al. Increase in proteins involved in mitochondrial fission, mitophagy, proteolysis and antioxidant response in type I endometrial cancer as an adaptive response to respiratory complex I deficiency. Biochem Biophys Res Commun. 2017;491(1):85–90. [PubMed: 28698145]

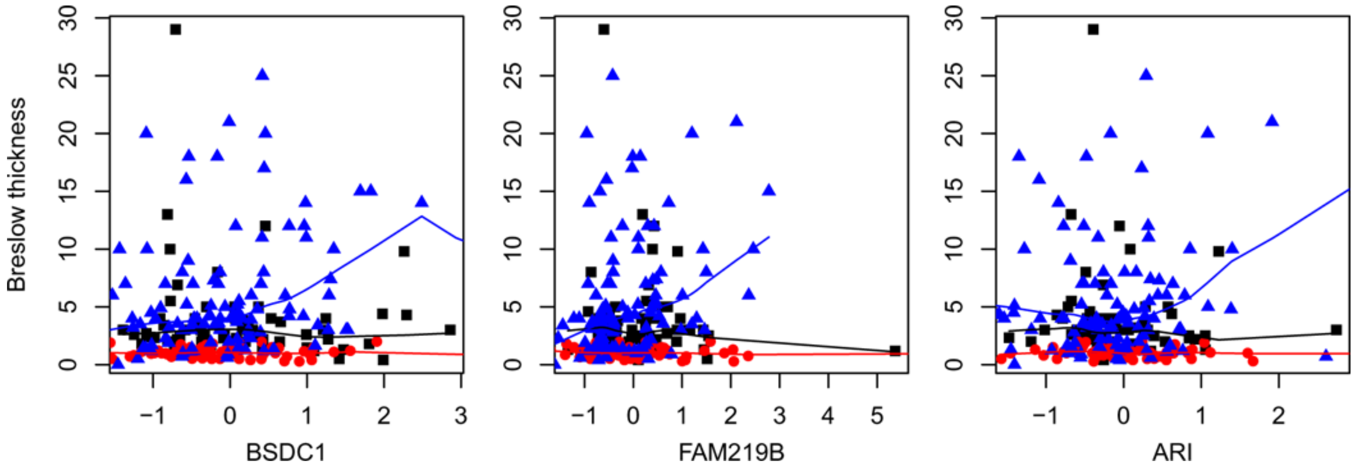

FIGURE 1.

Analysis of the TCGA melanoma data: Breslow thickness against the expressions of three genes. Three colors correspond to three cancer stages [Colour figure can be viewed at wileyonlinelibrary.com]

