frontiers in
CELL AND DEVELOPMENTAL BIOLOGY

ORIGINAL RESEARCH ARTICLE
published: 27 October 2014
doi: 10.3389/fcell.2014.00062

Integrative analysis of multiple diverse omics datasets by sparse group multitask regression

Dongdong Lin[1,2], Jigang Zhang[2,3], Jingyao Li[1,2], Hao He[2,3], Hong-Wen Deng[2,3] and Yu-Ping Wang[1,2,3]*

[1] Biomedical Engineering Department, Tulane University, New Orleans, LA, USA
[2] Center for Bioinformatics and Genomics, Tulane University, New Orleans, LA, USA
[3] Department of Biostatistics and Bioinformatics, Tulane University, New Orleans, LA, USA

Edited by:
Jennifer Elizabeth Dent, Quintiles, UK

Reviewed by:
Dougu Nam, Ulsan National Institute of Science and Technology, South Korea
Ayse Meric Ovacik, Merck, USA

***Correspondence:**
Yu-Ping Wang, Department of Biomedical Engineering, Biostatistics and Bioinformatics, Tulane University, New Orleans, LA 70118, USA
e-mail: wyp@tulane.edu

A variety of high throughput genome-wide assays enable the exploration of genetic risk factors underlying complex traits. Although these studies have remarkable impact on identifying susceptible biomarkers, they suffer from issues such as limited sample size and low reproducibility. Combining individual studies of different genetic levels/platforms has the promise to improve the power and consistency of biomarker identification. In this paper, we propose a novel integrative method, namely sparse group multitask regression, for integrating diverse omics datasets, platforms, and populations to identify risk genes/factors of complex diseases. This method combines multitask learning with sparse group regularization, which will: (1) treat the biomarker identification in each single study as a task and then combine them by multitask learning; (2) group variables from all studies for identifying significant genes; (3) enforce sparse constraint on groups of variables to overcome the "small sample, but large variables" problem. We introduce two sparse group penalties: sparse group lasso and sparse group ridge in our multitask model, and provide an effective algorithm for each model. In addition, we propose a significance test for the identification of potential risk genes. Two simulation studies are performed to evaluate the performance of our integrative method by comparing it with conventional meta-analysis method. The results show that our sparse group multitask method outperforms meta-analysis method significantly. In an application to our osteoporosis studies, 7 genes are identified as significant genes by our method and are found to have significant effects in other three independent studies for validation. The most significant gene SOD2 has been identified in our previous osteoporosis study involving the same expression dataset. Several other genes such as TREML2, HTR1E, and GLO1 are shown to be novel susceptible genes for osteoporosis, as confirmed from other studies.

Keywords: sparse regression, multitask learning, group lasso, significant test, osteoporosis

INTRODUCTION

Increasing amounts of high-throughput biological data have been collected to investigate the genetic mechanism underlying complex traits at different levels, e.g., genomics, transcriptomics, proteomics, and metabolomics. However, there are usually two bottlenecks for these genetic studies. One is availability of limited sample size due to the experimental cost. Small sample size can lead to the loss of detection power and the reduction of confidence on identified biomarkers. To analyze data with small sample size but large variables is still a challenging statistical problem (Hamid et al., 2009). The other is that biomarkers identified from these different studies often suffer from poor reproducibility. This issue could be caused by many factors such as differences on profiling techniques, demographic, and ancestral information of subjects, sample sizes, and quality control in these datasets (Phan et al., 2012; Song et al., 2012). To increase the power and consistency of biomarker identification, integrating the information of diverse biological datasets from different levels and platforms shows great promise and is highly demanded.

Methods for integration of diverse biological datasets include conventional meta-analysis and a variety of integrative approaches recently developed (Huttenhower et al., 2006; Liu et al., 2013). Meta-analysis is a statistical method to summarize the *p*-values or statistics (e.g., z-score) from each individual dataset (Evangelou and Ioannidis, 2013). There are a dozen of approaches for combing multiple *p*-values or statistics such as Fisher method. Meta-analysis is usually used to find common features across multiple datasets with different sample sizes and platforms but under the same hypothesis (Rhodes and Chinnaiyan, 2005). Recently, a number of integrative approaches have been developed, which are based on machine learning and statistical methods (Zhang et al., 2010; Kirk et al., 2012; Xiong et al., 2012). They can analyze multiple datasets from: (1) different platforms and levels but for the same subjects; (2) same platforms but different levels and subjects; (3) different platforms but for the same levels and subjects. They have been successfully used for various applications such as a single or a set of biomarker identification (Chen et al., 2013), gene-gene interaction prediction

(Troyanskaya et al., 2003), and genetic network construction (Balbin et al., 2013). The results in these studies demonstrate the advantage of integrating multiple diverse datasets over analyzing them individually.

In this work, we propose a novel method for integrating multiple datasets from different platforms, levels, and samples to identify common biomarkers (e.g., genes). The method was based on multitask regression model enforced with sparse group regularization, which can overcome the "small sample size, but large number of variables" problem. Multitask learning method has been successfully applied to medical imaging data fusion, where multiple types of images (e.g., CT, MRI) were combined for identifying susceptible brain regions and improving disease classification (Zhang and Shen, 2012). Among various sparse regularization terms, the use of sparse group penalty has been shown to outperform other penalties such as lasso in our previous study of pair-wise genomic data integration (Lin et al., 2013). In this study, we enforce two sparse group penalties [i.e., sparse group lasso (Friedman et al., 2010) and sparse group ridge (Chen et al., 2010a)] into the multitask regression model for data integration. We assume a regression model for each dataset as a task, and then multiple regression models will be considered as multiple tasks. Variables from all datasets will be grouped by specific units (e.g., genes). A sparse group penalty is introduced with the aims to (1) reduce dimensionality, i.e., removing a number of irrelevant genes; (2) perform group-wise feature selection, i.e., removing SNPs or expression measurements from the same gene. An effective algorithm based on alternative direction method (ADM) is proposed to solve the model. Based on the estimation of the model, a statistical test is constructed for the identification of potentially causal genes. We perform two simulation studies with both fixed and dynamic genetic effects to evaluate our sparse regression methods, which shows that our sparse group multitask regression model can increase the power of detecting risk genes by integrating multiple diverse datasets effectively. Real data analysis on four osteoporosis studies identifies some significant genes with highly susceptible to bone mineral density and osteoporosis.

MATERIALS AND METHODS

In this section, we will first introduce the sparse group multitask regression model and then propose an effective algorithm based on ADM to solve the model. Finally, a gene based statistical test is constructed to give the level of significance for each selected gene.

