Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Liu et al., J Biomet Biostat 2015, 6:2
DOI: 10.4172/2155-6180.1000228

Research Article Open Access

# A Penalized Regression Approach for Integrative Analysis in Genome-Wide Association Studies

**Liu J[1], Wang F[2], Gao X[3], Zhang H[4], Wan X[5] and Can Yang[6]***

[1]Centre of Quantitative Medicine, Duke-NUS Graduate Medical School, Singapore
[2]Department of Biostatistics and Epidemiology, University of Pennsylvania, USA
[3]Department of Ophthalmology and Visual Science, University of Illinois, Chicago, USA
[4]Department of Psychiatry, Yale University, USA
[5]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[6]Department of Mathematics, Hong Kong Baptist University, Hong Kong

## Abstract

Over one thousand genome-wide association studies (GWAS) have been conducted in the past decade. Increasing biological evidence suggests the polygenic genetic architecture of complex traits: a complex trait is affected by many risk variants with small or moderate effects jointly. Meanwhile, recent progress in GWAS suggests that complex human traits may share common genetic bases, which is known as "pleiotropy". To further improve statistical power of detecting risk genetic variants in GWAS, we propose a penalized regression method to analyze the GWAS dataset of primary interest by incorporating information from other related GWAS. The proposed method does not require the individual-level of genotype and phenotype data from other related GWAS, making it useful when only summary statistics are available. The key idea of the proposed approach is that related traits may share common genetic basis. Specifically, we propose a linear model for integrative analysis of multiple GWAS, in which risk genetic variants can be detected via identification of nonzero coefficients. Due to the pleiotropy effect, there exist genetic variants affecting multiple traits, which correspond to a consistent nonzero pattern of coefficients across multiple GWAS. To achieve this, we use a group Lasso penalty to identify this nonzero pattern in our model, and then develop an efficient algorithm based on the proximal gradient method. Simulation studies showed that the proposed approach had satisfactory performance. We applied the proposed method to analyze a body mass index (BMI) GWAS dataset from a European American (EA) population and achieved improvement over single GWAS analysis.

**Keywords:** Integrative analysis of GWAS; Penalized methods; Scaled group Lasso

## Introduction

Genome-wide association studies (GWAS) provide an unprecedented opportunity for identifying disease-associated genetic variants. Although disease associated SNPs at genome-wide significance level (e.g. P-value$<5 \times 10^{-8}$) were identified for some diseases [1-4], those identified SNPs only explained a small fraction of genetic contributions to the etiology of the diseases. This phenomenon is referred to as *"missing heritability"*. Rather than only using genome-wide significant SNPs, Yang et al. [5] showed that 45% of the variance for human height can be explained by using all of the genotyped common SNPs. This result suggests that most of the "missing heritability" is not missing but remains hidden in the genome: due to the limited sample size, many individual effects of genetic markers are too weak to pass the genome-wide significance level, and thus those risk genetic variants remain undiscovered. So far, people have found similar genetic architectures for many other complex diseases [4], such as psychiatric disorders [6,7], i.e., the phenotype is affected by many genetic variants with small or moderate effects, which are referred to as "polygenicity". The polygenicity of complex diseases is further supported by recent GWAS with larger sample sizes, in which more associated common SNPs with moderate effects have been identified (e.g., GWAS data from 34,840 patients and 114,981 healthy people are analyzed to understand the genetic architecture of type 2 diabetes [8]). However, large sample recruitment may be expensive and time-consuming.

For single GWAS analysis, many existing statistical methods have been proposed [9,10]. Among them, penalized regression methods [11-14] have drawn particular attention in GWAS. However, due to limited sample size of a single GWAS and polygenicity of a complex trait, many existing methods do not have enough power to uncover the remaining risk genetic variants. Recently, increasing evidence suggests that complex traits may share common genetic bases, which is known as "pleiotropy" [15-17]. A systematic investigation of pleiotropy [18] suggests that 16.9% of genes and 4.6% of SNPs have been reported to show pleiotropic effects. Therefore, it is possible to further improve statistical power in GWAS data analysis by integrating multiple GWAS. The difficulties of integrative analysis of GWAS mainly come from two aspects. First, a direct pool of samples from multiple GWAS is questionable due to heterogeneity in different studies. Second, some existing methods (e.g. [19]) require the availability of all genotype data from multiple GWAS, which could be practically difficult due to the privacy restrictions on sharing individual-level data.

In this work, we aim at improving statistical power of identifying associated markers for the given GWAS data by integrating information from other GWAS, where only the summary statistics rather than the geneotype data of some GWAS are needed. We propose a penalized method for integrating multiple GWAS (pIGWAS). The key idea of the proposed approach is that genetically related traits can share common genetic bases [18,20], which enables us to borrow information from some related GWAS when analyzing the trait of primary interest.

