BMC Bioinformatics

METHODOLOGY ARTICLE

Open Access

CrossMark

Identification of gene pairs through penalized regression subject to constraints

Rex Shen[1†], Lan Luo[2†] and Hui Jiang[2*] (iD)

Abstract

Background: This article concerns the identification of gene pairs or combinations of gene pairs associated with biological phenotype or clinical outcome, allowing for building predictive models that are not only robust to normalization but also easily validated and measured by qPCR techniques. However, given a small number of biological samples yet a large number of genes, this problem suffers from the difficulty of high computational complexity and imposes challenges to the accuracy of identification statistically.

Results: In this paper, we propose a parsimonious model representation and develop efficient algorithms for identification. Particularly, we derive an equivalent model subject to a sum-to-zero constraint in penalized linear regression, where the correspondence between nonzero coefficients in these models is established. Most importantly, it reduces the model complexity of the traditional approach from the quadratic order to the linear order in the number of candidate genes, while overcoming the difficulty of model nonidentifiablity. Computationally, we develop an algorithm using the alternating direction method of multipliers (ADMM) to deal with the constraint. Numerically, we demonstrate that the proposed method outperforms the traditional method in terms of the statistical accuracy. Moreover, we demonstrate that our ADMM algorithm is more computationally efficient than a coordinate descent algorithm with a local search. Finally, we illustrate the proposed method on a prostate cancer dataset to identify gene pairs that are associated with pre-operative prostate-specific antigen.

Conclusion: Our findings demonstrate the feasibility and utility of using gene pairs as biomarkers.

Keywords: Gene pair, Biomarker, Penalized regression, ADMM

Background

In biomedical research, gene identification has been critical towards understanding and predictive modeling, whose activities are associated with biological phenotype, disease status, or clinical outcome. These genes, referred to as biomarkers, are further utilized for predictive modeling to facilitate scientific investigation, clinical diagnosis, prognosis, and treatment developments. In this discovery process, the expression levels of candidate genes are measured through genomic techniques enabling thousands of genes simultaneously, permitting monitoring molecular variation on a genome-wide scale [1] and providing more precise and reliable diagnosis [2]. As widely used techniques, DNA microarray, parallel qPCR, and RNA-Seq measure gene expression at the mRNA level. Yet, two major issues emerge with regard to the utilization of gene expression. First, the number of genes greatly exceeds that of biological samples typically, with tens of thousands of genes in the presence of up to only a few hundred biological samples or observations. As a result, inference tends to be unstable, misleading, or even invalid due to high statistical uncertainty, in addition to extremely high cost of computation. This, in turn, demands reliable and accurate methods of identification. Second, prior to any analysis, raw gene expressions must be normalized to compensate for differences in labeling, sample preparation, and detection methods. A common practice focuses on normalization of each sample's raw expression based on remaining ones in the same dataset, known as between-sample normalization, often in the forms of sample-wise scaling in RNA-Seq data [3]. However, such a normalization requires recomputation when a new sample is

*Correspondence: jianghui@umich.edu
†Equal contributors
2Department of Biostatistics, University of Michigan, MI 48109 Ann Arbor, USA
Full list of author information is available at the end of the article

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 2 of 11

removed or added from the dataset, imposing computational challenges for large studies. Moreover, any analysis using selected genes based on one dataset may be sensitive to normalization, leading to non-generalizable and/or non-reproducible scientific findings [4].

To address the foregoing challenges, a modeling method based on gene pairs was first presented in the top-scoring pair (TSP) classifier by [5] and later implemented by [6]. Compared to predictors based on individual genes, gene pair-based predictors are more robust to normalization and have better predicting or classifying accuracy. Another advantage of gene-pair based predictive modeling is its ease of evaluation and validation by qPCR methods. Ideally, to use qPCR to measure a single gene's expression level, one applies the delta-Ct method [7], in which the differenced Ct values between the gene of interest and another housekeeping gene such as GAPDH measures gene expression level. However, between sample variation of a housekeeping gene may be large, imposing a great challenge [8]. In this sense, gene-pair based modeling removes the requirement of housekeeping genes since the differenced Ct values between the two genes of interest can be directly treated as a measurement. Consequently, the two genes in a gene pair serve as internal controls for each other. Due to all these advantages, gene pair-based predictors have been adopted in several cancer studies [9–11].

Despite the advantage of the gene pair approach, due to the combinatorial complexity, identifying the best gene pair, or best combinations of several gene pairs, is statistically and computationally challenging, from all the possible pairs from a pool of tens of thousands of genes. For instance, the TSP algorithm employs a direct search, whose running time grows quadratically in terms of the number of candidate genes. Although in practice one can first identify differentially expressed genes and then perform a restrictive search to these individual genes, such a two-step approach is no longer invariant to normalization and may miss informative pairs in which at most one gene is differentially expressed [5]. The computational problem is even more severe when more than one gene pair is sought, such as in *k*-TSP which involves exactly *k* top disjoint pairs in prediction [12].

Moreover, even though rank-based gene pair predictors like the TSP are robust to normalization, their utility in modeling complex data remains limited. One possible extension is to use ratios of gene expression levels as predictors and use regression models to select gene pairs. In recent years, regression models with penalties enforcing sparsity (such as the Lasso [13], SCAD [14], and TLP [15] penalties) have been widely used for variable selection, and many efficient algorithms have been proposed for fitting such models. One may employ such an approach by treating ratios of gene expression levels from all possible

gene pairs as candidate predictors. However, this amounts to a quadratic complexity in the number of candidate genes.

In this paper, we develop a new regression approach and an efficient algorithm for identifying gene pairs associated with biological phenotype or clinical outcome. we propose an equivalent model subject to a sum-to-zero constraint on regression coefficients, where the correspondence between nonzero coefficients in these models is established. The model of this type has been proposed for compositional data [16] and recently for reference point insensitive data [17]. One salient aspect is that this model is more parsimonious, involving only predictors linearly in the number of candidate genes. To deal with the constraint, we develop an efficient algorithm based on the alternating direction method of multipliers (ADMM) [18, 19], for identification and model parameter estimation. The new approach shares not only the benefit of simplicity in interpretation but also a linear complexity. Most importantly, the proposed method substantially improves the statistical accuracy and computation efficiency. Finally, in simulations, the method compares favorably against the traditional method in terms of the accuracy of identification, and our ADMM algorithm is more computationally efficient than a coordinate descent with local search (CD+LS) algorithm of [17].

Methods
High-dimensional linear regression
This section proposes predictive models based on combinations of ratios of gene expression levels on the ground that ratios of gene expression levels not only are robust to normalization but also can be easily validated and measured by qPCR techniques.