SPARSE GROUP MULTITASK REGRESSION MODEL

We assume T independent datasets collected from K levels of genomic data (e.g., SNP, mRNA) with $P_k (k = 1, \ldots, K)$ platforms (e.g., Affymetrix, Illumina) for each level, and thus $T = \sum_{k=1}^{K} P_k$. The number of observations in each dataset is denoted by n_i, $i = 1, \ldots T$. Sample sizes could also be different due to the diversity of protocols in each experiment. The measurement matrix of each experiment is denoted by $X^{(i)} \in R^{n_i \times d_i}$, $i = 1, \ldots, T$, where $\mathrm{d_i}$ is the dimension of features in the i-th dataset, and usually $d_i >> n_i$. These features (e.g., SNPs and mRNA expression probes) are annotated to the genes and we assume that the genes in different datasets are the same, denoted by $G = \{G_i | i = 1, \ldots Q\}$. For example, all SNPs and mRNA expressions are tested for the same set of genes G. To reduce scale differences among different levels and platforms, the features in $X^{(i)}$s will be normalized to have zero mean and unit standard deviation. The phenotypic response in each dataset is $Y^{(i)} \in R^{n_i}$, $i = 1, \ldots\ T$, which can be binary or quantitative trait. The study is to identify biomarkers shared by different experiments for the same phenotype. The coefficient matrix for the regression model is denoted by $\boldsymbol{C} = \left[C^{(1)'}, C^{(2)'}, \ldots, C^{(T)'}\right]'$, where $C^{(i)} \in R^{d_i}$ is the coefficient vector of the i-th model $Link\left(Y^{(i)}\right) = X^{(i)} C^{(i)}$, and $Link(.)$ is the known link function.

Multitask learning is adopted in this study for identifying the shared biomarkers across a set of distinct but correlated tasks for better accuracy. In this context, each regression model for an experiment under different level and/or platform is considered as a task. For the sake of simplicity, we assume a linear regression model for each experiment with quantitative trait (i.e., link function will be the identity matrix). The loss function for each model $L^{(i)}\left(X^{(i)}, C^{(i)}\right)$ can be derived from the negative log likelihood function and thus the total loss function for the multitask regression model is $L(X, C) = \sum_{i=1}^{T} L^{(i)}\left(X^{(i)},\ C^{(i)}\right)$.

Many conventional regression methods become ineffective for processing the large scale biological data, which usually have small sample sizes and large number of features. This issue can be addressed by introducing sparse penalty in the model. We propose a sparse multitask regression model as follows:

$$min_{\boldsymbol{C}} L(\boldsymbol{X}, \boldsymbol{C}) + \Phi(\boldsymbol{C}) \qquad (1)$$

where $\Phi(\boldsymbol{C})$ is the sparse penalty function. Two popular penalties are used: sparse group lasso and sparse group ridge, and the corresponding models are denoted by multitask-sglasso and multitask-sgridge, respectively. For multitask-sglasso, $\Phi(\boldsymbol{C}) = \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_2 + \lambda_2 \|\boldsymbol{C}\|_{1,\ 1}$, where $\boldsymbol{C}_{\{k \in G_q\}}$ indicates a subset of vector $\boldsymbol{C}$ corresponding to the set of features annotated to gene G_q from Ttypes of datasets and $\|C\|_{1,\ 1} = \sum_{i=1}^{T} \sum_{k=1}^{d_i} \left|C^{(i,k)}\right|$ is the l-1 norm on $\boldsymbol{C}$. This sparse group lasso penalty groups features from all datasets based on genes to perform gene level selection. The l-1 norm penalty on $\boldsymbol{C}$ can further remove those irrelevant features from each gene. This bi-level feature selection penalty has been proven to outperform several other single level sparse penalties such as lasso, group lasso, and elastic net for feature identification. For multitask-sgridge, a composite sparse penalty, i.e., group ridge penalty $\Phi(\boldsymbol{C}) = \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_1^2$, is imposed on $\boldsymbol{C}$ to perform bi-level feature selection, where the features are also grouped by genes. The penalty uses the inner l-1 norm penalty on $\boldsymbol{C}_{\{k \in G_q\}}$ to achieve the sparsity within each gene while the outer ridge penalty to perform ridge regression at the gene level. This group ridge penalty has also been found to give higher power in identifying causal genes in high dimensional genomic dataset than other single level sparse penalties (Chen et al., 2010a).

In this study, we adopt these two bi-level penalties in our multitask regression models to integrate multiple diverse genomic datasets for gene-based test. Specifically, these two sparse group multitask regression models are formulated as follows:

$$\text{Multitask-sglasso: } min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_2 + \lambda_2 \|\boldsymbol{C}\|_{1,\,1} \tag{2}$$

$$\text{Multitask-sgridge: } min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_1^2 \tag{3}$$

where ω_is are the weights for the loss function of different levels of datasets, and δ_js are the weights accounting for the sample size differences among the experiments of the same type of datasets. To be more specific, ω_is reflect the prior knowledge on the importance of different levels of measurements, e.g., SNP, gene expression, and proteomics. We choose $\omega_i = 1$, $i = 1, 2, l \ldots K$ in this work, assuming that all levels of measurements contain the same important genetic information. Larger sample size is expected to provide more reliable significance test on biomarkers; therefore, the weight for the experiment under the j-th platform to measure the i-th level of genomic data is given by $\delta_j = \frac{n_j}{\sum_{j=1}^{P_i} n_j}$, $j \in P_i$, where λ_1, λ_2, and λ are the tuning parameters to control the sparsity of genes and the number of features in the models.

It could be noted that our sparse multitask regression model can be taken as the generalization of those existing sparse regression models to the representation of multiple datasets from different levels and/or platforms. For example, when $K = 1$, $P = 1$, it is sparse regression model for single dataset as used in Chen et al. (2010a) and Simon et al. (2013); when $K = 1$, $P > 1$, it can be reduced to sparse model on multiple datasets at the same level but from different platforms, similar to the work in Ma et al. (2011); when $K > 1$, $P = 1$, it can work for multiple datasets at different levels.

SOLUTION ALGORITHM BY ALTERNATIVE DIRECT METHOD (ADM)

Although both (2) and (3) are convex optimization problem with global solutions, the non-smoothness and the composite norms still cause difficulties in solving the optimization. Several algorithms have been studied to address such an issue for single task regression models, e.g., second-order cone programming (SOCP) algorithm (Candes and Romberg, 2005), spectral projected gradient method (SPGL1) (van den Berg et al., 2008), accelerated gradient method (SLEP) (Liu et al., 2009), block-coordinate descent algorithm and SpaRSA (Wright et al., 2009). In sparse multitask regression model, since the loss function is separable, these algorithms are still applicable but expensive in computations. In this study, we apply ADM to solve sparse multitask regression model. ADM uses the splitting strategy to decompose the optimization problem into several easily solvable ones and updates the variable in each subproblem iteratively until the convergence is achieved. It has been successfully applied to solve many convex or non-convex optimization problems, such as lasso (Yang and Zhang, 2011), total variation regularization (Esser, 2009), matrix decomposition and our recent work on sparse low rank decomposition (Dongdong et al., 2013). Deng et al. compared ADM with several other algorithms and found that ADM outperformed others with more robustness and faster computation (Deng et al., 2013).