***Corresponding author:** Can Yang, Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, Tel:(852) 3411-7339; Fax:(852) 3411-5811; E-mail: eeyang@hkbu.edu.hk

**Received** February 13, 2015; **Accepted** May 22, 2015; **Published** May 29, 2015

**Citation:** Liu J, Wang F, Gao X, Zhang H, Wan X, et al. (2015) A Penalized Regression Approach for Integrative Analysis in Genome-Wide Association Studies. J Biomet Biostat 6: 228. doi:10.4172/2155-6180.1000228

Specifically, we propose a novel loss function that combines multiple GWAS together and use a group-Lasso penalty to integrate information from different GWAS. We further derive a gradient-based algorithm to efficiently optimize the model parameters. Based on both simulation study and real data analysis, we showed that the proposed method had advantage over single-GWAS analysis.

## Material and Method

### Model

Suppose we have q GWAS, in which we have complete data for GWAS 1, including its genotype data and phenotype data, and only have summary statistics (marginal regression coefficients) for the rest q-1 of GWAS. There are p SNPs shared by all q GWAS. Let $y_1 \in \mathbb{R}^{n_1}$ and $X_1 \in \mathbb{R}^{n_1 \times p}$ be the phenotype vector and genotype matrix of GWAS 1, respectively, where n1 is the sample size of GWAS 1. Let $z_k \in \mathbb{R}^p$ be the vector of marginal regression coefficients from k$^{th}$ GWAS, k=2, . . . , q.

Consider a linear model between the phenotype $y_1$ and genotype $X_1$ of GWAS 1,

$$y_1 = X_1 b_1 + e_1,$$

$$e_1 \sim N\left(0,\ \sigma_1^2 I\right),$$

Where $b_1 \in \mathbb{R}^p$ is the coefficient vector, $e_1 \in \mathbb{R}^{n_1}$ is the error term $\sigma_1^2$ denotes the noise variance. For the rest of GWAS with only summary statistics, we assume that $z_k \in \mathbb{R}^p$ is an estimate of the true effect size $b_k \in \mathbb{R}^p$ with noise $e_k \in \mathbb{R}^p$, i.e.,

$$z_k = b_k + e_k,$$

$$e_1 \sim N\left(0,\ \sigma_1^2 I\right),$$

where $\sigma_k^2$ denotes the noise variance for the k-th GWAS, k=2, . . . , q. Here we use a simple example to illustrate the key idea. Suppose the vectors of the true effect sizes from three GWAS are given as follows:

$$b_1 = (0.1,\ 0.4,\ 0.3,\ 0.8,\ 0,\ \ldots,\ 0)'$$

$$b_2 = (0.3,\ 0.2,\ 0.5,\ 0.6,\ 0,\ \ldots,\ 0)',$$

$$b_3 = (-0.5,\ 0.1,\ 0.7,\ 0.9,\ 0,\ \ldots,\ 0)'.$$

The joint analysis of $b_1$, $b_2$ and $b_3$ can improve the statistical power of identifying risk variants as the same loci consistently produce non-zero effect sizes among different studies. Therefore, we propose to consider all the effect sizes of the same variant as a group. For this toy example, we have $B_1 = (0.1,\ 0.3,\ -0.5)'$ as the vector of the true effect sizes of the first group, $B_1 = (0.4,\ 0.2,\ 0.1)'$ for the second group, and $B_j$ for the j-th group, j=1, .. . , p. For convenience, we denote $B = (b_1, b_2, \ldots, b_q)' \in \mathbb{R}^{q \times p}$ as the effect size matrix and $B_j \in \mathbb{R}^q$ is the j-th column.

To integrate information from multiple GWAS, we propose the following optimization problem

$$\min_{B,\sigma_1,\ldots,\sigma_k} \frac{\|y_1 - X_1 b_1\|^2}{2n_1\sigma_1} + \sum_{k=2}^{q} \frac{\|z_k - b_k\|^2}{2p\sigma_k} + \sum_{k=1}^{q} \frac{\sigma_k}{2} + \gamma \sum_{j=1}^{p} |B_j|,$$

Where γ is the regularization parameter controlling the sparsity of B, ||.|| denote the l2-norm of a vector. The proposed optimization is closely related to the scaled Lasso problem [21]. Here we emphasize on integration of information from multiple GWAS and use the group penalty to achieve this goal.

### Algorithm

Now we present our algorithm for parameter estimation in the above model. Noticing that objective function (1) is jointly convex in (B, $\sigma_1, \ldots, \sigma_k$), it is very convenient for us to use an alternating strategy in optimization. For fixed values of $\sigma_k$, k=1, . . . , q, we optimize (1) with respect to B. Then we update $\sigma_k$ (k=1, . . . , q) using the current fitted B. The details of the algorithm are given below.

Fixing $\sigma_k = \hat{\sigma}_k (k = 1, \ldots, q)$, the optimization problem becomes

$$\min_B \frac{\|y_1 - X_1 b_1\|^2}{2n_1\hat{\sigma}_1} + \sum_{k=2}^{q} \frac{\|z_k - b_k\|^2}{2p\hat{\sigma}_k} + \gamma \sum_{j=1}^{p} |B_j|. \quad (2)$$

Since $\sum_{j=1}^{p} |B_j|$ is non-differentiable, we adopt the proximal gradient method [22].