Given p predictors $(x_1, \ldots, x_p)$ measuring the expression levels of p genes $(g_1, \ldots, g_p)$, we consider informative second-order interactions defined by pairwise ratios $\{x_j/x_k, 1 \le j < k \le p\}$ of $(x_1, \ldots, x_p)$ with respect to a continuous response such as the pre-operative prostate-specific antigen level measured from prostate cancer patients, as demonstrated in the "Results" section. It is assumed that there are only a small number (i.e., much smaller than p) of informative genes. Now consider a regression model in which response Y depends on a predictor vector $\boldsymbol{x}$ in a linear fashion:

$$Y = f(\boldsymbol{z}) + \epsilon \equiv \boldsymbol{\alpha}^T \boldsymbol{z} + \epsilon; \quad \epsilon \sim N(0, \sigma^2); \qquad (1)$$

where $\boldsymbol{\alpha} = (\alpha_{12}, \alpha_{13}, \ldots, \alpha_{p-1p})^T$ and $\boldsymbol{z} = (\log(x_1/x_2), \log(x_1/x_3), \ldots, \log(x_{p-1}/x_p))^T$ are $q = \frac{p(p-1)}{2}$-dimensional vectors of regression coefficients and predictors, and ϵ is random error that is independent of $\boldsymbol{z}$. For convenience, for $i < j$, we let $\alpha_{ji} = -\alpha_{ij}$. In Eq. (1), primary reasons for the logarithm of the pairwise ratios $\{x_j/x_k, 1 \le j <$

Shen *et al. BMC Bioinformatics*   (2017) 18:466

Page 3 of 11

$k \leq p$} are two-fold. First, it stabilizes the variance of gene expression levels so that Eq. (1) is suitable. In fact, the logarithm transformation is widely used in the literature on gene expression modeling [20]. Second, it facilitates an efficient model fitting algorithm to be introduced subsequently. Our objective is to identify nonzero coefficients of $\boldsymbol{\alpha}$ corresponding to informative gene pairs based on gene expression.

There are several challenges for identification of informative ratios within the framework of Eq. (1), in which p may greatly exceed the sample size n, known as high-dimensional regression. Normally, one may apply a feature selection method such as the Lasso [13] for this task. Unfortunately, however, high-dimensionality of Eq. (1) impedes the accuracy of feature selection in the presence of noise in addition to computational cost, which are roughly proportional to p^2. To overcome these difficulties, we propose an alternative yet equivalent model of Eq. (1) through a more parsimonious representation involving one linear constraint.

The next proposition says that $f(\boldsymbol{z})$ in Eq. (1) has an equivalent representation with only p-variables. In a sense, it achieves the objective of dimensionality reduction.

Proposition 1 *The following equivalent form of $f(\boldsymbol{z})$ in Eq.* (1) *is as follows:*

$$f(\boldsymbol{z}) = \sum_{j=1}^{p} \beta_j \log x_j, \quad \beta_j = \sum_{k \neq j} \alpha_{jk}. \qquad (2)$$

Importantly, $\sum_{j=1}^{p} \beta_j = 0$.

Based on Proposition 1, we derive an equivalent model of Eq. (1):

$$Y = \boldsymbol{\beta}^T \tilde{\boldsymbol{x}} + \epsilon, \quad \sum_{j=1}^{p} \beta_j = 0, \quad \epsilon \sim N(0, \sigma^2); \qquad (3)$$

where $\tilde{\boldsymbol{x}} = (\log x_1, \ldots, \log x_p)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Most critically, the correspondence between coefficients of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is established by Eq. (2), where Eq. (1) and Eq. (3) can have different number of nonzero coefficients, which is because of the reparametrization and the sum-to-zero constraint. For instance, suppose that $\alpha_{12} = 3$, $\alpha_{23} = -2$, and $\alpha_{ij} = 0$ otherwise in Eq. (1), then $\beta_1 = 3$, $\beta_2 = -5$, $\beta_3 = 2$, and $\beta_j = 0$ otherwise in Eq. (3). Model Eq. (3) has been proposed for compositional data [16] and recently also for reference point insensitive data [17]. Here [16] established model selection consistency and bounds for the resulting estimator.

In contrast to Eq. (1), Eq. (3) contains only p instead of $\frac{p(p-1)}{2}$ predictors, subject to the sum-to-zero constraint

for the regression coefficients. In other words, model Eq. (3) is more parsimonious than model Eq. (1) in terms of the number of active parameters in a model. As a result, there can not be a one-to-one correspondence between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. It is shown in Eq. (2) that the value of $\boldsymbol{\beta}$ in Eq. (3) is uniquely determined by that of $\boldsymbol{\alpha}$ in Eq. (1). The inverse does not hold – many values of $\boldsymbol{\alpha}$ in Eq. (1) correspond to the same value of $\boldsymbol{\beta}$ in Eq. (3). The non-existence of one-to-one correspondence between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is due to the fact that model Eq. (1) is largely non-identifiable. In fact, for any cycle formed by sequence $i_1, i_2, \ldots, i_k, i_{k+1} = i_1$, we can add any constant c to the $\alpha_{i_j i_{j+1}}$'s formed by adjacent indices without changing the model. That is, we can construct $\boldsymbol{\alpha}'$ where $\alpha'_{i_j i_{j+1}} = \alpha_{i_j i_{j+1}} + c$ for $j = 1, \ldots, k$ and $\alpha'_{ij} = \alpha_{ij}$ otherwise and model Eq. (1) with $\boldsymbol{\alpha}$ is equivalent to that with $\boldsymbol{\alpha}'$. Therefore, when we obtain a solution $\boldsymbol{\beta}$ by solving Eq. (3), due to the argument above, there are an infinite number of $\boldsymbol{\alpha}$ that are related to $\boldsymbol{\beta}$ through Eq. (2). Among them, we would like to choose the "simplest" one. In this paper, we define the "simplest" $\boldsymbol{\alpha}$ to be the one(s) with the minimum L_1 norm, where the L_1-norm of a vector $\boldsymbol{y} = (y_1, \ldots, y_p)$ is $||\boldsymbol{y}||_1 = \sum_{i=1}^{p} |y_i|$. In practice, given an estimate of $\boldsymbol{\beta}$ from (3), an estimate of $\boldsymbol{\alpha}$ can be obtained using Algorithm 1 below.

Algorithm 1 (Peeling) *Given an estimate of $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p \right)^T$ satisfying the sum-to-zero constraint $\sum_{j=1}^{p} \hat{\beta}_j = 0$, initialize $\tilde{\boldsymbol{\beta}}$ as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ as $\hat{\alpha}_{j,k} = 0$ for all $1 \leq j < k \leq p$.*
Step 1: Identify one positive and one negative $\tilde{\beta}_j$'s, say $\tilde{\beta}_{k_1} > 0$ and $\tilde{\beta}_{k_2} < 0$, where k_1 and k_2 are two distinct indices from $\{1, \cdots, p\}$. For instance, $\tilde{\beta}_{k_1}$ and $\tilde{\beta}_{k_2}$ can be taken as the most positive and most negative (ties can be broken arbitrarily) $\tilde{\beta}_j$'s. This can always proceed as long as not all $\tilde{\beta}_j$'s are zero.
Step 2: Set $\hat{\alpha}_{k_1 k_2} = \min \left(|\tilde{\beta}_{k_1}|, |\tilde{\beta}_{k_2}| \right)$. For instance, $\tilde{\beta}_1 = 1.5$ and $\tilde{\beta}_2 = -0.5$, then set $\hat{\alpha}_{12} = 0.5$.
Step 3: Subtract $\hat{\alpha}_{k_1 k_2}$ from $\tilde{\beta}_{k_1}$ and $-\hat{\alpha}_{k_1 k_2}$ from $\tilde{\beta}_{k_2}$ to make one of them zero, that is, $\tilde{\beta}_{k_1} \leftarrow \tilde{\beta}_{k_1} - \hat{\alpha}_{k_1 k_2}$ and $\tilde{\beta}_{k_2} \leftarrow \tilde{\beta}_{k_2} + \hat{\alpha}_{k_1 k_2}$. In the previous example, $\tilde{\beta}_1 \leftarrow 1$ and $\tilde{\beta}_2 \leftarrow 0$.
Step 4: Repeat Steps 1–3 until all components of $\tilde{\boldsymbol{\beta}}$ become zero.