Taking the model in (2) for example, we use ADM to split the penalties and transform (2) into the following optimization:

$$min_{\boldsymbol{C}} \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V_1}_{\{k \in G_q\}} \right\|_2 + \lambda_2 \|\boldsymbol{V_2}\|_{1,1} \tag{4}$$
$$s.t.\ \boldsymbol{C} = \boldsymbol{V_1}, \boldsymbol{C} = \boldsymbol{V_2}$$

where $\boldsymbol{V_1}, \boldsymbol{V_2}$ are two variables making the loss function separable. The augmented Lagrange function can be derived as

$$\begin{aligned} & L\left(\boldsymbol{C}, \boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{D_1}, \boldsymbol{D_2}, \lambda_1, \lambda_2, \mu, \rho\right) \\ & = \sum_{i=1}^{K} \omega_i \sum_{j=1}^{P_i} \delta_j \left\| \boldsymbol{Y}^{(i,j)} - \boldsymbol{X}^{(i,j)} \boldsymbol{C}^{(i,j)} \right\|_F^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V_1}_{\{k \in G_q\}} \right\|_2 \\ & + \lambda_2 \|\boldsymbol{V_2}\|_{1,1} + \frac{\rho}{2} \|\boldsymbol{C} - \boldsymbol{V_1} - \boldsymbol{D_1}\|_2^2 + \frac{\rho}{2} \|\boldsymbol{C} - \boldsymbol{V_2} - \boldsymbol{D_2}\|_2^2 \end{aligned} \tag{5}$$

where ρ is augmentedLagrangian parameter which can be updated iteratively; $\boldsymbol{D_1}$,$\boldsymbol{D_2}$ are the Lagrange multipliers to approximate the residuals between $\boldsymbol{C}$ and $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, respectively. Since the objective function and constraints are both separable and convex, ADM method is effective to solve $\{\boldsymbol{C}, \boldsymbol{V_1}, \boldsymbol{V_2}, \boldsymbol{D_1}, \boldsymbol{D_2}\}$ sequentially. We present the algorithm for solving multitask-sglasso by ADM in **Table 1**.

Remark 1. We decouple (2) into several small convex optimization problems. Step 3 is a regular least square estimation on matrix $\boldsymbol{C}$, where an analytical solution can be derived. Step 4 is a classical sparse group lasso minimization, which can be solved efficiently by block coordinate decent in Sprechmann et al. (2011). Step 5 is a simple lasso problem, which can also be solved by soft-thresholding. The division of complex optimization into

Table 1 | Algorithm of solving multitask-sglasso by ADM.

```
1  Initialization: $k = 0$, choose $\lambda_1, \lambda_2, \mu, \rho, > 0, \boldsymbol{V}_1^0, \boldsymbol{V}_2^0, \boldsymbol{D}_1^0, \boldsymbol{D}_2^0$
2  Repeat:
3      $\boldsymbol{C}^{k+1} \leftarrow argmin_A L\left(\boldsymbol{C}, \boldsymbol{V}_1^k, \boldsymbol{V}_2^k, \boldsymbol{D}_1^k, \boldsymbol{D}_2^k\right)$
4      $\boldsymbol{V}_1^{k+1} \leftarrow argmin_{V_1} L\left(\boldsymbol{C}^{k+1}, \boldsymbol{V}_1, \boldsymbol{V}_2^k, \boldsymbol{D}_1^k, \boldsymbol{D}_2^k\right)$
           $= \mathrm{argmin}_{V_1} \frac{\rho}{2} \left\| \boldsymbol{C}^{k+1} - \boldsymbol{V}_1 - \boldsymbol{D}_1^k \right\|_2^2 + \lambda_1 \sum_{q=1}^{Q} \left\| \boldsymbol{V_1}_{\{k \in G_q\}} \right\|_2$
5      $\boldsymbol{V}_2^{k+1} \leftarrow argmin_{V_2} L\left(\boldsymbol{C}^{k+1}, \boldsymbol{V}_1^{k+1}, \boldsymbol{V}_2, \boldsymbol{D}_1^k, \boldsymbol{D}_2^k\right)$
           $= argmin_{V_2} \frac{\rho}{2} \left\| \boldsymbol{C}^{k+1} - \boldsymbol{V}_2 - \boldsymbol{D}_2^k \right\|_2^2 + \lambda_2 \|\boldsymbol{V_2}\|_{1,\,1}$
6      Update Lagrange multipliers
       $\boldsymbol{D}_1^{k+1} \leftarrow \boldsymbol{D}_1^k - \boldsymbol{C}^{k+1} + \boldsymbol{V}_1^{k+1}$
       $\boldsymbol{D}_2^{k+1} \leftarrow \boldsymbol{D}_2^k - \boldsymbol{C}^{k+1} + \boldsymbol{V}_2^{k+1}$
7      Update iteration $k \leftarrow k + 1$
8  Until some stopping criterion is satisfied
```

several simple sub-optimizations will improve the efficiency of computation.

Remark 2. We adopt the stopping criterion as suggested by Boyd et al. (2010) that both primal res$_{\text{pri}}$ and dual res$_{\text{dual}}$ residuals must be small, i.e., $res_{pri} \leq \varepsilon_{pri}$, $res_{dual} \leq \varepsilon_{dual}$, where primal residual indicates the difference between $\boldsymbol{C}$ and $\boldsymbol{V_1}$ ($\boldsymbol{V_2}$) while dual residual measures the difference between $\boldsymbol{V_1}$ ($\boldsymbol{V_2}$) and the values at the last iteration.

Remark 3. The convergence rate depends on the choice of Lagrangian parameter ρ. Some studies adjust ρ based on primal and dual variables iteratively to speed up the convergence. In this work, we update ρ by keeping the ratio between primal and dual residual norms within a given interval, say [0.1, 10] until they both converge to zeros.

For optimization (3), it can similarly be transformed into ADM formulation where only one splitting variable (i.e., $\boldsymbol{V_1}$) is needed to separate (3) into two subproblems. The estimation of $\boldsymbol{V_1}$ at Step 4 can be replaced by:

$$\boldsymbol{V}_{\boldsymbol{1}}^{\boldsymbol{k+1}} \leftarrow argmin_{\boldsymbol{V_1}} \frac{\rho}{2} \left\| \boldsymbol{C}^{\boldsymbol{k+1}} - \boldsymbol{V_1} - \boldsymbol{D}_{\boldsymbol{1}}^{\mathbf{k}} \right\|_2^2 + \lambda \sum_{q=1}^{Q} \left\| \boldsymbol{C}_{\{k \in G_q\}} \right\|_1^2 \tag{6}$$

where soft-threshold can be used to get the solution.

STATISTICAL TEST

λ_1, λ_2, and λ are tuning parameters used to control the number of genes and features within a gene. The K-fold cross validation is widely used to select optimal values of these parameters. Briefly, the subjects are divided into k groups, where k−1 groups of subjects are used for estimating the coefficient matrix $\boldsymbol{C}$ and the rest group of subjects is used to calculate the prediction error by the estimated $\boldsymbol{C}$. We set λ_1, λ_2, and λ to $[10^{0.1}, 10^{0.2}, \ldots, 10^3]$ with 30 values. We search the 30 × 30 grid to find an optimal combination of $(\lambda_1^*, \lambda_2^*)$ for multitask-sglasso and similarly optimal value of λ^* for multitask-sgridge by 5-fold cross validation. Finally, the estimate of $\boldsymbol{C}$ can be calculated by the derived optimal parameters.