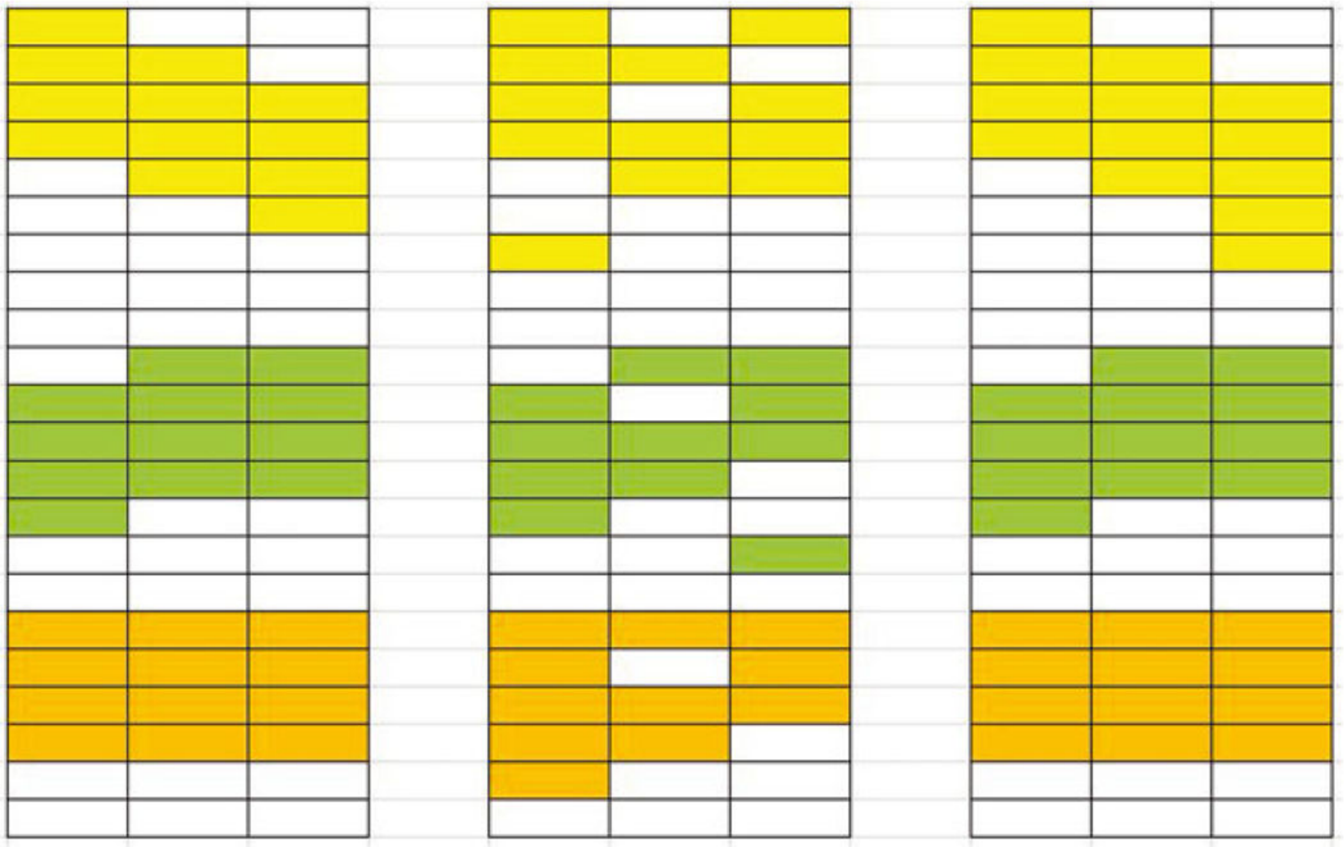

FIGURE 2.

Scheme of analysis. Left: true model structure. Middle: analysis by an alternative. Right: CoFu analysis [Colour figure can be viewed at wileyonlinelibrary.com]