Algorithm 1 terminates in at most p steps because the number of nonzeros in $\tilde{\boldsymbol{\beta}}$ decreases by either 1 or 2 after each iteration. Note that $\tilde{\beta}_{k_1}$ and $\tilde{\beta}_{k_2}$ identified in Step 1 may not be unique. Therefore it may lead to different $\hat{\boldsymbol{\alpha}}$'s. Importantly, this algorithm always yields a minimum L_1-norm $\boldsymbol{\alpha}$ estimate (see Proposition 5 later in this section).

The following two propositions characterize properties of such $\boldsymbol{\alpha}$ with respect to its representations.

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 4 of 11

Proposition 2 (Minimum L_1-norm of $\boldsymbol{\alpha}$) *Given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfying Eq.* (2), *the following conditions are equivalent:*

(A) *For all $\boldsymbol{\alpha}'$ satisfying* (2), $||\boldsymbol{\alpha}||_1 \le ||\boldsymbol{\alpha}'||_1$.
(B) *For all* $1 \le i, j, k \le p, i \ne j, j \ne k, \alpha_{ij}\alpha_{jk} \le 0$.
(C) $||\boldsymbol{\alpha}||_1 = \frac{1}{2}||\boldsymbol{\beta}||_1$.

Proposition 3 (Uniqueness of the representation of $\boldsymbol{\alpha}$) *Given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfying Eq.* (2), *the following conditions are equivalent:*

(D) *For all $\boldsymbol{\alpha}' \ne \boldsymbol{\alpha}$ satisfying Eq.* (2), $||\boldsymbol{\alpha}||_1 < ||\boldsymbol{\alpha}'||_1$.
(E) *The conditions in Proposition* 2 *are met by $\boldsymbol{\alpha}$. Furthermore, there does not exist distinct* (j_1, k_1, j_2, k_2) *such that $\alpha_{j_1 k_1} \ne 0$ and $\alpha_{j_2 k_2} \ne 0$ simultaneously.*
(F) *There exists j such that $|\beta_j| = \sum_{i \ne j} |\beta_i|$. Correspondingly, $\alpha_{ij} = \beta_i$ for all $i \ne j$ and $\alpha_{ik} = 0$ for all $i \ne j, k \ne j$.*

The following proposition establishes the relations between the numbers of nonzero elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different settings.

Proposition 4 *Assume that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfy Eq.* (2). *Let* $A = ||\boldsymbol{\alpha}||_0 = \sum_{1 \le j < k \le p} I(|\alpha_{jk}| \ne 0)$ *and* $B = ||\boldsymbol{\beta}||_0 = \sum_{j=1}^{p} I(|\beta_j| \ne 0)$ *denote the numbers of nonzero elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. Then,*

(G) $B \le 2A$.
(H) *If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfy the conditions in Proposition* 2, *then $2\sqrt{A} \le B \le 2A$.*
(I) *If $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ satisfy the conditions in Proposition* 3 *and $\boldsymbol{\alpha} \ne \boldsymbol{0}$ and $\boldsymbol{\beta} \ne \boldsymbol{0}$, then $B = A + 1$.*

In view of condition (H) in Proposition 4, if those conditions of Proposition 2 are met with $B = 2$ or 3, then those of Proposition 3 must be satisfied.

Proposition 5 *For any $\hat{\boldsymbol{\beta}}$ satisfying the sum-to-zero constraint, the corresponding $\hat{\boldsymbol{\alpha}}$ produced by Algorithm 1 satisfies Proposition* 2.

The proofs of the propositions are supplied in Appendix.

Constrained penalized likelihood

Given model Eq. (3), a random sample of n observations $(\tilde{\boldsymbol{x}}_i, Y_i)_{i=1}^{n}$ is obtained, based on which the log-likelihood function $l(\boldsymbol{\beta})$ can be written as $l(\boldsymbol{\beta}) = \frac{-1}{2\sigma^2} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}^T \tilde{\boldsymbol{x}}_i\right)^2$. In a high-dimensional situation, model Eq. (3) is overparameterized when $p > n$, and hence that $l(\boldsymbol{\beta})$ has multiple maximizers. Towards this end, we introduce a constrained penalized likelihood

as a generalization of the Lasso regression using L_1-regularization.

$$\min - l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|, \quad \text{subject to} \sum_{j=1}^{p} \beta_j = 0. \quad (4)$$

Minimization of Eq. (4) in $\boldsymbol{\beta}$ yields its minimizer $\hat{\boldsymbol{\beta}}$. Since the term σ^2 can be absorbed into the regularization coefficient λ in the penalized likelihood, we set $\sigma = 1$ in the objective function for simplicity.

In contrast to the Lasso problem, Eq. (4) has one additional linear constraint. The coordinate descent algorithm has been shown to be very efficient for solving L_1-penalized problems [21] since the nondifferentiable L_1 penalty is separable with respect to β_j's. However, the sum-to-zero constraint destroys the separability so that the coordinate descent algorithm can no longer guarantee convergence. In [17], the authors proposed adding additional diagonal moves and random local search to the coordinate descent algorithm, which improves the chance for convergence.

To deal with this convex optimization subject to linear constraints, we develop an algorithm using the alternating direction method of multipliers (ADMM) [18, 19] to solve iteratively, see Appendix for details. In each iteration, we derive an analytic updating formula to expedite convergence of ADMM, and convergence is guaranteed by a result in Section 3.1 of [18]. We compare our ADMM algorithm with the algorithm proposed in [17] in the "Results" section.

Results

Comparison of ADMM and CD+LS algorithms

This section compares our ADMM algorithm with a coordinate descent with local search (CD+LS) algorithm of [17] for Eq. (4) with respect to computation efficiency through one simulated example. The CD+LS algorithm is implemented in R package zeroSum (version 1.0.4, https://github.com/rehbergT/zeroSum). In this example, we consider correlated predictors, that is, $\tilde{\boldsymbol{x}}_i$'s are drawn iid from $N(0, \boldsymbol{V})$ and are independent of ϵ_i's that are sampled from $N(0, 1)$, and $\boldsymbol{V}$ is a $p \times p$ matrix whose ijth element is $0.5^{|i-j|}$. Moreover, the true β_j's are drawn iid from $N(0, 1)$ and then centered to have a sum zero, and λ is fixed as 1. Then, their rates for successful convergence and running times are recorded for the ADMM and the CD+LS algorithms over 20 simulations. Particularly, a precision or tolerance error of 10^{-10} is used for both algorithms. Successful convergence is reported if a solution from an algorithm satisfies the sum-to-zero constraint with a tolerance error of 10^{-8}, and both the solution and its objective value are no further than 10^{-8} to the optimal solution and its corresponding objective value in terms of the L_2-distance. Here, the optimal solution is defined as the one, among

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 5 of 11

the two solutions from the two algorithms, having the minimal objective value and satisfying the sum-to-zero constraint.

Four different settings are compared, ranging from low- to high-dimensional situations. As showed in Table 1, the proposed ADMM algorithm outperforms the CD+LS algorithm with respect to both convergence guarantee and running time.

Comparison of the proposed method and the Lasso

This section examines effectiveness of the proposed method through simulated examples. Specifically, the proposed method is compared with the Lasso in terms of predictive accuracy and identification of the true model, where the Lasso is implemented in R package glmnet [21].