To test the significance of identified biomarkers with non-zeros coefficients at $\boldsymbol{C}$, we construct a gene based statistical test to measure the strength and significance of the association between genes and phenotype across experiments from different platforms and levels. For the i-th gene G_i, $\left\{\widehat{\boldsymbol{C}}_i^{(j)} | j = 1, 2, \ldots, T\right\}$ indicates the corresponding coefficient vector estimated from the j-th experiment, denoted by $\left[\hat{C}_{i,1}^{(j)}, \hat{C}_{i,2}^{(j)}, \ldots, \hat{C}_{i,m_i}^{(j)}\right]$, where m_i is the number of features annotated to gene G$_i$ in the j-th experimental dataset. The null hypothesis is there is no association between the i-th gene and phenotype in all T experiments, denoted by $H_0 : \left[\hat{C}_i^{(1)\prime}, \hat{C}_i^{(2)\prime}, \ldots, \hat{C}_i^{(T)\prime}\right]^{\prime} = 0$, vs. the alternative hypothesis $H_A : \hat{C}_i^{(k)} \neq 0,\ k = 1, 2, \ldots, T$ for some k. To test the hypothesis, we summarize the coefficients of the i-th gene on all datasets as follows.

$$\hat{S}_i = \sqrt{\sum_{j=1}^{T} \left\| \hat{C}_i^{(j)} \right\|_2^2} \tag{7}$$

where $\hat{S}_i$, $i = 1, 2, \ldots, Q$ is the statistical value on all Q genes. Due to different number of features included in different genes, an adjustment for gene size is necessary. A permutation based approach is used to reduce the potential bias due to varying gene size. The standardized gene level statistic is given by

$$\tilde{S}_i = \frac{\hat{S}_i - \hat{S}_i^0}{\widehat{\sigma}_i} \tag{8}$$

where $\hat{S}_i^0$ and $\widehat{\sigma}_i$ are the mean and standard deviation of the i-th gene under the null hypothesis. Samples are permuted B times to construct null distribution of $\hat{S}_i$, denoted by $\widehat{\Gamma}_i^0 = \{\hat{S}_{i,j}^0 | j = 1, 2, \ldots, B\}$. $\hat{S}_i^0$ and $\widehat{\sigma}_i$ are then estimated based on permutation data. Since all $\hat{S}_{i,j}^0$ have been normalized, we could pool all $\widehat{\Gamma}_i^0$ into a set $\Gamma^0 = \{\widehat{\Gamma}_i^0 | i = 1, 2, \ldots, Q\}$ as the estimated null distribution. Therefore, the gene-level p-value of the i-th gene can be calculated by

$$p_i = \frac{\#\ of\ \{\Gamma^0 \geq \hat{S}_i\}}{\#\ of\ \{\Gamma^0\}} \tag{9}$$

SIMULATION

To evaluate the performance of our proposed integrative method for identifying biomarkers, we simulated two levels of measurements: SNP and gene expression, and assigned different sample size for each dataset.

For each simulation, we generated 3 SNP datasets and 3 gene expression datasets. The sample sizes were 600, 400, and 200 for SNP data and 70, 50, and 30 for gene expression, respectively. 200 genes were simulated in each dataset. To mimic the linkage disequilibrium (LD) structure among SNPs, we chose a chromosome, chromosome 22, from HapMap CEU panel with phase III data and sample subjects by software HAPGEN2 (Su et al., 2011). Those SNPs were kept after the following filters were applied: (1) Minor allele frequency (MAF) at least 5%; and (2) Hardy-Weinberg Equilibrium (HWE) with significant level less than 0.001. We generated a dataset consisting of 15,235 SNPs which were assigned to 576 genes as the gene pool. Assuming an additive genetic model, each SNP was recorded as the count of minor allele (denoted as A) at that locus and thereby was valued by 0 (homozygote of major allele, aa), 1 (heterozygote, Aa) and 2 (homozygote of minor allele, AA). 200 genes including more than 10 SNPs were randomly selected from the pool, of which 20 genes were chosen as causal genes and 2 SNPs with MAF from uniform distribution (Unif) (0.15, 0.25) from each causal gene were further used to induce causal genetic effects on gene expression. The number of SNPs from 200 selected genes was randomly set from Unif(10,100) and those non-causal SNPs in each gene were selected from pooled SNPs.

We used SNP data to generate gene expression and phenotype data, referring to the similar method in Huang et al. (2014). Three SNP datasets with 70, 50, and 30 subjects were first simulated, as described in the method section. For each causal gene, e.g., gene i, the expression value G_i was derived from the causal SNPs in this gene by

$$G_i = \sum_{j=1}^{n} SNP_{causal}^{j} \beta_j + \varepsilon \tag{10}$$

where n was the number of causal SNPs included in G_i; and β_j indicated the effect of the j-th causal SNP(SNP_{causal}^{j}) on G_i. We set β value from Unif(1, 1.2) and noise ε from normal distribution $N(0, 1)$. The other non-causal gene expression values were generated by multivariate normal distribution $N(0, \Sigma)$, where Σ was the covariance matrix of gene expressions, and the expressions of gene i and j have correlation coefficient $0.3^{|i-j|}$. Based on the simulated gene expression, the phenotype was generated by the following formula:

$$logit\{Pr(Y_i = 1)\} = \sum_{j=1}^{m} G_{causal}^{j} \tau_j + \varepsilon' \tag{11}$$

where m was the number of causal genes, i.e., $m = 20$ in this study; G_{causal}^{j} was gene expression for the j-th causal gene and τ_j was the corresponding effects on the outcome. The logit function was used to generate binary outcome. The identity function can be used if the quantitative phenotype was used. ε' was non-genetic variable, which was assumed to follow normal distribution $N(0, 1)$.

RESULTS

SYNTHETIC DATA

We assessed the performance of the two proposed sparse multitask models- multitask sglasso and multitask sgridge-on each single dataset and all datasets, respectively, and also compared them with widely used meta-analysis on three SNP datasets (meta-SNP) and three gene expression datasets (meta-EXP). Meta-analysis was implemented by the software MetaL (Willer et al., 2010).

Simulation 1: Fixed effect of causal genes in diverse dataset

In this simulation, we studied the scenario that the effects of causal genes across diverse datasets were fixed, i.e., $\tau_j^1 = \tau_j^2 = \cdots = \tau_j^6$, $i = 1, 2, \ldots, m$, which indicated a causal gene had the same effect on all datasets. For m casual genes, first, we set a baseline vector $\boldsymbol{\eta} \in R^m$ from Unif(0.2, 2) and Unif(−2, −0.2). Next, to evaluate the performance of different methods on identifying casual genes under different levels of effects, a factor $\delta = 0,\ 0.2,\ 0.4,\ 0.6,\ 0.8,\ 1.0$ was multiplied by $\boldsymbol{\eta}$ to have the final value of gene effects $\boldsymbol{\tau} = \boldsymbol{\eta} \times \delta$. 50 replicates were performed and $B = 500$ permutations in each replicate were implemented to calculate empirical p-value of sparse multitask models. Finally, we compared the results of the following eight cases: multitask-sglasso on three expression datasets, three SNP datasets, and all six datasets; multitask-sgridge on three expression datasets, three SNP datasets and all six datasets; meta-analysis on three SNP datasets and three expression datasets.