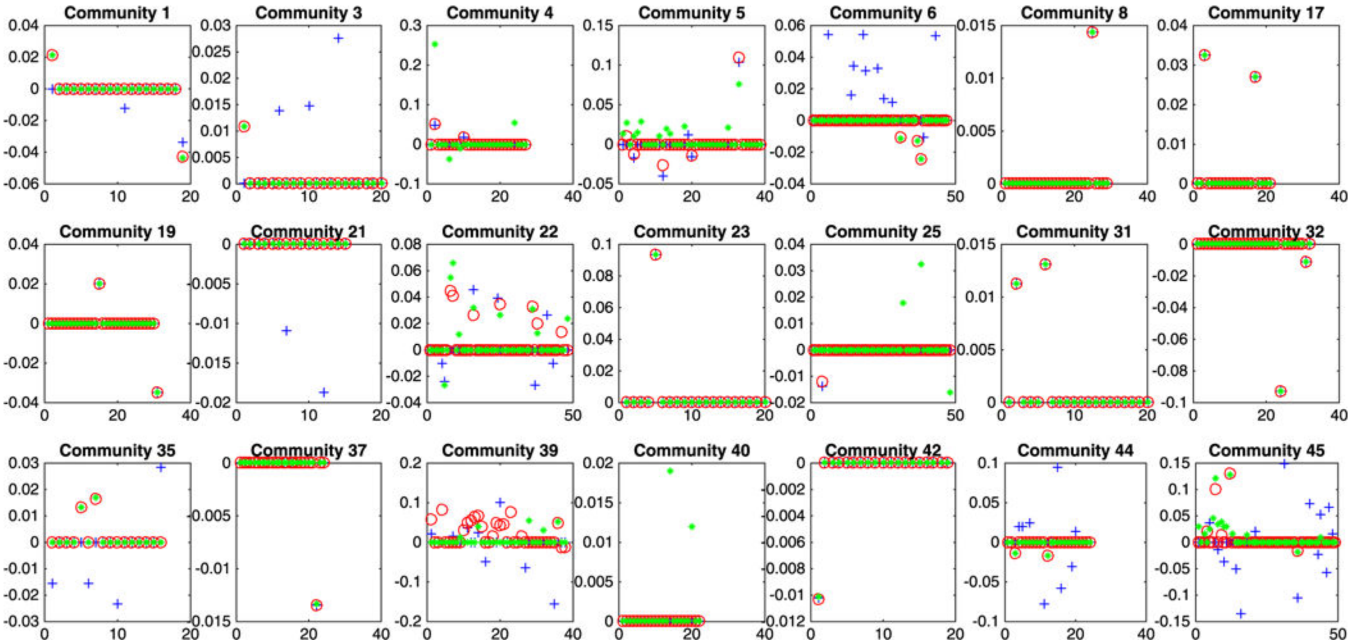

FIGURE 3.

Analysis of the TCGA SKCM data. Blue crosses correspond to stage I, red circles to stage II, and green filled circles to stage III and IV [Colour figure can be viewed at wileyonlinelibrary.com]