Simulated examples are generated with correlation structures as to be analyzed. These simulations are designed to examine various operating characteristics of the proposed method with respect to (p, n), noise level σ^2, and correlation structures among predictors in Eqs. (1) and (3). For tuning, λ is searched over 100 grid points that are uniformly spaced (in the log-scale) between 10^4 and 10^{-2}. An independent testing dataset with 1000 randomly generated data points are used to find the optimal λ which minimizes the mean squared error (MSE).

For performance metrics, an independent validation dataset with 1000 randomly generated data points are used to evaluate the performance of the fitted model in terms of mean squared error (MSE) and R^2. To assess robustness of the approaches under data normalization, we randomly add sample-wise shifts from $N(0, 1)$ to the validation dataset. Furthermore, we consider other two metrics for parameter estimation and the quality of identification of zero elements of true $\boldsymbol{\alpha}$ in Eq. (1) and $\boldsymbol{\beta}$ in Eq. (3). For parameter estimation, we use the relative error

(RE) for estimating the true regression coefficients $\boldsymbol{\gamma}^0 = \left(\gamma_1^0, \ldots, \gamma_p^0\right)^T$, defined as

$$\text{RE} = \frac{||\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0||_2}{||\boldsymbol{\gamma}^0||_2},$$

where $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)^T$ are estimated regression coefficients. This metric allows for accounting for different scales between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For accuracy of identification, we use the false identification rate (FR), defined as

$$\text{FR} = 1 - \frac{|\tilde{\Gamma}| \cap |\Gamma^0|}{|\Gamma^0|},$$

where $\Gamma^0 = \{j | \gamma_j^0 \neq 0\}$ and $\tilde{\Gamma} = \{j | \tilde{\gamma}_j' \neq 0\}$, with $\tilde{\gamma}'$ being a truncated version of $\tilde{\gamma}$ such that only the coefficients with the $|\Gamma^0|$ largest absolute values are retained, and all others are zeroed out. Our simulated example concerns correlation structures among predictors. In Eqs. (1) and (3), $\log \boldsymbol{x}_i$'s are iid from $N(0, \boldsymbol{V})$ and are independent of ϵ_i's that are iid from $N(0, \sigma^2)$, and $\boldsymbol{V}$ is a $p \times p$ matrix whose ijth element is $0.5^{|i-j|}$, $\boldsymbol{\alpha} = (\alpha_{12}, \alpha_{13}, \ldots, \alpha_{p-1p})^T$. Three settings for $\boldsymbol{\alpha}$ are considered:

1) $\alpha_{12} = 1$, $\alpha_{13} = 0.5$, $\alpha_{24} = 0.5$, and $\alpha_{jk} = 0$ otherwise, which does not satisfy the conditions defined in Proposition 2 with $\beta_1 = 1.5$, $\beta_2 = \beta_3 = \beta_4 = -0.5$, and $\beta_j = 0$ for $j \geq 5$.
2) $\alpha_{12} = 1$, $\alpha_{13} = 0.5$, $\alpha_{24} = -0.5$, and $\alpha_{jk} = 0$ otherwise, which satisfies the conditions defined in Proposition 2 but does not satisfy the conditions defined in Proposition 3 with $\beta_1 = 1.5$, $\beta_2 = -1.5$, $\beta_3 = -0.5$, $\beta_4 = 0.5$, and $\beta_j = 0$ for $j \geq 5$.
3) $\alpha_{12} = 1$, $\alpha_{13} = 0.5$, $\alpha_{14} = 0.5$, and $\alpha_{jk} = 0$ otherwise, which satisfies the conditions defined in Proposition 3 with $\beta_1 = 2$, $\beta_2 = -1$, $\beta_3 = \beta_4 = -0.5$, and $\beta_j = 0$ for $j \geq 5$.

The proposed method is compared with the Lasso in two models, corresponding to the gene-pair level design matrix $\boldsymbol{z}$ in Eq. (1) and $\tilde{\boldsymbol{x}}$ in Eq. (3) (without the sum-to-zero constraint), referred to as Lasso 1 and Lasso 2, respectively. These three methods are examined on simple and difficult situations correspondingly with $(p, n, \sigma) = (25, 50, 0.5)$, $(100, 25, 0.2)$. Then the values of MSE, R^2, RE, and *FR* are reported. As suggested in Table 2, the proposed method performs better or the same compared with Lasso 1 and Lasso 2 in terms of accuracy and robustness across all the six situations. The improved performance is attributed to a reduced number of candidate parameters in Eq. (1) than Eq. (3), as well as to the sum-to-zero constraint introduced in Eq. (3). Interestingly, the false identification rates of the proposed method are almost zero in three low-noise setting of $(p, n, \sigma) = (20, 50, 0.5)$ regardless if the conditions in Propositions 2 and 3 are met, and are small in the other

Table 1 Comparison of ADMM and CD+LS algorithms

		% Success Rate		Time (s)	
p	n	ADMM	CD+LS	ADMM	CD+LS
20	50	100.0	100.0	0.03	0.0
		(0.0)	(0.0)	(0.01)	(0.0)
100	100	100.0	100.0	0.06	1.3
		(0.0)	(0.0)	(0.00)	(0.1)
100	500	100.0	80.0	0.14	0.6
		(0.0)	(9.2)	(0.04)	(0.2)
200	1000	100.0	80.0	0.54	3.9
		(0.0)	(9.2)	(0.12)	(0.4)

Sample means (standard errors in parentheses) of rates for successful convergence (in percentages), and running times (in seconds), based on 20 simulation replications, for the proposed ADMM and CD+LS algorithms

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 6 of 11

Table 2 Comparison of the proposed method and the Lasso in simulations

Setting	(p, n, σ)	Method	MSE	R^2	RE	FR
1	(20, 50, 0.5)	Lasso 1	0.29	0.89	0.77	0.33
			(0.01)	(0.00)	(0.01)	(0.00)
		Lasso 2	0.31	0.88	0.14	0.00
			(0.01)	(0.00)	(0.01)	(0.00)
		Proposed	0.29	0.89	0.13	0.00
			(0.01)	(0.00)	(0.01)	(0.00)
1	(100, 25, 0.2)	Lasso 1	0.19	0.92	0.80	0.37
			(0.07)	(0.03)	(0.02)	(0.02)
		Lasso 2	0.24	0.89	0.22	0.06
			(0.08)	(0.03)	(0.03)	(0.03)
		Proposed	0.19	0.92	0.21	0.04
			(0.07)	(0.03)	(0.03)	(0.03)
2	(20, 50, 0.5)	Lasso 1	0.32	0.89	0.44	0.20
			(0.01)	(0.00)	(0.02)	(0.04)
		Lasso 2	0.36	0.88	0.14	0.00
			(0.02)	(0.01)	(0.01)	(0.00)
		Proposed	0.32	0.89	0.14	0.01
			(0.01)	(0.00)	(0.01)	(0.01)
2	(100, 25, 0.2)	Lasso 1	0.42	0.84	0.52	0.37
			(0.07)	(0.02)	(0.03)	(0.06)
		Lasso 2	0.63	0.76	0.32	0.22
			(0.09)	(0.03)	(0.03)	(0.05)
		Proposed	0.41	0.85	0.31	0.24
			(0.07)	(0.02)	(0.03)	(0.05)
3	(20, 50, 0.5)	Lasso 1	0.30	0.93	0.18	0.00
			(0.01)	(0.00)	(0.02)	(0.00)
		Lasso 2	0.33	0.92	0.11	0.00
			(0.01)	(0.00)	(0.01)	(0.00)
		Proposed	0.30	0.93	0.11	0.00
			(0.01)	(0.00)	(0.01)	(0.00)
3	(100, 25, 0.2)	Lasso 1	0.15	0.96	0.25	0.00
			(0.02)	(0.01)	(0.03)	(0.00)
		Lasso 2	0.26	0.93	0.16	0.01
			(0.08)	(0.02)	(0.02)	(0.01)
		Proposed	0.15	0.96	0.15	0.00
			(0.02)	(0.01)	(0.01)	(0.00)