Let

$f(B) = \frac{\|y_1 - X_1 b_1\|^2}{2n_1\hat{\sigma}_1} + \sum_{k=2}^{q} \frac{\|z_k - b_k\|^2}{2p\hat{\sigma}_k}$ and $g(B) = \gamma \sum_{j=1}^{p} |B_j|$. The proximal gradient algorithm solves optimization problem (2) iteratively using the proximal operator of g(B):

$$B^{(m)} = \text{prox}_{1/\tau}(g)\left(B^{(m-1)} - \frac{1}{\tau}\nabla f(\mathrm{B}^{(m-1)})\right) = \arg\min_B \left(g(B) + \frac{\tau}{2}\left\|B - \left(B^{(m-1)} - \frac{1}{\tau}\nabla f(B^{(m-1)})\right)\right\|^2\right) \quad (3)$$

where the superscript m indicates the m$^{th}$ iteration, τ is the Lipschitz constant of f (B) and

$$\nabla f(B^{(m-1)}) = \left(\left(\frac{X_1'X_1}{n_1\hat{\sigma}_1} b_1^{(m-1)} - \frac{X_1'y_1}{n_1\hat{\sigma}_1}\right), \frac{(b_2^{(m-1)} - z_2)}{p\hat{\sigma}_2}, \ldots, \frac{(b_q^{(m-1)} - z_q)}{p\hat{\sigma}_q}\right)'.$$

Is the gradient of f (B) evaluated at B$^{(m-1)}$. Note that $B^{(m-1)} - \frac{1}{\tau}\nabla f(\mathrm{B}^{(m-1)})$ is q × p matrix which does not involve the optimization variable B. Let its $G_j^{(m)}$ denote the j-th column of $B^{(m-1)} - \frac{1}{\tau}\nabla f(\mathrm{B}^{(m-1)})$. Then optimization problem (3) can be rewritten into p separate optimization problems and be solved analytically:

$$\begin{aligned} B_j^{(m)} &= \arg\min_{B_j}\left(g(B_j) + \frac{\tau}{2}\left\|B_j - G_j^{(m)}\right\|^2\right) \\ &= \arg\min_{B_j}\left(\gamma \|B_j\| + \frac{\tau}{2}\left\|B_j - G_j^{(m)}\right\|^2\right) \\ &= \left(1 - \frac{\gamma}{\tau\|G_j^{(m)}\|}\right)_+ G_j^{(m)} n \\ &= S(G_j^{(m)}; \frac{\gamma}{\tau}) \end{aligned} \quad (4)$$

To further accelerate the convergence of the above proximal gradient algorithm, we use the accelerated proximal gradient algorithm (APG) [22], where two points {B$^{(m-1)}$, B$^{(m-2)}$} are employed to find the optimal solution for a fixed value of tuning parameter γ [23]. The detail of the APG algorithm is given in Algorithm 1.

**Algorithm 1:** Accelerated proximal gradient algorithm (APG)

Given a value of the tuning parameter γ, we first initialize Lipschitz constant $\tau = \max\left(\lambda_{max}\left(\frac{X_1'X_1}{n_1\hat{\sigma}_1}\right), \frac{1}{p\hat{\sigma}_2}, \ldots, \frac{1}{p\hat{\sigma}_q}\right)$ and $t_1$=1, where $\lambda_{max}\left(\frac{X_1'X_1}{n_1\hat{\sigma}_1}\right)$ is the maximum eigenvalue of matrix $\frac{X_1'X_1}{n_1\hat{\sigma}_1}$

for m ≥ 1 do

1. $G_j^{(m)} = \left[\widetilde{B}^{(m-1)} - \frac{1}{\tau}\nabla f(\widetilde{B}^{(m-1)})\right]_j, j = 1,\ldots,p$

2. $B_j^{(m)} = S\left(G_j^{(m)}, \frac{\gamma}{\tau}\right), for \;\; j = 1,\ldots,p$

3. $t_{(m+1)} = \frac{1+\sqrt{1+4t_{(m)}^2}}{2}$

4. $\widetilde{B}_j^{(m)} = B_j^{(m)} + \frac{t_{(m)}-1}{t_{(m+1)}}(B_j^{(m)} - B_j^{(m-1)}), \;\; for \;\; j = 1,\ldots,p$

end

Let $\hat{B} = (\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_q)'$ be the solution of Algorithm 1. Fixing B at $\hat{B}$, we can take the derivative of (1) with respect to $\sigma_k$, k=1, . . . , q and set them to zeros, yielding the following updating equations:

$$\hat{\sigma}_1 = \frac{1}{\sqrt{n_1}}||\, y_1 - X_1\hat{b}_1||, \tag{5}$$

$$\hat{\sigma}_k = \frac{1}{\sqrt{p}}||\, z_k - \hat{b}_k||, \; k = 2,\ldots,q. \tag{6}$$

Based on the above alternative optimization strategy, we now summarize the overall working algorithm in Algorithm 2.

**Algorithm 2:** Working Algorithm to Solve (1)

We first initialize $\hat{\sigma}_k^{(0)}, k = 1,\ldots,q$ using the null model.

for l ≥ 1 do

1. Using Algorithm 1 with $\hat{\sigma}_k^{(l-1)}, k = 2,\ldots,q$ to optimize (2) that results $\hat{b}_k^{(l)}, k = 1,\ldots,q$.