Figure 1 shows the comparison result of a set of methods under different values of δ, i.e., [0, 0.2, 0.4, 0.6, 0.8, 1.0]. The ROC curves were plotted using the false positive rate against true positive rate by varying the p-value threshold from 10^{-4} to 1. It could be seen that all methods had similar performance when there were no effective causal genes in all datasets (i.e., $\delta = 0$). When the effects of causal genes (i.e., δ) increase, i.e., more variability of phenotypes could be explained by genetic variants, multitask-sglasso method shows better performance by removing the irrelevant genes with improved signal to noise ratio. When δ was greater than 0.2, multitask-sglasso methods on SNP,

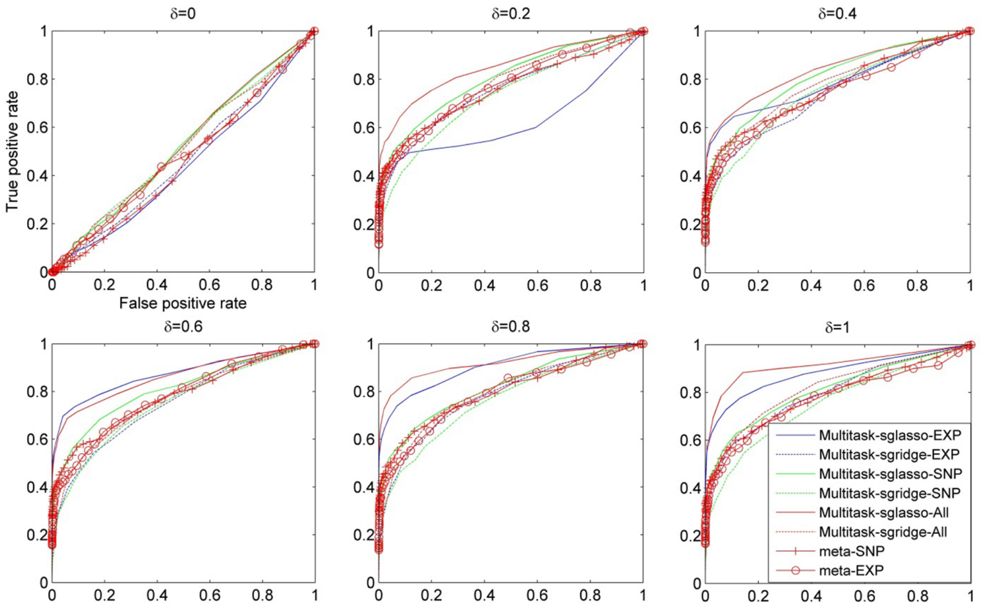

FIGURE 1 | The ROC curves for the comparison of eight cases: sparse multitask-sglasso and multitask-sgridge methods on three SNP datasets, expression datasets and all datasets, and meta-analysis on SNP and expression datasets, respectively.

expression and both datasets significantly outperformed the other methods. This indicates that Multitask-sglasso method showed better performance by integrating all datasets than that of using only one level of data. In addition, when δ was greater than 0.4, multitask-sglasso method using only SNP or expression datasets still gave higher power than meta-analysis method. Multitask-sgridge method had less power than multitask-sglasso method and only showed better performance than meta-analysis method when causal genes have high effect sizes.

Simulation 2: Dynamic effects of causal genes in diverse datasets

In this simulation, we consider the situation that a causal gene has different effects at different levels and platforms. This is more likely to happen for real datasets since multiple datasets are usually generated from different studies with different study protocols, profiling techniques, and experimental platforms, leading to dynamic effect sizes of casual genes. We aimed to compare the performance of our sparse multitask methods with meta-analysis for biomarker identification in this dynamic case. Six datasets were generated with the same sample size and causal genes as those in the first simulation study. We simulated the dynamic effects of causal genes at different datasets by setting $\tau_j \sim N\left(\eta, \sigma^2\right), i = 1, 2, \ldots, 6$, where η was fixed effect as described above, and σ was standard deviation indicating the dynamic effect of genes across datasets. We changed the value of σ from 0 to 1 with the interval of 0.2 to show different extent of heterogeneity of causal genes across diverse datasets. 50 replicates were averaged to draw the ROC curve for comparison.

Figure 2 showed the comparison result of eight cases under dynamic effect models with variance of causal genes varying from 0 to 1. When $\sigma = 0$, the models reduced to the ones with fixed effects. When σ was greater than 0.4, sparse multitask-sglasso method on SNP, expression and both datasets significantly outperformed other methods in identifying casual genes. Except for sparse multitask-sglasso method, we can also see that the performance of sparse multitask-sgridge on all datasets was better than meta-analysis methods, which indicated the advantage of multitask method for integrating diverse datasets.

REAL DATA ANALYSIS

In this study, we took advantage of 3 gene expression datasets and 1 GWAS dataset with bone mineral density (BMD) measurements from our previous studies. The cohort I of gene expression data contained 80 Caucasian females, including 40 high and 40 low hip subjects (Chen et al., 2010b). The cohort II of gene expression data contained 19 Caucasian females, including 10 high and 9 low hip BMD subjects (Liu et al., 2005). The cohort III of gene expression data contained 26 Chinese females, all premenopausal and including 14 high and 12 low hip BMD subjects (Lei et al., 2009). For the GWAS dataset, SNP data were obtained using Affymetrix 500K arrays on 1,000 unrelated homogeneous Caucasians. After a suite of quality control procedures were performed, the SNP set for subsequent analysis contained 379,319 SNPs, yielding an average marker spacing of ~7.9 kb throughout the human genome (Xiong et al., 2009).

We combined gene expression and SNP datasets to identify those risk genes of BMD by our sparse multitask-sglasso integrative method. We chose one chromosome 6 containing the largest number of genes to perform gene-based analysis. 504 genes were included in the chromosome. More details in each dataset were given in **Table 2**.