Sample means (standard errors in parentheses) of mean squared error (MSE), R^2, relative error (RE) and false identification rate (FR), based on 20 simulation replications, for the proposed method and the Lasso

three settings. In contrast, Lasso 1 has a higher relative error and false identification rate. While Lasso 2 has similar relative error and false identification rate as the proposed method, it has higher MSE and lower R^2 in all settings, due to its non-robustness to sample-wise

scaling without the sum-to-zero constraint. Overall, all three methods perform better in the low-noise situation of $(p, n, \sigma) = (20, 50, 0.5)$ than the high-noise situation of $(p, n, \sigma) = (100, 25, 0.2)$. Across the three settings of $\boldsymbol{\alpha}$, the performance of the proposed method is rather stable. However, Lasso 1 performs much worse for settings in which $\boldsymbol{\alpha}$ fails to satisfy the conditions in Proposition 3, corresponding to non-uniqueness of the representation of $\boldsymbol{\alpha}$. Most critically, when $\boldsymbol{\alpha}$ satisfies the conditions in Proposition 3, the proposed method continues to outperform its counterpart in terms of the performance metrics. Overall, the proposed method achieves our objective.

An application to a real RNA-Seq dataset
This section applies the proposed method to a prostate adenocarcinoma (PRAD) RNA-Seq dataset published as part of The Cancer Genome Atlas (TCGA) project [22]. Particularly, we identify gene pairs that are associated with pre-operative prostate-specific antigen (PSA), an important risk factor for prostate cancer. Towards this end, we download normalized gene expression data from the TCGA data portal (https://tcga-data.nci.nih.gov/docs/publications/tcga/). As described by TCGA, tissue samples from 333 PRAD patients were sequenced using the Illumina sequencing instruments. While raw sequencing reads were processed and analyzed using the SeqWare Pipeline 0.7.0 and MapspliceRSEM workflow 0.7 developed by the University of North Carolina, read alignment was performed using MapSplice [23] to the human reference genome, and gene expression levels were estimated using RSEM [24] with gene annotation GAF 2.1, and further normalized so that the upper quartile count is 1,000 in each sample. All these steps were performed by the TCGA consortium.

In our experiment, the normalized RSEM gene expression estimates are used, excluding samples with missing pre-operative PSA values and genes for which the average normalized expression level is lower than 10. This prepossessing step yields $p = 15,382$ genes and $n = 187$ samples. Furthermore, we run Pearson correlation tests for each gene between log-transformed expression levels and log-transformed pre-operative PSA levels, and exclude genes with false discovery rate (FDR) values (calculated using the Benjamini-Hochberg [25] method based on the p-values from the Pearson correlation tests) larger than 0.01. Consequently, only 520 genes are retained in the analysis, on which we fit model Eq. (3) using the proposed ADMM algorithm.

To visualize the selection result, we display the solution paths of the model fitting. As shown in Fig. 1, the first pair of genes entering the model are PTPRR and KRT15. While PTPRR is a member of the protein tyrosine phosphatase (PTP) family, which is known to be related with prostate cancer [26, 27], KRT15 is a member of the keratin gene

Shen *et al. BMC Bioinformatics*   (2017) 18:466

Page 7 of 11

Fig. 1 Solution paths of the model fitting with $p = 520$ genes
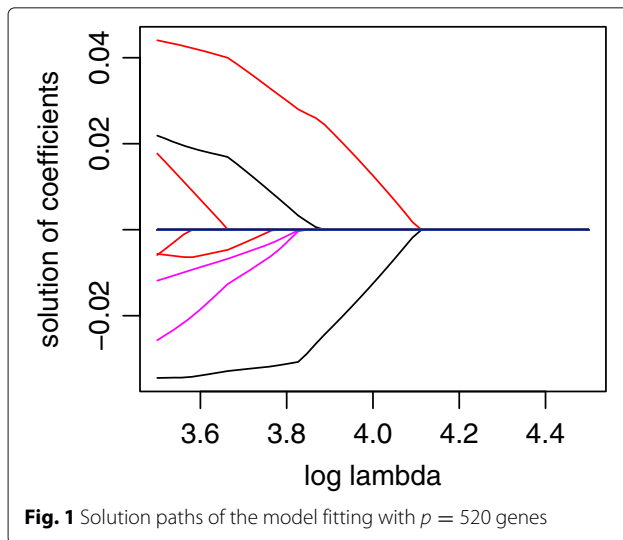
family, which is known to be associated with breast cancer [28] and lung cancer [29]. Interestingly, we find no publication record on PubMed with keywords such as "KRT15 AND PSA" or "KRT15 AND prostate". By correlating log expression levels and log PSA levels in the 187 patients, we find that both PTPRR and KRT15 are significantly correlated with PSA levels ($r = 0.28$ and $p < 10^{-4}$ for PTPRR, $r = -0.33$ and $p < 10^{-5}$ for KRT15). Not surprisingly, their log-ratio is even more strongly correlated with log PSA levels ($r = 0.41$ and $p < 10^{-8}$), demonstrating the potential of using gene pairs as biomarkers.

The other selected genes are HIST1H1E, LRAT, LCN2, KCNN4, RHOU, and EPHA5 in the order of them entering the model, among which LRAT [30], LCN2 [31], RHOU [32], and EPHA5 [33] are known to link to prostate cancer, and HIST1H1E and KCNN4 are connected to myeloma [34] and pancreatic cancer [35], respectively.

To demonstrate the scalability of our proposed method, we employ the proposed ADMM algorithm to fit Eq. (3) with all the $p = 15,382$ genes without pre-screening. In this situation, the first pair of genes entering the model are BCL8 and KRT15, where BCL8 is known to be associated with lymphoma [36]. The other selected genes are PTPRR, LRAT and LCN2, which are very similar to the foregoing results based on pre-screening. By comparison, fitting the corresponding model Eq. (1) using a standard Lasso algorithm, such as glmnet [21], would be practically prohibitive, which requires storing a design matrix of $\frac{p(p-1)}{2} \times n \approx 22 \times 10^9$ elements. To further demonstrate robustness of the proposed method with respect to data normalization, we randomly scale the gene expression levels, both along the gene dimension and the sample dimension, mimicking the gene length normalization and the library size or sequencing depth normalization, respectively. Numerically, we find that the solution path

fitted using the randomly scaled data is always identical to that fitted using the original data.

Discussion

Model Eq. (3) has been proposed for compositional data [16] and recently for reference point insensitive data [17]. In this article, we explore Eq. (3) for the identification of gene pairs as biomarkers, enjoying robustness to sample-wise scaling normalization (which is a common practice for RNA-Seq data) and simplicity of validation and measurement by qPCR techniques. Through Propositions 1–4, we establish the relationship between models Eq. (1) and Eq. (3). Additionally, we develop an efficient ADMM algorithm for solving model Eq. (3), which is guaranteed to converge and is shown to be highly competitive in terms of computational efficiency.