2. With $\hat{b}_k^{(l)}, k = 1,\ldots,q$ we can update $\sigma_k$, k=1, . . . , q using (5).

end

For the tuning parameter γ, we searched for optimal settings using a five-fold cross validation to search the best γ in $[\epsilon\gamma_{max}, \gamma_{max}]$, where E=0.05 in our experiment, and $\gamma_{max}$ is the minimum γ such that all the elements in B are estimated to be zero. A sequence of 100 γ values is generated equally in the log-space of $[\epsilon\gamma_{max}, \gamma_{max}]$. The optimal γ is chosen according to the criteria that the minimum prediction error in primary GWAS is selected.

## Simulation Study

We conducted a simulation study to evaluate the performance of the proposed method. For comparison, we also considered scaled Lasso on one GWAS with genotype data. We simulated two sets of genotype data, one for GWAS with genotype and one for summary statistics of GWAS. In the simulation study, we set $n_1$=500 with $n_2$=500 or $n_2$=2000 for the sample sizes of two GWAS, while we set the number of SNPs to be p=5000 or p=10000. We considered the auto-regressive correlation (AR) and block AR. For AR, SNP j and k have correlation coefficient $\rho^{|j-k|}$. For block AR, we set block size to be 20 equally distributed over all SNPs and the correlation coefficient for SNP j and k within a block is set to $\rho^{|j-k|}$ and 0 otherwise. We considered three scenarios with ρ=0.2, 0.5 and 0.8, corresponding to weak, moderate, and strong correlations, respectively. SNPs in the simulation study were generated with a two-stage procedure [11]. First, we drew the predictor vector xi from a p- dimensional multivariate normal distribution under different correlation structure. Then, the genotype of the ith SNP was set to be 0, 1, or 2 according to whether $x_{ij}$<-c,-c ≤ $x_{ij}$ ≤ c, or $x_{ij}$>c. The cutoff point c was the first quartile of a standard normal distribution. In this simulation, we considered that the first ten SNPs were associated with the trait in both datasets and regression coefficients were generated under normal distribution in the way that signal-to-noise ratio was controlled at 1:1 (corresponding to heritability=50%). To be specific, we first generated genotype data using the way described above. Then we normalized the genotype such that $\sum_i^n x_{ij} = 0$ and $\text{std}(x_j) = 1/\sqrt{p}$. Finally, we generated the quantitative trait using the linear model,

$y_k = X_k b_k + e_k$, k=1, . . . , q,

where $e_k$ was the error term under normal distribution with mean zero. In both correlation structures, there were ten trait-associated markers and in block AR, each of the ten blocks contained one trait-associated marker. For the second dataset, we applied the single-marker analysis and obtained the summary statistics for the integration, in which we only used this partial information without knowledge of genotype data.

In total, there were twenty-four scenarios with different combinations of correlation structures (AR and block AR), the sample size of the second GWAS $n_2$ and the total number of SNPs p for comparing the proposed method with the scaled Lasso. We used area under the curve (AUC) to show the selection performance. We also used the square of correlation coefficients ($r^2$) of observed values and predictive values, based on cross-validation. The results are shown in Figure 1. As indicated by AUC and $r^2$, the proposed method has better selection and prediction performance than scaled Lasso. The selection and prediction performance of the proposed method can further improve as the sample size of the second GWAS $n_2$ increase from 500 to 2000, which indicates the proposed pIGWAS method is able to effectively integrate additional information.

## Real Data Analysis

We applied our pIGWAS method to the quantitative trait-body mass index (BMI). We primarily used European American (EA) samples from two GWAS-Study of Addiction: Genetics and Environment (SAGE) and the Collaborative Study on the Genetics of Alco- holism (COGA). The summary statistics of height were downloaded from the web- site of Genetic Investigation of Anthropometric Traits (GIANT) consortium (http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files) [24]. After quality control, there were 656,848 SNPs with minor allele frequency (MAF) ≥ 0.01 and p-value ≥ 0.001 for Hardy-Weinberg equilibrium test in both GWAS data. For summary statistics from GIANT, we used SNPs with no missing values in MAF. Overall, there were 619,651 SNPs used in all chromosomes satisfying the pruning criteria in genotype data and existing non-missing values of MAF in summary statistics. The Manhattan plots for $\log_{10}$ p-value and $\hat{\beta}$ using conventional marginal analysis for GWAS data are given in Figure 2. Obviously, there is not a rich set of findings using SAGE and COGA EA samples for phenotype BMI (the upper panel of Figure 2) compared to the meta-analysis conducted in [25] (the lower panel of Figure 2).