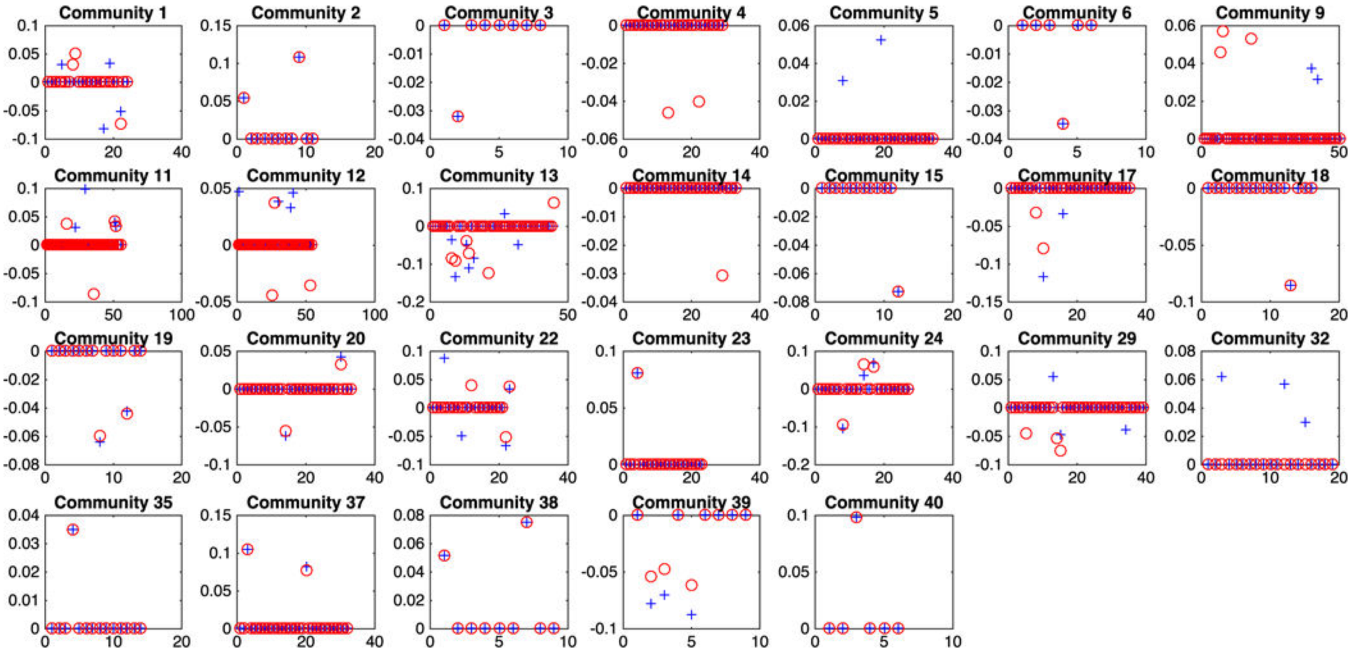

FIGURE 4.

Analysis of the TCGA lung cancer data. Blue crosses correspond to LUAD, and red circles to LUSC. [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1

Simulation under the linear regression model: mean(sd) of area under the curve for effect identification. All nonzero coefficients are equal to 0.5

		***r* = 100**			***r* = 150**	
	Structured	**Unstructured**	**Independence**	**Structured**	**Unstructured**	**Independence**
(ρ_a,ρ_b,ρ_n) = (0.1,0,0.9)						
P.Lasso	0.619(0.014)	0.593(0.015)	0.62(0.013)	0.578(0.017)	0.574(0.013)	0.579(0.018)
S.Lasso	0.77(0.009)	0.833(0.01)	0.761(0.012)	0.695(0.015)	0.777(0.01)	0.685(0.012)
CoFu	0.771(0.01)	0.813(0.012)	0.753(0.012)	0.694(0.012)	0.762(0.012)	0.676(0.013)
(ρ_a,ρ_b,ρ_n)= (0.1,0.9,0)						
P.Lasso	0.732(0.015)	0.743(0.02)	0.725(0.015)	0.621(0.02)	0.703(0.012)	0.646(0.021)
S.Lasso	0.781(0.013)	0.85(0.012)	0.782(0.012)	0.689(0.013)	0.787(0.007)	0.687(0.013)
CoFu	0.827(0.013)	0.891(0.008)	0.814(0.009)	0.73(0.012)	0.8(0.012)	0.716(0.015)
(ρ_a,ρ_b,ρ_n)= (0.2, 0.6, 0.2)						
P.Lasso	0.752(0.017)	0.719(0.023)	0.744(0.017)	0.689(0.018)	0.666(0.014)	0.68(0.016)
S.Lasso	0.772(0.011)	0.847(0.011)	0.76(0.011)	0.699(0.012)	0.786(0.008)	0.689(0.01)
CoFu	0.838(0.012)	0.878(0.007)	0.826(0.01)	0.762(0.008)	0.803(0.01)	0.753(0.011)
(ρ_a,ρ_b,ρ_n) = (0.4, 0.1, 0.5)						
P.Lasso	0.746(0.012)	0.741(0.012)	0.739(0.012)	0.698(0.013)	0.684(0.01)	0.69(0.014)
S.Lasso	0.775(0.007)	0.831(0.013)	0.752(0.013)	0.691(0.016)	0.775(0.008)	0.681(0.017)
CoFu	0.823(0.009)	0.865(0.011)	0.817(0.008)	0.756(0.017)	0.795(0.011)	0.746(0.016)
(ρ_a,ρ_b,ρ_n)= (0.5, 0.5, 0)						
P.Lasso	0.885(0.013)	0.866(0.007)	0.885(0.01)	0.82(0.019)	0.799(0.017)	0.812(0.017)
S.Lasso	0.776(0.011)	0.834(0.009)	0.764(0.015)	0.692(0.015)	0.781(0.009)	0.69(0.009)
CoFu	0.894(0.011)	0.893(0.005)	0.891(0.011)	0.817(0.009)	0.839(0.011)	0.816(0.01)
(ρ_a,ρ_b,ρ_n)= (0.6, 0.2, 0.2)						
P.Lasso	0.878(0.011)	0.847(0.013)	0.873(0.009)	0.831(0.022)	0.779(0.016)	0.809(0.024)
S.Lasso	0.762(0.012)	0.831(0.011)	0.76(0.013)	0.692(0.016)	0.777(0.009)	0.684(0.018)
CoFu	0.887(0.011)	0.888(0.009)	0.881(0.008)	0.822(0.017)	0.829(0.011)	0.827(0.019)
(ρ_a,ρ_b,ρ_n)= (0.9, 0, 0.1)						
P.Lasso	0.89(0.014)	0.925(0.02)	0.872(0.015)	0.832(0.017)	0.884(0.012)	0.828(0.02)
S.Lasso	0.768(0.009)	0.845(0.012)	0.774(0.012)	0.691(0.013)	0.785(0.008)	0.685(0.015)
CoFu	0.876(0.013)	0.899(0.008)	0.859(0.01)	0.805(0.012)	0.864(0.012)	0.798(0.014)