One interesting yet important issue of model Eq. (3) is determination of the value of $\boldsymbol{\alpha}$. One proposal is choosing the $\boldsymbol{\alpha}$ to minimize the L_0-norm instead of L_1-norm. However, in such a situation, it remains unclear what kind of conditions as those in Propositions 2, 3 and 4 may be. Furthermore, minimization of the L_0-norm in $\boldsymbol{\alpha}$ continues to be challenging by itself due to non-convexity and discontinuity of the L_0-function. Therefore, our approach based on the L_1-norm gives rise to convex minimization, which is easier to manage.

One important aspect of model Eq. (3) is that it enables to identify gene pairs in an unbiased manner without any prior knowledge of the known biology of the disease. However, in some situations, information regarding the disease of interest is available from previous studies. Then the prior knowledge needs to be integrated for gene pair identification so that more probable subset of genes or pathways should have a higher chance of being selected. This can be accomplished through weighted regularization with weights $\{\lambda_k\}_{k=1}^p$, with a large weight corresponding to a small chance of being selected. Moreover, in some other situations, gene pairs are constrained in that gene pairs are formed only between relevant genes from the same pathway. This can be achieved by replacing the Lasso penalty by either a (sparse) group Lasso penalty [37] and/or the simple equality constraint by a set of constraints, each corresponding to a given pathway of interest. Finally, some non-convex penalties such as the SCAD penalty [14] and the TLP penalty [15] can be used as opposed to the Lasso penalty to achieve a higher accuracy of selection at an expense of computation.

For a large-scale problem, an ADMM algorithm may have a linear convergence rate. To expedite convergence, an inexact ADMM algorithm may be useful in our setting, which has been shown recently to lead to a substantial improvement over the standard ADMM algorithm [19]. Furthermore, parallelization of our ADMM algorithm may achieve further scalability, which is one advantage of

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 8 of 11

ADMM algorithms over many other optimization techniques [18].

One extension of Eq. (1) is generalized linear models such as logistic regression or other predictive models like support vector machine. In such a situation, the proposed method for Eq. (1) is directly applicable to gene pair identification with some modifications. Further investigation is necessary.

Conclusion

In conclusion, the experimental results demonstrate that gene pairs can be used as robust biomarkers which can tolerate sample-wise scaling normalization. Furthermore, using L_1 penalized regression with equality constraints, the model fitting can be formulated as a convex optimization problem which can be solved efficiently using the proposed ADMM algorithm. This approach has the potential to discover novel and reliable biomarkers for biological or clinical studies.

Appendix
Proofs of propositions
Proof of Proposition 1: From Eq. (1), we have

$$f(\boldsymbol{z}) = \sum_{j=1}^{p} \sum_{k=j+1}^{p} \alpha_{jk} (\log x_j - \log x_k)$$

$$= \sum_{j=1}^{p} \log x_j \left(\sum_{k=j+1}^{p} \alpha_{jk} - \sum_{k=1}^{j-1} \alpha_{kj} \right)$$

$$= \sum_{j=1}^{p} \log x_j \sum_{k \neq j} \alpha_{jk}$$

$$= \sum_{j=1}^{p} \beta_j \log x_j.$$

Furthermore

$$\sum_{j=1}^{p} \beta_j = \sum_{j=1}^{p} \sum_{k \neq j} \alpha_{jk}$$

$$= \sum_{j=1}^{p} \sum_{k=j+1}^{p} \alpha_{jk} - \sum_{j=1}^{p} \sum_{k=1}^{j} \alpha_{kj}$$

$$= \sum_{j=1}^{p} \sum_{k=j+1}^{p} \alpha_{jk} - \sum_{k=1}^{p} \sum_{j=k+1}^{p} \alpha_{kj}$$

$$= 0.$$

This completes the proof.

Proof of Proposition 2: We show (A)$\Rightarrow$(B), (B)$\Rightarrow$(C) and (C)$\Rightarrow$(A), respectively.

(A)$\Rightarrow$(B): We prove by contradiction. Without loss of generality, assume that $\alpha_{12} \geq \alpha_{23} > 0$. Then, we construct $\boldsymbol{\alpha}'$ such that $\alpha'_{12} = \alpha_{12} - \alpha_{23}, \alpha'_{23} = 0, \alpha'_{13} = -\alpha_{13} + \alpha_{23} >$

0 and $\alpha'_{ij} = \alpha_{ij}$ otherwise. Easily, $\boldsymbol{\alpha}'$ satisfies Eq. (2) and $||\boldsymbol{\alpha}'||_1 - ||\boldsymbol{\alpha}||_1 = -2\alpha_{13} < 0$, which contradicts (A).

(B)$\Rightarrow$(C): By (B), α_{ij} and α_{jk} always have opposite signs. This, together with the definition that $\alpha_{ji} = -\alpha_{ij}$, α_{ji} and α_{jk} always have the same sign, where 0 can be regarded as an arbitrary sign. Therefore, from Eq. (2), we have

$$\frac{1}{2} ||\boldsymbol{\beta}||_1 = \frac{1}{2} \sum_{j=1}^{p} |\beta_j| = \frac{1}{2} \sum_{j=1}^{p} \left| \sum_{k \neq j} \alpha_{jk} \right|$$

$$= \frac{1}{2} \sum_{j=1}^{p} \sum_{k \neq j} |\alpha_{jk}| = ||\boldsymbol{\alpha}||_1.$$

(C)$\Rightarrow$(A): For any $\boldsymbol{\alpha}'$ satisfying Eq. (2), we have

$$||\boldsymbol{\alpha}||_1 = \frac{1}{2} ||\boldsymbol{\beta}||_1 = \frac{1}{2} \sum_{j=1}^{p} |\beta_j| = \frac{1}{2} \sum_{j=1}^{p} \left| \sum_{k \neq j} \alpha'_{jk} \right|$$

$$\leq \frac{1}{2} \sum_{j=1}^{p} \sum_{k \neq j} |\alpha'_{jk}| = ||\boldsymbol{\alpha}'||_1.$$

This completes the proof.

Proof of Proposition 3: We show (D)$\Rightarrow$(E), (E)$\Rightarrow$(F) and (F)$\Rightarrow$(D), respectively.

(D)$\Rightarrow$(E): As in the proof of Proposition 2, we assume, without loss of generality, that $\alpha_{24} \geq \alpha_{13} > 0$. Then, we construct $\boldsymbol{\alpha}'$ such that $\alpha'_{13} = 0$, $\alpha'_{14} = \alpha_{14} + \alpha_{13}$, $\alpha'_{23} = \alpha_{23} + \alpha_{13}, \alpha'_{24} = \alpha_{24} - \alpha_{13}$ and $\alpha'_{ij} = \alpha_{ij}$ otherwise. Easily, $\boldsymbol{\alpha}'$ also satisfies Eq. (2), and $||\boldsymbol{\alpha}'||_1 \leq ||\boldsymbol{\alpha}||_1$, which contradicts (D).

(E)$\Rightarrow$(F): From (E), there exists a j such that $\alpha_{ik} = 0$ for all $i \neq j$ and $k \neq j$. From Eq. (2) and (B) in Proposition 2, $\beta_i = \sum_{k \neq i} \alpha_{ik} = \alpha_{ij}$, for all $i \neq j$, and $|\beta_j| = \left| \sum_{k \neq j} \alpha_{jk} \right| = \sum_{k \neq j} |\alpha_{jk}| = \sum_{k \neq j} |\beta_k|$.