First, we performed pIGWAS using combined GWAS data of SAGE and COGA with summary statistics from GIANT. Specifically, we used the EA samples containing genotype data with corresponding summary statistics implemented with the proposed method. The analysis was conducted chromosome by chromosome. We also evaluated the relative stability of the selected SNPs using random sampling [26]. Specifically, we randomly sampled 80% of the subjects and applied pIGWAS to identify associated SNPs. This process was repeated 100 times. For each SNP, we were able to calculate the proportion of times

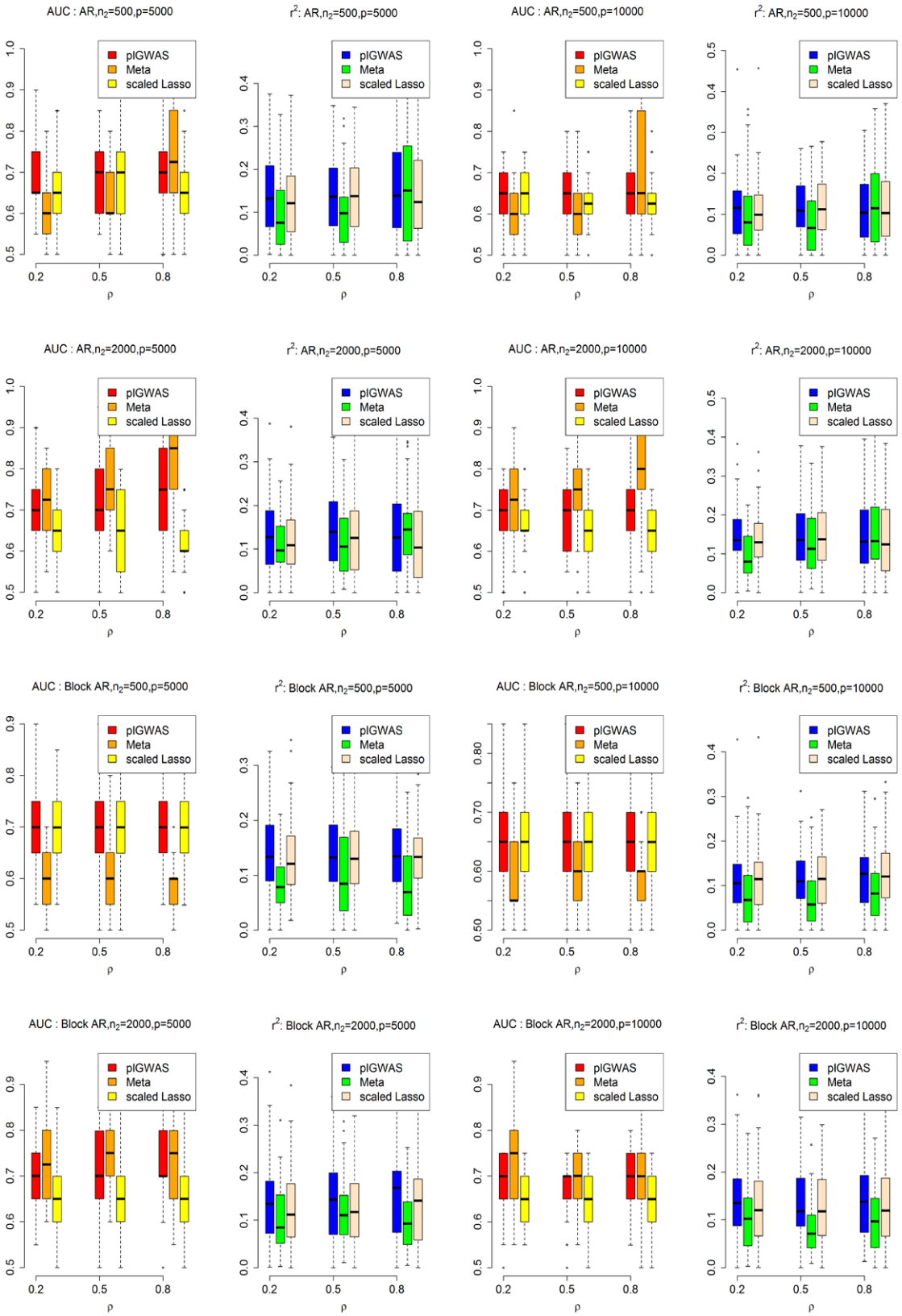


**Figure 1:** Boxplots of areas under the curve (AUC) and r2 under different combinations of $n_2$, p and correlation structure (AR, Block AR).