We applied sparse multitask-sglasso method to SNP, gene expression and both datasets, respectively. To compare with meta-analysis, two gene expression datasets with the same level and experimental platforms, EXP-19 and EXP-80, were used for meta-analysis, denoted by meta-Exp. The most significant expression

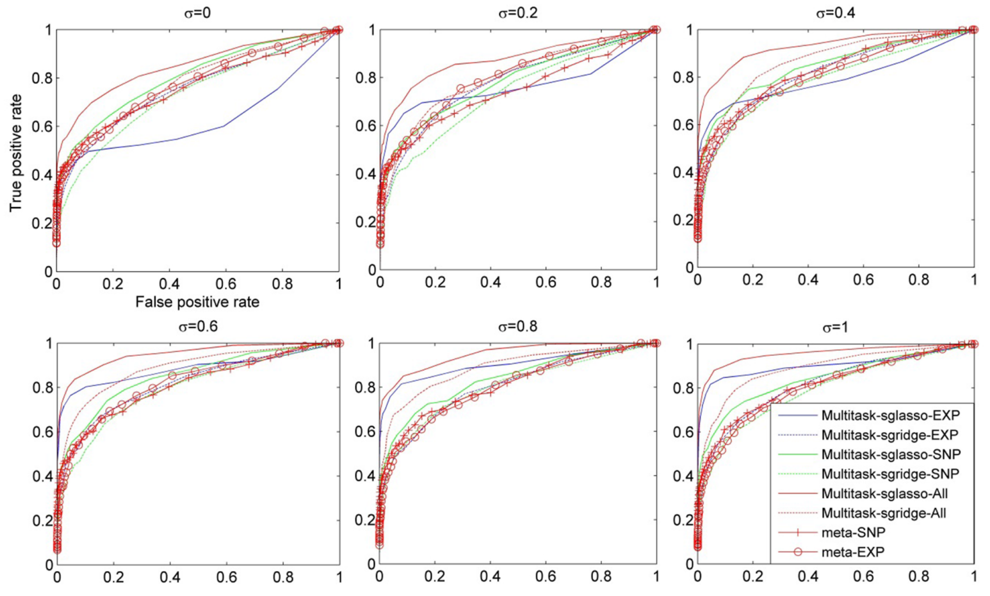

FIGURE 2 | The comparison of eight methods on three SNP and three expression datasets simulated with the dynamic model. The variance of effect size of causal genes is set to normal distribution with variance varying from 0 to 1 at an interval of 0.2.

Table 2 | A summary of four datasets from different levels and platforms used in this analysis.

Data type	Platform	Gene	Genetic variants	Sample
SNP	Affymetrix 500K	504	10685	1000
Gene expression	HG-U133A	504	874	19
Gene expression	HG-U133A	504	1225	26
Gene expression	HG-U133A-Plus_2.0	504	874	80

measurement in each gene was chosen to represent significance level of the gene. **Figure 3** shows the Venn diagram of gene list by three methods: multitask-sglasso on all gene expression datasets, multitask-sglasso on all gene expression and SNP datasets, and meta-analysis on two expression datasets under the significant threshold 0.05. We could see that there were 45 genes shared by meta-Exp and multitask-sglasso on three expression datasets; 10 genes overlapped by meta-Exp and multitask-sglasso on both SNP and expression datasets; and three genes ("GPR116," "HLA-DMB," "PHACTR1") identified by all methods. The small overlapping between multitask-sglasso Exp and multitask sglasso SNP + Exp is due to the use of additional information from large sample size of SNP dataset.

Table 3 lists 7 top significant genes identified and sorted by their *p*-values from sparse multitask-sglasso method on all datasets and the corresponding *p*-values by meta-analysis. Note that the *p*-values of the same gene usually were different in different studies. For example, SOD2 had much lower *p*-values in SNP and EXP-26 datasets than those in other datasets. This difference showed the dynamic effects of genes across diverse datasets with different levels and platforms. There are three genes ("TREML2," " HTR1E," and "GLO1") shared by sparse multitask-sglasso method on all of datasets and meta-Exp. Except for gene TREML2, the *p*-values of genes derived from all datasets were lower than those from the other methods, indicating higher level of significance given by our multitask method. The relatively smaller *p*-values of these genes in SNP data were due to the large sample size of SNP dataset, which will give more confidence on the findings.

To further evaluate the significance of identified genes by multitask-sglasso, we performed gene level meta-analysis on three independent BMD studies for validation, more details were shown in supplementary data. The result (Table S1) listed the *p*-values of 24 identified genes based on single studies and meta-analysis. Most of these genes showed significant effects on BMD ($p < 0.01$), indicating the effectiveness of our sparse multitask regression method in identifying genetic risk factors.

Three shared genes ("TREML2," "HTR1E," and "GLO1") may have important biological functions related to BMD associated with osteoporosis. TREML2 (also known as TLT-2) was located in a gene cluster on chromosome 6 with the single Ig variable (IgV) domain activating receptors TREM1 and TREM2, while these TREM receptor families were found to participate in the process of bone homeostasis by controlling the rate of osteoclastogenesis and regulating the differentiation of osteoclasts (Klesney-Tait et al., 2006; Otero et al., 2012). HTR1E was recently identified to contain SNPs significantly associated with a linear combination of multiple osteoporosis-related phenotypes including BMD (Karasik et al., 2012). GLO1, as a binding protein of methyl-gerfelin (M-GFN), was found to be able to result in the inhibition of osteoclastogenesis (Kawatani et al., 2008). Besides these three common genes, our method was also able to identify other osteoporosis-susceptible genes but was undetectable by meta-analysis. For instance, SOD2 has been identified as the gene susceptible to osteoporosis in our previous integrative analysis of mRNA, SNP, and protein data (Deng et al., 2011). It may play a significant role in BMD variation and pathogenesis of osteoporosis. HDAC2, as a member of histone deacetylases (HDACs), was found to play a critical role in bone development and biology (McGee-Lawrence and Westendorf, 2011). These genes were missed out with meta-analysis but can be detected with our proposed method, showing improved sensitivity.

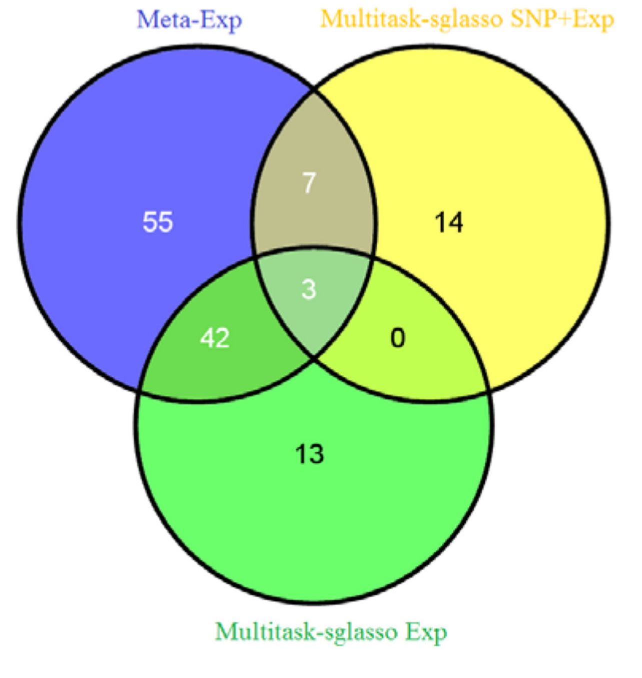

FIGURE 3 | The Venn diagram of identified genes by three methods: meta-analysis on EXP-19 and EXP-80 datasets, multitask-sglasso on all three expression datasets and multitask-sglasso on all gene expression and SNP datasets.

CONCLUSION AND DISCUSSION

In this work, we proposed a multi-omics integration method, i.e., sparse group multitask regression model, which can integrate multiple genomic datasets from different levels, platforms, and subjects for gene based analysis. An efficient computational algorithm based on ADM was provided for its solution. The performance of the model was compared with meta-analysis in simulation datasets. The simulation results showed that our sparse group multitask regression model can increase the power of detecting risk genes by integrating multiple diverse datasets effectively. In particular, multitask-sglasso model outperformed

Table 3 | The top 7 identified genes and their *p*-values by sparse multitask-sglasso method in bone mineral density studies.