(F)$\Rightarrow$(D): Suppose that $\boldsymbol{\alpha}'$ satisfies Eq. (2) and the conditions in Proposition 2. From (B), we know for all $k \neq j$, α'_{jk} have the same sign. From Eq. (2), (C) and (F), we have

$$\frac{1}{2} ||\boldsymbol{\beta}||_1 = |\beta_j| = \left| \sum_{k \neq j} \alpha'_{jk} \right| = \sum_{k \neq j} |\alpha'_{jk}| \leq ||\boldsymbol{\alpha}'||_1 = \frac{1}{2} ||\boldsymbol{\beta}||_1$$

Therefore, we must have $\sum_{k \neq j} |\alpha'_{jk}| = ||\boldsymbol{\alpha}'||_1$. That is, $\alpha'_{ik} = 0$ for all $i \neq j, k \neq j$. Furthermore, for any $i \neq j$, we have $\beta_i = \sum_{k \neq i} \alpha'_{ik} = \alpha'_{ij}$. Therefore, $\boldsymbol{\alpha}' = \boldsymbol{\alpha}$, implying the uniqueness of $\boldsymbol{\alpha}$. This completes the proof.

Proof of Proposition 4:

(G): By Eq. (2), $\beta_j \neq 0$ only if at least for some $k \neq j$, $\alpha_{jk} \neq 0$. Based on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we can construct an undirected graph $G = (V, E)$ with p vertices such that there is an edge between vertices i and j if and only if $\alpha_{ij} \neq 0$. We know that β_j can not be nonzero unless vertex V_j has a degree of at least 1. Since the total number of edges is

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 9 of 11

A, we know that $B \leq \sum_{j=1}^{p} I(degree(V_j) > 0) \leq \sum_{j=1}^{p} degree(V_j) = 2A$.

(H): Suppose α satisfies the conditions defined in Proposition 2. If $\alpha_{ij} \neq 0$, let α_{ij} be the weight associated with edge connecting V_i and V_j. By condition (B), for any cycle in the graph formed by sequence $i_1, i_2, \ldots, i_k, i_{k+1} = i_1$, we know that weights associated with adjacent edges (i.e., $\alpha_{i_{j-1}i_j}$ and $\alpha_{i_ji_{j+1}}$) always have opposite signs. Therefore, the number of edges in the cycle has to be an even number, which means that the graph has to be a bipartite graph. It is then easy to see that $A \leq \left(\frac{B}{2}\right)^2$ for such graphs. That is, $2\sqrt{A} \leq B$.

(I): Suppose $\alpha \neq 0$ satisfies the conditions defined in Proposition 3. By condition (F), $B = A + 1$. This completes the proof.

Proof of Proposition 5: It suffices to show that $\hat{\alpha}$ satisfies (C) of Proposition 2. Note that $\tilde{\beta}$ in Algorithm 1 satisfies the sum-to-zero constraint at each step of iteration before termination. In the beginning, $\|\hat{\alpha}\|_1 = 0$ and $\|\tilde{\beta}\|_1 = \|\hat{\beta}\|_1$. After each iteration, $\|\hat{\alpha}\|_1$ is increased by $\hat{\alpha}_{k_1 k_2} = \min(|\tilde{\beta}_{k_1}|, |\tilde{\beta}_{k_2}|)$, and $\|\tilde{\beta}\|_1$ is decreased by $2\hat{\alpha}_{k_1 k_2}$. In the end, $\|\tilde{\beta}\|_1 = 0$, and therefore $\|\hat{\alpha}\|_1 = \frac{1}{2}\|\hat{\beta}\|_1$. This completes the proof.

ADMM algorithm for solving Eq. (4)

We adopt the notation in [18] and reformulate Eq. (4) as

$$
\begin{aligned}
&\min (1/2)\|Ax - b\|_2^2 + \lambda \|z\|_1 \\
&\text{subject to } Cx = d, x - z = 0.
\end{aligned}
\tag{5}
$$

where x are the parameters of interest. If $C = 1^T$ and $d = 0$, we will have all coefficients sum up to 0. When there is an intercept in the model, we can add a scalar 0 as the first element in C, meaning that we do not have constraint on the intercept. Similarly, as a convention we also do not penalize the intercept. Denoting $B = \begin{bmatrix} C \\ I \end{bmatrix}$, $D = \begin{bmatrix} 0 \\ -I \end{bmatrix}$, and $d = \begin{bmatrix} d \\ 0 \end{bmatrix}$, the two equality constraints can be simplified as $Bx + Dz = d$. To use the ADMM algorithm [18], we form the augmented Lagrangian

$$
\begin{aligned}
L_\rho(x, z, y) = {} & (1/2)\|Ax - b\|_2^2 + \lambda \|z\|_1 \\
& + (\rho/2)\|Bx + Dz - d + u\|_2^2 - (\rho/2)\|u\|_2^2,
\end{aligned}
$$

with $u^k = (1/\rho)y^k = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}^p$ where u_1 is a scalar and $u_2 \in \mathbb{R}^{p \times 1}$.

Let $E = \begin{bmatrix} A \\ \sqrt{\rho} C \end{bmatrix}$, the ADMM algorithm consists of the following iterations

$$
x^{k+1} := (E^T E + \rho I)^{-1} \left(A^T b + \rho \left(z^k - C^T u_1^k + C^T d - u_2^k \right) \right)
$$

$$
z^{k+1} := S_{\lambda/\rho} \left(x^{k+1} + u_2^k \right)
$$

$$
u^{k+1} := u^k + \left(Bx^{k+1} + Dz^{k+1} - d \right).
$$

The x update can be accelerated by caching an initial factorization. Suppose the dimension of E is $m \times p$. If $m < p$, we cache the factorization of $I + \rho E E^T$ (with dimension $m \times m$) and use the matrix inversion lemma

$$
\left(\rho I + E^T E \right)^{-1} = I/\rho - E^T \left(I + 1/\rho E E^T \right)^{-1} E/\rho^2
$$

to update x. Otherwise, we cache the factorization of $\rho I + E^T E$ (with dimension $p \times p$) and use back- and forward- solve to update x directly. The iteration stops when the primal and dual residuals are smaller than their corresponding tolerances,

$$
\|r^{k+1}\|_2 \leq \epsilon^{pri} \quad \text{and} \quad \|s^{k+1}\|_2 \leq \epsilon^{dual},
$$

where

$$
r^{k+1} = Bx^{k+1} + Dz^{k+1} - d,
$$

$$
s^{k+1} = \rho B^T D \left(z^{k+1} - z^k \right),
$$

$$
\epsilon^{pri} = \sqrt{p+1} \epsilon^{abs} + \epsilon^{rel} \max \left\{ \|Bx^{k+1}\|_2, \| - z^{k+1}\|_2, \|d\|_2 \right\},
$$

$$
\epsilon^{dual} = \sqrt{p} \epsilon^{abs} + \epsilon^{rel} \|\rho B^T u^{k+1}\|_2.
$$

Usually, the relative stopping criteria is chosen to be $\epsilon^{rel} = 10^{-4}$, and the choice of absolute stopping criteria ϵ^{abs} depends on the scale of the variable values. See Boyd et al. [18] for details. To compute a solution path for a decreasing sequence of λ values, we adopt the approach in Friedman et al. [21] and use warm starts for each λ value. The sequence of λ values are either provided by the user, or we begin with $\lambda_{max} = \|A^T b\|_\infty$ for which all the coefficients are equal to 0. We set $\lambda_{min} = \epsilon_\lambda \lambda_{max}$, where ϵ_λ is a small value, such as 0.01, and generate a decreasing sequence of 100 λ values from λ_{max} to λ_{min} on the log-scale.