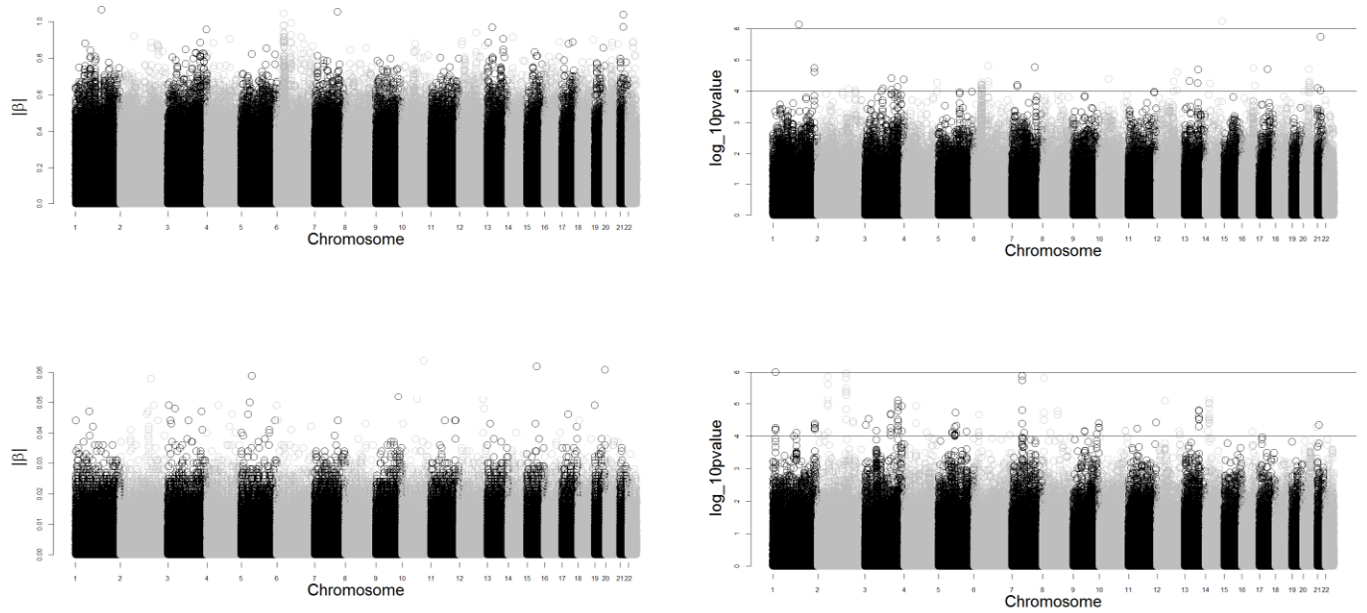


**Figure 2:** Manhattan plots of SAGE-COGA and GIANT. Upper panel: Manhattan plots of the summary statistics (coefficient β and p-value) from SAGE-COGA. Lower panel: Man- hattan plots of the summary statistics (coefficient β and p-value) from GAINT.