TABLE 2

Simulation under the linear regression model: mean(sd) of area under the curve for community differentiation. All nonzero coefficients are equal to 0.5

		***r* = 100**			***r* = 150**	
	Structured	**Unstructured**	**Independence**	**Structured**	**Unstructured**	**Independence**
(ρ_a,ρ_b,ρ_n) = (0.1,0,0.9)						
S.Lasso	0.578(0.035)	0.54(0.097)	0.571(0.066)	0.565(0.078)	0.515(0.105)	0.537(0.068)
CoFu	0.739(0.033)	0.762(0.038)	0.724(0.042)	0.711(0.042)	0.74(0.042)	0.679(0.039)
(ρ_a,ρ_b,ρ_n) = (0.1,0.9,0)						
S.Lasso	0.543(0.057)	0.54(0.092)	0.53(0.066)	0.496(0.088)	0.468(0.105)	0.484(0.071)
CoFu	0.712(0.038)	0.762(0.036)	0.724(0.043)	0.692(0.042)	0.691(0.062)	0.656(0.052)
(ρ_a,ρ_b,ρ_n) = (0.2,0.6,0.2)						
S.Lasso	0.573(0.082)	0.54(0.105)	0.549(0.074)	0.517(0.066)	0.52(0.119)	0.509(0.081)
CoFu	0.744(0.032)	0.721(0.05)	0.704(0.048)	0.679(0.049)	0.71(0.053)	0.679(0.042)
(ρ_a,ρ_b,ρ_n) = (0.4,0.1,0.5)						
S.Lasso	0.565(0.043)	0.541(0.091)	0.557(0.061)	0.538(0.063)	0.531(0.1)	0.552(0.06)
CoFu	0.749(0.031)	0.746(0.032)	0.752(0.038)	0.724(0.049)	0.737(0.066)	0.73(0.057)
(ρ_a,ρ_b,ρ_n) = (0.5,0.5,0)						
S.Lasso	0.533(0.048)	0.49(0.103)	0.552(0.062)	0.476(0.099)	0.492(0.124)	0.513(0.084)
CoFu	0.777(0.043)	0.771(0.029)	0.779(0.046)	0.728(0.038)	0.724(0.036)	0.729(0.03)
$(\rho_a,\rho_b,\rho n)$= (0.6,0.2,0.2)						
S.Lasso	0.515(0.036)	0.5(0.096)	0.53(0.043)	0.438(0.093)	0.477(0.1)	0.486(0.087)
CoFu	0.755(0.053)	0.755(0.041)	0.779(0.04)	0.685(0.051)	0.721(0.046)	0.694(0.033)
(ρ_a,ρ_b,ρ_n) = (0.9,0,0.1)						
S.Lasso	0.518(0.066)	0.498(0.098)	0.54(0.07)	0.466(0.078)	0.441(0.098)	0.45(0.076)
CoFu	0.965(0.042)	0.965(0.048)	0.975(0.042)	0.947(0.045)	0.907(0.048)	0.961(0.048)