Methods Gene ID	SNP	EXP-19	EXP-26	EXP-80	EXP_all	Meta-EXP	SNP + EXP
SOD2	0.0021	0.9136	0.0017	0.9566	0.7152	0.0752	0.0016
TREML2*	0.0014	0.1295	0.5243	0.1648	0.1665	0.0312	0.0018
HTR1E*	0.0030	0.4062	0.3481	0.0963	0.0750	0.0203	0.0023
HDAC2	0.0067	0.0089	0.1118	0.4382	0.4360	0.0553	0.0032
HCRTR2	0.0045	0.1074	0.5972	0.3293	0.3282	0.6297	0.0044
MUT	0.0073	0.2173	0.7665	0.9763	0.9910	0.571	0.0055
GLO1*	0.0084	0.0651	0.6182	0.1012	0.1298	0.0183	0.0073

** Genes identified by both meta-Exp and sparse multitask-sglasso on all datasets.*

meta-analysis method in simulations on genes with both fixed and dynamic effects. Our real data analysis on osteoporosis studies identified significant genes but missed by meta-analysis, and these genes were reported to be highly susceptible to BMD and osteoporosis. Overall, the advantages of our sparse group multitask regression method for biomarker identification from multiple omics datasets include: (1) it can combine diverse and complementary omic datasets without; (2) group the features by gene or gene set to account for the group structures in data (e.g., LD structure, co-expression, and genetic regulatory network); (3) remove irrelevant genes and/or features within a gene simultaneously.

Our proposed sparse multitask regression model provided a general framework for integrative analysis of diverse datasets. To fuse multiple diverse datasets, we considered the regression on each single dataset as a single task and then combined all single tasks into the model. Two sets of parameters were used in the model. ω_is were used to weight object functions (i.e., data fitting term at each level) different levels, while δ_j were used for different platforms. Similar to other works, we set ω to be equal by assuming each level of genetic data contains the same information (Ma et al., 2011). We assign δ_j to the data from different platforms by their sample sizes (Wilson and Lipsey, 2001). Other methods can also be applied to estimating weights such as Kaplan–Meier estimate (Liu et al., 2013) and inverse variance (Wilson and Lipsey, 2001). In order to account for the group effects and reduce the large number of features, we used two group sparse penalties in our multitask regression models, i.e., sparse group lasso and sparse group ridge, respectively. These penalties can perform feature selection at both group level and individual for multiple dataset levels, showing better performance than those of using lasso and group lasso penalties for single dataset analysis. Similar regression models were also recently proposed for using two-level sparse group penalties such as group bridge and group MCP (Huang et al., 2012). Ma et al. has recently applied these penalties in regression model for cancer studies to identify those risk oncology genes by integrating multiple expression level datasets from different cancer studies (Liu et al., 2013). Chen et al. has also compared and found that sparse group ridge outperformed group bridge penalty in single dataset regression model (Chen et al., 2010c). However, no study has been performed to compare them for multiple dataset integration and further work is needed in this direction.

WEB SOURCES

The gene expression datasets from three cohorts can be accessed in GEO database (http://www.ncbi.nlm.nih.gov/geo/) with the following accession numbers: 19 Caucasians BMD study (GSE2208), 26 Chinese study (GSE7158), and 80 Caucasians study (GSE56815).

ACKNOWLEDGMENTS

Yu-Ping Wang was partially supported by NIH grants (R01MH104680, R01GM109068). Hong-Wen Deng was partially supported by NIH grants (P50AR055081, R01AG026564, R01AR059781, R01AR050496, and R01AR057049) and Edward G. Schlieder Endowment. This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or NHLBI. The FHS datasets used for the analyses described in this manuscript were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap through dbGaP accession phs000007.v14.p6.

FUNDING SOURCE

The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute NHLBI in collaboration with Boston University (Contract No. N01-HC-25195). Funding for SHARe Affymetrix genotyping was provided by NHLBI Contract N02-HL-64278. SHARe Illumina genotyping was provided under an agreement between Illumina and Boston University. Funding support for the Framingham Bone Mineral Density datasets was provided by NIH grants R01AR/AG 41398.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fcell.2014.00062/abstract

REFERENCES

Balbin, O. A., Prensner, J. R., Sahu, A., Yocum, A., Shankar, S., Malik, R., et al. (2013). Reconstructing targetable pathways in lung cancer by integrating diverse omics data. *Nat. Commun.* 4, 2617. doi: 10.1038/ncomms3617

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2010). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* 3, 1–122. doi: 10.1561/2200000016

Candes, E., and Romberg, J. (2005). *l1-Magic: Recovery of Sparse Signals via Convex Programming*. Available online at: http://users.ece.gatech.edu/justin/l1magic/downloads/l1magic.pdf

Chen, L., Hutter, C., Potter, J. D., Liu, Y., Prentice, R. L., and Peters, U. L. H. (2010c). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86, 860–871. doi: 10.1016/j.ajhg.2010.04.014

Chen, L. S., Hutter, C. M., Potter, J. D., Liu, Y., Prentice, R. L., Peters, U., et al. (2010a). Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* 86, 860–871. doi: 10.1016/j.ajhg.2010.04.014

Chen, X. D., Xiao, P., Lei, S. F., Liu, Y. Z., Guo, Y. F., Deng, F. Y., et al. (2010b). Gene expression profiling in monocytes and SNP association suggest the importance of the STAT1 gene for osteoporosis in both Chinese and Caucasians. *J. Bone Miner. Res.* 25, 339–355. doi: 10.1359/jbmr.090724

Chen, Y., Wu, X., and Jiang, R. (2013). Integrating human omics data to prioritize candidate genes. *BMC Med. Genomics* 6:57. doi: 10.1186/1755-8794-6-57

Deng, F. Y., Lei, S. F., Chen, X. D., Tan, L. J., Zhu, X. Z., and Deng, H. W. (2011). An integrative study ascertained SOD2 as a susceptibility gene for osteoporosis in Chinese. *J. Bone Miner. Res.* 26, 2695–2701. doi: 10.1002/jbmr.471

Deng, W., Yin, W., and Zhang, Y. (2013). "Group sparse optimization by alternating direction method," in *SPIE Optical Engineering+ Applications: 2013: International Society for Optics and Photonics; 88580R-88580R-88515* (San Diego, CA).

Dongdong, L., Hao, H., Jingyao, L., Hong-Wen, D., Calhoun, V. D., and Yu-Ping, W. (2013). "Network-based investigation of genetic modules associated with functional brain networks in schizophrenia," in *2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Beijing), 9–16.

Esser, E. (2009). *Applications of Lagrangian-Based Alternating Direction Methods and Connections to Split Bregman*. Technical Report 09-31. Berkeley, CA: University of California.

Evangelou, E., and Ioannidis, J. P. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14, 379–389. doi: 10.1038/nrg3472

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv*:1001.0736.

Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M., and Beyene, J. (2009). Data integration in genetics and genomics: methods and challenges. *Hum. Genomics Proteomics* 2009:869093. doi: 10.4061/2009/869093

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Stat. Sci.* 27, 481–499. doi: 10.1214/12-STS392

Huang, Y. T., Vanderweele, T. J., and Lin, X. (2014). Joint analysis of Snp and Gene expression data in genetic association studies of complex diseases. *Ann. Appl. Stat.* 8, 352–376. doi: 10.1214/13-AOAS690

Huttenhower, C., Hibbs, M., Myers, C., and Troyanskaya, O. G. (2006). A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics* 22, 2890–2897. doi: 10.1093/bioinformatics/btl492

Karasik, D., Cheung, C. L., Zhou, Y., Cupples, L. A., Kiel, D. P., and Demissie, S. (2012). Genome−wide association of an integrated osteoporosis−related phenotype: is there evidence for pleiotropic genes? *J. Bone Miner. Res.* 27, 319–330. doi: 10.1002/jbmr.563

Kawatani, M., Okumura, H., Honda, K., Kanoh, N., Muroi, M., Dohmae, N., et al. (2008). The identification of an osteoclastogenesis inhibitor through the inhibition of glyoxalase I. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11691–11696. doi: 10.1073/pnas.0712239105

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z., and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 28, 3290–3297. doi: 10.1093/bioinformatics/bts595

Klesney-Tait, J., Turnbull, I. R., and Colonna, M. (2006). The TREM receptor family and signal integration. *Nat. Immunol.* 7, 1266–1273. doi: 10.1038/ni1411

Lei, S. F., Wu, S., Li, L. M., Deng, F. Y., Xiao, S. M., Jiang, C., et al. (2009). An *in vivo* genome wide gene expression study of circulating monocytes suggested GBP1, STAT1 and CXCL10 as novel risk genes for the differentiation of peak bone mass. *Bone* 44, 1010–1014. doi: 10.1016/j.bone.2008.05.016

Lin, D., Zhang, J., Li, J., Calhoun, V. D., Deng, H. W., and Wang, Y. P. (2013). Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics* 14:245. doi: 10.1186/1471-2105-14-245

Liu, J., Huang, J., Xie, Y., and Ma, S. (2013). Sparse group penalized integrative analysis of multiple cancer prognosis datasets. *Genet. Res. (Camb).* 95, 68–77. doi: 10.1017/S0016672313000086

Liu, J., Ji, S., and Ye, J. (2009). *SLEP: Sparse learning with efficient projections. Arizona State University 6*. Available online at: http://www.public.asu.edu/~jye02/Software/SLEP

Liu, Y. Z., Dvornyk, V., Lu, Y., Shen, H., Lappe, J. M., Recker, R. R., et al. (2005). A novel pathophysiological mechanism for osteoporosis suggested by an *in vivo* gene expression study of circulating monocytes. *J. Biol. Chem.* 280, 29011–29016. doi: 10.1074/jbc.M501164200

Ma, S., Huang, J., and Song, X. (2011). Integrative analysis and variable selection with multiple high-dimensional data sets. *Biostatistics* 12, 763–775. doi: 10.1093/biostatistics/kxr004

McGee-Lawrence, M. E., and Westendorf, J. J. (2011). Histone deacetylases in skeletal development and bone mass maintenance. *Gene* 474, 1–11. doi: 10.1016/j.gene.2010.12.003

Otero, K., Shinohara, M., Zhao, H., Cella, M., Gilfillan, S., Colucci, A., et al. (2012). TREM2 and beta-catenin regulate bone homeostasis by controlling the rate of osteoclastogenesis. *J. Immunol.* 188, 2612–2621. doi: 10.4049/jimmunol.1102836

Phan, J. H., Quo, C. F., Cheng, C., and Wang, M. D. (2012). Multiscale integration of -omic, imaging, and clinical data in biomedical informatics. *IEEE Rev. Biomed. Eng.* 5, 74–87. doi: 10.1109/RBME.2012.2212427

Rhodes, D. R., and Chinnaiyan, A. M. (2005). Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37(Suppl.), S31–S37. doi: 10.1038/ng1570

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Stat.* 22, 231–245. doi: 10.1080/10618600.2012.681250

Song, R., Huang, J., and Ma, S. (2012). Integrative prescreening in analysis of multiple cancer genomic studies. *BMC Bioinformatics* 13:168. doi: 10.1186/1471-2105-13-168

Sprechmann, P., Ramirez, I., Sapiro, G., and Eldar, Y. C. (2011). C-HiLasso: a collaborative hierarchical sparse modeling framework. *IEEE Trans. Signal Process.* 59, 4183–4198. doi: 10.1109/TSP.2011.2157912

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* 27, 2304–2305. doi: 10.1093/bioinformatics/btr341

Troyanskaya, O. G., Dolinski, K., Owen, A. B., Altman, R. B., and Botstein, D. (2003). A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci. U.S.A.* 100, 8348–8353. doi: 10.1073/pnas.0832373100

van den Berg, E., Schmidt, M., Friedlander, M. P., and Murphy, K. (2008). *Group Sparsity via Linear-Time Projection*. Vancouver, BC: Dept Comput Sci, Univ British Columbia.

Willer, C. J., Li, Y., and Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191. doi: 10.1093/bioinformatics/btq340

Wilson, D. B., and Lipsey, M. (2001). *Practical Meta-Analysis*. *Оригинал презен тации* Available online at: http://www.mason.gmu.edu.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. (2009). Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.* 57, 2479–2493. doi: 10.1109/TSP.2009.2016892

Xiong, D. H., Liu, X. G., Guo, Y. F., Tan, L. J., Wang, L., Sha, B. Y., et al. (2009). Genome-wide association and follow-up replication studies identified ADAMTS18 and TGFBR3 as bone mass candidate genes in different ethnic groups. *Am. J. Hum. Genet.* 84, 388–398. doi: 10.1016/j.ajhg.2009.01.025

Xiong, Q., Ancona, N., Hauser, E. R., Mukherjee, S., and Furey, T. S. (2012). Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. *Genome Res.* 22, 386–397. doi: 10.1101/gr.124370.111

Yang, J., and Zhang, Y. (2011). Alternating direction algorithms for \ell_1-problems in compressive sensing. *SIAM J. Sci. Comput.* 33, 250–278. doi: 10.1137/090777761

Zhang, D., and Shen, D. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 895–907. doi: 10.1016/j.neuroimage.2011.09.069

Zhang, K., Gray, J. W., and Parvin, B. (2010). Sparse multitask regression for identifying common mechanism of response to therapeutic targets. *Bioinformatics* 26, i97–i105. doi: 10.1093/bioinformatics/btq181

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2014; accepted: 01 October 2014; published online: 27 October 2014.

Citation: Lin D, Zhang J, Li J, He H, Deng H-W and Wang Y-P (2014) Integrative analysis of multiple diverse omics datasets by sparse group multitask regression. Front. Cell Dev. Biol. ***2****:62. doi: 10.3389/fcell.2014.00062*

This article was submitted to Systems Biology, a section of the journal Frontiers in Cell and Developmental Biology.