| SNP | Chr | Position | Gene[a] | Band | pIGWAS | | Scaled Lasso | |
|---|---|---|---|---|---|---|---|---|
| rs11588887 | 1 | 157983786 | CRP | q23.2 | 1.39 | 0.99 | 3.49 | 0.99 |
| rs1254207 | 1 | 234434850 | GPR137B | q42.3 | -0.36 | 0.84 | -0.78 | 0.75 |
| rs2477586 | 1 | 234451564 | ERO1LB | q42.3 | -0.32 | 0.81 | -0.61 | 0.73 |
| rs1860875 | 7 | 21078168 | RPL23P8 | p15.3 | | | 0.23 | 0.70 |
| rs2390470 | 7 | 21088622 | RPL23P8 | p15.3 | | | -0.06 | 0.34 |
| rs2192300 | 7 | 121194239 | PTPRZ1 | q31.32 | | | -0.14 | 0.81 |
| rs1054611 | 12 | 10061428 | CLEC12B | p13.2 | 0.10 | 0.78 | 0.09 | 0.66 |
| rs4565970 | 12 | 81715248 | TMTC2 | q21.31 | 0.03 | 0.56 | 0.01 | 0.52 |
| rs12370680 | 12 | 83436479 | MIR548T | q21.31 | 0.25 | 0.87 | 0.35 | 0.83 |
| rs4393415 | 12 | 102464392 | STAB2 | q23.3 | 0.10 | 0.72 | 0.13 | 0.60 |
| rs4405407 | 12 | 102466587 | STAB2 | q23.3 | 0.04 | 0.60 | 0.03 | 0.40 |
| rs1336850 | 13 | 22680577 | SGCG | q12.12 | | | -1.19 | 0.86 |
| rs622227 | 13 | 27937214 | FLT1 | q12.3 | | | 1.00 | 0.78 |
| rs1058214 | 13 | 38885772 | LHFP | q13.3 | | | 1.32 | 0.80 |
| rs7323630 | 13 | 39747343 | LINC00548 | q14.11 | | | -0.16 | 0.59 |
| rs4473069 | 13 | 43009971 | ENOX1 | q14.11 | | | -0.03 | 0.40 |
| rs2786712 | 13 | 44192569 | LINC00330 | q14.11 | | | -0.18 | 0.56 |
| rs1330476 | 13 | 81854308 | SLITRK1 | q31.1 | -0.04 | 0.61 | | |
| rs9531358 | 13 | 81964898 | SLITRK1 | q31.1 | -0.14 | 0.89 | -3.55 | 1.00 |
| rs2777825 | 13 | 83099526 | SLITRK1 | q31.1 | | | -1.20 | 0.83 |
| rs9531489 | 13 | 83152986 | SLITRK1 | q31.1 | | | -0.57 | 0.62 |
| rs9546479 | 13 | 83154395 | SLITRK1 | q31.1 | | | -0.03 | 0.33 |
| rs9319013 | 13 | 83522913 | MIR548F1 | q31.1 | | | -0.29 | 0.55 |
| rs723576 | 13 | 95002092 | CLDN10 | q32.1 | | | -0.11 | 0.47 |
| rs1547166 | 13 | 95071328 | DZIP1 | q32.1 | | | -0.37 | 0.46 |
| rs7338545 | 13 | 95073552 | DZIP1 | q32.1 | | | -0.16 | 0.39 |
| rs8018440 | 14 | 32981820 | NPAS3 | q13.1 | 0.25 | 0.71 | 0.03 | 0.52 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| rs4903707 | 14 | 39909619 | FBXO33 | q21.1 | 0.57 | 0.78 | 0.38 | 0.73 |
| rs9944120 | 14 | 40082198 | FBXO33 | q21.1 | 0.04 | 0.51 | | |
| rs7149526 | 14 | 80098349 | CEP128 | q31.1 | 0.00 | 0.47 | | |
| rs1951980 | 14 | 95322780 | TCL1A | q32.13 | -1.47 | 1.00 | -1.54 | 0.97 |
| rs1345300 | 16 | 8671929 | ABAT | p13.2 | -0.10 | 0.68 | | |
| rs2283557 | 16 | 24273067 | CACNG3 | p12.1 | -0.01 | 0.63 | | |
| rs4784651 | 16 | 54869275 | GO1 | q13 | 0.15 | 0.42 | | |
| rs9922112 | 16 | 54872188 | GO1 | q13 | 0.10 | 0.39 | | |
| rs2587878 | 16 | 54872860 | GO1 | q13 | -0.16 | 0.46 | -0.02 | 0.48 |
| rs8047093 | 16 | 59619195 | CDH8 | q21 | 0.09 | 0.64 | | |
| rs8044561 | 16 | 70108642 | CHST4 | q22.3 | 0.18 | 0.28 | | |
| rs310334 | 16 | 70131048 | CHST4 0.11769305 | q22.3 | 0.12 | 0.29 | | |
| rs2432524 | 16 | 70141834 | CHST4 | q22.3 | 1.36 | 0.92 | 1.66 | 0.97 |
| rs8056272 | 16 | 72267976 | LOC100506172 | q22.3 | 0.04 | 0.50 | | |
| rs12928065 | 16 | 77094975 | WWOX | q23.1 | -0.13 | 0.61 | -0.02 | 0.52 |
| rs933374 | 17 | 13742206 | CDRT15P1 | p12 | 0.04 | 0.69 | 0.16 | 0.65 |
| rs9889937 | 17 | 18512091 | FOXO3B | p11.2 | 0.00 | 0.58 | 0.01 | 0.58 |
| rs4792855 | 17 | 40815480 | ARHGAP27 | q21.31 | -0.18 | 0.87 | -0.48 | 0.87 |
| rs17673185 | 17 | 48822712 | MIR548AJ2 | q22 | -0.02 | 0.57 | -0.12 | 0.55 |
| rs7265169 | 20 | 312747 | TRIB3 | p13 | 0.09 | 0.63 | | |
| rs459012 | 20 | 410008 | CSNK2A1 | p13 | 0.25 | 0.59 | | |
| rs6053384 | 20 | 5354093 | LINC00658 | p12.3 | 0.83 | 0.82 | 0.75 | 0.80 |
| rs1555669 | 20 | 12598312 | SPTLC3 | p12.1 | 0.01 | 0.41 | | |
| rs6074541 | 20 | 12926517 | SPTLC3 | p12.1 | -0.01 | 0.46 | | |
| rs6081333 | 20 | 18660916 | DTD1 | p11.23 | -0.38 | 0.72 | -0.16 | 0.57 |
| rs2067845 | 20 | 19446645 | SLC24A3 | p11.23 | 0.88 | 0.91 | 0.83 | 0.81 |
| rs6035387 | 20 | 19524045 | SLC24A3 | p11.23 | 0.22 | 0.46 | 0.03 | 0.42 |
| rs6515030 | 20 | 19529688 | SLC24A3 | p11.23 | 0.39 | 0.67 | 0.23 | 0.51 |
| rs942990 | 20 | 19533661 | SLC24A3 | p11.23 | 0.07 | 0.32 | | |
| rs199575 | 20 | 19902601 | RIN2 | p11.23 | -0.11 | 0.73 | | |
| rs56916 | 20 | 19936892 | RIN2 | p11.23 | 1.15 | 0.90 | 1.26 | 0.82 |
| rs199572 | 20 | 19940313 | RIN2 | p11.23 | 0.17 | 0.42 | | |
| rs200175 | 20 | 19949483 | A20 | p11.23 | 0.45 | 0.54 | 0.21 | 0.35 |
| rs6050359 | 20 | 25070945 | LOC284798 | p11.21 | 0.33 | 0.50 | 0.10 | 0.41 |
| rs6050372 | 20 | 25081225 | LOC284798 | p11.21 | 0.77 | 0.81 | 0.84 | 0.88 |
| rs6050418 | 20 | 25118643 | LOC284798 | p11.21 | 0.48 | 0.62 | 0.36 | 0.63 |
| rs3787076 | 20 | 25143018 | ENTPD6 | p11.21 | 0.37 | 0.68 | 0.07 | 0.57 |
| rs2073077 | 20 | 25143913 | ENTPD6 | p11.21 | 0.13 | 0.54 | | |
| rs6022419 | 20 | 36083479 | TTI1 | q11.23 | -0.08 | 0.67 | | |
| rs6030352 | 20 | 40658434 | PTPRT | q12 | 0.66 | 0.87 | 0.52 | 0.71 |
| rs1010310 | 20 | 44268451 | CDH22 | q13.12 | -0.02 | 0.48 | | |
| rs846743 | 20 | 48768050 | PARD6B | q13.13 | -0.13 | 0.59 | | |
| rs6021702 | 20 | 50145309 | ZFP64 | q13.2 | -0.13 | 0.64 | | |
| rs7268780 | 20 | 56735571 | STX16-NPEPL1 | q13.32 | 0.03 | 0.65 | | |
| rs2823209 | 21 | 15586648 | NRIP1 | q21.1 | 0.24 | 0.66 | 0.22 | 0.54 |
| rs2823216 | 21 | 15591805 | NRIP1 | q21.1 | 0.34 | 0.74 | 0.30 | 0.60 |
| rs463370 | 21 | 30177240 | GRIK1 | q21.3 | 1.40 | 0.95 | 1.51 | 1.00 |