Shen *et al. BMC Bioinformatics* (2017) 18:466

Page 10 of 11

Availability of data and materials
The datasets and R programs for reproducing the results in this paper are available at http://www-personal.umich.edu/~jianghui/gene-pair/.

Authors' contributions
RS conducted the experiments and involved in the methodology development. LL developed and implemented the algorithm. HJ conceived the study, developed the algorithm and conducted the experiments. All authors drafted, read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]The Blake School, MN 55403 Minneapolis, USA. [2]Department of Biostatistics, University of Michigan, MI 48109 Ann Arbor, USA.

References
1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science. 1995;270(5235):467.
2. Quackenbush J. Microarray analysis and tumor classification. N Engl J Med. 2006;354(23):2463–72.
3. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. Brief Bioinform. 2013;14(6):671–83.
4. Patil P, Bachant-Winner PO, Haibe-Kains B, Leek JT. Test set bias affects reproducibility of gene signatures. Bioinformatics. 2015;31(14):2318–23.
5. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mrna comparisons. Stat Appl Genet Mol Biol. 2004;3(1):1–19.
6. Leek JT. The tspair package for finding top scoring pair classifiers in r. Bioinformatics. 2009;25(9):1203–04.
7. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative pcr and the 2-$\delta\delta$ct method. methods. 2001;25(4):402–8.
8. Arukwe A. Toxicological housekeeping genes: do they really keep the house?. Environ Sci Technol. 2006;40(24):7944–9.
9. Ma XJ, Wang Z, Ryan PD, Isakoff SJ, Barmettler A, Fuller A, Muir B, Mohapatra G, Salunga R, Tuggle JT, et al. A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen. Cancer cell. 2004;5(6):607–16.
10. Raponi M, Lancet JE, Fan H, Dossey L, Lee G, Gojo I, Feldman EJ, Gotlib J, Morris LE, Greenberg PL, et al. A 2-gene classifier for predicting response to the farnesyltransferase inhibitor tipifarnib in acute myeloid leukemia. Blood. 2008;111(5):2589–96.
11. Price ND, Trent J, El-Naggar AK, Cogdell D, Taylor E, Hunt KK, Pollock RE, Hood L, Shmulevich I, Zhang W. Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas. Proc Natl Acad Sci. 2007;104(9):3414–19.
12. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005;21(20):3896–904.
13. Tibshirani R. Regression shrinkage and selection via the lasso. R Stat Soc Ser B Methodol J. 1996;58(1):267–88.
14. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–60.
15. Shen X, Pan W, Zhu Y. Likelihood-based selection and sharp parameter estimation. J Am Stat Assoc. 2012;107(497):223–32.
16. Lin W, Shi P, Feng R, Li H, et al. Variable selection in regression with compositional covariates. Biometrika. 2014;101(4):785–97.
17. Altenbuchinger M, Rehberg T, Zacharias H, Stämmler F, Dettmer K, Weber D, Hiergeist A, Gessner A, Holler E, Oefner PJ, et al. Reference point insensitive molecular data analysis. Bioinformatics. 2016;598:1–122.
18. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends® Mach Learn. 2011;3(1):1–122.
19. Wang H, Banerjee A. Bregman alternating direction method of multipliers. In: Advances in Neural Information Processing Systems. Curan Associates, Inc.; 2014. p. 2816–24.
20. Smyth G. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3(1):1–25.
21. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1.
22. The Cancer Genome Atlas Research Network: The molecular taxonomy of primary prostate cancer. Cell. 2015;163(4):1011–25.
23. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38(18):178–8.
24. Li B, Dewey CN. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. BMC bioinformatics. 2011;12(1):1.
25. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
26. Rand KA, Rohland N, Tandon A, Stram A, Sheng X, Do R, Pasaniuc B, Allen A, Quinque D, Mallick S, et al. Whole-exome sequencing of over 4100 men of african ancestry and prostate cancer risk. Hum Mol Genet. 2016;25(2):371–81.
27. Munkley J, Lafferty NP, Kalna G, Robson CN, Leung HY, Rajan P, Elliott DJ. Androgen-regulation of the protein tyrosine phosphatase ptprr activates erk1/2 signalling in prostate cancer cells. BMC cancer. 2015;15(1):9.
28. Gelfand R, Vernet D, Bruhn KW, Sarkissyan S, Heber D, Vadgama JV, Gonzalez-Cadavid NF. Long-term exposure of mcf-7 breast cancer cells to ethanol stimulates oncogenic features. Int J Oncol. 2017;50(1):49–65.
29. Boyero L, Sánchez-Palencia A, Miranda-León MT, Hernández-Escobar F, Gómez-Capilla JA, Fárez-Vidal ME. Survival, classifications, and desmosomal plaque genes in non-small cell lung cancer. Int J Med Sci. 2013;10(9):1166.
30. Guo X, Knudsen BS, Peehl DM, Ruiz A, Bok D, Rando RR, Rhim JS, Nanus DM, Gudas LJ. Retinol metabolism and lecithin: retinol acyltransferase levels are reduced in cultured human prostate cancer cells and tissue specimens. Cancer Res. 2002;62(6):1654–61.
31. Sunil VR, Patel KJ, Nilsen-Hamilton M, Heck DE, Laskin JD, Laskin DL. Acute endotoxemia is associated with upregulation of lipocalin 24p3/lcn2 in lung and liver. Exp Mol Pathol. 2007;83(2):177–87.
32. Alinezhad S, Väänänen RM, Tallgrén T, Perez IM, Jambor I, Aronen H, Kähkönen E, Ettala O, Syvänen K, Nees M, et al. Stratification of aggressive prostate cancer from indolent disease-prospective controlled trial utilizing expression of 11 genes in apparently benign tissue. Urologic Oncol. 2016;34(6):255–15. Elsevier.
33. Li S, Zhu Y, Ma C, Qiu Z, Zhang X, Kang Z, Wu Z, Wang H, Xu X, Zhang H, et al. Downregulation of epha5 by promoter methylation in human prostate cancer. BMC cancer. 15(1):18.
34. Walker BA, Boyle EM, Wardell CP, Murison A, Begum DB, Dahir NM, Proszek PZ, Johnson DC, Kaiser MF, Melchor L, et al. Mutational spectrum, copy number changes, and outcome: results of a sequencing study of patients with newly diagnosed myeloma. J Clin Oncol. 2015;33(33):3911–20.

Shen *et al. BMC Bioinformatics*   (2017) 18:466

Page 11 of 11

35. Bonito B, Sauter DP, Schwab A, Djamgoz MA, Novak I. Kca3. 1 (ik)
     modulates pancreatic cancer cell migration, invasion and proliferation:
     anomalous effects on tram-34. Pflügers Archiv-European J Physiol.
     2016;468(11-12):1865–75.
36. Dyomin VG, Rao PH, Dalla-Favera R, Chaganti R. Bcl8, a novel gene
     involved in translocations affecting band 15q11–13 in diffuse large-cell
     lymphoma. Proc. Natl Acad Sci. 1997;94(11):5728–32.
37. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J
     Comput Graphical Stat. 2013;22(2):231–45.