[a] Gene names that SNPs belong to or are closest to.

**Table 1:** SNPs selected incorporating summary statistics from public available source by using pIGWAS and SNPs selected using scaled Lasso for GWAS with genotype.

that the SNP was associated with the trait out of 100 samplings, i.e., the observed occurrence index (OOI). For comparison, we conducted single data analysis of GWAS with EA samples using scaled Lasso, and evaluated its relative stability of the selected markers using the OOI. The associated markers identified by integrative and single data analysis were listed in Table 1. The average OOI of SNPs selected by pIGWAS is 0.649 while that of SNPs selected by scaled Lasso is 0.648, suggesting a limited improvement of pIGWAS over scaled Lasso on the COGA-SAGE data set. There may be two reasons for this. First, the GWAS signals of BMI from the COGA-SAGE may be too weak to be distinguished from noise (Figure 1). Second, the pleiotropic effects between BMI and height may not be strong enough to boost the power of pIGWAS. It is expected that pIGWAS could achieve a better performance in presence of well-powered GWAS signals and pleiotropy information.

## Conclusion

GWAS suffer from low statistical power due to the individual weak effects of genetic variants. In this study, we proposed a statistical approach to jointly analyzing primary GWAS data with summary statistics together from other source. The key idea of the proposed approach lies on the existence of pleiotropic effects of genetic variants, which allows us to borrow information from genetically related GWAS. Specifically, we proposed a novel penalized regression that combines multiple GWAS together. The computationally efficient algorithm is derived for optimizing the model parameters. Based on both simulation study and real data analysis, we demonstrated the advantages of the proposed method over single-GWAS analysis.

### Acknowledgment

This study was supported by National Institutes of Health grants R01EY022651, Hong Kong Research grant HKBU12202114, Hong Kong Baptist University FRG2/13-14/005 and Duke- NUS Graduate Medical School WBS: R-913-200-098-263.

### References

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.
2. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-266.
3. Allen HL, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.
4. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90: 7-24.
5. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565-569.
6. Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature genetics 45: 984-994.
7. Yang C, Li C, Kranzler HR, Farrer LA, Zhao H, et al. (2014) Exploring the genetic architecture of alcohol dependence in African-Americans via analysis of a genomewide set of common variants. Hum Genet 133: 617-624.
8. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44: 981-990.
9. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781-791.
10. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681-690.
11. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714-721.
12. Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2375-2382.
13. Yang C, Wan X, Yang Q, Xue H, Yu W (2010) Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. BMC Bioinformatics 11 Suppl 1: S18.
14. Liu J, Huang J, Ma S, Wang K (2013) Incorporating group correlations in genome-wide association studies using smoothed group Lasso. Biostatistics 14: 205-219.
15. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet 14: 483-495.
16. Li C, Yang C, Gelernter J, Zhao H (2014) Improving genetic risk prediction by leveraging pleiotropy. Hum Genet 133: 639-650.
17. Chung D, Yang C, Li C, Gelernter J, Zhao H (2014) GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet 10: e1004787.
18. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, et al. (2011) Abundant pleiotropy in human complex diseases and traits. Am J Hum Genet 89: 607-618.
19. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR (2012) Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics 28: 2540-2542.
20. Andreassen OA, Djurovic S, Thompson WK, Schork AJ, Kendler KS, et al. (2013) Improved detection of common variants associated with schizophrenia by leveraging pleiotropy with cardiovascular-disease risk factors. The American Journal of Human Genetics 92: 197-209.
21. Sun T, Zhang CH (2012) Scaled sparse linear regression. Biometrika 99: 879-898.
22. Beck A, Teboulle M (2009) Gradient-based algorithms with applications to signal recovery. Convex Optimization in Signal Processing and Communications.
23. Edenberg HJ (2002) The collaborative study on the genetics of alcoholism: an update. Alcohol Res Health 26: 214-218.
24. Yang J, Loos RJ, Powell JE, Medland SE, Speliotes EK, et al. (2012) FTO genotype is associated with phenotypic variability of body mass index. Nature 490: 267-272.
25. Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, et al. (2013) Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. Nature genetics 45: 501-512.
26. Huang J, Ma S (2010) Variable selection in the accelerated failure time model via the bridge method. Lifetime Data Anal 16: 176-195.